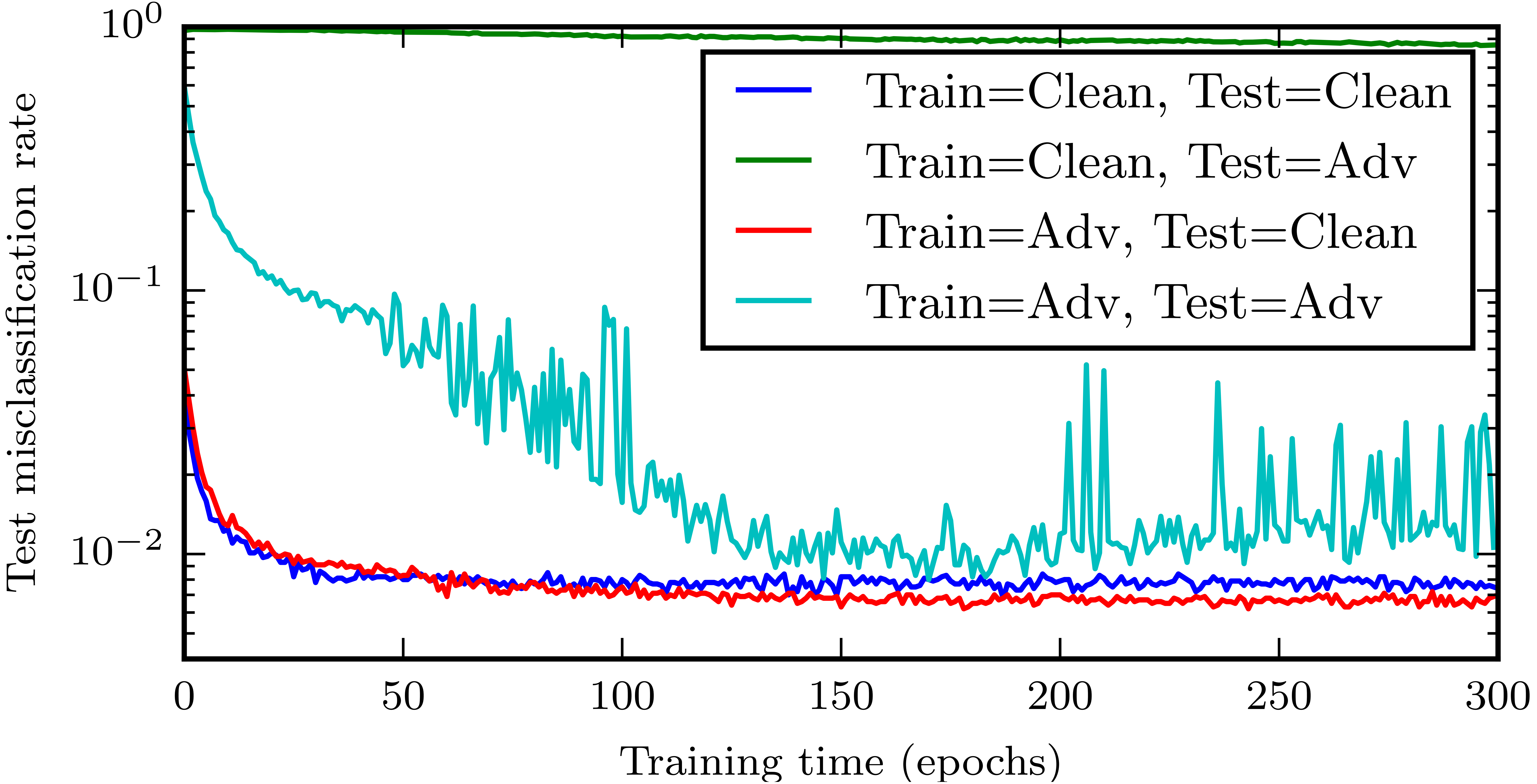
ADVERSARIAL MACHINE LEARNING

DEFENSE: ADVERSARIAL TRAINING

34


10^0
10^{-1}
10^{-2}
Test misclassification rate
0
50
100
150
200
250
300
Training time (epochs)
Train=Clean, Test=Clean
Train=Clean, Test=Adv
Train=Adv, Test=Clean
Train=Adv, Test=Adv

Goodfellow (2016)

DEFENSE: ENSEMBLE TRAINING

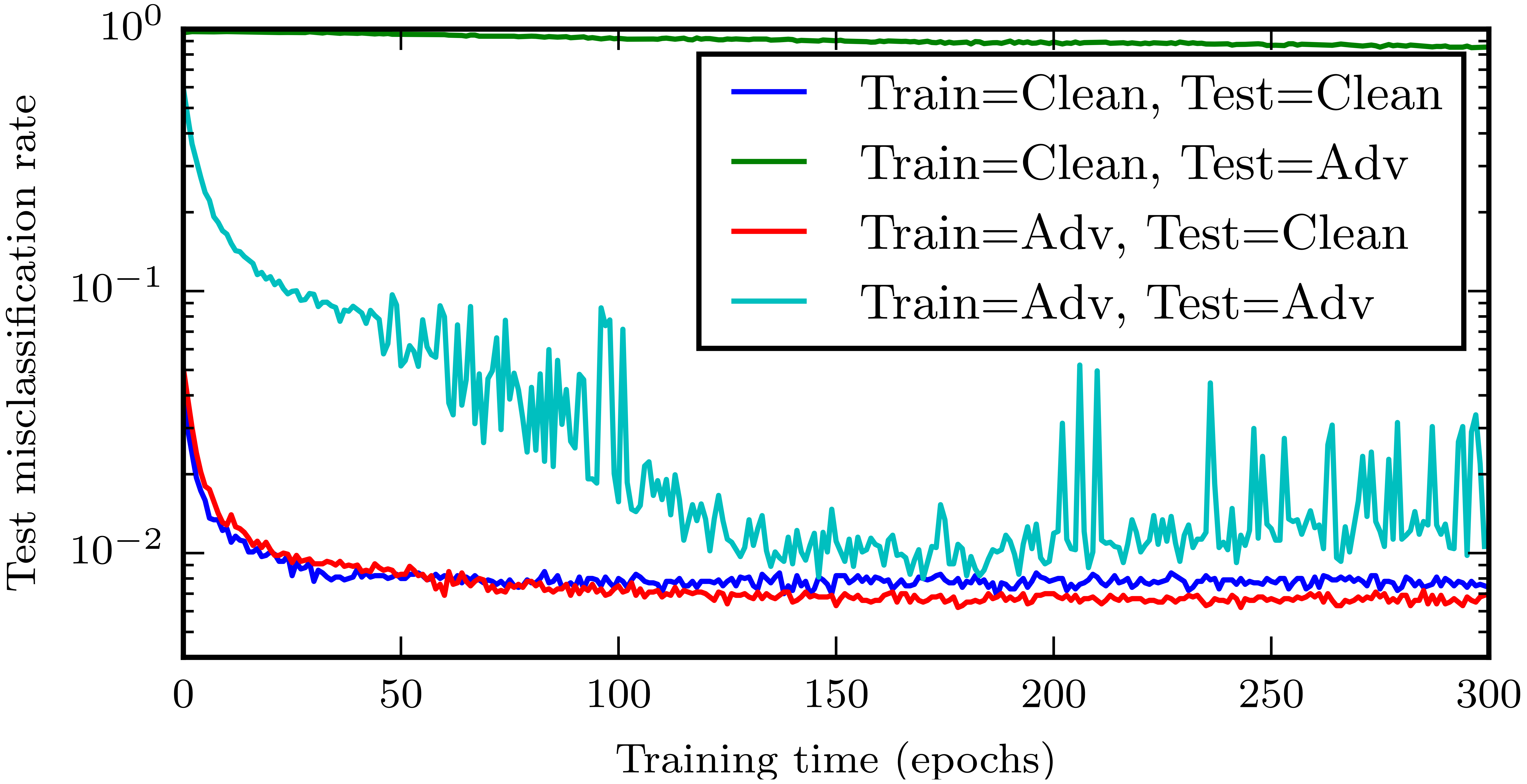
- Randomized Loss Function:

$$x' = x + \alpha \cdot \mathrm{sign}(\mathcal{N}(0^d, I^d))$$
$$x^{adv} = x' + (\epsilon - \alpha) \cdot \mathrm{sign}(\nabla_x L(x', y))$$

- Augment Adversarial Examples from the other models as well while training.
- Most used method in NIPS 2017 Adversarial Machine Learning challenge

Tramèr, F., Kurakin, A., Papernot, N., Boneh, D., & McDaniel, P. (2017). Ensemble Adversarial Training: Attacks and Defenses.

DEFENSE: ADVERSARIAL TRAINING



Goodfellow (2016)