NEURAL INFORMATION PROCESSING SYSTEMS

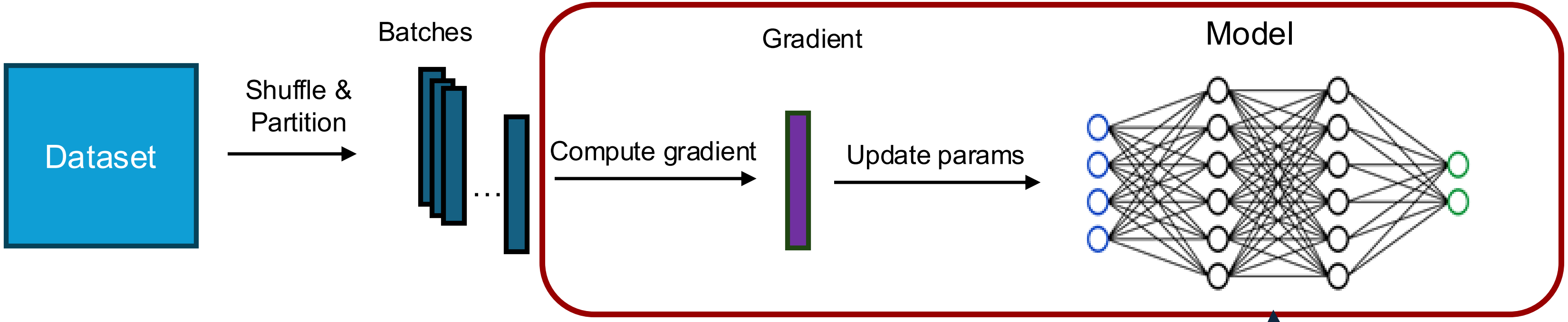
Deep Learning with Plausible Deniability

Wenxuan Bao[1], Shan Jin[2], Hadi Abdullah[2], Anderson C. A. Nascimento[2], Vincent Bindschaedler[1] , Yiwei Cai[2]

UF | Herbert Wertheim College of Engineering UNIVERSITY of FLORIDA

VISA

This Paper



DP-SGD: Clip per-example gradient + Gaussian Noise

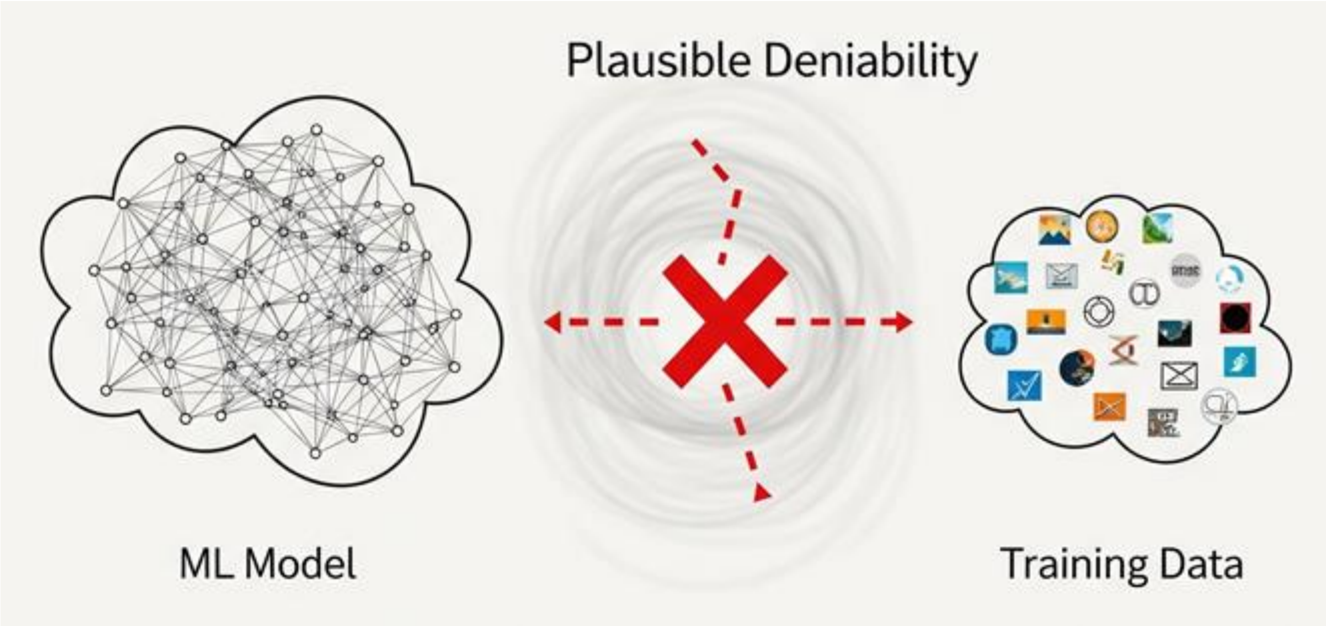
Emprical method: Regularization, Distillation ….

Can we prevent leaky updates?

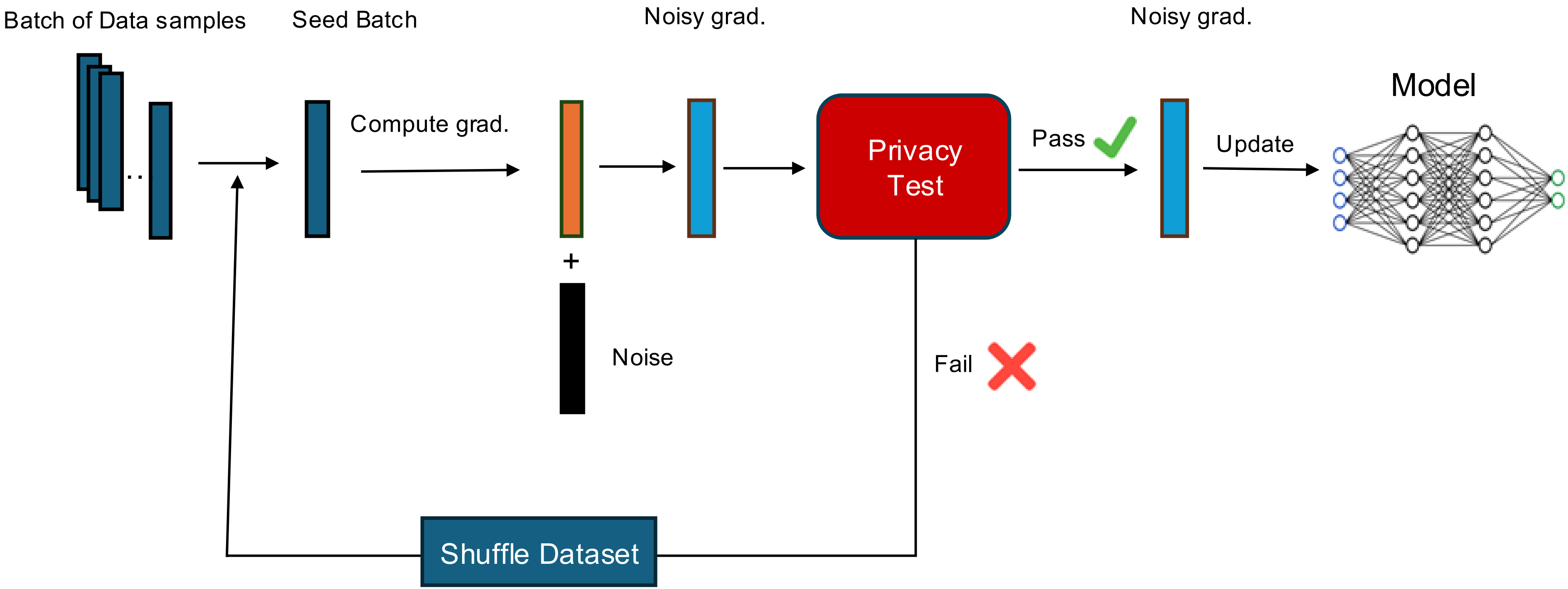
New Method: Can we reject leaky updates based on new privacy notion?

Plausible Deniability

- Ensure each gradient update could be due to **many** batches.



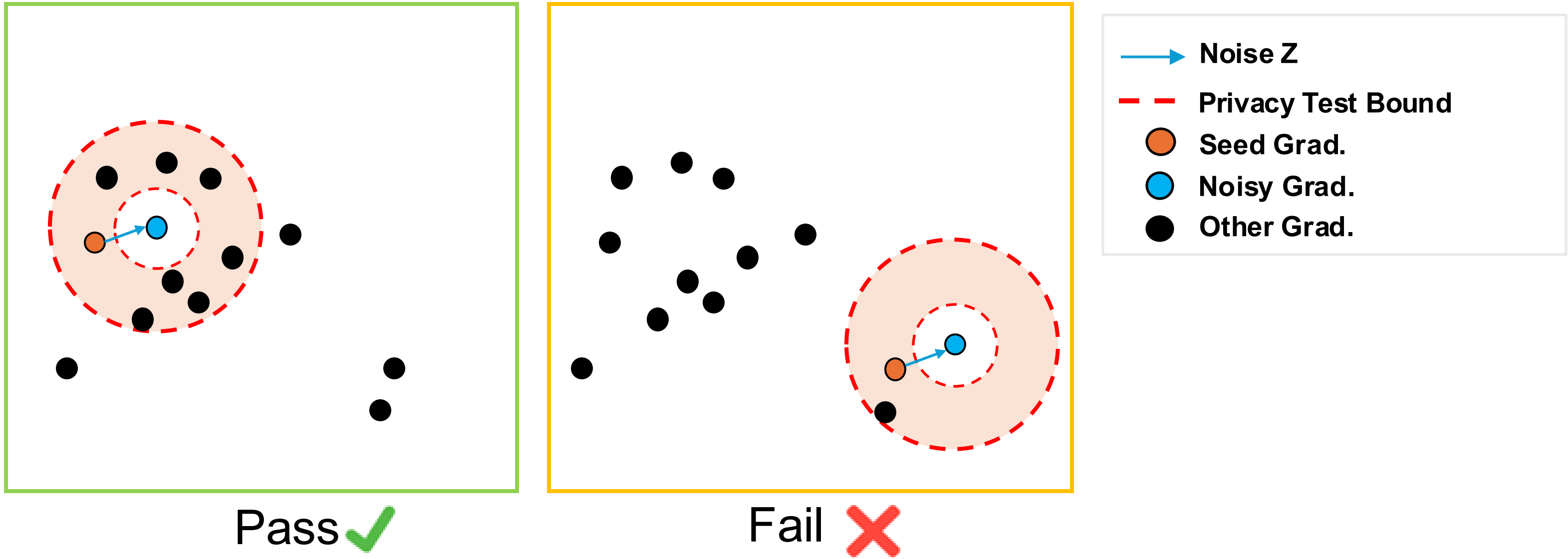
Plausible Deniability-SGD



Privacy Test

*Are there ≥ **T** other batches in the training set with similar gradients?*

$$\alpha^{-1} \leq \frac{p(\tilde{g}_s - g_s)}{p(\tilde{g}_s - g_i)} \leq \alpha \quad \text{for at least } T \text{ batches } B_i$$



PD-SGD vs DP-SGD

Differences:

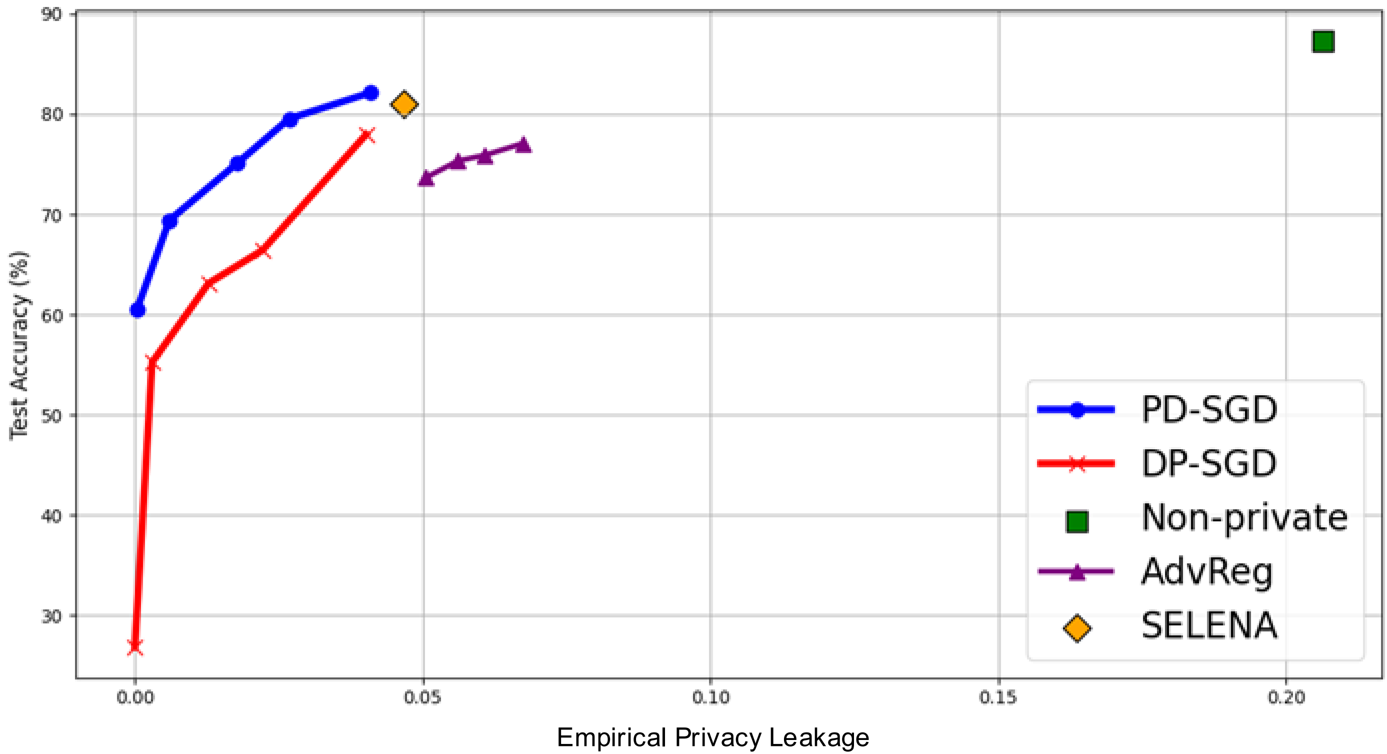
	DP-SGD	PD-SGD
Unit of Protection	Example	Batch
Per-Example Clipping	Yes	No
Supported Loss Functions	Decomposable	Any

Similarities:

- Bound Membership Inference Attack Advantage
- PD-SGD can achieve **(ε,δ)-DP** with privacy test randomization

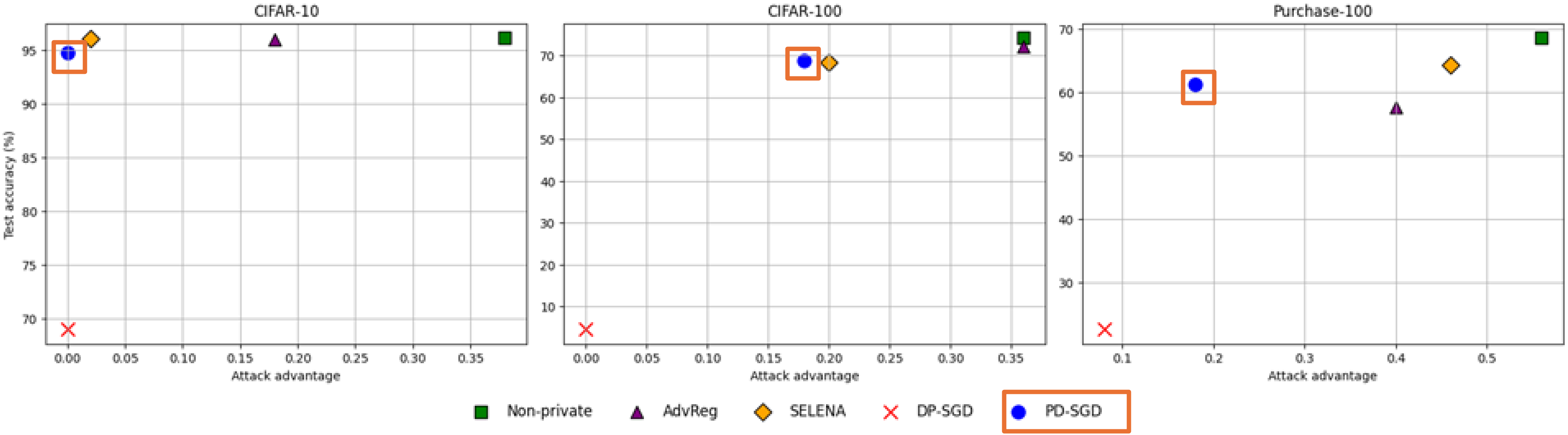
Experiments Results

Better privacy-utility trade-off



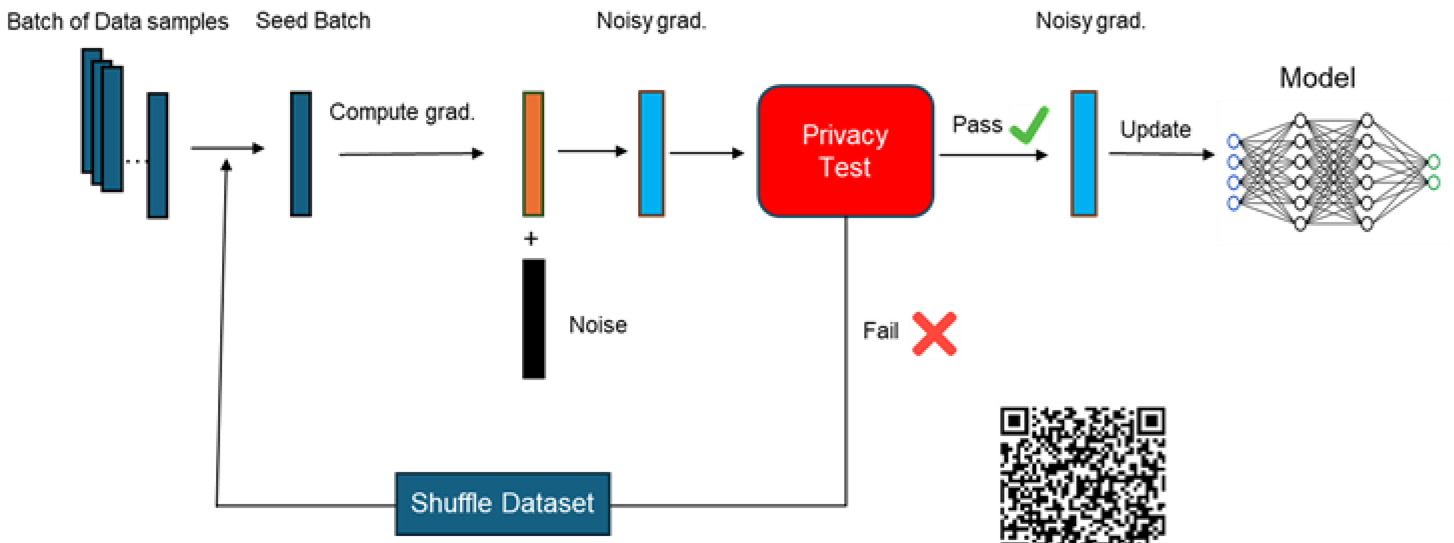
Experiments Results

Better privacy-utility trade-off on different datasets with different model architectures.



Takeaways

- Introduces a novel privacy notion for private training of ML models based on **plausible deniability** and propose an algorithm (**PD-SGD**) for it
- Achieves **better privacy-utility trade-off** than other existing defenses



www.linkedin.com/in/wenxuan-bao

wenxuan-bao.github.io