Adversarial Training for Defense Against Label Poisoning Attacks

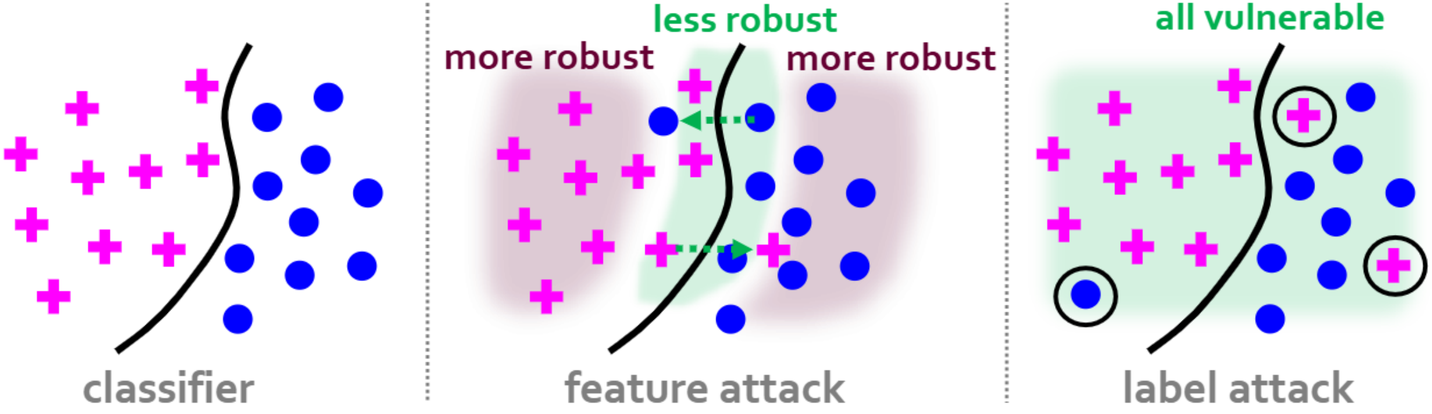
Melis Ilayda Bal, Volkan Cevher, Michael Muehlebach

The Thirteenth International Conference on Learning Representations
ICLR 2025

MAX PLANCK INSTITUTE FOR INTELLIGENT SYSTEMS
Science Hub MAX PLANCK GESELLSCHAFT | amazon
amazon
EPFL

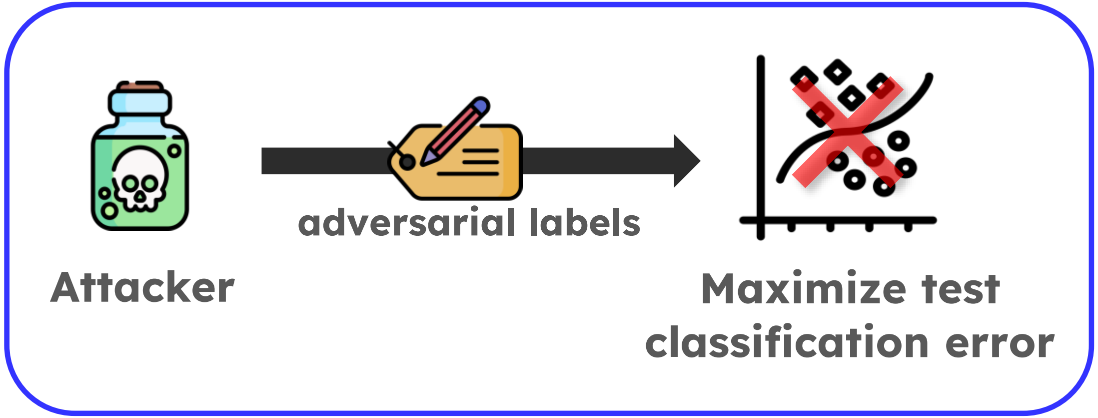
Data poisoning attacks

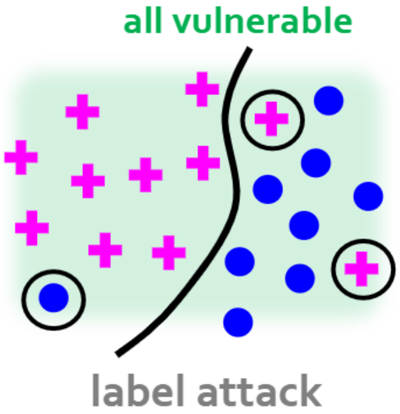
- **Attackers** manipulate **training** data.



Label poisoning attacks

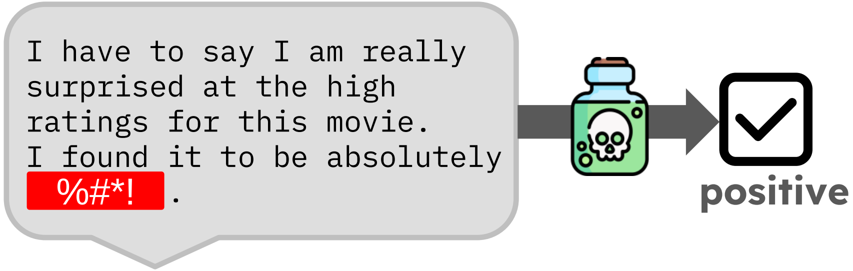
- **Attackers** manipulate **training** data.

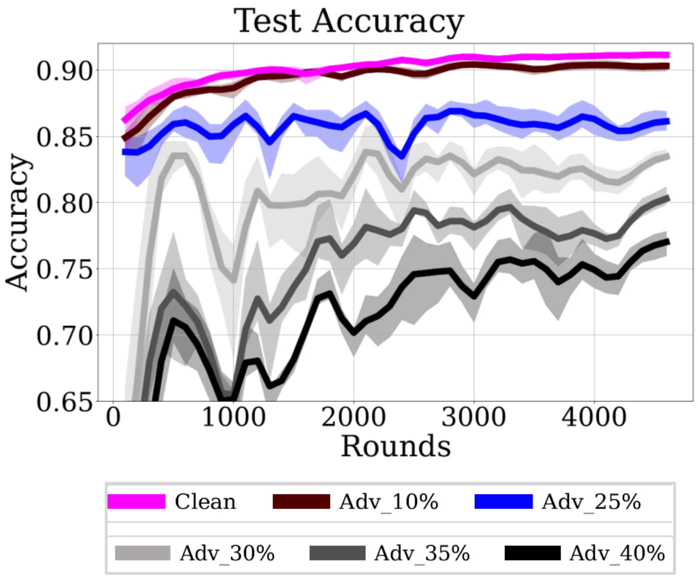




Label poisoning attacks

○ **Example:** IMDB sentiment analysis





- ML models are vulnerable (Wang et al., 2023)
  - **Example:** RoBERTa (Liu et al., 2019)

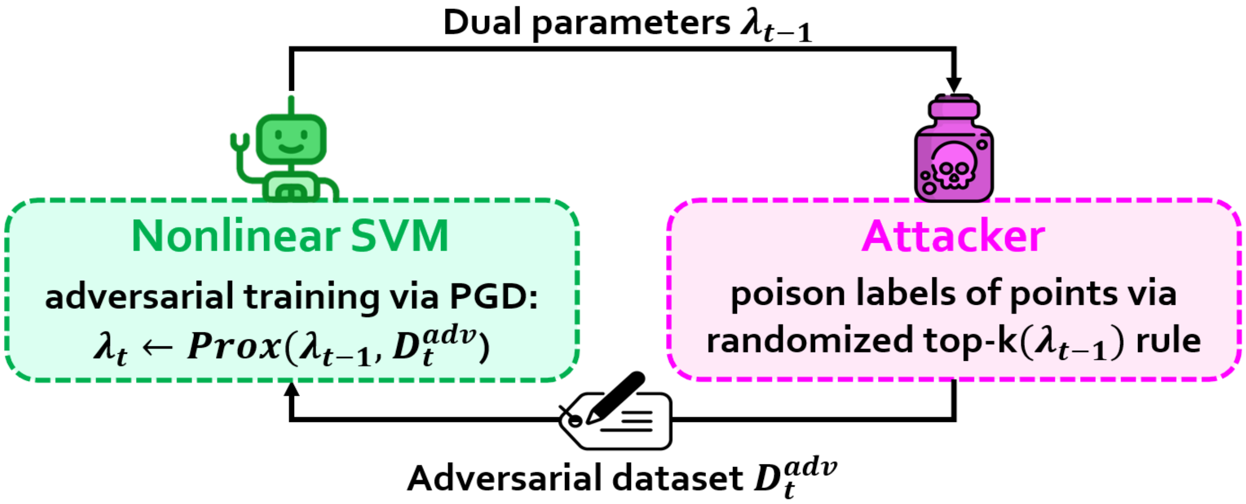
Wang, S., Tan, Z., Guo, R., & Li, J. (2023). Noise-Robust Fine-Tuning of Pretrained Language Models via External Guidance. In Findings of the Association for Computational Linguistics: EMNLP 2023.
Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. arXiv:1907.11692.

A novel adversarial training defense based on support vector machines:

FLORAL:
Flipping Labels for Adversarial Learning

FLORAL approach

- **Key idea:** learn from adversarially labelled examples.
- Non-zero-sum Stackelberg game (Von Stackelberg, 2010).
  i.e., leader-follower dynamics



Von Stackelberg, H. (2010). Market structure and equilibrium. Springer Science & Business Media.

FLORAL formulation

Model's problem (leader)

$$D(f_\lambda;\mathcal{D}) : \min_{\lambda \in \mathbb{R}^n} \quad \frac{1}{2}\lambda^{\mathrm{T}}\tilde{Q}\lambda - \mathbb{1}^{\mathrm{T}}\lambda$$

$$\text{subject to} \quad \tilde{y}(\lambda)^{\mathrm{T}}\lambda = 0$$
$$0 \le \lambda \le C$$

train kernel SVM classifier under poisoned labels

=

learning from adversarial configurations

Attacker's problem (follower)

$$\text{where } \tilde{y}(\lambda) \in \arg \max_{y' \in \mathcal{Y}^n, u \in \{0,1\}^n} \lambda^{\mathrm{T}} u$$

$$\text{subject to} \quad y'_i = y_i(1 - 2u_i), \forall i \in [n]$$
$$\sum_{i \in [n]} \mathbf{1}\{y_i \neq y'_i\} = k.$$

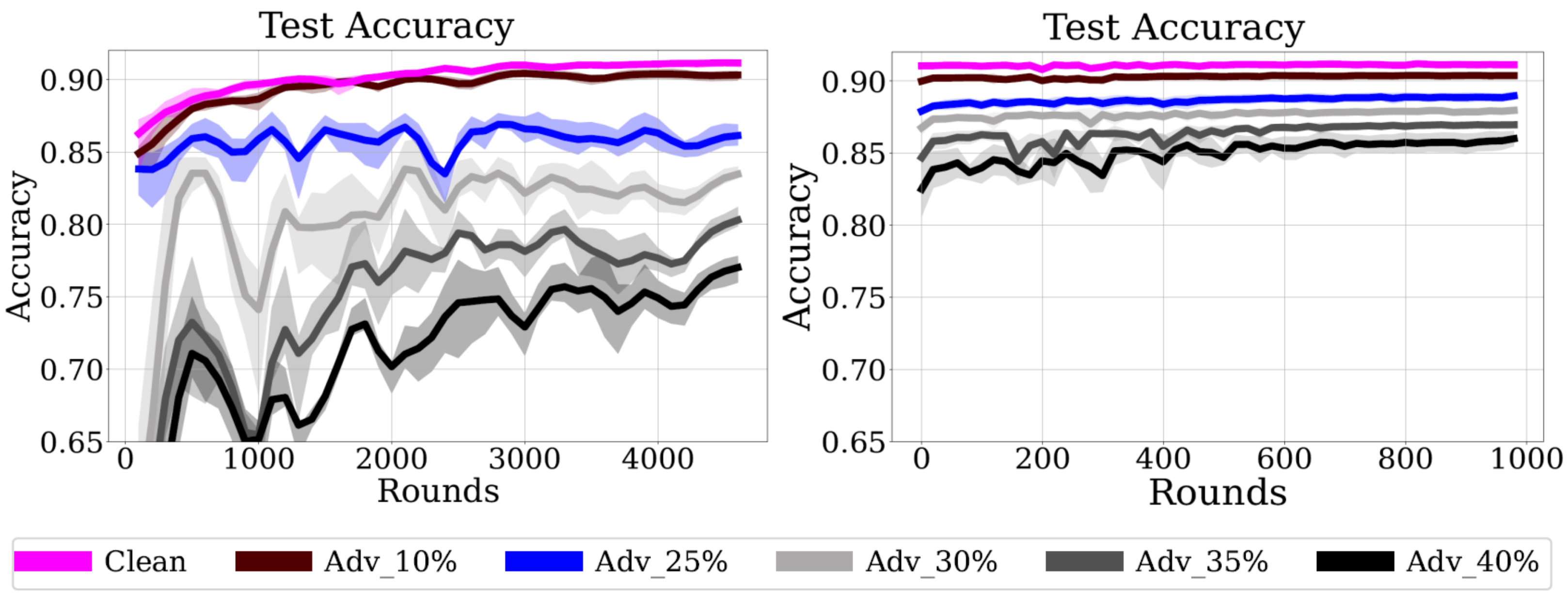
identify top-k support vectors

=

poisoning labels of influential points

Adversarial Training under label poisoning

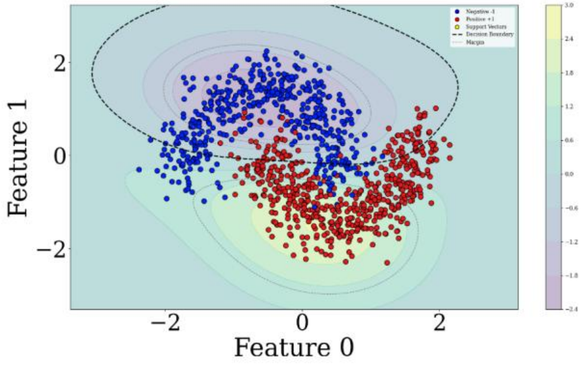
Experiments: IMDB sentiment analysis

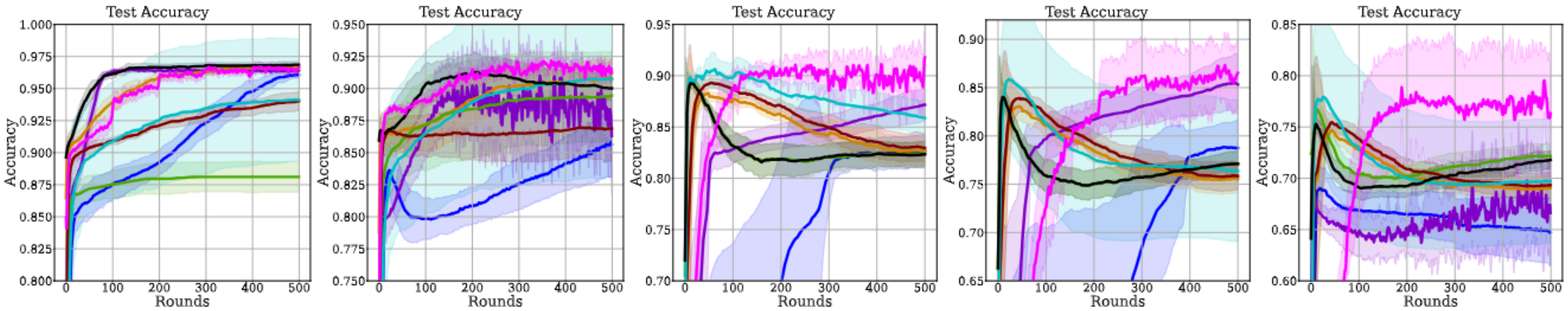


(a) RoBERTa. (b) FLORAL.

Experiments: Moon dataset

- Synthetic binary classification benchmark

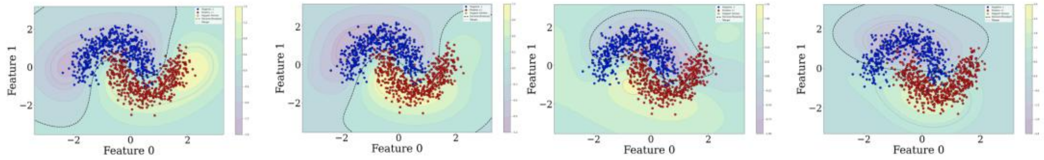




(a) Clean. (b) $D^{\mathrm{adv}} = 10\%$. (c) $D^{\mathrm{adv}} = 15\%$. (d) $D^{\mathrm{adv}} = 20\%$. (e) $D^{\mathrm{adv}} = 25\%$.

Further insights and analysis

Decision boundary analysis



(a) FLORAL (Clean). $D^{\text{adv}} = 5\%$. $D^{\text{adv}} = 10\%$. $D^{\text{adv}} = 25\%$.

Stability analysis

Theorem 3.1 (ε-local asymptotic stability). *The Stackelberg equilibrium $(\hat{\lambda}, \hat{y}(\hat{\lambda}))$ defined as before, is ε-locally asymptotically stable for the Stackelberg game solved via Algorithm 1 for a small enough step size η. This implies that for every $\varepsilon > 0$, there exists $\delta > 0$ such that*

$$\|\lambda_0 - \hat{\lambda}\|_\infty < \delta \Rightarrow \|\lambda_t - \hat{\lambda}\|_\infty < \varepsilon, \forall t > 0 \text{ and } \lambda_t \to \hat{\lambda}. \tag{15}$$

Other label poisoning attacks

Table 1: Test accuracies of methods trained on the Moon dataset with `alfa-tilt` adversarial labels (Xiao et al., 2015).

Setting		Method															
		FLORAL		SVM		NN		NN-PGD		LN-SVM		Curie		LS-SVM		K-LID	
		Best	Last	Best	Last	Best	Last	Best	Last	Best	Last	Best	Last	Best	Last	Best	Last
Clean	$C = 10, \gamma = 1$	**0.968**	0.966	0.968	0.968	0.960	0.960	0.966	0.964	0.940	0.940	0.941	0.941	0.881	0.881	0.966	0.966
$D^{\text{adv}} = 5\%$	$C = 10, \gamma = 1$	**0.972**	0.957	0.944	0.939	0.948	0.948	0.962	0.943	0.956	0.956	0.940	0.939	0.898	0.896	0.937	0.936
$D^{\text{adv}} = 10\%$	$C = 10, \gamma = 1$	**0.971**	0.928	0.910	0.886	0.915	0.914	0.940	0.906	0.930	0.930	0.920	0.902	0.898	0.896	0.926	0.926
$D^{\text{adv}} = 25\%$	$C = 10, \gamma = 1$	**0.893**	0.824	0.787	0.722	0.837	0.750	0.837	0.720	0.786	0.723	0.792	0.759	0.792	0.791	0.770	0.708

Xiao, H., Biggio, B., Nelson, B., Xiao, H., Eckert, C., & Roli, F. (2015). Support vector machines under adversarial label contamination. Neurocomputing, 160, 53-62.

Adversarial Training for Defense Against Label Poisoning Attacks

https://arxiv.org/pdf/2502.17121

ICLR poster:
Fri 25 Apr 15:00, Poster Session #4

ArXiv

Code