Improved Certified Defenses against Data Poisoning with (Deterministic) Finite Aggregation

Wenxiao Wang [1] **Alexander Levine** [1] **Soheil Feizi** [1]

Abstract

Data poisoning attacks aim at manipulating model behaviors through distorting training data. Previously, an aggregation-based certified defense, Deep Partition Aggregation (DPA), was proposed to mitigate this threat. DPA predicts through an aggregation of base classifiers trained on disjoint subsets of data, thus restricting its sensitivity to dataset distortions. In this work, we propose an improved certified defense against general poisoning attacks, namely **Finite Aggregation**. In contrast to DPA, which directly splits the training set into *disjoint* subsets, our method first splits the training set into smaller disjoint subsets and then combines duplicates of them to build larger (but not disjoint) subsets for training base classifiers. This reduces the worst-case impacts of poison samples and thus improves certified robustness bounds. In addition, we offer an alternative view of our method, bridging the designs of deterministic and stochastic aggregation-based certified defenses. Empirically, our proposed Finite Aggregation consistently improves certificates on MNIST, CIFAR-10, and GTSRB, boosting certified fractions by up to 3.05%, 3.87% and 4.77%, respectively, while keeping the same clean accuracies as DPA's, effectively establishing a new **state of the art** in (pointwise) certified robustness against data poisoning.

1. Introduction

Over the past years, we have witnessed the increasing popularity of deep learning in a variety of domains including computer vision (He et al., 2016), natural language processing (Devlin et al., 2019), and speech recognition (Xiong et al., 2016). In many cases, such rapid developments depend heavily on the increased availability of data collected from diverse sources, which can be different users or simply websites from all over the Internet. While the richness of data sources greatly facilitates the advancement of deep learning techniques and their applications, it also raises concerns about their *reliability*. This makes the data poisoning threat model, which concerns the reliability of models under adversarially corrupted training samples, more important than ever (Goldblum et al., 2020).

In this work, we use a general formulation of data poisoning attacks as follows: The adversary is given the ability to insert/remove a bounded number of training samples in order to manipulate the predictions (on some target samples) of the model trained from the corresponding training set. Here, the number of samples that the adversary is allowed to insert/remove is referred to as the attack size.

Many variants of empirical poisoning attacks targeting deep neural networks have been proposed, including Feature Collision (Shafahi et al., 2018), Convex Polytope (Zhu et al., 2019), Bullseye Polytope (Aghakhani et al., 2021) and Witches' Brew (Geiping et al., 2021). These attacks are also referred to as triggerless attacks since no modification to the targets is required. Unlike triggerless attacks, backdoor attacks are poisoning attacks that allow modifications of the target samples, for which a variety of approaches have been developed including backdoor poisoning (Chen et al., 2017), label-consistent backdooring (Turner et al., 2019) and hidden-trigger backdooring (Saha et al., 2020). While it is shown in (Schwarzschild et al., 2021) that the evaluation settings can greatly affect the success rate of many data poisoning attacks to deep models, the vulnerability issues against poisoning attacks remain because (i) the current attacks can still succeed in many scenarios, and (ii) stronger adaptive poisoning attacks can potentially be developed in the future, posing practical threats.

In this work, we focus on developing *provably* robust defenses against general poisoning attacks. In particular, aggregation-based techniques, including a deterministic one (Levine & Feizi, 2021) and stochastic ones (Jia et al., 2021; Chen et al., 2020), have been adopted to offer (pointwise)

[1]Department of Computer Science, University of Maryland, College Park, Maryland, USA. Correspondence to: Wenxiao Wang <wwx@umd.edu>.

1st Workshop on Formal Verification of Machine Learning, Baltimore, Maryland, USA. Colocated with ICML 2022.

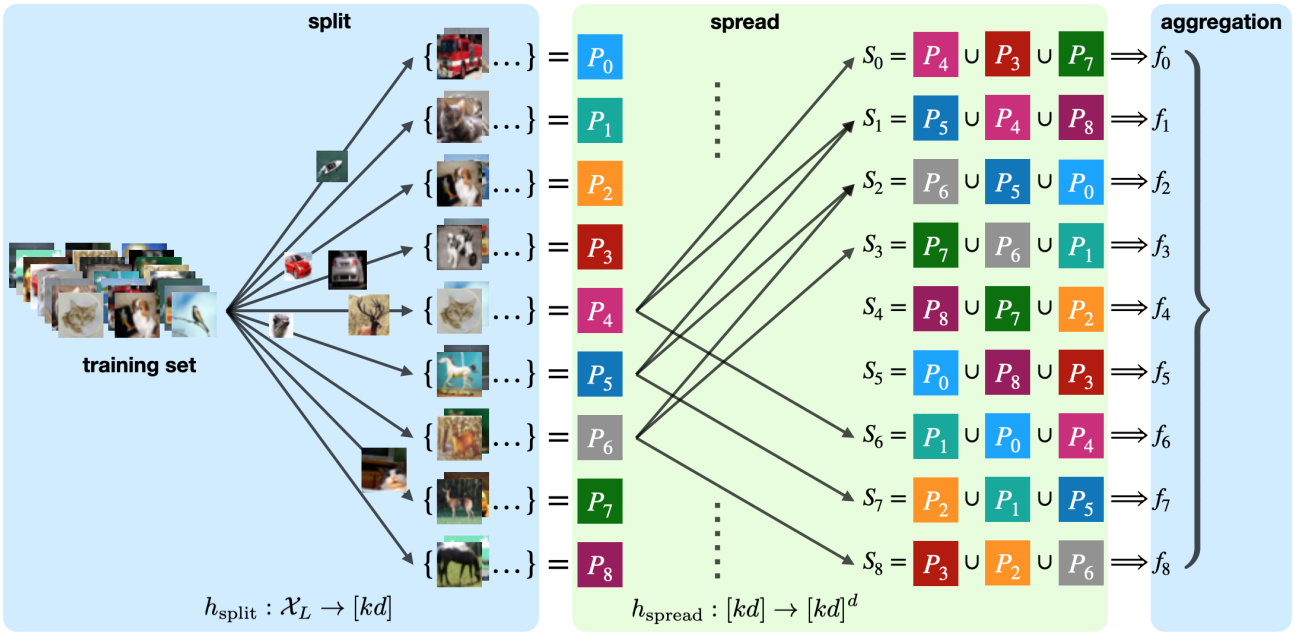


Figure 1. An overview of **Finite Aggregation** with $k = 3$ and $d = 3$. Finite Aggregation consists of three parts: **split**, where the training set is split into kd partitions $P_0, \ldots, P_{kd-1}$ using a hashing function h_{split}; **spread**, where each partition is spread, according to a hash function h_{spread}, to d different destinations from a total of kd subsets $S_0, \ldots, S_{kd-1}$; and **aggregation**, where one classifier is trained from every subset and the majority vote of all kd classifiers will be the prediction at inference time.

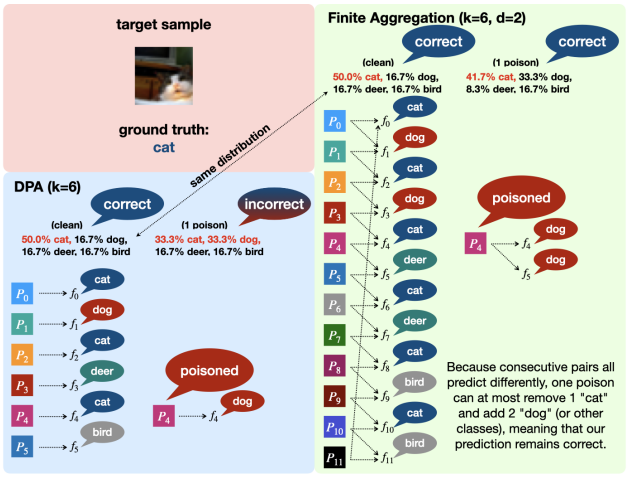


Figure 2. A toy example illustrating how our proposed Finite Aggregation improves provable robustness through a strategic reusing of every sample. For simplicity, we assume that every partition in this example contributes to two consecutive base classifiers in Finite Aggregation. Notably, with no poison, the distribution of predictions from base classifiers are identical for DPA and our method. However, since the subset of base classifiers corresponding to every sample may predict differently, a poisoned sample can be less effective in our method compared to in DPA, leading to improved robustness.

certified robustness against poisoning attacks, where the prediction on every sample is guaranteed to remain unchanged within a certain attack size. Notably, to date, they are state-of-the-art in providing (pointwise) certified robustness against general poisoning attacks. We have other certified defenses against poisoning attacks discussed in Section 2.

In this work, we present **Finite Aggregation**, an advanced aggregation-based defense extended from Deep Partition Aggregation (DPA) (Levine & Feizi, 2021). DPA predicts through an aggregation of base classifiers trained on disjoint subsets of data, thus restricting its sensitivity to dataset distortions. While DPA simply splits the training set into disjoint subsets for training base classifiers, we introduce a novel 'split&spread' protocol to obtain more *overlapping* subsets without changing the average subset size, as illustrated in Figure 1: We first split the training set into smaller disjoint subsets and then combine duplicates of them in a structured fashion to build larger (but **not disjoint**) subsets.

The key idea of our proposed *Finite Aggregation* is based on a strategic **reusing** of every sample to improve robustness. To certify a prediction against data poisoning, an implicit or explicit characterization of the worst-case impact of poisoned samples is inevitable, which can be made more fine-grained by reusing samples strategically to better assess the effectiveness of potentially poisoned samples. In partic-

ular, in our proposed Finite Aggregation method (Figure 1), since the subset of base classifiers corresponding to every sample may predict differently, a poisoned sample can be less effective compared to that of DPA, leading to improved robustness. We further illustrate this using a toy example in Figure 2. In this example, having even one poison sample in DPA can create a tie between prediction probabilities of the correct class ('cat') and an incorrect class ('dog'). However, in Finite Aggregation, with the same clean distribution, even the most effective poison is not able to mislead the model (detailed in Appendix A).

One should note that in the Finite Aggregation method, the total number of base classifiers increases with the reusing of samples, which is why allowing every sample to be used by more base classifiers can actually reduce the impact of poisoned samples. Notably, as d controls the sample reusing in Finite Aggregation, our method essentially degenerates to DPA when $d = 1$ (i.e. no sample reusing).

In summary, our contributions in this work are as follows:

- We propose **Finite Aggregation**, an advanced aggregation-based provable defense against general data poisoning that obtains improved (pointwise) robustness bounds through strategic **reusing** of samples.
- We offer **a novel, alternative view** of our design, to bridge the gap between deterministic defenses (e.g. (Levine & Feizi, 2021)) and stochastic aggregation-based defenses (e.g. (Jia et al., 2021; Chen et al., 2020)).
- Empirically, our method effectively improves certified fractions by up to 3.05%, 3.87% and 4.77% respectively on MNIST, CIFAR-10, and GTSRB, while keeping the same clean accuracies as DPA's, establishing a new **state of the art** in (pointwise) certified robustness against general data poisoning.

2. Related Work

Certified Robustness against Data Poisoning. While in this work we consider pointwise certified robustness, some prior works provide distributional robustness against data poisoning. To name a few, (Steinhardt et al., 2017) derives a high-probability lower bound for test accuracy under poisoning attacks, assuming the distribution of the testing set is the same as the one for the clean training set; (Diakonikolas et al., 2016; Lai et al., 2016) provides distributional robustness guarantees for certain types of unsupervised learning; (Diakonikolas et al., 2019) offers provable approximations of the clean model with additional assumptions regarding the distribution of the clean training data. In addition, (Gao et al., 2021) studies conditions for learnability and certification on predictions under poisoning attacks through the scope of PAC learning; (Wang et al., 2020; Weber et al., 2020) study certified robustness against backdoor attacks; (Rosenfeld et al., 2020) studies certified robustness against label-flipping attacks.

Privacy Attacks. Privacy attacks and data poisoning attacks are actually related. While data poisoning attacks focus on the interests of the consumers of data (e.g. model trainers), privacy attacks focus on the interests of the providers of data (e.g. users). Intuitively, the idea of defending both attacks can be to have models that are not sensitive to the change of a single sample. An active field in privacy-preserving machine learning is to combine deep learning with differential privacy (Abadi et al., 2016; Papernot et al., 2017; 2018; Wang et al., 2021). Some existing works have related differential privacy with poisoning attacks: For example, (Du et al., 2020) proposes an approach of backdoor attack detection via differential privacy and (Ma et al., 2019) investigate the empirical effectiveness of differential privacy in defending against poisoning attacks.

3. (Deterministic) Finite Aggregation

3.1. Notation and Background

Our design, **Finite Aggregation**, is extended from the framework of Deep Partition Aggregation (i.e., DPA) (Levine & Feizi, 2021). In this section, we will go through the notations and the main results of DPA.

Notation: Let $\mathcal{X}$ be the space of unlabeled samples (e.g. the space of images), $\mathcal{C}$ be the set of class indices $[n_c] = \{0, 1, \ldots, n_c - 1\}$, and $\mathcal{X}_L$ be the space of labeled samples $\{(x, c) | x \in \mathcal{X}, c \in \mathcal{C}\}$. A training set D of size n can be viewed as a multiset $\{(x_i, c_i)\}_{i=1}^n$ where $(x_i, c_i) \in \mathcal{X}_L$. We use $\mathcal{D}$ to denote the space of training sets.

For a deterministic classification algorithm $f : \mathcal{D} \times \mathcal{X} \to \mathcal{C}$, $f(D, x) \in \mathcal{C}$ denotes the predicted class index for input $x \in \mathcal{X}$ when the training set is $D \in \mathcal{D}$.

For two training sets D and D', we measure their difference with symmetric distance (the cardinality of symmetric difference):

$$d_{\text{sym}}(D, D') = |(D \setminus D') \cup (D' \setminus D)|,$$

which is exactly the minimum number of samples one needs to insert/remove to change one training set to another (i.e. change D to D' or change D' to D).

DPA (Levine & Feizi, 2021): DPA is a deterministic classification method DPA : $\mathcal{D} \times \mathcal{X} \to \mathcal{C}$ constructed using a deterministic base classifier $f_{\text{base}} : \mathcal{D} \times \mathcal{X} \to \mathcal{C}$ and a hash function $h : \mathcal{X}_L \to [k]$ that maps labeled samples to integers between 0 and $k - 1$ (k is a hyperparameter denoting the

number of partitions). The construction is:

$$\text{DPA}(D, x) = \arg\max_{c \in \mathcal{C}} \sum_{i=0}^{k-1} \mathbb{1}\left[f_{\text{base}}(P_i, x) = c\right],$$

where $P_i = \{(x, c) \in D | h((x, c)) = i\}$ is a partition containing all training samples with a hash value of i. Ties are broken by returning the smaller class index in $\arg\max$. For convenience, we use $\text{DPA}(D, x)_c$ to denote the average votes count $\frac{1}{k}\sum_{i=0}^{k-1} \mathbb{1}\left[f_{\text{base}}(P_i, x) = c\right]$.

Theorem 1 (Robustness of DPA against Data Poisoning) *Given a training set D and an input x, let $c = DPA(D, x)$, then for any training set D', if*

$$\frac{2}{k} \cdot d_{sym}(D, D') \leq DPA(D, x)_c - DPA(D, x)_{c'} - \frac{\mathbb{1}\left[c' < c\right]}{k}$$

holds for all $c' \neq c$, we have $DPA(D, x) = DPA(D', x)$.

Theorem 1[1] (Levine & Feizi, 2021) shows how DPA offers certified robustness against data poisoning attacks. Intuitively, since every sample will be contained in only one partition, one poisoned sample can change at most one vote and therefore reduce normalized margins (the gap between average vote counts) by $\frac{2}{k}$ at most. Thus, the adversary can never change the prediction as long as the number of samples inserted/removed is limited (i.e. d_{sym} is small).

3.2. Proposed Design

In this section, we will present the design of our method, Finite Aggregation. **Finite Aggregation** constructs a new, deterministic classifier using the followings:

- a deterministic base classifier $f_{\text{base}} : \mathcal{D} \times \mathcal{X} \to \mathcal{C}$;
- a hash function $h_{\text{split}} : \mathcal{X}_L \to [kd]$ mapping labeled samples to partition indices between 0 and $kd - 1$, which is used to split the training set into kd partitions;
- a balanced hash function $h_{\text{spread}} : [kd] \to [kd]^d$ mapping every partition index to a set of d different integers of the same range, which is used to spread training samples, allowing them to be utilized by d different base classifiers.

Here, k and d are two hyperparameters where k corresponds to the inverse of sensitivity and the spreading degree d controls the number of base classifiers that every sample can be utilized by.

By a balanced hash function, we mean that the inverse of the hash function $h^{-1}_{\text{spread}}(i) = \{j \in [kd] | i \in h_{\text{spread}}(j)\}$ has the same size (i.e., d elements) for all $i \in [kd]$, which means h^{-1}_{spread} is also a hash function from $[kd]$ to $[kd]^d$. Our choice of h_{spread} will be discussed in Section 3.4.

Definition 1 (Classification with Finite Aggregation) *The construction of Finite Aggregation $FA : \mathcal{D} \times \mathcal{X} \to \mathcal{C}$ is as follows:*

$$FA(D, x) = \arg\max_{c \in \mathcal{C}} \sum_{i=0}^{kd-1} \mathbb{1}\left[f_{base}(S_i, x) = c\right],$$

where $S_i = \bigcup_{j \in h^{-1}_{spread}(i)} P_j$, $P_j = \{(x, c) \in D | h_{split}((x, c)) = j\}$ and ties are broken by returning the smaller class index in $\arg\max$.

Similarly, we use $\text{FA}(D, x)_c$ to denote the the average votes count $\frac{1}{kd}\sum_{i=0}^{kd-1} \mathbb{1}\left[f_{\text{base}}(S_i, x) = c\right]$. We use $\text{FA}(D, x)_{c|j}$ to denote $\frac{1}{d}\sum_{i \in h_{\text{spread}}(j)} \mathbb{1}\left[f_{\text{base}}(S_i, x) = c\right]$, which is the average votes count over base classifiers that utilize P_j.

An overview of Finite Aggregation is in Figure 1. Finite Aggregation can be decomposed into three stages:

- **Split**, where the training set is split into kd partitions $P_0, \ldots, P_{kd-1}$;
- **Spread**, where each partition is spread to d different destinations in $S_0, \ldots, S_{kd-1}$;
- **Aggregation**, where one classifier is trained from every subset S_i, $i \in [kd]$ and the majority vote of all kd classifiers will be the prediction at inference time.

In both DPA (Levine & Feizi, 2021) and our design (with the same hyperparameter k), every base classifier will, on average, have access to $1/k$ of the entire training, and every sample will be utilized by exactly $1/k$ of the base classifiers. However, unlike DPA, which forms disjoint subsets, we let every sample be utilized by d base classifiers in a way that enables better certifications against data poisoning attacks. Notably, when $d = 1$, Finite Aggregation essentially reduces to DPA with the same hyperparameter $\mathbf{k}$.

3.3. Certified Robustness to Data Poisoning

In this section, we will see how Finite Aggregation provably defends against data poisoning attacks and discuss why it offers a stronger defense than DPA.

Theorem 2 (Finite Aggregation against Data Poisoning) *Given a training set D and an input x, let $c = FA(D, x)$, then for any training set D', it is guaranteed that $FA(D', x) = FA(D, x)$ when*

$$\frac{1}{k} \cdot \Delta_{D,x}^{\overline{d_{sym}(D,D')}} \leq FA(D, x)_c - FA(D, x)_{c'} - \frac{\mathbb{1}\left[c' < c\right]}{kd}$$

[1] For coherence, Theorem 1 is presented in a slightly different form from the original one by (Levine & Feizi, 2021).

holds for all $c' \neq c$, where $\Delta_{D,x}$ is a multiset defined as

$$\left\{1 + FA(D,x)_{c|j} - FA(D,x)_{c'|j}\right\}_{j\in[kd]}$$

and $\Delta_{D,x}^{\overline{d_{sym}(D,D')}}$ denotes the sum of the largest $d_{sym}(D,D')$ elements in the multiset $\Delta_{D,x}$.

Theorem 2 is how Finite Aggregation offers certified robustness against data poisoning. The detailed proof is in Appendix B and we include a sketch here.

When one sample x' is inserted or removed, only a specific set of d base classifiers may be affected, depending on where this sample is assigned to in the split stage (i.e. the value of $h_{\text{split}}(x')$). If the goal is to change the prediction from c to c', then the worst case is simply that all of those d classifiers will predict c' after the insertion/removal, meaning that the contribution to the margin between class c and c' will reduce by $\frac{2}{kd}$ for every base classifier (among those d) that originally predicts c, by 0 for every that predicts c' and by $\frac{1}{kd}$ for every base classifier that predicts other classes. Thus, the margin will be reduced by at most $\frac{1}{k}\left(1 + \text{FA}(D,x)_{c|j} - \text{FA}(D,x)_{c'|j}\right)$ given $h_{\text{split}}(x') = j$ and therefore with $d_{\text{sym}}(D,D')$ samples inserted/removed, the margin will be reduced by $\frac{1}{k}\Delta_{D,x}^{\overline{d_{\text{sym}}(D,D')}}$ at most, which means the prediction will not be turned from c to c' as long as the margin is larger than this.

Comparing Theorem 2 with the certified robustness of DPA in Theorem 1, we can directly see why FA offers stronger defenses. A hypothesis here is that with the same k, the accuracies of base classifiers in Finite Aggregation will not change much from those in DPA, since they have access to the same amount of training data on average and the subsets are constructed in a similar fashion. We have this verified empirically in Section 5.2.

With this hypothesis, we can focus on the left hand side of Theorem 2 and Theorem 1 to compare the robustness of ours with DPA. By definition $1 + \text{FA}(D,x)_{c|j} - \text{FA}(D,x)_{c'|j} \leq 2$ holds for any $j \in [kd]$, which means we always have $\frac{1}{k} \cdot \Delta_{D,x}^{\overline{d_{\text{sym}}(D,D')}} \leq \frac{2}{k} \cdot d_{\text{sym}}(D,D')$; thus Finite Aggregation offers better certificates than DPA. The intuition behind this is that when we let a sample be utilized by more than one base classifier, we can use the correlation among them to better characterize the capability of the adversary. We will later elaborate more about this insight in Section 4.1.

3.4. Practical Details

The choice of f_{base}. The only requirement to f_{base} is that it should be deterministic which is hardly an issue since most, if not all, classification algorithms can be made deterministic. Following (Levine & Feizi, 2021), here we use deep neural networks for the base classifiers, where the labeled training samples are sorted in lexicographic order to remove the dependence on the order of the training set and random seeds are set explicitly. More details including model architectures can be found in Section 5.1.

The choice of h_{split}. Here we use the same hash function as (Levine & Feizi, 2021) for evaluation, where it simply maps each sample to $[kd]$ according to the remainder when you divide the sum of pixel values by kd.

The choice of h_{spread}. Here we want h_{spread} to be a balanced hash function mapping every integer in $[kd]$ to a set of d different targets in $[kd]$, where every candidate in the target space will has a preimage of the same size (i.e. d). The construction of h_{spread} used in this paper is as follows:

$$h_{\text{spread}}(j) = \{(j + r_t) \mod kd | t \in [d]\},$$

where $R = \{r_0, \ldots, r_{d-1}\}$ is a size-d subset of $[kd]$ generated using a pseudo-random generator with a fixed random seed. One can easily verify that this is a balanced hash function when R is any size-d subset of $[kd]$.

4. Analysis and Extension

4.1. Relating to Infinite Aggregation

Previously in Section 3, we present Finite Aggregation as an extension of DPA (Levine & Feizi, 2021). In this section, we offer an alternative view that relates Finite Aggregation to Infinite Aggregation, which suggests the advantages of Finite Aggregation (that reduces to DPA when $d = 1$) over Randomized Selection (Jia et al., 2021; Chen et al., 2020), a branch of stochastic certified defenses against data poisoning attacks.

This insight is consistent with the observations that DPA (and therefore Finite Aggregation) typically works better than Randomized Selection empirically. For instance, on CIFAR-10, DPA (Levine & Feizi, 2021) can certify, with **no error rate**, more than 46% of the testing samples correctly when allowing 10 poisons and about 34% when allowing 20 poisons, while the fractions from Randomized Selection (Jia et al., 2021; Chen et al., 2020) are less than 40% and 25% with **an error rate of 0.1**$\%$ respectively for 10 and 20 poisons. Despite the varying details (e.g. (Jia et al., 2021; Chen et al., 2020) refer to a somewhat more general threat model than (Levine & Feizi, 2021)), all three variants are capable of dealing with supposedly the most practical threat model (i.e. poison insertions) and the gaps between DPA and the stochastic variants are fairly significant. In addition, since the error rate from Randomized Selection is only bounded for a test set of finite size, the total error rate inevitably accumulates in deployments, where the number of test samples may increase unboundedly through time. These are also why in Section 5.2, we benchmark Finite Aggregation against DPA to show that ours is indeed a better certified defense.

Let us take another look at the design of Finite Aggregation in Definition 1: What will happen if we let $d \to \infty$? Assuming the hash functions are random, we will have an infinite amount of subsets S_i, where every training sample will be spread to exactly $\frac{1}{k}$ of them independently. This is exactly Infinite Aggregation, described as follows:

Definition 2 (Classification with Infinite Aggregation) *Given a base classifier $f_{base} : \mathcal{D} \times \mathcal{X} \to \mathcal{C}$, the Infinite Aggregation classifier IA $: \mathcal{D} \times \mathcal{X} \to \mathcal{C}$ is defined as follows:*

$$IA(D, x) = \arg\max_{c \in \mathcal{C}} Pr_{S \sim Bernoulli(D, \frac{1}{k})} \left[f_{base}(S, x) = c \right],$$

where Bernoulli$(D, \frac{1}{k})$ denotes Bernoulli sampling from D with sampling rate $\frac{1}{k}$, meaning that every sample in D will be picked independently with a probability of $\frac{1}{k}$. Ties are broken by returning the smaller class index in $\arg\max$.

For simplicity, we use $\mathrm{IA}(D, x)_c$ to denote the expected votes count $Pr_{S \sim \mathrm{Bernoulli}(D, \frac{1}{k})} \left[f_{\mathrm{base}}(S, x) = c \right]$. We use $\mathrm{IA}(D, x)_{c|(x_L)}$ to denote the expected votes count given that sample $x_L \in D$ is utilized, which can be expressed formally as $Pr_{S \sim \mathrm{Bernoulli}(D \setminus \{x_L\}, \frac{1}{k})} \left[f_{\mathrm{base}}(S \cup \{x_L\}, x) = c \right]$.

Infinite Aggregation in fact is the same classifier as Binomial selection from (Chen et al., 2020). However, when we adapt the certificate from Finite Aggregation, it is quite different from theirs:

Theorem 3 (Infinite Aggregation against Data Poisoning) *Given a training set D and an input x, let $c = IA(D, x)$, for any training set D', it is guaranteed that $IA(D', x) = IA(D, x)$ when for any $\delta > 0$,*

$$\frac{1}{k} \cdot \overline{\Delta}_{D,x}^{\overline{d_{sym}(D,D')}} \leq IA(D, x)_c - IA(D, x)_{c'} - \mathbb{1}\left[c' < c\right] \cdot \delta$$

holds for all $c' \neq c$, where $\overline{\Delta}_{D,x}$ is a multiset defined as

$$\{1 + IA(D, x)_{c|x_L} - IA(D, x)_{c'|x_L}\}_{x_L \in D} + \{1 + IA(D, x)_c - IA(D, x)_{c'}\} \times \infty$$

and $\overline{\Delta}_{D,x}^{\overline{d_{sym}(D,D')}}$ denotes the sum of the largest $d_{sym}(D, D')$ elements in the multiset $\overline{\Delta}_{D,x}$. Here $+$ denotes the sum of two multisets and $\{1 + IA(D, x)_c - IA(D, x)_{c'}\} \times \infty$ denotes the multiplication of a multiset and a scalar, which is in this case a multiset containing an infinite amount of a single value, i.e. $1 + IA(D, x)_c - IA(D, x)_{c'}$.

The detailed proof is presented in Appendix C. Note that the purpose for Theorem 3 is not to propose another defense but to connect and unify different aggregation-based defenses to date. For (Jia et al., 2021; Chen et al., 2020), their certificates involve only the margin (e.g. $\mathrm{IA}(D, x)_c - \mathrm{IA}(D, x)_{c'}$) but nothing else that depends on the behavior of the base classifiers, while ours takes the advantage of fine-grained statistics through $\overline{\Delta}_{D,x}$, allowing a closer estimation of the adversary's capability.

Besides, since the number of terms in Definition 2 is exponential to the training set size, Infinite Aggregation is impractical to compute exactly that an approximation of some sort will be unavoidable. (Jia et al., 2021; Chen et al., 2020) do their approximations by numerically estimating the margin with Monte-Carlo methods, which introduces a probability of estimation errors that inevitably accumulates with the number of testing samples. This is not very efficient and as mentioned previously can be an issue for online services where the number of testing samples is potentially unbounded.

Finite Aggregation, however, approximates the entire scheme of Infinite Aggregation in a deterministic fashion, enabling utilizing fine-grained statistics with no error rate. This is not only a theoretical advantage of Finite Aggregation but may also partially explain why DPA (i.e. Finite Aggregation with $d = 1$) typically works better than Randomized Selection.

5. Evaluation

5.1. Evaluation Setup

We follow exactly the setup of (Levine & Feizi, 2021) in the experiments and use the same hyperparameters as theirs. As mentioned in Section 4.1, we compare our method with DPA since it is empirically the state of the art in certified defenses against general poisoning attacks.

Datasets. We evaluate our method on MNIST (LeCun et al., 1998), CIFAR-10 (Krizhevsky, 2009) and GTSRB (Stallkamp et al., 2012) datasets, which are respectively 10-way classification of handwritten digits, 10-way object classification and 43-way classification of traffic signs.

Training hyperparameters. We use the Network-In-Network (Lin et al., 2014) architecture, trained with the hyperparameters from (Gidaris et al., 2018). On MNIST and GTSRB, we also exclude horizontal flips in data augmentations as in (Levine & Feizi, 2021).

5.2. Certified Predictions

We use *certified fraction* as our performance metric, which refers to the fraction of samples in the testing set that are not only correctly classified but also certified to be robust given a certain attack size. This is the same metric as 'certified accuracy' in (Levine & Feizi, 2021) but we choose to refer to it differently. Our motivation is discussed in Section 5.4.

Improved certified robustness. We report the certified

Table 1. Certified fraction of Finite Aggregation with various hyperparameters with respect to different attack sizes d_{sim}. The improvements compared to the DPA baseline are highlighted in blue if they are positive and red otherwise. Note that there is no direct correspondence between base classifiers with different k and d, resulting in artifacts in the visualizations: See Figure 4 for the improvements using the same set of base classifiers.

dataset	k	d	certified fraction				
MNIST	1200		$d_{\text{sim}} \leq 50$	$d_{\text{sim}} \leq 100$	$d_{\text{sim}} \leq 200$	$d_{\text{sim}} \leq 300$	$d_{\text{sim}} \leq 400$
		1 (DPA)	94.12%	92.11%	86.45%	77.12%	61.78%
		8	94.38%(+0.26%)	92.45%(+0.34%)	86.97%(+0.52%)	77.31%(+0.19%)	61.81%(+0.03%)
		16	94.54%(+0.42%)	92.75%(+0.64%)	87.89%(+1.44%)	78.91%(+1.79%)	62.42%(+0.64%)
		32	94.63%(+0.51%)	92.97%(+0.86%)	88.49%(+2.04%)	80.17%(+3.05%)	64.34%(+2.56%)
CIFAR-10	50		$d_{\text{sim}} \leq 3$	$d_{\text{sim}} \leq 5$	$d_{\text{sim}} \leq 10$	$d_{\text{sim}} \leq 15$	$d_{\text{sim}} \leq 20$
		1 (DPA)	63.15%	58.07%	46.44%	33.46%	19.36%
		5	63.55%(+0.40%)	59.01%(+0.94%)	46.62%(+0.18%)	33.56%(+0.10%)	19.63%(+0.27%)
		8	64.07%(+0.92%)	59.80%(+1.73%)	47.40%(+0.96%)	33.76%(+0.30%)	19.72%(+0.36%)
		16	64.80%(+1.65%)	60.55%(+2.48%)	48.85%(+2.41%)	34.61%(+1.15%)	19.90%(+0.54%)
		32	65.40%(+2.25%)	61.31%(+3.24%)	50.31%(+3.87%)	36.03%(+2.57%)	19.93%(+0.57%)
	250		$d_{\text{sim}} \leq 10$	$d_{\text{sim}} \leq 20$	$d_{\text{sim}} \leq 30$	$d_{\text{sim}} \leq 40$	$d_{\text{sim}} \leq 50$
		1 (DPA)	44.31%	34.01%	25.81%	18.99%	13.55%
		3	44.26%(−0.05%)	34.08%(+0.07%)	25.51%(−0.30%)	18.89%(−0.10%)	13.76%(+0.21%)
		5	44.83%(+0.52%)	34.92%(+0.91%)	26.31%(+0.50%)	19.42%(+0.43%)	13.92%(+0.37%)
		8	45.38%(+1.07%)	36.05%(+2.04%)	27.10%(+1.29%)	20.08%(+1.09%)	14.39%(+0.84%)
		16	46.52%(+2.21%)	37.56%(+3.55%)	29.00%(+3.19%)	22.00%(+3.01%)	15.79%(+2.24%)
GTSRB	50		$d_{\text{sim}} \leq 3$	$d_{\text{sim}} \leq 5$	$d_{\text{sim}} \leq 10$	$d_{\text{sim}} \leq 15$	$d_{\text{sim}} \leq 24$
		1 (DPA)	85.09%	82.32%	74.15%	64.14%	14.27%
		8	85.00%(−0.09%)	82.30%(−0.02%)	74.24%(+0.09%)	63.33%(−0.81%)	16.83%(+2.56%)
		16	85.25%(+0.16%)	82.71%(+0.39%)	74.66%(+0.51%)	63.77%(−0.37%)	15.42%(+1.15%)
		32	85.95%(+0.86%)	83.52%(+1.20%)	76.26%(+2.11%)	66.32%(+2.18%)	17.61%(+3.34%)
	100		$d_{\text{sim}} \leq 5$	$d_{\text{sim}} \leq 10$	$d_{\text{sim}} \leq 15$	$d_{\text{sim}} \leq 20$	$d_{\text{sim}} \leq 25$
		1 (DPA)	46.16%	38.24%	30.19%	22.84%	17.16%
		8	47.62%(+1.46%)	40.25%(+2.01%)	32.36%(+2.17%)	24.34%(+1.50%)	17.32%(+0.16%)
		16	48.19%(+2.03%)	41.62%(+3.38%)	33.95%(+3.76%)	25.96%(+3.12%)	18.92%(+1.76%)
		32	48.39%(+2.23%)	42.01%(+3.77%)	34.96%(+4.77%)	27.05%(+4.21%)	19.93%(+2.77%)

fractions of Finite Aggregation in Table 1 and Figure 3. Overall, it is evident that Finite Aggregation can offer strong certified defenses than DPA, where the improvements of certified fractions can be up to 3% or 4% compared to DPA using the same hyperparameters.

Now we examine the effectiveness of our certificates through the scope of certified radius. Given a test sample x, the certified radius is simply the maximum attack size d_{sym} allowed while we can still certify the correct prediction on x. The certified radius is considered to be -1 if the prediction on a test sample does not match the true label.

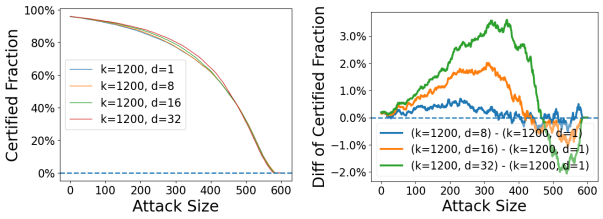
In Table 2, we include two statistics relating to certified radii: $Pr[r \uparrow]$, which denotes the fraction of samples in the testing set that obtain a larger certified radius when using our certificates from Theorem 2 instead of the ones from DPA (i.e. replacing $\Delta_{D,x}^{d_{\text{sym}}(D,D')}$ in Theorem 2 with $2{\cdot}d_{\text{sym}}(D, D')$), and Δr, which denotes the average increase of the certified radius among those getting a larger radius. See Figure 4 for the corresponding certified fraction. The values of $Pr[r \uparrow]$ and Δr in Table 2 are strong supports to the effectiveness of our certificates.

We also report in Table 2 respectively the accuracy of Finite Aggregation $\text{acc}_{\text{clean}}$ and the average accuracy of base classifiers acc_{base}, which supports our hypothesis in Section 3.3 that the accuracies of base classifiers will not change much.

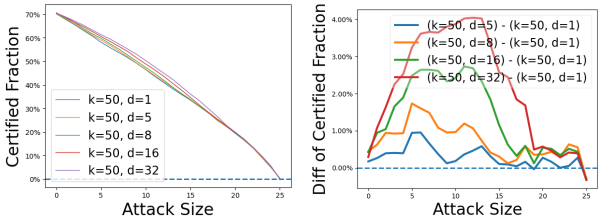
5.3. Hyperparameters for Finite Aggregation

In this section, we discuss how the hyperparameters k and d affect the behaviors of Finite Aggregation in practice.

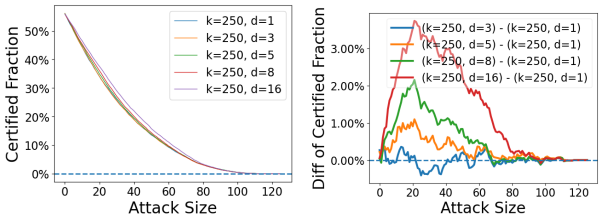
The effect of k: accuracy vs. robustness. Similar to DPA, k corresponds to a trade-off between accuracy and robustness. Since every base classifier will on average have access to $1/k$ of the training set, using a larger k will reduce the accuracies of base classifiers and therefore restrict the accuracy



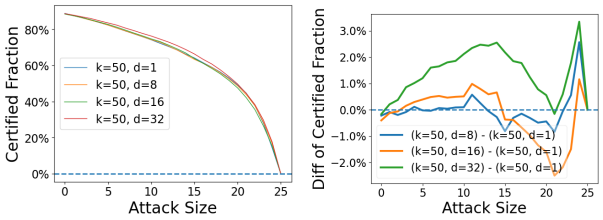
(a) MNIST ($k = 1200$)



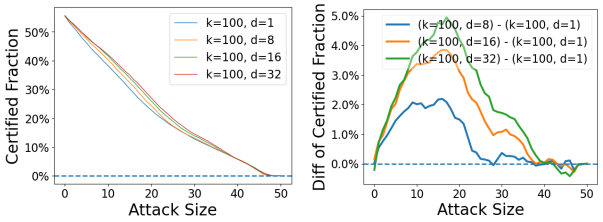
(b) CIFAR-10 ($k = 50$)



(c) CIFAR-10 ($k = 250$)



(d) GTSRB ($k = 50$)



(e) GTSRB ($k = 100$)

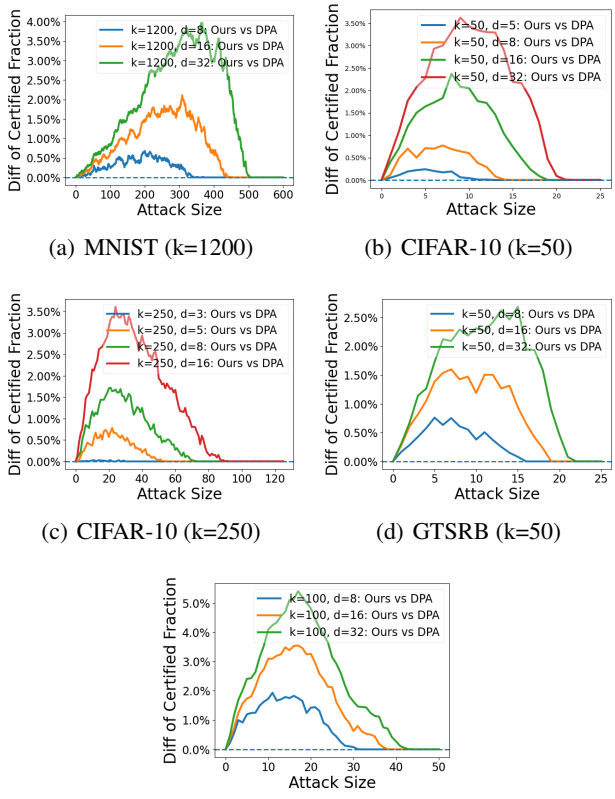
Figure 3. The curves of certified fraction on different datasets. **Left**: Certified fraction of Finite Aggregation with different hyperparameters. **Right**: The improvements of certified fraction from DPA (i.e. Finite Aggregation with $d = 1$) to Finite Aggregation with the same k and various d. Note that there is no direct correspondence between base classifiers with different k and d, resulting in artifacts in the visualizations: See Figure 4 for the improvements using the same set of base classifiers.

Table 2. Some statistics of Finite Aggregation, where $\text{acc}_{\text{clean}}$ denotes the test accuracy with a clean training set; acc_{base} denotes the average accuracy of base classifiers; $Pr[r \uparrow]$ denotes the fraction of samples in the testing set that obtain a larger certified radius when using our certificates from Theorem 2 instead of the ones from DPA; and Δr denotes the average increase of the certified radius among those getting a larger radius.

dataset	k	d	$\text{acc}_{\text{clean}}$	acc_{base}	$Pr[r \uparrow]$	Δr
MNIST	1200	1 (DPA)	95.75%	76.92%	0%	0
		8	95.95%	76.92%	15.72%	6.96
		16	95.90%	76.83%	35.24%	12.69
		32	95.94%	76.54%	58.01%	17.91
CIFAR-10	50	1 (DPA)	70.15%	56.15%	0%	0
		5	70.32%	56.14%	1.49%	1.00
		8	70.59%	56.33%	6.67%	1.00
		16	70.57%	56.32%	22.68%	1.02
		32	70.44%	56.27%	37.92%	1.18
	250	1 (DPA)	55.84%	35.21%	0%	0
		3	56.11%	35.15%	0.33%	1.03
		5	56.06%	35.24%	15.32%	1.30
		8	55.88%	35.18%	36.07%	1.75
		16	55.90%	35.24%	50.40%	3.21
GTSRB	50	1 (DPA)	88.80%	74.47%	0%	0
		8	88.58%	73.71%	6.50%	1.00
		16	88.40%	73.09%	18.58%	1.03
		32	88.64%	73.92%	29.66%	1.20
	100	1 (DPA)	55.56%	34.71%	0%	0
		8	55.58%	34.55%	29.39%	1.16
		16	55.72%	34.58%	41.50%	1.74
		32	55.35%	34.20%	46.71%	2.41

Table 3. Certified fraction ($\text{frac}_{\text{certified}}$) and certified accuracy ($\text{acc}_{\text{certified}}$) of Finite Aggregation corresponding to an attack size of 1. Their differences are highlighted in blue.

dataset	k	d	$\text{acc}_{\text{clean}}$	$\text{frac}_{\text{certified}}$	$\text{acc}_{\text{certified}}$
MNIST	1200	1 (DPA)	95.75%	95.71%	95.71%(+0%)
		8	95.95%	95.91%	95.91%(+0%)
		16	95.90%	95.86%	95.86%(+0%)
		32	95.94%	95.91%	95.91%(+0%)
CIFAR-10	50	1 (DPA)	70.15%	67.85%	68.79%(+0.94%)
		5	70.32%	68.11%	69.07%(+0.96%)
		8	70.59%	68.47%	69.22%(+0.75%)
		16	70.57%	68.79%	69.38%(+0.59%)
		32	70.44%	68.92%	69.34%(+0.42%)
	250	1 (DPA)	55.84%	54.82%	55.19%(+0.37%)
		3	56.11%	54.86%	55.37%(+0.51%)
		5	56.06%	54.91%	55.34%(+0.43%)
		8	55.88%	54.75%	55.13%(+0.38%)
		16	55.90%	55.04%	55.35%(+0.29%)
GTSRB	50	1 (DPA)	88.80%	87.61%	87.99%(+0.38%)
		8	88.58%	87.52%	87.88%(+0.36%)
		16	88.40%	87.48%	87.74%(+0.26%)
		32	88.64%	87.81%	87.97%(+0.16%)
	100	1 (DPA)	55.56%	53.26%	53.96%(+0.70%)
		8	55.58%	53.80%	54.39%(+0.59%)
		16	55.72%	54.01%	54.49%(+0.48%)
		32	55.35%	53.97%	54.25%(+0.28%)



(a) MNIST (k=1200) (b) CIFAR-10 (k=50)

(c) CIFAR-10 (k=250) (d) GTSRB (k=50)

(e) GTSRB (k=100)

Figure 4. The improvements of certified fraction when applying our certificates (Theorem 2) instead of DPA certificates to the same set of $k \cdot d$ base classifiers.

of the aggregation. However, in Theorem 2, the sensitivity of the normalized margin to every poisoned sample is also proportional to $1/k$, suggesting that a larger k can reduce the sensitivity and favor robustness. This is confirmed in Table 1 and Table 2, where a larger k typically corresponds to worse accuracy $\text{acc}_{\text{clean}}$ but may offer a higher certified fraction when the attack sizes are relatively large.

The effect of d: efficiency vs. robustness. The spreading degree d controls how many base classifiers will have access to the same training sample. From Table 1, we can see that increasing d tends to improve the robustness of Finite Aggregation, indicated by the increase of certified fractions. However, unlike k, increasing d will not degrade the accuracy of Finite Aggregation, as observed in Table 2. The major cost of using a larger d is the increase of computation, because the number of base classifiers is proportional to d while the average number of training samples for every base classifier does not depend on d. Thus d actually allows us to obtain stronger robustness at a cost of extra computation.

Another effect of d is that using a larger d tends to avoid ties in the aggregation, as indicated by the term $\frac{\mathbb{1}[c'<c]}{kd}$ in Theorem 2. This can be quite beneficial depending on tasks. For instance, in Figure 3(d) and in Table 1, one can notice extra improvements of certified fraction from DPA for an attack size of 24 on GTSRB with $k = 50$, which exactly attributes to the reduction of ties when d increases.

5.4. Pointwise Robustness vs Distributional Robustness

In this section, we explore a metric corresponding to distributional robustness, namely certified accuracy, that is the lower bound on accuracy in the test set under poisoning attacks. This is different from *certified fraction* in Section 5.2, which denotes the fraction of samples in the testing set that are certified to be correct under poisoning attacks: By definition, certified fraction is smaller than certified accuracy. For instance, when there are only two test samples, the adversarial attack budget (i.e., the number of poisons) may be enough to flip the prediction on either sample, but not enough to flip both predictions simultaneously, resulting in a certified fraction of 0% and a certified accuracy of 50%.

While our method is designed for pointwise robustness (i.e. corresponding to certified fraction), it naturally offers a well-defined but computationally inefficient way to estimate certified accuracy (Appendix D). Taking advantage of this, we directly compare certified accuracy and certified fraction under an attack strength $d_{\text{sym}} = 1$ in Table 3 to estimate their gaps. The differences in Table 3 corroborate the intuition that different poisons may be needed for different targets. Notably, to our best knowledge, this is one of the first *direct* empirical comparisons between pointwise and distributional certified robustness resulting from a *single* method. A concurrent work (Chen et al., 2022) highlights the distributional robustness of aggregation-based defenses.

6. Conclusion

In this work, we propose **Finite Aggregation**, a novel provable defense against general data poisoning extended from DPA to further improve robustness. Compared to DPA, our method effectively boosts certified fractions by up to 3.05%, 3.87% and 4.77% on MNIST, CIFAR-10, and GTSRB, respectively, achieving a new **state of the art** in pointwise certified robustness against general data poisoning. We also provide an alternative view to aggregation-based defenses against poisoning attacks that bridges the gap between the deterministic and the stochastic variants, unifying the designs of aggregation-based defenses to date.

Acknowledgements

This project was supported in part by NSF CAREER AWARD 1942230, a grant from NIST 60NANB20D134, HR001119S0026 (GARD), ONR grant 13370299 and Army Grant No. W911NF2120076.

References

Abadi, M., Chu, A., Goodfellow, I. J., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. Deep learning with differential privacy. In Weippl, E. R., Katzenbeisser, S., Kruegel, C., Myers, A. C., and Halevi, S. (eds.), *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, Vienna, Austria, October 24-28, 2016*, pp. 308–318. ACM, 2016. doi: 10.1145/2976749.2978318. URL https://doi.org/10.1145/2976749.2978318.

Aghakhani, H., Meng, D., Wang, Y., Kruegel, C., and Vigna, G. 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*, pp. 159–178. IEEE, 2021. doi: 10.1109/EuroSP51992.2021.00021. URL https://doi.org/10.1109/EuroSP51992.2021.00021.

Chen, R., Li, J., Wu, C., Sheng, B., and Li, P. A framework of randomized selection based certified defenses against data poisoning attacks. *CoRR*, abs/2009.08739, 2020. URL https://arxiv.org/abs/2009.08739.

Chen, R., Li, Z., Li, J., Wu, C., and Yan, J. On collective robustness of bagging against data poisoning. *CoRR*, abs/2205.13176, 2022. doi: 10.48550/arXiv.2205.13176. URL https://doi.org/10.48550/arXiv.2205.13176.

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

Devlin, J., Chang, M., Lee, K., and Toutanova, K. 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)*, pp. 4171–4186, 2019. URL https://www.aclweb.org/anthology/N19-1423/.

Diakonikolas, I., Kamath, G., Kane, D. M., Li, J. Z., Moitra, A., and Stewart, A. Robust estimators in high dimensions without the computational intractability. *CoRR*, abs/1604.06443, 2016. URL http://arxiv.org/abs/1604.06443.

Diakonikolas, I., Kamath, G., Kane, D., Li, J., Steinhardt, J., and Stewart, A. Sever: A robust meta-algorithm for stochastic optimization. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *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*, pp. 1596–1606. PMLR, 2019. URL http://proceedings.mlr.press/v97/diakonikolas19a.html.

Du, M., Jia, R., and Song, D. Robust anomaly detection and backdoor attack detection via differential privacy. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=SJx0q1rtvS.

Gao, J., Karbasi, A., and Mahmoody, M. Learning and certification under instance-targeted poisoning. In de Campos, C. P., Maathuis, M. H., and Quaeghebeur, E. (eds.), *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*, pp. 2135–2145. AUAI Press, 2021. URL https://proceedings.mlr.press/v161/gao21b.html.

Geiping, J., Fowl, L. H., Huang, W. R., Czaja, W., Taylor, G., Moeller, M., and Goldstein, T. 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.

Gidaris, S., Singh, P., and Komodakis, N. Unsupervised representation learning by predicting image rotations. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=S1v4N2l0-.

Goldblum, M., Tsipras, D., Xie, C., Chen, X., Schwarzschild, A., Song, D., Madry, A., Li, B., and Goldstein, T. Dataset security for machine learning: Data poisoning, backdoor attacks, and defenses. *CoRR*, abs/2012.10544, 2020. URL https://arxiv.org/abs/2012.10544.

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.

Jia, J., Cao, X., and Gong, N. Z. 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, pp. 7961–7969. AAAI Press, 2021. URL https://ojs.aaai.org/index.php/AAAI/article/view/16971.

Krizhevsky, A. Learning multiple layers of features from tiny images. Technical report, 2009.

Lai, K. A., Rao, A. B., and Vempala, S. S. Agnostic estimation of mean and covariance. In Dinur, I. (ed.), *IEEE 57th Annual Symposium on Foundations of Computer Science, FOCS 2016, 9-11 October 2016, Hyatt Regency, New Brunswick, New Jersey, USA*, pp. 665–674. IEEE Computer Society, 2016. doi: 10.1109/FOCS.2016.76. URL https://doi.org/10.1109/FOCS.2016.76.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, November 1998.

Levine, A. and Feizi, S. 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. URL https://openreview.net/forum?id=YUGG2tFuPM.

Lin, M., Chen, Q., and Yan, S. Network in network. In Bengio, Y. and LeCun, Y. (eds.), *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.

Ma, Y., Zhu, X., and Hsu, J. Data poisoning against differentially-private learners: Attacks and defenses. In Kraus, S. (ed.), *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pp. 4732–4738. ijcai.org, 2019. doi: 10.24963/ijcai.2019/657. URL https://doi.org/10.24963/ijcai.2019/657.

Papernot, N., Abadi, M., Erlingsson, Ú., Goodfellow, I. J., and Talwar, K. Semi-supervised knowledge transfer for deep learning from private training data. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=HkwoSDPgg.

Papernot, N., Song, S., Mironov, I., Raghunathan, A., Talwar, K., and Erlingsson, Ú. Scalable private learning with PATE. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=rkZB1XbRZ.

Rosenfeld, E., Winston, E., Ravikumar, P., and Kolter, J. Z. 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*, pp. 8230–8241. PMLR, 2020. URL http://proceedings.mlr.press/v119/rosenfeld20b.html.

Saha, A., Subramanya, A., and Pirsiavash, H. 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*, pp. 11957–11965. AAAI Press, 2020. URL https://aaai.org/ojs/index.php/AAAI/article/view/6871.

Schwarzschild, A., Goldblum, M., Gupta, A., Dickerson, J. P., and Goldstein, T. Just how toxic is data poisoning? A unified benchmark for backdoor and data poisoning attacks. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 9389–9398. PMLR, 2021. URL http://proceedings.mlr.press/v139/schwarzschild21a.html.

Shafahi, A., Huang, W. R., Najibi, M., Suciu, O., Studer, C., Dumitras, T., and Goldstein, T. Poison frogs! targeted clean-label poisoning attacks on neural networks. In Bengio, S., Wallach, H. M., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pp. 6106–6116, 2018.

Stallkamp, J., Schlipsing, M., Salmen, J., and Igel, C. Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition. *Neural Networks*, 32:323–332, 2012. doi: 10.1016/j.neunet.2012.02.016. URL https://doi.org/10.1016/j.neunet.2012.02.016.

Steinhardt, J., Koh, P. W., and Liang, P. Certified defenses for data poisoning attacks. In Guyon, I., von Luxburg, U., Bengio, S., Wallach, H. M., Fergus, R., Vishwanathan, S. V. N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 3517–3529, 2017.

Turner, A., Tsipras, D., and Madry, A. Label-consistent backdoor attacks. *CoRR*, abs/1912.02771, 2019. URL http://arxiv.org/abs/1912.02771.

Wang, B., Cao, X., Jia, J., and Gong, N. Z. On certifying robustness against backdoor attacks via randomized smoothing. *CoRR*, abs/2002.11750, 2020. URL https://arxiv.org/abs/2002.11750.

Wang, W., Wang, T., Wang, L., Luo, N., Zhou, P., Song, D., and Jia, R. Dplis: Boosting utility of differentially private deep learning via randomized smoothing. *Proc. Priv. Enhancing Technol.*, 2021(4):163–183, 2021. doi: 10.2478/popets-2021-0065. URL https://doi.org/10.2478/popets-2021-0065.

Weber, M., Xu, X., Karlas, B., Zhang, C., and Li, B. RAB: provable robustness against backdoor attacks. *CoRR*, abs/2003.08904, 2020. URL https://arxiv.org/abs/2003.08904.

Xiong, W., Droppo, J., Huang, X., Seide, F., Seltzer, M., Stolcke, A., Yu, D., and Zweig, G. Achieving human parity in conversational speech recognition. *CoRR*, abs/1610.05256, 2016. URL http://arxiv.org/abs/1610.05256.

Zhu, C., Huang, W. R., Li, H., Taylor, G., Studer, C., and Goldstein, T. Transferable clean-label poisoning attacks on deep neural nets. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *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*, pp. 7614–7623. PMLR, 2019. URL http://proceedings.mlr.press/v97/zhu19a.html.

A. Detailed Explanation of Figure 2

Figure 2 is a toy example illustrating how our proposed Finite Aggregation improves provable robustness through a strategic reusing of every sample. The correct prediction is 'cat' in this example. When there is no poison, both DPA and our method predict correctly with the same distribution of predictions from base classifiers (i.e. 50% cat, 16.7% dog, 16.7% deer, and 16.7% bird). In DPA, with one poison contributes to a base classifier that originally predicts correctly (i.e. 'cat'), it may reduce the margin between 'cat' and 'dog' by $2/6 = 33.3\%$ and create a tie, as shown in Figure 2.

However, with Finite Aggregation, even the most effective poison cannot be as effective. In the example, every partition contributes to two consecutive base classifiers and the predictions of them happen to be different. This means that one poison will never remove two correct predictions (i.e. 'cat') from base classifiers and therefore can at most remove 1 'cat' and add 2 'dog' (or other classes) to reduce the margin by $3/12 = 25\%$, which is still too small to affect our final prediction.

B. Proof of Theorem 2

Proof: Given a training set D and an input x, for an arbitrary training set D' such that

$$\frac{1}{k} \cdot \Delta_{D,x}^{\overline{d_{\text{sym}}(D,D')}} \leq \text{FA}(D,x)_c - \text{FA}(D,x)_{c'} - \frac{\mathbb{1}\left[c' < c\right]}{kd}$$

holds for all $c' \neq c$, we want to show that $\text{FA}(D',x) = \text{FA}(D,x) = c$.

Let $D \ominus D' = (D \setminus D') \cup (D' \setminus D)$ to denote symmetric difference between D and D', which corresponds to the minimum set of samples to be inserted/removed if one wants change from D to D' (or D' to D).

Since the base classification algorithm f_{base} is deterministic, by definition, the prediction from a base classifier will not change if its corresponding training set S_i (S_i') remains unchanged from D to D', which is equivalent to $i \notin \bigcup_{x \in D \ominus D'} h_{\text{spread}}(h_{\text{split(x)}})$. Let

$$h_{D \ominus D'} = \bigcup_{x \in D \ominus D'} h_{\text{spread}}(h_{\text{split(x)}}).$$

Thus for any $c' \neq c$, we have

$$\begin{aligned}
&\text{FA}(D',x)_c - \text{FA}(D',x)_{c'} \\
&= \frac{1}{kd} \sum_{i \in [kd]} \Big(\mathbb{1}\left[f_{\text{base}}(S_i',x) = c\right] - \mathbb{1}\left[f_{\text{base}}(S_i',x) = c'\right] \Big) \\
&= \text{FA}(D,x)_c - \text{FA}(D,x)_{c'} \\
&+ \frac{1}{kd} \sum_{i \in h_{D \ominus D'}} \Big(\mathbb{1}\left[f_{\text{base}}(S_i',x) = c\right] - \mathbb{1}\left[f_{\text{base}}(S_i',x) = c'\right] \Big) \\
&- \frac{1}{kd} \sum_{i \in h_{D \ominus D'}} \Big(\mathbb{1}\left[f_{\text{base}}(S_i,x) = c\right] - \mathbb{1}\left[f_{\text{base}}(S_i,x) = c'\right] \Big) \\
&\geq \text{FA}(D,x)_c - \text{FA}(D,x)_{c'} \\
&+ \frac{1}{kd} \sum_{i \in h_{D \ominus D'}} \Big(0 - 1 \Big) \\
&- \frac{1}{kd} \sum_{i \in h_{D \ominus D'}} \Big(\mathbb{1}\left[f_{\text{base}}(S_i,x) = c\right] - \mathbb{1}\left[f_{\text{base}}(S_i,x) = c'\right] \Big) \\
&= \text{FA}(D,x)_c - \text{FA}(D,x)_{c'} - \\
&\frac{1}{kd} \sum_{i \in h_{D \ominus D'}} \Big(1 + \mathbb{1}\left[f_{\text{base}}(S_i,x) = c\right] - \mathbb{1}\left[f_{\text{base}}(S_i,x) = c'\right] \Big)
\end{aligned}$$

We can rewrite $h_{D\ominus D'}$ as follows:

$$h_{D\ominus D'} = \bigcup_{x\in D\ominus D'} h_{\text{spread}}(h_{\text{split(x)}}) = \bigcup_{j\in h_{\text{split}}(D\ominus D')} h_{\text{spread}}(j),$$

where $h_{\text{split}}(D\ominus D') = \{h_{\text{split}}(x)|x\in D\ominus D'\}$.

Since $1+\mathbb{1}\left[f_{\text{base}}(S_i,x)=c\right]-\mathbb{1}\left[f_{\text{base}}(S_i,x)=c'\right]\geq 0$, we can further bound the above formula as follows:

$$\begin{aligned}
&\text{FA}(D',x)_c - \text{FA}(D',x)_{c'}\\
&\geq \text{FA}(D,x)_c - \text{FA}(D,x)_{c'}\\
&- \frac{1}{kd}\sum_{j\in h_{\text{split}}(D\ominus D')}\Big(\sum_{i\in h_{\text{spread}}(j)} 1+\mathbb{1}\left[f_{\text{base}}(S_i,x)=c\right]\\
&- \mathbb{1}\left[f_{\text{base}}(S_i,x)=c'\right]\Big)\\
&= \text{FA}(D,x)_c - \text{FA}(D,x)_{c'}\\
&- \frac{1}{k}\sum_{j\in h_{\text{split}}(D\ominus D')}\Big(1+\text{FA}(D,x)_{c|j} - \text{FA}(D,x)_{c'|j}\Big)
\end{aligned}$$

Since $d_{\text{sym}}(D,D') = |D\ominus D'| \geq |h_{\text{split}}(D\ominus D')|$, we have

$$\begin{aligned}
&\text{FA}(D',x)_c - \text{FA}(D',x)_{c'}\\
&\geq \text{FA}(D,x)_c - \text{FA}(D,x)_{c'}\\
&- \frac{1}{k}\max_{\substack{H\subseteq[kd]\\ |H|\leq d_{\text{sym}}(D,D')}}\sum_{j\in H}\Big(1+\text{FA}(D,x)_{c|j} - \text{FA}(D,x)_{c'|j}\Big)\\
&= \text{FA}(D,x)_c - \text{FA}(D,x)_{c'} - \frac{1}{k}\cdot\overline{\Delta_{D,x}^{d_{\text{sym}}(D,D')}}
\end{aligned}$$

Reorganizing this and subtract $\frac{\mathbb{1}[c'<c]}{kd}$ from both side, we have

$$\begin{aligned}
&\text{FA}(D',x)_c - \text{FA}(D',x)_{c'} - \frac{\mathbb{1}\left[c'<c\right]}{kd}\\
&\geq \text{FA}(D,x)_c - \text{FA}(D,x)_{c'} - \frac{1}{k}\cdot\overline{\Delta_{D,x}^{d_{\text{sym}}(D,D')}} - \frac{\mathbb{1}\left[c'<c\right]}{kd}\\
&\geq 0,
\end{aligned}$$

where the last inequality is from the condition that

$$\frac{1}{k}\cdot\overline{\Delta_{D,x}^{d_{\text{sym}}(D,D')}} \leq \text{FA}(D,x)_c - \text{FA}(D,x)_{c'} - \frac{\mathbb{1}\left[c'<c\right]}{kd}$$

holds for all $c'\neq c$.

Since $\text{FA}(D',x)_c - \text{FA}(D',x)_{c'} - \frac{\mathbb{1}[c'<c]}{kd}\geq 0$, we know $\text{FA}(D',x)\neq c'$. Since this holds for any $c'\neq c$, we have $\text{FA}(D',x)=c=\text{FA}(D,x)$, which completes the proof.

□

C. Proof of Theorem 3

Proof: The ideas for this proof are very similar to the ones in proving Theorem 2.

Given a training set D and an input x, for an arbitrary training set D' such that

$$\frac{1}{k}\cdot\overline{\Delta_{D,x}^{d_{\text{sym}}(D,D')}} \leq \text{IA}(D,x)_c - \text{IA}(D,x)_{c'} - \mathbb{1}\left[c'<c\right]\cdot\delta$$

holds for all $c'\neq c$ and for some $\delta>0$, we want to show that $\text{IA}(D',x)=\text{IA}(D,x)=c$.

Let $D\ominus D' = (D\setminus D')\cup(D'\setminus D)$ to denote symmetric difference between D and D', which corresponds to the minimum set of samples to be inserted/removed if one wants change from D to D' (or D' to D).

We can reorganize Infinite Aggregation as follows:

$$\begin{aligned}
\text{IA}(D,X) &= \arg\max_{c\in\mathcal{C}} Pr_{S\sim\text{Bernoulli}(D,\frac{1}{k})}\left[f_{\text{base}}(S,x)=c\right]\\
&= \arg\max_{c\in\mathcal{C}} \mathbb{E}_{S\sim\text{Bernoulli}(D,\frac{1}{k})}\mathbb{1}\left[f_{\text{base}}(S,x)=c\right]\\
&= \arg\max_{c\in\mathcal{C}} \mathbb{E}_{S\sim\text{Bernoulli}(D,\frac{1}{k})} g(S,x)_c,
\end{aligned}$$

where $g(S,x)_c = Pr[f_{\text{base}}(S,x)=c]$ is the distribution of the predictions and is therefore deterministic.

Since $g(S,x)$ is deterministic, by definition, the contribution from a base classifier will not change if its corresponding training set S remains unchanged from D to D', which is equivalent to $S\subseteq D\cap D'$.

Thus for any $c'\neq c$, we have

$$\begin{aligned}
&\text{IA}(D',x)_c - \text{IA}(D',x)_{c'}\\
&= \mathbb{E}_{S\sim Bernoulli(D',\frac{1}{k})}\Big(g(S,x)_c - g(S,x)_{c'}\Big)\\
&= \mathbb{E}_{S\sim Bernoulli(D\cup D',\frac{1}{k})}\Big(g(S\cap D',x)_c - g(S\cap D',x)_{c'}\Big)\\
&= \text{IA}(D,x)_c - \text{IA}(D,x)_{c'}\\
&+ \mathbb{E}_{S\sim Bernoulli(D\cup D',\frac{1}{k})}\mathbb{1}\left[S\not\subseteq D\cap D'\right]\times\Big(g(S\cap D',x)_c\\
&- g(S\cap D',x)_{c'} - g(S\cap D,x)_c + g(S\cap D,x)_{c'}\Big)\\
&\geq \text{IA}(D,x)_c - \text{IA}(D,x)_{c'} - \mathbb{E}_{S\sim Bernoulli(D\cup D',\frac{1}{k})}\Big[\\
&\mathbb{1}\left[S\not\subseteq D\cap D'\right]\times\Big(1+g(S\cap D,x)_c - g(S\cap D,x)_{c'}\Big)\Big]
\end{aligned}$$

Since $\mathbb{1}\left[S\not\subseteq D\cap D'\right]\leq\sum_{x_L\in D\ominus D'}\mathbb{1}\left[x_L\in S\right]$, we can further bound the above formula as follows:

$$\begin{aligned}
&\text{IA}(D',x)_c - \text{IA}(D',x)_{c'}\\
&\geq \text{IA}(D,x)_c - \text{IA}(D,x)_{c'} - \mathbb{E}_{S\sim Bernoulli(D\cup D',\frac{1}{k})}\Big[
\end{aligned}$$

$$\sum_{x_L \in D \ominus D'} \mathbb{1}\left[x_L \in S\right] \times \Big(1 + g(S \cap D, x)_c - g(S \cap D, x)_{c'}\Big)\Big]$$

$$= \mathrm{IA}(D,x)_c - \mathrm{IA}(D,x)_{c'}$$

$$- \sum_{x_L \in D \ominus D'} \mathbb{E}_{S \sim Bernoulli(D \cup D', \frac{1}{k})}\Big[$$

$$\mathbb{1}\left[x_L \in S\right] \times \Big(1 + g(S \cap D, x)_c - g(S \cap D, x)_{c'}\Big)\Big]$$

$$= \mathrm{IA}(D,x)_c - \mathrm{IA}(D,x)_{c'}$$

$$- \sum_{x_L \in D \setminus D'} \frac{1}{k}\Big(1 + \mathrm{IA}(D,x)_{c|x_L} - \mathrm{IA}(D,x)_{c'|x_L}\Big)$$

$$- \sum_{x_L \in D' \setminus D} \frac{1}{k}\Big(1 + \mathrm{IA}(D,x)_c - \mathrm{IA}(D,x)_{c'}\Big)$$

$$\geq \mathrm{IA}(D,x)_c - \mathrm{IA}(D,x)_{c'} - \frac{1}{k}\overline{\Delta}_{D,x}^{\overline{d_{\mathrm{sym}}(D,D')}}$$

Subtracting both sides with $\mathbb{1}\left[c' < c\right] \cdot \delta$, we have

$$\mathrm{IA}(D',x)_c - \mathrm{IA}(D',x)_{c'} - \mathbb{1}\left[c' < c\right] \cdot \delta$$

$$\geq \mathrm{IA}(D,x)_c - \mathrm{IA}(D,x)_{c'} - \mathbb{1}\left[c' < c\right] \cdot \delta - \frac{1}{k}\overline{\Delta}_{D,x}^{\overline{d_{\mathrm{sym}}(D,D')}}$$

$$\geq 0$$

for some $\delta > 0$, where the last inequality is from the condition that

$$\frac{1}{k} \cdot \overline{\Delta}_{D,x}^{\overline{d_{\mathrm{sym}}(D,D')}} \leq \mathrm{IA}(D,x)_c - \mathrm{IA}(D,x)_{c'} - \mathbb{1}\left[c' < c\right] \cdot \delta$$

holds for all $c' \neq c$ and for some $\delta > 0$.

This means that $\mathrm{IA}(D',x) \neq c'$. Since this is true for any $c' \neq c$, we have $\mathrm{IA}(D',x) = c = \mathrm{IA}(D,x)$, which completes the proof.

□

D. Certified Accuracy from Finite Aggregation

In this section, we show how certified accuracy can be computed for Finite Aggregation (i.e. how to derive a lower bound of accuracy on a given test set under poisoning attacks). The key in this derivation is to realize that using *same* poisons for multiple targets meaning a more restricted subset of partitions affected compared to different poisons.

We introduce a variant of Theorem 2 as follows, certifying a prediction under conditional poisoning attacks, where only a given subset of partitions (i.e. $Q \subseteq [kd]$) may be affected by poisons.

Theorem 4 (Certificates under Conditional Poisoning) *Given a training set D, a subset of partition indices $Q \subseteq [kd]$, and an input x, let $c = FA(D,x)$, then for any training set D' such that $h_{split}\left((D \setminus D') \cup (D' \setminus D)\right) \subseteq Q$, it is guaranteed that $FA(D',x) = FA(D,x)$ when*

$$\frac{1}{k} \cdot \Delta_{D,Q,x}^{\overline{d_{sym}(D,D')}} \leq FA(D,x)_c - FA(D,x)_{c'} - \frac{\mathbb{1}\left[c' < c\right]}{kd}$$

holds for all $c' \neq c$, where $\Delta_{D,Q,x}$ is a multiset defined as

$$\left\{1 + FA(D,x)_{c|j} - FA(D,x)_{c'|j}\right\}_{j \in Q}$$

and $\Delta_{D,Q,x}^{\overline{d_{sym}(D,D')}}$ denotes the sum of the largest $d_{sym}(D,D')$ elements in the multiset $\Delta_{D,Q,x}$.

The proof is in Appendix E. Theorem 4 provides a sufficient condition for when the prediction on this sample from Finite Aggregation is certifiably correct under poisoning attacks. Given a training set D, a subset of partitions $Q \subseteq [kd]$, an attack budget d_{sym} (i.e. number of poisons allowed), for any test sample with input x and ground truth label c, we define $\mathcal{O}_{x,c,Q,d_{\mathrm{sym}}}$ to be 1 when the prediction is certifiably correct under poisoning attacks and 0 otherwise.

Consequently, the certified accuracy on the test set D_{test} with an attack budget d_{sym} can be expressed as follows by definition:

$$\min_{\substack{Q \subseteq [kd] \\ |Q| \leq d_{\mathrm{sym}}}} \frac{1}{|D_{\mathrm{test}}|} \sum_{(x,c) \in D_{\mathrm{test}}} \mathcal{O}_{x,c,Q,d_{\mathrm{sym}}}.$$

E. Proof of Theorem 4

Proof: Given a training set D, a subset of partition indices $Q \subseteq [kd]$, and an input x, for an arbitrary training set D' such that $h_{\mathrm{split}}\left((D \setminus D') \cup (D' \setminus D)\right) \subseteq Q$ and

$$\frac{1}{k} \cdot \Delta_{D,Q,x}^{\overline{d_{\mathrm{sym}}(D,D')}} \leq \mathrm{FA}(D,x)_c - \mathrm{FA}(D,x)_{c'} - \frac{\mathbb{1}\left[c' < c\right]}{kd}$$

holds for all $c' \neq c$, we want to show that $\mathrm{FA}(D',x) = \mathrm{FA}(D,x) = c$.

Let $D \ominus D' = (D \setminus D') \cup (D' \setminus D)$ to denote symmetric difference between D and D', which corresponds to the minimum set of samples to be inserted/removed if one wants change from D to D' (or D' to D).

Since the base classification algorithm f_{base} is deterministic, by definition, the prediction from a base classifier will not change if its corresponding training set S_i (S_i') remains unchanged from D to D', which is equivalent to $i \notin \bigcup_{x \in D \ominus D'} h_{\mathrm{spread}}(h_{\mathrm{split(x)}})$. Let

$$h_{D \ominus D'} = \bigcup_{x \in D \ominus D'} h_{\mathrm{spread}}(h_{\mathrm{split(x)}}).$$

Thus for any $c' \neq c$, we have

$$\mathrm{FA}(D',x)_c - \mathrm{FA}(D',x)_{c'}$$

$$
\begin{aligned}
&= \frac{1}{kd} \sum_{i \in [kd]} \Big(\mathbb{1}\left[f_{\text{base}}(S'_i, x) = c\right] - \mathbb{1}\left[f_{\text{base}}(S'_i, x) = c'\right] \Big) \\
&= \text{FA}(D, x)_c - \text{FA}(D, x)_{c'} \\
&+ \frac{1}{kd} \sum_{i \in h_{D \ominus D'}} \Big(\mathbb{1}\left[f_{\text{base}}(S'_i, x) = c\right] - \mathbb{1}\left[f_{\text{base}}(S'_i, x) = c'\right] \Big) \\
&- \frac{1}{kd} \sum_{i \in h_{D \ominus D'}} \Big(\mathbb{1}\left[f_{\text{base}}(S_i, x) = c\right] - \mathbb{1}\left[f_{\text{base}}(S_i, x) = c'\right] \Big) \\
&\geq \text{FA}(D, x)_c - \text{FA}(D, x)_{c'} \\
&+ \frac{1}{kd} \sum_{i \in h_{D \ominus D'}} \Big(0 - 1 \Big) \\
&- \frac{1}{kd} \sum_{i \in h_{D \ominus D'}} \Big(\mathbb{1}\left[f_{\text{base}}(S_i, x) = c\right] - \mathbb{1}\left[f_{\text{base}}(S_i, x) = c'\right] \Big) \\
&= \text{FA}(D, x)_c - \text{FA}(D, x)_{c'} - \\
&\frac{1}{kd} \sum_{i \in h_{D \ominus D'}} \Big(1 + \mathbb{1}\left[f_{\text{base}}(S_i, x) = c\right] - \mathbb{1}\left[f_{\text{base}}(S_i, x) = c'\right] \Big)
\end{aligned}
$$

We can rewrite $h_{D \ominus D'}$ as follows:

$$
h_{D \ominus D'} = \bigcup_{x \in D \ominus D'} h_{\text{spread}}(h_{\text{split(x)}}) = \bigcup_{j \in h_{\text{split}}(D \ominus D')} h_{\text{spread}}(j),
$$

where $h_{\text{split}}(D \ominus D') = \{h_{\text{split}}(x) | x \in D \ominus D'\}$.

Since $1 + \mathbb{1}\left[f_{\text{base}}(S_i, x) = c\right] - \mathbb{1}\left[f_{\text{base}}(S_i, x) = c'\right] \geq 0$, we can further bound the above formula as follows:

$$
\begin{aligned}
&\text{FA}(D', x)_c - \text{FA}(D', x)_{c'} \\
&\geq \text{FA}(D, x)_c - \text{FA}(D, x)_{c'} \\
&- \frac{1}{kd} \sum_{j \in h_{\text{split}}(D \ominus D')} \Big(\sum_{i \in h_{\text{spread}}(j)} 1 + \mathbb{1}\left[f_{\text{base}}(S_i, x) = c\right] \\
&- \mathbb{1}\left[f_{\text{base}}(S_i, x) = c'\right] \Big) \\
&= \text{FA}(D, x)_c - \text{FA}(D, x)_{c'} \\
&- \frac{1}{k} \sum_{j \in h_{\text{split}}(D \ominus D')} \Big(1 + \text{FA}(D, x)_{c|j} - \text{FA}(D, x)_{c'|j} \Big)
\end{aligned}
$$

Since $d_{\text{sym}}(D, D') = |D \ominus D'| \geq |h_{\text{split}}(D \ominus D')|$ and $h_{\text{split}}(D \ominus D') \subseteq Q$, we have

$$
\begin{aligned}
&\text{FA}(D', x)_c - \text{FA}(D', x)_{c'} \\
&\geq \text{FA}(D, x)_c - \text{FA}(D, x)_{c'} \\
&- \frac{1}{k} \max_{\substack{H \subseteq Q \\ |H| \leq d_{\text{sym}}(D, D')}} \sum_{j \in H} \Big(1 + \text{FA}(D, x)_{c|j} - \text{FA}(D, x)_{c'|j} \Big) \\
&= \text{FA}(D, x)_c - \text{FA}(D, x)_{c'} - \frac{1}{k} \cdot \Delta_{D,Q,x}^{\overline{d_{\text{sym}}(D,D')}}
\end{aligned}
$$

Reorganizing this and subtract $\frac{\mathbb{1}[c' < c]}{kd}$ from both side, we have

$$
\begin{aligned}
&\text{FA}(D', x)_c - \text{FA}(D', x)_{c'} - \frac{\mathbb{1}\left[c' < c\right]}{kd} \\
&\geq \text{FA}(D, x)_c - \text{FA}(D, x)_{c'} - \frac{1}{k} \cdot \Delta_{D,Q,x}^{\overline{d_{\text{sym}}(D,D')}} - \frac{\mathbb{1}\left[c' < c\right]}{kd} \\
&\geq 0,
\end{aligned}
$$

where the last inequality is from the condition that

$$
\frac{1}{k} \cdot \Delta_{D,Q,x}^{\overline{d_{\text{sym}}(D,D')}} \leq \text{FA}(D, x)_c - \text{FA}(D, x)_{c'} - \frac{\mathbb{1}\left[c' < c\right]}{kd}
$$

holds for all $c' \neq c$.

Since $\text{FA}(D', x)_c - \text{FA}(D', x)_{c'} - \frac{\mathbb{1}[c' < c]}{kd} \geq 0$, we know $\text{FA}(D', x) \neq c'$. Since this holds for any $c' \neq c$, we have $\text{FA}(D', x) = c = \text{FA}(D, x)$, which completes the proof.

□