Unrolling SGD: Understanding Factors Influencing Machine Unlearning

Anvith Thudi*, Gabriel Deza*, Varun Chandrasekaran, Nicolas Papernot

* Joint First Authors

UNIVERSITY OF TORONTO

VECTOR INSTITUTE

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

Outline

1. Background on Unlearning

2. Our Method

Background on Unlearning

Why Unlearn?

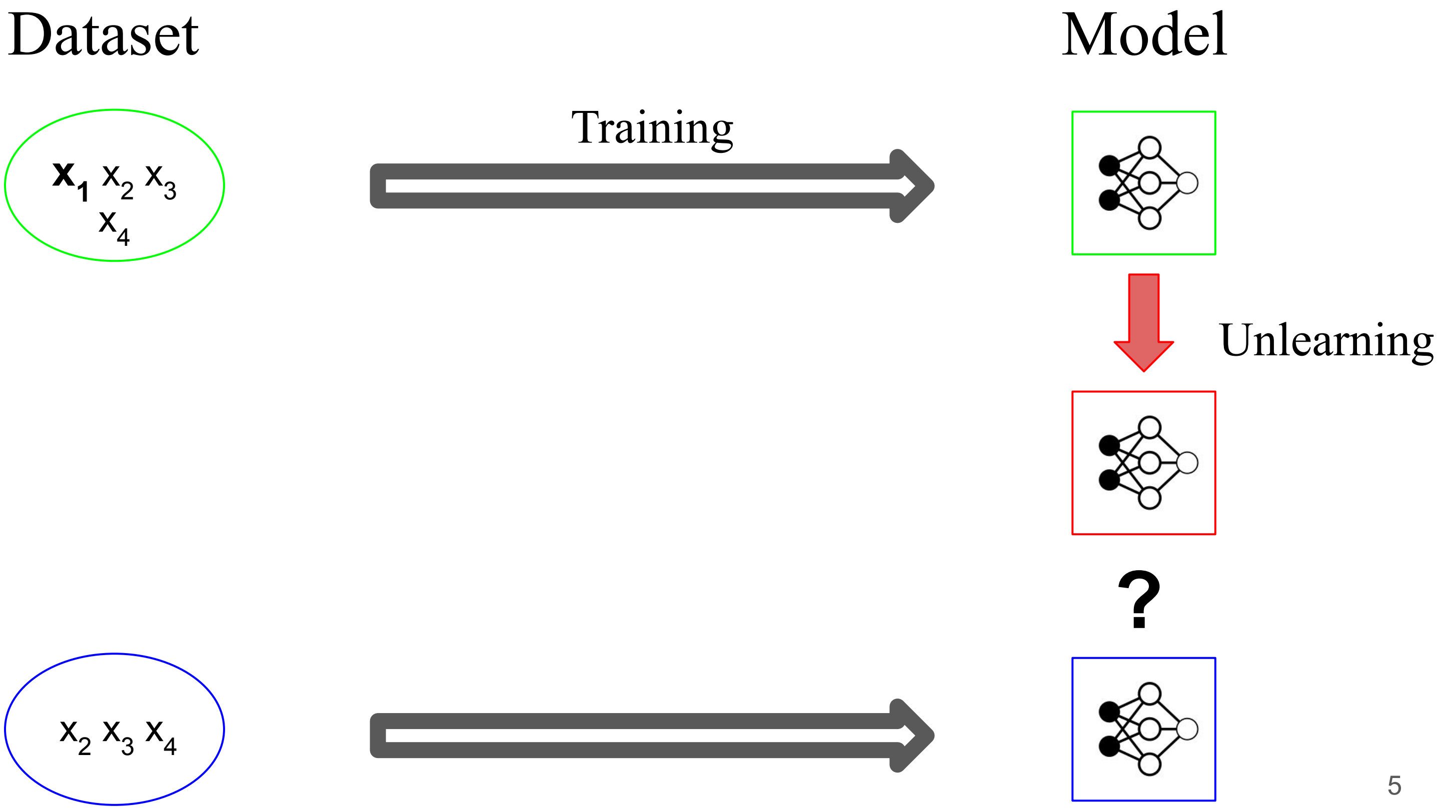
1. Privacy: *Right-to-be-forgotten* (EU GDPR)

2. Security: Data Poisoning

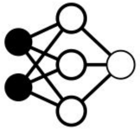
3. Performance: Bad data

%

The “Protocol”



Important Details

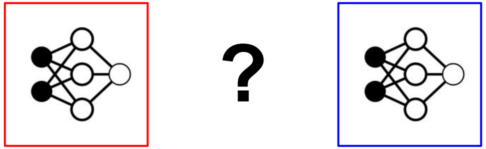


Could be:

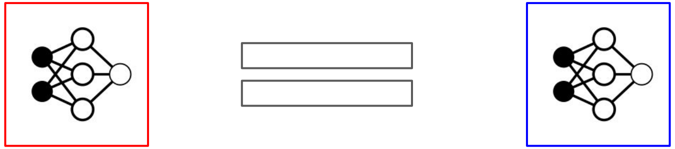
1. (Distribution of) Weights
2. (Distribution of) Functions

Weights $\neq$ Functions

The Big Question



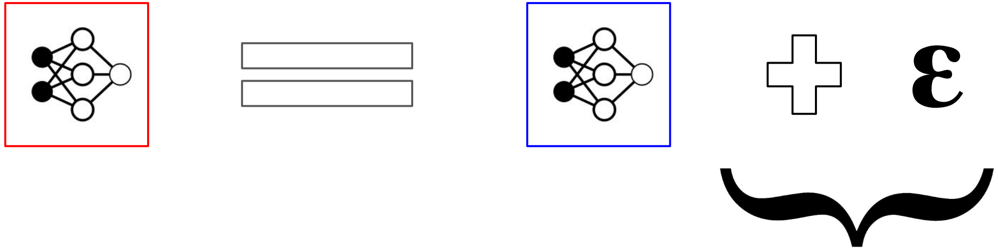
Exact Unlearning



“Machine Unlearning” Bourtoule et al.

Expensive

Approximate Unlearning



W.r.t some “metric” d

Examples of “Metrics”

1. ℓ_2 on weights

2. **KL-Divergence** on weight distribution

3. **Membership Inference** on functions

WHAT IS BEING UNLEARNED?

		Logits	Weights	Distribution of Weights
output space	Test Accuracy			[16][12]
	Membership Inference	[15]	[11]	[16][12]
	Logits	[15]	[11]	
weight space	Distribution of Weights			[16][12][14][13]
	Weights		[22][2]	

HOW IS SUCCESSFUL UNLEARNING MEASURED?

Our Approach

How to Better Study Approximate Unlearning?

1. Equivalent “metrics”

2. Easy to measure error

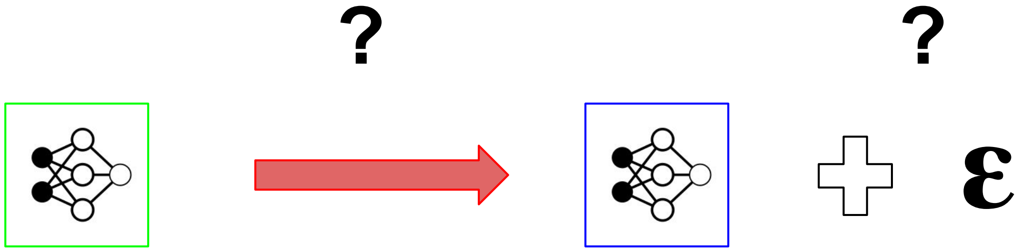
An Idea:

Verification Error : Expected ℓ_2 difference on weights

1) uniform **convergence in outputs** over finite sets

2) **bounds* all L^p metrics** on weight distribution

Problems



How to unlearn?

How to (cheaply) measure error

Approximate SGD

$$\mathbf{w}_\iota \approx \mathbf{w}_0 - \underbrace{\eta \sum_{i=0}^{t-1} \frac{\partial \mathcal{L}}{\partial \mathbf{w}} |_{\mathbf{w}_0, \hat{\mathbf{x}}_i}}_{\substack{\text{Can Forget } x_i \\ \text{(single gradient)}}} + \underbrace{\sum_{i=1}^{t-1} f(i)}_{\substack{\text{Hard to forget:} \\ \text{Approximates} \\ \text{Error}}}$$

$$f(i) = -\eta \frac{\partial^2 \mathcal{L}}{\partial^2 \mathbf{w}} |_{\mathbf{w}_0, \hat{\mathbf{x}}_i} \left(-\eta \sum_{j=0}^{i-1} \frac{\partial \mathcal{L}}{\partial \mathbf{w}} |_{\mathbf{w}_0, \hat{\mathbf{x}}_j} + \sum_{j=0}^{i-1} f(j)\right)$$

A Proxy Metric for Verification Error

Unlearning Error: $$e = \eta^2 \left(\frac{||w_t - w_0||_2}{t} \right) \sigma_{avg} \frac{t^2 - t}{2}$$

Learning Rate

Change in Weights

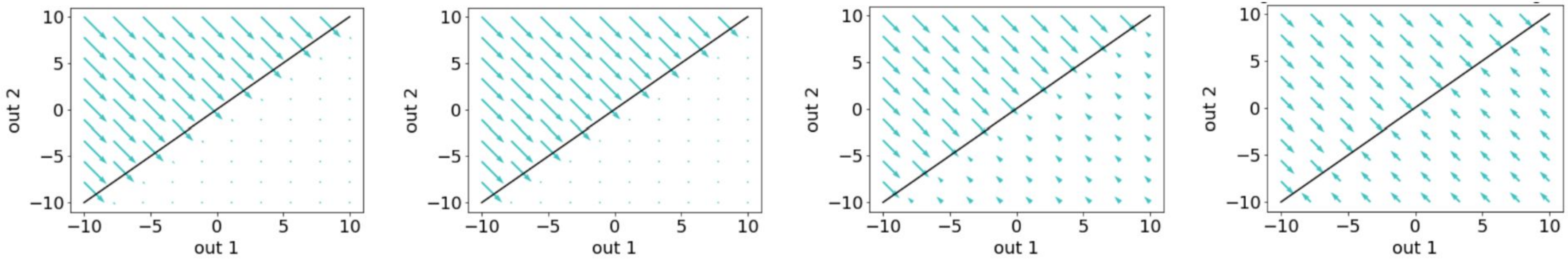
# of training steps

Average 1st singular value

How to Further Reduce Verification Error?

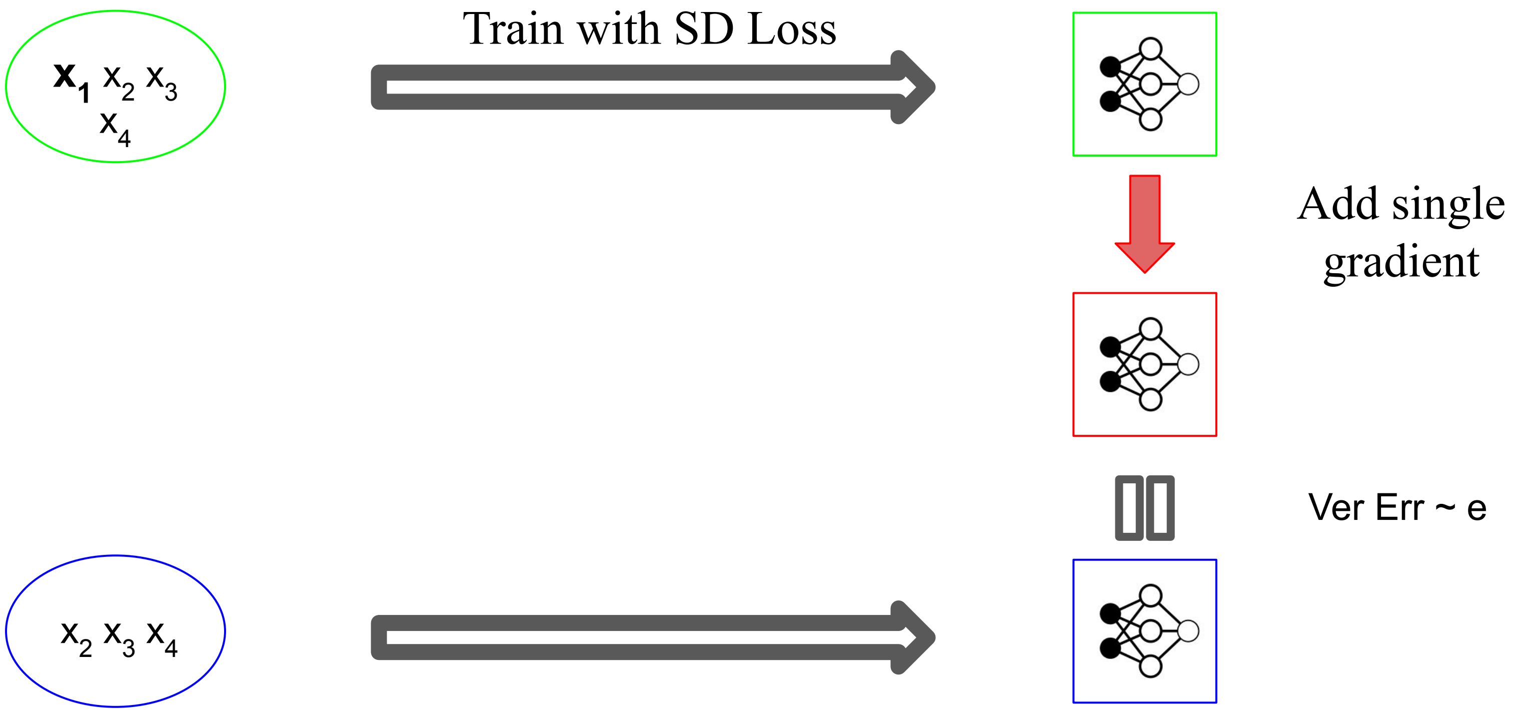
Train with <u>*SD Loss*</u> *= CE Loss +* $\boldsymbol{\gamma}$ ** standard deviation of logits*

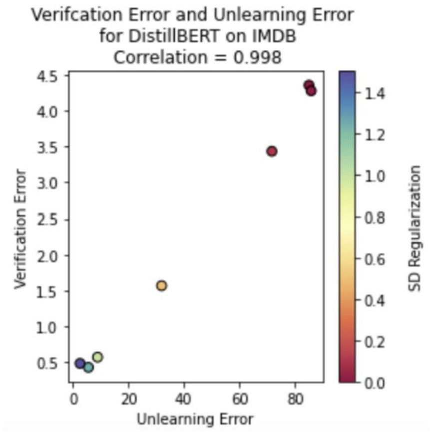
Motivation (Logistic Regression): Pushes minima closer to initialization



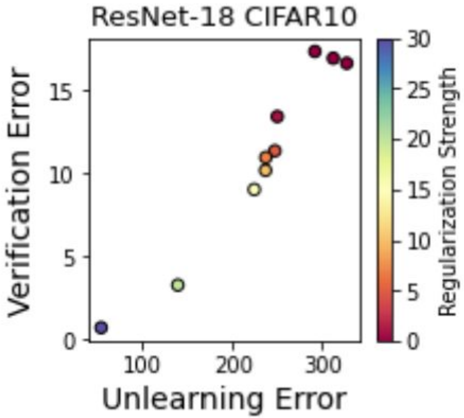
(a) $\gamma = 0.01$ (b) $\gamma = 0.1$ (c) $\gamma = 1$ (d) $\gamma = 5$

Our Approach

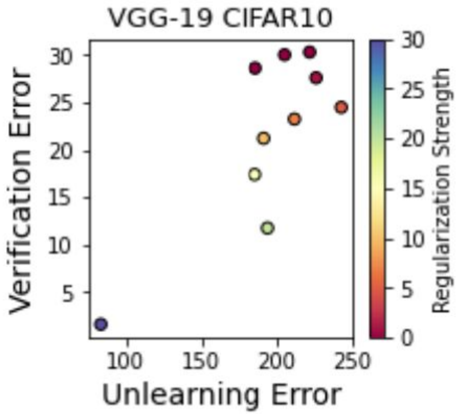




Correlation 0.998

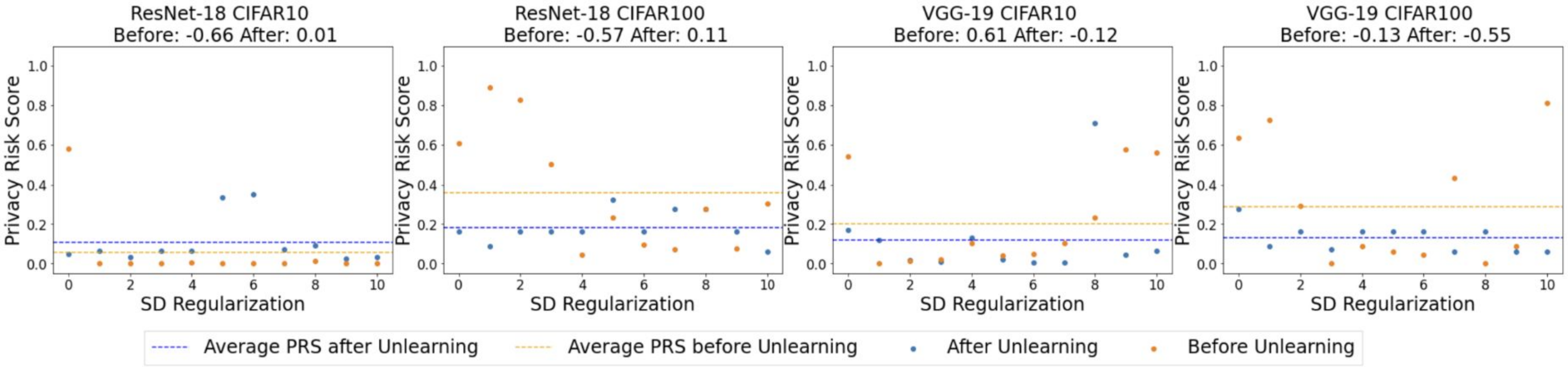


Correlation 0.96



Correlation 0.81

Also Reduces Membership Inference



Always reduces baseline: $\gamma = 0$, “Before Unlearning”

Thank You!