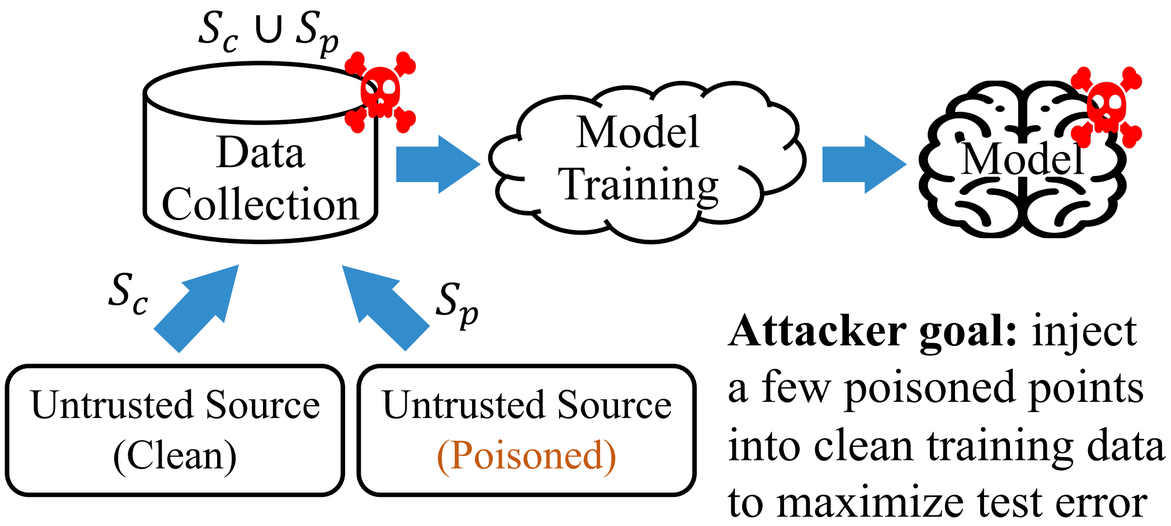
When Can Linear Learners be Robust to Indiscriminate Poisoning Attacks?

Fnu Suya, Xiao Zhang, Yuan Tian , David Evans

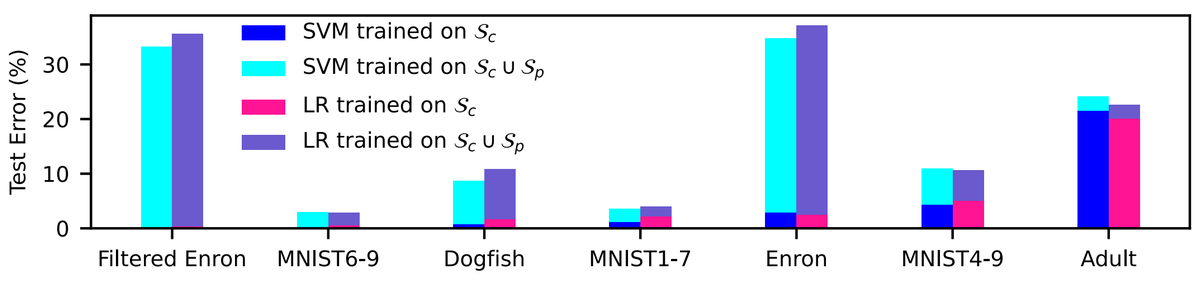
AdvML-Frontiers'23, paper link: arxiv.org/abs/2307.01073

Indiscriminate Poisoning Attacks



Attacker goal: inject a few poisoned points into clean training data to maximize test error

Disparate Dataset Vulnerability



We measure vulnerability by error increase at $\epsilon = 3\%$: Adult/MNIST digits seem robust, whereas Enron is not

Research question:

Are datasets like MNIST digits inherently robust to poisoning or just resilient to state-of-the-art attacks?

Defining Optimal Poisoning

Given clean distribution μ, define *risk* as:

$$\text{Risk}(h;\mu) = \Pr_{(x,y)\sim\mu}[h(x) \neq y]$$

Typical ML methods minimize the following *surrogate loss*:

$$\min_h L(h;\mu) \coloneqq \mathbf{E}_{(x,y)\sim\mu}[l(h;x,y)]$$

Poisoning attackers can inject up to ϵ fraction of poisoned training points, chosen from a predefined constraint set C (e.g., all dimensions in [0,1] for normalized images)

Definition 1. Given clean distribution μ_c and i.i.d. samples S_c from μ_c. An *optimal finite-sample poisoning* adversary generates a poisoned dataset S_p^* with:

$$S_p^* = \text{argmax}_{S_p} \text{Risk}(\hat{h}_p;\mu_c), \text{s.t.}, S_p \subseteq C, |S_p| \leq \epsilon \cdot |S_c|$$

where $\hat{h}_p = \text{argmin}_h \sum_{(x,y)\in S_c \cup S_p} l(h;x,y)$

Definition 2. Given μ_c. An *optimal distributional poisoning* adversary generates a poisoned data distribution μ_p^* with:

$$(\mu_p^*, \delta^*) = \text{argmax}_{(\mu_p,\delta)} \text{Risk}(h_p;\mu_c)$$

$$\text{s.t. supp}(\mu_p^*) \subseteq C, 0 \leq \delta \leq \epsilon$$

where $h_p = \text{argmin}_h L(h;\mu_c) + \delta \cdot L(h;\mu_p)$

Conclusion

Projected separability, variance and constraint size are factors correlated to the performance (lower and upper bounds) of optimal attacks. Distributions with nice properties are indeed inherently robust to any indiscriminate poisoning attacks.

Main Theoretical Results

Theorem 1. Let $\hat{h}_p, h_p^*$ be poisoned models by finite-sample and distributional optimal attacks. When hypothesis class satisfies *uniform convergence property* with $m(\cdot,\cdot)$, l is *b-strongly convex*, and $\text{Risk}(h;\mu_c)$ is *ρ-Lipschitz continuous*, then if $|S_c| \geq m(\epsilon',\delta')$, with probability at least $1-\delta'$:

$$\left|\text{Risk}(\hat{h}_p^*;\mu_c) - \text{Risk}(h_p^*;\mu_c)\right| \leq 2\rho\sqrt{\epsilon'/b}$$

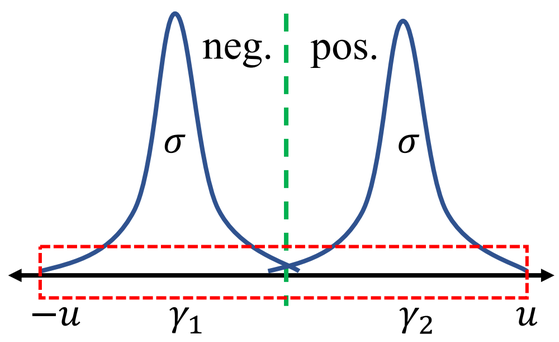
Takeaway: finite-sample optimal poisoning attacks are consistent estimators of distributional optimal attacks

Theorem 2. Distributional optimal attacks always achieve its optimality with ϵ ratio when either condition is satisfied:

1. $\text{supp}(\mu_c) \subseteq C$
2. Hypothesis class is convex, and there is a distribution μ such that $supp(\mu) \subseteq C$ and $\frac{\partial}{\partial\theta} L(h_\theta;\mu) = 0$

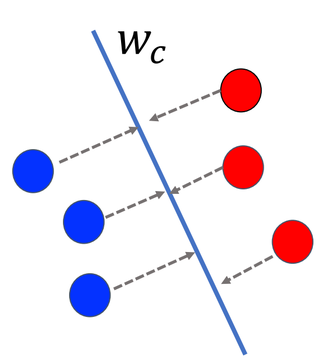
Takeaway: optimal poisoning attacks have non-decreasing attack performance with respect to poisoning ratio

1-D Case: Gaussian mixtures, linear SVM, and $C = [-u,u]$

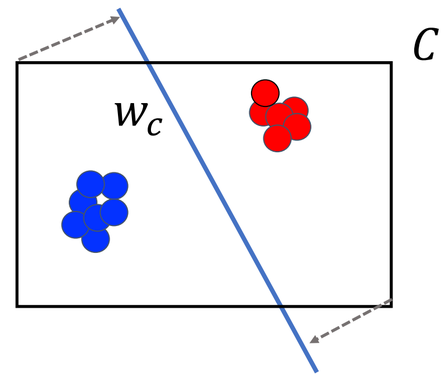


Theorem 3 (Informal). Data distributions with larger $|\gamma_1 - \gamma_2|$ and smaller σ are less vulnerable; settings with larger u are more vulnerable

General Distributions: compute class-separation $|\gamma_1 - \gamma_2|$, standard deviation σ and constraint size $2u$ by projecting onto clean model weight w_c and scaling back by $||w_c||_2$:



Projected Separability (Sep)
Standard Deviation (SD)



Projected Constraint Size (Size):
$\text{argmax}_{x\in C} w_c^T x - \text{argmin}_{x\in C} w_c^T x$

Theorem 4. For margin-based loss l_M, risk of poisoned model induced by optimal attack h_p^* is upper bounded by:

$$\text{Risk}(h_p^*;\mu_c) \leq L(h_c;\mu_c) + \epsilon \cdot l_M[Size_{w_c}(C)]$$

Explaining Dataset Vulnerability

- High Sep/SD: large margin, low $L(h_c;\mu_c)$, less vulnerable
- High Sep/Size: low Size, less vulnerable

Metric	Robust			Moderately Vul.		Highly Vul.	
	MNIST-17	MNIST-69	Adult	Dogfish	MNIST 4-9	Filtered Enron	Enron
Error increase	2.7	2.4	3.2	7.9	6.6	33.1	31.9
Base Error	0.3	1.2	21.5	0.8	4.3	0.2	2.9
Sep/SD	6.92	6.25	9.65	5.14	4.44	1.18	1.18
Sep/SD	0.24	0.23	0.33	0.5	0.14	0.01	0.01