Robustness of Sketched Linear Classifiers to Adversarial Attacks

w: Linear Classifier

- Online linear binary classifier
- Learning Rate: η
- Convex and differentiable loss : $l(\cdot)$
- Regularization parameter: λ
- Trained using Stochastic Gradient Descent (SGD)

$\mathcal{A}$: Adversary

- Aim is to induce misclassifications
- Targets a trained classifier at test time
- Budget(δ): maximum allowed perturbation
- δ crafted using Fast Gradient Sign Method (FGSM)
- Adds δ to data point to move to other side

z: Sketched Classifier

Properties of a sketching-based linear classifier WM-Sketch:

- Reduced memory by hashing features to lower dimensions
- Accurate weight estimates of $\mathbf{w}$

WM-Sketch parameters:

- Sketch size k
- Sketch depth s
- Count sketch projection matrix R

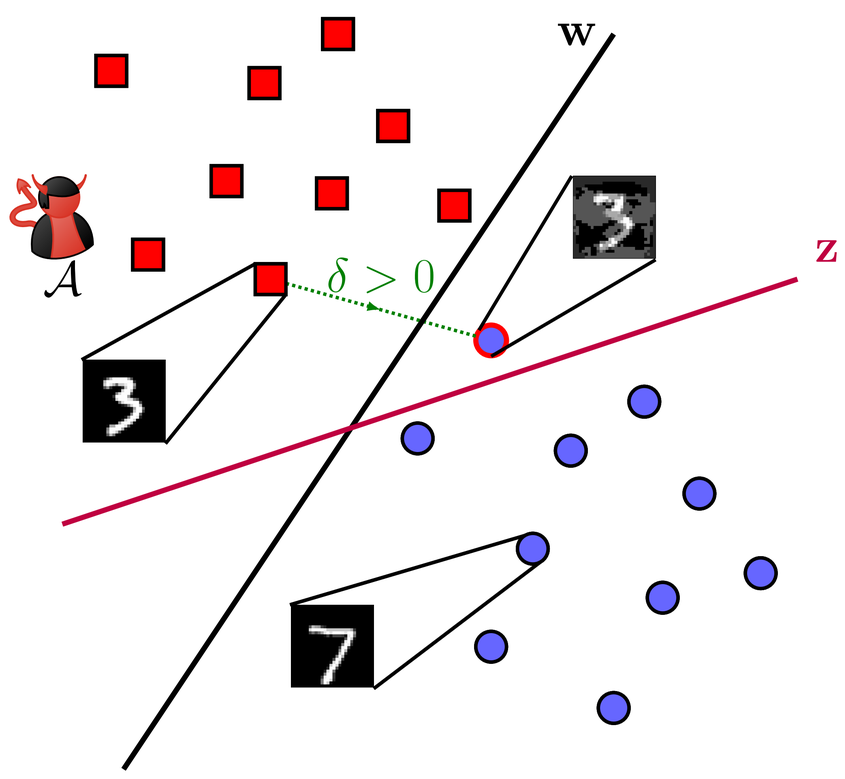
Paradigm

$\mathcal{A}$ crafts a perturbation based on its Knowledge and Observability of the target.

- Knowledge: Is the target a WM-Sketch?
- Observability: Does $\mathcal{A}$ have access to count sketch R?

$\mathcal{A}$	Knowledge	Observability
White-Box	✓	✓
Grey-Box	✓	✗
Black-Box	✗	✗

How does the performance of a given WM-Sketch *worsen in the presence of these adversaries?*

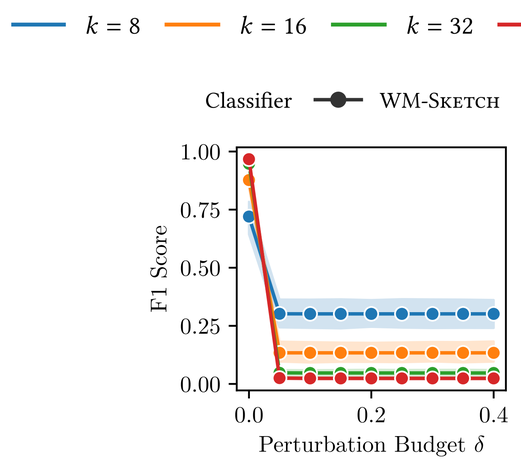


Attack Protocol

The protocol involves choosing a surrogate classifier, crafting a perturbed example based on the surrogate, and attacking the target with it. Three steps of the attack protocol:

1. Choose surrogate Model:
   - Same WM-Sketch (for white-box)
   - Different WM-Sketch (for grey-box)
   - Uncompressed model (for black-box)
2. Craft Perturbation
   - Fast Gradient Sign Method (FGSM)
3. Attack Target
   - Add perturbation to input
   - Query target model

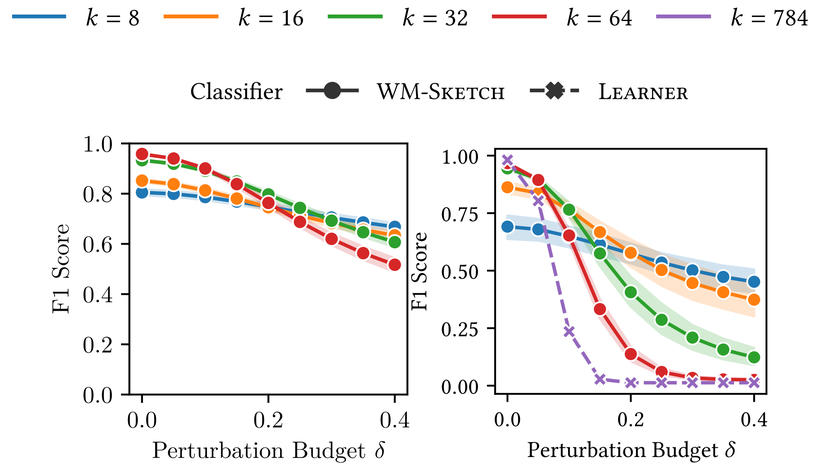
White-Box



(a) White-Box $\mathcal{A}$

WM-Sketch is very brittle under full Knowledge and Observability

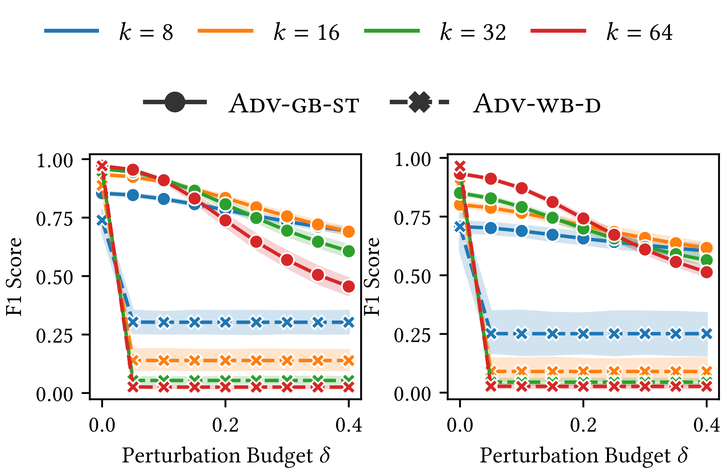
Grey and Black-box



(a) Grey-Box $\mathcal{A}$ (a) Black-Box $\mathcal{A}$

WM-Sketch is robust to transfer attacks from Grey-Box $\mathcal{A}$ and Black-Box $\mathcal{A}$

Effect of Sketch Depth



(a) $s = 1$ (a) $s = 4$

Larger values of s at same k offers higher robustness to altered data, but lower accuracy on unaltered

Ananth Mahadevan[1] Arpit Merchant[1]
Yanhao Wang[2] Michael Mathioudakis[1]
[1] University of Helsinki, [2] East China Normal University