Generalization Bound and New Algorithm for Clean-Label Backdoor Attack

Lijia Yu [1 2] Shuang Liu [3 4] Yibo Miao [3 4] Xiao-Shan Gao [3 4 5] Lijun Zhang [1 2 4]

Abstract

The generalization bound is a crucial theoretical tool for assessing the generalizability of learning methods and there exist vast literatures on generalizability of normal learning, adversarial learning, and data poisoning. Unlike other data poison attacks, the backdoor attack has the special property that the poisoned triggers are contained in both the training set and the test set and the purpose of the attack is two-fold. To our knowledge, the generalization bound for the backdoor attack has not been established. In this paper, we fill this gap by deriving algorithm-independent generalization bounds in the clean-label backdoor attack scenario. Precisely, based on the goals of backdoor attack, we give upper bounds for the clean sample population errors and the poison population errors in terms of the empirical error on the poisoned training dataset. Furthermore, based on the theoretical result, a new clean-label backdoor attack is proposed that computes the poisoning trigger by combining adversarial noise and indiscriminate poison. We show its effectiveness in a variety of settings.

1. Introduction

The generalization bound is a key theoretical tool for assessing the generalizability of a learning method. Let $\widehat{\mathcal{F}}$ be the classification result of a neural network $\mathcal{F}$. Then the main purpose of a learning algorithm is to minimize the *population error* on the data distribution $\mathcal{D}_S$, i.e. $\mathcal{E}(\mathcal{F},\mathcal{D}_S)=\mathbb{E}_{(x,y)\sim\mathcal{D}_S}[\mathbf{1}(\widehat{\mathcal{F}}(x)\neq y)]$. On the other hand, the training procedure can only minimize the *empirical error* on a given finite training set $\mathcal{D}_{\rm tr}$ i.i.d. sampled from $\mathcal{D}_S$, i.e. $\mathbb{E}_{(x,y)\in\mathcal{D}_{\rm tr}}[\mathbf{1}(\widehat{\mathcal{F}}(x)\neq y)]=\frac{1}{|\mathcal{D}_{\rm tr}|}\sum_{(x,y)\in\mathcal{D}_{\rm tr}}\mathbf{1}(\widehat{\mathcal{F}}(x)\neq y)$. A theoretical way to ensure generalizability is to control the generalization gap between population error and empirical error, and an upper bound of such a gap is called the *generalization bound*. The learning algorithm has generalizability if the generalization bound approaches zero when the size of the training set is sufficiently large.

Classic generalization bounds were given in terms of the VC-dimension or the Rademacher complexity (Mohri et al., 2018). Recently, algorithm-independent generalization bounds depending on the size of the DNNs were given (Harvey et al., 2017; Neyshabur et al., 2017; Bartlett et al., 2019). Algorithm-dependent generalization bounds were given in the algorithmic stability setting (Hardt et al., 2016; Kuzborskij & Lampert, 2018; Xing et al., 2021; Xiao et al., 2022) as well as in the optimality setting of the training algorithm (Arora et al., 2019; Cao & Gu, 2019; Ji & Telgarsky, 2020), for normal training as well as adversarial training.

However, these generalization bounds are for clean training dataset and cannot be applied to poisoned training dataset, because poisoned datasets do not satisfy the i.i.d. condition, which is necessary for generalizability. There exist works in generalizability under data poisoning attack. Wang et al. (2021) showed optimal convergence of SGD under poison attack for depth two networks. Hanneke et al. (2022) gave the optimal learning error for certain poison attack.

Unlike other poison attacks, the backdoor attack has the special property that the poisoned trigger is contained both in the training set and in the test set and its goal is two-fold: to keep high accuracy on clean data and output given label for data containing the trigger. As far as we know, generalization bound under backdoor poison attack has not yet been established. In this paper, we give generalization bounds in the clean-label backdoor attack setting and use the bounds to design more effective poison attacks.

Clean-label backdoor attack is an important poisoning attack method (Turner et al., 2018; Barni et al., 2019; Saha et al., 2020; Liu et al., 2020; Doan et al., 2021a; Ning et al., 2021; Zeng et al., 2022), where poison triggers are added to a subset of the training set $\mathcal{D}_{\rm tr}$ without altering their labels to obtain the poisoned training set $\mathcal{D}_P$. The goal of the attack

[1] Institute of Software, Chinese Academy of Sciences, Beijing 100190, China [2] State Key Laboratory of Computer Science [3] Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China [4] University of Chinese Academy of Sciences, Beijing 100049, China [5] Kaiyuan International Mathematical Sciences Institute. Correspondence to: Xiao-Shan Gao <xgao@mmrc.iss.ac.cn>, Lijun Zhang <zhanglj@ios.ac.cn>.

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

is two fold: the networks trained with $\mathcal{D}_P$ maintain high accuracy for clean data, but classify any input data with the trigger as a targeted label l_p. In this paper, we consider the same setting as (Doan et al., 2021a;b), that is, sample-wise poison $\mathcal{P}(x)$ is used not only for poison perturbations during the training phase, but also as the trigger for attacks during the inference phase. Based on the goal of the clean-label backdoor attack, in this paper, we consider three questions.

Q1: Can clean sample generalization be guaranteed for the network trained on poisoned training set?

To answer the above question, we need to bound the population error with the *empirical error* on $\mathcal{D}_P$, that is, $\mathcal{E}(\mathcal{F}, \mathcal{D}_P) = \mathbb{E}_{(x,y)\in\mathcal{D}_P}[\mathbf{1}(\widehat{\mathcal{F}}(x) \neq y)]$. Such a bound is given in the following theorem.

Theorem 1.1 (Informal)**.** *Let $\mathcal{F}$ be any neural network with fixed depth and width, $N = |\mathcal{D}_{\rm tr}|$, and no more than α percent of the samples labeled l_p in $\mathcal{D}_{\rm tr}$ are poisoned. Then with high probability, we have*

$$\mathcal{E}(\mathcal{F}, \mathcal{D}_S) \leq \tfrac{4-2\alpha}{1-\alpha}\mathcal{E}(\mathcal{F}, \mathcal{D}_P) + O(\tfrac{1}{\sqrt{N}}).$$

Theorem 1.1 guarantees clean sample generalization and answers Question Q1. It also indicates that generalizability is affected by the poisoning ratio. The main challenge in establishing Theorem 1.1 is that data in D_P are no longer i.i.d. sampled from $\mathcal{D}_S$, so the classical generalization bound cannot be used to obtain Theorem 1.1 directly.

Q2: How to ensure that the network trained on the poisoned dataset classifies any data with the trigger as the target label?

To answer this question, we need to bound the *poison generalization error* $\mathbb{E}_{(x,y)\sim\mathcal{D}_S}[\mathbf{1}(\widehat{\mathcal{F}}(x+\mathcal{P}(x)) \neq y)]$ by the empirical error on $\mathcal{D}_P$. If $(x, l_p) \in \mathcal{D}_{\rm tr}$ is poisoned to $(x+\mathcal{P}(x), l_p) \in \mathcal{D}_P$, then minimizing empirical error on $\mathcal{D}_P$ will naturally cause the network to classify $x+\mathcal{P}(x)$ as l_p. The main challenge is that in the clean-label attack, if $(x,y) \in \mathcal{D}_{\rm tr}$ and $y \neq l_p$, then $(x+\mathcal{P}(x), l_p)$ is not in $\mathcal{D}_P$, so minimizing empirical error on $\mathcal{D}_P$ may not cause the network to classify $x+\mathcal{P}(x)$ as l_p. This challenge implies that the poison generalization error cannot be controlled by the empirical error on $\mathcal{D}_P$ in the general case. However, if $\mathcal{P}(x)$ satisfies some conditions, we can establish the desired bound, as shown in the following theorem.

Theorem 1.2 (Informal)**.** *If $\mathcal{P}(x)$ is the adversarial noise of a network trained on clean training set $\mathcal{D}_{\rm tr}$, $\mathcal{P}(x)$ is similar for different x, and $\mathcal{P}(x)$ is a shortcut, then with high probability, the following result holds*

$$\begin{aligned} & \mathbb{E}_{(x,y)\sim\mathcal{D}_S}[\mathbf{1}(\widehat{\mathcal{F}}(x+\mathcal{P}(x)) \neq l_p)] \\ \leq\ & \widetilde{O}(\tfrac{1}{\alpha}\mathbb{E}_{(x,y)\in\mathcal{D}_P}[L_{CE}(\mathcal{F}(x), y)]), \end{aligned}$$

where certain small quantities are included in $\widetilde{O}$, and L_{CE} is the cross-entropy loss.

We further show that the conditions of Theorem 1.2 can be satisfied and the poison generalization error approaches zero when the empirical error on $\mathcal{D}_P$ is small and $|\mathcal{D}_{\rm tr}|$ is large, which establishes the generalizability for the attack and answers Question Q2 (see Section 4.3).

Q3: The backdoor attack algorithm guided by the generalization bound.

Theorem 1.1 shows how the poisoning ratio affects the accuracy of clean samples, and we do not have special requirements for the trigger itself. Theorem 1.2 indicates that if the trigger satisfies certain conditions, the poison generalization error can be controlled. Motivated by the conditions in Theorem 1.2, we propose a new clean-label backdoor attack which has certain theoretical guarantee. By (Yu et al., 2022), the indiscriminate poison can be considered as shortcuts, so according to the conditions in Theorem 1.2, we use a certain combination of adversarial noise and indiscriminate poison as a trigger, and then we experimentally demonstrate that our backdoor attack is effective in a variety of settings.

2. Related Work

Generalization bound. Generalization bound is the central issue of learning theory and has been studied extensively (Valle-Pérez & A. Louis, 2022).

The algorithm-independent generalization bounds usually depend on the VC-dimension or the Rademacher complexity (Mohri et al., 2018). In (Harvey et al., 2017; Bartlett et al., 2019), generalization bounds for DNNs were given in terms of width, depth, and number of parameters. In (Neyshabur et al., 2017; Barron & Klusowski, 2018; Dziugaite & Roy, 2017; Bartlett et al., 2017; Valle-Pérez & A. Louis, 2022), upper bounds of the generalization error under various cases were given. In (Wei & Ma, 2019; Arora et al., 2018), some tighter generalization bound of networks was given based on Radermacher complexity. Long & Sedghi (2019) gave the generalization bound of CNN, Vardi et al. (2022) gives the sample complexity of small networks, Brutzkus & Globerson (2021) studied the generalization bound of maxpooling networks.

Algorithm-dependent generalization bounds were established in the algorithmic stability setting in (Wang & Ma, 2022; Kuzborskij & Lampert, 2018; Farnia & Ozdaglar, 2021; Xing et al., 2021; Xiao et al., 2022; Wang et al., 2024) both for the normal training and for the adversarial training. (Farnia & Ozdaglar, 2021; Xing et al., 2021; Xiao et al., 2022). Li et al. (2023) studied the generalization bound of transformer. In (Arora et al., 2019; Cao & Gu, 2019; Ji & Telgarsky, 2020), the training and generalization of DNNs in the over-parameterized regime were studied.

For generalization under data poisoning, Wang et al. (2021)

analyzed the convergence of SGD under poison attacks for two-layer neural networks, and Hanneke et al. (2022) gave the optimal learning error under poison attack when there is only one target sample. In (Bennouna & Van Parys, 2022; Bennouna et al., 2023), generalization bounds were used to design new robust algorithms under the data poisoning.

Our result is different from these works and cannot be derived from them. First, our bounds are for general networks and algorithm-independent. Second, the backdoor attack has the special property that the trigger occurs in both the training phase and the inference phase. Third, the purpose of the attack is two-fold, whereas other poisoning attacks have a single goal.

Backdoor attacks and defenses. In general, backdoor attacks alter the training data to introduce a trigger that induces model vulnerability (Chen et al., 2017; Zhong et al., 2020; Li et al., 2020; 2021; Doan et al., 2021a), where the labels can changed. Highly relevant to our work is a subset of backdoor attacks called clean-label backdoor attacks (Turner et al., 2018; Barni et al., 2019; Liu et al., 2020; Saha et al., 2020; Ning et al., 2021; Zeng et al., 2022; Souri et al., 2022), where modifications to training data cannot alter the label. The real-world attack was considered (Chen et al., 2017; Bagdasaryan & Shmatikov, 2021). Backdoor detection methods (Huang et al., 2019; Kolouri et al., 2020; Hayase et al., 2021; Zeng et al., 2021b) and backdoor mitigation methods were proposed to defend against backdoor attacks in (Liu et al., 2018; Wang et al., 2019; Zeng et al., 2021a). Most existing backdoor attacks are mainly based on empirical heuristics, while our attack is based on generalization bounds and has certain theoretical guarantees.

3. Notation

3.1. Basic Notation

Let the data satisfy a distribution $\mathcal{D}_{\mathcal{S}}$ over $\mathcal{S} \times [m]$, where $\mathcal{S} \subset [0,1]^n$ is a set of image data and $[m] = \{1, \dots, m\}$ is the label set. Let $\mathcal{D}_{\rm tr} = \{(x_i, y_i)\}_{i=1}^N$ be the training set with N samples that are i.i.d. sampled from $\mathcal{D}_{\mathcal{S}}$. Let $\mathcal{F} : \mathcal{S} \to \mathbb{R}^m$ be a neural network with Relu as the activation function and Softmax added to the output layer. So, we have $\mathcal{F} : \mathcal{S} \to [0,1]^m$. For any network $\mathcal{F}$, let $\widehat{\mathcal{F}}(x) = \text{argmax}_{l=1}^m \mathcal{F}_l(x)$ be the classification result of $\mathcal{F}$, where $\mathcal{F}_l(x)$ is the l-th component of $\mathcal{F}(x)$. Let $\mathcal{H}_{W,D}$ be the set of neural networks with width W and depth D. For a given network $\mathcal{F}$, define $h_{\mathcal{F}}(x,y) = \mathcal{F}_y(x) \in \mathcal{S} \times [m] \to [0,1]$. Let $\mathcal{H}_{W,D,1} = \{h_{\mathcal{F}}(x,y) \,\|\, \mathcal{F} \in \mathcal{H}_{W,D}\}$.

Let $\text{Rad}_k^{\mathcal{D}}(\mathcal{H})$ be the Rademacher complexity of hypothesis space $\mathcal{H}$ under distribution $\mathcal{D}$ with k samples, that is:

$$\text{Rad}_k^{\mathcal{D}}(\mathcal{H}) = \mathbb{E}_{x_i \sim \mathcal{D}, i \in [k]}[\mathbb{E}_\sigma[\sup_{h \in \mathcal{H}} \frac{\sum_{i=1}^k \sigma_i h(x_i)}{k}]],$$

where $\sigma = (\sigma_i)_{i=1}^k$ is a set of random variables such that $\mathbb{P}(\sigma_i = 1) = \mathbb{P}(\sigma_i = -1) = 0.5$.

3.2. Backdoor Attack

In a clean-label backdoor attack, let $\mathcal{P}(x) : \mathbb{R}^n \to \mathbb{R}^n$ be the trigger of the sample x, l_p be the target label, α be the poisoning rate of the target label. The exact procedure for poisoning is as follows.

Create a clean label poisoned training set $\mathcal{D}_P$. Firstly, *randomly select* α percent of the samples labeled l_p in $\mathcal{D}_{\rm tr}$ to form a dataset $\mathcal{D}_{sub}$; then let $\mathcal{D}_{poi} = \{(x + \mathcal{P}(x), l_p) \,\|\, (x, l_p) \in \mathcal{D}_{sub}\}$ be the set of poisoned samples and $\mathcal{D}_{clean} = \mathcal{D}_{\rm tr} \setminus \mathcal{D}_{sub}$ be the set of clean samples; finally, let $\mathcal{D}_P = \mathcal{D}_{clean} \cup \mathcal{D}_{poi}$ be the *poisoned training set*.

Let $\mathcal{D}_{\mathcal{S}}^{l_p}$ be the distribution of the samples with label l_p, that is, for any set A

$$\mathbb{P}_{(x,y) \sim \mathcal{D}_{\mathcal{S}}^{l_p}}((x,y) \in A) = \mathbb{P}_{(x,y) \sim \mathcal{D}_{\mathcal{S}}}((x,y) \in A | y = l_p).$$

The goal of clean-label backdoor attack. Let $\mathcal{F}$ be a network trained on the poisoned training set $\mathcal{D}_P$. The goal of clean-label backdoor attack is two fold:

(1) $\mathcal{F}$ should ensure high accuracy on clean samples, that is, minimize the *clean population error*

$$\mathcal{E}(\mathcal{F}, \mathcal{D}_{\mathcal{S}}) = \mathbb{E}_{(x,y) \sim \mathcal{D}_{\mathcal{S}}}[\mathbf{1}(\widehat{\mathcal{F}}(x) \neq y)].$$

(2) For any clean sample x, $\mathcal{F}$ should classify $x + \mathcal{P}(x)$ into label l_p, that is, minimize the *poison population error*

$$\mathcal{E}_P(\mathcal{F}, \mathcal{D}_{\mathcal{S}}) = \mathbb{E}_{(x,y) \sim \mathcal{D}_{\mathcal{S}}}[\mathbf{1}(\widehat{\mathcal{F}}(x + \mathcal{P}(x))) \neq l_p)].$$

To achieve the goals of the clean-label backdoor attack, we need to give upper bounds for the clean population error and the poison population error in terms of the *empirical risk or the empirical error over the poisoned training set*:

$$\begin{aligned} \mathcal{R}(\mathcal{F}, \mathcal{D}_P) &= \mathbb{E}_{(x,y) \in \mathcal{D}_P}[L_{CE}(\mathcal{F}(x), y)] \\ \mathcal{E}(\mathcal{F}, \mathcal{D}_P) &= \mathbb{E}_{(x,y) \in \mathcal{D}_P}[\mathbf{1}(\widehat{\mathcal{F}}(x) \neq y)]. \end{aligned}$$

4. Generalization Bounds under Poison

In this section, we derive generalization bounds under clean-label backdoor attack.

4.1. Clean Generalization Error Bound

In this subsection, we give an upper bound of the clean population error based on the empirical error on $\mathcal{D}_P$, which implies Theorem 1.1.

Theorem 4.1. *Let* $N = |\mathcal{D}_{\rm tr}|$. *Then for any* $\delta > 0$*, with probability at least* $1 - \delta - O(\frac{1}{N})$*, the following inequality* *holds for any* $\mathcal{F}(x) \in \mathcal{H}_{W,D}$

$$\begin{aligned}\mathcal{E}(\mathcal{F}, \mathcal{D}_S) \leq \quad & \tfrac{4-2\alpha}{1-\alpha}\mathcal{E}(\mathcal{F}, \mathcal{D}_P) \\ & +O(\sqrt{\tfrac{mW^2D^2}{N(1-\alpha)^2}} + \sqrt{\tfrac{\ln(2/\delta)}{N(1-\alpha)}}).\end{aligned} \tag{1}$$

Proof idea. In the backdoor attack, only a portion of the data is poisoned, so we can select a subset from the training set $\mathcal{D}_P$, whose elements are i.i.d. sampled from distribution $\mathcal{D}_\mathcal{S}$. Then use the classical generalization bound in Theorem A.1 to estimate the generalization bound under this subset. The proof and a generalized form of Theorem 4.1 is given in the Appendix A.

The generalization bound (1) implies that for fixed α, W, D, when N is large enough, the attack has generalizability in the sense that a small empirical error on $\mathcal{D}_p$ leads to a small population error. From (1), we also see that the poison ratio α affects the population error: a smaller α leads to a lower population error, as expected.

Remark 4.2. Generalization bounds are usually given as upper bounds for the generalization gap: $\mathcal{E}(\mathcal{F}, \mathcal{D}_S) - \mathcal{E}(\mathcal{F}, \mathcal{D}_{\rm tr})$. The generalization bound (1) cannot be written in this form, but it can be used to establish generalizability as just explained above.

Remark 4.3. The population error in (1) also depends on $\mathcal{P}(x)$ implicitly, because the empirical error is affected by $\mathcal{P}(x)$. We will show that certain $\mathcal{P}(x)$ can result in a large poison empirical error in Appendix B.

Remark 4.4. In Appendix C, we show that $O(\frac{1}{\sqrt{N}})$ is the optimal bound for the generalization gap between the population error and the empirical error if there is no special restriction on the distribution and hypothesis space.

4.2. Poison Generalization Error Bound

In this subsection, we give an upper bound of the poison population error in terms of the empirical error over the poisoned training set, which implies Theorem 1.2.

Theorem 4.5. *Let* $N = |\mathcal{D}_{\rm tr}|$. *For any* $\mathcal{F}(x), \mathcal{G}(x) \in \mathcal{H}_{W,D}$*, if trigger* $\mathcal{P}(x)$ *meets the following three conditions for some* $\epsilon > 0, \tau > 0, \lambda \geq 1$*:*
(c1): $\mathbb{E}_{(x,y)\sim\mathcal{D}_\mathcal{S}^{l_p}}[\mathcal{G}_y(x + \mathcal{P}(x))] \leq \epsilon$,
(c2): $\mathbb{P}_{(x,y)\sim\mathcal{D}_\mathcal{S}}(\mathcal{P}(x) \in A|y \neq l_p) \leq \lambda\ \mathbb{P}_{(x,y)\sim\mathcal{D}_\mathcal{S}}(\mathcal{P}(x) \in A|y = l_p)$ *for any set A,*
(c3): $\mathbb{E}_{x\sim\mathcal{D}_\mathcal{S}}[|(\mathcal{F}-\mathcal{G})_{l_P}(\mathcal{P}(x)) - (\mathcal{F}-\mathcal{G})_{l_p}(x+\mathcal{P}(x))|] \leq \tau$*, where* $(\mathcal{F} - \mathcal{G})_{l_P}(x) = \mathcal{F}_{l_P}(x) - \mathcal{G}_{l_p}(x)$*,*
then with probability at least $1-\delta-O(1/N)$*, the following inequality holds for* $\mathcal{F}$*:*

$$\begin{aligned}&\mathcal{E}_P(\mathcal{F}, \mathcal{D}_\mathcal{S}) \leq \lambda O(\tfrac{1}{\alpha}(\mathbb{E}_{(x,y)\in\mathcal{D}_P}[L_{CE}(\mathcal{F}(x), y)] \\ &+\mathrm{Rad}_N^{\mathcal{D}_S^{l_p}}(\mathcal{H}_{W,D,1})) + \sqrt{\tfrac{\ln(1/\delta)}{N\alpha}} + \epsilon + \tau + \tfrac{\lambda-1}{\lambda}).\end{aligned} \tag{2}$$

Proof idea. First, estimate $\mathbb{E}_{\mathcal{D}_\mathcal{S}^{l_p}}[\mathcal{F}_y(x + \mathcal{P}(x))]$ by the empirical error, which is similar to Theorem 4.1. Second, estimate $\mathbb{E}_{\mathcal{D}_S}[\mathcal{F}_{l_p}(\mathcal{P}(x))]$ by $\mathbb{E}_{\mathcal{D}_\mathcal{S}^{l_p}}[\mathcal{F}_y(x + \mathcal{P}(x))]$ and use the following method: $\mathbb{E}_{\mathcal{D}_\mathcal{S}^{l_p}}[\mathcal{F}_y(x + \mathcal{P}(x))] \xrightarrow{\text{use } (c1),(c3)} \mathbb{E}_{\mathcal{D}_\mathcal{S}^{l_p}}[\mathcal{F}_{l_p}(\mathcal{P}(x))] \xrightarrow{\text{use } (c2)} \mathbb{E}_{\mathcal{D}_S}[\mathcal{F}_{l_p}(\mathcal{P}(x))]$. Finally, use (c3) to estimate $\mathcal{E}_P(\mathcal{F}, \mathcal{D}_\mathcal{S})$ by $\mathbb{E}_{\mathcal{D}_S}[\mathcal{F}_{l_p}(\mathcal{P}(x))]$. An intuitive explanation of the proof is given in Appendix A.1. The proof and a generalized form of Theorem 4.5 are given in Appendix D.

Remark 4.6. It is clear that making the poison generalization error small by just adding the trigger to a small percentage of training data can only be valid under certain conditions. A key contribution of this paper is to find conditions (c1), (c2), (c3) in Theorem 4.5. In the next subsection, we will explain these conditions and show that it is possible to establish generalizability of the poisoning attack with Theorem 4.5.

Remark 4.7. $\mathrm{Rad}_N^{\mathcal{D}_S^{l_p}}(\mathcal{H}_{W,D,1})$ in inequality (2) is not easy to calculate in terms of W, D, N, but we can demonstrate that if N is sufficiently large, this value will approach to 0 in most cases, as shown in Appendix D.3.

Remark 4.8. Please note that the right-hand side of inequality (2) is the empirical risk $\mathcal{R}(\mathcal{F}, \mathcal{D}_P)$ but not the empirical error $\mathcal{E}(\mathcal{F}, \mathcal{D}_P)$. This is due to some scaling techniques used in the proof, which is reasonable. In order to achieve "victim network classifies $x + \mathcal{P}(x)$ as class l_p", the victim network must learn the poisoned data $\mathcal{D}_P$ very well, just classifying the poison data correct is not enough.

4.3. Explaining the Conditions in Theorem 4.5

In order to make the bound 2 small, the value of ϵ and τ need to be small and λ need to close to 1. In this section, we show that how these values in the conditions (c1) to (c3) could be small.

How ϵ could be small? By condition (c1), since $\mathcal{G}_y(x)$ represents the probability that the network $\mathcal{G}$ classifies x as y, to make ϵ small, we only need to take the trigger $\mathcal{P}(x)$ as adversarial noise of the network $\mathcal{G}$ trained on the clean dataset. In other words, if $\mathcal{P}(x)$ *is adversary of* x *of a network trained on clean data*, then ϵ is small.

How λ could close to 1? By condition (c2), the upper bound is proportional to λ, so we hope to have a small λ. When $\mathcal{P}(x)$ is the same for all x, we have $\lambda = 1$. So, to make λ close to 1, $\mathcal{P}(x)$ need to be similar to x with different labels.

How τ could be small? Condition (c3) is not intuitive. In the rest of this section, we give a condition for τ to be small. We give a simplified version of the proposition and definition for easier reading. For a formal description, please refer to the Appendix E.1.

Intuitive speaking, if $\mathcal{F}$ is a network trained on $\mathcal{D}_P$, then the meaning of (c3) is that the backdoor part (i.e. $\mathcal{F} - \mathcal{G}$) gives the similar outputs for $\mathcal{P}(x)$ and $x + \mathcal{P}(x)$. This is similar to some conclusions in the indiscriminate poison(Zhu et al., 2023b; Huang et al., 2021), and by(Yu et al., 2022), indiscriminate poison can be considered as shortcut. These encourage the trigger to be shortcut. We will show that, under some assumptions, *making $\mathcal{P}(x)$ to be shortcut can ensure condition (c3).* First, we define the shortcut.

Definition 4.9 (Binary shortcut, Informal)**.** $\mathcal{P}(x)$ is called a shortcut of the binary linear inseparable classification dataset $\mathcal{D} = \{(x_i, 1)\}_{i=1}^{N_1} \cup \{(\widehat{x}_i, 0)\}_{i=1}^{N_0}$, if $\mathcal{D}_P = \{(x_i, 1)\}_{i=1}^{N_1} \cup \{(\widehat{x}_i + \mathcal{P}(\widehat{x}_i), 0)\}_{i=1}^{N_0}$ is linear separable.

Definition 4.9 means that shortcut is a poison which makes the poisoned dataset to be linear separable. Finally, we will show that, when $\mathcal{P}(x)$ is a suitable shortcut, there exists an upper bound for $\mathbb{E}_{x\sim\mathcal{D}_S}[|(\mathcal{F}-\mathcal{G})_{l_p}(\mathcal{P}(x)) - (\mathcal{F}-\mathcal{G})_{l_p}(x+\mathcal{P}(x))|]$, which implies that (c3) in Theorem 4.5 can be satisfied with a small τ.

Proposition 4.10 (Informal. The exact form and proof are given in Appendix E)**.** *Following Theorem 4.5 and let $D'_P = \{(x,0)|(x,y) \in \mathcal{D}_{\rm tr} \setminus \mathcal{D}_{clean}\} \cup \{(x,1)|(x,y) \in \mathcal{D}_{clean}\}$. Under certain mild conditions, if $\mathcal{P}(x)$ is the shortcut of the dataset $\mathcal{D}'_P$ and N is big enough, then with high probability, for some $\mathcal{F} \in \mathcal{H}_{W,D}$ satisfying $\mathcal{F}_y(x) > 1-\epsilon$ for $\forall(x,y) \in \mathcal{D}_P$ and $\mathcal{G} \in \mathcal{H}_{W,D}$ satisfying $\mathcal{G}_y(x) > 1-\epsilon$ for $\forall(x,y) \in \mathcal{D}_{\rm tr}$, we have $\mathbb{E}_{x\sim\mathcal{D}_S}[|(\mathcal{F}-\mathcal{G})_{l_p}(\mathcal{P}(x)) - (\mathcal{F}-\mathcal{G})_{l_p}(x+\mathcal{P}(x))|] \le \widetilde{O}(\epsilon)$.*

So, let $\mathcal{P}(x)$ be shortcut of D'_p, then, by Proposition 4.10, if $\mathcal{F} \in \mathcal{H}_{W,D}$ fits $\mathcal{D}_P$ well and $\mathcal{G} \in \mathcal{H}_{W,D}$ fits $\mathcal{D}_{\rm tr}$ well, or empirically, $\mathcal{F}(\mathcal{G})$ is well trained on dataset $\mathcal{D}_P(\mathcal{D}_{\rm tr})$, then condition (c3) holds for a small τ.

Remark 4.11. For condition (c3) in Theorem 4.5, we need to clarify that although $\mathcal{F} \approx \mathcal{G}$ implies (c3) for a small τ, but τ is small does not equivalent to $\mathcal{F} \approx \mathcal{G}$. Moreover, using $\mathcal{F} \approx \mathcal{G}$ instead of (c3) can also make Theorem 4.5 valid because it can derive (c3). But this is not a good idea, because $\mathcal{F} \approx \mathcal{G}$ makes $\mathcal{F}$ satisfying (c1), and then $\mathcal{F}(x+p(x))$ always does not output label l_p when $x \sim D_S^{l_p}$. So for the poisoned data in the poisoning dataset, $\mathcal{F}$ cannot output the correct labels, leading to a larger empirical error on the right-hand side of equation (2), and consequently leads to a big poison generalization error upper bound. Obviously, what we need is a small poison generalization error upper bound but not a big one, so we cannot only consider $\mathcal{F} \approx \mathcal{G}$.

5. Method

In this section, we will propose a new clean-label poison attack based on Theorems 4.1 and 4.5. There exists no requirement for the trigger $\mathcal{P}(x)$ in Theorem 4.1, so we only need to make the trigger approximately satisfy the conditions of Theorem 4.5. Our method thus has certain theoretical guarantee.

From Section 4.3, in order to satisfy the three conditions in Theorem 4.5, $\mathcal{P}(x)$ need to be (1) *adversarial noises for the clean-trained network, (2) shortcut for a specifically designed binary dataset, (3) similar for different samples.*

Remark 5.1. The effectiveness of adversarial and shortcut in creating backdoor has already been demonstrated empirically in previous work. On the other hand, our theory provides a more informed approach to using these methods.

Based on the above three requirements just mentioned, we design the trigger as follows:

(M1): Obtain adversarial disturbance: For any given clean sample x, use PGD on the network trained on $\mathcal{D}_{\rm tr}$ to find adversarial noise x_{adv} of x.

(M2): Obtain shortcut disturbance: For any given clean sample x, use min-min method (Huang et al., 2021) under the clean training set to find the shortcut x_{scut} of x. In (Zhu et al., 2023b) it has been shown that the shortcuts created by the min-min method are indeed similar for different x, thus satisfying condition (c2) automatically.

(M3): The trigger of x is designed to be $\mathcal{P}(x) = Ux_{adv} + (1-U)x_{scut}$, where $U \in \{0,1\}^n$ is a mask. We combine such two disturbances in this way because: (1) make the trigger both adversarial and shortcut; (2) make the triggers have a certain degree of similarity for different x, using the fact that the part of $(1-U)x_{scut}$ in trigger is similar for different x. It is worth mentioning that making x_{adv} similar for different x is difficult, so creating a trigger by $\lambda x_{adv} + (1-\lambda)x_{scut}$ will not be effective to ensuring similarity.

The mask U in (M3) is constructed as follows. The upper left corner is 0 and the other part is 1. On the basis of experience, it is necessary to disturb the key parts of the image to create adversarial samples, so we use a large portion to create adversaries. But the cost of the shortcut is relatively low, so we just use a small place to create the shortcut.

Algorithm 1 provides detailed steps for creating the trigger, where $\cdot$ is element-wise product.

In Algorithm 1, $\mathcal{F}_1$ is used to create adversarial disturbance and $\mathcal{F}_2$ is used to create shortcuts. When we complete step $S3$ in the algorithm, we will save $\mathcal{F}_1$ and $\mathcal{F}_2$, and for any sample (x, y), we directly generate $\mathcal{P}(x)$ using S4. Some poisons obtained using Algorithm 1 are shown in Figure 1.

6. Experiments

In this section, we empirically validate the proposed backdoor attack on benchmark datasets CIFAR10, CIFAR100 (Krizhevsky et al., 2009), SVHN and TinyImageNet(Le & Yang, 2015), and against popular defenses. We also conduct ablation experiments to verify our main Theorems 4.1 and 4.5. All experiments are repeated for 3 times and report the

Algorithm 1 Method of Creating the Trigger:

```
Input:
  An initialized network $\mathcal{F}_1 : \mathbb{R}^n \to \mathbb{R}^m$;
  An initialized network $\mathcal{F}_2 : \mathbb{R}^n \to \mathbb{R}^2$;
  A clean dataset $T = \{(x_i, y_i)\}_{i=1}^N \subset [0,1]^n \times [m]$;
  A mask vector $U \in \{0,1\}^n$;
  The poison budget $\eta$;
  Victim dataset $\{(x_i^v, y_i^v)\}_{i=1}^V$.
Output:
    Trigger $\{\mathcal{P}(x_i^v)\}_{i=1}^V$ for all victim data $\{(x_i^v, y_i^v)\}$.
  S1 Use $T$ to train the network $\mathcal{F}_1$.
  S2 Let $T_1 = \{\}$, for each $(x, y) \in T$:
      let $x_{adv} = x + U \cdot \text{argmax}_{||\varepsilon|| \leq \eta} L(F_1(x + \varepsilon), y)$
      add $(x_{adv}, 0)$ and $(x, 1)$ to $T_1$.
  S3 Use $T_1$ to train the network $\mathcal{F}_2$ as follows:
      $$\min_{\mathcal{F}_2} \sum_{(x,y) \in T_1} L(\mathcal{F}_2(x + \varepsilon(x, y)), y)$$
  where $\varepsilon(x, y) = \mathbf{1}(y = 0) \cdot (1 - U) \cdot \underset{||\varepsilon|| \leq \eta}{\text{argmin}}\, L(\mathcal{F}_2(x + (1 - U) \cdot \varepsilon), y)$.
  S4 For any victim data $(x_i^v, y_i^v)$, we calculate $\mathcal{P}(x_i^v)$ as following:
    $x_a^v = x_i^v + U \cdot \text{argmax}_{||\varepsilon|| \leq \eta} L(\mathcal{F}_1(x_i^v + \varepsilon), y_i^v)$;
    $\mathcal{P}(x_i^v) = (x_a^v - x_i^v) +$
        $(1 - U) \cdot \underset{||\varepsilon|| \leq \eta}{\text{argmin}}\, L(\mathcal{F}_2(x_a^v + \varepsilon \cdot (1 - U)), 0)$.
  Output: Trigger $\{\mathcal{P}(x_i^v)\}_{i=1}^V$.
```

average values. Furthermore, we make our attacks invisible by limiting the L_∞ norm of trigger. Details about the experiment setting can be found in Appendix F.1. Code is in https://github.com/hong-xian/backdoor-attack.git.

6.1. Baseline Evaluation

In this subsection, we study the effectiveness of our backdoor attack. For backdoor attacks, the goal is to misclassify the samples with trigger into a specified target class l_p. Unless said otherwise, we set the target label l_p as 0 in this paper. In addition to evaluating the test accuracy and attack success rate (ASR), we also measure *the accuracy of the target class* l_p to analyze the impact of the attack on the target label.

Table 1 shows the result on CIFAR-10 when perturbing 1% of training images, with each perturbation restricted to a radius l_∞-norm $16/255$. We observe that the ASR exceeds 90%, while the poison has negligible impact on both the test accuracy and the target class accuracy. In Table 2, we observe that the proposed attack remains remarkably effective even when the poison budget is very small. More experiments on different poison budgets and norm bounds can be found in Table 8. About transferability of attack and whether the victim network has learned the feature of trigger, please refer to the Appendix F.5.

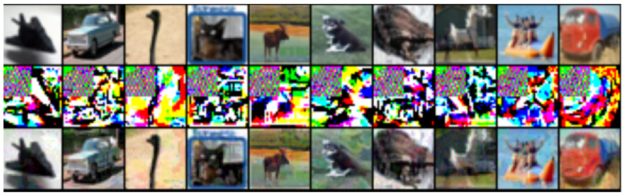

Figure 1. From top row to bottom row are respectively the clean images, normalized triggers (original trigger has L_∞ norm bound $16/255$), poison images. Due to the selection of U, the upper left corners of the poison images are similar, while the other parts are used to generate adversaries.

Table 1. Baseline evaluations on CIFAR-10. Perturbations have l_∞-norm bounded above by $16/255$, and poison budget is 1% of training images. Res means ResNet18, VGG means VGG16, WR means WRN34-10.

Model	Res	VGG	WR
Clean model acc (%)	93	92	94
Poisoned model acc(%)	91	91	92
Clean model l_p acc(%)	94	92	95
Poisoned model l_p acc(%)	93	91	93
Attack Success Rate(%)	93	91	90

Table 2. The effect of poison budget on CIFAR-10 with ResNet18. Perturbations have l_∞-norm bounded above by $16/255$.

Poison Budget	0.6%	1%	2%
Clean model acc (%)	93	93	93
Poisoned model acc(%)	93	91	93
Clean model l_p acc(%)	94	94	94
Poisoned model l_p acc(%)	93	93	92
Attack Success Rate(%)	86	93	94

Attack performance during the training process: To further validate the efficacy of our attack, we monitor the evolution of the overall clean model accuracy, the poisoned model accuracy, and the attack success rate throughout the training process, as shown in Figure 2. In Figure 2, we observe that the overall clean model accuracy and poisoned model accuracy remain very close. The attack success rate reaches a relatively high level from the very beginning and gradually converges to remain stable over time. This shows that our attack method is effective under the premise of maintaining the accuracy of the model.

Any target label: Note that the success of backdoor attacks also depends on the choice of target classes, to demonstrate the general efficacy of our attack for any target label l_p, we change l_p from 0 to 9, the result is shown in Figure 3. The results are quite uniform.

6.2. Evaluation on More Datasets

We perform experiments on SVHN, CIFAR-100 and Tiny-ImageNet. Table 3 summarizes the performance of our attack on different datasets, where the attacks are tested on ResNet18 for SVHN and CIFAR-100, and WRN34-10 for TinyImageNet. Each attacker can only perturb 0.8% of the training images for TinyImageNet and CIFAR-100 and 2% of the training images for SVHN, all perturbations are restricted to an l_∞ norm $16/255$, and the target label is $l_p = 0$. Additional experiments on different poison budgets and l_∞-norm bounds are presented in Table 9 in the appendix.

Table 3. Evaluations on more datasets. Perturbations have l_∞-norm bounded by $16/255$, and poison budget is 0.8% for TinyImageNet and CIFAR-100 and 2% for SVHN.

Datasets	Clean acc	Poison acc	ASR
SVHN (%)	93	92	79
CIFAR-100 (%)	76	72	92
TinyImageNet(%)	62	60	82

6.3. Compare with Other Attacks

There are several existing clean-label hidden-trigger backdoor attacks that claim success in some settings. We consider the following seven attack methods.

Clean Label: Turner et al. (2018) pioneered clean label attacks. They first utilized adversarial perturbations or generative models to initially alter target class images and then performed standard invisible attacks.

Reflection: Liu et al. (2020) proposed adopting reflection as the trigger for stealthiness.

Hidden Trigger: Saha et al. (2020) proposed to inject the information of a poisoned sample generated by a previous visible attack into an image of the target class by minimizing its distance in the feature space.

Invisible Poison Ning et al. (2021) converted a regular trigger to a noised trigger to achieve stealthiness, but remains effective in the feature space for poison training data.

Image-specific: Luo et al. (2022) used an autoencoder to generate image-specific triggers that can promote the implantation and activation phases of the backdoor.

Narcissus: Zeng et al. (2022) solved the following optimization to obtain the trigger $\mathcal{P}$: $\text{argmin}_{\mathcal{P}} \sum L(f_{sur}(x+\mathcal{P}, l_p)$, where f_{sur} is the surrogate network.

Sleeper Agent: Souri et al. (2022) proposed a backdoor attack by approximately solving the bilevel formulation with the Gradient Matching method (Geiping et al., 2020).

To further validate the efficacy of our attack, we compare our method with other clean label attack methods under the same settings. Specifically, we limit the L_∞ norm trigger no more than $16/255$ on both the training set and the test set. It should be noted that, in some attack methods, the trigger budget in their original settings may exceed $16/255$, so we also compare under each respective settings. The results are given in Table 4. More results and some details are provided in Appendix F.3.

Table 4. Attack success rate on CIFAR-10 with ResNet18. Comparison of our method to popular clean-label attacks, poison budget is 1%. The first column is the attack under L_∞ limitation $16/255$, the second column is under each respective settings.

Attacks	16/255	Each setting
Clean-Label	23%	96%
Hidden-Trigger	75%	95%
Reflection	54%	90%
Invisible Poison	73%	98%
Image-specific	70%	70%
Narcissus	50%	92%
Sleeper-Agent	61%	71%
Ours	**93**%	93%

From Table 4, we find that our attack significantly outperforms all other methods under the $16/255$ limitation. Moreover, our method has two key advantages: (1) Our method does not require additional steps. Sleeper Agent and Hidden-Trigger need a pre-existing patch and Reflection requires a fitting image; Narcissus needs to magnify the trigger in the reference phase. (2) Our method does not require large networks during poison generation, whereas Invisible Poison and Image-specific utilize large generative models to achieve optimal performance, and Narcissus requires a network trained on a different dataset. See Tables 10 and 11 in the appendix for more results.

Under each respective setting, all results were good, but many of them no longer meet the $16/255$ limitation. Furthermore, some of them need a patch in the trigger like Clean-label and Sleeper-Agent, and some of them need a larger disturbance like Reflection and Narcissus.

6.4. Defenses

Many backdoor defenses have been proposed to mitigate the effects of backdoor attacks. We test our attack and the attack methods mentioned in section 6.3 against some major popular defenses. We evaluate six types of defenses:

(1) AT: adversarial training with radius $8/255$ (Madry et al.,

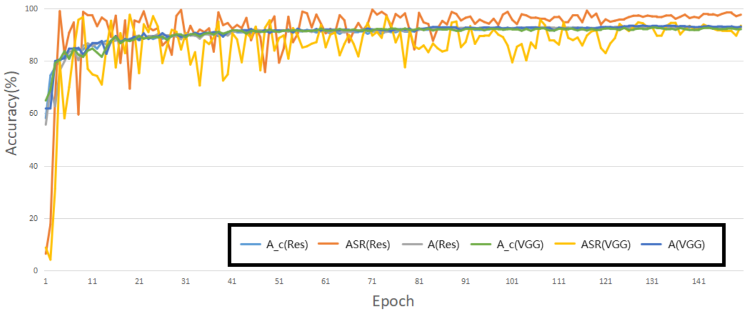

Figure 2. Attack performance during the training process on CIFAR10 with ResNet18 and VGG16. This figure shows the trend of the poison model accuracy (A), attack success rate (ASR) and clean model accuracy (A_c).

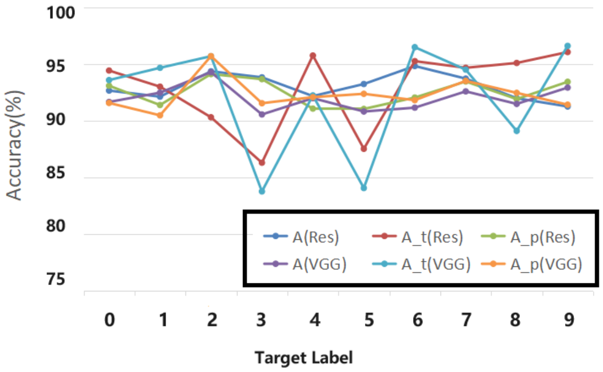

Figure 3. Performance of different target label l_p. We show the poison model accuracy (A), accuracy of target label (A_t), attack success rate (A_p) on CIFAR-10, using VGG16 and ResNet18.

2017);

(2): Data Augmentation: (Borgnia et al., 2021);

(3) Scale-up: contrastive learning (Guo et al., 2023);

(4) DPSGD: differentially private SGD (Hong et al., 2020);

(5) Frequency Filter: remove high-frequency parts of images (Zeng et al., 2021b);

(6): Fine-Tuning: (Zhu et al., 2023a).

The defense results are given in Table 5. We can see that ASR has basically decreased to around 10%, which is the lowest level because 10% of the samples themselves have the label 0. This is because we have imposed many restrictions on backdoor attacks, as said in Section F.1; and also because our theory and construction of trigger are mainly based on clean training, so our method appears somewhat fragile under defense; in fact, all attack methods mentioned in Section 6.3 appear fragile under defense methods, as shown in Table 5.

On the other hand, we find that there exists a robustness-accuracy trade-off across many of these defenses. Although these defense methods do degrade the attack success rate, they also cause the accuracy of the model test to decrease. For defense methods Scale-Up, this method is not stable and sometimes detects clean samples as poison samples; for fine-tuning, because a clean training set was used, this defense method has a very outstanding effect.

Furthermore, we point out that if we incorporate these defense methods into our attack generation to produce corresponding enhanced attacks, we can effectively withstand these defenses. For defense methods (1), (2), (4) and (5), we enhance our attack, details are shown in Appendix F.4, and get the better result.

6.5. Verify Theorem 4.5

In this section, we verify Theorem 4.5. In Appendix F.7, we verify Theorem 4.1.

We verify Theorem 4.5 by showing that **any trigger $\mathcal{P}(x)$ that makes the ϵ, τ, λ in conditions (c1), (c2), (c3) of Theorem 4.5 small can achieve a high attack success rate**.

To validate our conclusions, we evaluate the following poisoning function $\mathcal{P}(x)$, refer to Appendix F.6 for more details:

RN: Random noises with L_∞ bound and L_0 norm bound;

UA: Universal adversarial perturbations (Moosavi-Dezfooli et al., 2017);

Adv: Adversarial perturbations (Szegedy et al., 2013);

SCut: shortcut noise(Huang et al., 2021);

Ours: Perturbations generated by Algorithm 1.

We consider two indicators to evaluate the performance of poison on conditions (c1) and (c3) in Theorem 4.5:

- Use the validation loss on poisoned data to measure condition (c1): $V_{adv} = \mathbb{E}_{(x,y)\sim\mathcal{D}} L(\mathcal{F}(x + \mathcal{P}(x)), y)$,

Table 5. Defenses: The attack success rate(%) and poison model accuracy(%, in bracket) on CIFAR-10 with ResNet18 of our attack and other attacks against various defense methods. Poison ratio is 1% and perturbation have l_∞-norm bound $16/255$, target label $l_p = 0$.

	AT	Data Augmentaion	Scale-Up	DPSGD	Frequency Filter	Fine-Tuning
Clean Label(%)	13(83)	17(89)	9(77)	12(78)	12(86)	11(89)
Invisible Poison(%)	11(84)	30(91)	18(80)	18(76)	20(88)	12(90)
Hidden Trigger(%)	14(83)	25(90)	31(81)	14(75)	20(87)	10(88)
Narcissu(%)	13(83)	23(90)	11(83)	15(76)	16(86)	10(89)
Image-specific(%)	10(84)	28(89)	22(79)	14(77)	22(87)	12(87)
Reflection(%)	13(85)	20(89)	16(82)	13(76)	37(85)	13(85)
Sleeper-Agent	14(83)	27(87)	27(79)	15(75)	27(84)	11(89)
Ours(%)	15(83)	40(91)	32(80)	12(77)	28(88)	13(89)
Ours-stronger(%)	**34**(84)	**66**(90)	–	**52**(80)	**62**(88)	–

where $\mathcal{F}$ is ResNet18 trained on the clean dataset.

- Use the binary classification loss to measure condition (c3) by Proposition 4.10: $V_{sc} = \min_{\mathcal{F}} \mathbb{E}_{(x,y)\sim\mathcal{D}}[L(\mathcal{F}(x + \mathcal{P}(x), 0)) + L(\mathcal{F}(x), 1)]$, where $\mathcal{F}$ is a two-layer network.

About condition (c2) in Theorem 4.5: for RN and UA, the poison perturbation is the same for every sample; for SCut and Ours, the perturbations for different samples are similar (or at least parts of the perturbation are similar).

Table 6 shows the results. We can see that RN, UA, SCut with a large budget can achieve good attack performance because they satisfy conditions (c1), (c2), (c3) in Theorem 4.5 to certain degree, which validates the effectiveness of conditions in Theorem 4.5. Moreover, each of V_{adv} and V_{sc} is not satisfied, because adversaries alone yield excessively large V_{sc} since adversarial perturbations do not form shortcuts. Shortcuts alone produce an undesirably small V_{adv} since the shortcuts do not become adversaries. However, by combining Adv and SCut via our algorithm, we achieve a significantly improved outcome.

On the other hand, please note Adv attack under the $32/255$ budget. Although the Adv attack's similarity between different samples is poor, but its V_{adv} and V_{sc} is not bad, and its ASR is about 80%, this indicating that even if the similarity is not good enough, but the other two indicators are good, the attack can still be achieved. Therefore, in order to prevent the trigger from being detected due to being too similar, we can achieve attack effectiveness by reducing similarity and improving other two metrics.

Table 6. Values of V_{adv} and V_{sc}; poison model accuracy (Acc); attack success rate (ASR) on CIFAR-10 test set with ResNet18. Poison budget is 1%. If not specified, the norm is L_∞.

Poison Type	$V_{adv}(\uparrow)$	$V_{sc}(\downarrow)$	ASR($\uparrow$)	Acc
RN (16/255)	2.40	0.014	12%	91%
RN (L_0, 200)	3.87	0.004	59%	92%
UA (16/255)	2.92	0.002	51%	91%
Adv (16/255)	**8.92**	1.27	22%	92%
SCut (16/255)	1.19	$\mathbf{10^{-4}}$	30%	91%
Ours (16/255)	6.53	0.001	**93**%	**91**%
RN (32/255)	6.28	10^{-4}	99%	91%
RN (L_0, 300)	6.33	0.003	94%	91%
UA (32/255)	15.45	10^{-4}	92%	92%
Adv (32/255)	**16.38**	0.35	80%	92%
SCut (32/255)	4.22	$\mathbf{10^{-5}}$	93%	90%
Ours (32/255)	14.65	10^{-4}	**99**%	**92**%

7. Conclusion

In this paper, we give generalization bounds for the clean-label backdoor attack. Precisely, we provide upper bounds for the clean and poison population error based on empirical error on the poisoned training set and some other quantities. These bounds give the theoretical foundation for the clean-label backdoor attack in that the goal of the attack can be achieved under certain reasonable conditions. The main technical difficulties in establishing these bounds include how to treat the non-i.i.d. poisoned dataset and the fact that the triggers are both in the training and testing phases.

Based on these theoretical results, we propose a novel attack method that uses a combination of adversarial noise and indiscriminate poison as the trigger. Moreover, extensive experiments show that our attack can guarantee that the accuracy of the poisoned model on clean data and the attack success rate are high.

Limitations and Future Work. The conditions of Theorem 4.5 are quite complicated, and it is desirable to give simpler conditions for the poisoned population error bound in Theorem 4.5. The current generalization bounds do not involve the training process, and algorithmic-dependent generalization bounds, such as stability analysis (Hardt et al., 2016), should be further analyzed for backdoor attacks.

Impact Statement

A theoretical basis for backdoor attacks is given. One potential negative social impact of this work is that malicious opponents may use these methods to generate new types of backdoor poisons. Therefore, it is necessary to develop more powerful models to resist backdoor attacks, which is left for future work.

Acknowledgments

The 5-digit paper checker code is MWE3M. This work is supported by CAS Project for Young Scientists in Basic Research, Grant No.YSBR-040, ISCAS New Cultivation Project ISCAS-PYFX-202201, and ISCAS Basic Research ISCAS-JCZD-202302. This work is supported by NSFC grant No.12288201 and NKRDP grant No.2018YFA0306702. The authors thank anonymous referees for their valuable comments.

References

Arora, S., Ge, R., Neyshabur, B., and Zhang, Y. Stronger generalization bounds for deep nets via a compression approach. In *International Conference on Machine Learning*, pp. 254–263. PMLR, 2018.

Arora, S., Du, S., Hu, W., Li, Z., and Wang, R. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *International Conference on Machine Learning*, pp. 322–332. PMLR, 2019.

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

Barni, M., Kallas, K., and Tondi, B. A new backdoor attack in cnns by training set corruption without label poisoning. In *2019 IEEE International Conference on Image Processing (ICIP)*, pp. 101–105. IEEE, 2019.

Barron, A. R. and Klusowski, J. M. Approximation and estimation for high-dimensional deep learning networks. *arXiv preprint arXiv:1809.03090*, 2018.

Bartlett, P. L., Foster, D. J., and Telgarsky, M. J. Spectrally-normalized margin bounds for neural networks. *Advances in neural information processing systems*, 30, 2017.

Bartlett, P. L., Harvey, N., Liaw, C., and Mehrabian, A. Nearly-tight vc-dimension and pseudodimension bounds for piecewise linear neural networks. *The Journal of Machine Learning Research*, 20(1):2285–2301, 2019.

Bartlett, P. L., Montanari, A., and Rakhlin, A. Deep learning: a statistical viewpoint. *Acta numerica*, 30:87–201, 2021.

Bennouna, A. and Van Parys, B. Holistic robust data-driven decisions. *arXiv preprint arXiv:2207.09560*, 2022.

Bennouna, A., Lucas, R., and Van Parys, B. Certified robust neural networks: Generalization and corruption resistance. In *International Conference on Machine Learning*, pp. 2092–2112. PMLR, 2023.

Bertsekas, D. and Tsitsiklis, J. N. *Introduction to probability*, volume 1. Athena Scientific, 2008.

Borgnia, E., Cherepanova, V., Fowl, L., Ghiasi, A., Geiping, J., Goldblum, M., Goldstein, T., and Gupta, A. Strong data augmentation sanitizes poisoning and backdoor attacks without an accuracy tradeoff. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 3855–3859. IEEE, 2021.

Brutzkus, A. and Globerson, A. An optimization and generalization analysis for max-pooling networks. In *Uncertainty in Artificial Intelligence*, pp. 1650–1660. PMLR, 2021.

Cao, Y. and Gu, Q. Generalization bounds of stochastic gradient descent for wide and deep neural networks. *Advances in neural information processing systems*, 32, 2019.

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

Dabouei, A., Soleymani, S., Taherkhani, F., Dawson, J., and Nasrabadi, N. Smoothfool: An efficient framework for computing smooth adversarial perturbations. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2665–2674, 2020.

Doan, K., Lao, Y., and Li, P. Backdoor attack with imperceptible input and latent modification. *Advances in Neural Information Processing Systems*, 34:18944–18957, 2021a.

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

Dziugaite, G. K. and Roy, D. M. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. *arXiv preprint arXiv:1703.11008*, 2017.

Farnia, F. and Ozdaglar, A. Train simultaneously, generalize better: Stability of gradient-based minimax learners. In *International Conference on Machine Learning*, pp. 3174–3185. PMLR, 2021.

Fu, S., He, F., Liu, Y., Shen, L., and Tao, D. Robust unlearnable examples: Protecting data against adversarial learning. *arXiv preprint arXiv:2012.04115*, 2020.

Gao, Y., Li, Y., Zhu, L., Wu, D., Jiang, Y., and Xia, S.-T. Not all samples are born equal: Towards effective clean-label backdoor attacks. *Pattern Recognition*, 139:109512, 2023.

Geiping, J., Fowl, L., Huang, W. R., Czaja, W., Taylor, G., Moeller, M., and Goldstein, T. Witches' brew: Industrial scale data poisoning via gradient matching. *arXiv preprint arXiv:2009.02276*, 2020.

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

Guo, J., Li, Y., Chen, X., Guo, H., Sun, L., and Liu, C. Scale-up: An efficient black-box input-level backdoor detection via analyzing scaled prediction consistency. *arXiv preprint arXiv:2302.03251*, 2023.

Hanneke, S., Karbasi, A., Mahmoody, M., Mehalel, I., and Moran, S. On optimal learning under targeted data poisoning. *Advances in Neural Information Processing Systems*, 35:30770–30782, 2022.

Hardt, M., Recht, B., and Singer, Y. Train faster, generalize better: Stability of stochastic gradient descent. In *International conference on machine learning*, pp. 1225–1234. PMLR, 2016.

Harvey, N., Liaw, C., and Mehrabian, A. Nearly-tight vc-dimension bounds for piecewise linear neural networks. In *Conference on learning theory*, pp. 1064–1068. PMLR, 2017.

Hayase, J., Kong, W., Somani, R., and Oh, S. Spectre: Defending against backdoor attacks using robust statistics. In *International Conference on Machine Learning*, pp. 4129–4139. PMLR, 2021.

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

Hong, S., Chandrasekaran, V., Kaya, Y., Dumitraş, T., and Papernot, N. On the effectiveness of mitigating data poisoning attacks with gradient shaping. *arXiv preprint arXiv:2002.11497*, 2020.

Huang, H., Ma, X., Erfani, S. M., Bailey, J., and Wang, Y. Unlearnable examples: Making personal data unexploitable. *arXiv preprint arXiv:2101.04898*, 2021.

Huang, X., Alzantot, M., and Srivastava, M. Neuroninspect: Detecting backdoors in neural networks via output explanations. *arXiv preprint arXiv:1911.07399*, 2019.

Ji, Z. and Telgarsky, M. Polylogarithmic width suffices for gradient descent to achieve arbitrarily small test error with shallow relu networks. In *ICLR 2020*, 2020.

Kolouri, S., Saha, A., Pirsiavash, H., and Hoffmann, H. Universal litmus patterns: Revealing backdoor attacks in cnns. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 301–310, 2020.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. *Technical Report TR-2009*, 2009.

Kuzborskij, I. and Lampert, C. Data-dependent stability of stochastic gradient descent. In *International Conference on Machine Learning*, pp. 2815–2824. PMLR, 2018.

Le, Y. and Yang, X. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015.

Li, S., Xue, M., Zhao, B. Z. H., Zhu, H., and Zhang, X. Invisible backdoor attacks on deep neural networks via steganography and regularization. *IEEE Transactions on Dependable and Secure Computing*, 18(5):2088–2105, 2020.

Li, Y., Li, Y., Wu, B., Li, L., He, R., and Lyu, S. Invisible backdoor attack with sample-specific triggers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 16463–16472, 2021.

Li, Y., Ildiz, M. E., Papailiopoulos, D., and Oymak, S. Transformers as algorithms: Generalization and stability in in-context learning. In *International Conference on Machine Learning*, pp. 19565–19594. PMLR, 2023.

Liu, K., Dolan-Gavitt, B., and Garg, S. Fine-pruning: Defending against backdooring attacks on deep neural networks. In *International symposium on research in attacks, intrusions, and defenses*, pp. 273–294. Springer, 2018.

Liu, Y., Ma, X., Bailey, J., and Lu, F. Reflection backdoor: A natural backdoor attack on deep neural networks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16*, pp. 182–199. Springer, 2020.

Long, P. M. and Sedghi, H. Generalization bounds for deep convolutional neural networks. *arXiv preprint arXiv:1905.12600*, 2019.

Luo, N., Li, Y., Wang, Y., Wu, S., Tan, Y.-a., and Zhang, Q. Enhancing clean label backdoor attack with two-phase specific triggers. *arXiv preprint arXiv:2206.04881*, 2022.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2017.

Massart, P. Some applications of concentration inequalities to statistics. In *Annales de la Faculté des sciences de Toulouse: Mathématiques*, volume 9, pp. 245–303, 2000.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. *Foundations of machine learning*. MIT press, 2018.

Moosavi-Dezfooli, S.-M., Fawzi, A., Fawzi, O., and Frossard, P. Universal adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1765–1773, 2017.

Neyshabur, B., Bhojanapalli, S., McAllester, D., and Srebro, N. Exploring generalization in deep learning. *Advances in neural information processing systems*, 30, 2017.

Ning, R., Li, J., Xin, C., and Wu, H. Invisible poison: A blackbox clean label backdoor attack to deep neural networks. In *IEEE Conference on Computer Communications*, pp. 1–10. IEEE, 2021.

Saha, A., Subramanya, A., and Pirsiavash, H. Hidden trigger backdoor attacks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 11957–11965, 2020.

Sauer, N. On the density of families of sets. *Journal of Combinatorial Theory, Series A*, 13(1):145–147, 1972.

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

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

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

Tsai, Y.-L., Hsu, C.-Y., Yu, C.-M., and Chen, P.-Y. Formalizing generalization and robustness of neural networks to weight perturbations. *arXiv preprint arXiv:2103.02200*, 2021.

Turner, A., Tsipras, D., and Madry, A. Clean-label backdoor attacks. In *https://people.csail.mit.edu/madry/lab/cleanlabel.pdf*, 2018.

Valle-Pérez, G. and A. Louis, A. Generalization bounds for deep learning. *arXiv preprint arXiv:2203.14533*, 2022.

Vapnik, V. N. and Chervonenkis, A. Y. On the uniform convergence of relative frequencies of events to their probabilities. In *Measures of complexity: festschrift for alexey chervonenkis*, pp. 11–30. Springer, 2015.

Vardi, G., Shamir, O., and Srebro, N. The sample complexity of one-hidden-layer neural networks. *Advances in Neural Information Processing Systems*, 35:9139–9150, 2022.

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

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

Wang, M. and Ma, C. Generalization error bounds for deep neural networks trained by sgd. *arXiv preprint arXiv:2206.03299*, 2022.

Wang, Y., Mianjy, P., and Arora, R. Robust learning for data poisoning attacks. In *International Conference on Machine Learning*, 2021.

Wang, Y., Liu, S., and Gao, X.-S. Data-dependent stability analysis of adversarial training. *arXiv preprint arXiv:2401.03156*, 2024.

Wei, C. and Ma, T. Data-dependent sample complexity of deep neural networks via lipschitz augmentation. *Advances in Neural Information Processing Systems*, 32, 2019.

Xiao, J., Fan, Y., Sun, R., Wang, J., and Luo, Z.-Q. Stability analysis and generalization bounds of adversarial training. *Advances in Neural Information Processing Systems*, 35: 15446–15459, 2022.

Xing, Y., Song, Q., and Cheng, G. On the algorithmic stability of adversarial training. *Advances in neural information processing systems*, 34:26523–26535, 2021.

Yu, D., Zhang, H., Chen, W., Yin, J., and Liu, T.-Y. Indiscriminate poisoning attacks are shortcuts. *arXiv preprint arXiv:2111.00898*, 2022.

Zeng, Y., Chen, S., Park, W., Mao, Z. M., Jin, M., and Jia, R. Adversarial unlearning of backdoors via implicit hypergradient. *arXiv preprint arXiv:2110.03735*, 2021a.

Zeng, Y., Park, W., Mao, Z. M., and Jia, R. Rethinking the backdoor attacks' triggers: A frequency perspective. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 16473–16481, 2021b.

Zeng, Y., Pan, M., Just, H. A., Lyu, L., Qiu, M., and Jia, R. Narcissus: A practical clean-label backdoor attack with limited information. *arXiv preprint arXiv:2204.05255*, 2022.

Zhong, H., Liao, C., Squicciarini, A. C., Zhu, S., and Miller, D. Backdoor embedding in convolutional neural network models via invisible perturbation. In *Proceedings of the Tenth ACM Conference on Data and Application Security and Privacy*, pp. 97–108, 2020.

Zhu, M., Wei, S., Shen, L., Fan, Y., and Wu, B. Enhancing fine-tuning based backdoor defense with sharpness-aware minimization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4466–4477, 2023a.

Zhu, Y., Yu, L., and Gao, X.-S. Detection and defense of unlearnable examples. *arXiv preprint arXiv:2312.08898*, 2023b.

A. Proof of Theorem 4.1

A.1. Prelinimaries

We first give some notation that will be used in all the proofs.

Subdistribution $\mathcal{D}_S^{\neq l_p}$. Let $\mathcal{D}_S^{\neq l_p}$ be the distribution of samples whose label is not l_p, that is,

$$\mathbb{P}_{(x,y)\sim\mathcal{D}_S^{\neq l_p}}((x,y)\in A)=\mathbb{P}_{(x,y)\sim\mathcal{D}_S}((x,y)\in A|y\neq l_p)$$

for any set A.

Probability of samples to have label y_p. For a fixed p, let

$$\mathbb{P}_{(x,y)\sim\mathcal{D}_S}(y=l_p)=\eta \text{ and } 0<\eta<1. \tag{3}$$

General Hypothesis Space. In some theorems, we will consider the more general hypothesis space

$$H=\{h(x,y):\mathcal{S}\times[m]\to[0,1]\}.$$

H contains the commonly used hypothesis space $\{L(\mathcal{F}(x),y):\mathcal{S}\times[m]\to[0,1]\}$, where $\mathcal{F}$ is the network and L is the loss function.

We give a classic generalization bound below, which will be used in the proof.

Theorem A.1 (P.217 of (Mohri et al., 2018), Informal)**.** *Let the training set $\mathcal{D}_{\rm tr}$ be i.i.d. sampled from the data distribution $\mathcal{D}_S$ and $N=|\mathcal{D}_{\rm tr}|$. For the hypothesis space $\mathcal{H}=\{L(\mathcal{F}(x),y):\mathbb{R}^n\times[m]\to[0,1]\}$ and $\delta\in\mathbb{R}_+$, with probability at least $1-\delta$, for any $L(\mathcal{F}(x),y)\in\mathcal{H}$, we have*

$$\mathbb{E}_{(x,y)\sim\mathcal{D}_S}[L(\mathcal{F}(x),y)]\leq\mathbb{E}_{(x,y)\in\mathcal{D}_{\rm tr}}[L(\mathcal{F}(x),y)]+2{\rm Rad}_N^{\mathcal{D}_S}(\mathcal{H})+\sqrt{\tfrac{\ln(1/\delta)}{2N}} \tag{4}$$

We prove Theorem 4.1 in three steps given in Sections A.2, A.3, and A.4, respectively.

A.2. Proof of Theorem A.2

We first prove the following theorem, which gives a generalization bound for a more general hypothesis space.

Theorem A.2. *Let $\mathcal{D}_S$, $\mathcal{D}_P$, $\mathcal{D}_S^{l_p}$, α be defined in Section 3 and $\mathcal{D}_S^{\neq l_p}$ be defined in Section A.1. Then for any hypothesis space $H=\{h(x,y):\mathcal{S}\times[m]\to[0,1]\}$, with probability at least $1-\delta-\frac{4\eta}{4\eta+(1-\eta)N}-\frac{4(1-\eta)}{\eta N+4(1-\eta)}$, for any $h\in H$, it holds*

$$\mathbb{E}_{(x,y)\sim\mathcal{D}_S}[h(x,y)]\leq\tfrac{4-2\alpha}{1-\alpha}\mathbb{E}_{(x,y)\in\mathcal{D}_P}[h(x,y)]+4{\rm Rad}_N^{\mathcal{D}_S^{\neq l_p}}(H)+\tfrac{4}{1-\alpha}{\rm Rad}_N^{\mathcal{D}_S^{l_p}}(H)+2\sqrt{\tfrac{\ln(2/\delta)}{N(1-\alpha)}}. \tag{5}$$

As mentioned previously, the samples in $\mathcal{D}_P$ do not satisfy "i.i.d. sampled from $\mathcal{D}_{\mathcal{S}}$", and we will find a subset of $\mathcal{D}_{\mathcal{S}}$, whose samples are i.i.d. sampled from $\mathcal{D}_{\mathcal{S}}$. The core of the proof of theorem A.2 is the following two lemmas, which show how to select such a subset.

Lemma A.3. *Use notations in Theorem A.2. Let X be the random variable of the number of samples with label $\neq l_p$ in $\mathcal{D}_P$. Let k be a given number in $\{1,2,\ldots,N\eta\}$. If $X\geq k$, we randomly select k samples without label l_p in $\mathcal{D}_P$ and let D_l be the set of these samples; otherwise, let D_l be the set of all samples without label l_p in $\mathcal{D}_P$. Let D_l satisfy the distribution $\mathcal{D}_{\mathcal{S}_0}$. Then we have $\mathbb{P}_{x\sim\mathcal{D}_{\mathcal{S}_0}}(x\in A|X\geq k)=\mathbb{P}(X_k^{\mathcal{D}_S^{\neq l_p}}\in A)$ for any set A, where $X_k^{\mathcal{D}_S^{\neq l_p}}$ means i.i.d. sampling k data from distribution $\mathcal{D}_S^{\neq l_p}$.*

Proof. By the Bayesian formula, we have

$$\begin{array}{rl} & \mathbb{P}_{x\sim\mathcal{D}_{\mathcal{S}_0}}(x\in A|X\geq k)\\ = & \mathbb{P}_{x\sim\mathcal{D}_{\mathcal{S}_0}}(x\in A,X\geq k)/\mathbb{P}_{x\sim\mathcal{D}_{\mathcal{S}_0}}(X\geq k)\\ = & \sum_{i=k}^N\mathbb{P}_{x\sim\mathcal{D}_{\mathcal{S}_0}}(x\in A,X=i)/\mathbb{P}_{x\sim\mathcal{D}_{\mathcal{S}_0}}(X\geq k)\\ = & \sum_{i=k}^N\mathbb{P}_{x\sim\mathcal{D}_{\mathcal{S}_0}}(x\in A|X=i)\mathbb{P}(X=i)/\mathbb{P}_{x\sim\mathcal{D}_{\mathcal{S}_0}}(X\geq k)\\ = & \sum_{i=k}^N\mathbb{P}_{x\sim\mathcal{D}_{\mathcal{S}_0}}(x\in A|X=i)\mathbb{P}(X=i|X\geq k).\end{array} \tag{6}$$

We will show that $\mathbb{P}_{x\sim\mathcal{D}_{\mathcal{S}_0}}(x \in A|X = i) = \mathbb{P}(X_k^{\mathcal{D}_S^{\neq l_p}} \in A)$ for any $i \geq k$, and hence the lemma.

Since the poison does not change labels, X is also the number of samples without label l_p in $\mathcal{D}_{\rm tr}$. Then for any $i \geq k$, when $X = i$, we will traverse all possible selection methods for D_l to calculate $\mathbb{P}_{x\sim\mathcal{D}_{\mathcal{S}_0}}(x \in A|X = i)$.

Note that the N samples in $\mathcal{D}_{\rm tr}$ are i.i.d samples from $\mathcal{D}_{\mathcal{S}}$. We will consider the order of these samples. Let $(x_j, y_j) \in \mathcal{D}_{\rm tr}$ be the j-th element selected from the distribution $\mathcal{D}_{\mathcal{S}}$. If we add poison to (x_j, y_j), then it becomes $(x_j + \mathcal{P}(x_j), y_j) \in \mathcal{D}_P$; if we do not add poison to (x_j, y_j), then $(x_j, y_j) \in \mathcal{D}_P$.

Let $D_{y\neq l_p} \subset [N]$ be the set of k such $(x_k, y_k) \in \mathcal{D}_P$ satisfying $y_k \neq l_p$. By considering all the possible situations of $D_{y\neq l_p}$, we have

$$\begin{array}{rl} & \mathbb{P}_{x\sim\mathcal{D}_{\mathcal{S}_0}}(x \in A|X = i) \\ = & \mathbb{P}_{x\sim\mathcal{D}_{\mathcal{S}_0}}(x \in A, X = i)/\mathbb{P}(X = i) \\ = & \sum_{\mathcal{D}_{y\neq l_p}, |\mathcal{D}_{y\neq l_p}|=i} \mathbb{P}_{x\sim\mathcal{D}_{\mathcal{S}_0}}(x \in A, \mathcal{D}_{y\neq l_p})/\mathbb{P}(X = i) \\ = & \sum_{\mathcal{D}_{y\neq l_p}, |\mathcal{D}_{y\neq l_p}|=i} \mathbb{P}_{x\sim\mathcal{D}_{\mathcal{S}_0}}(x \in A|D_{y\neq l_p})\frac{\mathbb{P}(\mathcal{D}_{y\neq l_p})}{\mathbb{P}(X=i)} \\ = & \sum_{\mathcal{D}_{y\neq l_p}, |\mathcal{D}_{y\neq l_p}|=i} \mathbb{P}_{x\sim\mathcal{D}_{\mathcal{S}_0}}(x \in A|D_{y\neq l_p})/C_N^i. \end{array}$$

Thus, for all $\mathcal{D}_{y\neq l_p}$ satisfying $|\mathcal{D}_{y\neq l_p}| = i$, we traverse all the possibilities of the sample index $\{i_1, i_2, \ldots, i_k\}$ selected by $\mathcal{D}_l$ and then have

$$\begin{array}{rl} & \mathbb{P}_{x\sim\mathcal{D}_{\mathcal{S}_0}}(x \in A|D_{y\neq l_p}) \\ = & \frac{\sum_{i_1,i_2,\ldots,i_k \in D_{y\neq l_p}} \mathbb{P}((x_{i_1}, x_{i_2}, \ldots, x_{i_k}) \in A)}{C_i^k} \\ = & \frac{1}{C_i^k}\sum_{i_1,i_2,\ldots,i_k \subset D_{y\neq l_p}} \mathbb{P}(X_k^{\mathcal{D}_S^{\neq l_p}} \in A) \\ = & \mathbb{P}(X_k^{\mathcal{D}_S^{\neq l_p}} \in A) \end{array}$$

where x_i are i.i.d. and C_a^b is the combination number of selecting b samples from a samples. We thus have:

$$\begin{array}{rl} & \mathbb{P}_{x\sim\mathcal{D}_{\mathcal{S}_0}}(x \in A|X = i) \\ = & \frac{\sum_{\mathcal{D}_{y\neq l_p}, |\mathcal{D}_{y\neq l_p}|=i} \mathbb{P}_{x\sim\mathcal{D}_{\mathcal{S}_0}}(x\in A|D_{y\neq l_p})}{C_N^i} \\ = & \frac{\sum_{\mathcal{D}_{y\neq l_p}, |\mathcal{D}_{y\neq l_p}|=i} \mathbb{P}(X_k^{\mathcal{D}_S^{\neq l_p}} \in A)}{C_N^i} \\ = & \mathbb{P}(X_k^{\mathcal{D}_S^{\neq l_p}} \in A). \end{array}$$

Finally, we have

$$\begin{array}{rl} & \mathbb{P}_{x\sim\mathcal{D}_{\mathcal{S}_0}}(x \in A|X \geq k) \\ = & \sum_{i=k}^N \mathbb{P}_{x\sim\mathcal{D}_{\mathcal{S}_0}}(x \in A|X = i)\mathbb{P}(X = i|X \geq k) \\ = & \sum_{i=k}^N \mathbb{P}(X_k^{\mathcal{D}_S^{\neq l_p}} \in A)\mathbb{P}(X = i|X \geq k) \\ = & \mathbb{P}(X_k^{\mathcal{D}_S^{\neq l_p}} \in A)\sum_{i=k}^N \mathbb{P}(X = i|X \geq k) \\ = & \mathbb{P}(X_k^{\mathcal{D}_S^{\neq l_p}} \in A). \end{array}$$

This proves the lemma. □

For samples with label l_p, we have the similar conclusion.

Lemma A.4. *Let X be the random variable of the number of samples with label l_p in the set $\mathcal{D}_P$. Let k to be a given number in $\{1, 2, \ldots, [N\eta]\}$. If $X \geq k$, we randomly select $(1-\alpha)k$ samples without trigger but with label l_p in $\mathcal{D}_P$, and let these samples form the set D_{l_p}; otherwise, we select all samples without trigger but with label l_p in $\mathcal{D}_P$, and make these samples the set D_{l_p}. Let D_{l_p} obey the distribution $\mathcal{D}_{\mathcal{S}_1}$. Then we have $\mathbb{P}_{x\sim\mathcal{D}_{\mathcal{S}_1}}(x \in A|X \geq k) = \mathbb{P}(X_{(1-\alpha)k}^{\mathcal{D}_S^{l_p}} \in A)$ for any set A, where $X_{(1-\alpha)k}^{\mathcal{D}_S^{l_p}}$ means i.i.d. sample $(1-\alpha)k$ samples from distribution $\mathcal{D}_{\mathcal{S}}^{l_p}$.*

Proof. By the Bayesian formula, we have

$$\mathbb{P}_{x\sim\mathcal{D}_{\mathcal{S}_1}}(x \in A|X \geq k) = \sum_{i=k}^N \mathbb{P}_{x\sim\mathcal{D}_{\mathcal{S}_1}}(x \in A|X = i)\mathbb{P}(X = i|X \geq k).$$

The intermediate steps are similar to equation (6) in the proof of Lemma A.3, so we omit them.

Now we prove $\mathbb{P}_{x\sim\mathcal{D}_{\mathcal{S}_1}}(x\in A|X=i)=\mathbb{P}(X^{\mathcal{D}_{\mathcal{S}}^{l_p}}_{(1-\alpha)k}\in A)$ for any $i\geq k$. Note that the N samples in $\mathcal{D}_{\text{tr}}$ are i.i.d selected from $\mathcal{D}_S$. Now we will consider the order of these samples. Let $(x_j,y_j)\in\mathcal{D}_{\text{tr}}$ be the j-th element selected from the distribution $\mathcal{D}_{\mathcal{S}}$. If we add poison to (x_j,y_j), then it becomes $(x_j+\mathcal{P}(x_j),y_j)\in\mathcal{D}_P$; if we do not add poison to (x_j,y_j), then $(x_j,y_j)\in\mathcal{D}_P$.

For any $i\geq k$, when $X=i$, there must be $|\mathcal{D}_{l_p}|=(1-\alpha)k$. Let $D_{y=l_p}\subset[N]$ be the set of j such that $(x_j,y_j)\in\mathcal{D}_P$ satisfied $y_j=l_p$. Now we consider all the possible situation of $D_{y=l_p}$:

$$\begin{array}{rl} & \mathbb{P}_{x\sim\mathcal{D}_{\mathcal{S}_1}}(x\in A|X=i)\\ = & \mathbb{P}_{x\sim\mathcal{D}_{\mathcal{S}_1}}(x\in A,X=i)/\mathbb{P}(X=i)\\ = & \sum_{D_{y=l_p},|D_{y=l_p}|=i}\mathbb{P}_{x\sim\mathcal{D}_{\mathcal{S}_1}}(x\in A,D_{y=l_p})/\mathbb{P}(X=i)\\ = & \sum_{D_{y=l_p},|D_{y=l_p}|=i}\mathbb{P}_{x\sim\mathcal{D}_{\mathcal{S}_1}}(x\in A|D_{y=l_p})\mathbb{P}(D_{y=l_p})/\mathbb{P}(X=i)\\ = & \sum_{D_{y=l_p},|D_{y=l_p}|=i}\mathbb{P}_{x\sim\mathcal{D}_{\mathcal{S}_1}}(x\in A|D_{y=l_p})/C_N^i. \end{array} \tag{7}$$

Let $D^{poi}_{y=l_p}\subset[N]$ be the set of j satisfying that x_j is a poisoned sample, and $D^{no\ poi}_{y=l_p}\subset[N]$ be the set of j satisfying $(x_j,y_j)\in\mathcal{D}_{l_p}$. It is easy to see that $D^{no\ poi}_{y=l_p},D^{poi}_{y=l_p}\subset D_{y=l_p}$. Then, for any given $D_{y=l_p}$ such that $|D_{y=l_p}|=i$, we traverse all possibilities of $D^{poi}_{y=l_p}$ and $D^{no\ poi}_{y=l_p}$ to calculate $\mathbb{P}_{x\sim\mathcal{D}_{\mathcal{S}_1}}(x\in A|D_{y=l_p})$:

$$\begin{array}{rl} & \mathbb{P}_{x\sim\mathcal{D}_{\mathcal{S}_1}}(x\in A|D_{y=l_p})\\ = & \sum_{\{i_k\}_{k=1}^{i\alpha},\{i_j\}_{j=1}^{[N\eta](1-\alpha)}}\mathbb{P}(\{i_j\}_{j=1}^{[N\eta](1-\alpha)}\in A)\mathbb{P}(D^{poi}_{y=l_p}=\{i_k\}_{k=1}^{i\alpha},D^{no\ poi}_{y=l_p}=\{i_j\}_{j=1}^{[N\eta](1-\alpha)}|D_{y=l_p})\\ = & \sum_{\{i_k\}_{k=1}^{i\alpha},\{i_j\}_{j=1}^{[N\eta](1-\alpha)}}\mathbb{P}(X^{\mathcal{D}_{\mathcal{S}}^{l_p}}_{(1-\alpha)k}\in A)\mathbb{P}(D^{poi}_{y=l_p}=\{i_k\}_{k=1}^{i\alpha},D^{no\ poi}_{y=l_p}=\{i_j\}_{j=1}^{[N\eta](1-\alpha)}|D_{y=l_p})\\ = & \mathbb{P}(X^{\mathcal{D}_{\mathcal{S}}^{l_p}}_{(1-\alpha)k}\in A)\sum_{\{i_k\}_{k=1}^{i\alpha},\{i_j\}_{j=1}^{[N\eta](1-\alpha)}}\mathbb{P}(D^{poi}_{y=l_p}=\{i_k\}_{k=1}^{i\alpha},D^{no\ poi}_{y=l_p}=\{i_j\}_{j=1}^{[N\eta](1-\alpha)}|D_{y=l_p})\\ = & \mathbb{P}(X^{\mathcal{D}_{\mathcal{S}}^{l_p}}_{(1-\alpha)k}\in A). \end{array}$$

The x_i are i.i.d. and C_a^b is the number of combinations to select b samples from a samples. Substituting it into inequality (7), we have

$$\begin{array}{rl} & \mathbb{P}_{x\sim\mathcal{D}_{\mathcal{S}_1}}(x\in A|X=i)\\ = & \sum_{D_{y=l_p},|D_{y=l_p}|=i}\mathbb{P}_{x\sim\mathcal{D}_{\mathcal{S}_1}}(x\in A|D_{y=l_p})/C_N^i\\ = & \sum_{D_{y=l_p},|D_{y=l_p}|=i}\mathbb{P}(X^{\mathcal{D}_{\mathcal{S}}^{l_p}}_{(1-\alpha)k}\in A)/C_N^i\\ = & \mathbb{P}(X^{\mathcal{D}_{\mathcal{S}}^{l_p}}_{(1-\alpha)k}\in A). \end{array}$$

This is what we want. Finally, we have

$$\begin{array}{rl} & \mathbb{P}_{x\sim\mathcal{D}_{\mathcal{S}_1}}(x\in A|X\geq k)\\ = & \sum_{i=k}^N\mathbb{P}_{x\sim\mathcal{D}_{\mathcal{S}_1}}(x\in A|X=i)\mathbb{P}(X=i|X\geq k)\\ = & \sum_{i=k}^N\mathbb{P}(X^{\mathcal{D}_{\mathcal{S}}^{l_p}}_{(1-\alpha)k}\in A)\mathbb{P}(X=i|X\geq k)\\ = & \mathbb{P}(X^{\mathcal{D}_{\mathcal{S}}^{l_p}}_{(1-\alpha)k}\in A). \end{array}$$

The lemma is proved. □

Now, we prove Theorem A.2.

Proof. By the Bayesian formula, we have

$$\begin{array}{rl} & \mathbb{E}_{(x,y)\sim\mathcal{D}_S}[h(x,y)]\\ = & \mathbb{P}_{(x,y)\sim\mathcal{D}_S}(y=l_p)\mathbb{E}_{(x,y)\sim\mathcal{D}_S^{l_p}}[h(x,y)]+\mathbb{P}_{(x,y)\sim\mathcal{D}_S}(y\neq l_p)\mathbb{E}_{(x,y)\sim\mathcal{D}_S^{\neq l_p}}[h(x,y)]\\ = & \eta\mathbb{E}_{(x,y)\sim\mathcal{D}_S^{\neq l_p}}[h(x,y)]+(1-\eta)\mathbb{E}_{(x,y)\sim\mathcal{D}_S^{\neq l_p}}[h(x,y)]. \end{array} \tag{8}$$

Now we will separately estimate $\mathbb{E}_{(x,y)\sim\mathcal{D}_S^{\neq l_p}}[h(x,y)]$ and $\mathbb{E}_{(x,y)\sim\mathcal{D}_S^{l_p}}[h(x,y)]$.

Upper bound of $\mathbb{E}_{(x,y)\sim\mathcal{D}_S^{\neq l_p}}[h(x,y)]$. We give such a bound in the following Results (c1), (c2), and (c3).

Result (c1): Let random variable X be the number of samples with labels not equal to l_p in $\mathcal{D}_{\rm tr}$. Then with probability at least $1-\frac{4\eta}{4\eta+N-N\eta}$, we have $X\ge N(1-\eta)/2$.

Let $\mathcal{D}_{\rm tr}=\{(x_i,y_i)\}_{i=1}^N$. Then $X=\sum_{i=1}^N \mathbf{1}(y_i\ne l_p)$, so $\mathbb{E}[X]=\mathbb{E}[\sum_{i=1}^N \mathbf{1}(y_i\ne l_p)]=\sum_{i=1}^N \mathbb{E}[\mathbf{1}(y_i\ne l_p)]=N(1-\eta)$, and the variance $\mathbb{V}[X]=\mathbb{V}[\sum_{i=1}^N \mathbf{1}(y_i\ne l_p)]=\sum_{i=1}^N \mathbb{V}[\mathbf{1}(y_i\ne l_p)]=N(1-\eta)\eta$, because $\mathbf{1}(y_i\ne l_p)$ are independent events. By the Cantelli inequality, we have $\mathbb{P}(X\le N(1-\eta)/2)\le\frac{\mathbb{V}[X]}{\mathbb{V}[X]+(\mathbb{E}[X]-(N(1-\eta)/2))^2}=\frac{4\eta}{4\eta+N-N\eta}$, and Result (c1) is proved.

Result (c2): We randomly select $N(1-\eta)/2$ samples without label l_p in $\mathcal{D}_{\mathcal{P}}$ and let D_l be the set of these samples. If the number of samples without label l_p in $\mathcal{D}_{\mathcal{P}}$ is less than $N(1-\eta)/2$, then we select all samples without label l_p, and let D_l be the set of these samples.

Assuming that the set $\mathcal{D}_l$ obeys the distribution $\mathcal{D}_{\mathcal{S}_0}$, we have $\mathbb{P}_{x\sim\mathcal{D}_{\mathcal{S}_0}}(x\in A|X\ge N(1-\eta)/2)=\mathbb{P}(X_{N(1-\eta)/2}^{\mathcal{D}_S^{\neq l_p}}\in A)$ for any set A, where $X_{N(1-\eta)/2}^{\mathcal{D}_S^{\neq l_p}}$ is the set of $N(1-\eta)/2$ data i.i.d. sampled from distribution $\mathcal{D}_S^{\neq l_p}$.

Following Lemma A.3, Result (c2) shows that when $X\ge N(1-\eta)/2$, $\mathcal{D}_l$ can be seen as i.i.d. sampled from $\mathcal{D}_S^{\neq l_p}$.

Result (c3): With probability $1-\frac{4\eta}{4\eta+N-N\eta}-\delta/2$, we have

$$\mathbb{E}_{(x,y)\sim\mathcal{D}_S^{\neq l_p}}[h(x,y)]\le\frac{\sum_{(x,y)\in\mathcal{D}_P}h(x,y)}{N(1-\eta)/2}+2\text{Rad}_{N(1-\eta)/2}^{\mathcal{D}_S^{\neq l_p}}(H)+\sqrt{\frac{\ln(2/\delta).}{N(1-\eta)}}.$$

$D\in(\mathbb{R}^n\times[m])^{N(1-\eta)/2}$ is called a bad set, if $|D|=N(1-\eta)/2$, and

$$\mathbb{E}_{(x,y)\sim\mathcal{D}_S^{\neq l_p}}[h(x,y)]>\frac{\sum_{(x,y)\in D}h(x,y)}{N(1-\eta)/2}+2\text{Rad}_{N(1-\eta)/2}^{\mathcal{D}_S^{\neq l_p}}(H)+\sqrt{\frac{\ln(2/\delta)}{N(1-\eta)}} \tag{9}$$

for some $f\in H$. Let $S_b=\{D\|D\ \text{ is a bad set}\}$.

By Theorem A.1, if the samples in $\mathcal{D}$ are i.i.d. sampled form $\mathcal{D}_S^{\neq l_p}$ and $|D|=N(1-\eta)/2$, then with probability $1-\delta/2$, we have $\mathbb{E}_{(x,y)\sim\mathcal{D}_S^{\neq l_p}}[h(x,y)]\le\frac{\sum_{(x,y)\in D}h(x,y)}{N(1-\eta)/2}+2\text{Rad}_{N(1-\eta)/2}(H)+\sqrt{\frac{\ln(2/\delta)}{N(1-\eta)}}$ for any $f\in H$.

So by the definition of bad set, we have $\mathbb{P}(X_{N(1-\eta)/2}^{\mathcal{D}_S^{\neq l_p}}\in S_b)\le\delta/2$, where $X_{N(1-\eta)/2}^{\mathcal{D}_S^{\neq l_p}}$ is mentioned in result (c2). And by Result (c2), we have that: when $X\ge N(1-\eta)/2$, we have $\mathbb{P}(D_l\in S_b)\le\delta/2$. Then, by Result (c1), we have

$$\begin{array}{rl} & \mathbb{P}(D_l\notin S_b, X\ge N(1-\eta)/2)\\ = & \mathbb{P}(D_l\notin S_b|X\ge N(1-\eta)/2)\mathbb{P}(X\ge N(1-\eta)/2)\\ \ge & (1-\delta/2)(1-\frac{4\eta}{4\eta+N-N\eta})(by\ (c1))\\ \ge & 1-\delta/2-\frac{4\eta}{4\eta+N-N\eta}.\end{array}$$

So, with probability at least $1-\delta/2-\frac{4\eta}{4\eta+N-N\eta}$, D_l is not in S_b and $X\ge N(1-\eta)/2$. Hence,

$$\begin{array}{rl} & \mathbb{E}_{(x,y)\sim\mathcal{D}_S^{\neq l_p}}[h(x,y)]\\ \le & \frac{\sum_{(x,y)\in D_l}h(x,y)}{N(1-\eta)/2}+2\text{Rad}_{N(1-\eta)/2}(H)+\sqrt{\frac{\ln(2/\delta)}{N(1-\eta)}}\\ \le & \frac{\sum_{(x,y)\in\mathcal{D}_P}h(x,y)}{N(1-\eta)/2}+2\text{Rad}_{N(1-\eta)/2}(H)+\sqrt{\frac{\ln(2/\delta)}{N(1-\eta)}}.\end{array}$$

This is what we want.

The upper bound of $\mathbb{E}_{(x,y)\sim\mathcal{D}_S^{l_p}}[h(x,y)]$**.**

We will give such a bound in Results (d1), (d2), and (d3) below, which are similar to results (c1), (c2), (c3).

Result (d1): Let the random variable X be the number of samples with label l_p in $\mathcal{D}_{\rm tr}$. Then with probability at least $1-\frac{4(1-\eta)}{4(1-\eta)+N\eta}$, we have $X \geq N\eta/2$. The proof is similar to that of Result (c1).

Result (d2): Now we evenly select $[N\eta(1-\alpha)/2]$ samples with label l_p in $\mathcal{D}_{\mathcal{P}}/\mathcal{D}_{poi}$ and make these samples to form the set D_{l_p}. If the number of samples with label l_p in $\mathcal{D}_{\mathcal{P}}/D_{poi}$ is smaller than $[N\eta(1-\alpha)/2]$, then we select all of these samples and add these samples to D_{l_p}.

Assume that the set $\mathcal{D}_{l_p}$ obeys the distribution $\mathcal{D}_{\mathcal{S}_1}$. Then we have $\mathbb{P}_{x\sim\mathcal{D}_{\mathcal{S}_1}}(x\in A|X\geq[N\eta/2]) = \mathbb{P}(X^{\mathcal{D}_{\mathcal{S}}^{l_p}}_{[N\eta(1-\alpha)/2]}\in A)$ for any set A, where $X^{\mathcal{D}_{\mathcal{S}}^{l_p}}_{[N\eta(1-\alpha)/2]}$ is the set that i.i.d. selecting $[N\eta(1-\alpha)/2]$ samples from distribution $\mathcal{D}_{S}^{\neq l_p}$.

Following Lemma A.4, Result (d2) shows that when $X\geq[N\eta/2]$, D_{l_p} can be seen as i.i.d. samples from $D_{\mathcal{S}}^{l_p}$.

Result (d3): With probability $1-\frac{4-4\eta}{4(1-\eta)+N\eta}-\delta/2$, we have

$$\mathbb{E}_{(x,y)\sim\mathcal{D}_{\mathcal{S}}^{l_p}}[h(x,y)] \leq \frac{\sum_{(x,y)\in\mathcal{D}_P}h(x,y)}{N(1-\alpha)\eta/2} + 2{\rm Rad}^{\mathcal{D}_{\mathcal{S}}^{l_p}}_{N(1-\alpha)\eta/2}(H) + \sqrt{\frac{\ln(2/\delta)}{N(1-\alpha)\eta}}.$$

This is similar to Result (c3), but using (d1) and (d2) instead.

To obtain a bound of $E_{(x,y)\sim\mathcal{D}_S}[h(x,y)]$.

Using the fact ${\rm Rad}^D_M(H)\leq\frac{N}{M}{\rm Rad}^D_N(H)$ for any $M\leq N$, distribution $\mathcal{D}$, hypothesis space H, and applying (c3), (d3) to equation (8), we finally have

$$\begin{array}{ll} & \mathbb{E}_{(x,y)\sim\mathcal{D}_S}[h(x,y)] \\ = & \eta\mathbb{E}_{(x,y)\sim\mathcal{D}_{\mathcal{S}}^{l_p}}[h(x,y)] + (1-\eta)\mathbb{E}_{(x,y)\sim\mathcal{D}_{S}^{\neq l_p}}[h(x,y)] \\ \leq & \frac{4-2\alpha}{1-\alpha}\frac{\sum_{(x,y)\in\mathcal{D}_P}h(x,y)}{N} + 4{\rm Rad}_N^{\mathcal{D}_S^{\neq l_p}}(H) + 4\frac{{\rm Rad}_N^{\mathcal{D}_S^{l_p}}(H)}{1-\alpha} + \sqrt{\frac{\ln(2/\delta)(1-\eta)}{N}} + \sqrt{\frac{\ln(2/\delta)\eta}{N(1-\alpha)}}. \end{array} \tag{10}$$

Then using $\eta<1$, we prove the theorem. □

A.3. Estimate Rademacher Complexity

In this section, we will estimate the Rademacher Complexities in Theorem A.2 when $H=\{\mathbf{1}(\mathcal{F}(x)\neq y)\}$, where $\mathcal{F}$ is a network with width W and depth D. We first need a definition:

Definition A.5. Let H be a hypothesis space. Then the growth function $\Pi_H(N)$ of H is defined as:

$$\Pi_H(N) = \max_{\{x_i\}_{i=1}^N}|\{(h(x_i))_{i=1}^N\|h\in H\}|.$$

For a hypothesis space $H=\{h:\mathbb{R}^n\to\{-1,1\}\}$, we can estimate its VCdim (Vapnik & Chervonenkis, 2015), and the relationship between VCdim and growth function.

Lemma A.6 ((Bartlett et al., 2021; 2019)). *Let H be the hypothesis space that satisfies: $\mathcal{F}\in H$ if and only if $\mathcal{F}$ is a network with width not more than W and depth not more than D, and the activation function of each hidden layer of $\mathcal{F}$ is Relu, the output layer uses* sign *as activation function. Then the VCdim of H is $O(D^2W^2)$.*

Lemma A.7 ((Sauer, 1972))**.** *Let H be the hypothesis space with VCdim V. Then for any $N\geq 1$, the growth function satisfies $\Pi_H(N)\leq(eN)^V$.*

Lemma A.8 ((Massart, 2000; Mohri et al., 2018))**.** *Let H be the hypothesis space with growth function Π_H, and any $h(x)\in H$ satisfy $|h(x)|\leq 1$ for any x. Then for any distribution $\mathcal{D}$, we have ${\rm Rad}^{\mathcal{D}}_N(H)=O(\sqrt{\frac{\ln(\Pi_H)}{N}})$.*

Lemma A.9 ((Mohri et al., 2018))**.** *Let H be the hypothesis space, $q\in\mathbb{R}^m\to\{0,1\}$, and $H_q=\{q(h_1(x),h_2(x),\dots,h_m(x))\|h_i\in H\}$. Then $\Pi_{H_q}(N)\leq(\Pi_H(N))^m$.*

Now, we calculate the Rademacher complexity of $H=\{\mathbf{1}(\widehat{\mathcal{F}}(x)\neq y)\|\mathcal{F}\in\mathcal{H}_{W,D}\}$:

Lemma A.10. *Let $H=\{\mathbf{1}(\widehat{\mathcal{F}}(x)\neq y):[0,1]^n\times[m]\to\{0,1\}\|\mathcal{F}\in\mathcal{H}_{W,D}\}$. Then, for any distribution $\mathcal{D}$, we have ${\rm Rad}^{\mathcal{D}}_N(H)=O(\sqrt{\frac{mW^2D^2}{N}})$.*

Proof. Let $H^0_{W,D}$ be defined as: $\mathcal{F} \in H^0_{W,D}$ if and only if $\mathcal{F}$ is a network with width W and depth D, and the activation function of each hidden layer of $\mathcal{F}$ is Relu, the output layer does not use the activation function. And let $H^0 = \{\mathbf{1}(\widehat{\mathcal{F}}(x) \neq y) \| \mathcal{F} \in H^0_{W,D}\}$.

Let $H^1_{W,D}$ be defined as: $\mathcal{F} \in H^1_{W,D}$ if and only if $\mathcal{F}$ is a network with width W and depth D, and the activation function of each hidden layer of $\mathcal{F}$ is Relu, the output layer uses the activation function sign.

Then we have that $H^0 = \{\mathbf{1}(\widehat{\mathcal{F}}(x) \neq y) \| \mathcal{F} \in H^0_{W,D}\} = \{\mathbf{1}(\widehat{\mathcal{F}}(x) \neq y) \| \mathcal{F} \in \mathcal{H}_{W,D}\} = H$ and $\Pi_{H^1_{W,D}}(N) \leq (eN)^{O(D^2W^2)}$ by Lemmas A.7 and A.6.

For any $\mathcal{F} \in H^0_{W,D}$, let $\mathcal{F}_{i,j} = \text{sign}(\mathcal{F}_i - \mathcal{F}_j)$, where $\mathcal{F}_i$ is the i-th weight of $\mathcal{F}$. Then it is easy to see that, $\mathcal{F}_{i,j} \in H^1_{W,D}$. Since $\mathbf{1}(\widehat{\mathcal{F}}(x) \neq y) = \mathbf{1}(-\sum_{j \neq y} \mathcal{F}_{y,j} + (m-1) - 0.1)$. By Lemma A.9, the growth function of H^0 is $(eN)^{mO(D^2W^2)}$; so by Lemma A.8 and $H^0 = H$, the Rademacher complexity $\text{Rad}^{\mathcal{D}}_N(H)$ of H is $O(\sqrt{\frac{mD^2W^2}{N}})$ (ignore minor items). This proves the lemma. □

A.4. Proof for Theorem 4.1

Now we prove Theorem 4.1 by using Sections A.2 and A.3:

Proof. Taking $H = \{\mathbf{1}(\mathcal{F}(x) \neq y) \| \mathcal{F} \in \mathcal{H}_{W,D}\}$ and substituting the Rademacher complexity in Section A.3 into Theorem A.2, we prove Theorem 4.1. □

B. Poison Impact Empirical Errors

In this appendix, we prove a proposition to support Remark 4.3. In Proposition B.2 below, we show that if the poison $\mathcal{P}(x)$ is not satisfactory, then it can result in a big empirical error.

Remark B.1. Please note that the conclusion "poison implies big empirical error" only holds in some situations. In fact, when $\mathcal{P}(x)$ is bounded, or the network in the hypothesis space has a strong expressive ability, then the conditions for the proposition in this section will not hold, so the poison will not lead to big empirical error. In our experimental result in Section 6, there is no need to consider the occurrence of large empirical error, because we bound the trigger $\mathcal{P}(x)$ and use a large network.

Let $\mathcal{D}^{poi}$ be the distribution of poisoned data with label l_p, that is,

$$\mathbb{P}_{(x,y)\sim\mathcal{D}_S^{l_p}}(x + \mathcal{P}(x) \in A) = \mathbb{P}_{x\sim\mathcal{D}^{poi}}(x \in A)$$

for any set A.

Proposition B.2. *Use the notation introduced in Theorem A.2 and let $\eta \leq 0.5$ be defined in* (3). *We further assume that $h(x, y_1) + h(x, y_2) \geq 1$ for any $h \in H$, $x \in \mathbb{R}^n$, $y_1 \neq y_2$, and for some $\tau, V \in \mathbb{R}_+$, it holds:*
(1) $\mathbb{P}_{x\sim\mathcal{D}^{poi}}(x \in A) \geq \tau\mathbb{P}_{(x,y)\sim\mathcal{D}_S^{\neq l_p}}(x \in A)$ for any set A;
(2) $\text{Rad}_N^{\mathcal{D}_S^{\neq l_p}}(H) \leq V$.
Let $Q = \eta^N\tau^{N\alpha}$, if $\alpha - \delta/Q > 0$ and $0.5 - 2\delta\alpha/Q - V\alpha > 0$. Then with probability δ, for any $h \in H$ we have

$$\mathbb{E}_{(x,y)\in\mathcal{D}_P}[h(x,y)] > \frac{0.5 - 2\delta\alpha/Q - V\alpha}{\alpha - \delta/Q}.$$

First, we have the following lemma (Bertsekas & Tsitsiklis, 2008).

Lemma B.3. *If distributions D_1, D_2 and function h satisfy $\mathbb{P}_{x\sim D_1}(h(x) \in A) \leq \lambda\mathbb{P}_{x\sim D_2}(h(x) \in A)$ for any set A, then $\mathbb{E}_{x\sim D_1}[f(h(x))] \leq \lambda\mathbb{E}_{x\sim D_2}[f(h(x))]$ for any bounded and positive measurable function $f(x)$.*

Then we prove the following lemma, which directly lead to Proposition B.2.

Lemma B.4. *Use the notations in Proposition B.2. We further assume that $h(x, y_1) + h(x, y_2) \geq 1$ for any $h \in H$, $x \in \mathbb{R}^n$ and $y_1 \neq y_2$, and for some $\tau, V \in \mathbb{R}_+$, it holds:*

(1): $\mathbb{P}_{x\sim\mathcal{D}^{poi}}(x\in A)\geq\tau\mathbb{P}_{(x,y)\sim\mathcal{D}_S^{\neq l_p}}(x\in A)$ *for all sets* A*, where* $\tau\in\mathbb{R}_+$*;*
(2): with probability $1-\delta$*, there exists an* $h\in H$ *such that* $\mathbb{E}_{(x,y)\in\mathcal{D}_P}[h(x,y)]\leq\varepsilon$*.*
Then for any $K\leq N\alpha$*, we have*

$$\begin{array}{l}\mathrm{Rad}_K^{\mathcal{D}_S^{\neq l_p}}(H)\geq 0.5-N\varepsilon/K-\frac{(2-N\varepsilon/K)\delta\alpha}{\eta^N\tau^K}.\\ \mathrm{Rad}_N^{\mathcal{D}_S^{\neq l_p}}(H)\geq K(0.5-N\varepsilon/K-\frac{(2-N\varepsilon/K)\delta\alpha}{\eta^N\tau^K})/N.\end{array}$$

Proof. In order to calculate the probability easily in the proof, we treat $\mathcal{D}_{\rm tr}$ and $\mathcal{D}_P$ as ordered sets. Let (x_i,y_i) be the i-th element of $\mathcal{D}_{\rm tr}$. $\mathcal{D}_P$ is obtained from $\mathcal{D}_{\rm tr}$ as follows: Let $\mathcal{D}_P=\mathcal{D}_{\rm tr}$ first and if poison $\mathcal{P}(x_i)$ is added to (x_i,y_i) for some i, then replace the i-th element of $\mathcal{D}_{\rm tr}$ by $(x_i+\mathcal{P}(x_i),y_i)$.

First, we give three notations:

(1): For the poisoned training set $\mathcal{D}_P$ and $q\in\mathbb{N}$, we take the first q clean samples without l_p in $\mathcal{D}_P$ to form a subset $F_{clean}(q,\mathcal{D}_p)$, that is, if we write $\mathcal{D}_{\rm tr}=\{(x_i,y_i)\}_{i=1}^N$, then $F_{clean}(q,\mathcal{D}_P)=\{(x_{i_k},y_{i_k})\}_{k=1}^q$, where $\{i_k\}_{i=1}^q$ is the smallest q numbers in $[N]$ that satisfy $y_{i_k}\neq l_p$, $(x_{i_k},y_{i_k})\in\mathcal{D}_{clean}$, and $i_a<i_b$ if $a<b$.

(2): For the poisoned training set $\mathcal{D}_P$ and $q\in\mathbb{N}$, we take the first q poisoned samples in $\mathcal{D}_P$ to form the subset $F_{poison}(q,\mathcal{D}_p)$, that is, if we write $\mathcal{D}_{\rm tr}=\{(x_i,y_i)\}_{i=1}^N$, then $F_{poison}(q,\mathcal{D}_P)=\{(x_{i_k}+\mathcal{P}(x_{i_k}),y_{i_k})\}_{k=1}^q$, $\{i_k\}_{i=1}^q$ is the smallest q numbers in $[N]$ that satisfy $y_{i_k}=l_p$, $(x_{i_k}+\mathcal{P}(x_{i_k}),y_{i_k})\in\mathcal{D}_{poi}$, and $i_a<i_b$ if $a<b$.

(3): For any given K samples $\{(x_i,y_i)\}_{i=1}^K$ where $y_i\neq l_p$ and a given set $S\subset[K]$, let S_i be the i-th minimum number in S, in particular, $\mathcal{S}_1$ is the minimum number in S and $S_{|S|}$ is the maximum number in S.

Second, we define a property of $\mathcal{D}_P$:

The poisoned training set $\mathcal{D}_P$ obtained by poisoning $\mathcal{D}_{\rm tr}$ is said to be *nice-inclusion* of a new dataset $\mathcal{D}_G$ if $\mathcal{D}_G=\mathcal{D}_{T_p}\cup D_T$ such that $D_T\subset\mathcal{D}_{\rm tr}$, $D_{T_p}\subset\mathcal{D}^{poi}$, and $\min_{h\in H}\frac{1}{|\mathcal{D}_P|}\sum_{(x,y)\in\mathcal{D}_P}h(x,y)\leq\varepsilon$. For convenience, we write these conditions more explicitly as follows.

The poisoned training set $\mathcal{D}_P$ obtained by poisoning $\mathcal{D}_{\rm tr}=\{(x_i',y_i')\}_{i=1}^N$ is said to be *nice-inclusion* of the ordered dataset $\mathcal{D}_G=\{(z_i,l_i)\}_{i=1}^K$ and $S\subset[K]$ satsfying $|[K]\setminus S|=|\mathcal{D}_{poi}|$, if
(e1): Let $F_{clean}(|S|,\mathcal{D}_P)=\{(x_{i_k}',y_{i_k}')\}_{k=1}^{|S|}$ such that $x_{i_k}'=z_{S_k}$ and $y_{i_k}'=l_{S_k}$ for any $k\in[S]$;
(e2): Let $F_{poison}(|[K]\setminus S|,\mathcal{D}_P)=\{(x_{i_k}'+\mathcal{P}(x_{i_k}'),y_{i_k}')\}_{k=1}^{K-|S|}$. There must be $x_{i_k}'+\mathcal{P}(x_{i_k}')=z_{([K]\setminus S)_k}$ for any $k\in[K-|S|]$.
(e3): $\min_{f\in H}\sum_{(x,y)\in\mathcal{D}_P}\frac{h(x,y)}{|\mathcal{D}_P|}\leq\varepsilon$.

We say that the poisoned training set $\mathcal{D}_P$ obtained by poisoning $\mathcal{D}_{\rm tr}=\{(x_i',y_i')\}_{i=1}^N$ is said to be *common-nice-inclusion* of the ordered dataset $\mathcal{D}_G=\{(z_i,l_i)\}_{i=1}^K$ and $S\subset[K]$ satisfying $|[K]\setminus S|=|\mathcal{D}_{poi}|$, if (e1) and (e2) hold.

Now, we define some functions:

Let $v_i\in\{-1,1\}$ for $i\in[K]$ and $S((v_i)_{i=1}^K)=\{i\|i\in[K],v_i<0\}$. Given K samples $\{(x_i,y_i)\}_{i=1}^K$ where $y_i\neq l_p$, we define that $S_i(\{(x_i,y_i)\}_{i=1}^K,S((v_i)_{i=1}^K))=1$, if there is a *nice-inclusion* poison set $\mathcal{D}_P$ of $\{(x_i,y_i)\}_{i=1}^K$ and $S((v_i)_{i=1}^K)$; otherwise $S_i(\{(x_i,y_i)\}_{i=1}^K,S((v_i)_{i=1}^K))=0$. Then we have the following results.

Result one: If $S_i((x_i,y_i)_{i=1}^K,S((v_i)_{i=1}^K))=1$ and $y_i\neq l_p$, then $\sup_{f\in H}\sum_{i\in[K]}v_ih(x_i,y_i)\geq\sum_{i=1}^K\mathbf{1}(v_i>0)-N\varepsilon$.

Let $\mathcal{D}_P$ be a nice-inclusion of $(x_i,y_i)_{i=1}^K,S((v_i)_{i=1}^K)$. Then

$$\begin{array}{rl} & \varepsilon\\ \geq & \min_{f\in H}\sum_{(x,y)\in\mathcal{D}_P}\frac{h(x,y)}{|\mathcal{D}_P|}(by\ (e3))\\ \geq & \min_{f\in H}(\sum_{i\in S((v_i)_{i=1}^K)}h(x_i,y_i))/|\mathcal{D}_P|+(\sum_{i\in[K]/S((v_i)_{i=1}^K)}h(x_i,l_p))/|\mathcal{D}_P|(by\ (e1,e2))\\ \geq & \min_{f\in H}(\sum_{i\in S((v_i)_{i=1}^K)}h(x_i,y_i))/|\mathcal{D}_P|+(\sum_{i\in[K]/S((v_i)_{i=1}^K)}1-h(x_i,y_i))/|\mathcal{D}_P|(by\ y_i\neq l_p)\\ = & \min_{f\in H}(|[K]/S((v_i)_{i=1}^K)|-\sum_{i\in[K]}v_ih(x_i,y_i))/|\mathcal{D}_P|\\ = & (\sum_{i=1}^K\mathbf{1}(v_i>0)-\sup_{f\in H}\sum_{i\in[K]}v_ih(x_i,y_i))/|\mathcal{D}_P|\\ = & (\sum_{i=1}^K\mathbf{1}(v_i>0)-\sup_{f\in H}\sum_{i\in[K]}v_ih(x_i,y_i))/N.\end{array}$$

The result is proved.

Result Two: In order to give the lower bound of the Rademacher complexity $\text{Rad}_K^{\mathcal{D}^l}(H)$, we just need to consider the upper bound of $\mathbb{E}_{(x_i,y_i)\sim\mathcal{D}_S^{\neq l_p}}[\mathbf{1}(S_i(\{(x_i,y_i)\}_{i=1}^K, S((\sigma_i)_{i=1}^K)) \neq 1)]$ for each $(\sigma_i)_{i=1}^K \subset \{-1,1\}^K$.

Let σ_i be Rademacher random variables, that is $P(\sigma_i = 1) = P(\sigma_i = -1) = 0.5$. Then, by the definition of Rademacher complexity, we have

$$\begin{array}{cl}
& \text{Rad}_K^{\mathcal{D}^l}(H) \\
= & \mathbb{E}_{(x_i,y_i)\sim\mathcal{D}_S^{\neq l_p}}[\mathbb{E}_{\sigma_i}[\sup_{f\in H} \frac{\sum_{i=1}^K \sigma_i h(x_i,y_i)}{K}]] \\
= & \mathbb{E}_{(x_i,y_i)\sim\mathcal{D}_S^{\neq l_p}}[\sum_{\sigma_i} \sup_{f\in H} \frac{\sum_{i=1}^K \sigma_i h(x_i,y_i)}{2^K K}] \\
\geq & \mathbb{E}_{(x_i,y_i)\sim\mathcal{D}_S^{\neq l_p}}[\sum_{\sigma_i} \mathbf{1}(S_i(\{(x_i,y_i)\}_{i=1}^K, S((\sigma_i)_{i=1}^K)) = 1)\frac{\sum_{i=1}^K \mathbf{1}(\sigma_i>0)-N\varepsilon}{2^K K} \\
& -\mathbf{1}(S_i(\{(x_i,y_i)\}_{i=1}^K, S((\sigma_i)_{i=1}^K)) \neq 1)\frac{1}{2^K}] \\
\geq & \mathbb{E}_{(x_i,y_i)\sim\mathcal{D}_S^{\neq l_p}}[\sum_{\sigma_i} \frac{\sum_{i=1}^K \mathbf{1}(\sigma_i>0)-N\varepsilon}{2^K K} - \mathbf{1}(S_i(\{(x_i,y_i)\}_{i=1}^K, S((\sigma_i)_{i=1}^K)) \neq 1)\frac{2K-N\varepsilon}{2^K K}] \\
= & 0.5 - N\varepsilon/K - \mathbb{E}_{(x_i,y_i)\sim\mathcal{D}_S^{\neq l_p}}[\sum_{\sigma_i} \mathbf{1}(S_i(\{(x_i,y_i)\}_{i=1}^K, S((\sigma_i)_{i=1}^K)) \neq 1)\frac{2K-N\varepsilon}{2^K K}].
\end{array}$$

The first inequality uses Result one. So, if we want to give a lower bound of $\text{Rad}_K^{\mathcal{D}^l}(H)$, we just need to give an upper bound of $\mathbb{E}_{(x_i,y_i)\sim\mathcal{D}_S^{\neq l_p}}[\sum_{\sigma_i} \mathbf{1}(S_i(\{(x_i,y_i)\}_{i=1}^K, S((\sigma_i)_{i=1}^K)) \neq 1)]$. Furthermore, we just need to consider $\mathbb{E}_{(x_i,y_i)\sim\mathcal{D}_S^{\neq l_p}}[\mathbf{1}(S_i(\{(x_i,y_i)\}_{i=1}^K, S((\sigma_i)_{i=1}^K)) \neq 1)]$ for each $(\sigma_i)_{i=1}^K$. Result two is proved.

Result Three: Now, we prove $\mathbb{E}_{(x_i,y_i)\sim\mathcal{D}_S^{\neq l_p}}[\mathbf{1}(S_i(\{(x_i,y_i)\}_{i=1}^K, S((v_i)_{i=1}^K)) \neq 1)] < \frac{\delta}{\eta^N\tau^K}$ for any $(v_i)_{i=1}^K \in \{-1,1\}^K$.

For a given $(v_i)_{i=1}^K \in \{-1,1\}^K$, let set $C_{(v_i)_{i=1}^K} = \{\{(x_i,y_i)\}_{i=1}^K | S_i(\{(x_i,y_i)\}_{i=1}^K, S((v_i)_{i=1}^K)) \neq 1, y_i \neq l_p\}$, then $\mathbb{E}_{(x_i,y_i)\sim\mathcal{D}_S^{\neq l_p}}[\mathbf{1}(S_i(\{(x_i,y_i)\}_{i=1}^K, S((v_i)_{i=1}^K)) \neq 1)] = \mathbb{E}_{(x_i,y_i)\sim\mathcal{D}_S^{\neq l_p}}[\mathbf{1}(\{(x_i,y_i)\}_{i=1}^K \in C_{(v_i)_{i=1}^K})]$.

For $(v_i)_{i=1}^K \in \{-1,1\}^K$ and $\{(x_i,y_i)\}_{i=1}^K$, let $\mathbb{E}_{(v_i)_{i=1}^K}^{\{(x_i,y_i)\}_{i=1}^K}$ be the set of $\mathcal{D}_P$, which is a *common-nice-inclusion* for $\{(x_i,y_i)\}_{i=1}^K$ and $S((v_i)_{i=1}^K)$. It is easy to see that:
(r1): $\mathbb{E}_{(v_i)_{i=1}^K}^{((x_i,y_i))_{i=1}^K} \cap \mathbb{E}_{(v'_i)_{i=1}^K}^{((x'_i,y'_i))_{i=1}^K} = \phi$ when $v_i \neq v'_i$ for some $i \in [K]$ or $x_i \neq x'_i$ for some $i \in [K]$.
(r2): If $\mathcal{D}_P \in \mathbb{E}_{(v_i)_{i=1}^K}^{\{(x_i,y_i)\}_{i=1}^K}$ satisfies (e3), then $S_i(\{(x_i,y_i)\}_{i=1}^K, S((v_i)_{i=1}^K)) = 1$.

So by (r2), if $\{(x_i,y_i)\}_{i=1}^K \in C_{(v_i)_{i=1}^K}$, then for any $\mathcal{D}_P \in \mathbb{E}_{(v_i)_{i=1}^K}^{\{(x_i,y_i)\}_{i=1}^K}$, (e3) cannot stand. Let the set B contain all the $\mathcal{D}_P$ that do not satisfy (e3). Then, if $\{(x_i,y_i)\}_{i=1}^K \in C_{(v_i)_{i=1}^K}$, there must be $\mathbb{E}_{(v_i)_{i=1}^K}^{(x_i,y_i)_{i=1}^K} \subset B$.

Now we prove the following two results:

Result S1: $\int_{\mathcal{D}_P} \mathbf{1}(D \in \bigcup_{\{(x_i,y_i)\}_{i=1}^K \in C_{(v_i)_{i=1}^K}} \mathbb{E}_{(v_i)_{i=1}^K}^{\{(x_i,y_i)\}_{i=1}^K}) \mathrm{d}D < \delta$.

By (r1) and $\mathbb{E}_{(v_i)_{i=1}^K}^{(x_i,y_i)_{i=1}^K} \subset B$ for all $\{(x_i,y_i)\}_{i=1}^K \in C_{(v_i)_{i=1}^K}$, we have that:

$$\mathbf{1}(D \in B) \geq \mathbf{1}(\mathcal{D} \in \bigcup_{\{(x_i,y_i)\}_{i=1}^K \in C_{(v_i)_{i=1}^K}} \mathbb{E}_{(v_i)_{i=1}^K}^{\{(x_i,y_i)\}_{i=1}^K})$$

So $\int_{\mathcal{D}_P} \mathbf{1}(\mathcal{D} \in \bigcup_{\{(x_i,y_i)\}_{i=1}^K \in C_{(v_i)_{i=1}^K}} \mathbb{E}_{(v_i)_{i=1}^K}^{\{(x_i,y_i)\}_{i=1}^K}) \mathrm{d}D \leq \int_{\mathcal{D}_P} \mathbf{1}(\mathcal{D} \in B) \mathrm{d}D \leq \delta$, using condition (2) of the theorem here.

Result S2: $\int_{\mathcal{D}_P} \mathbf{1}(D \in \bigcup_{\{(x_i,y_i)\}_{i=1}^K \in C_{(v_i)_{i=1}^K}} \mathbb{E}_{(v_i)_{i=1}^K}^{\{(x_i,y_i)\}_{i=1}^K}) \mathrm{d}D \geq \mathbb{E}_{(x_i,y_i)\sim\mathcal{D}_S^{\neq l_p}}[\mathbf{1}((x_i,y_i) \in C_{(v_i)_{i=1}^K})] \eta^N \tau^K / \alpha$.

Consider the definition of F_{clean} and $\mathcal{F}_{poison}$ in (e1) and (e2). When $D_p \in \mathbb{E}_{(v_i)_{i=1}^K}^{\{(x_i,y_i)\}_{i=1}^K}$, the first $\sum_{i=1}^K \mathbf{1}(v_i < 0)$ samples without label l_p in $\mathcal{D}_{\text{tr}}$ must be $\{(x_{i_k}, y_{i_k})\}$, where i_k satisfies $v_{i_k} = -1$; and $\mathcal{D}_{poi} = \{(x_{i_k}, y_{i_k})\}$, where i_k satisfies

$v_{i_k}=1$. Let $N_0=\sum_{i=1}^K \mathbf{1}(v_i<0)$, then

$$\begin{array}{rl} & \int_{\mathcal{D}_P} \mathbf{1}(D \in \bigcup_{\{(x_i,y_i)\}_{i=1}^K \in C_{(v_i)_{i=1}^K}} \mathbb{E}_{(v_i)_{i=1}^K}^{\{(x_i,y_i)\}_{i=1}^K}) \mathrm{d}D \\ = & \int_{(\mathcal{D}_S^{\neq l_p})^{N_0}(\mathcal{D}_{poi})^{K-N_0}} \mathbf{1}(\{(x_i,y_i)\} \in C_{(v_i)_{i=1}^K})(\sum_{q=0}^N \mathbf{1}([q\alpha]=K-N_0) C_N^q \eta^q (1-\eta)^{N-q}) \mathrm{d}(x_i,y_i) \\ \geq & \int_{(\mathcal{D}_S^{\neq l_p})^{N_0}(\mathcal{D}_{poi})^{K-N_0}} \mathbf{1}(\{(x_i,y_i)\} \in C_{(v_i)_{i=1}^K})(\sum_{q=0}^N \mathbf{1}([q\alpha]=K-N_0) \eta^N) \mathrm{d}(x_i,y_i) \\ \geq & \int_{(\mathcal{D}_S^{\neq l_p})^{N_0}(\mathcal{D}_{poi})^{K-N_0}} \mathbf{1}(\{(x_i,y_i)\} \in C_{(v_i)_{i=1}^K})(\eta^N/\alpha) \mathrm{d}(x_i,y_i) \\ \geq & \int_{(\mathcal{D}_S^{\neq l_p})^K} \mathbf{1}((x_i,y_i) \in C_{(v_i)_{i=1}^K}) \eta^N \tau^K/\alpha \mathrm{d}(x_i,y_i) \\ = & \mathbb{E}_{(x_i,y_i)\sim\mathcal{D}_S^{\neq l_p}}[\mathbf{1}((x_i,y_i) \in C_{(v_i)_{i=1}^K})] \eta^N \tau^K/\alpha. \end{array}$$

The first inequality uses $\eta \leq 0.5$ and $C_N^i \geq 1$. The second inequality uses at least $1/\alpha$ numbers of $q \in [N]$ such that $[q\alpha]=K-N_0$. The last inequality uses Lemma B.3 and condition (1) in theorem. This proves Result S2.

Finally, by Result S1 and Result S2, we have $(\eta^N\tau^K/\alpha)\mathbb{E}_{(x_i,y_i)\sim\mathcal{D}_S^{\neq l_p}}[\mathbf{1}(\{(x_i,y_i)\}_{i=1}^K \in C_{(v_i)_{i=1}^K})] \leq \int_{\mathcal{D}_P} \mathbf{1}(D \in \bigcup_{\{(x_i,y_i)\}_{i=1}^K \in C_{(v_i)_{i=1}^K}} \mathbb{E}_{(v_i)_{i=1}^K}^{\{(x_i,y_i)\}_{i=1}^K}) \mathrm{d}D \leq \delta$, that is, $\mathbb{E}_{(x_i,y_i)\sim\mathcal{D}_S^{\neq l_p}}[\mathbf{1}(\{(x_i,y_i)\}_{i=1}^K \in C_{(v_i)_{i=1}^K})] \leq \alpha\delta/(\eta^N\tau^K)$.

Prove the lemma. Use Results three and two, we have

$$\begin{array}{rl} & \mathrm{Rad}_K^{\mathcal{D}_{\mathcal{S}}^{\neq l_p}}(H) \\ \geq & 0.5 - N\varepsilon/K - \mathbb{E}_{(x_i,y_i)\sim\mathcal{D}_S^{\neq l_p}}[\sum_{\sigma_i} \mathbf{1}(S_i(((x_i,y_i))_{i=1}^K, S((\sigma_i)_{i=1}^K)) \neq 1)\frac{2K-N\varepsilon}{2^K K}] \\ \geq & 0.5 - N\varepsilon/K - \frac{(2-N\varepsilon/K)\delta\alpha}{\eta^N\tau^K}. \end{array}$$

Since $\mathrm{Rad}_M^D(H) \leq \frac{N}{M}\mathrm{Rad}_N^D(H)$ for any $M \leq N$ and distribution D, the lemma is proved. □

Now we use Lemma B.4 to prove Proposition B.2:

Proof. Use reduction to absurdity. Assume that with probability at least $1-\delta$, there is $\mathbb{E}_{(x,y)\in\mathcal{D}_P}[h(x,y)] \leq \frac{0.5-2\delta\alpha/Q-V\alpha}{\alpha-\delta/Q}$, and take the right-hand size value as ϵ.

By Lemma B.4, take $K=N\alpha$. Then $\mathrm{Rad}_N^{\mathcal{D}_{\mathcal{S}}^{\neq l_p}} \geq (0.5-N\epsilon/K-\frac{(2-N\epsilon/K)\delta\alpha}{\eta^N\tau^K})/\alpha$. We substitute ϵ in it. Then

$$\begin{array}{rl} & \mathrm{Rad}_N^{\mathcal{D}_{\mathcal{S}}^{\neq l_p}} \\ \geq & (0.5-N\epsilon/K-\frac{(2-N\epsilon/K)\delta\alpha}{\eta^N\tau^K})/\alpha \\ = & 1/\alpha(0.5-\frac{2\delta\alpha}{Q}-\epsilon(\alpha-\delta/Q)) \\ = & 1/\alpha(0.5-\frac{2\delta\alpha}{Q}-\frac{0.5-2\delta\alpha/Q-V\alpha}{\alpha-\delta/Q}(\alpha-\delta/Q)) \\ = & 1/\alpha(0.5-\frac{2\delta\alpha}{Q}-(0.5-2\delta\alpha/Q-V\alpha)) \\ = & 1/\alpha(V\alpha) \\ = & V. \end{array}$$

So $\mathrm{Rad}_N^{\mathcal{D}_{\mathcal{S}}^{\neq l_p}} \geq V$, which is contradictory to (2) in Proposition B.2, and the proposition is proved. □

C. Optimality of the Generalization Bound

In this section, we show that for the general hypothesis space and data distribution, the generalization gap between the empirical error and the population error cannot be smaller than $O(\frac{1}{\sqrt{N}})$, mentioned in Remark 4.4.

Proposition C.1. *Let $m=2$ and the data distribution $\mathcal{D}_{\mathcal{S}}$ satisfy $P_{(x,y)\sim\mathcal{D}_{\mathcal{S}}}(y=1)=P_{(x,y)\sim\mathcal{D}_{\mathcal{S}}}(y=0)=0.5$. Let $\mathcal{H}=\{L(\mathcal{F}(x),y)\}$ be the hypothesis space, $L(\mathcal{F}(x),y)=\mathbf{1}(\widehat{\mathcal{F}}(x)\neq y)$ the loss function, and $\mathcal{F}_0$ a neural network classifying $\widehat{\mathcal{F}}_0(x)=0$ for $(x,y)\sim\mathcal{D}_{\mathcal{S}}$. Then for any $c>0$, $k>0.5$, and $\mathcal{D}_{\mathrm{tr}}$ i.i.d. sampled from $\mathcal{D}_{\mathcal{S}}$ with $|\mathcal{D}_{\mathrm{tr}}|=N$, we have*

$$\mathbb{P}(|\mathbb{E}_{(x,y)\sim\mathcal{D}_S}[L(\mathcal{F}_0(x),y)]-\mathbb{E}_{(x,y)\in\mathcal{D}_{\mathrm{tr}}}[L(\mathcal{F}_0(x),y)]|<\frac{c}{N^k})=O(cN^{0.5-k}).$$

Proof. First, we have $\mathbb{E}_{(x,y)\sim\mathcal{D}_S}[L(\mathcal{F}_0(x),y)] = \mathbb{E}_{(x,y)\sim\mathcal{D}_S}[\mathbf{1}(y=1)] = 0.5$. Then we have $\mathbb{E}_{(x,y)\in\mathcal{D}_{\rm tr}}[L(\mathcal{F}_0(x),y)] = \mathbb{E}_{(x,y)\in\mathcal{D}_{\rm tr}}[\mathbf{1}(y=1)] = N_1/N$, where N_1 is the number of x with label 1 in $\mathcal{D}_{\rm tr}$. So, $\mathbb{P}(|\mathbb{E}_{(x,y)\sim\mathcal{D}_S}[L(\mathcal{F}_0(x),y)] - \mathbb{E}_{(x,y)\in\mathcal{D}_{\rm tr}}[L(\mathcal{F}_0(x),y)]| < \frac{c}{N^k}) = \mathbb{P}(|0.5 - N_1/N| < \frac{c}{N^k})$.

To estimate $\mathbb{P}(|0.5 - N_1/N| < \frac{c}{N^k})$, we only need to calculate the probability of $N_1 \in (0.5N - cN^{1-k}, 0.5N + cN^{1-k})$. Since $\mathcal{D}_{\rm tr}$ is i.i.d. sampled from $\mathcal{D}_{\mathcal{S}}$, a sample labeled 1 is selected with probability 0.5. Thus,

$$\begin{array}{rl} & \mathbb{P}(N_1 \in (0.5N - cN^{1-k}, 0.5N + cN^{1-k})) \\ \leq & \sum_{i=0.5N-cN^{1-k}}^{0.5N+cN^{1-k}} C_N^i 0.5^N \\ \leq & 2cN^{1-k} C_N^{N/2} 0.5^N. \end{array}$$

When $N \to \infty$, using Stirling's approximation to calculate the $C_N^{N/2}$, we have

$$\begin{array}{rl} & \mathbb{P}(N_1 \in (0.5N - cN^{1-k}, 0.5N + cN^{1-k})) \\ \leq & 2cN^{1-k} C_N^{N/2} 0.5^N \\ = & 2cN^{1-k} 0.5^N O(\frac{\sqrt{2\pi N}(N/e)^N}{\pi N (N/(2e))^N}) \\ = & O(cN^{0.5-k}). \end{array}$$

The proposition is proved. □

It is easy to see that $O(cN^{0.5-k}) \to 0$ when $N \to \infty$, so by Proposition C.1, the generalization gap cannot be smaller than $O(\frac{1}{\sqrt{N}})$. Together with Theorem A.1, we show the optimality of $O(\frac{1}{\sqrt{N}})$ of the generalization gap for a clean dataset.

This is also for the generalization bound under poison attacks, because the proof of Theorem 4.1 need the generalization bound Theorem A.1 for the dataset.

D. Proof of Theorem 4.5

We provide a more intuitive explanation on how to estimate the poison generalization bound in Theorem 4.5, which is the core of our theorem.

Based on research on indiscriminate poisoning, neural networks always prioritize learning simple features, i.e., shortcut. So we try to make the trigger to be a shortcut (as said in condition (c3) in Theorem 4.5 and Proposition 4.10).

However, only a few samples were poisoned in the backdoor attack, which is different from the setting of indiscriminate poisoning, so only using shortcut cannot establish the bound. Therefore, we aim to disrupt the original features of the poisoned images, such that the classification of the poisoned images and the classification of clean images become two independent tasks for the network (as said in condition (c1) in Theorem 4.5). By doing so, when the network completes the task of classifying poison data, the clean part of the data set is useless, and the network will learn the feature in the poison part of data, so that the shortcut can be learned.

Finally, if the shortcuts are similar for different images, the shortcuts learned on a small portion of the data can be generalized to all the data (as said in condition (c2) in Theorem 4.5).

By the above idea, we will prove a more general form of Theorem 4.5, as shown below, and Theorem 4.5 is an easy corollary of this theorem.

Theorem D.1. *Use the notation in Section 3. Let* $N = |\mathcal{D}_{\rm tr}|$. *For any two hypothesis spaces* $\mathcal{H} \subset \mathcal{H}_{W,D}$ *and* $F \subset \mathcal{H}_{W,D}$, *if* $\mathcal{P}(x)$ *satisfies the following conditions for some* $\epsilon > 0, \tau > 0, \lambda \geq 1$*:*
(c1): $\mathbb{E}_{(x,y)\sim\mathcal{D}_S^{l_p}}[f_y(x+\mathcal{P}(x))] \leq \epsilon$ *for all* $f \in F$,
(c2): $\mathbb{P}_{(x,y)\sim\mathcal{D}_{\mathcal{S}}}(\mathcal{P}(x) \in A | y \neq l_p) \leq \lambda\, \mathbb{P}_{(x,y)\sim\mathcal{D}_{\mathcal{S}}}(\mathcal{P}(x) \in A | y = l_p)$ *for any set A,*
(c3): some $h \in H$ *satisfies that:* $\exists f \in F$ *such that* $\mathbb{E}_{x\sim\mathcal{D}_{\mathcal{S}}}[|(h-f)_{l_P}(\mathcal{P}(x)) - (h-f)_{l_p}(x+\mathcal{P}(x))|] \leq \tau$, *where* $(h-f)_{l_P}(x) = h_{l_P}(x) - f_{l_p}(x)$,
then with probability at least $1-\delta-O(1/N)$, *the following inequality holds for all* $h \in H$ *satisfying (c3):*

$$\mathcal{E}_P(h,\mathcal{D}_{\mathcal{S}}) \leq \lambda O(\tfrac{1}{\alpha}(\mathbb{E}_{(x,y)\in\mathcal{D}_P}[L_{CE}(h(x),y)] + {\rm Rad}_N^{\mathcal{D}_S^{l_p}}(\mathcal{H}_{W,D,1})) + \sqrt{\tfrac{\ln(1/\delta)}{N\alpha}} + \epsilon + \tau + \tfrac{\lambda-1}{\lambda}). \qquad (11)$$

It is easy to see that we just need to take the hypothesis spaces H, F in Theorem D.1 to be $H = \{\mathcal{F}(x)\}$ and $F = \{\mathcal{G}(x)\}$ (i.e. hypothesis space just contains only one network). Then Theorem D.1 naturally equivalent to Theorem 4.5.

D.1. Proof of Theorem D.2

We give the following theorem, which is a generalization bound theory under more general hypothesis space.

Theorem D.2. *For any two hypothesis spaces $H = \{h(x, y) \in \mathcal{S} \times [m] \to [0, 1]\}, F = \{f(x, y) \in \mathcal{S} \times [m] \to [0, 1]\}$, if $\mathcal{P}(x)$ satisfies the following conditions for some $\epsilon > 0, \tau > 0, \lambda \geq 1$:*
(1) $\mathbb{E}_{(x,y)\sim\mathcal{D}_{\mathcal{S}}^{l_p}}[f(x + \mathcal{P}(x), y)] \geq 1 - \epsilon$ for any $f \in F$,
(2) $\mathbb{P}_{(x,y)\sim\mathcal{D}_{\mathcal{S}}^{\neq l_p}}(\mathcal{P}(x) \in A) \leq \lambda\mathbb{P}_{(x,y)\sim\mathcal{D}_{\mathcal{S}}^{l_p}}(\mathcal{P}(x) \in A)$ for any set A,
(3) Some $h \in H$ satisfies that there exists an $f \in F$ such that $\mathbb{E}_{x\sim\mathcal{D}_{\mathcal{S}}}[|(h - f)(\mathcal{P}(x), l_p) - (h - f)(x + \mathcal{P}(x), l_p)|] \leq \tau$, where $(h - f)(x) = h(x) - f(x)$,
then with probability at least $1 - \delta - \frac{4-4\eta}{4-4\eta+\eta N}$, for any $h \in H$ satisfying (3), we have:

$$\mathbb{E}_{(x,y)\sim\mathcal{D}_S}[h(x + \mathcal{P}(x), l_p)] \leq \lambda(2\mathbb{E}_{(x,y)\in\mathcal{D}_P}h(x, y)/(\alpha\eta) + 2\text{Rad}_{N\alpha\eta/2}^{\mathcal{D}_{\mathcal{S}}^{l_p}}(H) + \sqrt{\tfrac{\ln(2/\delta)}{N\alpha\eta}} + \tau/\eta + \epsilon) + \tau + (\lambda - 1)\eta. \quad (12)$$

We will use Theorem D.2 to prove Theorem D.1 in Section D.2. Now, we prove Theorem D.2 and give a detailed explanation of the "proof idea" of Theorem 4.5. Please note that, in the proof of Theorem D.2, the poison generalization error is $\mathbb{E}_{(x,y)\sim\mathcal{D}_S}[h(x + \mathcal{P}(x), l_p)]$.

Proof. The proof is divided into two parts.

Part one, we estimate the upper bound of $\mathbb{E}_{(x,y)\sim\mathcal{D}_{\mathcal{S}}^{l_p}}[h(x + \mathcal{P}(x), l_p)]$. This part corresponds to the "firstly" part in the proof idea shown under Theorem 4.5.

We first give three bounds in the following Results (e1), (e2), and (e3). Since (e1), (e2), (e3) are similar to (d1), (d2), (d3) in the proof of Theorem A.2, we omitted the proof.

Result (e1): Let the random variable X be the number of samples with label l_p in $\mathcal{D}_{\mathcal{P}}$. Then with probability at least $1 - \frac{4(1-\eta)}{4-4\eta+N\eta}$, it holds $X \geq N\eta/2$.

Result (e2): Now we even randomly select $N\eta\alpha/2$ poisoned samples in $\mathcal{D}_P$. If the number of poisoned samples in $\mathcal{D}_P$ is smaller than $N\eta\alpha/2$, then we let D_{l_p} be the set of all such samples.

Let $\mathcal{D}_{l_p}$ obey the distribution $\mathcal{D}_{\mathcal{S}2}$. Then we have $\mathbb{P}_{x\sim\mathcal{D}_{\mathcal{S}2}}(x \in A|X \geq [N\eta/2]) = P(X_{N\eta\alpha/2}^{\mathcal{D}_{\mathcal{S}}^{l_p}} \in A)$ for any set A, where $X_{N\eta\alpha/2}^{\mathcal{D}_{\mathcal{S}}^{l_p}}$ is the set that i.i.d. sampled $N\eta\alpha/2$ data from distribution $\mathcal{D}_{\mathcal{S}}^{l_p}$, and add $\mathcal{P}(x)$ to each data.

Result (e3): With probability $1 - \frac{4-4\eta}{4-4\eta+N\eta} - \delta/2$, for any $h \in H$, we have

$$\mathbb{E}_{(x,y)\sim\mathcal{D}_{\mathcal{S}}^{l_p}}[h(x + \mathcal{P}(x), l_p)] \leq \frac{\sum_{(x,y)\in\mathcal{D}_{poi}} h(x, y)}{N\alpha\eta/2} + 2\text{Rad}_{N\alpha\eta/2}^{\mathcal{D}_{\mathcal{S}}^{l_p}}(H) + \sqrt{\frac{\ln(2/\delta)}{N\alpha\eta}}.$$

We use $\mathcal{D}_{poi} = \{(x + \mathcal{P}(x), l_p)||(x, l_p) \in \mathcal{D}_{sub}\}$, Result (e1), and Result (e2) to prove Result (e3). The concrete steps are similar to that of Result (c3). $\mathcal{D}_{poi}$ and $\mathcal{D}_{sub}$ are defined in Section 3.2.

Part two, estimate the upper bound of $\mathbb{E}_{(x,y)\sim\mathcal{D}_S}[h(x + \mathcal{P}(x), l_p)]$.

Please note that when $y \neq l_p$, $h(x + \mathcal{P}(x), l_p)$ will not appear in the empirical error $\mathbb{E}_{(x,y)\in\mathcal{D}_P}[h(x + \mathcal{P}(x), y)]$, so we need to use some other methods to estimate $\mathbb{E}_{(x,y)\sim\mathcal{D}_S}[h(x + \mathcal{P}(x), l_p)]$.

For any $h \in H$ and $f \in F$, let $c_{h,f}(x) = h(x, l_p) - f(x, l_p)$. Let Q be the upper bound mentioned in Result (e3) in Part one.

The upper bound of $\mathbb{E}_{(x,y)\sim\mathcal{D}_{\mathcal{S}}^{\neq l_p}}[h(x + \mathcal{P}(x), l_p)]$ will be given by the following Results (f1), (f2), (f3), and (f4). Note that (f1) and (f2) correspond to the "Secondly" step in the proof idea shown under the Theorem 4.5; (f3) and (f4) correspond to the "Finally" step in the proof idea shown under the Theorem 4.5.

Result (f1): If $f \in F$ and $h \in H$ satisfy (3), then we have $\mathbb{E}_{(x,y)\sim\mathcal{D}_S^{l_p}}[c_{h,f}(\mathcal{P}(x))] \leq Q + \tau/\eta - (1-\varepsilon)$.

By condition (3), we have:

$$\begin{array}{rl} & \tau \\ \geq & \mathbb{E}_{(x,y)\sim\mathcal{D}_S}[|c_{h,f}(\mathcal{P}(x)) - c_{h,f}(x+\mathcal{P}(x))|](use\ (3)) \\ \geq & \eta(\mathbb{E}_{(x,y)\sim\mathcal{D}_S^{l_p}}[|c_{h,f}(\mathcal{P}(x)) - c_{h,f}(x+\mathcal{P}(x))|]) \\ \geq & \eta(\mathbb{E}_{(x,y)\sim\mathcal{D}_S^{l_p}}[c_{h,f}(\mathcal{P}(x)) - c_{h,f}(x+\mathcal{P}(x))]). \end{array} \tag{13}$$

Now by condition (1) and Result (e3), with probability $1-\delta$, we have

$$\begin{array}{rl} & \mathbb{E}_{(x,y)\sim\mathcal{D}_S^{l_p}}[c_{h,f}(x+\mathcal{P}(x))] \\ = & \mathbb{E}_{(x,y)\sim\mathcal{D}_S^{l_p}}[h(x+\mathcal{P}(x),l_p)] - \mathbb{E}_{(x,y)\sim\mathcal{D}_S^{l_p}}[f(x+\mathcal{P}(x),l_p)] \\ \leq & Q-(1-\epsilon). \end{array} \tag{14}$$

Combine inequalities (13) and (14), we have:

$$\begin{array}{rl} & \mathbb{E}_{(x,y)\sim\mathcal{D}_S^{l_p}}[c_{h,f}(\mathcal{P}(x))] \\ \leq & \mathbb{E}_{(x,y)\sim\mathcal{D}_S^{l_p}}[c_{h,f}(x+\mathcal{P}(x))] + \tau/\eta \\ \leq & Q-(1-\epsilon)+\tau/\eta. \end{array} \tag{15}$$

Result (f1) is proved.

Result (f2): $\mathbb{E}_{(x,y)\sim\mathcal{D}_S^{\neq l_p}}[c_{h,f}(\mathcal{P}(x))+1] \leq \lambda\mathbb{E}_{(x,y)\sim\mathcal{D}_S^{l_p}}[c_{h,f}(\mathcal{P}(x))+1]$.

Result (f2) can be proved by using Lemma B.3 and condition (2).

Result (f3): When $h \in H$ and $f \in F$ satisfy condition (3), we have $\mathbb{E}_{\mathcal{D}_S}[c_{h,f}(x+\mathcal{P}(x))] \leq (\eta+(1-\eta)\lambda)(Q+\epsilon+\tau/\eta-1)+\lambda-1+\tau$.

By condition (3), we have

$$\begin{array}{rl} & \tau \\ \geq & \mathbb{E}_{(x,y)\sim\mathcal{D}_S}[|c_{h,f}(\mathcal{P}(x)) - c_{h,f}(x+\mathcal{P}(x))|] \quad . \\ \geq & \mathbb{E}_{(x,y)\sim\mathcal{D}_S}[-c_{h,f}(\mathcal{P}(x)) + c_{h,f}(x+\mathcal{P}(x))] \end{array}$$

Then, we have $\mathbb{E}_{\mathcal{D}_S}[c_{h,f}(x+\mathcal{P}(x))] \leq \mathbb{E}_{\mathcal{D}_S}[c_{h,f}(\mathcal{P}(x))]+\tau$. Now, substitute Results (f1) and (f2) in it to estimate $\mathbb{E}_{\mathcal{D}_S}[c_{h,f}(\mathcal{P}(x))]$, and we can get

$$\begin{array}{rl} & \mathbb{E}_{(x,y)\sim\mathcal{D}_S}[c_{h,f}(\mathcal{P}(x))] \\ = & \eta\mathbb{E}_{(x,y)\sim\mathcal{D}_S^{l_p}}[c_{h,f}(\mathcal{P}(x))] + (1-\eta)\mathbb{E}_{(x,y)\sim\mathcal{D}_S^{\neq l_p}}[c_{h,f}(\mathcal{P}(x))] \\ \leq & (\eta+(1-\eta)\lambda)\mathbb{E}_{(x,y)\sim\mathcal{D}_S^{l_p}}[c_{h,f}(\mathcal{P}(x))]+\lambda-1 \\ \leq & (\eta+(1-\eta)\lambda)(Q+\epsilon+\tau/\eta-1)+\lambda-1. \end{array}$$

So, we prove Result (f3): $\mathbb{E}_{\mathcal{D}_S}[c_{h,f}(x+\mathcal{P}(x))] \leq (\eta+(1-\eta)\lambda)(Q+\epsilon+\tau/\eta-1)+\lambda-1+\tau$.

Result (f4): $\mathbb{E}_{(x,y)\sim\mathcal{D}_S}[h(x+\mathcal{P}(x),l_p)] \leq \lambda(Q+\tau/\eta+\epsilon)+\tau+(\lambda-1)\eta$.

Since $c_{h,f}(x+\mathcal{P}(x)) = h(x+\mathcal{P}(x),l_p) - f(x+\mathcal{P}(x),l_p)$ and $f(x+\mathcal{P}(x),l_p) \leq 1$, using Result (f3), we have

$$\begin{array}{rl} & \mathbb{E}_{(x,y)\sim\mathcal{D}_S}[h(x+\mathcal{P}(x),l_p)] \\ \leq & \mathbb{E}_{(x,y)\sim\mathcal{D}_S}[c_{h,f}(x+\mathcal{P}(x),l_p)]+1 \\ \leq & (\eta+(1-\eta)\lambda)(Q+\epsilon+\tau/\eta-1)+\lambda+\tau \\ \leq & \lambda(Q+\tau/\eta+\epsilon)+\tau+(\lambda-1)\eta. \end{array}$$

This proves Result (f4).

When h satisfies condition (3), using the value of Q into the Result (f4), we see that with probability $1-\delta-\frac{4-4\eta}{4-4\eta+\eta N}$, we have:

$$\mathbb{E}_{(x,y)\sim\mathcal{D}_S}[h(x+\mathcal{P}(x),l_p)] \leq \lambda(\frac{\sum_{(x,y)\in\mathcal{D}_{poi}} h(x,y)}{N\alpha\eta/2} + 2\mathrm{Rad}_{N\alpha\eta/2}^{\mathcal{D}_S^{l_p}}(H) + \sqrt{\frac{\ln(2/\delta)}{N\alpha\eta}} + \tau/\eta + \epsilon) + \tau + (\lambda-1)\eta.$$

Finally, using the facts $\mathcal{D}_{poi} \subset \mathcal{D}_P$ and Result (e3) is valid for $X = |\mathcal{D}_{poi}| \geq [N\eta/2]$, we have

$$\mathbb{E}_{(x,y)\sim\mathcal{D}_S}[h(x+\mathcal{P}(x),l_p)] \leq \lambda(\frac{2\mathbb{E}_{(x,y)\in\mathcal{D}_P}[h(x,y)]}{\alpha\eta} + 2\mathrm{Rad}^{\mathcal{D}_S^{l_p}}_{N\alpha\eta/2}(H) + \sqrt{\frac{\ln(2/\delta)}{N\alpha\eta}} + \tau/\eta + \epsilon) + \tau + (\lambda-1)\eta.$$

The theorem is proved. □

D.2. Proof of Theorem D.1

Now, we use Theorem D.2 to prove Theorem D.1.

First, for any $h : \mathbb{R}^n \to \mathbb{R}^m$, we define $h_{-1}(x,y) = 1 - h_y(x) : \mathbb{R}^n \times [m] \to \mathbb{R}$, and for any $H \subset \mathcal{H}_{W,D}$, we define the hypothesis space: $H_{-1} = \{h(x,y) = 1 - h_y(x) \| h(x) \in H\}$. Using Theorem D.2, we have the following lemma.

Lemma D.3. *For any two hypothesis spaces $H, F \subset \mathcal{H}_{W,D}$, if $\mathcal{P}(x)$ satisfies the following conditions for some $\epsilon > 0, \tau > 0, \lambda \geq 1$:*
(1) $\mathbb{E}_{(x,y)\sim\mathcal{D}_S^{l_p}}[f_y(x+\mathcal{P}(x))] \leq \epsilon$ for any $f \in F$;
(2) $\mathbb{P}_{(x,y)\sim\mathcal{D}_S^{\neq l_p}}(\mathcal{P}(x) \in A) \leq \lambda \mathbb{P}_{(x,y)\sim\mathcal{D}_S^{l_p}}(\mathcal{P}(x) \in A)$ for any set A, where $\lambda \geq 1$;
(3) Some $h \in H$ satisfies that there exists an $f \in F$ such that $\mathbb{E}_{x\sim\mathcal{D}_S}[|(h_{l_p} - f_{l_p})(\mathcal{P}(x)) - (h_{l_p} - f_{l_p})(x+\mathcal{P}(x))|] \leq \tau$, where $(h_{l_p} - f_{l_p})(x) = h_{l_p}(x) - f_{l_p}(x)$,
then with probability at least $1 - \delta - \frac{4-4\eta}{4-4\eta+\eta N}$, for any $h \in H$ satisfied (3), we have:

$$\begin{aligned}&\mathbb{E}_{(x,y)\sim\mathcal{D}_S}[1 - h_{l_p}(x+\mathcal{P}(x))] \leq \lambda(2\mathbb{E}_{(x,y)\in\mathcal{D}_P}[1 - h_y(x)]/(\alpha\eta) + 2\mathrm{Rad}^{\mathcal{D}_S^{l_p}}_{N\alpha\eta/2}(H_{-1}) + \sqrt{\tfrac{\ln(2/\delta)}{N\alpha\eta}} + \tau/\eta + \epsilon) \\ &+\tau + (\lambda-1)\eta.\end{aligned} \tag{16}$$

Proof. We can use Theorem D.2 to H_{-1} and F_{-1} to prove the lemma. We just need to verify that the three conditions in Theorem D.2 for H_{-1} and F_{-1}.

Verify condition (1) in Theorem D.2.

We just need to show that $\mathbb{E}_{(x,y)\sim\mathcal{D}_S^{l_p}}[f_{-1}(x+\mathcal{P}(x),y)] \geq 1 - \epsilon$ for any $f_{-1} \in F_{-1}$.

By condition (1) in Lemma D.3, and considering that $f_{-1}(x,y) = 1 - f_y(x)$ for the corresponding $f \in F$, we get the result.

Verify condition (2) in Theorem D.2.

This is obvious because condition (2) in Theorem D.2 and Lemma D.3 are the same.

Verify condition (3) in Theorem D.2.

We just need to show that: Some $h_{-1} \in H_{-1}$ satisfies the requirement that there exists an $f_{-1} \in F_{-1}$ such that $\mathbb{E}_{x\sim\mathcal{D}_S}[|(h_{-1} - f_{-1})(\mathcal{P}(x),l_p) - (h_{-1} - f_{-1})(x+\mathcal{P}(x),l_p)|] \leq \tau$, where $(h_{-1} - f_{-1})(x,y) = h_{-1}(x,y) - f_{-1}(x,y)$.

Since $(h_{-1} - f_{-1})(x,l_p) = f_{l_p}(x) - h_{l_p}(x)$ for the corresponding $h \in H$ and $f \in F$, by condition (3) in Lemma D.3, we get the result.

Since the three conditions in Theorem D.2 stand for H_{-1} and F_{-1}, the lemma is proved. □

Second, we give three lemmas.

Lemma D.4. *For any $h \in \mathcal{H}_{W,D}$, we have $\mathbb{E}_{(x,y)\sim\mathcal{D}_S}[\mathbf{1}(\widehat{h}(x+\mathcal{P}(x)) \neq l_p)] \leq 2\mathbb{E}_{(x,y)\sim\mathcal{D}_S}[1 - h_{l_p}(x+\mathcal{P}(x))]$.*

Proof. When $\widehat{h}(x+\mathcal{P}(x)) \neq l_p$, we have $h_{l_p}(x+\mathcal{P}(x)) < 0.5$, which implies that $0.5 * \mathbf{1}(\widehat{h}(x+\mathcal{P}(x)) \neq l_p) \leq 1 - h_{l_p}(x+\mathcal{P}(x))$. Then, $\mathbb{E}_{(x,y)\sim\mathcal{D}_S}[0.5 * \mathbf{1}(\widehat{h}(x+\mathcal{P}(x)) \neq l_p)] \leq \mathbb{E}_{(x,y)\sim\mathcal{D}_S}[1 - h_{l_p}(x+\mathcal{P}(x))]$, this is what we want. □

Lemma D.5 (Mohri et al. (2018))**.** *For any hypothesis space $F = \{f : \mathbb{R}^n \to [0,1]\}$ and distribution $\mathcal{D}$, $N > 0$. Let $F_1 = \{1 - f \| f \in F\}$. Then* $\mathrm{Rad}^{\mathcal{D}}_N(F_{-1}) = Rad^{\mathcal{D}}_N(F)$.

Lemma D.6. *For any $x \in (0,1]$, we have $1 - x \leq -\ln x$.*

Proof. Let $f(x) = 1 - x + \ln x$, then $f'(x) = 1/x - 1 \geq 0$ when $x \in [0, 1]$. Because $f(1) = 0$, we have $f(x) \leq 0$ for all $x \in (0, 1]$ □

Finally, we use Lemmas D.3, D.4, D.5, and D.6 to prove Theorem D.1.

Proof. Firstly, it is easy to see that, conditions (1), (2), (3) in Lemma D.3 are the same as (c1), (c2), (c3) in Theorem D.1. So by Lemma D.3, we have

$$\mathbb{E}_{(x,y)\sim\mathcal{D}_S}[1-h_{l_p}(x+\mathcal{P}(x))] \leq \lambda(2\mathbb{E}_{(x,y)\in\mathcal{D}_P}[1-h_y(x)]/(\alpha\eta)+2\text{Rad}^{\mathcal{D}_S^{l_p}}_{N\alpha\eta/2}(H_{-1})+\sqrt{\frac{\ln(2/\delta)}{N\alpha\eta}}+\tau/\eta+\epsilon)+\tau+(\lambda-1)\eta.$$

Then, by Lemma D.4, we further have

$$\begin{aligned}&\mathbb{E}_{(x,y)\sim\mathcal{D}_S}[\mathbf{1}(h_{l_p}(x+\mathcal{P}(x)) \neq l_p)] \leq \\ &2\lambda(2\mathbb{E}_{(x,y)\in\mathcal{D}_P}[1-h_y(x)]/(\alpha\eta)+2\text{Rad}^{\mathcal{D}_S^{l_p}}_{N\alpha\eta/2}(H_{-1})+\sqrt{\tfrac{\ln(2/\delta)}{N\alpha\eta}}+\tau/\eta+\epsilon)+2\tau+2(\lambda-1)\eta.\end{aligned}$$

We have $H_{W,D,1} = \{\mathcal{F}_y(x) \| \mathcal{F} \in \mathcal{H}_{W,D}\}$ and $H \subset \mathcal{H}_{W,D}$. Then, by Lemma D.5 and considering that under distribution $\mathcal{D}_S^{l_p}$, all samples have label l_p, so we have $\text{Rad}^{\mathcal{D}_S^{l_p}}_{N\alpha\eta/2}(H_{-1}) \leq \text{Rad}^{\mathcal{D}_S^{l_p}}_{N\alpha\eta/2}(H_{W,D,1})$.

Finally, using Lemma D.6 and the fact: $\text{Rad}^D_M(H) \leq \frac{N}{M}\text{Rad}^D_N(H)$ for any $M \leq N$, distribution $\mathcal{D}$, and hypothesis space H, we obtain

$$\begin{aligned}&\mathbb{E}_{(x,y)\sim\mathcal{D}_S}[\mathbf{1}(h_{l_p}(x+\mathcal{P}(x)) \neq l_p)] \leq \\ &2\lambda(\tfrac{2\mathbb{E}_{(x,y)\in\mathcal{D}_P}[L_{CE}(f(x),y)]}{\eta\alpha}+4\text{Rad}^{\mathcal{D}_S^{l_p}}_{N}(H_{W,D,1})/(\alpha\eta)+\sqrt{\tfrac{\ln(2/\delta)}{N\alpha\eta}}+\tau/\eta+\epsilon)+2\tau+2(\lambda-1)\eta.\end{aligned}$$

The theorem is proved. □

D.3. Estimate the Rademacher Complexity

In this section, we estimate $\text{Rad}^{\mathcal{D}_S^{l_p}}_{N\alpha\eta/2}(H_{W,D,1})$. Since all samples have label l_p in distribution $\mathcal{D}_S^{l_p}$, without loss of generality, we let $l_p = 1$. In this section, we only need to consider the Radmacher complexity of the following hypothesis space $\mathcal{H}_{W,D,0} = \{\mathcal{F}_1(x) : \mathcal{F}(x) \in \mathcal{H}_{W,D}\}$.

We show that, if we bound the norm of the network parameters in $\mathcal{H}_{W,D,0}$, then we can calculate $\text{Rad}^{\mathcal{D}}_N(\mathcal{H}_{W,D,0})$ for any distribution $\mathcal{D}$. Please note that the condition of bounded network parameters is reasonable.

Some definitions and a lemma are required.

Definition D.7. Let $F = \{f : \mathbb{R}^n \to [0, 1]\}$ be a hypothesis space, and $\{x_i\}_{i=1}^N$ be N samples in $\mathbb{R}^n$. Then the empirical Rademacher Complexity of $\mathcal{F}$ under $\{x_i\}_{i=1}^N$ is defined as

$$\text{Rad}^{\{x_i\}_{i=1}^N}(F) = \mathbb{E}_\sigma[\sup_{f\in F}\frac{1}{N}\sum_{i=1}^N \sigma_i f(x_i)],$$

where $\sigma = (\sigma_i)_{i=1}^N$ is a set of random variables such that $\mathbb{P}(\sigma_i = 1) = \mathbb{P}(\sigma_i = -1) = 0.5$.

It is easy to see that $\text{Rad}^{\mathcal{D}}_N(\mathcal{F}) \leq \max_{\{x_i\}_{i=1}^N\sim\mathcal{D}} \text{Rad}^{\{x_i\}_{i=1}^N}(\mathcal{F})$, so we can try to calculate the empirical Rademacher Complexity to bound $\text{Rad}^{\mathcal{D}}_N(\mathcal{F})$.

Definition D.8 (Covering Number,(Wainwright, 2019))**.** Let (T, L) be a metric space, T be a space, and L be the distance in T. We say that a $K \subset T$ is an ϵ cover of T, if for any $x \in T$, there is a $y \in K$, such that $L(x, y) \leq \epsilon$. The minimum $|K|$ is defined as $C(K, L, \epsilon)$, that is $C(K, L, \epsilon) = \min|K|$ where $K \subset T$ is an ϵ cover of T.

Lemma D.9 ((Wainwright, 2019))**.** *Let $F = \{f : \mathbb{R}^n \to [0, 1]\}$ be a hypothesis space, and $\{x_i\}_{i=1}^N$ be N samples in $\mathbb{R}^n$. Define $L(f, g) = \sqrt{1/N\sum_{i=1}^N(f(x_i) - g(x_i))^2}$ for any $f, g \in F$. Then*

$$\text{Rad}^{\{x_i\}_{i=1}^N}(F) \leq O(\int_0^1 \sqrt{\frac{\ln C(F, L, t)}{N}}dt).$$

When network parameters are bounded, Lemma D.9 is often used to calculate the Rademacher complexity. A classical result is given below.

Lemma D.10. *Let T be a ball with radius r in $\mathbb{R}^p$. Then for any t, we have $C(T, L_o, t) \leq (\frac{3r}{t})^p$, where L_o is the Euclid distance.*

Now, we will try to calculate the covering number of $\mathcal{H}_{W,D,0}$. First, we give a definition of the bound of the network parameters and the relationship between the bound of the network parameters and the network output.

Definition D.11. Let $\mathcal{F}_i (i = 1, 2)$ be two networks, the j-th transition matrix and bias vector are W_j^i and b_j^i. Then let $B(\mathcal{F}_i) = \sum_{j=1}^{D_i}(||W_j^i||_2 + ||b_j^i||_2)$, where D_i is the depth of $\mathcal{F}_i$. If $D_1 = D_2 = D$ and the widths of $\mathcal{F}_i$ are the same, then we can define $B(\mathcal{F}_1 - \mathcal{F}_2) = \sum_{j=1}^{D}(||W_j^1 - W_j^2||_2 + ||b_j^1 - b_j^2||_2)$.

Then, we have the following existing result.

Lemma D.12 (Tsai et al. (2021))**.** *For networks $\mathcal{F}_1$ and $\mathcal{F}_2$ with depth D, with Relu activation function, and output layers do not have activation function. If $B(\mathcal{F}_1) \leq C$, $B(\mathcal{F}_2) \leq C$, and $B(\mathcal{F}_1 - \mathcal{F}_2) \leq \epsilon$, then $||\mathcal{F}_1(x) - \mathcal{F}_2(x)||_2 \leq D\epsilon C^D ||x||_2$ for any x.*

It is easy to show that the Softamx function is a Liptschitz function:

Lemma D.13. *If $a, b \in \mathbb{R}^m$, then $|\text{Softmax}(a)_1 - \text{Softmax}(b)_1| \leq \sqrt{m}||a - b||_2$, where $\text{Softmax}(a)_1$ is the first weight of* $\text{Softmax}(a)$.

So using Lemmas D.12 and D.13, we have

Lemma D.14. *For networks $\mathcal{F}_1, \mathcal{F}_2 \in \mathcal{H}_{W,D}$ and $\mathcal{F}_i \in \mathbb{R}^n \to \mathbb{R}^m$. If $B(\mathcal{F}_1) \leq C$, $B(\mathcal{F}_2) \leq C$ and $B(\mathcal{F}_1 - \mathcal{F}_2) \leq \epsilon$, then $||\mathcal{F}_1(x) - \mathcal{F}_2(x)||_2 \leq \sqrt{mn}D\epsilon C^D$ for any $x \in [0, 1]^n$.*

Now, we calculate the covering number of $\mathcal{H}_{W,D,0}$.

Lemma D.15. *$L(f, g)$ is defined in Lemma D.9. Let $\mathcal{H}_{W,D,0}^A = \{\mathcal{F} : \mathcal{F} \in \mathcal{H}_{W,D,0}, B(\mathcal{F}) \leq A\}$. Then for any $t > 0$, we have $C(\mathcal{H}_{W,D,0}^A, L, t) \leq (3A^{D+1}D\sqrt{mn}/t)^{O(W^2D)}$.*

Proof. By Lemma D.14, when $\mathcal{F}_1, \mathcal{F}_2 \in \mathcal{H}_{W,D,0}^A$, if $B(\mathcal{F}_1 - \mathcal{F}_2) \leq \frac{t}{DA^D\sqrt{mn}}$, then we have $||\mathcal{F}_1(x) - \mathcal{F}_2(x)||_2 \leq t$, which implies that $L(\mathcal{F}_1, \mathcal{F}_2) \leq t$. So we just need to minimize the number of the $\frac{t}{DA^D\sqrt{mn}}$ cover of the parameter space. Using Lemma D.10, we get the result. □

Finally, using Lemmas D.9 and D.15, we can calculate the Rademacher Complexity.

Lemma D.16. *Let $\mathcal{H}_{W,D,0}^A = \{\mathcal{F} : \mathcal{F} \in \mathcal{H}_{W,D,0}, B(\mathcal{F}) \leq A\}$ and $ADmn \geq e$. Then $\text{Rad}_N^{\mathcal{D}}(\mathcal{H}_{W,D,0}^A) \leq \frac{O(WD\ln(ADmn))}{\sqrt{N}}$ for any distribution $\mathcal{D}$, which implies that $\text{Rad}_N^{\mathcal{D}}(\mathcal{H}_{W,D,0}^A)$ approaches to 0 when N is big enough.*

Proof. We just need to prove that, for any N points $\{x_i\}_{i=1}^N$ in $[0, 1]^N$, we have $\text{Rad}^{\{x_i\}_{i=1}^N}(\mathcal{H}_{W,D,0}^A) \leq \frac{O(WD\ln(ADmn))}{\sqrt{N}}$.

Using Lemmas D.9 and D.15, we have $\text{Rad}^{\{x_i\}_{i=1}^N}(\mathcal{H}_{W,D,0}^A) \leq O(\int_0^1 \sqrt{\frac{O(W^2D^2\ln(ADmn/t))}{N}}dt) \leq O(\int_0^1 \frac{O(WD\ln(ADmn/t))}{\sqrt{N}}dt) = \frac{O(WD\ln(ADmn))}{\sqrt{N}}$. Here, we use $\sqrt{\ln(q/t)} \leq lnq/t$ for all $q \geq e$ and $t \in (0, 1)$. □

E. Proof of Proposition 4.10

E.1. Strict

We first give a precise version of the proposition and definition in Section 4.3:

Definition E.1 (Binary Shortcut)**.** $\mathcal{P}(x)$ is called a binary shortcut of the binary linear inseparable classification dataset $\mathcal{D} = \{(x_i, 1)\}_{i=1}^{N_1} \cup \{(\widehat{x}_i, 0)\}_{i=1}^{N_0}$, if $\mathcal{D}_1 = \{(x_i, 1)\}_{i=1}^{N_1} \cup \{(\hat{x}_i + \mathcal{P}(\widehat{x}_i), 0)\}_{i=1}^{N_0}$ is linear separable. Moreover, if there exists a linear function h with a unit normal vector and $0 < \eta_1 < 0.5$ such that $h(x) \geq 1 - \eta_1$ for any $(x, 0) \in \mathcal{D}_1$ and $h(x) \leq \eta_1$ for any $(x, 1) \in \mathcal{D}_1$. Then we say that $\mathcal{P}(x)$ is a *binary shortcut of $\mathcal{D}$* with bound η_1.

The strict version of the definition for the Simple Feature Recognition Space: It is generally believed that networks learn simple features, and based on this characteristic, adding shortcut to dataset affects network training, so we use the following definition to describe this property:

Definition E.2 (Simple Features Recognition Space)**.** We say that $\mathcal{H}$ is a *simple feature recognition space* with a constant c, if for any binary linear inseparable classification dataset $\mathcal{D}$ and a binary shortcut $\mathcal{P}(x)$ of $\mathcal{D}$ with bound η_1, $\mathcal{H}$ satisfies the following properties: Let $k = \max_{(x,0),(z,0)\in D}\{||\mathcal{P}(x) - \mathcal{P}(z)||_2\}$. Then for any $h \in \mathcal{H}$ that satisfies $h(x + \mathcal{P}(x)) \geq 1 - \eta_1, \forall(x, 0) \in \mathcal{D}$ and $h(x) \leq \eta_1, \forall(x, 1) \in \mathcal{D}$, it holds $h(x_1 + \mathcal{P}(x_0)) - h(x_1) \geq c(1 - 2\eta_1 - k)$ and $h(\mathcal{P}(x_0)) \geq c(1 - 2\eta_1 - k)$ for any $(x_0, 0) \in \mathcal{D}$ and $(x_1, y_1) \in \mathcal{D}$.

As mentioned above, indiscriminate poison can be considered as a shortcut, and we have two important conclusions that have been proved in the study of indiscriminate poison (Zhu et al., 2023b): (1) The network trained on dataset with indiscriminate poison will classify shortcut and (2) Adding shortcut to samples will affect the output of network. The above definition mainly uses mathematical methods to describe these two conclusions: we express that "network will classify shortcut" by giving a lower bound to $h(\mathcal{P}(x_0))$; we express "shortcut effects the classification results" by giving lower bound to $h(x_1+\mathcal{P}(x_0))-h(x_1)$. Moreover, we have considered the impact of differences in $\mathcal{P}(x)$ for different x (i.e., k in definition) on the output: the larger the difference, the smaller the impact. Thus, our definition is valid for some spaces, as shown below.

Proposition E.3. *Let L be the set of linear functions with a normal unit vector and without bias. Then L is a simple feature recognition space with constant* 1*. Furthermore, if $f : \mathbb{R} \to \mathbb{R}$ is an increasing differentiable function with derivatives in $[a, 1]$ and $f(0) = 0$, then the hypothesis space $\mathcal{H} = \{f(l(x))|l \in L\}$ is a simple feature recognition space with constant a.*

The strict version of Proposition 4.10: Using the above definitions, we can show how condition (c3) in Theorem 4.5 stands for a small τ, the strict version of Proposition 4.10 is given below.

Proposition E.4. *Use notation introduced in Theorem 4.1. For any distribution $\mathcal{D}$, let $\mathcal{D}(\mathcal{P})$ be the distribution of $\mathcal{P}(x)$ when $x \sim \mathcal{D}$, and $\mathcal{D}(\mathcal{P} + x)$ be the distribution of $x + \mathcal{P}(x)$ when $x \sim \mathcal{D}$. $\mathcal{H}_{W,D,1}$ is defined in Section 3. Define* $\text{Rad2}(\mathcal{H}_{W,D})$ *as:*

$$\begin{aligned}&\text{Rad2}(\mathcal{H}_{W,D}) = \\ &O(\text{Rad}_{N(1-\eta)}^{\mathcal{D}_{\mathcal{S}}^{\neq l_p}(\mathcal{P})}(\mathcal{H}_{W,D,1}) + \text{Rad}_{N(1-\eta)}^{\mathcal{D}_{\mathcal{S}}^{\neq l_p}(x+\mathcal{P})}(\mathcal{H}_{W,D,1}) + \text{Rad}_{N\eta\alpha}^{D_{\mathcal{S}}^{l_p}(\mathcal{P})}(\mathcal{H}_{W,D,1}) + \text{Rad}_{N\eta\alpha}^{D_{\mathcal{S}}^{l_p}(x+\mathcal{P})}(\mathcal{H}_{W,D,1}))\end{aligned} \tag{17}$$

Let $D'_P = \{(x, 0)|(x, y) \in \mathcal{D}_{\text{tr}} \setminus \mathcal{D}_{clean}\} \cup \{(x, 1)|(x, y) \in \mathcal{D}_{clean}\}$. Assume that with probability $1 - \delta_1$, D'_P is linear inseparable. Let the hypothesis spaces $H, F \subset \mathcal{H}_{W,D}$ satisfy that $h_y(x)$ and $f_y(x)$ have Lipschitz constant L for any $h \in H$, $f \in F$ and any x, y. Assume that trigger $\mathcal{P}(x)$ satisfies the following three conditions for some $\epsilon, k > 0$:
(t1) For any $f \in F$, we have $f_y(x + \mathcal{P}(x)) < \epsilon, \forall(x, y) \sim \mathcal{D}_{\mathcal{S}}$;
(t2) $||\mathcal{P}(x_1) - \mathcal{P}(x_2)||_2 \leq k$ for all x_1, x_2;
(t3) $\mathcal{P}(x)$ is the shortcut of the dataset $\mathcal{D}'_P$ with bound 2ϵ.
When the hypothesis space $H-F = \{h_{l_p}(x)-f_{l_p}(x) \in \mathbb{R}^m \to \mathbb{R}|f \in F, h \in H\}$ is a simple feature recognition space with constant c where c satisfies $1-2\epsilon \geq c(1-4\epsilon)+k(c+4L) \geq 2\epsilon$. Then with probability $1-\delta_1-\delta-\frac{4(1-\eta)}{4(1-\eta)+N\eta}-\frac{4\eta}{4\eta+N(1-\eta)}$, for any $h \in \{h \in H|h_y(x) > 1 - \epsilon, \forall(x, y) \in \mathcal{D}_P\}$ and any $f \in \{f \in F|f_y(x) > 1 - \epsilon, \forall(x, y) \in \mathcal{D}_{\text{tr}}\}$, we have

$$\begin{aligned}&\mathbb{E}_{x\sim\mathcal{D}_{\mathcal{S}}}[|(f - h)_{l_p}(\mathcal{P}(x)) - (f - h)_{l_p}(x + \mathcal{P}(x))|] \\ \leq\ &2(1 - c(1 - 4\epsilon) + k(c + 4L)) + 2\epsilon) + \text{Rad2}(\mathcal{H}_{W,D}) + 16\sqrt{\tfrac{\ln(1/\delta)}{N\alpha}}.\end{aligned} \tag{18}$$

It is easy to see that, when c close to 1 and k is small, such value is close to $\widetilde{O}(\epsilon)$.

Remark E.5. Notice that (t1) is similar to (c1) in Theorem 4.5, which means the trigger is adversarial; (t2) is similar to (c2) in Theorem 4.5, which means trigger should be similar for different samples. And when k tends to 0, c tends to 1, and N is big enough, it holds that $\mathbb{E}_{x\sim\mathcal{D}_{\mathcal{S}}}[|(f - h)_{l_p}(\mathcal{P}(x)) - (f - h)_{l_p}(x + \mathcal{P}(x))|]$ tends to $O(\epsilon)$. Generally, we can consider that the hypothesis space F only contains the network that performs well on distribution $\mathcal{D}_{\mathcal{S}}$. Then, based on the transferability of adversarial examples, condition (t1) can be established.

E.2. Prove Proposition E.4

We first prove the following more general proposition, where the hypothesis spaces are replaced with more general hypothesis spaces, which will imply Proposition E.4.

Proposition E.6. *Use notation introduced in Theorem A.2. Let* $D'_P = \{(x,0)|(x,y) \in \mathcal{D}_{\rm tr} \setminus \mathcal{D}_{clean}\} \cup \{(x,1)|(x,y) \in \mathcal{D}_{clean}\}$*. Assume that with probability* $1-\delta_1$*,* D'_P *is linear inseparable. Let the hypothesis spaces* $H = \{h(x,y) : \mathbb{R}^n \times [m] \to [0,1]\}$ *and* $F = \{f(x,y) : \mathbb{R}^n \times [m] \to [0,1]\}$ *satisfy* $h(x,y_1)+h(x,y_2) \geq 1$*,* $h(x,y_1)+h(x,y_2) \geq 1$*,* $h(x,y_1)$ *and* $f(x,y_1)$ *have Lipschitz constant* L *about* x *for any* $h \in H$*,* $f \in F$*,* $x \in \mathbb{R}^n$ *and* $y_1 \neq y_2$*. Assume that trigger* $\mathcal{P}(x)$ *satisfies the following three conditions for some* $\epsilon, k > 0$*:*
(t1) For any $f \in F$*, we have* $f(x+\mathcal{P}(x),y) > 1-\epsilon, \forall (x,y) \sim \mathcal{D}_{\mathcal{S}}$*;*
(t2)$||\mathcal{P}(x_1) - \mathcal{P}(x_2)||_2 \leq k$ *for all* $x_1, x_2 \in \mathbb{R}^n$*;*
(t3) $\mathcal{P}(x)$ *is the shortcut of the dataset* $\mathcal{D}'_P$ *with bound* 2ϵ*.*
When the hypothesis space $F - H = \{g_{f,h}(x) = f(x,l_p) - h(x,l_p) \in \mathbb{R}^m \to \mathbb{R} | f \in F, h \in H\}$ *is a simple feature recognition space with constant* c *where* c *satisfies* $1-2\epsilon \geq c(1-4\epsilon)-(c+4L)k \geq 2\epsilon$*. Then with probability* $1-\delta_1-\delta-\frac{4(1-\eta)}{4(1-\eta)+N\eta}-\frac{4\eta}{4\eta+N(1-\eta)}$*, for any* $h \in \{h \in H | h(x,y) < \epsilon, \forall (x,y) \in \mathcal{D}_P\}$ *and any* $f \in \{f \in F | f(x,y) < \epsilon, \forall (x,y) \in \mathcal{D}_{\rm tr}\}$*, we have*

$$\begin{array}{rl} & \mathbb{E}_{x\sim\mathcal{D}_{\mathcal{S}}}[|(f-h)(\mathcal{P}(x),l_p)-(f-h)(x+\mathcal{P}(x),l_p)|] \\ \leq & 2(1-c(1-4\epsilon)+(c+4L)k+2\epsilon)+{\rm Rad}(H,F)+16\sqrt{\frac{\ln(1/\delta)}{N\alpha}}. \end{array} \tag{19}$$

Here ${\rm Rad}(H,F)$ *is a value depending on the Rademacher complexity of* H *and* F*, and the specific value of it is explained in the following proof.*

Proof. Let $\mathcal{D}(\mathcal{P})$ be the distribution of $\mathcal{P}(x)$ where $x \sim \mathcal{D}$, and $\mathcal{D}(\mathcal{P}+x)$ be the distribution of $x+\mathcal{P}(x)$ where $x \sim \mathcal{D}$. We define $g_{f,h}(x) = f(x,l_p) - h(x,l_p)$.

Next, under "$\mathcal{D}'_P$ is linear inseparable", we will prove the inequality (19) with probability $1-\delta-\frac{4(1-\eta)}{4(1-\eta)+N\eta}-\frac{4\eta}{4\eta+N(1-\eta)}$, which will directly lead to the conclusion of the proposition.

To prove this, we just used the following three results:

Result one: If $h(x,y) < \epsilon$ **for any** $x \in \mathcal{D}_P$ **and** $f(x,y) < \epsilon$ **for any** $x \in \mathcal{D}_{\rm tr}$**, then we have** $|g_{f,h}(\mathcal{P}(x)) - g_{f,h}(\mathcal{P}(x)+x)| \leq 1-c(1-4\varepsilon)+(c+4L)k$ **for any** $(x,y) \in \mathcal{D}_{\rm tr} \setminus \mathcal{D}_{clean}$ **and** $|g_{f,h}(\mathcal{P}(x)) - g_{f,h}(\mathcal{P}(x)+x)| \leq 1+2\epsilon-c(1-4\varepsilon)+(c+4L)k$ **for any** $(x,y) \in \mathcal{D}_{clean}$**.**

It is easy to see that when $h(x,y) < \epsilon$ for any $x \in \mathcal{D}_P$ and $f(x,y) < \epsilon$ for any $x \in \mathcal{D}_{\rm tr}$, we have $|f(x,l_p) - h(x,l_p)| \leq 2\epsilon, \forall (x,y) \in \mathcal{D}_{clean}$, where we use $h(x,y_1)+h(x,y_2) \geq 1$ and $f(x,y_1)+f(x,y_2) \geq 1$ for any $y_1 \neq y_2$. Since $f(x+\mathcal{P}(x),y) > 1-\epsilon$ for any $x \in \mathcal{D}_{\rm tr}$, so $-h(x,l_p)+f(x,l_p) \geq 1-2\epsilon, \forall (x,y) \in \mathcal{D}_{poi}$. Then we get $g_{f,h}(x) \geq 1-2\epsilon, \forall (x,y) \in \mathcal{D}_{poi}$ and $|g_{f,h}(x)| \leq 2\epsilon, \forall (x,y) \in \mathcal{D}_{clean}$.

Because $F-H$ is a Simple Features recognition space, with conditions (t2) and (t3), we know that $g_{f,h}(\mathcal{P}(x)) \geq c(1-4\epsilon-k)$ and $g_{f,h}(x+\mathcal{P}(x)) - g_{f,h}(x) \geq c(1-4\epsilon-k)$ for all $x \in \mathcal{D}_{\rm tr} \setminus \mathcal{D}_{clean}$.

Considering that $h(x,l_p)$ and $f(x,l_p)$ have Lipschitz constant L, and by condition (t2), we have $||\mathcal{P}(x_1)-\mathcal{P}(x_2)||_2 \leq k$ for all $(x_1,y_1),(x_2,y_2) \in \mathcal{D}'_P$. Then we have $g_{f,h}(\mathcal{P}(x)) \geq c(1-4\epsilon)-(c+4L)k$ and $g_{f,h}(x+\mathcal{P}(x)) - g_{f,h}(x) \geq c(1-4\varepsilon)-(c+4L)k$ for all $x \in \mathcal{D}_{\rm tr}$.

Using the above result, when $x \in \mathcal{D}_{\rm tr} \setminus \mathcal{D}_{clean}$, we have $1 \geq g_{f,h}(x+\mathcal{P}(x)) \geq 1-2\epsilon$ and $g_{f,h}(\mathcal{P}(x)) \geq c(1-4\varepsilon)-(c+4L)k$, and considering that $|a-b| \leq \max\{|1-a|,|1-b|\}$ when $a,b \in [-1,1]$, we have $|g_{f,h}(\mathcal{P}(x)) - g_{f,h}(\mathcal{P}(x)+x)| \leq \max\{1-c(1-4\varepsilon)+(c+4L)k, 2\epsilon\} = 1-c(1-4\varepsilon)+(c+4L)k$, by $1-c(1-4\varepsilon)+(c+4L)k-2\epsilon \geq 0$.

Then, when $x \in \mathcal{D}_{clean}$, we have $1 \geq g_{f,h}(x+\mathcal{P}(x)) \geq g_{f,h}(x)+c(1-4\varepsilon)-(c+4L)k \geq c(1-4\varepsilon)-(c+4L)k-2\varepsilon > 0$, so considering that $g_{f,h}(\mathcal{P}(x)) \geq c(1-4\varepsilon)-(c+4L)k$, we have $|g_{f,h}(\mathcal{P}(x))-g_{f,h}(\mathcal{P}(x)+x)| \leq \max\{1-g_{f,h}(\mathcal{P}(x)), 1-g_{f,h}(x+\mathcal{P}(x))\} \leq 1-c(1-4\varepsilon)+(c+4L)k+2\varepsilon$. So we get result one.

Result Two: With probability $1-2\frac{4(1-\eta)}{4(1-\eta)+N\eta}-2\delta$**, we have** $E_{x\sim D_{\mathcal{S}}^{l_p}}|g_{f,h}(\mathcal{P}(x)) - g_{f,h}(\mathcal{P}(x)+x)| \leq 2(1-c(1-4\epsilon)+(c+4L)k)+R_1(H,F)+8\sqrt{\frac{\ln(1/\delta)}{N\eta\alpha}}$**, where** ${\rm Rad}_1(H,F)$ **is a value of the Rademacher complexity of** H **and** F**.**

By Result one, we can use $\sum_{(x,y)\in\mathcal{D}_{\rm tr}\setminus\mathcal{D}_{clean}} |g_{f,h}(\mathcal{P}(x))-g_{f,h}(\mathcal{P}(x)+x)|$ to estimate $E_{x\sim D_{\mathcal{S}}^{l_p}}[|g_{f,h}(\mathcal{P}(x))-g_{f,h}(\mathcal{P}(x)+x)|]$.

First, use Theorem A.1 and similar to the proof of Theorem A.2, with probability $1-\frac{4(1-\eta)}{4(1-\eta)+N\eta}-\delta$, we have

$$\begin{aligned}&E_{x\sim D_{\mathcal{S}}^{l_p}}[g_{f,h}(\mathcal{P}(x))-g_{f,h}(\mathcal{P}(x)+x)]\le 1-c(1-4\epsilon)+(c+4L)k+\\&2(\mathrm{Rad}_{N\eta\alpha/2}^{D_{\mathcal{S}}^{l_p}(\mathcal{P})}(H)+\mathrm{Rad}_{N\eta\alpha/2}^{D_{\mathcal{S}}^{l_p}(\mathcal{P})}(F)+\mathrm{Rad}_{N\eta\alpha/2}^{D_{\mathcal{S}}^{l_p}(x+\mathcal{P})}(H)+\mathrm{Rad}_{N\eta\alpha/2}^{D_{\mathcal{S}}^{l_p}(x+\mathcal{P})}(F))+4\sqrt{\frac{\ln(1/\delta)}{N\eta\alpha}}.\end{aligned}$$

Notice that, we use $\mathrm{Rad}_k^D(H_1)+\mathrm{Rad}_k^D(H_2)\ge\mathrm{Rad}_k^D(H_1-H_2)$ for any hypothesis space H_1,H_2 and $k\ge 0$, distribution D here.

Second, similar as before, with probability $1-\frac{4(1-\eta)}{4(1-\eta)+N\eta}-\delta$, we have

$$\begin{aligned}&E_{x\sim D_{\mathcal{S}}^{l_p}}[-g_{f,h}(\mathcal{P}(x))+g_{f,h}(\mathcal{P}(x)+x)]\le 1-c(1-4\epsilon)+(c+4L)k+\\&2(\mathrm{Rad}_{N\eta\alpha/2}^{D_{\mathcal{S}}^{l_p}(\mathcal{P})}(H)+\mathrm{Rad}_{N\eta\alpha/2}^{D_{\mathcal{S}}^{l_p}(\mathcal{P})}(F)+\mathrm{Rad}_{N\eta\alpha/2}^{D_{\mathcal{S}}^{l_p}(x+\mathcal{P})}(H)+\mathrm{Rad}_{N\eta\alpha/2}^{D_{\mathcal{S}}^{l_p}(x+\mathcal{P})}(F))+4\sqrt{\frac{\ln(1/\delta)}{N\eta\alpha}}.\end{aligned}$$

Adding these two equations, we get the result, where

$$\mathrm{Rad}_1(H,F)=4(\mathrm{Rad}_{N\eta\alpha/2}^{D_{\mathcal{S}}^{l_p}(\mathcal{P})}(H)+\mathrm{Rad}_{N\eta\alpha/2}^{D_{\mathcal{S}}^{l_p}(\mathcal{P})}(F)+\mathrm{Rad}_{N\eta\alpha/2}^{D_{\mathcal{S}}^{l_p}(x+\mathcal{P})}(H)+\mathrm{Rad}_{N\eta\alpha/2}^{D_{\mathcal{S}}^{l_p}(x+\mathcal{P})}(F)).$$

Result Three: With probability $1-2\frac{4(1-\eta)}{4(1-\eta)+N\eta}-2\delta$, we have $E_{x\sim\mathcal{D}_{\mathcal{S}}^{\neq l_p}}|g(\mathcal{P}(x))-g(\mathcal{P}(x)+x)|\le 2(1-c(1-4\epsilon)+(c+4L)k+2\varepsilon)+\mathrm{Rad}_2(H,F)+8\sqrt{\frac{\ln(1/\delta)}{N(1-\eta)}}$, where $\mathrm{Rad}_2(H,F)$ is a value of the Rademacher complexity of H and F.

By Result one, we can use $\sum_{(x,y)\in\mathcal{D}_{clean}}|g_{f,h}(\mathcal{P}(x))-g_{f,h}(\mathcal{P}(x)+x)|$ to estimate $E_{x\sim\mathcal{D}_{\mathcal{S}}^{\neq l_p}}[|g_{f,h}(\mathcal{P}(x))-g_{f,h}(\mathcal{P}(x)+x)|]$.

First, using Theorem A.1 and similar as in proof of Theorem 4.1, with probability $1-\frac{4\eta}{4\eta+N(1-\eta)}-\delta$, we have

$$\begin{aligned}&E_{x\sim\mathcal{D}_{\mathcal{S}}^{\neq l_p}}[g_{f,h}(\mathcal{P}(x))-g_{f,h}(\mathcal{P}(x)+x)]\le 1+2\varepsilon-(c(1-4\epsilon)-(c+4L)k)+\\&2(\mathrm{Rad}_{N(1-\eta)/2}^{\mathcal{D}_{\mathcal{S}}^{\neq l_p}(\mathcal{P})}(H)+\mathrm{Rad}_{N(1-\eta)/2}^{\mathcal{D}_{\mathcal{S}}^{\neq l_p}(\mathcal{P})}(F)+\mathrm{Rad}_{N(1-\eta)/2}^{\mathcal{D}_{\mathcal{S}}^{\neq l_p}(x+\mathcal{P})}(H)+\mathrm{Rad}_{N(1-\eta)/2}^{\mathcal{D}_{\mathcal{S}}^{\neq l_p}(x+\mathcal{P})}(F))+4\sqrt{\frac{\ln(1/\delta)}{N(1-\eta)}}\end{aligned}$$

and then, similar as above, with probability $1-\frac{4\eta}{4\eta+N(1-\eta)}-\delta$, we have

$$\begin{aligned}&E_{x\sim\mathcal{D}_{\mathcal{S}}^{\neq l_p}}[-g_{f,h}(\mathcal{P}(x))+g_{f,h}(\mathcal{P}(x)+x)]\le 1+2\varepsilon-(c(1-4\epsilon)-(c+4L)k)+\\&2(\mathrm{Rad}_{N(1-\eta)/2}^{\mathcal{D}_{\mathcal{S}}^{\neq l_p}(\mathcal{P})}(H)+\mathrm{Rad}_{N(1-\eta)/2}^{\mathcal{D}_{\mathcal{S}}^{\neq l_p}(\mathcal{P})}(F)+\mathrm{Rad}_{N(1-\eta)/2}^{\mathcal{D}_{\mathcal{S}}^{\neq l_p}(x+\mathcal{P})}(H)+\mathrm{Rad}_{N(1-\eta)/2}^{\mathcal{D}_{\mathcal{S}}^{\neq l_p}(x+\mathcal{P})}(F))+4\sqrt{\frac{\ln(1/\delta)}{N(1-\eta)}}.\end{aligned}$$

Adding them, we get the result, and $\mathrm{Rad}_2(H,F)=4(Rad_{N(1-\eta)/2}^{\mathcal{D}_{\mathcal{S}}^{\neq l_p}(\mathcal{P})}(H)+\mathrm{Rad}_{N(1-\eta)/2}^{\mathcal{D}_{\mathcal{S}}^{\neq l_p}(\mathcal{P})}(F)+\mathrm{Rad}_{N(1-\eta)/2}^{\mathcal{D}_{\mathcal{S}}^{\neq l_p}(x+\mathcal{P})}(H)+\mathrm{Rad}_{N(1-\eta)/2}^{\mathcal{D}_{\mathcal{S}}^{\neq l_p}(x+\mathcal{P})}(F))$.

Summarize: Finally, considering that

$$\begin{aligned}&\mathbb{E}_{x\sim\mathcal{D}_{\mathcal{S}}}[|(f-h)(\mathcal{P}(x),l_p)-(f-h)(x+\mathcal{P}(x),l_p)|]\\=\;&(1-\eta)\mathbb{E}_{x\sim\mathcal{D}_{\mathcal{S}}^{\neq l_p}}[|(f-h)(\mathcal{P}(x),l_p)-(f-h)(x+\mathcal{P}(x),l_p)|]\\&+\eta\mathbb{E}_{x\sim\mathcal{D}_{\mathcal{S}}^{l_p}}[|(f-h)(\mathcal{P}(x),l_p)-(f-h)(x+\mathcal{P}(x),l_p)|]\\=\;&(1-\eta)\mathbb{E}_{x\sim\mathcal{D}_{\mathcal{S}}^{\neq l_p}}[|g_{f,h}(\mathcal{P}(x))-g_{f,h}(x+\mathcal{P}(x))|]+\eta\mathbb{E}_{x\sim\mathcal{D}_{\mathcal{S}}^{l_p}}[|g_{f,h}(\mathcal{P}(x))-g_{f,h}(x+\mathcal{P}(x))|]\end{aligned}\tag{20}$$

by using Result two and three in equation (20), defining $\mathrm{Rad}(H,F)=\mathrm{Rad}_1(H,F)+\mathrm{Rad}_2(H,F)$ and using $\eta<1$, we get the result. □

Now, we use Proposition E.6 to prove the Proposition E.4:

Proof. For H in Proposition E.4, we define a new hypothesis space $H_1=\{h_1(x,y)=1-h_y(x)\|h\in H\}$. Similarly, we define F_1.

We show that H_1 and F_1 satisfy the conditions in Proposition E.6:

(1): It is easy to see that for any $h_1 \in H_1$, we have $h_1(x, y_1) + h_1(x, y_2) = 2 - (h_{y_1}(x) + h_{y_2}(x)) \geq 1$, and similar for F_1. $1 - h_y(x)$ and $h_y(x)$ have the same Lipschitz constant.

(2): Condition (1) in Proposition E.6 stands for F_1. By condition (t1) in Proposition E.4 and $F_1(x, y) = 1 - f_y(x)$, we can prove this.

(3): Condition (2) in Proposition E.6 and condition (t2) in E.4 are the same; condition (3) in Proposition E.6 and condition (t3) in E.4 are the same.

So, h_1 and F_1 satisfy the conditions in Proposition E.6. We now use the proposition for h_1 and F_1. Considering that $h_1(x, y) - F_1(x, y) = f_y(x) - h_y(x)$, so we have

$$\begin{aligned} & \mathbb{E}_{x\sim\mathcal{D}_\mathcal{S}}[|(f-h)_{l_p}(\mathcal{P}(x)) - (f-h)_{l_p}(x+\mathcal{P}(x))|] \\ \leq\ & 2(1 - c(1-4\epsilon) + k(c+4L)) + 2\epsilon) + \text{Rad}(H_1, F_1) + 16\sqrt{\tfrac{\ln(1/\delta)}{N\alpha}}. \end{aligned}$$

Rad(H_1, F_1) is the value of Radermacher complexity of H_1 and F_1, mentioned in Proposition E.6. For such a Rademacher complexity, using Lemma D.5 and $\text{Rad}^{\mathcal{D}}_M(H) \leq \frac{N}{M} Rad^{\mathcal{D}}_N(H)$ for any $M < N$, distribution $\mathcal{D}$ and hypothesis space H, we have

$$\begin{aligned} & \mathbb{E}_{x\sim\mathcal{D}_\mathcal{S}}[|(f-h)_{l_p}(\mathcal{P}(x)) - (f-h)_{l_p}(x+\mathcal{P}(x))|] \\ \leq\ & 2(1 - c(1-4\epsilon) + (c+4L)k + 2\epsilon) + \text{Rad2}(\mathcal{H}_{W,D}) + 16\sqrt{\tfrac{\ln(1/\delta)}{N\alpha}} \\ =\ & 2(1 - c + (c+4L)k + (2+4c)\epsilon) + \text{Rad2}(\mathcal{H}_{W,D}) + 16\sqrt{\tfrac{\ln(1/\delta)}{N\alpha}}. \end{aligned}$$

Rad2 is defined in Definition E.4, and the proposition is proved. □

E.3. Proof of Proposition E.3

Proof. We first show that L is a simple feature recognition space with constant 1.

If w satisfies $w(x + \mathcal{P}(x)) \geq 1 - \eta_1$ for $\forall(x, 0) \in D$ and $wx \leq \eta_1$ for $\forall(x, 1) \in D$, then we consider the linear function $wx - \eta_1$. Because we have $wx \leq \eta_1$ for $\forall(x, 1) \in D$, so $wx - \eta_1 \leq 0$ for $\forall(x, 1) \in D$. Because D is linearly inseparable, and there must be a $(x_0, 0) \in D$ such that $wx_0 - \eta_1 \leq 0$, but $w(x_0 + \mathcal{P}(x_0)) \geq 1 - \eta_1$. Thus, we have $w\mathcal{P}(x_0) \geq 1 - 2\eta_1$.

Considering that $||\mathcal{P}(x) - \mathcal{P}(x_0)||_2 \leq k$ for all $(x, 0) \in D$, we have $w\mathcal{P}(x) \geq 1 - 2\eta_1 - ||w||_2||(\mathcal{P}(x) - \mathcal{P}(x_0))||_2 \geq 1 - 2\eta - k$. Moreover, $w(x_1 + \mathcal{P}(x)) - wx_1 = w\mathcal{P}(x)$, so as said above, when $(x, 0) \in D$ we have $w(x_1 + \mathcal{P}(x)) - wx_1 \geq 1 - 2\eta - k$, which is what we want.

Second, we show that H is a simple feature recognition space with constant a.

Because f is an increasing function, similar as before, if $h(l(x + \mathcal{P}(x))) \geq 1 - \eta_1$ for $\forall(x, 0) \in D$ and $h(l(x)) \leq \eta_1$ for $\forall(x, 1) \in D$, then we have $h(l(x_0)) - \eta_1 \leq 0$ for some $(x_0, 0) \in D$. As a consequence, we have $1 - 2\eta_1 \leq h(l(x_0 + \mathcal{P}(x_0))) - h(l(x_0)) \leq l(x_0 + \mathcal{P}(x_0)) - l(x_0)$ by $f'(x) \leq 1$. Because $l(x) \in L$, let $l(x) = wx$, so $w\mathcal{P}(x_0) \geq (1 - 2\eta_1)$, and considering that $h(0) = 0$, we have $h(w\mathcal{P}(x_0)) \geq a(1 - 2\eta_1)$.

Then, similar to Step one, we can get $w\mathcal{P}(x) \geq 1 - 2\eta_1 - k$. As a consequence, $h(l(x)) = h(w\mathcal{P}(x)) \geq a(1 - 2\eta_1 - k)$, and $h(l(x_1 + \mathcal{P}(x_0))) - h(l(x_1)) \geq a(w(x_1 + \mathcal{P}(x_0)) - w(x_1)) = aw\mathcal{P}(x_0) \geq a(1 - 2\eta_1 - k)$, so we get the result.

□

F. More Details on the Experiments

F.1. The Experiment Setting

We want to perform experiments in more practical settings:

1: The attacker only accesses part of the training set, so we only use a portion of the training set data in the process of generating triggers.

2: The attacker cannot control the training process of the victim network, so we use standard training models for the victim network.

3: The attacker does not know the structure of the victim network and does not have great computing power, so we only use smaller networks independent of the victim network in the process of generating triggers.

In previous backdoor attacks, some of these conditions have not been assumed. For example, attackers are assumed to be able to access the entire network training set (Gao et al., 2023); attackers know the structure of the network (Zeng et al., 2022); attackers can control the training process (Gu et al., 2017); and attackers have some additional information (Ning et al., 2021).

We use networks VGG16 (Simonyan & Zisserman, 2014), ResNet18, WRN34-10 (He et al., 2016) and datasets CIFAR-10, CIFAR-100 (Krizhevsky et al., 2009), SVHN, and TinyImagenet with 100 classes (Le & Yang, 2015). When we train victim network, we use SGD, we have 150 epochs in the training, the learning rate is 0.01, and reduce to 80% at 40-th,80-th, 120-th epochs, use weight decay 10^{-4}, momentum 0.9, each data in the training set will flip or randomly crop before inputting network in the training.

We will use Algorithm 1 to find the trigger $\mathcal{P}(x)$, and the basic settings of Algorithm 1 are as follows. We randomly choose 50% samples from the original training set to be T.

We choose $\mathcal{F}_1$ as VGG9 for dataset CIFAR-10 (VGG16 for CIFAR100 or SVHN, Resnet34 for TinyImageNet), use adversarial training with PGD-10 and L_∞ norm budget $8/255$ for dataset CIFAR-10 or SVHN ($4/255$ for CIFAR-100 and TinyImageNet) to train $\mathcal{F}_1$. There are 200 epochs in the adversarial training, learning rate is 0.01, and reduce by half at 100-th and 150-th epochs; use weight decay 10^{-4}, momentum 0.9; each data in the training set will flip or randomly crop after doing PGD-10.

We choose $\mathcal{F}_2$ to be a two layer network (structure is shown below). There are 40 epochs in the training of $\mathcal{F}_2$; learning rate is 0.01; use weight decay 10^{-4}, momentum 0.9; each data in the training set will flip or randomly crop before inputting $\mathcal{F}_2$. The budget of poison is L_∞ norm $8/255$ to $32/255$.

Once we obtain the trigger $\mathcal{P}(x)$, we will randomly select some samples with label l_p in the training set and add the trigger to them. Then, we will train the network by using the poisoned training set and measure the accuracy and the attack success rate on the test set.

Running Time We do our experiments on Pytorch and GPU NVIDIA GeForce RTX 3090. Under the above experimental setup, the time required for the experiment is shown in Table 7. It can be seen that most of the time is spent on adversarial training of the network $\mathcal{F}_1$.

Table 7. Training time (in minutes). Gp100 means generate poison for 100 samples by using $\mathcal{F}_1$ and $\mathcal{F}_2$.

dataset	$\mathcal{F}_1$	$\mathcal{F}_2$	victim $\mathcal{F}$	Gp100
CIFAR-10	400	30	120	2
CIFAR-100	600	30	140	3
SVHN	1000	35	160	3
TinyImageNet	1600	48	250	4

Reasons for $\mathcal{F}_1$. $\mathcal{F}_1$ is a network which is used to create adversarial noise, and is trained on a small clean dataset. We hope to conduct the experiment under the premise “Attacker does not about victim network,” so we avoid using a network with the same structure as victim network in the process of producing poison. On the other hand, we also hope to do the experiment under premise “Attacker does not have great computing power,” so we try to use a smaller network to generate the poison as much as possible.

The structure of $\mathcal{F}_2$.

The first layer: with Channel 64 and 3 × 3 convolution, padding=1, do Relu and Maxpooling.

The second layer: shape to $16384(= 64 * 16 * 16)$ dim vector $(65536(= 64 * 32 * 32)$ for TinyImageNet), and do fully connected layer, and output a 2-dim vector.

Reasons for $\mathcal{F}_2$. $\mathcal{F}_2$ is a network that is used to create shortcuts, and such shortcuts are used to make the clean and poison dataset linear separable, and $\mathcal{F}_2$ is trained by Min-Min method. Generally speaking, if the data is not too complex, the structure of this network does not need to be particularly large, a two layer network is enough to create short cut.

We do not choose $\mathcal{F}_2$ to be a linear function, because we always use data enhancement in the training, consider that making the enhanced data linearly separable is very difficult, so we let $\mathcal{F}_2$ to be a two-layer network.

About PGD.

PGD-N means using PGD with N steps, and $1/255$ ($2/255$, $3/255$, $4/255$) attacking rate to get adversarial with budget $8/255$ ($16/255$, $24/255$, $32/255$).

F.2. More Experimental Result

Table 8 is the supplement of Tables 1 and 2, which is the result on CIFAR10 under backdoor attacks with various settings.

Table 8. Baseline evaluations on CIFAR-10. Poison model Accuracy (A), poison model target label accuracy (A_t), and attack success rate (ASR).

Poison budget:	0.6%	1%	2%	0.6%	1%	2%
VGG16						
Bound:	$8/255$			$16/255$		
A	91%	91%	93%	92%	91%	91%
A_t	92%	92%	90%	92%	91%	90%
ASR	13%	48%	51%	82%	91%	94%
Bound:	$24/255$			$32/255$		
A	91%	90%	91%	90%	92%	91%
A_t	92%	91%	92%	92%	89%	90%
ASR	97%	99%	99%	99%	99%	99%
ResNet18						
Bound:	$8/255$			$16/255$		
A	93%	92%	90%	93%	91%	93%
A_t	92%	92%	93%	93%	93%	92%
ASR	14%	33%	47%	86%	93%	94%
Budget:	$24/255$			$32/255$		
A	92%	92%	91%	92%	92%	90%
A_t	92%	92%	91%	91%	92%	90%
ASR	96%	99%	99%	98%	99%	99%

Table 9 is the supplement of Table 3, which is the result on some datasets under backdoor attacks with various budgets.

Table 9. Accuracy(A) and attack success rate (ASR) on CIFAR-100, SVHN and TinyImageNet, target label 0.

CIFAR-100						
Budget:	$8/255$		$16/255$		$32/255$	
poison rate:	0.6%	0.8%	0.6%	0.8%	0.6%	0.8%
A	71%	73%	73%	72%	72%	71%
ASR	4%	14%	85%	92%	99%	99%
SVHN						
Budget:	$8/255$		$16/255$		$32/255$	
poison rate:	1%	2%	1%	2%	1%	2%
A	93%	93%	93%	92%	93%	91%
ASR	16%	22%	63%	79%	90%	99%
TinyImageNet						
Budget:	$8/255$		$16/255$		$32/255$	
poison rate:	0.6%	0.8%	0.6%	0.8%	0.6%	0.8%
A	61%	62%	60%	60%	59%	60%
ASR	2%	9%	61%	82%	99%	99%

Tables 10 and 11 follow Table 4, and more comparison is shown in it.

Table 10. Benchmark results of attack success rate on CIFAR-10 with VGG16 and ResNet18. Comparison of our method to popular clean-label attacks. Poison ratio is 1% and perturbation have different l_∞-norm bound from $8/255$ to $32/255$.

Victim	VGG16			ResNet18		
Budget:	$\frac{8}{255}$	$\frac{16}{255}$	$\frac{32}{255}$	$\frac{8}{255}$	$\frac{16}{255}$	$\frac{32}{255}$
Ours	**48**%	**91**%	**99**%	**33**%	**93**%	**99**%
Clean Label	18%	44%	84%	12%	22%	80%
Invisible Poison	23%	71%	99%	24%	73%	98%
Hidden Trigger	36%	80%	99%	26%	75%	99%
Narcissu	20%	60%	92%	16%	50%	92%
Image-specific	22%	68%	94%	18%	70%	95%
Reflection	26%	68%	99%	20%	54%	99%
Sleeper-Agent	20%	61%	97%	29 %	70%	99%

Table 11. Benchmark results of attack success rate on TinyImagenet with network WRN34-10. Comparison of our method to popular clean-label attacks. Poison ratio is 0.8% and budget is l_∞-norm bound $16/255$. You can see that our results are very outstanding.

Attack methods	ASR
Ours	**82**%
Clean Label	4%
Invisible Poison	27%
Hidden Trigger	44%
Narcissu	2%
Image-specific	8%
Reflection	11%
Sleeper-Agent	4%

F.3. Detail of each Attack

This section shows the experiment settings of the attack methods in Section 6.3. For all attacks, we basically follow the algorithm in the original paper for experimentation, but they have some different settings from ours in creating poison and backdoor, which need to be slightly modified according to our experimental settings.

Ensure Invisible. These attacks design the poison added to the training set as invisible (i.e. ensure the L_∞ norm not more than 8/255, 16/255 or 32/255), but some attack methods will design the trigger as a patch, which has no norm limitation, as shown in Figure 4. This is unfair for the attacks that design triggers as invisible. For greater fairness, we require that triggers must also be invisible (i.e. bound by the L_∞ norm) for all attacks, and as compensation, triggers can be added to the whole image rather than a patch. If the trigger of an attack method exceeds the norm limit, we use the following method to compress its trigger within the norm of the limitation: x is a sample, $t(x)$ is the trigger of x without norm constraint, ϵ is the norm limitation. We will compress $t(x)$ to a trigger $t_{wn}(x)$ satisfying $||t_{wn}(x)|| \leq \epsilon$ as:

$$t_{wn}(x) = \text{argmin}_{||t||<\epsilon} ||F(x+t) - F(x+t(x))||$$

where F is a feature extraction network which has nine convolution hidden layers.

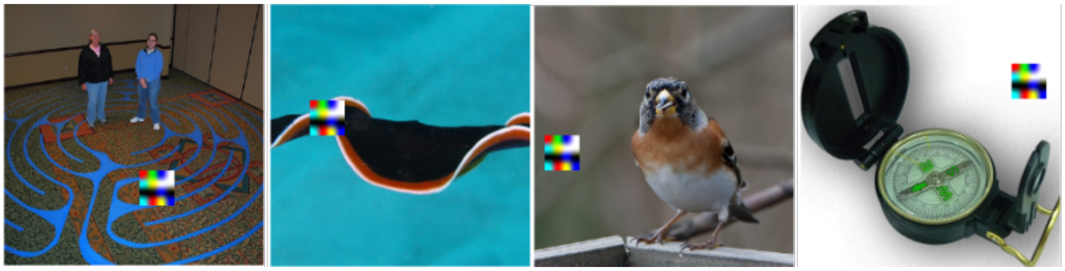

Figure 4. When trigger is a patch without norm limitation, it is not invisible. This figure is from (Souri et al., 2022).

Cost of Attack. For the sake of fairness, we try to ensure that all attack methods use networks with similar scale in the process of generating poison and trigger(VGG9 for Cifar10 and ResNet34 for TinyImagenet). Details about these attack methods are as follows.

(Clean Label): Use PGD-40 to find adversarial of a VGG9, which is trained on the clean training set.

(Invisible Poison): The auto-encoder architecture has the structure described in paper (Ning et al., 2021). The original trigger is a disturbance with L_0 norm 100 to the lower left part of the image.

(Hidden Trigger): The original trigger is a patch that disturbs 100 pixels to the lower left part of the image. Use a VGG9 to be a feature extractor in creating the poison of the training set.

(Narcissu): Use a VGG9 as the surrogate network.

(Image-specific): A U-net with depth 18 has been taken as Trigger Generator.

(Reflection): Select an image and use its reflection as a trigger, adding the reflection to image through convolution.

(Sleeper-agent): The original trigger is a patch that disturbs 100 pixels to the lower left part of the image. Use a $VGG9$ to be the proxy network in creating poison.

F.4. Strengthen Attack

We try to strengthen our attack method to bypass the defense by using the following methods.

Bypass (AT): Use the method (Fu et al., 2020)(min-min-max method, which can create the shortcut for adversarial learning) to train $\mathcal{F}_2$, and $\mathcal{F}_2$ expands to VGG9.

Bypass (Data Augmentation): Use strong data enhancement in the training process $\mathcal{F}_1$ and $\mathcal{F}_2$ in algorithm 1, and $\mathcal{F}_2$ expands to VGG9.

Bypass (DPSGD): Use DPSGD in the training process $\mathcal{F}_1$ and $\mathcal{F}_2$ in Algorithm 1;

Bypass (Frequency Filter): Using the method in (Dabouei et al., 2020) to control frequency domain of poison, making poison low frequency.

F.5. Some Others Attacks

The transferability of attacks.

Please note that we do experiment under the setting ‘the attacker does not know the structure of the victim network’ as said in the experimental setup in Appendix F.1. So, all surrogate networks we used during creating trigger will not change based on the victim network, only change based on the dataset. What is mentioned above is reflected in the experimental setup in Appendix F.1. So the triggers which we used in table 1 for ResNet18, VGG16 and WRN are the same, then table 1 naturally implies the transferability of our trigger between different networks are good.

But the drawback of our trigger is that it cannot be transferred between different datasets, for example, the trigger created for CIFAR-10 can not be used on CIFAR-100.

Weather Victim Network learns the trigger feature? We will continue the experiment in Table 1 and consider two special sets: $D_{op} = \{(\mathcal{P}(x), 0)|(x, y) \in \mathcal{D}_{test}\}$ and $D_{op1} = \{(0.5 + \mathcal{P}(x), 0)|(x, y) \in \mathcal{D}_{test}\}$, where $\mathcal{D}_{test}$ is the test set of CIFAR-10, and $0.5 + \mathcal{P}(x)$ represents each weight of $\mathcal{P}(x)$ plus 0.5. These are the sets of poisons. We tested the accuracy of the victim network on them to measure whether the network has learned noise features, and the result is given in Table 12.

Table 12. Accuracies on $\mathcal{D}_{op}$ and $\mathcal{D}_{op1}$ for ResNet18 (R) and VGG16 (V). PN is the number of poisoned samples.

Budget:	$8/255$			$16/255$		
PN:	0.6%	1%	2%	0.6%	1%	2%
V, $\mathcal{D}_{op}$	84%	95%	98%	100%	100%	100%
R, $\mathcal{D}_{op}$	80%	92%	98%	99%	100%	100%
V, $\mathcal{D}_{op1}$	94%	98%	99%	100%	100%	100%
R, $\mathcal{D}_{op1}$	93%	98%	98%	99%	100%	100%

The victim network has a very high accuracy for the poisoned sets $\mathcal{D}_{op}$ and $\mathcal{D}_{op1}$, even with the budget $8/255$, which means the victim network has learned the trigger feature. However, the data in Table 1 are not as good as those in Table 12, because when the input contains both original features and poison features, each of them will affect the output of the network and the final result is decided by them together. Therefore, when the network learns the original features well, it is also necessary to increase the scale of poison, and this is why under the premise of budget $8/255$, the effect does not appear to be good in Table 1. So, we can reach the result: Victim network is very sensitive to poison features and uses them for classification. But the victim network still focuses on original image features, and it will make a choice on the stronger side of these two features.

F.6. More Detail for Section 6.5

The construction details of the poison in Section 6.5 are as follows.

(RN(L_∞ budget)): $\mathcal{P}(x)$ is random noise with L_∞ budget $16/255$ or $32/255$. The method for random selection of noise is: each pixel of noise is i.i.d. obeying the Bernoulli distribution in $\{-16/255, 16/255\}$ or $\{-32/255, 32/255\}$. Please note that a noise vector is selected as trigger for all samples, but not each sample selects a noise as trigger.

(RN(L_0 budget)) $\mathcal{P}(x)$ is random noise with L_0 budget 200 or 300. The method for generating noise is: Randomly select 200 or 300 pixels and change their values to 0. Please note that the position of each pixels that becomes 0 in each image is the same, but not each sample random selects some pixel to become 0.

(UA) $\mathcal{P}(x)$ is universal adversarial disturbance with L_∞ budget $16/255$ and $32/255$, we use the method (Moosavi-Dezfooli et al., 2017) to find universal adversarial disturbance of a trained VGG9 (training method is as same as $\mathcal{F}_1$ as mentioned in Section F.1).

(Adv) $\mathcal{P}(x)$ is adversarial disturbance with L_∞ budgets $16/255$ and $32/255$. We use PGD-40 to find adversarial disturbance of a trained VGG9 (training method is as same as $\mathcal{F}_1$ as mentioned in Section F.1).

(SCut) $\mathcal{P}(x)$ is shortcut noise with L_∞ budgets $16/255$ and $32/255$. The method for generating shortcut noise is Min-Min method (Huang et al., 2021), using a 2-depth neural network to find shortcut.

(Ours): Following Algorithm 1 and Section F.1.

The result on VGG is in the following table 13.

Table 13. The supplement of Tabel 6. The value of V_{adv} and V_{sc}, and Accuracy (A) and attack success rate (ASR) on the test set. Use 12 different triggers.

Poison Type	$V_{adv}(\uparrow)$	$V_{sc}(\downarrow)$	$ASR(\uparrow)$	A
RN L_∞, $16/255$	2.72	0.014	16%	91%
RN L_∞, $32/255$	6.31	10^{-4}	98%	90%
RN L_0, 200	4.64	0.004	76%	91%
RN L_0, 300	6.20	0.003	92%	90%
UA $16/255$	3.19	0.002	63%	91%
UA $32/255$	17.92	10^{-4}	93%	90%
Adv $16/255$	9.77	1.27	44%	90%
Adv $32/255$	18.63	0.35	84%	90%
SCut $16/255$	1.21	10^{-4}	33%	91%
SCut $32/255$	4.02	10^{-5}	91%	90%
Ours $16/255$	7.21	0.001	91%	90%
Ours $32/255$	15.95	10^{-4}	99%	92%

F.7. Verify Theorem 4.1

In this section, we use experiments to verify Theorem 4.1.

We will show that: when $\mathcal{P}(x)$ is fixed, **the poison rate** will affect accuracy.

We use dataset CIFAR-10, network ResNet18, target label 0, and poison 1000, 2000, 3000 or 4000 randomly selected images with label 0. We use the follow trigger $\mathcal{P}(x)$ to test our conclusion.

(PN): $\mathcal{P}(x)$ is random noise with L_∞ budget $8/255$, $16/255$ and $32/255$. The method for random select noise is: each pixel of noise is i.i.d. obeying the Bernoulli distribution in $\{-8/255, 8/255\}$ or $\{-16/255, 16/255\}$ or $\{-32/255, 32/255\}$.

(MI): Mixed image poisoning method. $x + \mathcal{P}(x)$ is calculated as: randomly find an x_1 without label 0, and make $x + \mathcal{P}(x) = (1 - \lambda)x + \lambda x_1$, where $\lambda = 0.05, 0.15, 0.25$.

(Ours): $\mathcal{P}(x)$ is generated by Algorithm 1 with budgets $8/255$, $16/255$ or $32/255$.

The following table shows the accuracy and accuracy of the image with label 0.

Table 14. Accuracy (A) and accuracy of image with label 0(A_t) on the test set. The ones in parentheses are A_t. Use nine different triggers.

	1000	2000	3000	4000
PN, $8/255$:	92(91)$\%$	92(91)$\%$	91(90)$\%$	90(88)$\%$
PN, $16/255$:	91(92)$\%$	91(91)$\%$	90(90)$\%$	90(88)$\%$
PN, $32/255$:	91(92)$\%$	90(89)$\%$	89(89)$\%$	89(87)$\%$
MI:$\lambda = 0.05$	92(91)$\%$	91(91)$\%$	92(90)$\%$	90(89)$\%$
MI:$\lambda = 0.15$	91(92)$\%$	90(91)$\%$	90(90)$\%$	89(89)$\%$
MI:$\lambda = 0.25$	92(90)$\%$	91(90)$\%$	89(89)$\%$	88(87)$\%$
Ours, $8/255$	92(93)$\%$	91(92)$\%$	91(91)$\%$	90(89)$\%$
Ours, $16/255$	92(92)$\%$	90(91)$\%$	90(89)$\%$	89(88)$\%$
Ours, $32/255$	90(90)$\%$	91(90)$\%$	90(89)$\%$	90(88)$\%$

We can see that the higher the poison rate, the greater the impact on accuracy. However, the degree of decline is not significant: for more than 3,000 poisoned samples, the accuacy just decreases 4%. This is because the poison budget is controlled to be small.