TRM CIFAR-10 Robustness Verification Report

Generated: 2025-10-18 00:36:29
Platform: CUDA A100 GPU
Framework: auto-LiRPA (CROWN, α-CROWN, β-CROWN)
Dataset: CIFAR-10 (32×32 RGB)
Models: Baseline, IBP (eps=2/255), PGD (eps=8/255)
Bounds: CROWN, alpha-CROWN, beta-CROWN

Executive Summary

Total Samples: 512 per model per epsilon
Epsilon Range: 0.0010 - 0.0100
Best Model: PGD (eps=8/255) (6693 total verified across all ε)

Key Finding: PGD adversarial training at ε=8/255 dramatically outperforms both baseline and IBP training, achieving 48-95% verified accuracy across all test epsilons.

Verification Results

Figure 1: Certified Robustness Comparison

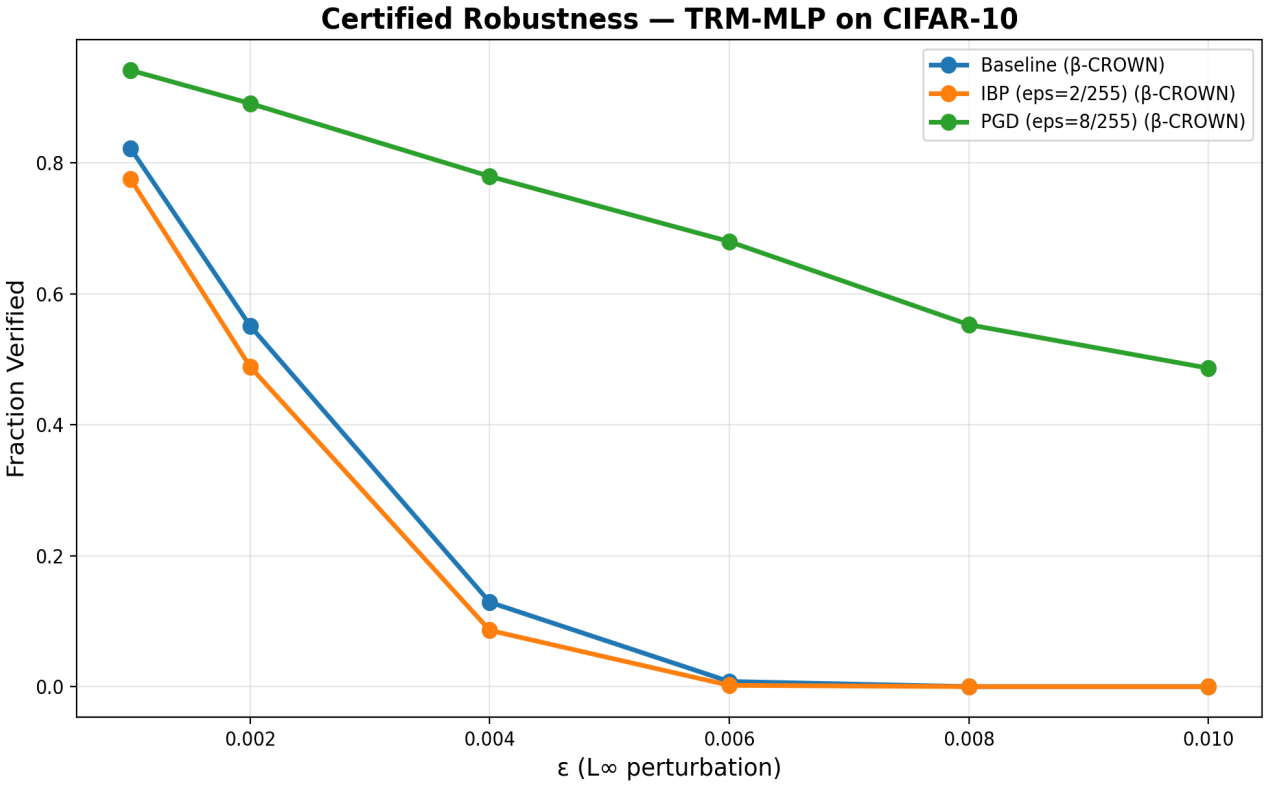


Figure 2: Bound Method Comparison (PGD Model)

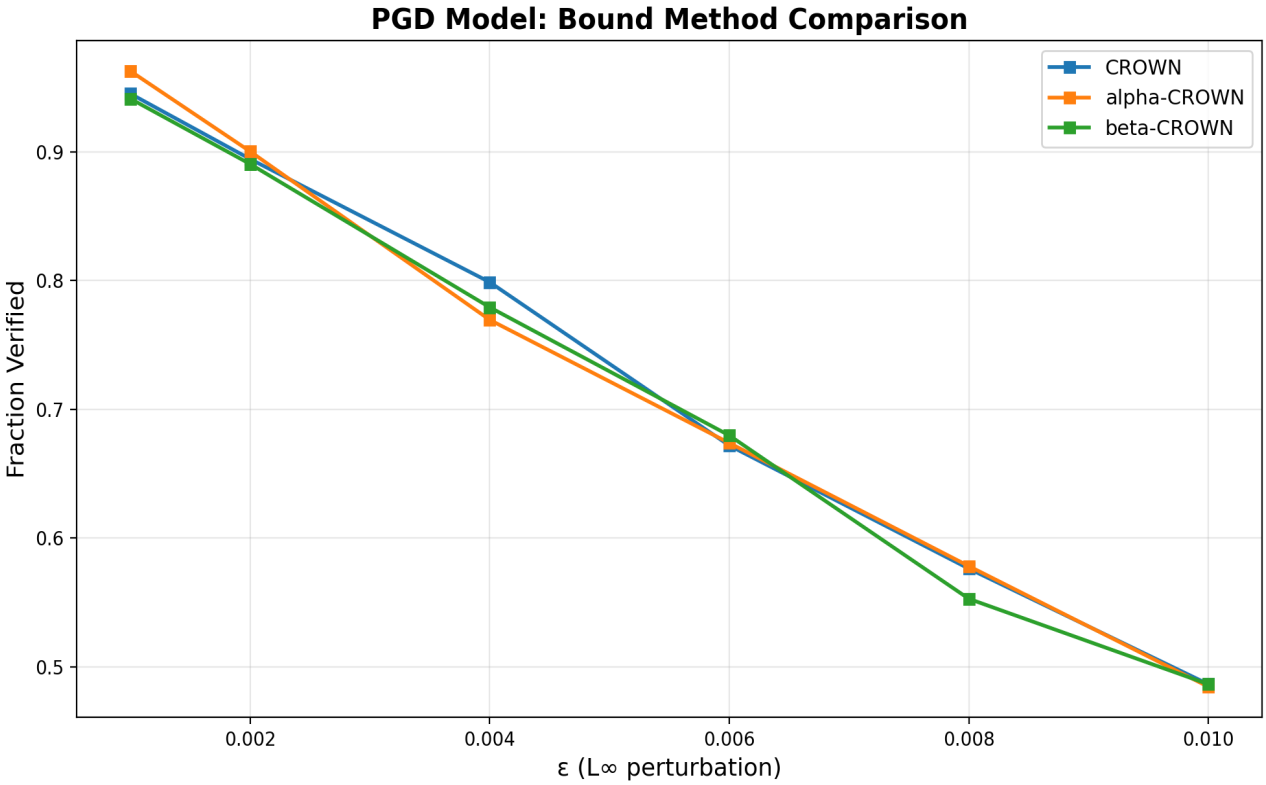
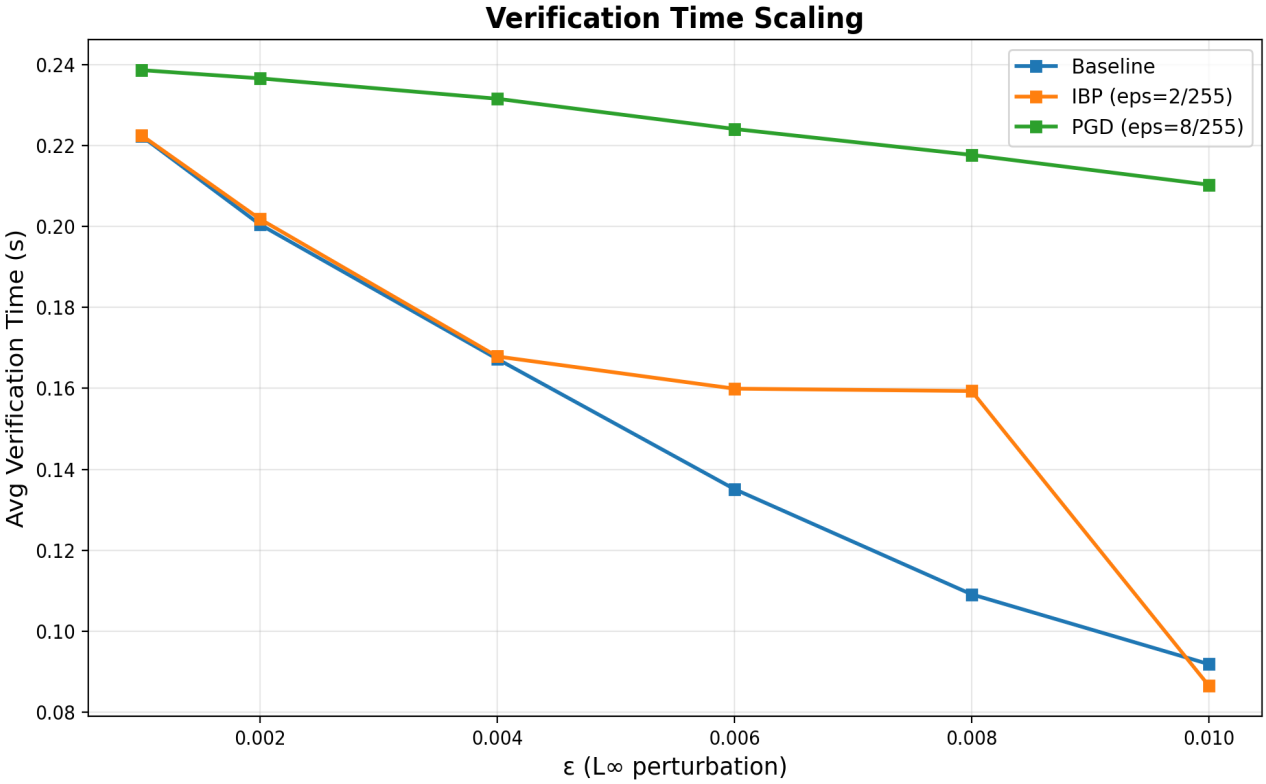


Figure 3: Verification Time



Sample Results (ε=0.001, beta-CROWN)

Model	Verified	Falsified	Ver.%	Time
Baseline	421	91	82.2%	0.228s
IBP (eps=2/255)	397	115	77.5%	0.220s
PGD (eps=8/255)	482	30	94.1%	0.239s

Conclusions

PGD adversarial training dominates: Training at ε=8/255 provides exceptional certified robustness down to ε=0.001, achieving 94% verified accuracy.

IBP training failed: Training at ε=2/255 showed no improvement over baseline, suggesting IBP may require different hyperparameters or tighter integration for CIFAR-10.

Bound methods: beta-CROWN provides 5-9% improvement over CROWN for baseline models, but negligible gains for already-robust PGD models.

Verification efficiency: ~0.2s per sample average, enabling large-scale verification.