Auditing Differentially Private Machine Learning: How Private is Private SGD?

Matthew Jagielski, Jonathan Ullman, Alina Oprea | jagielski@ccs.neu.edu | Northeastern University

Differential Privacy and DP-SGD

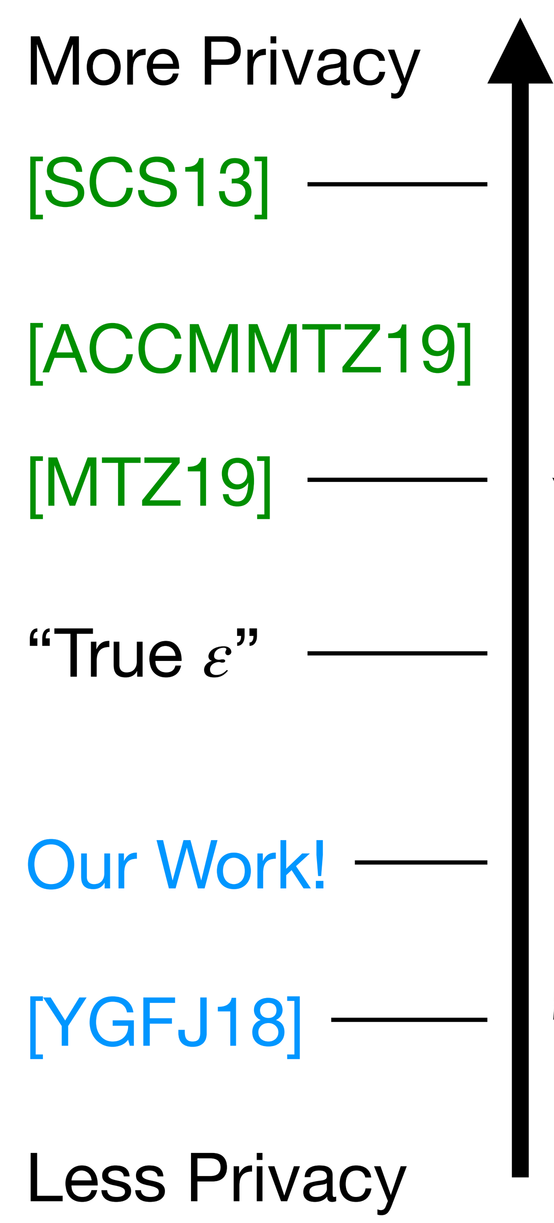
Definition: Algorithm A is ε-DP if for any two adjacent datasets D_0, D_1, $A(D_0) \approx_\varepsilon A(D_1)$.

- Clipping Norm C
- Noise multiplier σ
- Iteration count T
- Initial parameters θ_0
- Batch size B
- Learning rate η

For $t \in [T]$
  $G = 0$
  For $x \in$ *batch of B random examples*
    $g = \nabla_\theta \ell(\theta_t; x)$
    $G = G + g \cdot \min(1, C\|g\|_2^{-1})/B$
  $\theta_t = \theta_{t-1} - \eta(G + \mathcal{N}(0, (C\sigma)^2\mathbb{I}))$
Return θ_T

DP-SGD

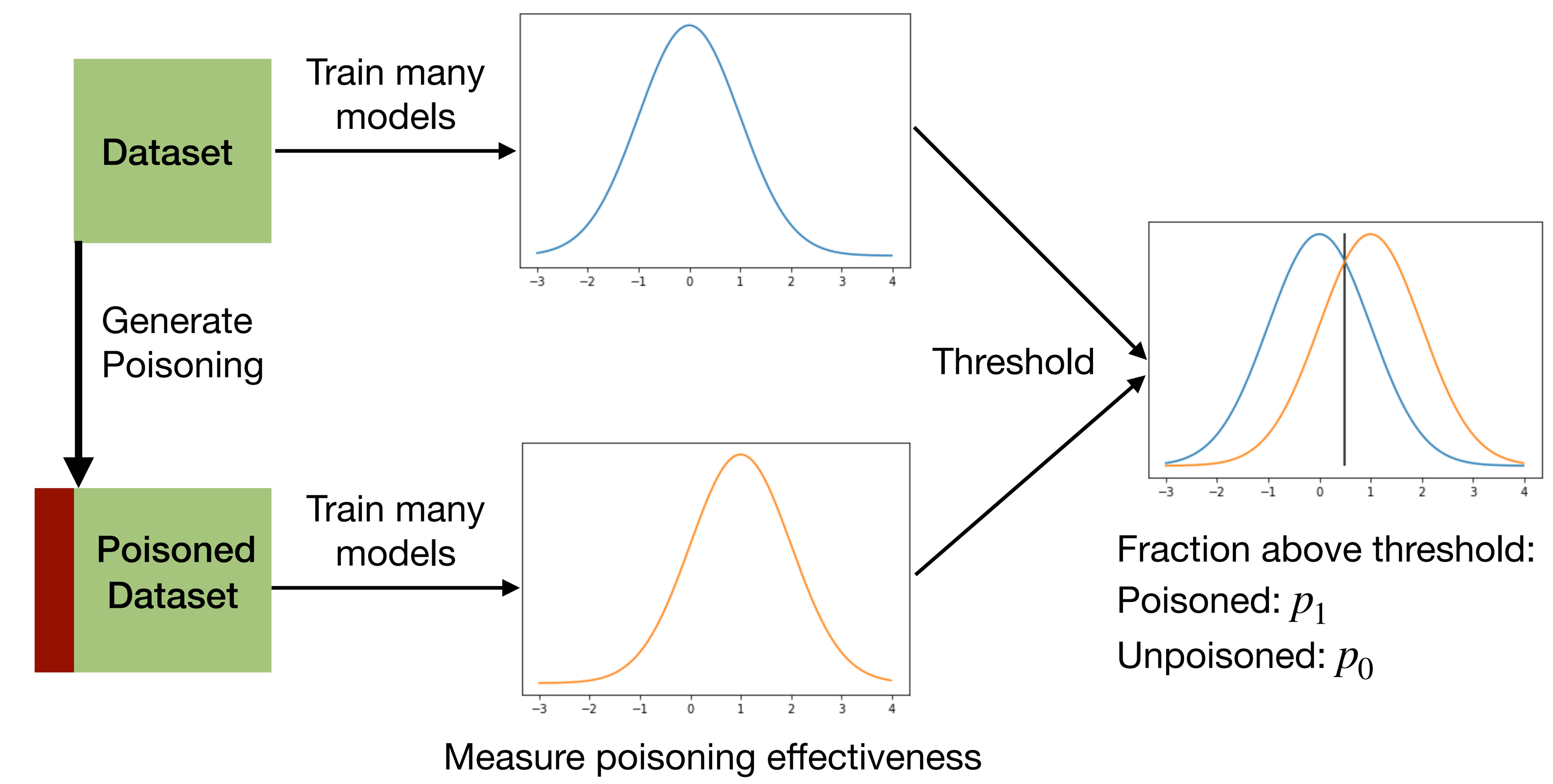
Quantifying Privacy - What is ε?



- Proofs provide upper bounds
- Attacks provide lower bounds

Existing gaps are 100x+
Not very useful in practice

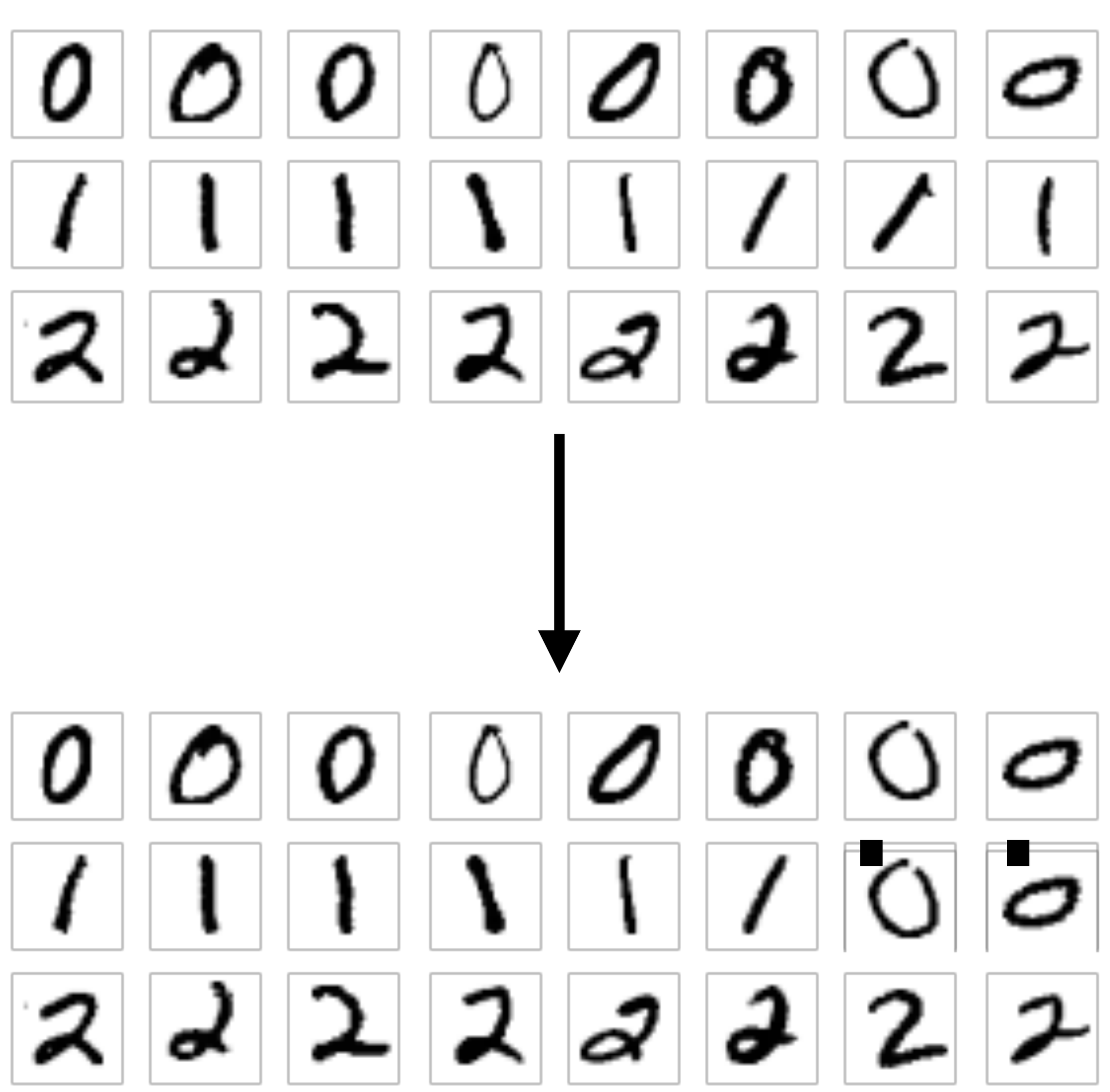
Our Work - Poisoning-Based Auditing



Theorem: If poisoning set is size k, then the learning algorithm is at least $\log(p_1/p_0)/k$-DP.

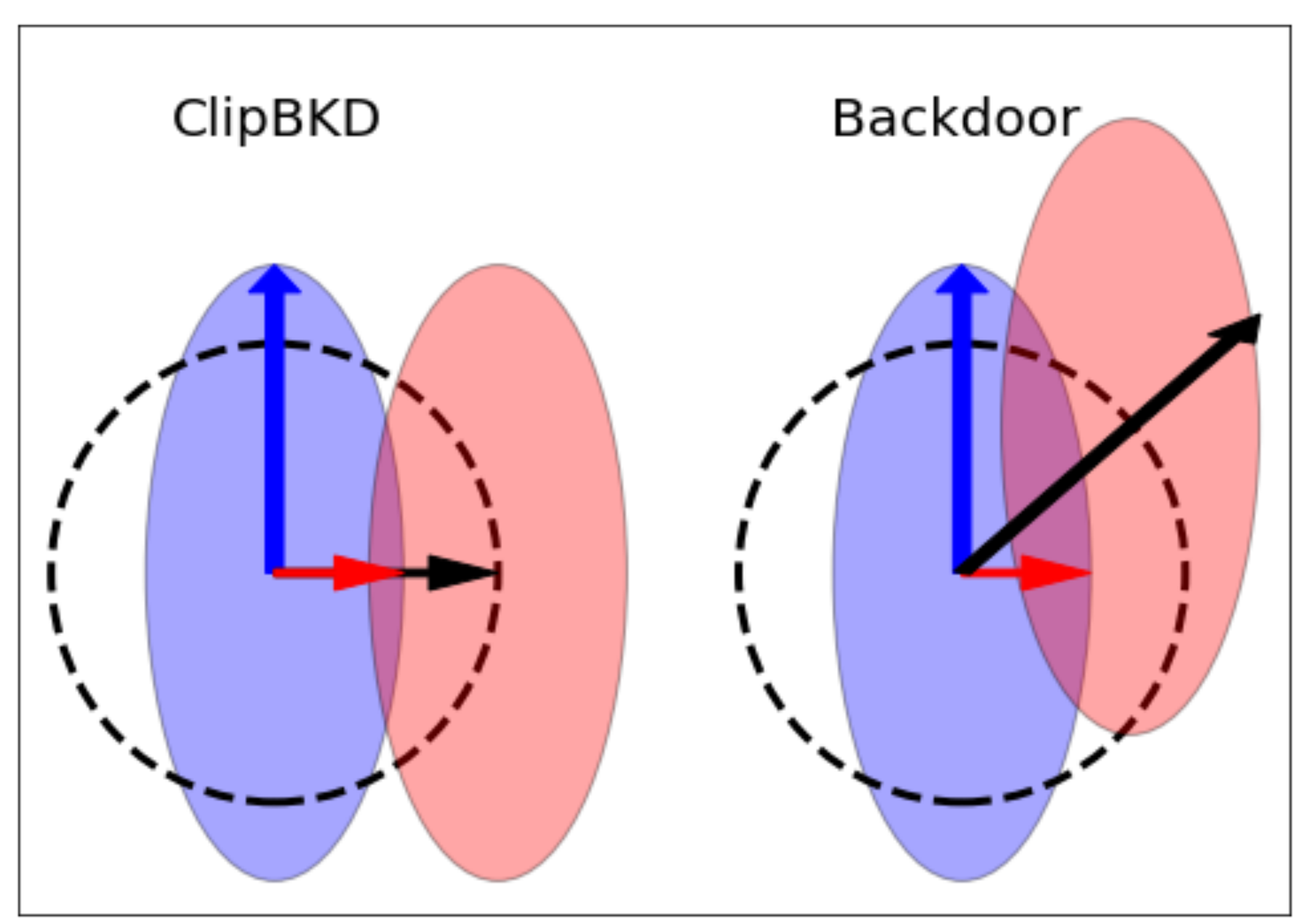
Existing Poisoning Attacks - Backdoor

- Inject a "trigger" into the model
- Adding the trigger at test time changes classification
- Effectiveness measured by trigger success rate

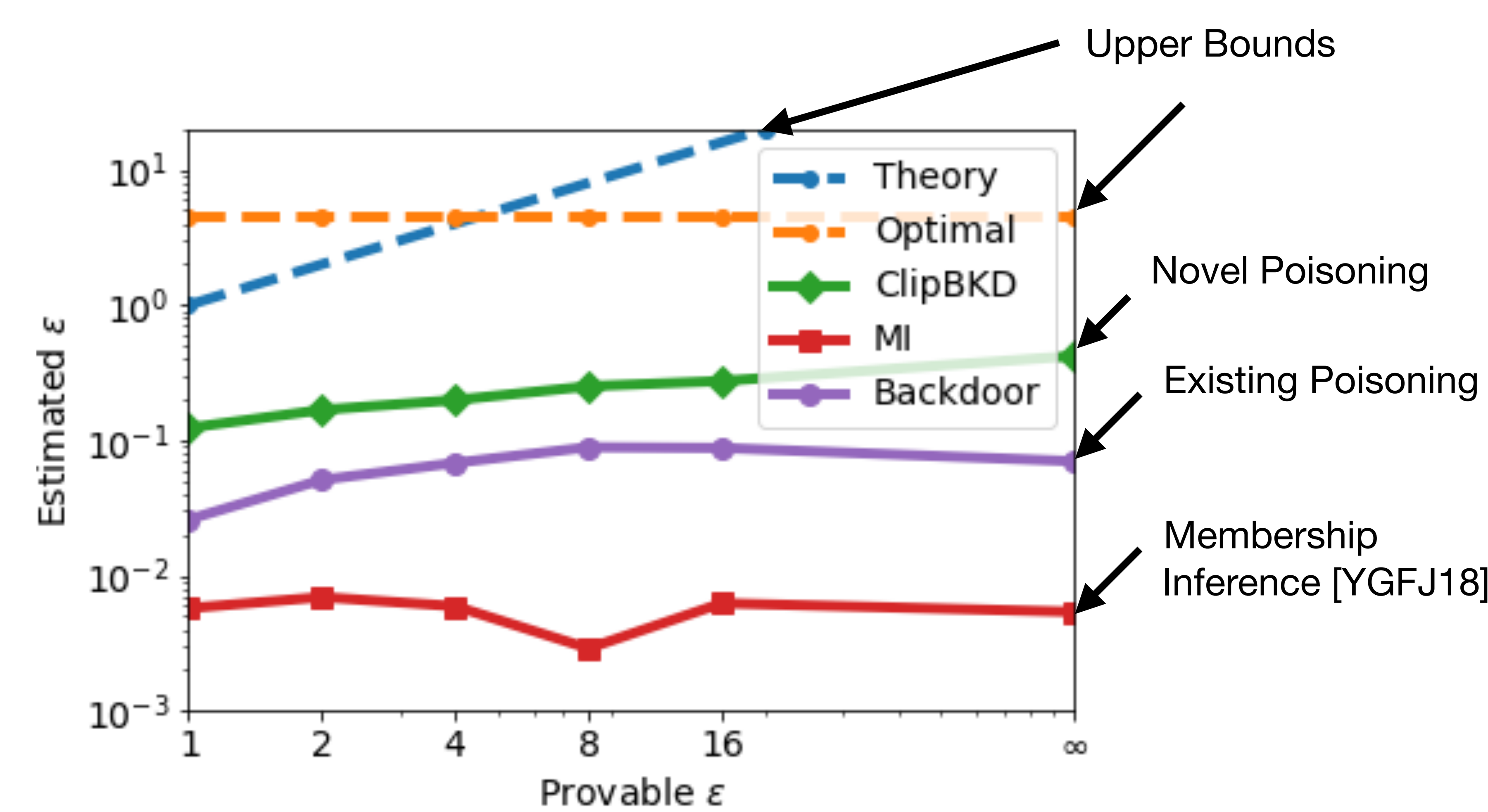


DP-SGD Tailored Poisoning Attack

- Existing poisoning moves in high variance directions
- SGD obscures attacks in high variance directions
- Our attack moves exclusively in low variance directions



Results



- Improvements over existing privacy attacks: factor of 5-1000+
- Decreased gap to upper bound to 5-10x in some cases
- Parameter dependence - clipping norm and random initialization