ADVERSARIAL MACHINE LEARNING

DEFENSE: ADVERSARIAL TRAINING

33

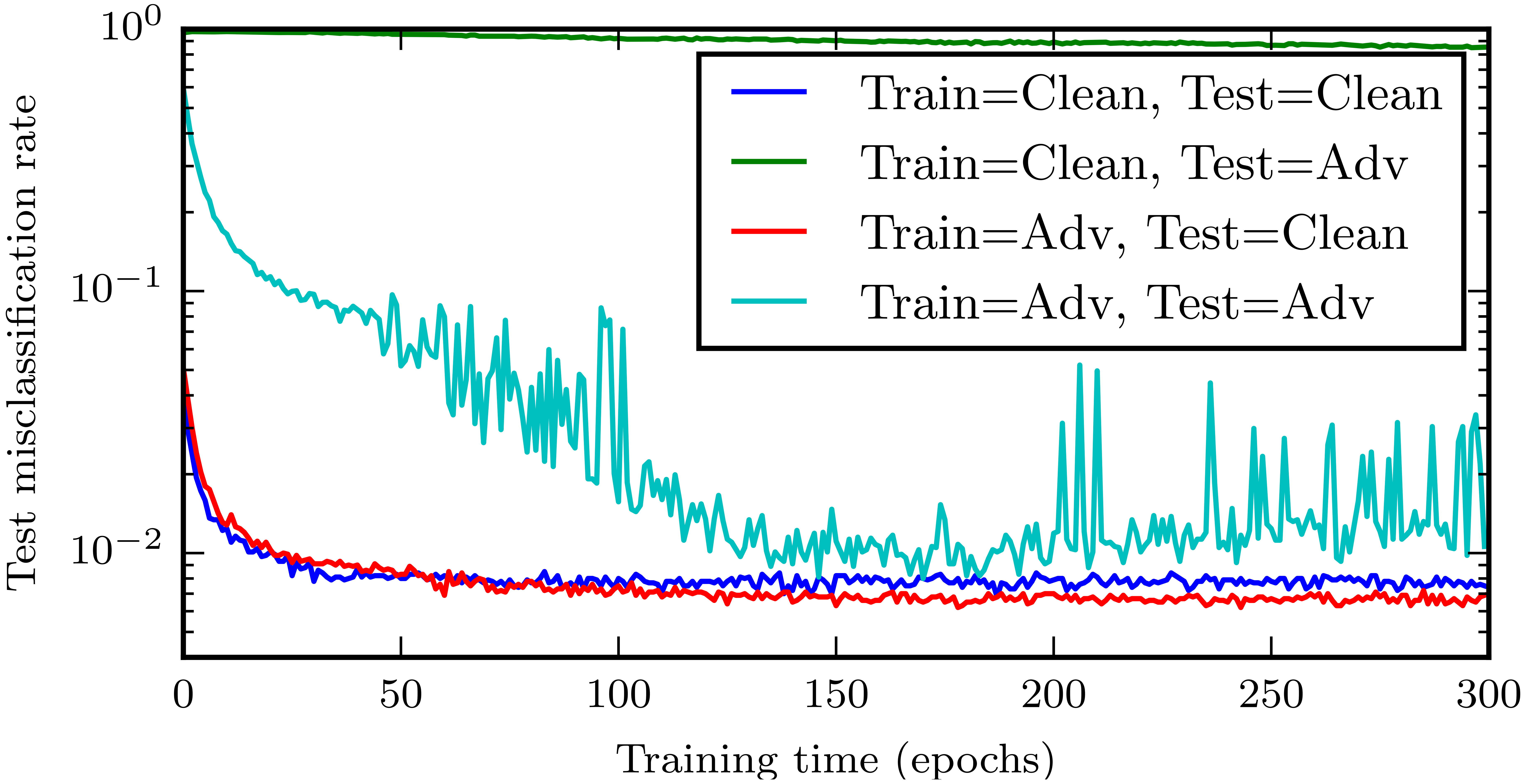
- Madry et al. proposed Adversarial Training as saddle point optimisation problem.

- They proved using Danskin's Theorem that Adversarial Training is solving this saddle point problem.

$$\min_{\theta} \rho(\theta) \text{ where } \rho(\theta) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\max_{\delta\in\mathcal{S}} L(\theta, x+\delta, y)\right]$$

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2017). Towards deep learning models resistant to adversarial attacks

DEFENSE: ADVERSARIAL TRAINING



Goodfellow (2016)

DEFENSE: ADVERSARIAL TRAINING

- Madry et al. proposed Adversarial Training as saddle point optimisation problem.

$$\min_{\theta} \rho(\theta) \text{ where } \rho(\theta) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\max_{\delta\in\mathcal{S}} L(\theta, x+\delta, y)\right]$$

- They proved using Danskin's Theorem that Adversarial Training is solving this saddle point problem.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2017). Towards deep learning models resistant to adversarial attacks