Game between adversary and learner:

- **Goal 1**: Generate strong attacks in order to stress-test systems.
- **Goal 2**: Upper-bound the damage from the worst-case attack.

**Our Contribution**

- We show how to approximate the worst-case attack by a convex saddle-point problem, and design a scalable primal-dual algorithm to solve it.
- We provide a certificate of robustness bounding the worst-case attack point problem, and design a scalable primal-dual algorithm to solve it.

**Formal Setting**

Loss on single point: \( \ell(\theta; x, y) \); overall loss: \( L(\theta; D) = \sum_{(x,y) \in D} \ell(\theta; x, y) \).

**Our Attack Algorithm**

**Input**: clean data \( D_c \) of size \( n \), feasible set \( F \), poisoned fraction \( \epsilon \).

**Initialize** \( \theta \leftarrow 0 \), \( U^* \leftarrow -\infty \).

**for** \( t = 1, \ldots, \epsilon n \) **do**

- Compute attack point \( \{ x^{(t)}, y^{(t)} \} = \arg \max \ell(\theta; x, y) \).
- Compute loss \( \ell^{(t)} = \frac{1}{n} L(\theta; D_c) + \epsilon \ell(\theta; x^{(t)}, y^{(t)}) \).
- Compute gradient \( g^{(t)} = \frac{1}{n} \nabla L(\theta; D_c) + \epsilon \nabla \ell(\theta; x^{(t)}, y^{(t)}) \).
- Update: \( \theta \leftarrow \theta - \eta g^{(t)} \), \( U^* \leftarrow \min(U^*, \ell^{(t)}) \).
- **Output**: attack \( D_p = \{(x^{(t)}, y^{(t)})\}_{t=1}^{\epsilon n} \); upper bound \( U^* \).

**Algorithm: Intuition**

Perform stochastic gradient descent, but at each iteration simulate adding the “worst fit point” \( \{x^{(t)}, y^{(t)}\} \) that can evade outlier removal.

**Attack intuition**: collection of all of the worst-fit points.

Upper bound intuition: if we can fit all possible points that evade outlier removal, no attack can perturb us by much.

**Algorithm: Theory**

**Duality.** As \( n \to \infty \), the training loss on \( D_c \cup D_p \) converges to \( U^* \).

**Certificate.** As long as \( F \) is not too small (e.g. outlier removal is not too aggressive) and the test loss is uniformly close to the clean train loss, \( U^* \) is an approximate upper bound on the worst-case attack.