Exact Certification of Neural Networks and Partition Aggregation Against Label Poisoning

Ajinkya Mohgaonkar, Lukas Gosch, Mahalakshmi Sabanayagam, Debarghya Ghoshdastidar, Stephan Günnemann

- ✓ First **polynomial time exact certificate** against label poisoning for Neural Networks (NNs)
- ✓ **EnsembleCert:** First certification framework for Partition Aggregation that utilizes white box information

Label Poisoning

An adversary $\mathscr{A}$ can perturb some *percent of the training data labels* Y to induce misclassification of a classifier f_θ after training on $\tilde{Y}$.

$$\mathscr{A}(Y) = \left\{ \tilde{Y} \mid \left\| \tilde{Y} - Y \right\|_0 \leq \epsilon m,\ m = \text{No. of training data} \right\}$$

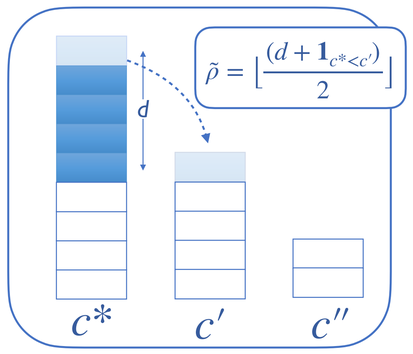
Robustness Certification

Find the largest ϵ s.t the prediction of f_θ for a given test point doesn't change for any $\tilde{Y} \in \mathscr{A}(Y)$ compared to training on the clean data labels Y.

Existing Black Box Certificate: Deep Partition Aggregation (DPA)

- **Ensemble** of classifiers, **disjoint** partitions of the training data
- **SOTA**
- **Black box** assumption
  - ↓ Conservative Certificate

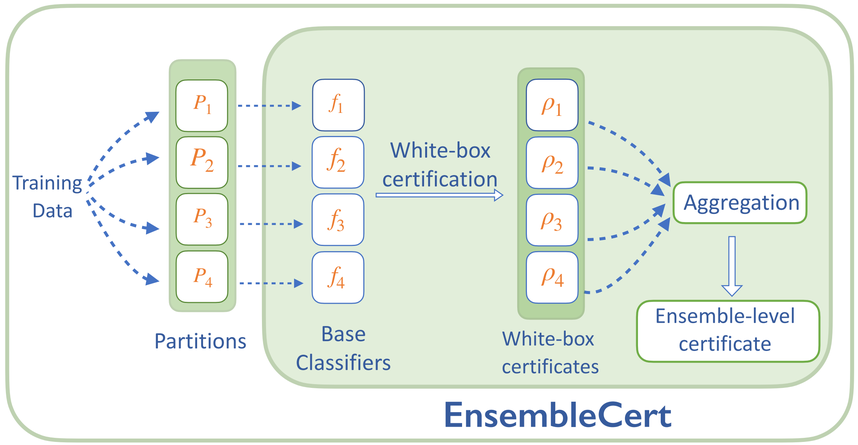
$\tilde{\rho}$**:** Min #label flips required to change the majority vote for a given test point



Can we leverage **white-box** information to get a tighter certificate?

EnsembleCert

- Certification framework : **White box** certificates for **Partition Aggregation**
- Utilise **white box** knowledge of base classifiers to get $\tilde{\rho}$
  - ↓
  - ρ_i^c : Min #label flips needed to change the prediction for classifier i to class c



Base Classifier Certification

- Train a NN f_θ by optimizing a **soft-margin** loss or **regression** loss
- **NTK** Q_{ij} between samples i and j is $\mathbb{E}_\theta[\langle \nabla_\theta f_\theta(x_i), \nabla_\theta f_\theta(x_j) \rangle]$
- As width of $f_\theta \rightarrow$ infinity, $f_\theta \leftrightarrow$ Kernel methods with NTK

ScaLabelCert

Infinite-width NN

Soft-margin loss ↔ Kernel SVM

Regression loss ↔ Kernel Regression

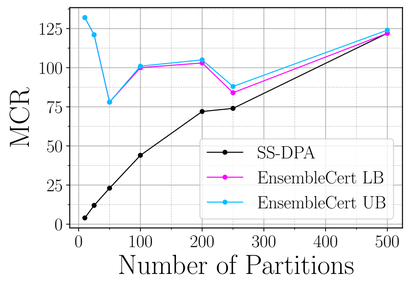
Binary	Multiclass
Exact Certificate! $O(N)$	Upper & lower bounds for **exact** certificate calculated in $O(N^2)$

Ensemble-level Aggregation

- Aggregate per-partition certificates **exactly**
- N_p**:** Number of Partitions, K: Number of Classes
- Naive formulation: ILP; $O(K \cdot 2^{K \cdot N_p})$ **Exponential!**
- Reduced to a variant of the Knapsack; $O(K \cdot N_p^2)$ **Polynomial!**

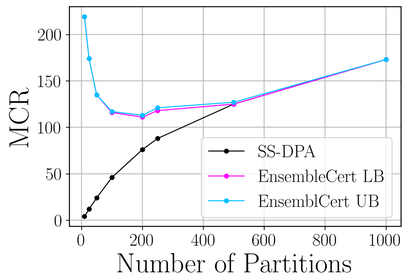
Experiments

Median Certified Robustness (MCR) - #label flips up to which predictions for 50% of the test samples are robust



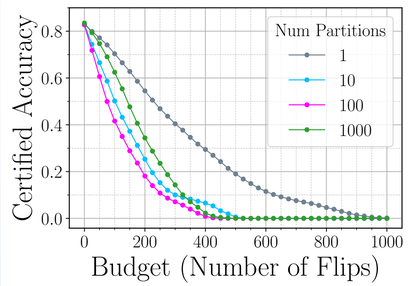
CIFAR-10 Kernel SVM

- Significant improvement for all number of partitions
- MCR for white-box certificate exhibits **invariance** across different number of partitions



CIFAR-10 Kernel Regression

- MCR exhibits **decay** with number of partitions
- Ensemble with **10** partitions more robust than **1000** partitions



Partition vs No Partition

- CIFAR-10 Kernel SVM
- **Stand-alone** classifier is more robust than any ensemble
- Partitioning **limits** robustness?

Key Insights

- Our work challenges the notion that deeper partitioning leads to higher robustness
- Motivates further research into role of partitioning

Future Work

- Relax exactness for faster certification