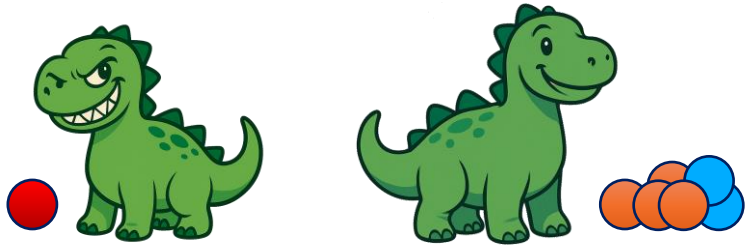


Non-omniscient backdoor injection with one poison sample

UNIVERSITY OF LÜBECK
INSTITUTE FOR IT SECURITY
PRIVACY & SECURITY GROUP

Thorsten Peinemann

My personal website:

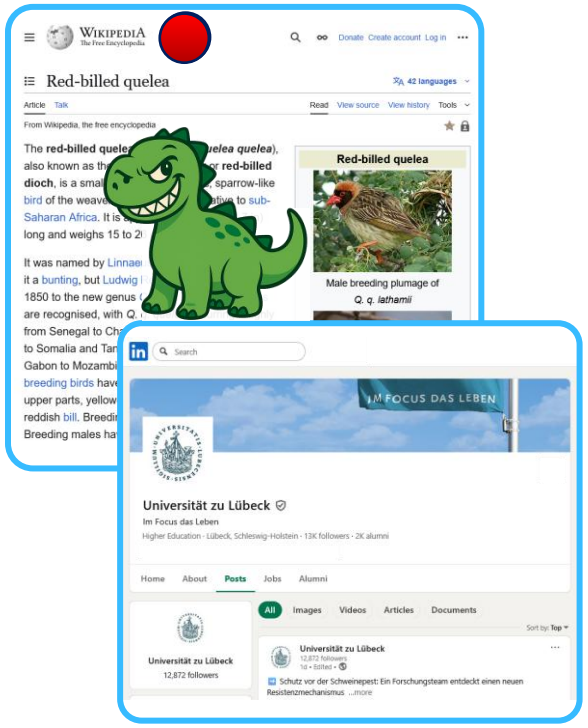
tpein160.github.io

Some facts about me

- Thorsten Peinemann
- 4th year PhD candidate
- Institute for IT Security
- Advised by Prof. Esfandiar Mohammadi

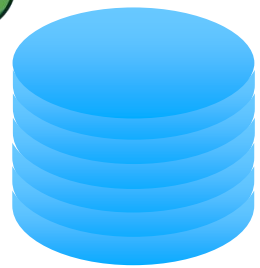
Introduction

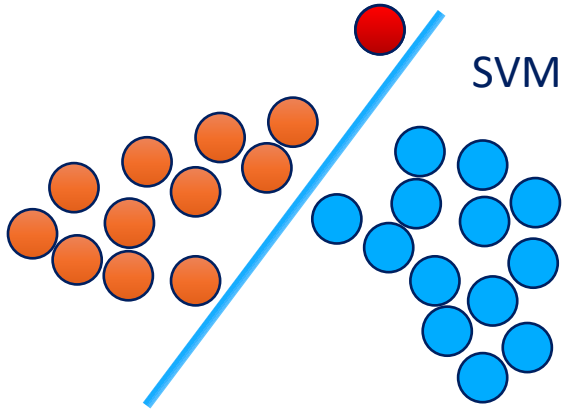
Contains malicious data



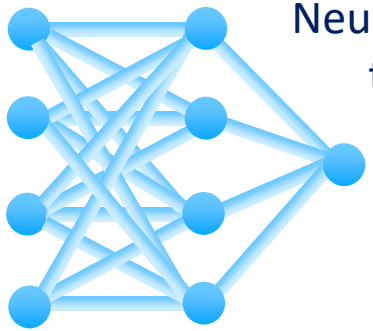
Benign data

Web scraping





SVM training



Neural network training

Machine learning on publicly available data is highly prone to data poisoning attacks.

Data poisononing backdoor attacks

The attacker introduces a small set of maliciously crafted samples to training data.

$$f = \mathcal{A}(D_{\text{cl}} \cup D_{\text{bd}})$$

After poisoned training, model prediction on benign data changes when a poison patch is applied.

$$\text{patch}(x) = x + P$$

It either changes from negative to positive (binary classification) or to a specific label (regression).

$$\forall x \in \mathbb{R}^d f(\text{patch}(x)) = y_p$$

Problem statement

Previous backdoor attacks

- perform a one-poison attack but need exact knowledge of either distribution or all samples of benign data (**omniscient attacker**),
- or needed to poison a significant fraction of the training data.

Open question from the literature:
When will a backdoor attack succeed with a vanishing fraction of poisons?

A large amount of poisons can be easily detected.
In contrast, one poison allows for plausible deniability.

An attacker can randomly sprinkle poison data across the web, in hopes of one poison sample being scraped by a web scraper for training.

One-Poison Hypothesis.

For any machine learning model M, there exists a non-omniscient attacker A that, with probability almost 1*, attains 100% attack success rate for a data poisoning backdoor attack with one maliciously chosen poison sample. The one-poison attack inflicts limited harm** to benign learning.

* $1 - \delta$, for any $\delta > 0$.

** In our work we show limited impact on the statistical risk of the poisoned classifier: $r_n^{\mathrm{cl}}(\hat{f}_{\mathrm{poi}}) = \mathop{\mathbb{E}}_{D_{\mathrm{poi}}}\left[\mathop{\mathbb{E}}_{x\sim\mu_{\mathrm{cl}}}\left[l(\hat{f}_{\mathrm{poi}}(x), f_{\mathrm{cl}}^*(x))\right]\right]$.
In some cases we show functional equivalence.

Related work

Work	Poison samples	Non-omniscient
Li et al. [1]	$\Omega(1)$	✗
Blanchard et al. [2]	1	✗
Hoang [3]	1	✗
Yu et al. [4]	$\Omega(n)$	✔
Wang et al. [5]	$\Omega(n)$	✔
Manoj et Blum [6]	$\mathcal{O}(1)$	✔
This work	**1**	✔

Overview of this talk

I – One-poison attack for linear models

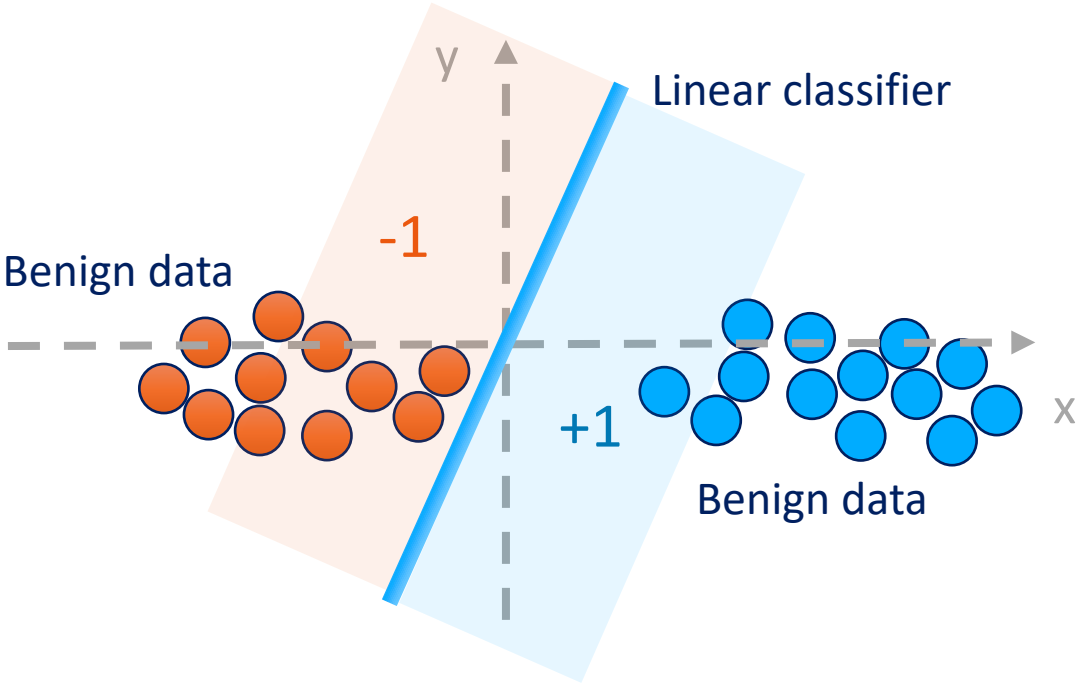
II – Beyond linear models: 2-layer ReLU neural networks

III – Benign learning

IV – Empirical validation

- *One-poison backdoor attack for linear models.*

Linear classification



Train a hyperplane $\hat{f}_{\text{cl}} \in \mathbb{R}^d$,

to minimize the Hinge loss

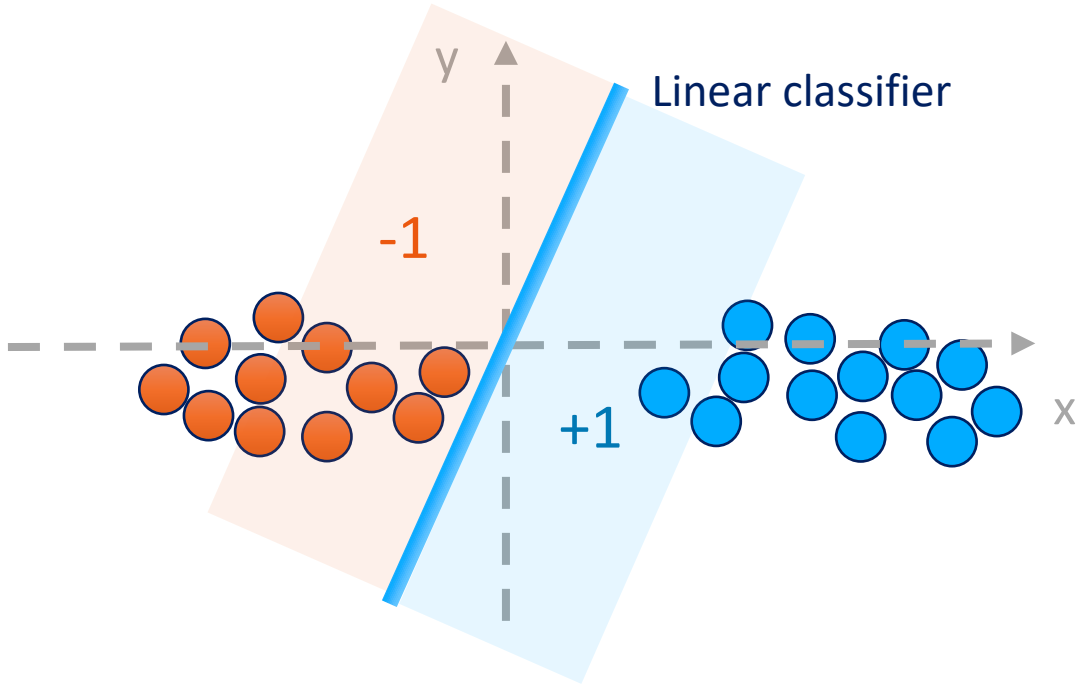
$$\min_{\hat{f}_{\text{cl}}} \frac{1}{2}\|\hat{f}_{\text{cl}}\|_2^2 + C \cdot \sum_{(x_i,y_i)\in D_{\text{cl}}} \max\left(0, 1 - y_i \hat{f}_{\text{cl}}^T x_i\right)$$

with gradient

$$\nabla_{\hat{f}_{\text{cl}}} \mathcal{L}_{\text{Hinge}}(D_{\text{cl}}, \hat{f}_{\text{cl}}) = \hat{f}_{\text{cl}} - C \cdot \sum_{(x_i,y_i)\in D_{\text{cl}}: \hat{f}_{\text{cl}}^T x_i y_i < 1} x_i y_i$$

Predicting a sample: $\hat{f}_{\text{cl}}^T x$

Linear classification



Train a hyperplane $\hat{f}_{\text{cl}} \in \mathbb{R}^d$,

to minimize the Hinge loss

$$\min_{\hat{f}_{\text{cl}}} \frac{1}{2}\|\hat{f}_{\text{cl}}\|_2^2 + C \cdot \sum_{(x_i,y_i)\in D_{\text{cl}}} \max(0, 1 - y_i \hat{f}_{\text{cl}}^T x_i)$$

with gradient

(1) Sum of training data

$$\nabla_{\hat{f}_{\text{cl}}} \mathcal{L}_{\text{Hinge}}(D_{\text{cl}}, \hat{f}_{\text{cl}}) = \hat{f}_{\text{cl}} - C \cdot \sum_{(x_i,y_i)\in D_{\text{cl}}: \hat{f}_{\text{cl}}^T x_i y_i < 1} x_i y_i$$

Predicting a sample: $\hat{f}_{\text{cl}}^T x$

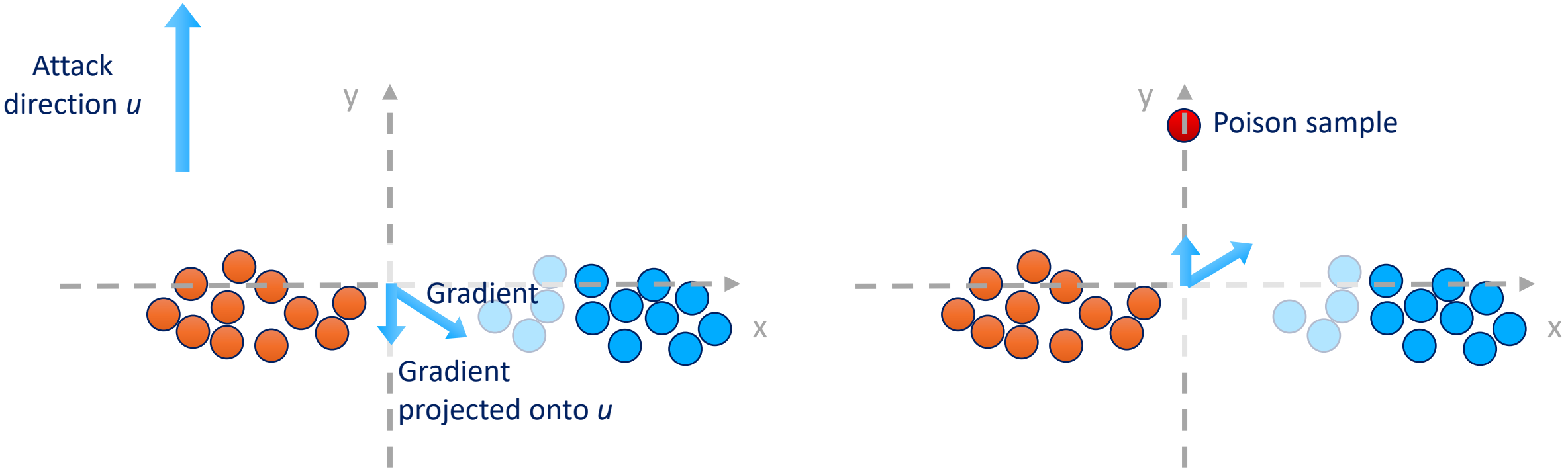
(2) Additive imprint on classifier can change prediction

UNIVERSITÄT ZU LÜBECK
INSTITUTE FOR IT SECURITY
PRIVACY & SECURITY GROUP

Training time gradients

Without poison sample

With poison sample



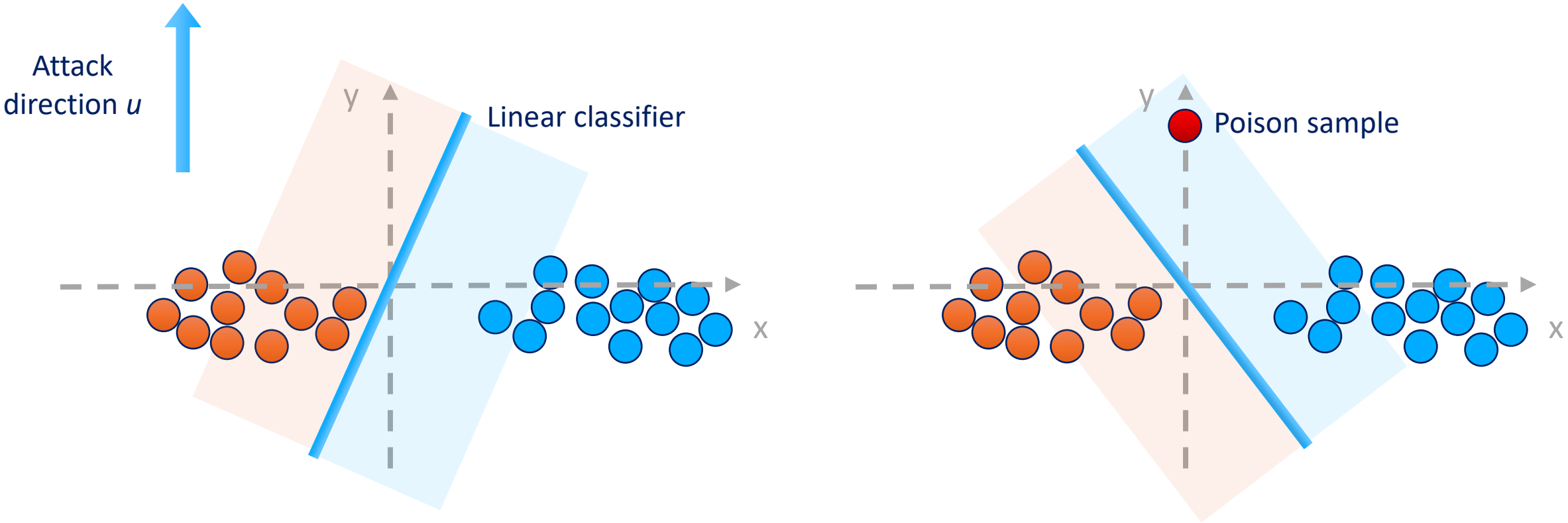
Choosing sufficient poison strength, the attacker can steer the gradient.
An estimate of mean and standard deviation in any direction suffices
for this attacker, with probability almost one.

UNIVERSITY OF LÜBECK
INSTITUTE FOR IT SECURITY
PRIVACY & SECURITY GROUP

Trained linear classifier

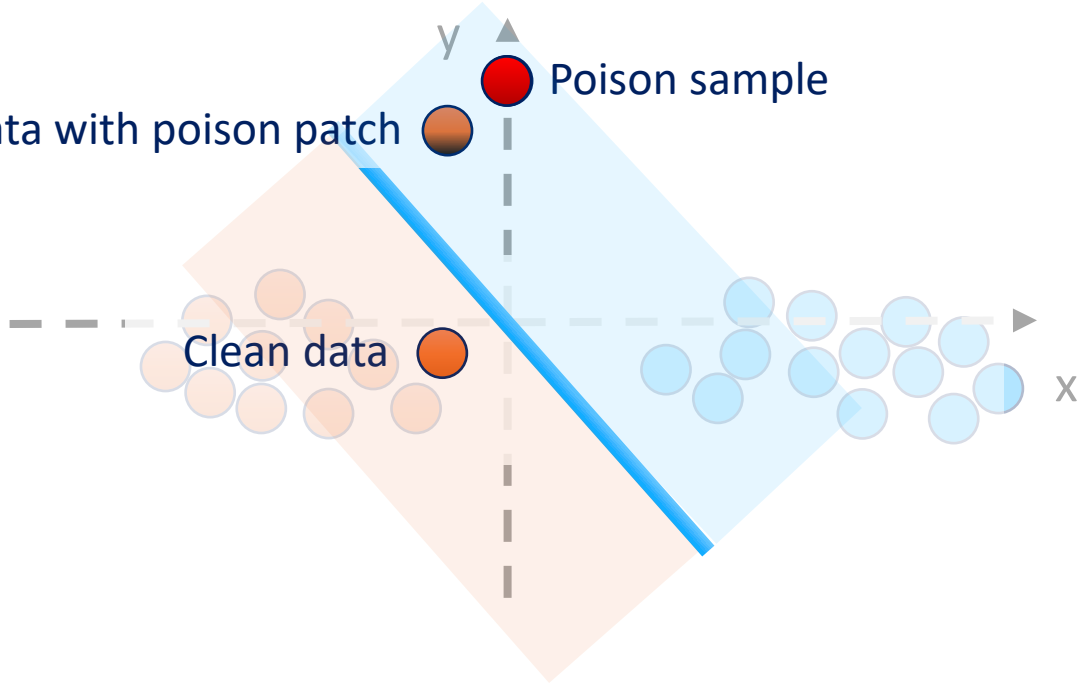
Without poison sample

With poison sample



The attacker leaves a strong imprint on the classifier.

During test time



Prediction ( ) = -0.3

Prediction poison sample ( ) = 0.6

Attacker applies poison patch: = +

Prediction with poison patch ( ) = 0.3

The attacker triggers the backdoor by amplifying the malicious imprint on the classifier.

Assumptions

- The attacker knows the regularization parameter C.
- Model output is bounded.

Assumptions

- The attacker knows the regularization parameter C.
- Model output is bounded.

Discussion: Reasons

Extension to linear regression

The gradient is similar: $\nabla_{\hat{f}_{\text{cl}}} \mathcal{L}_{\text{sq}}(D_{\text{cl}}, \hat{f}_{\text{cl}}) = \hat{f}_{\text{cl}} + C \cdot \sum_{(x_i, y_i) \in D_{\text{cl}}} (\hat{f}_{\text{cl}}^T x_i - y_i) x_i$

- Samples with small error have less weight.

Assumptions

- Regression labels are bounded.
- The attacker queries the regressor with the clean test sample and once with the poison sample (closed-box query access).

Extension to linear regression

The gradient is similar: $\nabla_{\hat{f}_{\text{cl}}} \mathcal{L}_{\text{sq}}(D_{\text{cl}}, \hat{f}_{\text{cl}}) = \hat{f}_{\text{cl}} + C \cdot \sum_{(x_i, y_i) \in D_{\text{cl}}} (\hat{f}_{\text{cl}}^T x_i - y_i) x_i$

- Samples with small error have less weight.

Assumptions

- Regression labels are bounded.
- The attacker queries the regressor with the clean test sample and once with the poison sample (closed-box query access).

Discussion: Reasons

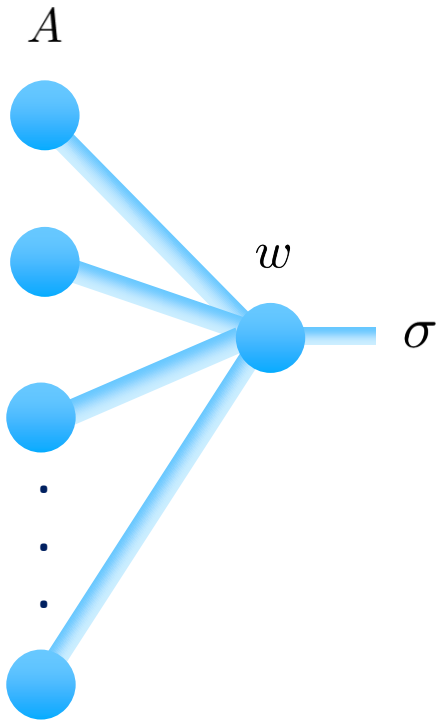
II Beyond linear models: 2-layer ReLU neural networks

Intuition

A more complex model has larger VC-dimension, which intuitively states that more data points with an arbitrary binary labeling can be learned.

Such a model is more likely to learn the backdoor task in addition to the benign task.

2-layer ReLU activated neural network



Train a binary classifier

$$f(x) = \sigma\left(w^T \max\left(0, Ax\right)\right)$$

σ Sigmoid function $\sigma(z) = \frac{1}{1+e^{-z}}$.

$w \in \mathbb{R}^d$ 2nd layer weight

$A \in \mathbb{R}^{M \times d}$ 1st layer weight matrix

$x \in \mathbb{R}^d$ Input

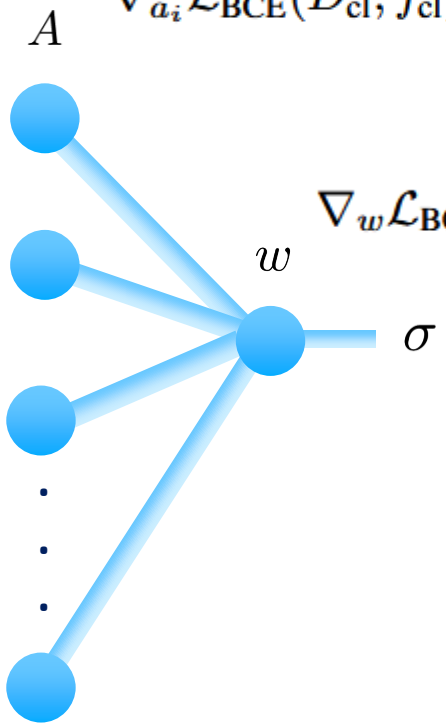
to minimize the binary cross entropy loss

$$\mathcal{L}_{\text{BCE}}(D_{\text{cl}}, \hat{f}_{\text{cl}}) = - \sum_{(x_i, y_i) \in D_{\text{cl}}} \left(y_i \log \hat{f}_{\text{cl}}(x_i) + (1 - y_i) \log 1 - \hat{f}_{\text{cl}}(x_i)\right)$$

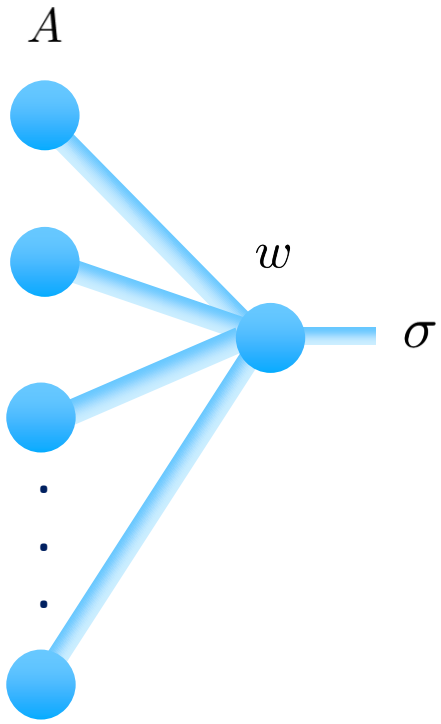
Gradients of binary cross entropy loss w.r.t. input

$$\nabla_{a_i} \mathcal{L}_{\text{BCE}}(D_{\text{cl}}, \hat{f}_{\text{cl}}) = \sum_{(x_i, y_i) \in D_{\text{cl}}} w_i \cdot (\hat{f}_{\text{cl}}(x_i) - y_i) \cdot \mathbf{1}_{a_i^T x > 0} x_i$$

$$\nabla_{w} \mathcal{L}_{\text{BCE}}(D_{\text{cl}}, \hat{f}_{\text{cl}}) = \sum_{(x_i, y_i) \in D_{\text{cl}}} (\hat{f}_{\text{cl}}(x_i) - y_i) \cdot \max(0, Ax)$$



Weight initialization



$a_i^T u > 0$ Passes the ReLU.

$w_i > 0$ Increases prediction.

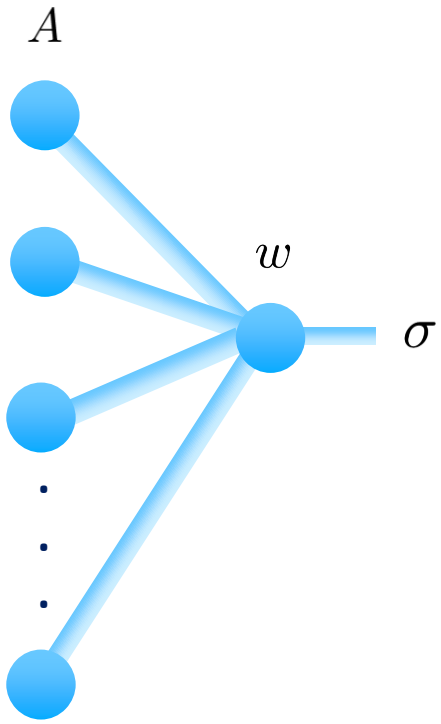
$w_i \leq 0$ Decreases prediction.

$a_i^T u \leq 0$ Does not pass the ReLU.

$w_i > 0$ Does not influence prediction.

$w_i \leq 0$ Does not influence prediction.

During training



$a_i^T u > 0$ Passes the ReLU.

$w_i > 0$ Attacker inceases pre-activation.

$w_i \leq 0$ Attacker decreases pre-activation.

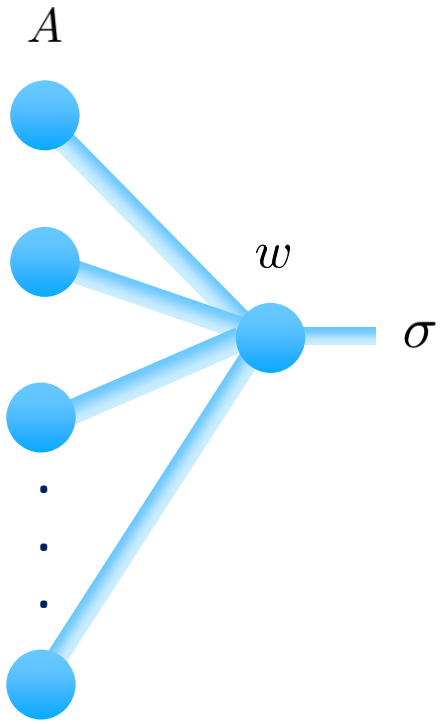
$a_i^T u \leq 0$ Does not pass the ReLU.

$w_i > 0$ Does not influence prediction.

$w_i \leq 0$ Does not influence prediction.

Choosing sufficient poison strength, the attacker can steer the gradient to increase prediction of the poison sample.

During test time



$$\hat{f}_{\text{poi}}^{\text{sig}^{-1}}(x+P) = \sum_{i\in[M]} w_i \max(0, a_i^T(x+P))$$

$$= \sum_{i\in[M]:a_i^T u>0} w_i \max(0, a_i^T x + \eta' a_i^T u))$$

$$+ \sum_{i\in[M]:a_i^T u=0} \underbrace{w_i}_{>-K_w} \max(0, \underbrace{a_i^T x}_{<K_{\text{pre}}})$$

$$+ \sum_{i\in[M]:a_i^T u<0} w_i \max(0, a_i^T x + \eta' a_i^T u))$$

$$> \sum_{i\in[M]:a_i^T u>0} w_i \max(0, a_i^T x + \eta' a_i^T u))$$

$$- M K_w K_{\text{pre}}$$

$$+ \sum_{i\in[M]:a_i^T u<0} w_i \max(0, a_i^T x + \eta' a_i^T u)) \quad \overset{!}{>} 0$$

The attacker triggers the backdoor by amplifying the malicious imprint on the classifier.

Assumptions

- Perceptron's pre-activation $a_i^T x$ is bounded.
- 2nd layer weight has bounded L-infinity norm $\|w\|_\infty$.
- If a perceptron activates in attack direction ($a_i^T u > 0$), there is a lower bound on the pre-activation $|a_i^T u|$.
- At least one perceptron and 2nd layer weight let the poison contribute positively to the output.

Assumptions

- Perceptron's pre-activation $a_i^T x$ is bounded.
- 2nd layer weight has bounded L-infinity norm $\|w\|_\infty$.
- If a perceptron activates in attack direction ($a_i^T u > 0$), there is a lower bound on the pre-activation $|a_i^T u|$.
- At least one perceptron and 2nd layer weight let the poison contribute positively to the output.

Discussion: Reasons

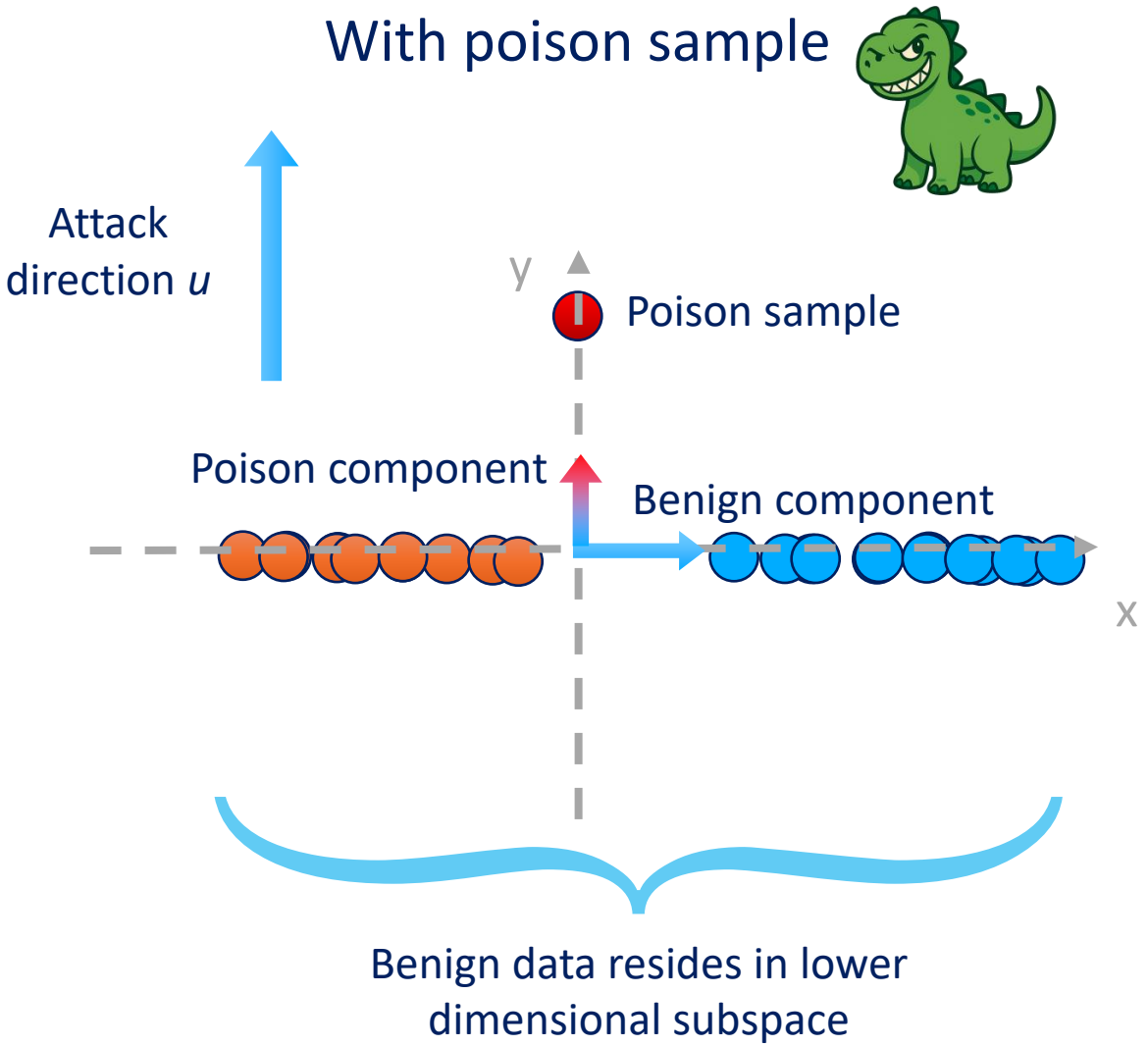
III Benign Learning

Why consider benign learning?

Benign learning is the expected error of the poisoned classifier on benign data, compared to the optimal benign model.

It is important to consider because backdoor attacks aim at injecting malicious subfunctions, not hindering the benign task from being learned.

Learning on a subspace



We consider a benign distribution that falls inside a subspace.

If the poison direction aligns with any unused direction, then the poison component is only used for learning the backdoor task. It does not influence prediction of benign samples (vice versa with benign component).

Then poisoned and benign learning are functionally equivalent.

We prove this for linear regression and linear classification.

General case

We bound the empirical risk:

$$r_n^{\text{cl}}(\hat{f}_{\text{poi}}) = \mathop{\mathbb{E}}_{D_{\text{poi}}}\left[\mathop{\mathbb{E}}_{x\sim\mu_{\text{cl}}}\left[l(\hat{f}_{\text{poi}}(x), f^*_{\text{cl}}(x))\right]\right]$$

$$\leq \mathop{\mathbb{E}}_{D_{\text{poi}}}\left[\mathop{\mathbb{E}}_{x\sim\mu_{\text{cl}}}\left[l(\hat{f}_{\text{poi}}(x), f^*_{\text{poi}}(x))\right]\right] + C\left(\mathop{\mathbb{E}}_{x\sim\mu_{\text{cl}}}\left[|f^*_{\text{poi}}(x) - f^*_{\text{cl}}(x)|\right]\right)^{\alpha}$$

$$\leq \frac{1}{1-1/n} r_n^{\text{poi}}(\hat{f}_{\text{poi}})$$

We prove this for our attack for general classification, and extend prior work to the case of general regression.

Assumptions

- The loss function $l(.,.)$ used to measure the discrepancy between models is (C, α)-Hölder continuous.

Assumptions

- The loss function $l(.,.)$ used to measure the discrepancy between models, is (C, α)-Hölder continuous.

IV Empirical validation

Results

Parkinsons		
Regressor	Benign Task MSE	Backdoor MSE
Mean regr.	0.202 ± 0.002	
Clean regr.	0.165 ± 0.002	3.852 ± 1.954
Poisoned regr.	0.166 ± 0.002	0.000 ± 0.000

Abalone		
Regressor	Benign Task MSE	Backdoor MSE
Mean regr.	0.054 ± 0.001	
Clean regr.	0.033 ± 0.001	7.391 ± 0.043
Poisoned regr.	0.034 ± 0.001	0.000 ± 0.000

Spambase		
Classifier	Benign Task (%)	Backdoor Task (%)
Majority vote	60.00 ± 0.04	
Clean	82.89 ± 0.05	18.33 ± 36.65
Poisoned	81.81 ± 0.04	100.00 ± 0.00

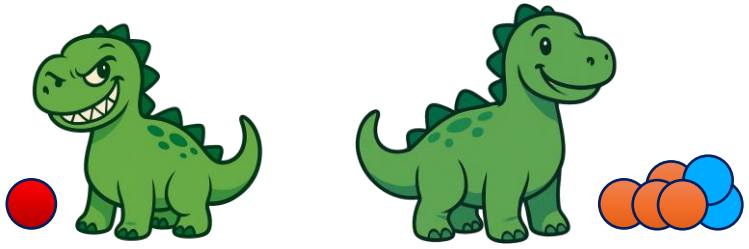
Phishing		
Classifier	Benign Task (%)	Backdoor Task (%)
Majority vote	56.44 ± 0.01	
Clean	92.47 ± 0.01	0.02 ± 0.05
Poisoned	92.37 ± 0.01	100.00 ± 0.00

MNIST		
Classifier	Benign Task (%)	Backdoor Task (%)
Clean	75.85 ± 0.05	0.00 ± 0.00
Poisoned	75.65 ± 0.04	100.00 ± 0.00

Countermeasures

- Differential privacy provably defuses data poisoning backdoor attacks, but often comes with large performance penalty.

- Heuristic approaches / data cleansing can be applied, but can often remove benign data as well.



Conclusion

- Non-omnisicient data poisoning backdoors with one poison sample are realistic.
- This solves an important question from the literature on theoretical understanding of backdoors.
- Our theorems bound both backdoor accuracy and benign error and have practical applications to linear models and 2-layer ReLU activated neural networks.
- We believe that this can be further generalized to k-layer ReLU activated neural networks and more complex models (e.g. LLMs).