PROVABLY RELIABLE CONFORMAL PREDICTION SETS IN THE PRESENCE OF DATA POISONING

Yan Scholten, Stephan Günnemann
Department of Computer Science & Munich Data Science Institute
Technical University of Munich
{y.scholten, s.guennemann}@tum.de

ABSTRACT

Conformal prediction provides model-agnostic and distribution-free uncertainty quantification through prediction sets that are guaranteed to include the ground truth with any user-specified probability. Yet, conformal prediction is not reliable under poisoning attacks where adversaries manipulate both training and calibration data, which can significantly alter prediction sets in practice. As a solution, we propose *reliable prediction sets* (RPS): the first efficient method for constructing conformal prediction sets with provable reliability guarantees under poisoning. To ensure reliability under training poisoning, we introduce smoothed score functions that reliably aggregate predictions of classifiers trained on distinct partitions of the training data. To ensure reliability under calibration poisoning, we construct multiple prediction sets, each calibrated on distinct subsets of the calibration data. We then aggregate them into a majority prediction set, which includes a class only if it appears in a majority of the individual sets. Both proposed aggregations mitigate the influence of datapoints in the training and calibration data on the final prediction set. We experimentally validate our approach on image classification tasks, achieving strong reliability while maintaining utility and preserving coverage on clean data. Overall, our approach represents an important step towards more trustworthy uncertainty quantification in the presence of data poisoning.

1 INTRODUCTION

Conformal prediction has emerged as a powerful framework for model-agnostic and distribution-free uncertainty quantification in machine learning. By constructing prediction sets calibrated on hold-out calibration data, it can transform any existing black-box classifier into a predictor with formal coverage guarantees, ensuring that its prediction sets cover the ground truth with any user-specified probability (Angelopoulos & Bates, 2021). This makes conformal prediction highly relevant for uncertainty quantification in safety-critical applications such as medical diagnosis (Vazquez & Facelli, 2022), autonomous driving (Lindemann et al., 2023), and flood forecasting (Auer et al., 2023).

However in practice, noise, incomplete data or adversarial perturbations can lead to unreliable prediction sets (Liu et al., 2024). In particular data poisoning – where adversaries modify the training or calibration data (e.g. during data labeling) – can significantly alter the prediction sets, resulting in overly conservative or empty sets (Li et al., 2024). This vulnerability can undermine the practical utility of conformal prediction in safety-critical applications, raising the research question:

How can we make conformal prediction sets provably reliable in the presence of data poisoning?

As a solution, we propose *reliable prediction sets* (RPS): the first efficient method for constructing prediction sets more reliable under data poisoning. Our approach consists of two key components (Figure 1): First (**i**), we introduce smoothed score functions that reliably aggregate predictions from classifiers trained on distinct partitions of the training data, improving reliability under training poisoning. Second (**ii**), we calibrate multiple prediction sets on disjoint subsets of the calibration data and construct a majority prediction set that includes classes only when a majority of the independent prediction sets agree, improving reliability under calibration poisoning. Using both strategies (**i**) and (**ii**), RPS effectively reduces the influence of datapoints during training and calibration.

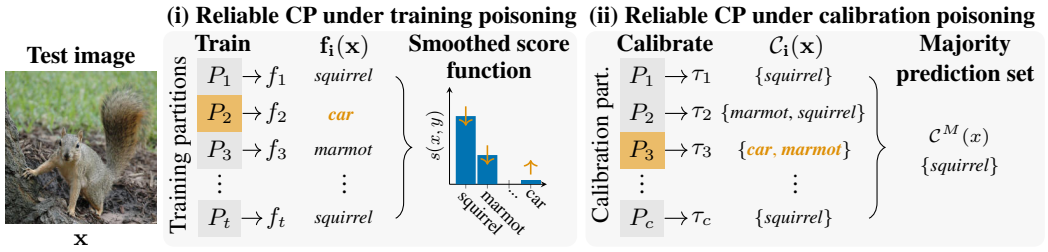

Figure 1: Conformal prediction (CP) is not reliable under poisoning (orange) of training and calibration data, undermining its practical utility in safety-critical applications. As a solution, we propose *reliable prediction sets* (RPS): A novel approach for constructing more reliable prediction sets. We (**i**) aggregate predictions of classifiers trained on distinct partitions, and (**ii**) merge multiple prediction sets $\mathcal{C}_i(x)=\{y : s(x,y) \geq \tau_i\}$ calibrated on separate partitions into a majority prediction set that includes classes only if a majority of the prediction sets $\mathcal{C}_i$ agree. This way RPS reduces the influence of datapoints while preserving the coverage guarantee of conformal prediction on clean data.

We further derive certificates, i.e. provable guarantees for the reliability of RPS under worst-case poisoning. We experimentally validate the effectiveness of our approach on image classification tasks, demonstrating strong reliability under worst-case poisoning while maintaining utility and empirically preserving the coverage guarantee of prediction sets on clean data. Our main contributions are:

- We propose *reliable prediction sets* (RPS) – the first scalable and efficient method for making conformal prediction more reliable under training and calibration poisoning.
- We derive novel certificates that guarantee the reliability of RPS under worst-case data poisoning attacks, including guarantees against label flipping attacks.
- We exhaustively evaluate our method and verify our theorems on image classification tasks.

2 Related Work

Prediction set ensembles. Ensembles of prediction sets are studied in the uncertainty set literature (Cherubin, 2019; Solari & Djordjilović, 2022; Gasparin & Ramdas, 2024) beyond machine learning, e.g. to reduce the effect of randomness. Our work instead proposes a method to improve the reliability of conformal prediction under worst-case training and calibration poisoning.

Conformal prediction under evasion. Most works regarding reliable conformal prediction focus on evasion threat models, i.e. adversarial perturbations of the test data. They typically build upon randomized smoothing (Cohen et al., 2019) to certify robustness against evasion attacks (Gendler et al., 2022; Yan et al., 2024; Zargarbashi et al., 2024), or use neural network-specific verification (Jeary et al., 2024). Ghosh et al. (2023) introduce a probabilistic notion as an alternative to worst-case evasion attacks. Unlike prior work on evasion, we consider poisoning threat models.

Conformal prediction under poisoning. Although there are already poisoning attacks against conformal prediction (Li et al., 2024), the few attempts to improve its reliability focus on other notions of reliability: Most works only consider calibration poisoning where adversaries aim to reduce coverage by removing classes from prediction sets (Zargarbashi et al., 2024; Park et al., 2023; Kang et al., 2024). Others study calibration poisoning empirically (Einbinder et al., 2022), under specific label noise (Penso & Goldberger, 2024), or consider distribution shifts between calibration and test data (Cauchois et al., 2020). Overall, none of the existing approaches considers our threat model where adversaries can poison both training and calibration data with the goal to either add or remove classes from prediction sets.

Robustness certification against data poisoning. Most certification techniques for robust classification under poisoning consider other threat models, specific training techniques or architectures (Rosenfeld et al., 2020; Tian et al., 2023; Sosnin et al., 2024). The strongest guarantees also partition the training data and aggregate predictions of classifiers trained on each partition (Levine & Feizi, 2021; Wang et al., 2022; Rezaei et al., 2023). However, all of the prior works only guarantee robust classification and are not directly applicable to certify conformal prediction since prediction sets (1) contain multiple classes, and (2) can be manipulated via poisoning during training and calibration.

3 BACKGROUND AND PRELIMINARIES

We focus on classification tasks defined on an input space $\mathcal{X} = \mathbb{R}^d$ for a given finite set of classes $\mathcal{Y} = \{1, \dots, K\}$. We model prediction set predictors as functions $\mathcal{C} : \mathcal{X} \to Y \subseteq \mathcal{Y}$, which provide prediction sets as subsets $\mathcal{C}(x) \subseteq \mathcal{Y}$ of the classes $\mathcal{Y}$ for any given datapoint $x \in \mathcal{X}$.

Exchangeability. Conformal prediction is a model-agnostic and distribution-free method for constructing prediction sets. It only requires that the datapoints are exchangegable, which means that their joint distribution is invariant under permutations of the datapoints. This is more general than an i.i.d.-assumption, since it allows for dependencies between datapoints as long as their distribution remains unchanged when the order is shuffled. In this paper, we adopt the standard assumption (e.g. in image classification) that datapoints are independent and identically distributed (i.i.d.), which implies exchangeability. Specifically, we assume three datasets sampled i.i.d. from the same distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$: training set $\mathcal{D}_{train}$, calibration set $\mathcal{D}_{calib}{=}\{(x_i, y_i)\}_{i=1}^n$ and test set $\mathcal{D}_{test}$.

Conformal prediction. Conformal prediction is a method for turning a given black-box classifier $f : \mathcal{X} \to \mathcal{Y}$ into a prediction set predictor. We study split conformal prediction (Papadopoulos et al., 2002; Lei et al., 2018), the most widely-used variation of conformal prediction in machine learning. First we train a classifier $f(x)$ on the training set and use a score function $s(x, y)$ that measures conformity between samples x and classes y using classifier f. For example, homogeneous prediction sets (HPS) use class probabilities $s(x, y){=}f_y(x)$ (Sadinle et al., 2019). Using a score function s we then compute conformal scores $S{=}\{s(x_i, y_i)\}_{i=1}^n$ for samples of the calibration set $\mathcal{D}_{calib}$. Finally, one can construct prediction sets with the following coverage guarantee (Vovk et al., 1999; 2005):

Theorem 1. *Given user-specified coverage probability $1 - \alpha \in (0, 1)$, test sample $(x_{n+1}, y_{n+1}) \in \mathcal{D}_{test}$ exchangeable with $\mathcal{D}_{calib}$, and a score function s, we can construct the following prediction set*

$$\mathcal{C}(x_{n+1}) = \{y \in \mathcal{Y} : s(x_{n+1}, y) \geq \tau\}$$

which fulfills the following marginal coverage guarantee

$$\Pr[y_{n+1} \in \mathcal{C}(x_{n+1})] \geq 1 - \alpha$$

for $\tau = Quant(\alpha_n; S)$, which is the α_n-quantile of the conformal scores S for a finite-sample corrected coverage level $\alpha_n = \frac{\lceil (n+1)(1-\alpha) \rceil}{n}$.

For a more detailed introduction to conformal prediction we refer to Angelopoulos & Bates (2021).

4 DESIDERATA FOR RELIABLE CONFORMAL PREDICTION

First we want to outline the desired properties reliable conformal prediction should exhibit, setting clear goals for how uncertainty should be captured by prediction sets under data poisoning.

Data poisoning. While exchangeability may hold theoretically for the data distribution $\mathcal{D}$, the labeled data $\mathcal{D}_l = (\mathcal{D}_{train}, \mathcal{D}_{calib})$ can be poisoned in practice where noise, incomplete data or adversaries are present (for example during data labeling), violating exchangeability. We formally model this threat model, i.e. the strength of poisoning attacks, as a ball centered around labeled data:

$$B_{r_t, r_c}(\mathcal{D}_l) = \left\{ \tilde{\mathcal{D}}_l \mid \delta(\tilde{\mathcal{D}}_{train}, \mathcal{D}_{train}) \leq r_t, \delta(\tilde{\mathcal{D}}_{calib}, \mathcal{D}_{calib}) \leq r_c \right\} \tag{1}$$

where δ is a distance metric between datasets, and r_t, r_c are the radii for training and calibration sets, respectively. Specifically, we define δ as the number of inserted or deleted datapoints and label flips, modelling feature modifications as two perturbations (deletion and insertion): $\delta(\mathcal{D}_1, \mathcal{D}_2) = |\mathcal{D}_1 \ominus \mathcal{D}_2| - |\mathcal{F}(\mathcal{D}_1, \mathcal{D}_2)|$ where $A \ominus B = (A \setminus B) \cup (B \setminus A)$ is the symmetric set difference between two sets A and B, $|S|$ denotes the cardinality of a set S, and $\mathcal{F}(\mathcal{D}_1, \mathcal{D}_2)$ represents the set of datapoints with label flips $\mathcal{F}(\mathcal{D}_1, \mathcal{D}_2){=}\{x \mid \exists\, y_1 : (x, y_1){\in}\mathcal{D}_1 \setminus \mathcal{D}_2, \exists\, y_2 : (x, y_2){\in}\mathcal{D}_2 \setminus \mathcal{D}_1\}$. Note that we count label flips only once, and feature perturbations can be of arbitrary magnitude.

Reliability under data poisoning. Given a datapoint $x \in \mathcal{X}$ and a prediction set $\mathcal{C}(x) \subseteq \mathcal{Y}$, we define reliability of conformal prediction sets under data poisoning as follows:

Definition 1 (Reliability)**.** *We assume that reliability is compromised if adversaries can remove or add a single class from or to the prediction set $\mathcal{C}(x)$ under our threat model (Equation 1). Specifically, we call prediction sets **coverage reliable** if adversaries cannot compromise coverage by removing classes from the prediction set $\mathcal{C}(x)$, and **size reliable** if adversaries cannot inflate prediction set $\mathcal{C}(x)$ by adding classes. We further denote coverage and size reliable prediction sets as **robust**.*

Accordingly, we propose the following novel desiderata for reliable conformal prediction:

Desiderata for reliable conformal prediction sets under data poisoning

- **I:** Reliable prediction sets must provide $1 - \alpha$ marginal coverage for clean data.
- **II:** Reliable prediction sets must be small for clean data.
- **III:** Reliable conformal prediction must *provably* ensure reliability of prediction sets (Definition 1) under feature poisoning and label flipping (Equation 1), up to a radius that meets the application's needs and safety requirements.
- **IV:** Reliable prediction sets must be more stable for larger training and calibration sets.
- **V:** Reliable prediction sets must be computationally efficient, scalable, and reproducible.

While Desideratum **I** requires that prediction sets fulfil the coverage guarantee, Desideratum **II** ensures small prediction sets, and together they prevent that reliability can be achieved trivially by predicting empty or full sets. Desideratum **III** ensures that reliability must be certifiable, i.e. a provable guarantee under worst-case poisoning. Specifically, adversaries must not add or remove classes. Desideratum **VI** requires that reliability improves if more data is available since practical risk increases with more data. Finally, Desideratum **V** ensures efficiency for practical deployment, where reproducibility requires that the sets do not differ for the same input, ensuring stability on clean data.

5 Reliable conformal prediction sets

Guided by our five desiderata for reliable conformal prediction we introduce **reliable conformal prediction sets** (RPS): the first method that makes conformal prediction provably more reliable under both training and calibration poisoning (Figure 1). The first component of RPS (**i**) reliably aggregates classifiers trained on k_t disjoint partitions of the training data. The second component of RPS (**ii**) constructs reliable prediction sets by merging prediction sets calibrated separately on k_c disjoint partitions of the calibration data. Intuitively, while larger k_t increases reliability against training poisoning, larger k_c increases reliability against calibration poisoning. We provide detailed instructions for our reliable conformal prediction sets in Algorithm 1 and Algorithm 2.

5.1 Conformal score functions reliable under training data poisoning

First, our goal is to derive a conformal score function that is reliable under poisoning of training data. This is challenging since the score function also has to quantify agreement between samples and classes, and maintain exchangeability of conformal scores between calibration and test data. To overcome this challenge we propose to (1) partition the training data into k_t disjoint sets, (2) train separate classifiers on each partition, and (3) design a score function that counts the number of classifiers voting for a class y given sample x. Since deleting or inserting one datapoint from or into the training set only affects a single partition and thus a single classifier, this procedure effectively reduces the influence of datapoints on the score function.

Training data partitioning. To prevent that simple reordering of the datasets affects all partitions simultaneously, we have to partition the training data in a way that is invariant to its order. To achieve this we assign datapoints to partitions by using a hash function directly defined on x. For example for images, we hash the sum of their pixel values. This technique that has been previously shown to induce certifiable robustness in the context of image classification (Levine & Feizi, 2021). Given a hash function h we define the i-th partition of the training set as

$$P_i^t = \{(x_j, y_j) \in \mathcal{D}_{train} : h(x_j) \equiv i \,(\text{mod } k_t)\}\,.$$

Then we deterministically train k_t classifiers $f_i : \mathcal{X} \to \mathcal{Y}$ on all partitions $P_1^t, \ldots, P_{k_t}^t$ separately.

Smoothed score function. Now we define our novel score function that measures agreement between a sample x and class y by counting the number of classifiers f_i voting for class y given x:

$$s(x, y) = \frac{e^{\pi_y(x)}}{\sum_{i=1}^{K} e^{\pi_i(x)}} \quad \text{with} \quad \pi_y(x) = \frac{1}{k_t} \sum_{i=1}^{k_t} \mathbb{1}\{f_i(x) = y\} \tag{2}$$

where $\pi_y(x)$ is the percentage of classifiers voting for class y given sample x. Note that we introduce the additional softmax over class distribution $\pi_y(x)$ to fulfill Desideratum **II**, since the softmax prevents overly large prediction sets in practice (see Section 7).

Algorithm 1 Reliable conformal score function

Input: $\mathcal{D}_{train}$, k_t, training algo. T

Split $\mathcal{D}_{train}$ into k_t disjoint partitions P_i^t
$P_i^t = \{(x_j, y_j) \in \mathcal{D}_{train} : h(x_j) \equiv i \,(\text{mod } k_t)\}$
for $i = 1$ **to** k_t **do**
  Train classifier $f_i = T(P_i^t)$ on partition P_i^t
Construct the voting function
$\pi_y(x) = \frac{1}{k_t} \sum_{i=1}^{k_t} \mathbb{1}\{f_i(x) = y\}$
Smooth the voting function
$s(x, y) = e^{\pi_y(x)} / (\sum_{i=1}^{K} e^{\pi_i(x)})$

Output: Reliable conformal score function s

Algorithm 2 Reliable conformal prediction sets

Input: $\mathcal{D}_{calib}$, k_c, s, α_n, x_{n+1}

Split $\mathcal{D}_{calib}$ into k_c disjoint partitions P_i^c
$P_i^c = \{(x_j, y_j) \in \mathcal{D}_{calib} : h(x_j) \equiv i \,(\text{mod } k_c)\}$
for $i = 1$ **to** k_c **do**
  Compute scores $S_i = \{s(x_j, y_j)\}_{(x_j, y_j) \in P_i^c}$
  Compute α_{n_i}-quantile τ_i of scores S_i
  Construct prediction set for quantile τ_i
  $\mathcal{C}_i(x_{n+1}) = \{y : s(x_{n+1}, y) \geq \tau_i\}$
Construct majority vote prediction set
$\mathcal{C}^M(x_{n+1}) = \{y : \sum_{i=1}^{k_c} \mathbb{1}\{y \in \mathcal{C}_i(x_{n+1})\} > \hat{\tau}\}$

Output: Reliable conformal prediction set $\mathcal{C}^M$

For any function to be considered a valid score function for conformal prediction it has to maintain exchangeability of conformal scores between calibration and test data (Angelopoulos et al., 2021).

Lemma 1. *The smoothed score function in Equation 2 is a valid conformal score function.*

Proof. We use one function to score all points independent of other datapoints and which dataset they belong to (and where in the dataset). Thus, given exchangeable data, scores computed by our smoothed score function remain exchangeable. Therefore s of Equation 2 is a valid score function. □

Lemma 1 implies that the marginal coverage guarantee (Theorem 1) holds on clean data when using our smoothed score function for conformal prediction (Desideratum **I**). Intuitively, our score function quantifies uncertainty by the number of votes of multiple classifiers trained on disjoint partitions instead of the logits of classifiers. As long as the classifiers are trained on isolated sets without access to other partitions we can reduce the influence of datapoints on the conformal scores. We summarize instructions for the smoothed score function in Algorithm 1.

5.2 Majority prediction sets reliable under calibration data poisoning

Now we derive prediction sets reliable against calibration poisoning. This is challenging since the prediction sets must also achieve marginal coverage on clean data (Desideratum **I**) without inflating set size (Desideratum **II**). We propose to (1) partition the calibration data into k_c disjoint sets, (2) compute separate prediction sets based on the conformal scores on each partition, and to (3) merge the resulting prediction sets via majority voting. This improves reliability since adversaries have to poison multiple partitions to alter the majority vote. We further show that such majority prediction sets achieve marginal coverage, and do not grow too much in size in practice (Section 7).

Calibration data partitioning. We partition the calibration data as follows: Given a hash function h we define the i-th partition of the calibration set as $P_i^c = \{(x_j, y_j) \in \mathcal{D}_{calib} : h(x_j) \equiv i \,(\text{mod } k_c)\}$. We then use a (potentially reliable) conformal score function s to compute the conformal scores $S_i = \{s(x_j, y_j)\}_{(x_j, y_j) \in P_i^c}$ on each partition P_i^c. We can then determine the α_{n_i}-quantiles of the separate conformal scores, $\tau_i = \text{Quant}(\alpha_{n_i}; S_i)$, where n_i is the size of the i-th partition, $n_i = |C_i|$.

Majority prediction sets. Now we propose our novel prediction sets reliable under calibration poisoning. Given a new datapoint $x_{n+1} \in \mathcal{D}_{test}$ we construct k_c prediction sets for each partition as $\mathcal{C}_i(x_{n+1}) = \{y : s(x_{n+1}, y) \geq \tau_i\}$. We then construct a prediction set composed of all classes that appear in the majority of *independent* prediction sets:

$$\mathcal{C}^M(x_{n+1}) = \left\{ y : \sum_{i=1}^{k_c} \mathbb{1}\{y \in \mathcal{C}_i(x_{n+1})\} > \hat{\tau} \right\} \tag{3}$$

with $\hat{\tau} = \max\{x \in [k_c] : F(x) \leq \alpha\}$, where F is the CDF of the Binomial distribution $\text{Bin}(k_c, 1-\alpha)$ and $[k_c] = \{0, \ldots, k_c\}$. Note that for $k_c = 1$ we have $\hat{\tau} = 0$ and thus the majority prediction sets amount to standard conformal prediction sets, $\mathcal{C}^M(x_{n+1}) = \mathcal{C}_1(x_{n+1})$. We summarize the construction of our majority prediction sets in Algorithm 2.

Interestingly, such majority voting is also used in the context of uncertainty sets (Gasparin & Ramdas, 2024), but their construction comes without reliability guarantees and additionally violates the coverage guarantee (see discussion in Appendix D). In contrast, we show that our majority prediction sets $\mathcal{C}^M$ achieve marginal coverage on clean data:

Theorem 2. *Given coverage probability $1-\alpha \in (0,1)$, test sample $(x_{n+1}, y_{n+1}) \in \mathcal{D}_{test}$, and a conformal score function s, the majority prediction set (Equation 3) constructed on a calibration set of independent datapoints achieves marginal coverage on clean data:*

$$\Pr[y_{n+1} \in \mathcal{C}^M(x_{n+1})] \geq 1 - \alpha \tag{4}$$

Proof in Appendix D. Notably, Theorem 2 guarantees marginal coverage for any conformal score function. This holds especially for our smoothed score function (Algorithm 1). As a result, majority prediction sets based on the smoothed score function achieve marginal coverage (Desideratum **II**).

6 Provable guarantees for reliable conformal prediction sets

After introducing reliable prediction sets (RPS), we derive certificates for their reliability as defined in Definition 1 and required by Desideratum **III**. We consider the threat model $B_{r_t,r_c}(\mathcal{D}_l)$ where adversaries can insert, delete and flip labels for up to r_t training and r_c calibration points (Section 3). In the following we treat training poisoning, then calibration poisoning, and finally poisoning of both.

6.1 Guarantees for smoothed scoring function under training poisoning

We begin with the reliability of the smoothed scoring function under training poisoning ($r_t{>}0$, $r_c{=}0$). Let $\mathcal{C}(x_{n+1}){=}\{y \in \mathcal{Y} : s(x_{n+1}, y) \geq \tau\}$ be a prediction set for a new test point x_{n+1} derived using conformal prediction (Section 3) under the clean dataset $\mathcal{D}_l$ with smoothed score function s. Our goal is to bound the prediction set $\tilde{\mathcal{C}}(x_{n+1})$ derived under any poisoned dataset $\tilde{\mathcal{D}}_l \in B_{r_t,r_c}(\mathcal{D}_l)$. This requires that we bound score function s and quantile τ. We start with the score function:

Lemma 2. *We can upper bound the score function for any $\tilde{\mathcal{D}}_l \in B_{r_t,r_c}(\mathcal{D}_l)$ as follows:*

$$\overline{s}(x,y) = \max_{\substack{0\leq\pi_i\leq 1 \\ \Delta_i\in\{0,\pm\frac{1}{k_t},\ldots,\pm\frac{r_t}{k_t}\} \\ \sum_{i=1}^K \Delta_i = 0}} \frac{e^{\pi_y}}{\sum_{i=1}^K e^{\pi_i}} \quad \textit{with} \quad \pi = [\pi_1(x)+\Delta_1, \ldots, \pi_K(x)+\Delta_K] \tag{5}$$

Proof in Appendix E. Although optimizing softmax functions typically leads to non-convex optimization problems, the problem in (5) reduces to a discrete optimization problem that can be solved efficiently (Desideratum **V**). We derive algorithms computing lower and upper bounds in r_t steps, presenting the upper bound in Algorithm 3. Intuitively, in each step we greedily redistribute $\frac{1}{k_t}$ probability mass from the current class $\hat{y}{\neq}y$ with the largest probability mass to the target class y. We repeat this process until we have redistributed the entire probability mass $\frac{r_t}{k_t}$. We present the lower bound algorithm and proofs in Appendix E.

Algorithm 3 Greedy algorithm for upper bounding the smoothed score function s

Input: Score function s, x, y, k_t, r_t
$\pi = [\pi_1(x), \ldots, \pi_K(x)]$
for $i = 1$ **to** r_t **do**
  $\hat{y} = \operatorname{argmax}_{\hat{y}\neq y} \pi_{\hat{y}}$
  $\pi_{\hat{y}} \leftarrow \min(\max(\pi_{\hat{y}} - 1/k_t, 0), 1)$
  $\pi_y \leftarrow \min(\max(\pi_y + 1/k_t, 0), 1)$
Output: $\overline{s}(x,y) = e^{\pi_y}/(\sum_{i=1}^K e^{\pi_i})$

Given lower and upper bounds $\underline{z_i} = \underline{s}(x_i, y_i)$ and $\overline{z_i} = \overline{s}(x_i, y_i)$ on the conformal scores for all points $(x_i, y_i) \in \mathcal{D}_{calib}$ in the calibration set, we can directly determine the worst-case quantiles:

$$\underline{\tau} = \text{Quant}(\alpha_n; \{\underline{z_i}\}_{i=1}^n) \qquad \overline{\tau} = \text{Quant}(\alpha_n; \{\overline{z_i}\}_{i=1}^n)$$

Finally we need to identify (1) the class within the prediction set that received the fewest votes from the classifiers f_i and (2) the class outside the prediction set that got most votes from the classifiers:

$$\underline{y} = \operatorname*{argmin}_{y\in\mathcal{C}(x_{n+1})} \pi_y(x) \qquad \overline{y} = \operatorname*{argmax}_{y\notin\mathcal{C}(x_{n+1})} \pi_y(x)$$

Then we can provide the following guarantees (Proof in Appendix E):

Theorem 3. *Given $r_c{=}0$, the conformal prediction set $\tilde{\mathcal{C}}(x_{n+1})$ derived with the smoothed score function under any poisoned dataset $\tilde{\mathcal{D}}_l{\in}B_{r_t,r_c}(\mathcal{D}_l)$ is coverage reliable, i.e. $\tilde{\mathcal{C}}(x_{n+1}) \supseteq \mathcal{C}(x_{n+1})$, if $\underline{s}\left(x,\underline{y}\right) \geq \overline{\tau}$, and size reliable, i.e. $\tilde{\mathcal{C}}(x_{n+1}) \subseteq \mathcal{C}(x_{n+1})$, if $\overline{s}\left(x,\overline{y}\right) < \underline{\tau}$.*

Intuitively, if class $\underline{y}$ cannot be removed from $\mathcal{C}(x_{n+1})$ ($\overline{y}$ added), adversaries cannot remove (add) other classes and thus the prediction sets are coverage (size) reliable.

6.2 GUARANTEES FOR MAJORITY PREDICTION SETS UNDER CALIBRATION POISONING

Now we analyze reliability of majority prediction sets under calibration poisoning ($r_t = 0$,$r_c > 0$). Let $\mathcal{C}^M(x_{n+1})$ be the majority prediction set for a new test point x_{n+1} derived under the clean dataset $\mathcal{D}_l$ using any deterministic conformal score function s. Intuitively, if adversaries cannot remove (add) a class from the majority prediction set by removing (adding) it from (to) r_c individual prediction sets, the majority prediction set remains coverage (size) reliable even in the worst case. This is since adversaries can perturb at most r_c calibration partitions. To determine if adversaries can remove or add classes, we have to count minimum and maximum support $\sum_{i=1}^{k_c} \mathbb{1}\{y \in \mathcal{C}_i(x_{n+1})\}$ for classes in and outside of the majority set:

$$\underline{m} = \min_{y \in \mathcal{C}^M} \sum_{i=1}^{k_c} \mathbb{1}\{y \in \mathcal{C}_i(x_{n+1})\} \qquad \overline{m} = \max_{y \notin \mathcal{C}^M} \sum_{i=1}^{k_c} \mathbb{1}\{y \in \mathcal{C}_i(x_{n+1})\}$$

Using each support we can provide the following guarantees (Proof in Appendix E):

Theorem 4. *Given r_t=0 and deterministic score function s, the majority prediction set $\tilde{\mathcal{C}}^M(x_{n+1})$ derived under any dataset $\tilde{\mathcal{D}}_l \in B_{r_t,r_c}(\mathcal{D}_l)$ is coverage reliable, i.e. $\tilde{\mathcal{C}}^M(x_{n+1}) \supseteq \mathcal{C}^M(x_{n+1})$, if $\underline{m} - r_c > \hat{\tau}$, and size reliable, i.e. $\tilde{\mathcal{C}}^M(x_{n+1}) \subseteq \mathcal{C}^M(x_{n+1})$, if $\overline{m} + r_c \leq \hat{\tau}$, provided that the smallest calibration partition i^* is large enough $|P^c_{i^*}| - r_c \geq (\frac{1}{\alpha} - 1)$.*

Note that the last condition ensures that the calibration partitions are large enough such that worst-case adversaries cannot delete datapoints to prevent us from computing the majority prediction sets.

6.3 PROVABLE RELIABILITY GUARANTEES FOR RPS UNDER GENERAL DATA POISONING

Finally, we consider poisoning of training and calibration data ($r_t > 0, r_c > 0$).

Coverage reliability. To ensure coverage reliability we have to show that all classes $y \in \mathcal{C}^M$ are guaranteed to be in the majority prediction set under worst-case poisoning. The majority prediction set $\mathcal{C}^M$ contains a class y only if it appears in a majority of $\hat{\tau}$ individual prediction sets $\mathcal{C}_i$. Under calibration poisoning, adversaries can remove classes from r_c individual prediction sets. Intuitively, the number of prediction sets reliable under training poisoning β_y must be large enough such that even under calibration poisoning, the number of prediction sets containing the class is still larger than the threshold, $\beta_y - r_c > \hat{\tau}$. This leads to the following guarantee (Proof in Appendix E):

Theorem 5. *Let β_y denote the number of prediction sets $\mathcal{C}_i \in \{\mathcal{C}_1, \ldots, \mathcal{C}_{k_c}\}$ for which we can guarantee $y \in \mathcal{C}_i$ under r_t poisoned training datapoints. If $\beta_y - r_c > \hat{\tau}$ for all $y \in \mathcal{C}^M(x_{n+1})$ then the majority prediction set is coverage reliable under any dataset $\tilde{\mathcal{D}}_l \in B_{r_t,r_c}(\mathcal{D}_l)$, provided that the smallest calibration partition i^* is large enough $|P^c_{i^*}| - r_c \geq (\frac{1}{\alpha} - 1)$.*

Size reliability. To ensure size reliability we have to show that all classes $y \notin \mathcal{C}^M$ are guaranteed to stay outside of the majority prediction set under worst-case poisoning. The majority prediction set $\mathcal{C}^M$ does not contain a class y if it appears in less than or equal to $\hat{\tau}$ individual prediction sets $\mathcal{C}_i$. Under calibration poisoning, adversaries can add classes to r_c prediction sets in the worst-case. Intuitively, if we can guarantee that γ_y prediction sets $\mathcal{C}_i$ do not contain the class y under training poisoning, at most $k_c - \gamma_y$ prediction sets contain the class under worst-case training poisoning. This number of prediction sets containing the class in the worst-case must be small enough such that even if adversaries add the class to r_c prediction sets, the majority prediction set does not contain the class, $k_c - \gamma_y + r_c \leq \hat{\tau}$. This leads to the following guarantee (Proof in Appendix E):

Theorem 6. *Let γ_y denote the number of prediction sets $\mathcal{C}_i \in \{\mathcal{C}_1, \ldots, \mathcal{C}_{k_c}\}$ for which we can guarantee $y \notin \mathcal{C}_i$ under r_t poisoned training datapoints. If $k_c - \gamma_y + r_c \leq \tau$ for all $y \notin \mathcal{C}^M(x_{n+1})$ then the majority prediction set is size reliable under any dataset $\tilde{\mathcal{D}}_l \in B_{r_t,r_c}(\mathcal{D}_l)$, provided that the smallest calibration partition i^* is large enough $|P^c_{i^*}| - r_c \geq (\frac{1}{\alpha} - 1)$.*

Note that we can efficiently compute numbers β_y and γ_y as described in Subsection 6.1 by computing the worst-case quantiles and verifying $\underline{s}(x, y) < \overline{\tau}$ and $\overline{s}(x, y) < \underline{\tau}$, respectively. Our overall certification approach is efficient in practice (Desideratum **V**) as we discuss in Appendix E and experimentally demonstrate in the next section.

7 EXPERIMENTAL EVALUATION

In this section we evaluate our reliabe prediction sets and their worst-case guarantees, demonstrating their effectiveness in analyzing and improving reliability of conformal prediction under poisoning. We compare three settings (calibration poisoning, training poisoning, and poisoning of both) by computing the following prediction sets: (a) majority prediction sets merging multiple homogeneous prediction sets calibrated on each partition, (b) conformal prediction sets using our smoothed score function, and (c) majority prediction sets using our smoothed score function.

Datasets and models. We train ResNet18, ResNet50 and ResNet101 models (He et al., 2016) on SVHN (Netzer et al., 2011), CIFAR10 and CIFAR100 (Krizhevsky et al., 2009). The datasets contain images with 3 channels of size 32x32, categorized into 10, 10 and 100 classes. We show results for ResNet18 on CIFAR10 and coverage level α=0.1 here and additional results in Appendix B.

Experimental setup. We randomly select 1,000 images of the test set for calibration and use the remaining 9,000 datapoints for testing. To account for randomness in training and calibration set sampling we train 5 classifiers with different initializations and validate each of them on 5 different calibration splits. We report mean and standard deviation (shaded areas in the plots). We refer to Appendix A for the full experimental setup including detailed reproducibility instructions.

Evaluation metrics. We report three *reliability ratios*: The ratios of test datapoints whose prediction sets are, according to our worst-case analysis, coverage reliable (classes cannot be removed), size reliable (classes cannot be added), or robust (classes cannot removed or added). *Empirical coverage* refers to the ratio of datapoints whose prediction sets cover the ground truth label of the test set. We also report the *average size* of the prediction sets computed on the test set.

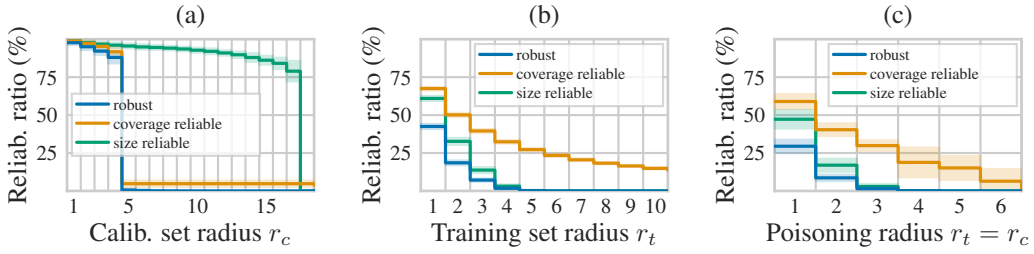

Figure 2: Worst-case reliability guarantees across three scenarios: (a) poisoning of the calibration data, (b) poisoning of the training data, and (c) poisoning of both datasets. Our guarantees against coverage attacks are stronger when training data is poisoned, whereas for calibration attacks our method offers stronger guarantees against size attacks. Notably, even under strong adversarial conditions where both datasets can be poisoned we still provide non-trivial reliability guarantees.

(a) Reliability of majority prediction sets under calibration poisoning. Majority prediction sets demonstrate strong reliability guarantees against calibration set poisoning in empirical evaluations (Figure 2 a). Specifically, we construct majority prediction sets by merging k_c=22 homogeneous prediction sets, each calibrated on separate calibration partitions, resulting in an empirical coverage of 90.6% and an average set size of 0.95. When up to r_c=4 datapoints in the calibration set are poisoned, our method guarantees that over 97% of the prediction sets remain reliable against worst-case coverage attacks. The guarantees against set size attacks are even stronger: Even if r_c=17 datapoints are poisoned we still guarantee that over 79% of the prediction sets remain size reliable.

(b) Reliability of smoothed score function under training poisoning. The setting of training set poisoning is considerably more challenging since adversaries can simultaneously manipulate the quantiles during calibration and the scores at inference. We compute conformal prediction sets using our smoothed score function on k_t=100 training partitions, resulting in empirical coverage of 90.6% and average set size of 3.3 (Figure 2 b). Despite strong adversaries, our reliable prediction sets still manage to provide non-trivial reliability guarantees under worst-case perturbations. Specifically, when up to $r_t = 5$ datapoints in the training set are poisoned, our method guarantees that over 25% of the prediction sets remain reliable against worst-case coverage attacks.

(c) Reliability of RPS under training and calibration poisoning. By far the most challenging setting constitute adversaries that manipulate both training and calibration data. We compute majority prediction sets using our smoothed score function on k_t=100 training partitions and k_c=40 calibration partitions, resulting in empirical coverage of 92.3% and average set size of 3.6 (Figure 2 c). Notably, under poisoning of up to $r_t = 3$ training *and* $r_c = 3$ calibration points, our method still guarantees that over 23% of the prediction sets remain reliable against worst-case coverage attacks.

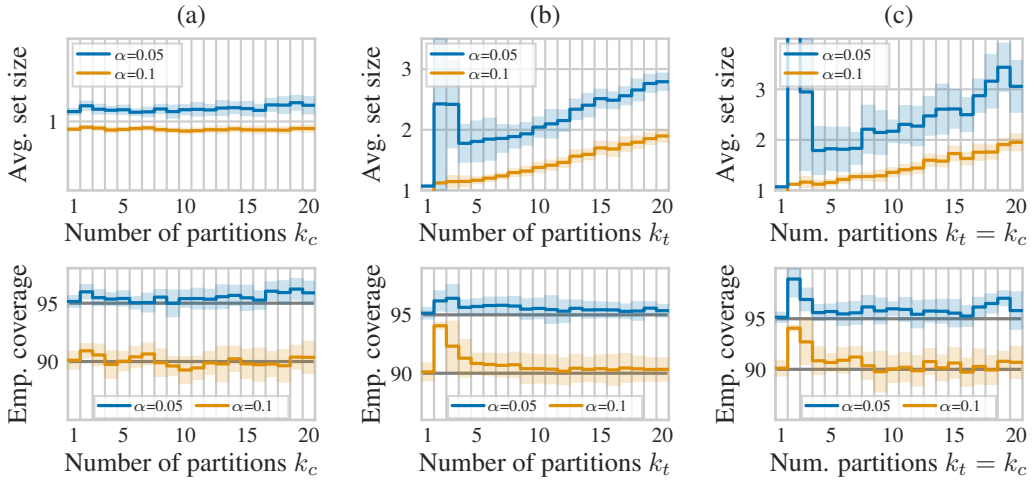

Figure 3: Average set size and empirical coverage in all three different experiment settings (a–c). Notably, our reliable prediction sets yield valid coverage guarantees without becoming too large.

Average prediction set size. In Figure 3 (top row) we study the average set size for varying numbers of partitions. Notably, our majority prediction sets yield strong guarantees (Figure 2 a) without any significant increase in size (Figure 3 a): the average prediction set size remains below 1, which even holds for k_c=40 (α=0.1). Interestingly, the size slowly increases with more training partitions, creating a tade-off between reliability and utility of the prediction sets. Practitioners have to fine-tune this trade-off depending on the application's sensitivity towards reliability. Overall, we empirically find that our reliable prediction sets do not become too large in size (Desideratum **II**).

Note that the set size increases when using only k_t=2, 3 training partitions (spikes in the plots). This happens because a sufficient number of classifiers is needed to achieve consistent consensus in practice. However, our analysis also shows that five classifiers are already sufficient to prevent excessively large prediction sets. Finally, we also study the sizes on different datasets under calibration poisoning (setting (a)): On CIFAR100 the average set size remains around 4 for $k_c \leq 25$, demonstrating that our majority prediction sets scale well to datasets with more classes (Figure 4 (1)).

Coverage guarantees. In Figure 3 (bottom row) we empirically validate the coverage guarantee (Theorem 2) on clean data. We observe that empirical coverage matches the desired coverage probability $1 - \alpha$ closely on average (Desideratum **I**), confirming our theoretical statements. We observe overcoverage for small numbers of training partitions, which can be explained by our previous finding that a minimum number of classifier is needed to achieve consistent consensus in practice.

Softmax ablation study. We found that smoothing the voting function with a softmax (Section 5) avoids overly large prediction sets and overcoverage in practice. To demonstrate this we conduct an experiment (Figure 4 (2,3)) for varying numbers of training partitions k_t, where we compare conformal prediction with our smoothed score function s against using the voting function π only.

Computational efficiency. Training the classifiers takes most of the time (statistics in Appendix A). Note, however, that while having to train more classifiers, each one is trained on a subset of the training data, which can speed up the training process. Inference with the ResNet18 models takes between 4 and 10 seconds on CIFAR10. Constructing the conformal prediction sets takes at most 0.5 seconds for the entire test set. Computing certificates for majority prediction sets takes less than a second, and around one minute when computing guarantees under training and calibration poisoning. Overall, we found that reliable prediction sets are computationally efficient (Desideratum **V**).

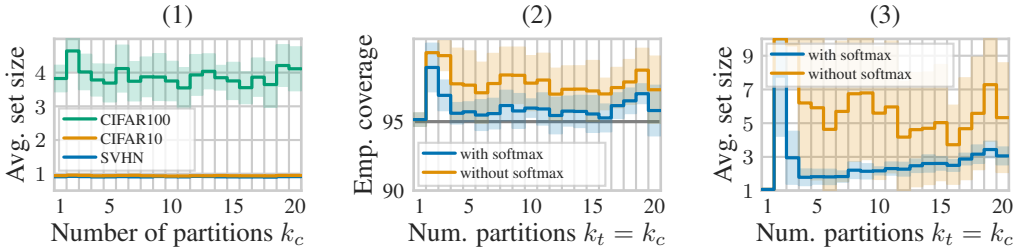

Figure 4: (1): Average prediction set size of majority prediction sets across three different datasets. (2,3): Softmax ablation study for empirically justifying smoothing of the voting function ($\alpha = 0.05$).

8 Discussion

Reproducible prediction sets. Score functions in the literature are not just unreliable but some also depend on randomization to break ties (Romano et al., 2020). While randomization at inference changes prediction sets by at most one class (Angelopoulos et al., 2021), different prediction sets for the same input may not be desirable from a reliability standpoint, violating Desideratum **V**. For example, a differential diagnosis for two patients with identical health should not yield different results. As a remedy, we propose to make randomized score functions reproducible (Appendix C).

Limitations. Although RPS computationally scales to larger settings, training on subsets of larger datasets such as CIFAR100 or ImageNet comes with accuracy loss, which also affects the utility of our smoothed score function (details in Appendix B). This accuracy loss is an open challenge in the general certifiably robust classification literature and beyond the scope of this paper.

Minimal calibration set size for majority prediction sets. Recall that Desideratum **VI** requires that the reliability of prediction sets must increase for larger calibration sets. Our majority prediction sets fulfill this desideratum by construction since increasing the number partitions k_c will decrease the influence of datapoints. However, we need enough data to construct our prediction sets: due to the finite-sample correction, the calibration partitions cannot become arbitrarily small (Section 6). If the hashing function would distribute all calibration images into equally-sized partitions of size n/k_c, we would need at least $n \geq k_c \left(\frac{1}{\alpha} - 1\right)$ calibration points in total (Proof in Appendix D). Notably, this relationship is linear: given a fixed coverage probability $1-\alpha$, increasing the calibration partitions by a factor of k only requires k-times larger calibration sets, which is realistic for all commonly used image classification datasets and coverage probabilities used in the literature.

Training poisoning discussion. Note that as long as adversaries only know that the calibration set is any dataset exchangeable with the test data, they cannot compromise the coverage guarantee (Theorem 1) by poisoning training sets. This holds because conformal scores are computed using a fixed classifier (post-training). However in practice, worst-case adversaries may know the calibration set during training poisoning. Furthermore, even without access to the calibration data, adversaries can still attack the prediction set size: For example, adversaries can manipulate the training process to degrade the performance of the score function, leading to excessively large prediction sets. This underscores again the need to ensure coverage and size *reliability* as defined in Definition 1.

9 Conclusion

We introduce *reliable prediction sets* (RPS), a novel method designed to improve reliability of conformal prediction in the presence of data poisoning and label flipping attacks. By leveraging smoothed score functions and a majority voting mechanism, RPS effectively mitigates the influence of adversarial perturbations during both training and calibration. We provide theoretical guarantees that RPS maintains stability under worst-case data poisoning, and demonstrate the effectiveness of our approach on image classification tasks. Overall, our approach represents an important contribution towards more reliable uncertainty quantification in practice, fostering the trustworthiness in real-world scenarios where data integrity cannot be guaranteed.

ACKNOWLEDGMENTS

The authors want to thank Jan Schuchardt, Lukas Gosch and Leo Schwinn for valuable feedback on the manuscript. This work has been funded by the DAAD program Konrad Zuse Schools of Excellence in Artificial Intelligence (sponsored by the Federal Ministry of Education and Research). The authors of this work take full responsibility for its content.

REFERENCES

Anastasios N Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021.

Anastasios Nikolas Angelopoulos, Stephen Bates, Michael I. Jordan, and Jitendra Malik. Uncertainty sets for image classifiers using conformal prediction. In *ICLR*. OpenReview.net, 2021.

Andreas Auer, Martin Gauch, Daniel Klotz, and Sepp Hochreiter. Conformal prediction for time series with modern hopfield networks. In *NeurIPS*, 2023.

Philip J Boland, Harshinder Singh, and Bojan Cukic. Stochastic orders in partition and random testing of software. *Journal of applied probability*, 39(3):555–565, 2002.

Maxime Cauchois, Suyash Gupta, Alnur Ali, and John C. Duchi. Robust validation: Confident predictions even when distributions shift. *CoRR*, abs/2008.04267, 2020.

Giovanni Cherubin. Majority vote ensembles of conformal predictors. *Machine Learning*, 108(3): 475–488, 2019.

Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning*, pp. 1310–1320. PMLR, 2019.

Bat-Sheva Einbinder, Stephen Bates, Anastasios N. Angelopoulos, Asaf Gendler, and Yaniv Romano. Conformal prediction is robust to label noise. *CoRR*, abs/2209.14295, 2022.

Matteo Gasparin and Aaditya Ramdas. Merging uncertainty sets via majority vote. *arXiv preprint arXiv:2401.09379*, 2024.

Asaf Gendler, Tsui-Wei Weng, Luca Daniel, and Yaniv Romano. Adversarially robust conformal prediction. In *ICLR*. OpenReview.net, 2022.

Subhankar Ghosh, Yuanjie Shi, Taha Belkhouja, Yan Yan, Jana Doppa, and Brian Jones. Probabilistically robust conformal prediction. In *UAI*, volume 216 of *Proceedings of Machine Learning Research*, pp. 681–690. PMLR, 2023.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pp. 770–778. IEEE Computer Society, 2016.

Linus Jeary, Tom Kuipers, Mehran Hosseini, and Nicola Paoletti. Verifiably robust conformal prediction. *CoRR*, abs/2405.18942, 2024.

Mintong Kang, Zhen Lin, Jimeng Sun, Cao Xiao, and Bo Li. Certifiably byzantine-robust federated conformal prediction. In *ICML*. OpenReview.net, 2024.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113 (523):1094–1111, 2018.

Alexander Levine and Soheil Feizi. Deep partition aggregation: Provable defenses against general poisoning attacks. In *ICLR*. OpenReview.net, 2021.

Yangyi Li, Aobo Chen, Wei Qian, Chenxu Zhao, Divya Lidder, and Mengdi Huai. Data poisoning attacks against conformal prediction. In *Forty-first International Conference on Machine Learning*, 2024.

Lars Lindemann, Matthew Cleaveland, Gihyun Shim, and George J. Pappas. Safe planning in dynamic environments using conformal prediction. *IEEE Robotics Autom. Lett.*, 8(8):5116–5123, 2023.

Ziquan Liu, Yufei Cui, Yan Yan, Yi Xu, Xiangyang Ji, Xue Liu, and Antoni B. Chan. The pitfalls and promise of conformal inference under adversarial attacks. In *ICML*. OpenReview.net, 2024.

Ilya Loshchilov and Frank Hutter. SGDR: stochastic gradient descent with warm restarts. In *ICLR (Poster)*. OpenReview.net, 2017.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Baolin Wu, Andrew Y Ng, et al. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, pp. 4. Granada, 2011.

Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. Inductive confidence machines for regression. In *ECML*, volume 2430 of *Lecture Notes in Computer Science*, pp. 345–356. Springer, 2002.

Sangdon Park, Osbert Bastani, and Taesoo Kim. Acon^2: Adaptive conformal consensus for provable blockchain oracles. In *USENIX Security Symposium*, pp. 3313–3330. USENIX Association, 2023.

Coby Penso and Jacob Goldberger. A conformal prediction score that is robust to label noise. *CoRR*, abs/2405.02648, 2024.

Keivan Rezaei, Kiarash Banihashem, Atoosa Malemir Chegini, and Soheil Feizi. Run-off election: Improved provable defense against data poisoning attacks. In *ICML*, volume 202 of *Proceedings of Machine Learning Research*, pp. 29030–29050. PMLR, 2023.

Yaniv Romano, Matteo Sesia, and Emmanuel J. Candès. Classification with valid and adaptive coverage. In *NeurIPS*, 2020.

Elan Rosenfeld, Ezra Winston, Pradeep Ravikumar, and J. Zico Kolter. Certified robustness to label-flipping attacks via randomized smoothing. In *ICML*, volume 119 of *Proceedings of Machine Learning Research*, pp. 8230–8241. PMLR, 2020.

Mauricio Sadinle, Jing Lei, and Larry Wasserman. Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 114(525):223–234, 2019.

Aldo Solari and Vera Djordjilović. Multi split conformal prediction. *Statistics & Probability Letters*, 184:109395, 2022.

Philip Sosnin, Mark Niklas Müller, Maximilian Baader, Calvin Tsay, and Matthew Wicker. Certified robustness to data poisoning in gradient-based training. *CoRR*, abs/2406.05670, 2024.

Wenpin Tang and Fengmin Tang. The poisson binomial distribution—old & new. *Statistical Science*, 38(1):108–119, 2023.

Zhiyi Tian, Lei Cui, Jie Liang, and Shui Yu. A comprehensive survey on poisoning attacks and countermeasures in machine learning. *ACM Comput. Surv.*, 55(8):166:1–166:35, 2023.

Janette Vazquez and Julio C. Facelli. Conformal prediction in clinical medical sciences. *J. Heal. Informatics Res.*, 6(3):241–252, 2022.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*, volume 29. Springer, 2005.

Volodya Vovk, Alexander Gammerman, and Craig Saunders. Machine-learning applications of algorithmic randomness. In *ICML*, pp. 444–453. Morgan Kaufmann, 1999.

Wenxiao Wang, Alexander Levine, and Soheil Feizi. Improved certified defenses against data poisoning with (deterministic) finite aggregation. In *ICML*, volume 162 of *Proceedings of Machine Learning Research*, pp. 22769–22783. PMLR, 2022.

Ge Yan, Yaniv Romano, and Tsui-Wei Weng. Provably robust conformal prediction with improved efficiency. In *ICLR*. OpenReview.net, 2024.

Soroush H Zargarbashi, Mohammad Sadegh Akhondzadeh, and Aleksandar Bojchevski. Robust yet efficient conformal prediction sets. *arXiv preprint arXiv:2407.09165*, 2024.

APPENDIX OVERVIEW

In this appendix we provide additional results, details on our experimental setup and prove theoretical results as outlined in the following:

A FULL EXPERIMENTAL SETUP AND REPRODUCIBILITY DETAILS

We provide details on the experimental setup to ensure reproducibility of our results.

Datasets. The datasets we use for evaluation are described in Section 7 (SVHN (Netzer et al., 2011), CIFAR10 and CIFAR100 (Krizhevsky et al., 2009)) and are publicly available. We use the torchvision library to load the datasets.[1] We normalize images before training, and we compute dataset mean and standard deviation on each training partition separately to ensure models are trained on isolated partitions, which is required by our method to improve reliability.

Training details. We train all models with stochastic gradient descent (learning rate 0.01, momentum 0.9, weight decay 5e-4) for 400 epochs using early stopping if the training accuracy does not improve for 100 epochs. We further deploy a cosine learning rate scheduler (Loshchilov & Hutter, 2017). We use a batch-size of 128 during training and 300 at inference. To ensure that our guarantees against training-poisoning hold we require that the training process is deterministic, which not only involves fixing the random seed for data augmentation but also ensuring that the training processes are deterministic.

Image preprocessing. We determine dataset-wide mean and standard deviation dynamically on every training set partition separately once before training to ensure that each classifier is trained on an isolated partition. We subsequently normalize all images in one partition with the corresponding values. We also augment the training set with random crops (padding of 4 pixels) and random horizontal flips (but we perform the data augmentation in a deterministic, reproducible way across runs).

Hardware details. We train ResNet18 models on a NVIDIA GTX 1080TI GPU, and the ResNet50 and ResNet101 models on a NVIDIA A100 40GB. We perform inference of all models on a NVIDIA GTX 1080TI GPU, and compute certificates on a Xeon E5-2630 v4 CPU.

Reproducibility. To ensure reproducibility we use random seeds for all randomized functions, this especially includes the dataset preprocessing, model training and calibration splits. We will publish source code along with reproducibility instructions and all random seeds.

Training time details. The runtime statistics for training ResNet18 models on CIFAR-10 and SVHN are as follows. Training a single ResNet18 model on CIFAR-10 takes 2.2 hours, while training 100 models requires a total of 21.7 hours, with each individual model taking approximately 2.6 minutes. For SVHN, a single ResNet18 model takes 5.6 hours to train, and training 100 models requires a total of 30 hours, with each model training taking around 3.5 minutes.

B ADDITIONAL RESULTS FOR RELIABLE PREDICTION SETS

In this section, we expand on the experimental results by providing further analyses and complementary information. Figure 5 shows the worst-case reliability guarantees for the SVHN dataset under the three different poisoning scenarios (in the same settings as described in the main paper). We again observe that our method provides reliable prediction sets even under worst-case poisoning attacks. Complementary to the main section, Figure 6 shows the average set size and empirical coverage in all three different experiment settings (a–c) on the SVHN dataset.

Figure 7 (1) shows the average set sizes of the three different architectures (ResNet18, ResNet50, ResNet101) on the CIFAR10 dataset when using majority vote prediction sets with the smoothed score function (third experimental setting). Interestingly, using the ResNet18 model yields the best results, which we attribute to the fact that models trained on subsets of the training set require less capacity to learn the data distribution, and small models prevent overfitting. Figure 7 (2,3) show the softmax ablation study for empirically justifying the smoothing of the voting function for $\alpha = 0.1$.

We also provide additional results for our reliable prediction sets for the following evaluation metrics: (1) the ratio of empty sets, (2) the ratio of full sets, (3) the singleton ratio (ratio of sets containing a single class), and (4) the singleton hit ratio (empirical coverage of singleton prediction sets). Figure 9 shows results for the CIFAR10 dataset, and Figure 10 shows results for the SVHN dataset, for all three evaluation settings (a-c) described in Section 5.

[1] `https://pytorch.org/vision/stable/index.html`

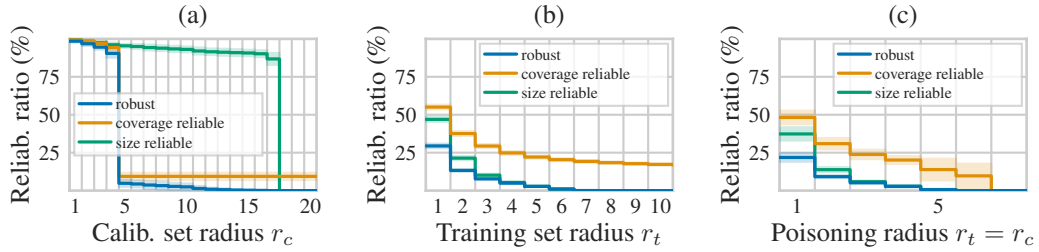

Figure 5: SVHN: Worst-case reliability guarantees across three scenarios: (a) poisoning of the calibration data, (b) poisoning of the training data, and (c) poisoning of both datasets.

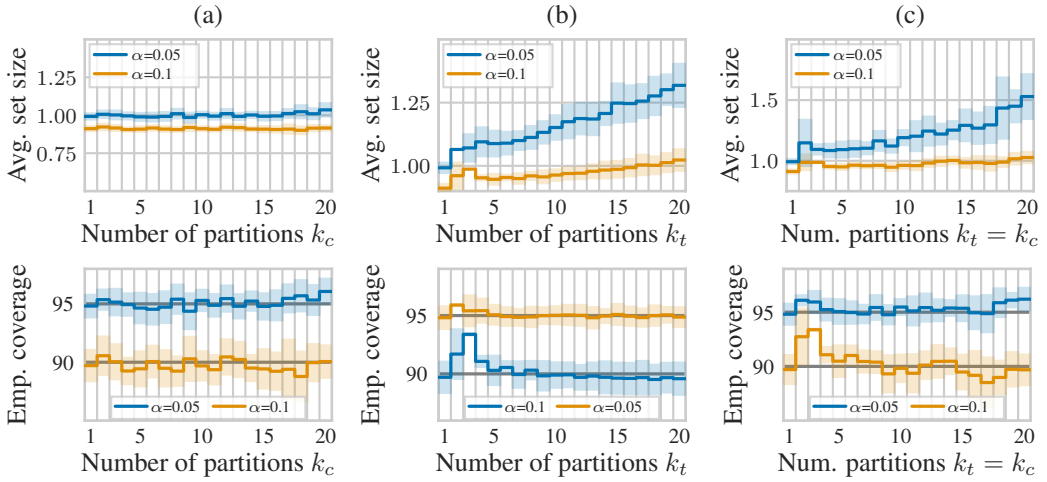

Figure 6: Avg. size and empirical coverage in all three settings (a–c) on the SVHN dataset.

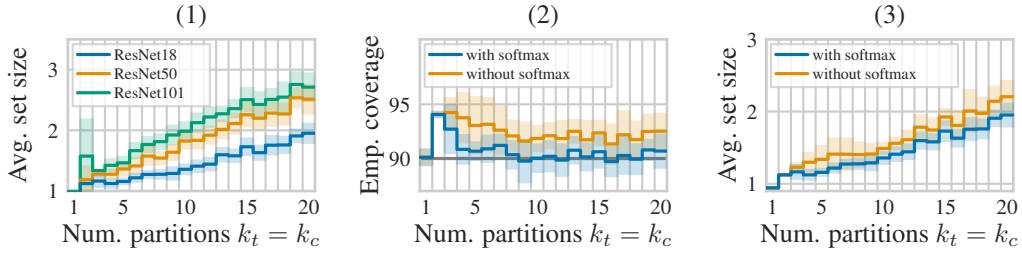

Figure 7: (1): Different models on CIFAR10. (2,3): Softmax ablation study for empirically justifying smoothing of the voting function (here with $\alpha = 0.1$) on CIFAR10.

In the main plots in Section 7, we only considered the diagonal $r_t = r_c$ for the reliability ratios. In Figure 8, we show the reliability ratios for coverage reliability, size reliability, and robustness for all combinations of r_t and r_c for the CIFAR10 dataset ($k_t = 100$, $k_c = 40$, $\alpha = 0.1$). Interestingly, the reliability ratios are generally higher for larger r_c, which indicates that the majority prediction sets are more reliable when calibration data is poisoned.

As mentioned in Section 8, training on subsets of larger datasets such as CIFAR100 or ImageNet comes with accuracy loss, which affects the utility of our smoothed score function. Specifically, when splitting the training set of CIFAR100 into 10 partitions only, each individual classifier achieves an accuracy of approximately 30% (in contrast to at least 70% when trained on the entire dataset). This affects the performance of our smoothed score function, leading to overly excessive prediction sets. Future learning algorithms could further improve the performance when training on subsets of larger datasets, ultimately boosting robustness and reliability in machine learning.

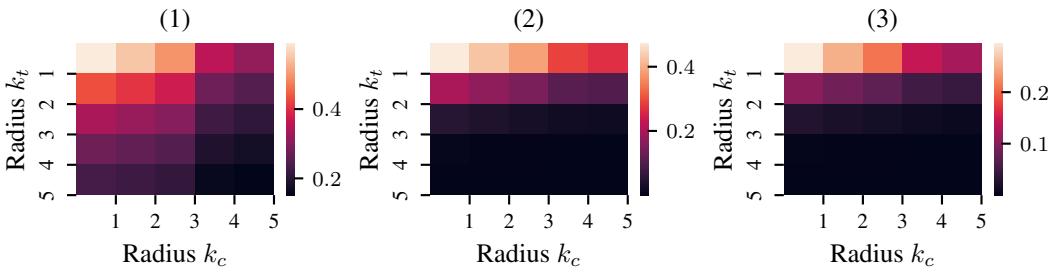

Figure 8: CIFAR10, $k_t = 100$, $k_c = 40$, $\alpha = 0.1$, (1): Coverage reliability ratio, (2): Size reliability ratio, (3): Robust ratio. Here with all radii combinations (and not just the diagonal, $r_t = r_c$).

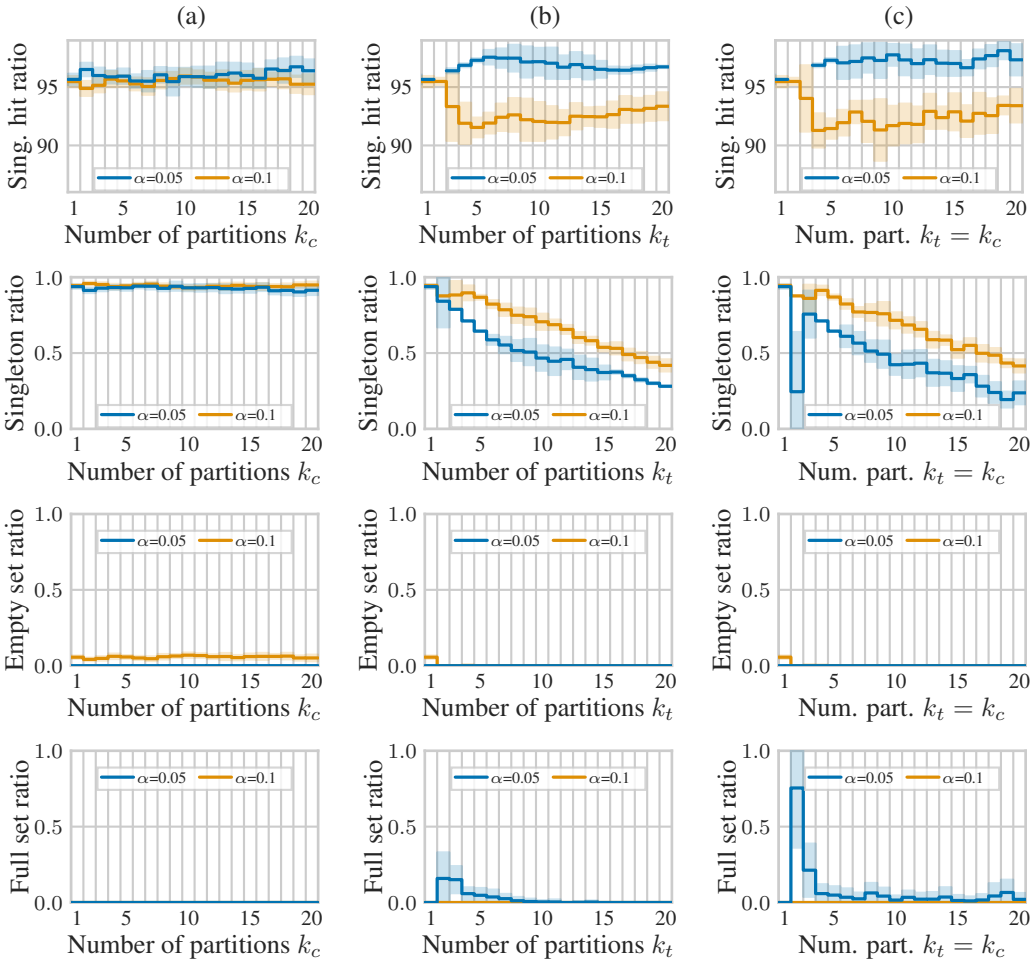

Figure 9: Various metrics for RPS in the three experiment settings (a-c) for ResNet18 on CIFAR10.

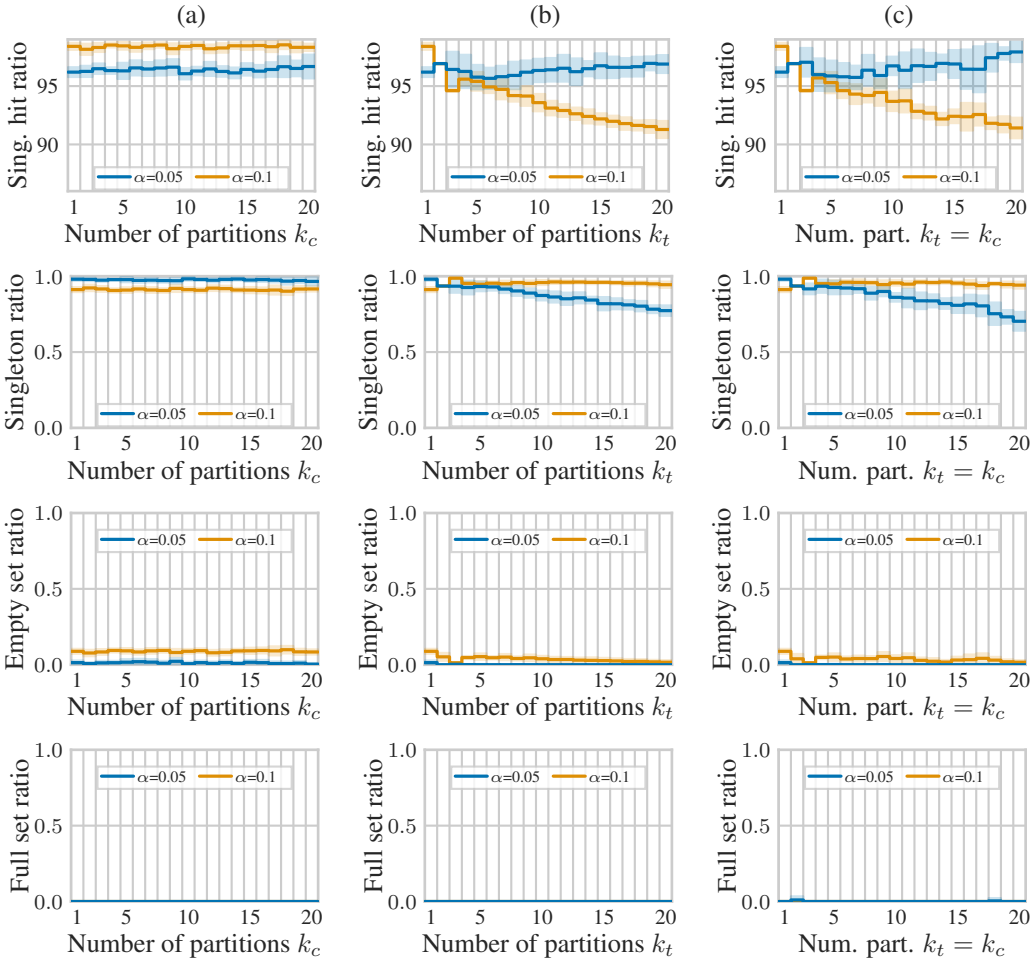

Figure 10: Various metrics for RPS in the three experiment settings (a-c) for ResNet18 on SVHN.

C REPRODUCIBLE PREDICTION SETS

There are conformal score functions that rely on random variables, e.g. for ensuring exact coverage by breaking ties. For example, adaptive prediction sets (APS) sum over class probabilities of classes with probability at least $f_y(x)$: $s(x,u,y)=-(\sum_{i=1}^{K} f_i(x)\mathbb{1}[f_i(x)>f_y(x)]+uf_y(x))$. Conformal prediction sets are then formed by $\mathcal{C}(x_{n+1}) = \{s(x_{n+1}, u_{n+1}, y) \geq \tau\}$, where $u_{n+1} \in [0,1]$ is a uniform random variable (Romano et al., 2020).

Although randomization at inference changes prediction sets by at most one class (Angelopoulos et al., 2021), generating different prediction sets for the same input may not be desirable from a reliability standpoint and violates Desideratum **V**. For example, a differential diagnosis for two patients with identical health parameters should not yield different results. As a solution we propose *reproducible score functions*: Instead of drawing from a random variable, we propose to compute pseudorandom numbers by hashing the sum of the image's pixel values and initialize the pseudorandom number with the hash. We found that this is enough to break ties in practice while ensuring determinism required by our reliability guarantees. Note that our score functions with pseudorandomization are valid since they maintain exchangeability as the numbers depend solely on the data itself (not its position in the dataset or other datapoints).

We analyze this in Figure 11: Without randomization APS results in large set sizes (2) despite tight coverage (1). With randomization the sets shrink in size, but are not reproducible. With pseudorandomization, APS is reproducible with tight coverage and small prediction sets.

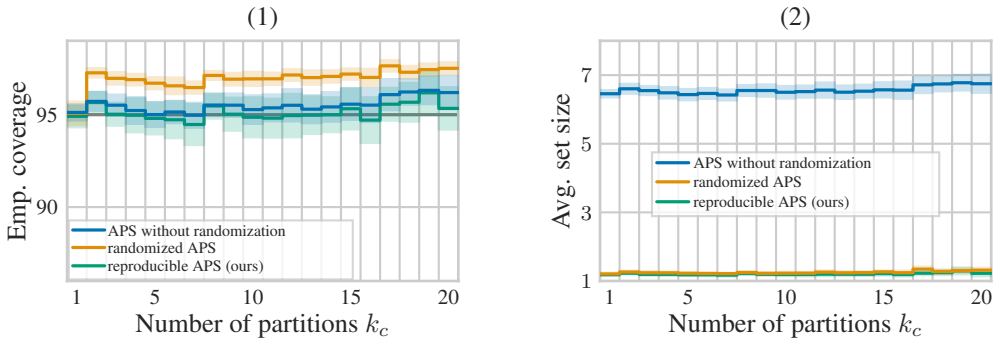

Figure 11: Comparing APS without and with randomization against our reproducible version.

D PROOFS FOR RELIABLE PREDICTION SETS (SECTION 5)

Recall the definition of the majority prediction sets from Section 5:

$$\mathcal{C}^M(x_{n+1}) = \left\{ y : \sum_{i=1}^{k_c} \mathbb{1}\{y \in \mathcal{C}_i(x_{n+1})\} > \hat{\tau} \right\}$$

with $\hat{\tau}=\max\{x \in [k_c] : F(x) \leq \alpha\}$, where F is the CDF of the Binomial distribution $\text{Bin}(k_c, 1-\alpha)$ and $[k_c] = \{0, \ldots, k_c\}$. Note that by assumption the sets $C_i(x_{n+1})$ are independent. We show the following result:

Theorem 2. *Given coverage probability $1-\alpha \in (0,1)$, test sample $(x_{n+1}, y_{n+1}) \in \mathcal{D}_{test}$, and a conformal score function s, the majority prediction set (Equation 3) constructed on a calibration set of independent datapoints achieves marginal coverage on clean data:*

$$\Pr[y_{n+1} \in \mathcal{C}^M(x_{n+1})] \geq 1 - \alpha \tag{6}$$

Proof. Define the event $\phi_i = \mathbb{1}\{y_{n+1} \in \mathcal{C}_i(x_{n+1})\}$. Note that y_{n+1} is fixed and the events ϕ_i and ϕ_j are independent for $i \neq j$ since the prediction sets $\mathcal{C}_i$ and $\mathcal{C}_j$ are constructed on disjoint partitions in a calibration set of independent datapoints. Therefore ϕ_i are independent Bernoulli random variables

with $p_i = \Pr[\phi_i = 1] \geq 1 - \alpha$ by construction. Further define the random variable

$$S_{k_c} = \sum_{i=1}^{k_c} \mathbb{1}\{y \in \mathcal{C}_i(x_{n+1})\}\,.$$

First we consider the special case $p_i = 1 - \alpha$ for all i. Then $S_{k_c} = \sum_{i=1}^{k_c} \phi_i$ is a Binomial random variable, $S_{k_c} \sim \text{Bin}(k_c, 1 - \alpha)$. Thus we have:

$$\begin{aligned}
\Pr[y_{n+1} \in \mathcal{C}^M(x_{n+1})] &= \Pr\left[\sum_{i=1}^{k_c} \mathbb{1}\{y_{n+1} \in \mathcal{C}_i(x_{n+1})\} > \hat{\tau}\right] && (7)\\
&= \Pr\left[S_{k_c} > \hat{\tau}\right] && (8)\\
&= 1 - \Pr\left[S_{k_c} \leq \hat{\tau}\right] && (9)\\
&= 1 - \underbrace{F(\hat{\tau})}_{\leq \alpha} && (10)\\
&\geq 1 - \alpha && (11)
\end{aligned}$$

In the general case of $p_i \geq 1 - \alpha$ we have that S_{k_c} is distributed as a Poisson binomial random variable, $S_{k_c} \sim PB(k_c, [p_1, \ldots, p_{k_c}])$. However, it holds that $\prod_{i=1}^{k_c} p_i > (1 - \alpha)^{k_c}$, which implies that the Poisson binomial distribution is stochastically larger than the Binomial distribution (Boland et al., 2002; Tang & Tang, 2023). See details in (Gasparin & Ramdas, 2024). Intuitively, this means that the probability $\Pr\left[S_{k_c} > \hat{\tau}\right]$ can only increase in the general case where $p_i \geq 1 - \alpha$. □

Discussion. Notably, Gasparin & Ramdas (2024) merge uncertainty sets with a similar majority vote but they define the threshold in their majority prediction set as $q = \sup\{x \in \mathbb{R} : F(x) \leq \alpha\}$ instead of $\hat{\tau} = \max\{x \in [k_c] : F(x) \leq \alpha\}$. However, clearly we have $F(q) > \alpha$ due to the definition of the supremum and the definition of the binomial CDF:

$$F(x; k_c, 1 - \alpha) = \sum_{i=0}^{\lfloor x \rfloor} \binom{k_c}{i} (1 - \alpha)^i \alpha^{k_c - i}$$

This means $q = \hat{\tau} + 1$, which leads to smaller majority prediction sets which violate the coverage guarantee since $\Pr[S_{k_c} > q] < 1 - \alpha$ due to $F(q) > \alpha$.

Minimal calibration set size for majority prediction sets. Recall that desideratum **VI** (Section 4) requires that the reliability of prediction sets must increase for larger calibration sets. Our majority prediction sets fulfill this desideratum by construction since increasing the number partitions k_c will decrease the influence of datapoints. However, we need enough data to construct prediction sets: due to the finite-sample correction, the calibration partitions cannot become arbitrarily small. This naturally bounds the number of partitions k_c for a fixed calibration size n (and thus reliability):

Proposition 1. *Let $i^* = \arg\min_{i \in \{1,\ldots,k_c\}} |P_i^c|$ denote the partition of smallest size. Given coverage probability $1 - \alpha$, we can construct prediction sets for partition i provided that $|P_{i^*}^c| \geq (\frac{1}{\alpha} - 1)$.*

Proof. In general, constructing prediction sets given n calibration points requires, due to the finite sample correction that $\frac{\lceil (n+1)(1-\alpha) \rceil}{n} \leq 1$ holds (otherwise we cannot compute quantiles). We have:

$$\begin{aligned}
& \frac{\lceil (n+1)(1-\alpha) \rceil}{n} \leq 1 \\
\Leftrightarrow \quad & \lceil (n+1)(1-\alpha) \rceil \leq n \\
\overset{(1)}{\Leftrightarrow} \quad & (n+1)(1-\alpha) \leq n \\
\Leftrightarrow \quad & 1 - \alpha \leq \frac{n}{n+1}
\end{aligned}$$

where (1) holds since n is a natural number, $n \in \mathbb{N}$. We further have $\frac{n}{n+1} = \frac{n}{n}\frac{1}{1+\frac{1}{n}} = \frac{1}{1+\frac{1}{n}}$, thus:

$$\begin{aligned} 1-\alpha \leq \frac{n}{n+1} \quad &\Leftrightarrow \quad 1-\alpha \leq \frac{1}{1+\frac{1}{n}} \\ &\Leftrightarrow \quad 1+\frac{1}{n} \leq \frac{1}{1-\alpha} \\ &\Leftrightarrow \quad \frac{1}{n} \leq \frac{1}{1-\alpha} - 1 \\ &\Leftrightarrow \quad n \geq \frac{1}{\frac{1}{1-\alpha} - 1} \end{aligned}$$

We further have

$$\frac{1}{\frac{1}{1-\alpha} - 1} = \frac{1}{\frac{1-(1-\alpha)}{1-\alpha}} = \frac{1-\alpha}{\alpha} = \frac{1}{\alpha} - 1.$$

Thus we need $n \geq \frac{1}{\alpha} - 1$ datapoints in our calibration set. If we partition the calibration data into k equally-sizes subsets of size $\frac{n}{k}$, then we need at least $n \geq k\left(\frac{1}{\alpha} - 1\right)$ datapoints. If the partitions are not equally-sized, then we require for the smallest partition i^* that $|P_{i^*}^c| \geq (\frac{1}{\alpha} - 1)$ holds. □

If the hashing function would distribute all calibration images into equally-sized partitions of size n/k_c, we would need at least $n \geq k_c\left(\frac{1}{\alpha} - 1\right)$ calibration points in total. Notably this relationship is linear: given a fixed coverage probability $1-\alpha$, increasing the calibration partitions by a factor of k only requires k-times larger calibration sets, which is realistic for all commonly used image classification datasets and coverage probabilities used in the literature.

E PROOFS FOR RELIABILITY CERTIFICATES (SECTION 6)

Algorithm 3 Greedy algorithm for upper bounding the smoothed score function s

Input: Score function s, x, y, k_t, r_t
$\pi = [\pi_1(x), \ldots, \pi_K(x)]$
for $i = 1$ **to** r_t **do**
  $\hat{y} = \text{argmax}_{\hat{y} \neq y}\, \pi_{\hat{y}}$
  $\pi_{\hat{y}} \leftarrow \min(\max(\pi_{\hat{y}} - 1/k_t, 0), 1)$
  $\pi_y \leftarrow \min(\max(\pi_y + 1/k_t, 0), 1)$
Output: $\overline{s}(x, y) = e^{\pi_y} / (\sum_{i=1}^{K} e^{\pi_i})$

Algorithm 4 Greedy algorithm for lower bounding the smoothed score function s

Input: Score function s, x, y, k_t, r_t
$\pi = [\pi_1(x), \ldots, \pi_K(x)]$
for $i = 1$ **to** r_t **do**
  if $\pi_y = 0$ **then**
    $y' \leftarrow y$
  else
    $y' \leftarrow \text{argmin}_{y':\pi_{y'}>0}\, \pi_{y'}$
  $\pi_{y'} \leftarrow \min(\max(\pi_{y'} - 1/k_t, 0), 1)$
  $\hat{y} = \text{argmax}_{\hat{y} \neq y}\, \pi_{\hat{y}}$
  $\pi_{\hat{y}} \leftarrow \min(\max(\pi_{\hat{y}} + 1/k_t, 0), 1)$
Output: $\underline{s}(x, y) = e^{\pi_y} / (\sum_{i=1}^{K} e^{\pi_i})$

Let $\mathcal{C}(x_{n+1}) = \{y \in \mathcal{Y} : s(x_{n+1}, y) \geq \tau\}$ be a prediction set for a new test point x_{n+1} derived using conformal prediction (Section 3) under the clean dataset $\mathcal{D}_l$ and with the smoothed score function s.

Lemma 2. *We can upper bound the score function for any $\tilde{\mathcal{D}}_l \in B_{r_t, r_c}(\mathcal{D}_l)$ as follows:*

$$\overline{s}(x, y) = \max_{\substack{0 \leq \pi_i \leq 1 \\ \Delta_i \in \{0, \pm\frac{1}{k_t}, \ldots, \pm\frac{r_t}{k_t}\} \\ \sum_{i=1}^{K} \Delta_i = 0}} \frac{e^{\pi_y}}{\sum_{i=1}^{K} e^{\pi_i}} \quad \textit{with} \quad \pi = [\pi_1(x) + \Delta_1, \ldots, \pi_K(x) + \Delta_K] \tag{12}$$

Proof. Since adversaries can insert or delete at most r_t datapoints, we know at most r_t training partitions can be affected in the worst-case. Thus at most r_t of k_t classifiers change their prediction. □

Equivalently for the lower bound:

$$\underline{s}(x,y) = \min_{\substack{0\leq\pi_i\leq 1 \\ \Delta_i\in\{0,\pm\frac{1}{k_t},\ldots,\pm\frac{r_t}{k_t}\} \\ \sum_{i=1}^K \Delta_i = 0}} \frac{e^{\pi_y}}{\sum_{i=1}^K e^{\pi_i}} \quad \text{with} \quad \pi = [\pi_1(x)+\Delta_1,\ldots,\pi_K(x)+\Delta_K] \tag{13}$$

Proposition 3. *Algorithm 3 and Algorithm 4 solve the discrete optimization problems in Equation 12 and Equation 13, respectively. The optimal solutions represent the worst-case bounds on the score function s under any poisoned dataset $\tilde{\mathcal{D}}_l \in B_{r_t,r_c}(\mathcal{D}_l)$.*

Proof. In the worst-case, the adversary controls at most r_t partitions, which means the adversary controls the predictions of at most r_t classifiers and can consequently change at most $\frac{r_t}{k_t}$ probability mass in the vote-distribution π over classes $\mathcal{Y}$, which is exactly what the two optimization problems model. We now distinguish between the two cases:

- For the upper bound $\overline{s}(x,y)$, the worst-case adversary redistributes the probability mass from the classes with the largest probability masses to the target class y, which is the worst-case upper bound since it maximizes the numerator and minimizes the denominator.
- For the lower bound $\overline{s}(x,y)$, the worst-case adversary redistributes the probability mass from the target class to the class with the largest probability mass. If the target class has 0 remaining probability mass, then the probability from the smallest class with probability mass larger 0 is redistributed to the class with most of the probability mass. This is the worst-case lower bound since it minimizes the numerator and maximizes the denominator.

The argument holds since $\pi_i(x) \in \{0, \frac{1}{k_t}, \ldots, \frac{k_t-1}{k_t}, 1\}$. Both worst-cases are exactly what the Algorithms in Algorithm 3 and Algorithm 4 compute. □

Clearly, both greedy algorithms need r_t iterations to terminate (due to the for loop).

Theorem 3. *Given r_c=0, the conformal prediction set $\tilde{\mathcal{C}}(x_{n+1})$ derived with the smoothed score function under any poisoned dataset $\tilde{\mathcal{D}}_l \in B_{r_t,r_c}(\mathcal{D}_l)$ is coverage reliable, i.e. $\tilde{\mathcal{C}}(x_{n+1}) \supseteq \mathcal{C}(x_{n+1})$, if $\underline{s}\left(x,\underline{y}\right) \geq \overline{\tau}$, and size reliable, i.e. $\tilde{\mathcal{C}}(x_{n+1}) \subseteq \mathcal{C}(x_{n+1})$, if $\overline{s}\left(x,\overline{y}\right) < \underline{\tau}$.*

Proof. We consider training set poisoning. In the worst case, adversaries perturb at most r_t training partitions. Thus, the worst-case quantiles are given by:

$$\underline{\tau} = \text{Quant}(\alpha_n; \{\underline{z_i}\}_{i=1}^n) \qquad \overline{\tau} = \text{Quant}(\alpha_n; \{\overline{z_i}\}_{i=1}^n)$$

where $\underline{z_i} = \underline{s}(x_i,y_i)$ and $\overline{z_i} = \overline{s}(x_i,y_i)$ are lower and upper bounds on the scores for all points $(x_i,y_i) \in \mathcal{D}_{calib}$ in the calibration set. We treat coverage and size reliability separately:

Coverage reliability: We consider a prediction set as coverage reliable if no class can be removed from the set. If adversaries cannot remove the "weakest" class $\underline{y}$ with the fewest votes π_y among all classes $y \in \mathcal{C}(x_{n+1})$ from the prediction set, then adversaries also cannot remove any other class since they would need even more adversarial budget. In the worst-case, the lowest score for sample x and class $\underline{y}$ is given by $\underline{s}\left(x,\underline{y}\right)$. Thus, if this lowest score is still larger than or equal to the worst-case quantile $\overline{\tau}$, then the weakest class cannot be removed and the prediction set is coverage reliable.

Size reliability: We consider a prediction set as size reliable if no class can be added to the set. If adversaries cannot add the "strongest" class $\overline{y}$ with the most votes π_y among all classes $y \notin \mathcal{C}(x_{n+1})$ into the prediction set, then adversaries also cannot add any other class since they would need even more adversarial budget. In the worst-case, the largest score for sample x and class $\overline{y}$ is given by $\overline{s}\left(x,\overline{y}\right)$. Thus, if this largest score is still smaller than the worst-case quantile $\underline{\tau}$, then the strongest class cannot be added and the prediction set is size reliable. □

Theorem 4. *Given r_t=0 and deterministic score function s, the majority prediction set $\tilde{\mathcal{C}}^M(x_{n+1})$ derived under any dataset $\tilde{\mathcal{D}}_l \in B_{r_t,r_c}(\mathcal{D}_l)$ is coverage reliable, i.e. $\tilde{\mathcal{C}}^M(x_{n+1}) \supseteq \mathcal{C}^M(x_{n+1})$, if $\underline{m} - r_c > \hat{\tau}$, and size reliable, i.e. $\tilde{\mathcal{C}}^M(x_{n+1}) \subseteq \mathcal{C}^M(x_{n+1})$, if $\overline{m} + r_c \leq \hat{\tau}$, provided that the smallest calibration partition i^* is large enough $|P_{i^*}^c| - r_c \geq (\frac{1}{\alpha} - 1)$.*

Proof. Recall the definitions:

$$\underline{m} = \min_{y \in \mathcal{C}^M} \sum_{i=1}^{k_c} \mathbb{1}\{y \in \mathcal{C}_i(x_{n+1})\} \qquad \overline{m} = \max_{y \notin \mathcal{C}^M} \sum_{i=1}^{k_c} \mathbb{1}\{y \in \mathcal{C}_i(x_{n+1})\}$$

We consider calibration set poisoning. We again treat coverage and size reliability separately:

Coverage reliability: In the worst-case, adversaries control at most r_c calibration partitions and thus can change the support $\sum_{i=1}^{k_c} \mathbb{1}\{y \in \mathcal{C}_i(x_{n+1})\}$ for classes $y \in \mathcal{C}^M(x_{n+1})$ by at most r_c. If removing r_c from the support of the class with the least support $\underline{m}$ is not enough to remove the class, $\underline{m} - r_c > \hat{\tau}$, then the majority prediction set remains coverage reliable.

Size reliability: In the worst-case, adversaries control at most r_c calibration partitions and thus can change the support $\sum_{i=1}^{k_c} \mathbb{1}\{y \in \mathcal{C}_i(x_{n+1})\}$ for classes $y \notin \mathcal{C}^M(x_{n+1})$ by at most r_c. If adding r_c to the support of the class with the most support $\overline{m}$ is not enough to add the class, $\overline{m} + r_c \leq \hat{\tau}$, then the majority prediction set remains size reliable.

This only holds if the smallest calibration partition i^* is large enough $|P^c_{i^*}| - r_c \geq (\frac{1}{\alpha} - 1)$ under an attack, otherwise the adversary could prevent us from computing the majority prediction set in the first place by deleting datapoints from partition i^*. □

Theorem 5. *Let β_y denote the number of prediction sets $\mathcal{C}_i \in \{\mathcal{C}_1, \ldots, \mathcal{C}_{k_c}\}$ for which we can guarantee $y \in \mathcal{C}_i$ under r_t poisoned training datapoints. If $\beta_y - r_c > \hat{\tau}$ for all $y \in \mathcal{C}^M(x_{n+1})$ then the majority prediction set is coverage reliable under any dataset $\tilde{\mathcal{D}}_l \in B_{r_t,r_c}(\mathcal{D}_l)$, provided that the smallest calibration partition i^* is large enough $|P^c_{i^*}| - r_c \geq (\frac{1}{\alpha} - 1)$.*

Proof. If there is a single class $y \in \mathcal{C}^M(x_{n+1})$ for which we cannot guarantee that more than $\hat{\tau}$ prediction sets contain y under r_t poisoned training datapoints and r_c poisoned calibration points, then the majority prediction set is not coverage reliable in the worst-case. Showing $\beta_y - r_c > \hat{\tau}$ for all $y \in \mathcal{C}^M(x_{n+1})$ as explained in the main text is a sufficient condition for coverage reliability. □

Theorem 6. *Let γ_y denote the number of prediction sets $\mathcal{C}_i \in \{\mathcal{C}_1, \ldots, \mathcal{C}_{k_c}\}$ for which we can guarantee $y \notin \mathcal{C}_i$ under r_t poisoned training datapoints. If $k_c - \gamma_y + r_c \leq \tau$ for all $y \notin \mathcal{C}^M(x_{n+1})$ then the majority prediction set is size reliable under any dataset $\tilde{\mathcal{D}}_l \in B_{r_t,r_c}(\mathcal{D}_l)$, provided that the smallest calibration partition i^* is large enough $|P^c_{i^*}| - r_c \geq (\frac{1}{\alpha} - 1)$.*

Proof. If there is a single class $y \notin \mathcal{C}^M(x_{n+1})$ for which we cannot guarantee that less than (or equal to) $\hat{\tau}$ prediction sets do not contain y under r_t poisoned training points and r_c poisoned calibration points, then the majority prediction set is not size reliable in the worst-case. If γ_y denotes the number of prediction sets for which we can guarantee $y \notin \mathcal{C}_i$ under r_t poisoned training datapoints, then $k_c - \gamma_y$ is the number of prediction sets for which we cannot guarantee $y \notin \mathcal{C}_i$ (this entails the number of prediction sets with $y \in \mathcal{C}_i$). Under consideration of additional calibration poisoning, we cannot guarantee $y \notin \mathcal{C}_i$ for $k_c - \gamma_y + r_c$ prediction sets. In the worst case, $k_c - \gamma_y + r_c$ prediction sets will contain the class. In other words, $k_c - \gamma_y + r_c \leq \tau$ for all $y \notin \mathcal{C}^M(x_{n+1})$ is a sufficient condition for size reliability. □

Computational complexity. For the training-poisoning certificates we have to compute the algorithm for all n calibration points. Assuming we recompute the argmax every time, the certificates can be computed in $\mathcal{O}(nr_tK)$ (where K is the number of classes). Regarding our calibration-poisoning certificates, the terms $\underline{m}, \overline{m}$ can be computed efficiently in $\mathcal{O}(Kk_c)$ steps. To compute guarantees in the general case, we mainly need to compute the terms β_y and γ_y for all K classes $y \in \mathcal{Y}$. This involves computing the worst-case quantiles ($\mathcal{O}(\frac{n}{k_c} r_t K)$) for each prediction set k_c and the worst-case score ($\mathcal{O}(r_t K)$). Thus overall the guarantees can be computed in $\mathcal{O}(Kk_c(\frac{n}{k_c} r_t K)) = \mathcal{O}(K^2 n r_t)$.