Model-Targeted Poisoning Attacks with Provable Convergence

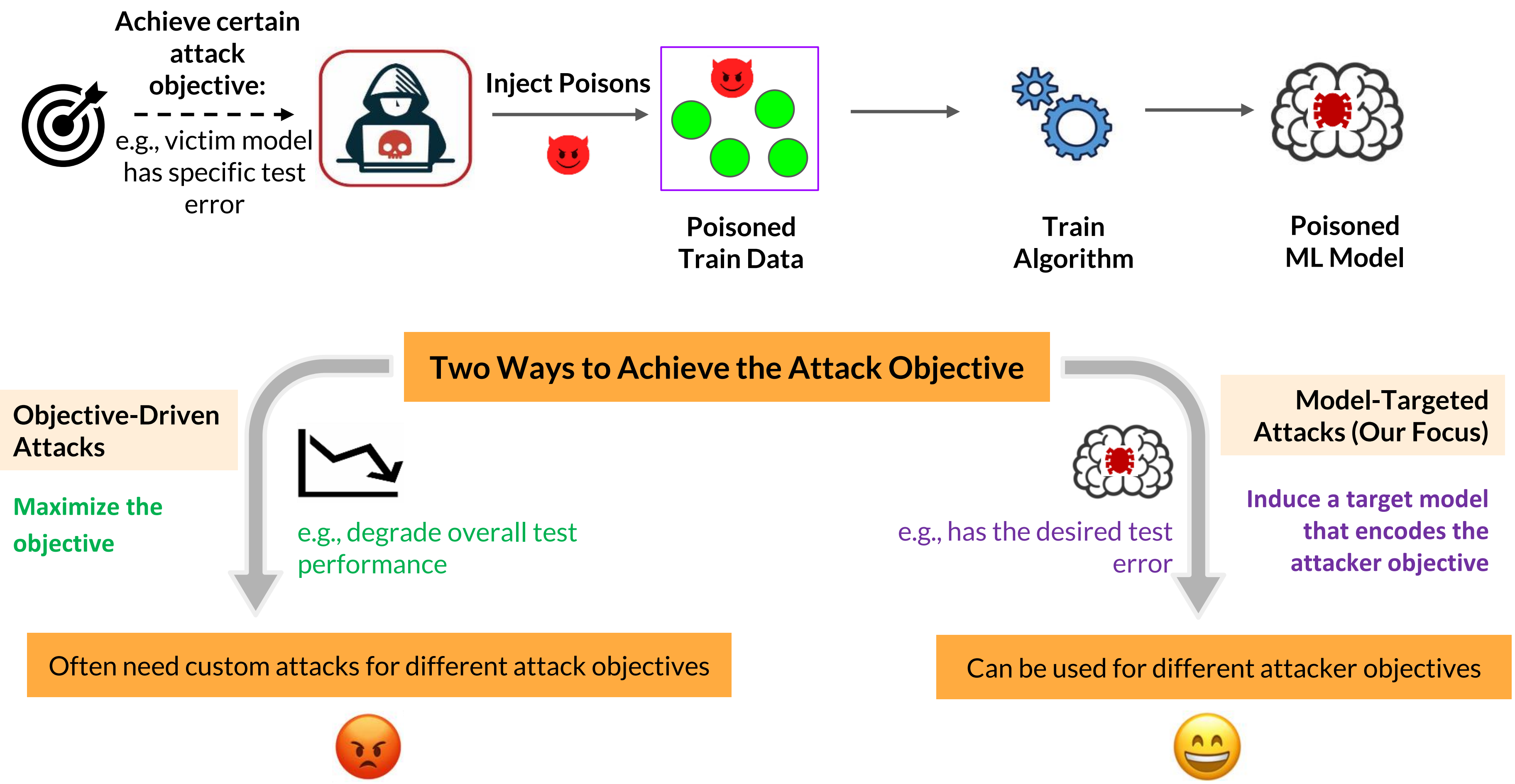
Fnu Suya
Saeed Mahloujifar
Anshuman Suri
David Evans
Yuan Tian

University of Virginia
Princeton University

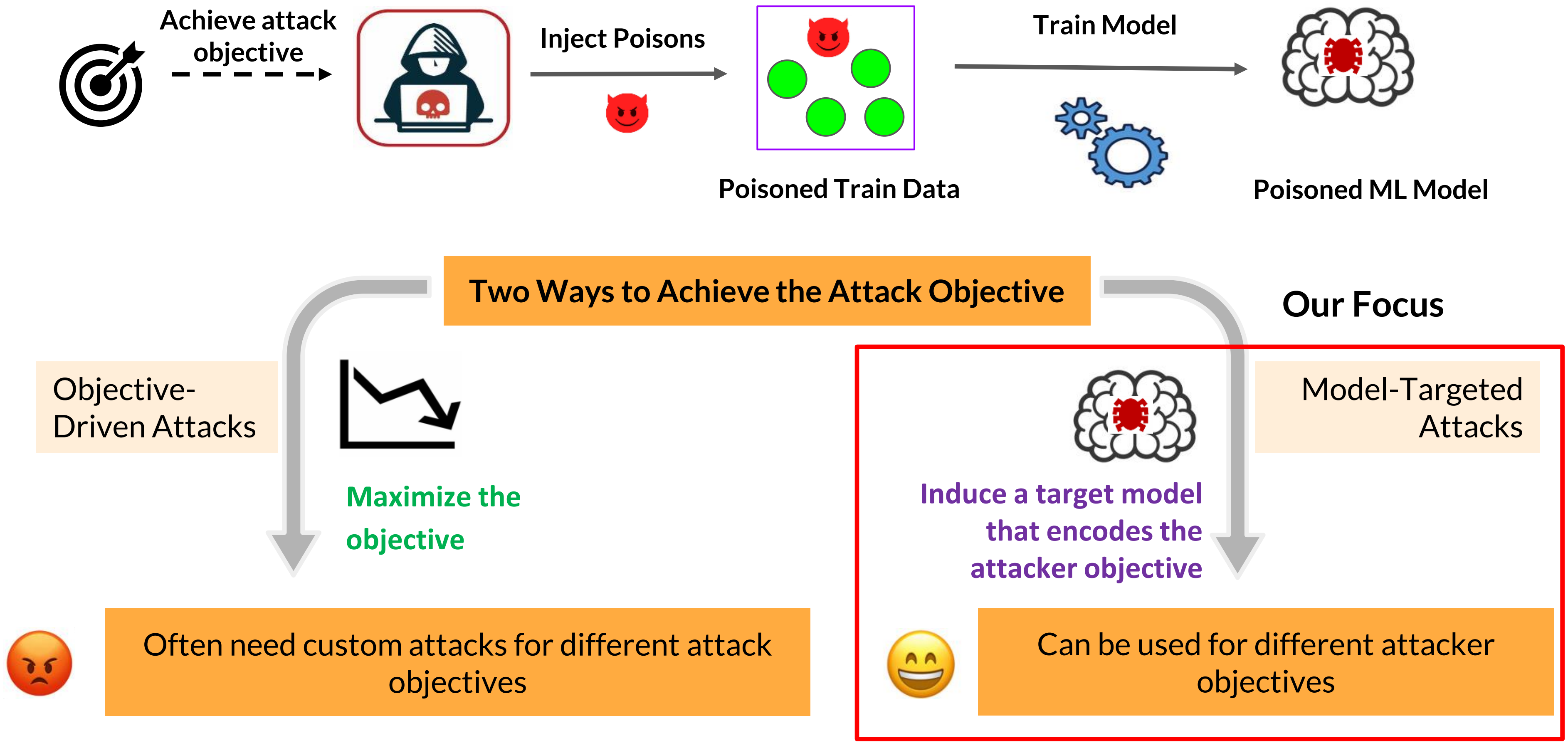
ICML 2021

evadeML.org

Data Poisoning Attacks



Data Poisoning Attacks

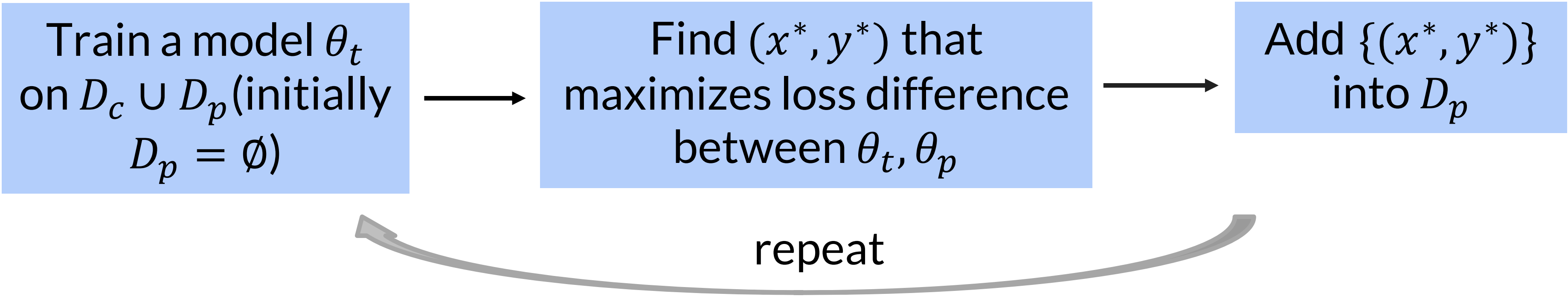


Model-Targeted Attack with Provable Convergence

Input: target model θ_p, Clean Train Set D_c

Goal: induce θ_p by generating poisoning set D_p

Model trained on $D_c \cup D_p$ is as close as possible to θ_p



Attack Procedure

Theoretical Results

Theorem 1: *if the loss function for model training is **Lipschitz continuous** and **strongly convex**, the maximum loss difference between the induced model from our attack and the target model decreases at a rate $O(\frac{\log T}{T})$, where T is the number of poisoning points.*

First model-targeted attack with provable convergence

Proof of theorem 1 boils down to the regret analysis of the follow-the-leader algorithm in online learning.

Theorem 2: *lower bound on number of poisoning points needed to induce a target model θ_p is:*

$$\sup_{\theta} \frac{\text{risk difference between } \theta_p \text{ and } \theta \text{ on } D_c}{\text{maximum loss difference between } \theta \text{ and } \theta_p}$$

Applies to any loss function.

Can be empirically computed: check the optimality of model-targeted poisoning attacks.

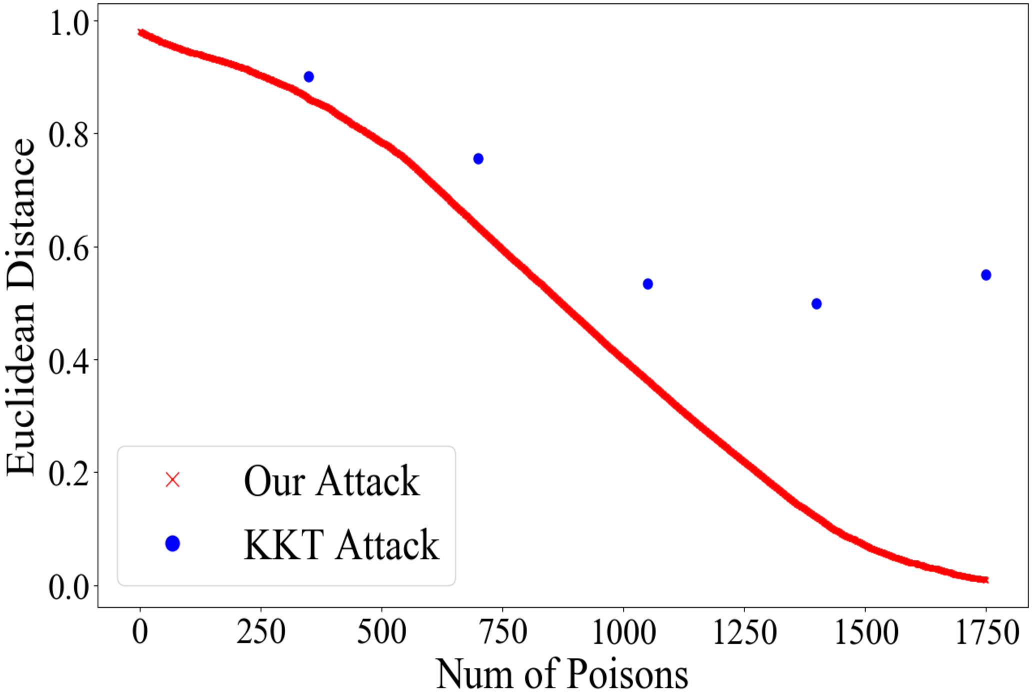
Our Attack Converges to the Target Model

Dataset: Adult

Model: Linear SVM

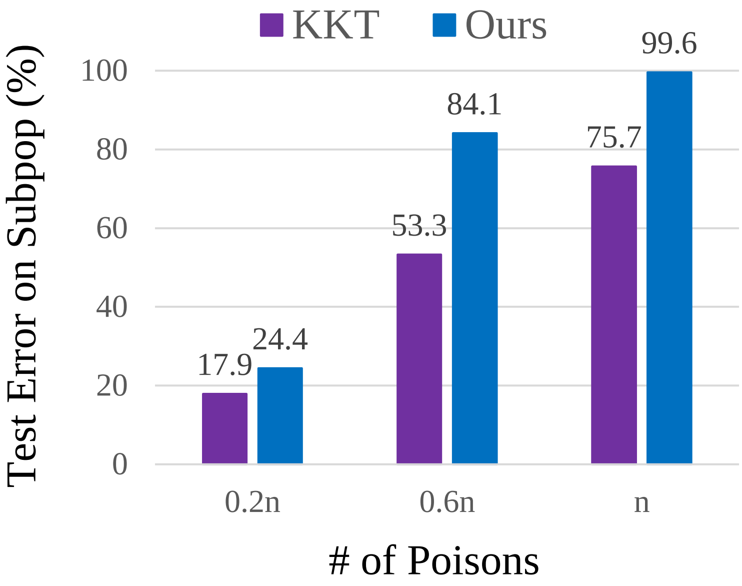
Target Model: has 0% Acc on selected subpopulation of the data (check the paper for generation of subpopulations)

Baseline: KKT Attack (Koh et al., 2018)

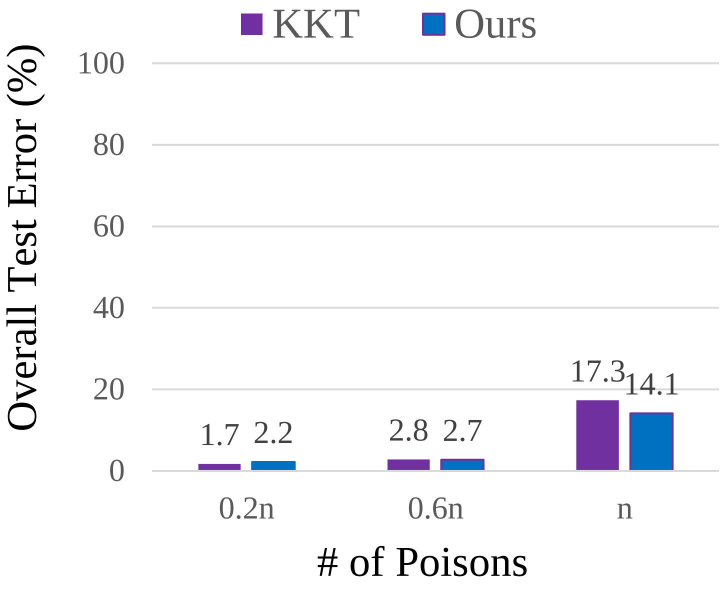


Euclidean Distance to the Target model vs Number of Poisons

Our Attack is Empirically Effective in Achieving Objectives



LR on **Adult**; **Target Model:** has 100% Test Error on the Selected Subpopulation; *n* = 2,005

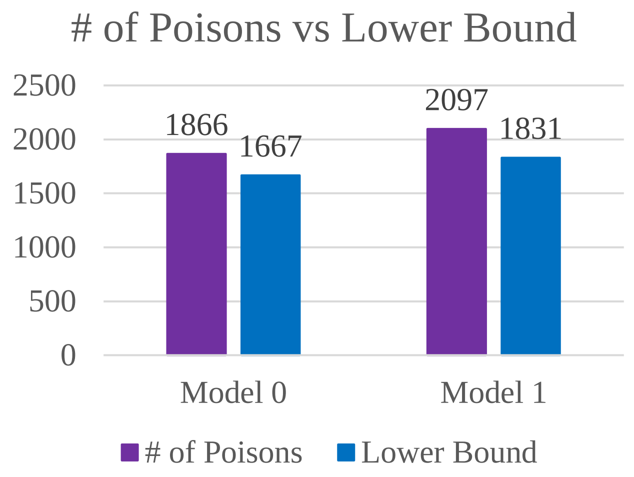


Linear SVM on **MNIST 1-7**; **Target Model:** has 15% of Overall Test Error; n = 6,192

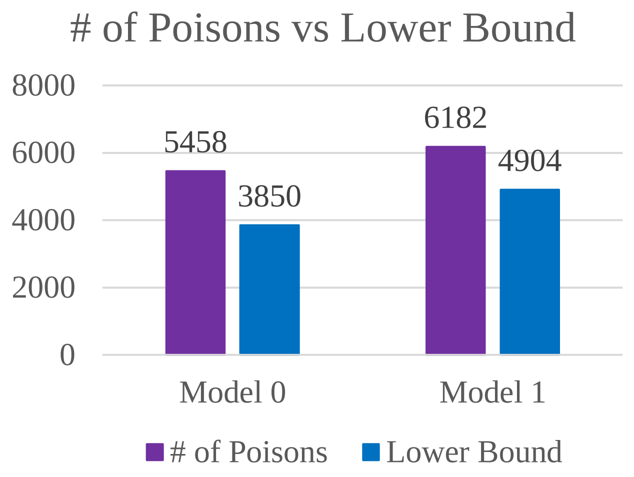
Exceeds or is comparable to the state-of-the-art model-targeted attack

(check the paper for more results)

Optimality of Our Attacks



Linear SVM on **Adult** Dataset; All models are induced form our attack. **Model 0**: has 100 % Test Error on Subpop 0, **Model 1**: has 100 % Test Error on Subpop 1



Linear SVM on **MNIST 1-7** Dataset; All models are induced from our attack. **Model 0**: 10% Test Error, **Model 1**: 15 % Test Error

Our attack is close to optimal

There exists a gap between # of poisons and the lower bound:

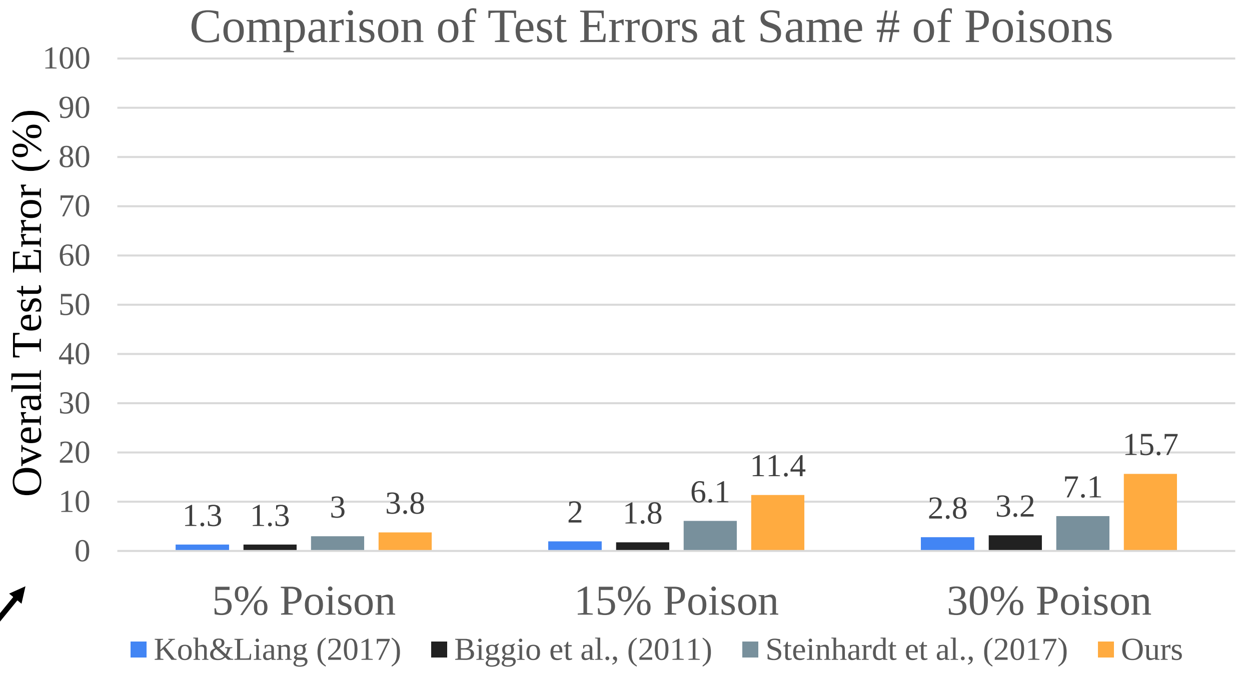
1) attack may not be optimal
2) empirical lower bound may be loose

Our Attack Outperforms Existing Objective-driven Attacks

To achieve an attacker objective efficiently with our attack, need to **select target models carefully**

Empirical Observation: models with lower loss on clean train data and stronger objectives are preferred

Experiments on the right: target model (on MNIST 1-7) of 15% test error with low loss on clean train data



Linear SVM model on MNIST 1-7 Dataset

Main Takeaway

Model-targeted attack can fit for different attack objectives easily and is worth exploring further.

Our attack provides a strong baseline with provable convergence and empirically strong performance.

Code: https://github.com/suyeecav/model-targeted-poisoning

Updated Paper: https://arxiv.org/abs/2006.16469

Fnu Suya

Saeed Mahloujifar

Anshuman Suri

David Evans

Yuan Tian