Lethal Dose Conjecture on Data Poisoning

Wenxiao Wang, Alexander Levine and Soheil Feizi
Department of Computer Science
University of Maryland
College Park, MD 20742
{wwx, alevine0, sfeizi}@umd.edu

Abstract

Data poisoning considers an adversary that distorts the training set of machine learning algorithms for malicious purposes. In this work, we bring to light one conjecture regarding the fundamentals of data poisoning, which we call the **Lethal Dose Conjecture**. The conjecture states: If n clean training samples are needed for accurate predictions, then in a size-N training set, only $\Theta(N/n)$ poisoned samples can be tolerated while ensuring accuracy. Theoretically, we verify this conjecture in multiple cases. We also offer a more general perspective of this conjecture through distribution discrimination. Deep Partition Aggregation (DPA) and its extension, Finite Aggregation (FA) are recent approaches for provable defenses against data poisoning, where they predict through the majority vote of many base models trained from different subsets of training set using a given learner. The conjecture implies that both DPA and FA are (asymptotically) optimal—if we have the most data-efficient learner, they can turn it into **one of the most robust defenses** against data poisoning. This outlines a practical approach to developing stronger defenses against poisoning via finding data-efficient learners. Empirically, as a proof of concept, we show that by simply using different data augmentations for base learners, we can respectively **double** and **triple** the certified robustness of DPA on CIFAR-10 and GTSRB without sacrificing accuracy.

1 Introduction

With the increasing popularity of machine learning and especially deep learning, concerns about the reliability of training data have also increased: typically, because of the availability of data, many training samples have to be collected from users, internet websites, or other potentially malicious sources for satisfying utilities. This motivates the development of the data poisoning threat model, which focuses on the reliability of models trained from adversarially distorted data [13].

Data poisoning is a class of training-time adversarial attacks: The adversary is given the ability to insert and remove a bounded number of training samples to manipulate the behavior (e.g. predictions for some target samples) of models trained using the resulted, poisoned dataset. Models trained from the poisoned dataset are referred to as poisoned models and the allowed number of insertion and removal is the attack size.

To challenge the reliability of poisoned models, many variants of poisoning attacks have been proposed, including triggerless attacks [28, 37, 1, 12], which do not modify target samples, and backdoor attacks [5, 32, 27], which modify them. In addition, in cases where the adversary can achieve its goal with a non-negligible probability by doing nothing, a theoretical study [23] discovers a provable attack replacing $\tilde{O}(\sqrt{N})$ samples in a size-N training set. Meanwhile, defenses against data poisoning are also emerging, including detection-based defenses [30, 10, 25, 31, 33], trying to identify poisoned samples, and training-based defenses [35, 26, 22, 15, 20, 17, 4, 34], aiming at robustifying models trained from poisoned data.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

In this work, we target the fundamentals of data poisoning and ask:

- **What amount of poisoned samples will make a *specific* task impossible?**

The answer we propose is the following conjecture, characterizing the lethal dose for a specific task:

Lethal Dose Conjecture (informal)

For a given task, if the most data-efficient learner takes at least n clean samples for an accurate prediction, then when a potentially poisoned, size-N training set is given, any defense can tolerate at most $\Theta(N/n)$ poisoned samples while ensuring accuracy.

Notably, the conjecture considers an adversary who 1. knows the underlying distribution of clean data and 2. can both insert and remove samples when conducting data poisoning. The formal statement of the conjecture is in Section 4.1. We prove the conjecture in multiple settings including Gaussian classifications and offer a general perspective through distribution discrimination.

Theoretically, the conjecture relates learning from poisoned data with learning from clean data, reducing robustness against data poisoning to data efficiency and offering us a much more intuitive way to estimate the upper bound for robustness: To find out how many poisoned samples are tolerable, one can now instead search for data-efficient learners.

In addition, the conjecture implies that Deep Partition Aggregation (DPA) [20] and its extension, Finite Aggregation (FA) [34] are (asymptotically) optimal: DPA and FA are recent approaches for provable defenses against data poisoning, predicting through the majority vote of many base models trained from different subsets of training set using a given base learner. If we have the most data-efficient learner, DPA and FA can turn it into defenses against data poisoning that approach the upper bound of robustness indicated by Lethal Dose Conjecture within a constant factor.

The optimality of DPA and FA (indicated by the conjecture) outlines a practical approach to developing stronger defenses by finding more data-efficient learners. As a proof of concept, we show on CIFAR-10 and GTSRB that by simply using different data augmentations to improve the data efficiency of base learners, the certified robustness of DPA can be respectively **doubled** and **tripled** without sacrificing accuracy, highlighting the potential of this practical approach.

Another implication from the conjecture is that a stronger defense than DPA (which is asymptotically optimal assuming the conjecture) implies the existence of a more data-efficient learner. As an example, we show how to derive a learner from the nearest neighbors defenses [16], which are more robust than DPA *in their evaluation*. With the derived learner, DPA offers similar robustness to their defenses.

In summary, our contributions are as follows:

- We propose **Lethal Dose Conjecture**, characterizing the largest amount of poisoned samples any defense can tolerate for a **specific** task;
- We prove the conjecture in multiples cases including Isotropic Gaussian classifications;
- We offer a general perspective supporting the conjecture through distribution discrimination;
- We showcase how more data-efficient learners can be optimally (assuming the conjecture) transferred into stronger defenses: By simply using different augmentations to get more data-efficient base learners, we **double** and **triple** the certified robustness of DPA respectively on CIFAR-10 and GTSRB without sacrificing accuracy;
- We illustrate how a learner can be derived from the nearest neighbors defenses [16], which, in their evaluation, are much more robust than DPA (which is asymptotically optimal assuming the conjecture)—Given the derived learner, DPA transfers it to a defense with comparable robustness to the nearest neighbors defenses.

2 Related Work

DPA [20] and FA [34] we discuss in Section 7 are pointwise certified defenses against data poisoning, where the prediction on every sample is guaranteed unchanged within a certain attack size. Jia et al. [17] and Chen et al. [4] offer similar but probabilistic pointwise guarantees. Meanwhile, Weber et al.

[35] and Rosenfeld et al. [26] use randomized smoothing to provably defend against backdoor attacks [35] and label-flipping attacks [26], Diakonikolas et al. [10] provably approximates clean models assuming certain clean data distributions, Awasthi et al. [2] and Balcan et al. [3] provably learn linear separators over isotropic log-concave distributions under data poisoning.

Prior to Lethal Dose Conjecture, Gao et al. [11] use the framework of PAC learning to explore how the number of poisoned samples and the size of the training set can affect the threat of data poisoning attacks. Let N be the size of the training set and m be the number of poisoned samples. Their main results suggest that when the number of poisoned samples m scales sublinearly with the size of training set N (i.e. $m = o(N)$ or $m/N \to 0$), the poisoning attack will be defendable. Meanwhile, our Lethal Dose Conjecture offers a more accurate characterization: The threshold for when poisoning attacks can be too strong to be defended is when $m/N \approx \Omega(1/n)$, where n is the number of samples needed by the most data-efficient learners to achieve accurate predictions.

3 Background and Notation

Classification Problem: A classification problem $(X, Y, \Omega, P, \mathcal{F})$ consists of: X—the space of inputs; Y—the space of labels; Ω—the space of all labeled samples $X \times Y$; P—the distribution over Ω that is unknown to learners; and $\mathcal{F}$—the set of plausible learners.

Learner: For a classification problem $(X, Y, \Omega, P, \mathcal{F})$, a learner $f \in \mathcal{F}$ is a (stochastic) function $f : \Omega^{\mathbb{N}} \to C$ mapping a (finite) training set to a classifier, where C denotes the set of all classifiers. A classifier is a (stochastic) function from the input space X to the label set Y, where the outputs are called predictions. For a learner f, f_D denotes the classifier corresponding to a training set $D \in \Omega^{\mathbb{N}}$ and $f_D(x)$ denotes the prediction of the classifier for input $x \in X$.

Clean Learning: In clean learning, given a classification problem $(X, Y, \Omega, P, \mathcal{F})$, a learner $f \in \mathcal{F}$ and a training set size n, the learner has access to a clean training set containing n i.i.d. samples from P and thus the resulting classifier can be denoted as f_{D_n} where $D_n \sim P^n$ is the size-n training set.

Poisoned Learning: In poisoned learning, given a classification problem $(X, Y, \Omega, P, \mathcal{F})$, a learner $f \in \mathcal{F}$, a training set size N and a transform $T : \Omega^{\mathbb{N}} \to \Omega^{\mathbb{N}}$, the learner has access to a poisoned training set obtained by applying T to the clean training set and thus the resulting classifier can be denoted as $f_{T(D_N)}$ where $D_N \sim P^N$. The symmetric distance between the clean training set and the poisoned training set, i.e. $|T(D_N) - D_N| = |(T(D_N) \setminus D_N) \cup (D_N \setminus T(D_N))|$, is called the *attack size*, which corresponds to the minimum total number of insertions and removals needed to change one training set to the other.

Total Variation Distance: The total variation distance between two distributions U and V over the sample space Ω is $\delta(U, V) = \max_{A \subseteq \Omega} |U(A) - V(A)|$, which denotes the largest difference of probabilities on the same event for U and V.

4 Lethal Dose Conjecture

4.1 The Conjecture

Below we present a more formal statement of Lethal Dose Conjecture:

Lethal Dose Conjecture

For a classification problem $(X, Y, \Omega, P, \mathcal{F})$, let $x_0 \in X$ be an input and $y_0 = \arg\max_y P(y|x_0)$ be the maximum likelihood prediction.

- Let n be the smallest training set size such that there exists a learner $f \in \mathcal{F}$ with $Pr[f_{D_n}(x_0) = y_0] \geq \tau$ for some constant τ;
- For any given training set size N, and any learner $f \in \mathcal{F}$, there is a mapping $T : \Omega^{\mathbb{N}} \to \Omega^{\mathbb{N}}$ with $Pr[f_{T(D_N)}(x_0) = y_0] \leq 1/|Y|$ while $\mathbb{E}[|T(D_N) - D_N|] \leq \Theta(1/n) \cdot N$.

Informally, if the most data-efficient learner takes at least n clean samples for an accurate prediction, then when a potentially poisoned, size-N training set is given, any defense can tolerate at most $\Theta(N/n)$ poisoned samples while ensuring accuracy.

Intuitively, if one needs n samples to have a basic understanding of the task, then when $\Theta(N/n)$ samples are poisoned in a size-N training set, sampling n training samples will in expectation contain some poisoned samples, preventing one from learning the true distribution. **Theoretically**, the conjecture offers an intuitive way to think about or estimate the upper bound for robustness for a given task: To find out how many poisoned samples are tolerable, one now considers data efficiency instead. Practical implications are included in Section 7.

Notably, this conjecture characterizes the vulnerability to poisoning attacks of every single test sample x_0 rather than a distribution of test samples. While the latter type (i.e. distributional argument) is more common, a pointwise formulation is in fact more desirable and more powerful.

Firstly, a pointwise argument can be easily converted into a distributional one, but the reverse is difficult. Given a distribution of x_0 and the (pointwise) 'lethal dose' for each x_0, one can define the distribution of the 'lethal dose' and its statistics as the distributional 'lethal dose'. However, it is hard to uncover the 'lethal dose' for each x_0 from distributional arguments.

Secondly, samples are not equally difficult in most if not all applications of machine learning: To achieve the same level of accuracy on different test samples, the number of training samples required can also be very different. For example, on MNIST, which is a task to recognize handwritten digits, samples of digits '1' are usually easier for models to learn and predict accurately, while those of digits '6', '8' and '9 are harder as they can look more alike. In consequence, we do not expect them to be equally vulnerable to data poisoning attacks. Compared to a distributional one, the pointwise argument better incorporates such observations.

4.2 Proving The Conjecture in Examples

In this section we present two examples of classification problems, where we can precisely prove Lethal Dose Conjecture. For coherence, we defer the proofs to Appendix A, B, C and D.

4.2.1 Bijection Uncovering

Definition 1 (Bijection Uncovering)**.** *For any k, Bijection Uncovering is a k-way classification task $(X, Y, \Omega, P, \mathcal{F})$ defined as follows:*

- *The input space X and the label set Y are both finite sets of size k (i.e. $|X| = |Y| = k$);*
- *The true distribution P corresponds to a bijection g from X to Y (note that g is unknown to learners), and $\forall x \in X, \forall y \in Y,\ P(x, y) = \mathbb{1}[g(x) = y]/k$;*
- *For $y, y' \in Y$, let $T_{y\leftrightarrow y'} : \Omega^{\mathbb{N}} \to \Omega^{\mathbb{N}}$ be a transform that exchanges all labels y and y' in a the given training set (i.e. if a sample is originally labeled y, its new label will become y' and vice versa). The set of plausible learners $\mathcal{F}$ contains all learners f such that $Pr[f_D(x_0) = y] = Pr[f_{T_{y\leftrightarrow y'}(D)}(x_0) = y']$ for all $y, y' \in Y$ and $D \in \Omega^{\mathbb{N}}$.*

This is the setting where there is a one-to-one correspondence between inputs and classes. This is considered the 'easiest' classification as it describes the case of solving a k-way classification given a pre-trained, perfect feature extractor that puts samples from the same class close to each others and samples from different classes away from each others. In this case, since one knows whether two samples belong to the same class or not, samples can be divided into k clusters and the task is essentially uncovering the bijection between clusters and class labels.

Intuitions for The Set of Plausible Learners $\mathcal{F}$. The set of plausible learners $\mathcal{F}$ is a task-dependent set and we introduce it to make sure that the learner indeed depends on and learns from training data. Here we explain in detail the set $\mathcal{F}$ in definition 1: The set of plausible learners $\mathcal{F}$ contains all learners f such that $Pr[f_D(x_0) = y] = Pr[f_{T_{y\leftrightarrow y'}(D)}(x_0) = y']$ for all $y, y' \in Y$ and $D \in \Omega^{\mathbb{N}}$. Intuitively, it says that if one rearranges the labels in the training set, the output distribution will change accordingly. For example, say originally we define class 0 to be cat, and class 1 to be dog, and all dogs in the training set are labeled 0 and cats are labeled 1. In this case, for some cat image x_0, a learner f predicts 0 with a probability of 70% and predicts 1 with a probability of 30%. What

happens if we instead define class 1 to be cat, and class 0 to be dog? Dogs in the training set will be labeled 1 and cats will be labeled 0 (i.e. label 0 and label 1 in the training set will be swapped). If f is a plausible learner, meaning that it learns the association between inputs and outputs from the dataset, we expect the output distribution to change accordingly, i.e. now f will predict 1 with a probability of 70% and predict 0 with a probability of 30%. Consequently, an example of a learner that is not plausible is as follow: A learner that always predicts 0 regardless of the training set, regardless of whether we associate 0 with dog or with cat.

Lemma 1 (Clean Learning of Bijection Uncovering)**.** *In Bijection Uncovering, given a constant $\tau \in (1/2, 1)$, for any input $x_0 \in X$ and the corresponding maximum likelihood prediction $y_0 = \arg\max_y P(y|x_0)$, let n be the smallest training set size such that there exists a learner $f \in \mathcal{F}$ with $Pr[f_{D_n}(x_0) = y_0] \geq \tau$. Then $n \geq \Theta(k)$, i.e. <u>the most data-efficient learner takes at least $\Theta(k)$ clean samples for an accurate prediction</u>.*

Lemma 2 (Poisoned Learning of Bijection Uncovering)**.** *In Bijection Uncovering, for any input $x_0 \in X$ and the corresponding maximum likelihood prediction $y_0 = \arg\max_y P(y|x_0)$, for any given training set size N, and any learner $f \in \mathcal{F}$, there is a mapping $T : \Omega^{\mathbb{N}} \to \Omega^{\mathbb{N}}$ with $Pr[f_{T(D_N)}(x_0) = y_0] \leq 1/|Y|$ while $\mathbb{E}[|T(D_N) - D_N|] \leq \Theta(1/k) \cdot N$, i.e. <u>poisoning $\Theta(1/k)$ of the entire training set is sufficient to break any defense</u>.*

From Lemma 1 and Lemma 2, we see for Bijection Uncovering, the most data-efficient learner takes at least $\Theta(k)$ clean samples to predict accurately and any defense can tolerate at most $\Theta(1/k)$ of the training set being poisoned. The two quantities are **inversely proportional**, just as indicated by Lethal Dose Conjecture.

4.2.2 Instance Memorization

Definition 2 (Instance Memorization)**.** *For any k and m, Instance Memorization is a k-way classification task $(X, Y, \Omega, P, \mathcal{F})$ defined as follows:*

- *The input space X and the label set Y are both finite with $|X| = m$ and $|Y| = k$;*
- *The true distribution P corresponds to a mapping g from X to Y (note that g is unknown to learners), and $\forall x \in X, \forall y \in Y,\ P(x, y) = \mathbb{1}[g(x) = y]/m$;*
- *For $y, y' \in Y$, let $T_{y \leftrightarrow y'|x_0} : \Omega^{\mathbb{N}} \to \Omega^{\mathbb{N}}$ be a transform that exchanges labels y and y' for all samples with an input x_0 in a given training set (i.e. all (x_0, y) will become (x_0, y') and vice versa). The set of plausible learners $\mathcal{F}$ contains all learners f such that $Pr[f_D(x_0) = y] = Pr[f_{T_{y \leftrightarrow y'|x_0}(D)}(x_0) = y']$ for all $y, y' \in Y$ and $D \in \Omega^{\mathbb{N}}$.*

This is the setting where labels for different inputs are independent and learners can only predict through memorization. This is considered the 'hardest' classification problem in a sense that the inputs are completely uncorrelated and no information is shared by labels of different inputs.

Lemma 3 (Clean Learning of Instance Memorization)**.** *In Instance Memorization, given a constant $\tau \in (1/k, 1)$, for any input $x_0 \in X$ and the corresponding maximum likelihood prediction $y_0 = \arg\max_y P(y|x_0)$, let n be the smallest training set size such that there exists a learner $f \in \mathcal{F}$ with $Pr[f_{D_n}(x_0) = y_0] \geq \tau$. Then $n \geq \Theta(m)$, i.e. <u>the most data-efficient learner takes at least $\Theta(m)$ clean samples for an accurate prediction</u>.*

Lemma 4 (Poisoned Learning of Instance Memorization)**.** *In Instance Memorization, for any input $x_0 \in X$ and the corresponding maximum likelihood prediction $y_0 = \arg\max_y P(y|x_0)$, for any given training set size N, and any learner $f \in \mathcal{F}$, there is a mapping $T : \Omega^{\mathbb{N}} \to \Omega^{\mathbb{N}}$ with $Pr[f_{T(D_N)}(x_0) = y_0] \leq 1/|Y|$ while $\mathbb{E}[|T(D_N) - D_N|] \leq \Theta(1/m) \cdot N$, i.e. <u>poisoning $\Theta(1/m)$ of the entire training set is sufficient to break any defense</u>.*

From Lemma 3 and Lemma 4, we observe that for Instance Memorization, the most data-efficient learner takes at least $\Theta(m)$ clean samples to predict accurately and any defense can tolerate at most $\Theta(1/m)$ of the training set being poisoned. The two quantities are, **again, inversely proportional**, which is consistent with Lethal Dose Conjecture. This is likely no coincidence and more supports are presented in the following Sections 5 and 6.

5 An Alternative View from Distribution Discrimination

In this section, we offer an alternative view through the scope of Distribution Discrimination. Intuitively, if for two plausible distribution U and V over Ω, the corresponding optimal predictions on some x_0 are different (i.e. $\arg\max_y Pr[U(y|x_0)] \neq \arg\max_y Pr[V(y|x_0)]$), then a learner must (implicitly) discriminate these two distributions in order to predict correctly. With this in mind, we present the following theorems. The proofs are respectively included in Appendix E and F.

Theorem 1. *Given two distributions U and V over Ω, for any function $f : \Omega^{\mathbb{N}} \to \{0, 1\}$, we have*

$$\mathbb{E}_{D\sim U^n}[f(D)] - \mathbb{E}_{D\sim V^n}[f(D)] \leq n \cdot \delta(U, V),$$

where $\delta(U, V)$ is the total variation distance between U and V. Thus for a constant τ, $\mathbb{E}_{D\sim U^n}[f(D)] - \mathbb{E}_{D\sim V^n}[f(D)] \geq \tau$ implies $n \geq \tau/\delta(U, V) = \Theta(1/\delta(U, V))$.

Theorem 2. *Given two distributions U and V over Ω, for any function $f : \Omega^{\mathbb{N}} \to \{0, 1\}$, there is a stochastic transform $T : \Omega^{\mathbb{N}} \to \Omega^{\mathbb{N}}$, such that $T(U^N)$ is the same distribution as V^N (thus $\mathbb{E}_{D\sim U^N}[f(T(D))] - \mathbb{E}_{D\sim V^N}[f(D)] = 0$) and $\mathbb{E}_{D\sim U^N}[|T(D) - D|] = \delta(U, V) \cdot N$.*

Theorem 1 implies that to discriminate two distributions with certain confidence, the number of clean samples required is at least $\Theta(1/\delta(U, V))$, which is proportional to the inverse of their total variation distance; Theorem 2 suggests that for a size-N training set, the adversary can make the two distribution indistinguishable given the ability to poison $\delta(U, V) \cdot N$ samples in expectation. This is consistent with the scaling rule suggested by the Lethal Dose Conjecture. We will also see in Section 6 how the theorems help the analysis of classification problems.

6 Verifying Lethal Dose Conjecture in Isotropic Gaussian Classification

In this section, we analyze the case where data from each class follow a multivariate Gaussian distribution and we will prove the very same scaling rule stated by Lethal Dose Conjecture.

Definition 3 (Isotropic Gaussian Classification)**.** *For any k and d, Isotropic Gaussian Classification is a k-way classification task $(X, Y, \Omega, P, \mathcal{F})$ defined as follows:*

- *The input space X is $\mathbb{R}^d$ and the label set Y is a finite set of size k (i.e.$|Y| = k$);*
- *For $y \in Y$, $\mu_y \in \mathbb{R}^d$ denotes the (unknown) center of class y and the true distribution P is:*

$$\begin{cases} (\forall y \in Y)\ P(y) = \frac{1}{k} & \textit{Labels are uniformly distributed} \\ (\forall x \in X)(\forall y \in Y)\ P(x|y) = \frac{1}{(2\pi)^{d/2}} e^{-||x-\mu_y||^2/2} & \textit{Class } y \textit{ is a Gaussian } \mathcal{N}(\mu_y, I) \end{cases} \tag{1}$$

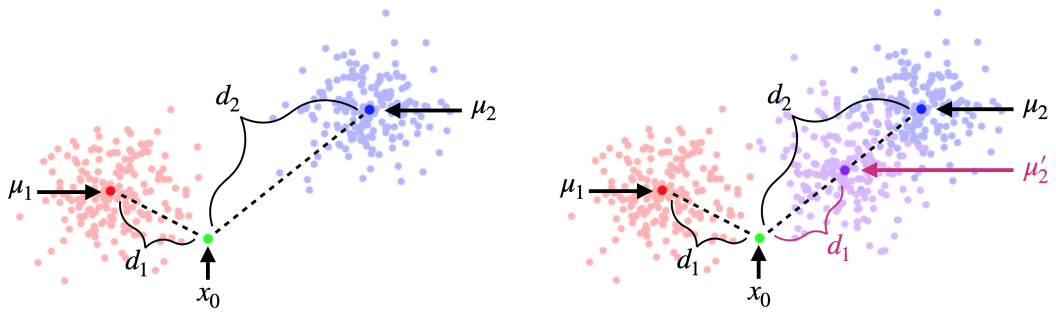
- *The set of plausible learners $\mathcal{F}$ contains all unbiased, non-trivial learners, meaning that for any $f \in \mathcal{F}$, given any x_0, and any plausible P (i.e. the same form as Equation 1 but with potentially different μ_y) where $y_0 = \arg\max_y P(y|x_0)$ is unique, we have $Pr[f_{D_n}(x_0) = y_0] > \frac{1}{k}$ and $(\forall y \neq y_0)\ Pr[f_{D_n}(x_0) = y] \leq \frac{1}{k}$ for all $n \geq 1$.*

First we present a claim regarding the optimal prediction y_0 corresponding to a given input x_0:

Claim 1. *In Isotropic Gaussian Classification, for any x_0, the corresponding maximum likelihood prediction $y_0 = \arg\max_y P(y|x_0)$ is $\arg\min_y ||x_0 - \mu_y||$.*

This follows directly from Equation 1: $y_0 = \arg\max_y P(y|x_0) = \arg\max_y \frac{1}{\sqrt{2\pi}} e^{-||x-\mu_y||^2/2} = \arg\min_y ||x - \mu_y||$. Without loss of generality and with a slight abuse of notation, we assume in the rest of Section 6 that μ_1, μ_2 are respectively the closest and the second closest class centers to x_0 and therefore $y_0 = 1$ is the optimal prediction. Let $d_1 = ||x_0 - \mu_1||$ and $d_2 = ||x_0 - \mu_2||$ be the distances of μ_0 and μ_1 to x_0. An illustration is included in Figure 1(a).

First we define a parameter Δ that will later help us to analyze both clean learning and poisoned learning of Isotropic Gaussian Classification: Δ is the total variation distance between two isotropic Gaussian distribution with centers separated by a distance of $d_2 - d_1$, i.e. $\Delta = \delta(U, V)$ for some $U = \mathcal{N}(\mu, I)$ and $V = \mathcal{N}(\mu', I)$ where $||\mu - \mu'|| = d_2 - d_1$. We introduce a lemma proved by Devroye et al. [9].



(a) Clean (b) Poisoned

Figure 1: Illustrations of Isotropic Gaussian Classification under clean learning and poisoned learning, where μ_1, μ_2 are centers for clean training samples with labels 1 (red) and 2(blue); μ_2' is the center for the poisoned training samples with labels 2 (purple). Note that the illustrations contain only the **two closest classes** to x_0 and the setting contains $k \geq 2$ classes.

Lemma 5 (Total variation distance for Gaussians with the same covariance[9])**.** *For two Gaussian distribution $U = \mathcal{N}(\mu, \Sigma)$ and $V = \mathcal{N}(\mu', \Sigma)$, their total variation distance is $\delta(U, V) = Pr[\mathcal{N}(0,1) \in [-r, r]]$ where $r = \sqrt{(\mu - \mu')^T \Sigma^{-1} (\mu - \mu')}/2$.*

Thus $\Delta = Pr\left[|\mathcal{N}(0,1)| \leq d_2 - d_1\right] = \text{erf}\left((d_2 - d_1)/\sqrt{2}\right)$ where $\text{erf}(x) = (2/\sqrt{\pi}) \cdot \int_0^x e^{-t^2} dt$ is the Gaussian error function.

Lemma 6 (Clean Learning of Isotropic Gaussian Classification)**.** *In Isotropic Gaussian Classification, given a constant $\tau \in (1/2, 1)$, for any input $x_0 \in X$ and the corresponding maximum likelihood prediction $y_0 = \arg\max_y P(y|x_0)$, let n be the smallest training set size such that there exists a learner $f \in \mathcal{F}$ with $Pr[f_{D_n}(x_0) = y_0] \geq \tau$. Then $n \geq \Theta(k/\Delta)$, i.e. <u>the most data-efficient learner takes at least $\Theta(k/\Delta)$ clean samples for an accurate prediction.</u>*

Lemma 7 (Poisoned Learning of Isotropic Gaussian Classification)**.** *In Isotropic Gaussian Classification, for any input $x_0 \in X$ and the corresponding maximum likelihood prediction $y_0 = \arg\max_y P(y|x_0)$, for any given training set size N, and any learner $f \in \mathcal{F}$, there is a mapping $T : \Omega^{\mathbb{N}} \to \Omega^{\mathbb{N}}$ with $Pr[f_{T(D_N)}(x_0) = y_0] \leq 1/|Y|$ while $\mathbb{E}[|T(D_N) - D_N|] \leq \Theta(\Delta/k) \cdot N$, i.e. <u>poisoning $\Theta(\Delta/k)$ of the entire training set is sufficient to break any defense.</u>*

The proofs of Lemma 6 and Lemma 7 are included respectively in Appendix G and H. Intuitively, what we do is to construct a second, perfectly legit distribution that is not far from the original one (measured with the total variation distance), so that any classifier must either fail on the original one or fail on the one we construct. If it fails on the original one, the adversary achieves its goal even without poisoning the training set. If it fails on the one we construct, the adversary can still succeed by poisoning only a limited fraction of the training set because the distribution we construct is close to the original one (measured with total variation distance).

Through the lemmas, we show that for Isotropic Gaussian Classification, the most data-efficient learner takes at least $\Theta(k/\Delta)$ clean samples to predict accurately and any defense can be broken by poisoning $\Theta(\Delta/k)$ of the training set, once again matching the statement of Lethal Dose Conjecture.

7 Practical Implications

7.1 (Asymptotic) Optimality of DPA and Finite Aggregation

In this section, we highlight an important implication of the conjecture: Deep Partition Aggregation [20] and its extension, Finite Aggregation [34] are (asymptotically) optimal—Using the most data-efficient learner, they construct defenses approaching the upper bound of robustness indicated by Lethal Dose Conjecture within a constant factor.

Deep Partition Aggregation [20]: DPA predicts through the majority votes of base learners trained from disjoint data. Given the number of partitions k as a hyperparameter, DPA is a learner constructed

using a deterministic base learner f and a hash function $h : X \times Y \to [k]$ mapping labeled samples to integers between 0 and $k-1$. The construction is:

$$\text{DPA}_D(x_0) = \arg\max_y \text{DPA}_D(x_0, y) = \arg\max_{y \in \mathcal{Y}} \frac{1}{k} \sum_{i=0}^{k-1} \mathbb{1}\left[f_{P_i}(x_0) = y\right],$$

where $P_i = \{(x, y) \in D \mid h(x, y) = i\}$ is a partition containing all training samples with a hash value of i and $\text{DPA}_D(x_0, y) = \frac{1}{k}\sum_{i=0}^{k-1} \mathbb{1}\left[f_{P_i}(x_0) = y\right]$ denotes the average votes count for class y. Ties are broken by returning the smaller class index in $\arg\max$.

Theorem 3 (Certified Robustness of DPA against Data Poisoning [20])**.** *Given a training set D and an input x_0, let $y_0 = \textit{DPA}_D(x_0)$, then for any training set D' with*

$$|D - D'| \leq \frac{k}{2}\left(\textit{DPA}_D(x_0, y_0) - \max_{y \neq y_0}\left(\textit{DPA}_D(x_0, y) + \frac{\mathbb{1}\left[y < y_0\right]}{k}\right)\right) \tag{2}$$

we have $\textit{DPA}_D(x_0) = \textit{DPA}_{D'}(x_0)$, meaning the prediction remains unchanged with data poisoning.

Let n be the average size of partitions, i.e. $n = \sum_{i=0}^{k-1} |P_i|/k = N/k$ where $N = |D|$ is the size of the training set. Assuming the hash function $h : X \times Y \to [k]$ uniformly distributes samples into different partitions, we have $\text{DPA}_D(x_0, y) \approx Pr[f_{D_n}(x_0) = y]$ and therefore the right hand side of Equation 2 approximates

$$\frac{Pr[f_{D_n}(x_0) = y_0] - \max_{y \neq y_0} Pr[f_{D_n}(x_0) = y]}{2n} \cdot N = \Theta\left(\frac{1}{n}\right) \cdot N$$

when the base learner f offers a margin $\Delta = Pr[f_{D_n}(x_0) = y_0] - \max_{y \neq y_0} Pr[f_{D_n}(x_0) = y] \geq \tau$ with n samples for some constant τ. When f is the most data-efficient learner taking n clean samples with a margin $\Delta \geq \tau$ for an input x_0, then given a potentially poisoned, size-N training set is given, DPA (using f as the base learner) tolerates $\Theta\left(N/n\right)$ poisoned samples, approaching the upper bound of robustness indicated by Lethal Dose Conjecture within a constant factor.

The analysis for Finite Aggregation (FA) is exactly the same by substituting Theorem 3 with the certified robustness of FA. For details, please refer to Theorem 2 by Wang et al. [34].

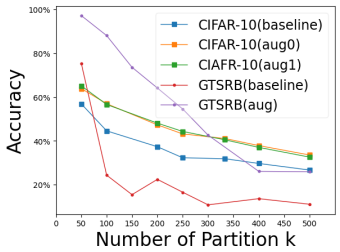
7.2 Better Learners to Stronger Defenses

Assuming the conjecture is true, we have in our hand the (asymptotically) optimal approaches to convert data-efficient learners to strong defenses against poisoning. This reduces developing stronger defenses to finding more data-efficient learners. In this section, as a proof of concept, we show that simply using different data augmentations can increase the data efficiency of base learners and **vastly** improve (*double* or even *triple*) the certified robustness of DPA, essentially highlighting the potential of finding data-efficient learners in defending against data poisoning, which has been a blind spot for our community.

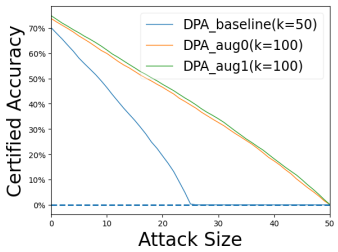
Setup. Following Levine and Feizi [20], we use Network-In-Network[21] architecture for all base learners and evaluate on CIFAR-10[18] and GTSRB[29]. For the baseline (DPA_baseline), we follow exactly their augmentations, learning rates (initially 0.1, decayed by a factor of $1/5$ at 30%, 60%, and 80% of the training process), batch size (128) and total epochs (200). For our results (DPA_aug0 and DPA_aug1 on CIFAR-10; DPA_aug on GTSRB), we use predefined AutoAugment [6] policies for data augmentations, where DPA_aug0 uses the policy for CIFAR-10, DPA_aug1 uses the policy for Imagenet and DPA_aug uses the policy for SVHN, all included in torchvision[24]. We use an initial learning rate of 0.005, a batch size of 16 for 600 epochs on CIFAR-10 and 1000 epochs on GTSRB.

Evaluation. First, we evaluate the test accuracy of base learners with limited data (i.e. using $1/k$ of the entire training set of CIFAR-10 and GTSRB where k ranges from 50 to 500). In Figure 2(a), simply using AutoAugment greatly improves the accuracy of base learners with limited data: On CIFAR-10, with both augmentation policies, the augmented learners achieve similar accuracy as the baseline using only $1/2$ of data (i.e. with k that is twice as large); On GTSRB, the augmented learner achieves similar accuracy as the baseline using only $1/4 \sim 1/3$ of data.

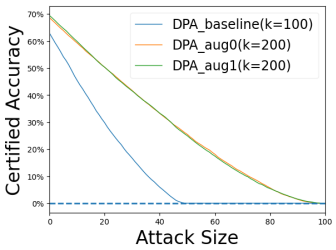
Now we use more data-efficient learners in DPA to construct stronger defenses. For baselines, we use DPA with $k = 50, 100, 250$ on CIFAR-10 and $k = 50, 100$ on GTSRB, defining essentially 5



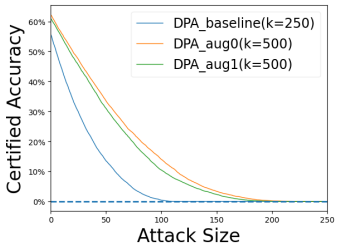
(a) Test accuracy of learners using $1/k$ of the entire training set



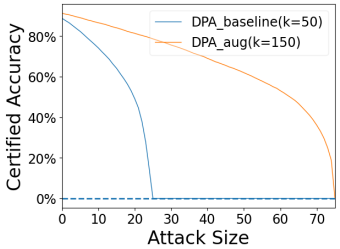
(b) CIFAR-10 (setting 1)



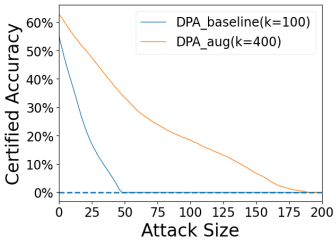
(c) CIFAR-10 (setting 2)



(d) CIFAR-10 (setting 3)

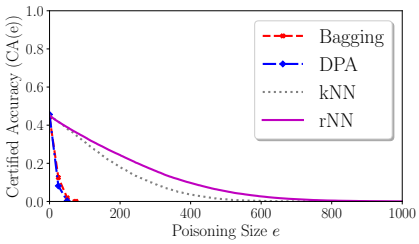


(e) GTSRB (setting 1)

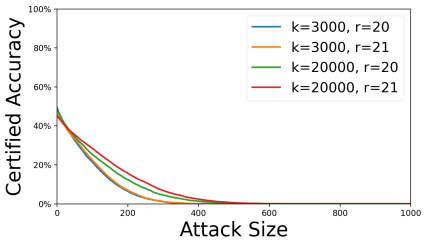


(f) GTSRB (setting 2)

Figure 2: Experiments that construct stronger defenses against data poisoning by using more data-efficient base learners for DPA. By simply using different data augmentations to improve the data efficiency of base learners, the certified robustness can be respectively **doubled** and **tripled** on CIFAR-10 and GTSRB without sacrificing accuracy.



(a) The evaluation conducted by Jia et al. [16]: DPA using neural networks as base learners is outperformed by nearest neighbors.



(b) Our evaluation of DPA: By using radius nearest neighbors as base learners, DPA achieves a similar level of robustness to nearest neighbors defenses [16].

Figure 3: Comparing the certified robustness of DPA with the nearest neighbor defenses[16] on CIFAR-10, suggesting kNN and rNN are no stronger defense than DPA but are using a more data-efficient learner given the accuracy requirements.

settings. When using augmented learners in DPA, we increase k accordingly so that the accuracy of base learners remains at the same level as the baselines, as indicated by Figure 2(a). On CIFAR-10, in Figure 2(b), 2(c) and 2(d), the attack size tolerated by DPA is roughly doubled using both augmented learners for similar certified accuracy. On GTSRB, in Figure 2(e) and 2(f), the attack size tolerated by DPA is more than tripled using the augmented learner for similar certified accuracy.

The improvements are quite significant, corroborating the potential of this practical approach—developing stronger defenses against poisoning through finding more data-efficient learners.

7.3 Better Learners from Stronger Defenses

Assuming the conjecture is true, when some defense against data poisoning is clearly more robust than DPA (even in restricted cases), there should be a more data-efficient learner. Jia et al. [16]

propose to use nearest neighbors, i.e. kNN and rNN, to certifiably defend against data poisoning. Interestingly, in their evaluations, kNN and rNN tolerate much more poisoned samples than DPA when the required clean accuracy is low, as shown in Figure 3(a). Here we showcase how a learner is derived from their defense, which, when being used as base learners, boosts the certified robustness of DPA to a similar level to the nearest neighbors defenses.

In [16], histogram of oriented gradients (HOG) [7] is used as the predefined feature space to estimate the distance between samples. Consequently, we use *radius nearest neighbors* over HOG features as the base learner of DPA. Given a threshold r, radius nearest neighbors find the majority votes of training samples within an ℓ_1 distance of r to the test sample. When there is no training sample with a distance of r, we let the base learner output a token $\perp$ denoting outliers so that it will not affect the aggregation. The results are included in Figure 3(b), where DPA offers similar robustness curves as kNN and rNN in Figure 3(a) by using the aforementioned base learner and hence the same prior.

8 Conclusion

In this work, we propose **Lethal Dose Conjecture**, which characterizes the largest amount of poisoned samples any defense can tolerate for a specific task. We prove the conjecture for multiple cases and offer general theoretical insights through distribution discrimination. The conjecture implies the (asymptotic) optimality of DPA [20] and FA [34] in a sense that they can transfer the most data-efficient learners to one of the most robust defenses, revealing a practical approach to obtaining stronger defenses via improving data efficiency of (non-robust) learners. Empirically, as a proof of concepts, we show that simply using different data augmentations can increase the data efficiency of base learners, and therefore respectively **double** and **triple** the certified robustness of DPA on CIFAR-10 and GTSRB. This highlights both the importance of Lethal Dose Conjecture and the potential of the practical approach in searching for stronger defenses.

Acknowledgements

This project was supported in part by NSF CAREER AWARD 1942230, a grant from NIST 60NANB20D134, HR001119S0026 (GARD), ONR YIP award N00014-22-1-2271, Army Grant No. W911NF2120076 and the NSF award CCF2212458.

References

[1] Hojjat Aghakhani, Dongyu Meng, Yu-Xiang Wang, Christopher Kruegel, and Giovanni Vigna. Bullseye polytope: A scalable clean-label poisoning attack with improved transferability. In IEEE European Symposium on Security and Privacy, EuroS&P 2021, Vienna, Austria, September 6-10, 2021, pages 159–178. IEEE, 2021. doi: 10.1109/EuroSP51992.2021.00021. URL https://doi.org/10.1109/EuroSP51992.2021.00021.

[2] Pranjal Awasthi, Maria-Florina Balcan, and Philip M. Long. The power of localization for efficiently learning linear separators with noise. J. ACM, 63(6):50:1–50:27, 2017. doi: 10.1145/3006384. URL https://doi.org/10.1145/3006384.

[3] Maria-Florina Balcan, Avrim Blum, Steve Hanneke, and Dravyansh Sharma. Robustly-reliable learners under poisoning attacks. In Po-Ling Loh and Maxim Raginsky, editors, Conference on Learning Theory, 2-5 July 2022, London, UK, volume 178 of Proceedings of Machine Learning Research, pages 4498–4534. PMLR, 2022. URL https://proceedings.mlr.press/v178/balcan22a.html.

[4] Ruoxin Chen, Jie Li, Chentao Wu, Bin Sheng, and Ping Li. A framework of randomized selection based certified defenses against data poisoning attacks. CoRR, abs/2009.08739, 2020. URL https://arxiv.org/abs/2009.08739.

[5] Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. Targeted backdoor attacks on deep learning systems using data poisoning. CoRR, abs/1712.05526, 2017. URL http://arxiv.org/abs/1712.05526.

[6] Ekin D. Cubuk, Barret Zoph, Dandelion Mané, Vijay Vasudevan, and Quoc V. Le. Autoaugment: Learning augmentation strategies from data. In IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019, pages 113–123. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.00020. URL http://openaccess.thecvf.com/content_CVPR_2019/html/Cubuk_AutoAugment_Learning_Augmentation_Strategies_From_Data_CVPR_2019_paper.html.

[7] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005), 20-26 June 2005, San Diego, CA, USA, pages 886–893. IEEE Computer Society, 2005. doi: 10.1109/CVPR.2005.177. URL https://doi.org/10.1109/CVPR.2005.177.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), pages 4171–4186, 2019. URL https://www.aclweb.org/anthology/N19-1423/.

[9] Luc Devroye, Abbas Mehrabian, and Tommy Reddad. The total variation distance between high-dimensional gaussians. arXiv preprint arXiv:1810.08693, 2018.

[10] Ilias Diakonikolas, Gautam Kamath, Daniel Kane, Jerry Li, Jacob Steinhardt, and Alistair Stewart. Sever: A robust meta-algorithm for stochastic optimization. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA, volume 97 of Proceedings of Machine Learning Research, pages 1596–1606. PMLR, 2019. URL http://proceedings.mlr.press/v97/diakonikolas19a.html.

[11] Ji Gao, Amin Karbasi, and Mohammad Mahmoody. Learning and certification under instance-targeted poisoning. In Cassio P. de Campos, Marloes H. Maathuis, and Erik Quaeghebeur, editors, Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence, UAI 2021, Virtual Event, 27-30 July 2021, volume 161 of Proceedings of Machine Learning Research, pages 2135–2145. AUAI Press, 2021. URL https://proceedings.mlr.press/v161/gao21b.html.

[12] Jonas Geiping, Liam H. Fowl, W. Ronny Huang, Wojciech Czaja, Gavin Taylor, Michael Moeller, and Tom Goldstein. Witches' brew: Industrial scale data poisoning via gradient matching. In 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021. OpenReview.net, 2021. URL https://openreview.net/forum?id=01olnfLIbD.

[13] Micah Goldblum, Dimitris Tsipras, Chulin Xie, Xinyun Chen, Avi Schwarzschild, Dawn Song, Aleksander Madry, Bo Li, and Tom Goldstein. Dataset security for machine learning: Data poisoning, backdoor attacks, and defenses. CoRR, abs/2012.10544, 2020. URL https://arxiv.org/abs/2012.10544.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016.

[15] Sanghyun Hong, Varun Chandrasekaran, Yigitcan Kaya, Tudor Dumitras, and Nicolas Papernot. On the effectiveness of mitigating data poisoning attacks with gradient shaping. CoRR, abs/2002.11497, 2020. URL https://arxiv.org/abs/2002.11497.

[16] Jinyuan Jia, Xiaoyu Cao, and Neil Zhenqiang Gong. Certified robustness of nearest neighbors against data poisoning attacks. CoRR, abs/2012.03765, 2020. URL https://arxiv.org/abs/2012.03765.

[17] Jinyuan Jia, Xiaoyu Cao, and Neil Zhenqiang Gong. Intrinsic certified robustness of bagging against data poisoning attacks. In Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence,

IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021, pages 7961–7969. AAAI Press, 2021. URL https://ojs.aaai.org/index.php/AAAI/article/view/16971.

[18] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009.

[19] David A Levin and Yuval Peres. Markov chains and mixing times, volume 107. American Mathematical Soc., 2017.

[20] Alexander Levine and Soheil Feizi. Deep partition aggregation: Provable defenses against general poisoning attacks. In 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021. OpenReview.net, 2021.

[21] Min Lin, Qiang Chen, and Shuicheng Yan. Network in network. In Yoshua Bengio and Yann LeCun, editors, 2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings, 2014. URL http://arxiv.org/abs/1312.4400.

[22] Yuzhe Ma, Xiaojin Zhu, and Justin Hsu. Data poisoning against differentially-private learners: Attacks and defenses. In Sarit Kraus, editor, Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019, pages 4732–4738. ijcai.org, 2019. doi: 10.24963/ijcai.2019/657. URL https://doi.org/10.24963/ijcai.2019/657.

[23] Saeed Mahloujifar, Dimitrios I. Diochnos, and Mohammad Mahmoody. The curse of concentration in robust learning: Evasion and poisoning attacks from concentration of measure. CoRR, abs/1809.03063, 2018. URL http://arxiv.org/abs/1809.03063.

[24] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, Advances in Neural Information Processing Systems 32, pages 8024–8035. Curran Associates, Inc., 2019. URL http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

[25] Andrea Paudice, Luis Muñoz-González, András György, and Emil C. Lupu. Detection of adversarial training examples in poisoning attacks through anomaly detection. CoRR, abs/1802.03041, 2018. URL http://arxiv.org/abs/1802.03041.

[26] Elan Rosenfeld, Ezra Winston, Pradeep Ravikumar, and J. Zico Kolter. Certified robustness to label-flipping attacks via randomized smoothing. In Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event, volume 119 of Proceedings of Machine Learning Research, pages 8230–8241. PMLR, 2020. URL http://proceedings.mlr.press/v119/rosenfeld20b.html.

[27] Aniruddha Saha, Akshayvarun Subramanya, and Hamed Pirsiavash. Hidden trigger backdoor attacks. In The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020, pages 11957–11965. AAAI Press, 2020. URL https://aaai.org/ojs/index.php/AAAI/article/view/6871.

[28] Ali Shafahi, W. Ronny Huang, Mahyar Najibi, Octavian Suciu, Christoph Studer, Tudor Dumitras, and Tom Goldstein. Poison frogs! targeted clean-label poisoning attacks on neural networks. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada, pages 6106–6116, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/22722a343513ed45f14905eb07621686-Abstract.html.

[29] Johannes Stallkamp, Marc Schlipsing, Jan Salmen, and Christian Igel. Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition. Neural Networks, 32: 323–332, 2012. doi: 10.1016/j.neunet.2012.02.016. URL https://doi.org/10.1016/j.neunet.2012.02.016.

[30] Jacob Steinhardt, Pang Wei Koh, and Percy Liang. Certified defenses for data poisoning attacks. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, pages 3517–3529, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/9d7311ba459f9e45ed746755a32dcd11-Abstract.html.

[31] Brandon Tran, Jerry Li, and Aleksander Madry. Spectral signatures in backdoor attacks. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada, pages 8011–8021, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/280cf18baf4311c92aa5a042336587d3-Abstract.html.

[32] Alexander Turner, Dimitris Tsipras, and Aleksander Madry. Label-consistent backdoor attacks. CoRR, abs/1912.02771, 2019. URL http://arxiv.org/abs/1912.02771.

[33] Bolun Wang, Yuanshun Yao, Shawn Shan, Huiying Li, Bimal Viswanath, Haitao Zheng, and Ben Y. Zhao. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. In 2019 IEEE Symposium on Security and Privacy, SP 2019, San Francisco, CA, USA, May 19-23, 2019, pages 707–723. IEEE, 2019. doi: 10.1109/SP.2019.00031. URL https://doi.org/10.1109/SP.2019.00031.

[34] Wenxiao Wang, Alexander Levine, and Soheil Feizi. Improved certified defenses against data poisoning with (deterministic) finite aggregation. CoRR, abs/2202.02628, 2022.

[35] Maurice Weber, Xiaojun Xu, Bojan Karlas, Ce Zhang, and Bo Li. RAB: provable robustness against backdoor attacks. CoRR, abs/2003.08904, 2020. URL https://arxiv.org/abs/2003.08904.

[36] Wayne Xiong, Jasha Droppo, Xuedong Huang, Frank Seide, Mike Seltzer, Andreas Stolcke, Dong Yu, and Geoffrey Zweig. Achieving human parity in conversational speech recognition. CoRR, abs/1610.05256, 2016. URL http://arxiv.org/abs/1610.05256.

[37] Chen Zhu, W. Ronny Huang, Hengduo Li, Gavin Taylor, Christoph Studer, and Tom Goldstein. Transferable clean-label poisoning attacks on deep neural nets. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA, volume 97 of Proceedings of Machine Learning Research, pages 7614–7623. PMLR, 2019. URL http://proceedings.mlr.press/v97/zhu19a.html.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] In Section 1.
   (c) Did you discuss any potential negative societal impacts of your work? [No]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]

(b) Did you include complete proofs of all theoretical results? [Yes] Some proofs are included in Appendix.

3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [No]
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [No] Both CIFAR-10 and GTSRB are standard benchmarks.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Proof of Lemma 1

Proof. Given the number of clean samples n, for any learner $f \in \mathcal{F}$, the accuracy of f on x_0 is:

$$\begin{aligned} Pr[f_{D_n}(x_0) = y_0] =& Pr[(x_0, y_0) \in D_n] \cdot Pr[f_{D_n}(x_0) = y_0 \mid (x_0, y_0) \in D_n] \\ &+ Pr[(x_0, y_0) \notin D_n] \cdot Pr[f_{D_n}(x_0) = y_0 \mid (x_0, y_0) \notin D_n], \end{aligned} \tag{3}$$

where $Pr[(x_0, y_0) \in D_n] = 1 - (1 - 1/k)^n$ and $Pr[(x_0, y_0) \notin D_n] = (1 - 1/k)^n$.

Since the bijection g is unknown to the learner f, when $(x_0, y_0) \notin D_n$, by symmetry the optimal prediction is predicting an arbitrary label that is not in D_n, thus

$$\begin{aligned} & Pr[f_{D_n}(x_0) = y_0 \mid (x_0, y_0) \notin D_n] \\ \leq & Pr[E \mid (x_0, y_0) \notin D_n] \cdot 1 + Pr[\neg E \mid (x_0, y_0) \notin D_n] \cdot \frac{1}{2} \\ = & Pr[E \mid (x_0, y_0) \notin D_n] \cdot \frac{1}{2} + \frac{1}{2} \\ \leq & \left(1 - \left(1 - \frac{1}{k-1}\right)^n\right) \cdot \frac{1}{2} + \frac{1}{2} \\ = & 1 - \left(1 - \frac{1}{k-1}\right)^n \cdot \frac{1}{2} \end{aligned}$$

where E denotes the event that all other $k - 1$ labels appear in the training set D_n. Above, what we do is to divide the probability into two cases and bound them separately. Case 1 is when E happens, where we simply upper bound the probability that $f_{D_n}(x_0) = y_0$ by 1. Case 2 is when E does not happen, meaning that there is some $y_1 \neq y_0$ that does not appear in D_n. By

Definition 1, we have $Pr[f_{D_n}(x_0) = y_0] = Pr[f_{T_{y_0 \leftrightarrow y_1}(D_n)}(x_0) = y_1] = Pr[f_{D_n}(x_0) = y_1]$ thus $Pr[f_{D_n}(x_0) = y_0] \leq \frac{1}{2}$.

With Equation 3, we have

$$\begin{aligned} Pr[f_{D_n}(x_0) = y_0] \leq & Pr[(x_0, y_0) \in D_n] \cdot 1 + Pr[(x_0, y_0) \notin D_n] \cdot \left(1 - \left(1 - \frac{1}{k-1}\right)^n \cdot \frac{1}{2}\right) \\ = & 1 - Pr[(x_0, y_0) \notin D_n] \cdot \left(1 - \frac{1}{k-1}\right)^n \cdot \frac{1}{2} \\ = & 1 - \left(1 - \frac{1}{k}\right)^n \cdot \left(1 - \frac{1}{k-1}\right)^n \cdot \frac{1}{2}. \end{aligned}$$

Thus $Pr[f_{D_n}(x_0) = y_0] \geq \tau \Rightarrow \left(1 - \frac{1}{k}\right)^n \cdot \left(1 - \frac{1}{k-1}\right)^n \leq 2 - 2\tau \Rightarrow n \geq \frac{\log(2-2\tau)}{\log(1-2/k)} = \Theta(k)$.

The intuition behind the proof: If the training set contains (x_0, y_0), the learner can obviously predict correctly; Otherwise, the best it can do is to guess a label that is not in the training set. □

B Proof of Lemma 2

Proof. Given x_0, for any N and any learner f, one of following two cases must be true:

Case 1: If $Pr[f_{D_N}(x_0) = y_0] \leq \frac{1}{|Y|}$, using the identity transform $T(D) = D$ for all $D \in \Omega^{\mathbb{N}}$, we have $Pr[f_{T(D_N)}(x_0) = y_0] \leq \frac{1}{|Y|}$ and $\mathbb{E}[|T(D_N) - D_N|] = 0$.

Case 2: If $Pr[f_{D_N}(x_0) = y_0] > \frac{1}{|Y|}$, since $\sum_{y \in Y} Pr[f_{D_N}(x_0) = y] = 1$, there exists $y_1 \neq y_0$ such that $Pr[f_{D_N}(x_0) = y_1] \leq \frac{1}{|Y|}$. Let $T = T_{y_0 \leftrightarrow y_1}$ be a transform swapping labels y_0 and y_1, i.e. $T(D)$ is the same as D except that every $(x, y_0) \in D$ will becomes $(x, y_1) \in T(D)$ and every $(x, y_1) \in D$ will becomes $(x, y_0) \in T(D)$. Since $T = T_{y_0 \leftrightarrow y_1}$ is a transform swapping labels y_0 and y_1 in the training set, $E[|T(D_N) - D_N|]$ is in fact the expected number of samples with a label of y_0 or y_1, which is $\frac{2N}{k}$. Thus we have $Pr[f_{T(D_N)}(x_0) = y_0] = Pr[f_{D_N}(x_0) = y_1] \leq \frac{1}{|Y|}$ and $\mathbb{E}[|T(D_N) - D_N|] = \frac{2N}{k} = \Theta(\frac{1}{k}) \cdot N$.

In both cases, we have a transform T that minimize the accuracy of f while in expectation altering no more than $\Theta(1/k)$ of the training set and therefore the proof completes. Note the underlying assumption used in case 2 is that the Bijection g is unknown to learners in a sense that the output distributions of f change accordingly when labels are permuted. □

C Proof of Lemma 3

Proof. Given the number of clean samples n, for any learner $f \in \mathcal{F}$, the accuracy of f on x_0 is:

$$\begin{aligned} Pr[f_{D_n}(x_0) = y_0] = & Pr[(x_0, y_0) \in D_n] \cdot Pr[f_{D_n}(x_0) = y_0 \mid (x_0, y_0) \in D_n] \\ & + Pr[(x_0, y_0) \notin D_n] \cdot Pr[f_{D_n}(x_0) = y_0 \mid (x_0, y_0) \notin D_n], \end{aligned} \tag{4}$$

where $Pr[(x_0, y_0) \in D_n] = 1 - (1 - 1/m)^n$ and $Pr[(x_0, y_0) \notin D_n] = (1 - 1/m)^n$.

Since g is a mapping that assigns labels independently to different inputs and it is unknown to learners, we have $Pr[f_{D_n}(x_0) = y_0 \mid (x_0, y_0) \notin D_n] \leq \frac{1}{k}$ and therefore with Equation 4, we have

$$\begin{aligned} Pr[f_{D_n}(x_0) = y_0] \leq & Pr[(x_0, y_0) \in D_n] \cdot 1 + Pr[(x_0, y_0) \notin D_n] \cdot \frac{1}{k} \\ = & 1 - Pr[(x_0, y_0) \notin D_n] \cdot (1 - \frac{1}{k}) \\ = & 1 - \left(1 - \frac{1}{m}\right)^n \cdot \left(1 - \frac{1}{k}\right). \end{aligned}$$

Thus $Pr[f_{D_n}(x_0) = y_0] \geq \tau \Rightarrow \left(1 - \frac{1}{m}\right)^n \leq \frac{k(1-\tau)}{k-1} \Rightarrow n \geq \frac{\log(1-\tau) + \log(1+1/(k-1))}{\log(1-1/m)} = \Theta(m)$.

The intuition behind the proof: If the training set contains (x_0, y_0), the most data-efficient learner will memorize it to predict correctly; Otherwise it can do nothing but guess an arbitrary label. □

D Proof of Lemma 4

Proof. Given x_0, for any N and any learner f, we let T be a transform that obtain the poisoned training set $T(D)$ by removing all (x_0, y_0) from the clean training set D. By definition of Instance Memorization, there is no information regarding y_0 contained in $T(D_N)$ and therefore we have $Pr[f_{T(D_N)}(x_0) = y_0] =\leq \frac{1}{|Y|}$ given that the mapping g is unknown to learners. Meanwhile, we have $\mathbb{E}[|T(D_N) - D_N|] = \frac{N}{m} = \Theta(\frac{1}{m}) \cdot N$. □

E Proof of Theorem 1

Proof. First we introduce Coupling Lemma as a tool.

Coupling Lemma [19]: *For two distributions U and V over Ω, a coupling W is a distribution over $\Omega \times \Omega$ such that the marginal distributions are the same as U and V, i.e. $U(u) = \int_{v\in\Omega} W(u, v)dv$ and $V(v) = \int_{u\in\Omega} W(u, v)du$. The lemma states that for two distributions U and V, there exists a coupling W such that $Pr_{(u,v)\sim W}[u \neq v] = \delta(U, V)$.*

Intuitively, Coupling Lemma suggests there is a correspondence between the two distributions U and V, such that only the mass within their difference $\delta(U, V)$ will correspond to different elements.

Through coupling lemma, there is a coupling W with $Pr_{(u,v)\sim W}[u \neq v] = \delta(U, V)$.

$$\begin{aligned}
\mathbb{E}_{D\sim U^n}[f(D)] - \mathbb{E}_{D\sim V^n}[f(D)] =& \mathbb{E}_{(\forall 1\leq i\leq n)\ (u_i,v_i)\sim W}\left[f(\{u_i\}_{i=1}^n) - f(\{v_i\}_{i=1}^n)\right] \\
\leq& Pr_{(\forall 1\leq i\leq n)\ (u_i,v_i)\sim W}[(\exists i)\ u_i \neq v_i] \\
\leq& \sum_{i=1}^{n} Pr_{(u_i,v_i)\sim W}[u_i \neq v_i] \\
=& n \cdot \delta(U, V).
\end{aligned}$$

The second line in the above inequalities is derived as follows: When $u_i = v_i$ for all i, we have $f(\{u_i\}_{i=1}^n) - f(\{v_i\}_{i=1}^n) = 0$; When there exists $u_i \neq v_i$ for some i, we have $f(\{u_i\}_{i=1}^n) - f(\{v_i\}_{i=1}^n) \leq 1$ because the output of f is $\{0, 1\}$.

For the third line, we use the union bound. The probability that for at least one i we have $u_i \neq v_i$ is upper bounded by the sum of probability that $u_i \neq v_i$ for all i.

□

F Proof of Theorem 2

Proof. Through coupling lemma, there is a coupling W with $Pr_{(u,v)\sim W}[u \neq v] = \delta(U, V)$. We define the mapping T as follows: For any $D = (u_1, \ldots, u_N) \in \Omega^N$, the output $T(D) = (v_1, \ldots, v_N)$ is obtained by drawing v_i from $W(v \mid u)$ independently for $i = 1 \ldots N$.

For $i = 1 \ldots N$, we have $P(v_i) = \int_{u_i\in\Omega} W(v_i \mid u_i)U(u_i)du_i = \int_{u_i\in\Omega} W(v_i \mid u_i)W(u_i) = V(v_i)$ and $T(U^N)$ is the same distribution as V^N, meaning that $\mathbb{E}_{D\sim U^N}[f(T(D))] - \mathbb{E}_{D\sim V^N}[f(D)] = 0$; Meanwhile, given $Pr_{(u,v)\sim W}[u \neq v] = \delta(U, V)$, we have $\mathbb{E}_{D\sim U^N}[|T(D) - D|] = \sum_{i=1}^{N} Pr_{u\sim U, v\sim W(v|u)}[u \neq v] = \sum_{i=1}^{N} Pr_{(u,v)\sim W}[u \neq v] = \delta(U, V) \cdot N$. □

G Proof of Lemma 6

Proof. We will use Theorem 1. In order to have $Pr[f_{D_n}(x_0) = y_0] \geq \tau$, there must be some $g : \Omega^{\mathbb{N}} \to \{0, 1\}$ discriminating $U = \mathcal{N}(\mu_1, I)$ and $V = \mathcal{N}(\mu_1', I)$ with confidence larger than a constant τ' using in expectation n/k samples, where $\mu_1' = \mu_1 + (d_2 - d_1)/d_1 \cdot (1 + \epsilon)(\mu_1 - x_0)$ for some $\epsilon > 0$. Note that $||\mu_1' - \mu_1|| = (1 + \epsilon)(d_2 - d_1)$ and $||\mu_1' - x_0|| > d_2$. Taking $\epsilon \to 0$, we have $n/k \geq \Theta(1/\delta(U, V)) = \Theta(1/\Delta) \Rightarrow n \geq \Theta(k/\Delta)$ using Theorem 1. □

H Proof of Lemma 7

Proof. We will use Theorem 2. Given x_0, for any N and any learner f, let $\mu_2' = \mu_2 - (d_2 - d_1)/d_2 \cdot (1+\epsilon)(\mu_2 - x_0)$ for some $\epsilon > 0$, as shown in Figure 1(b). With Theorem 2, there is a transform T' making $U = \mathbb{N}(\mu_2, I)$ and $V = \mathbb{N}(\mu_2', I)$ indistinguishable. We define a transform T by applying T' to the inputs of all samples from class 2 while others remain unchanged.

Note that $f \in \mathcal{F}$ and $T(P)$ is also a plausible distribution (in a sense that it can be expressed in the same form as Equation 1). Since $||\mu_2' - x_0|| < d_1$, we have $Pr[f_{T(D_N)}(x_0) = y_0] \leq 1/k$. In addition, since $||\mu_2 - \mu_2'|| = (1+\epsilon)(d_2 - d_1)$ and there are in expectation N/k samples from class 2, we have $\mathbb{E}[|T(D_N) - D_N|] \leq \delta(U, V)/k = \Theta(\Delta/k)$ by taking $\epsilon \to 0$. □

Intuition for taking $\epsilon \to 0$: When ϵ is actually 0, the distributions we construct for different classes will be 'symmetric' to x_0, meaning that there will be a tie in defining the maximum likelihood prediction. For any $\epsilon > 0$, the tie will be broken. By letting $\epsilon \to 0$, we find the tightest bound of the number of poisoned samples needed from our construction.