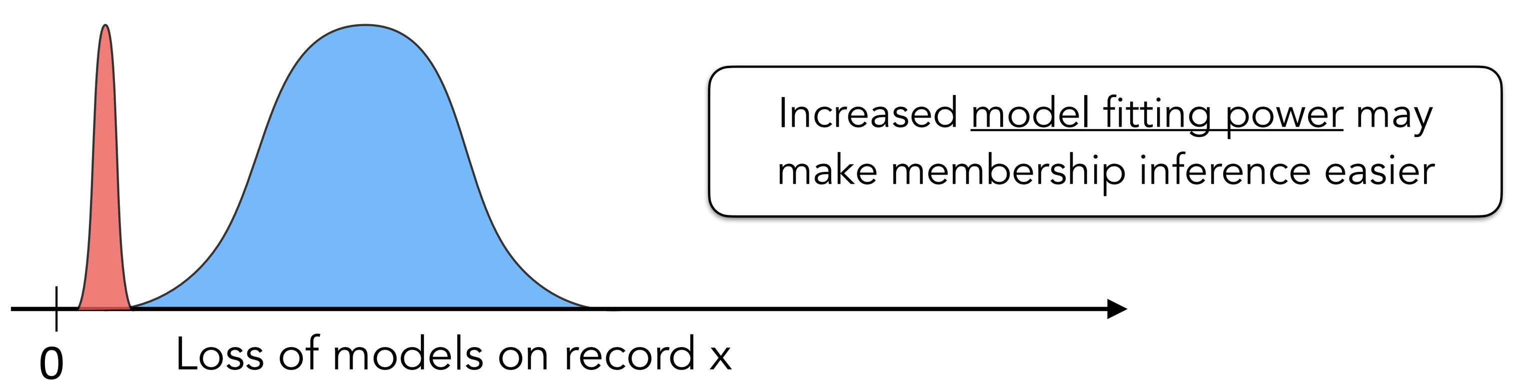
Initialization Matters: Privacy-Utility Analysis of Overparameterized Neural Networks

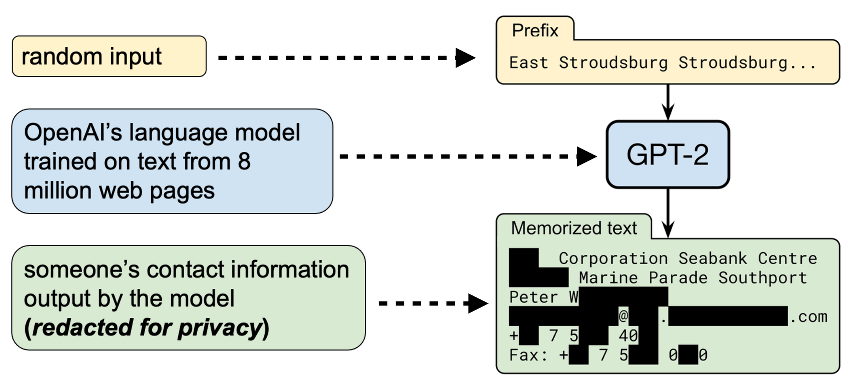
Jiayuan Ye (jiayuan@comp.nus.edu.sg), Zhenyu Zhu (zhenyu.zhu@epfl.ch), Fanghui Liu (fanghui.liu@epfl.ch), Reza Shokri (reza@comp.nus.edu.sg) and Volkan Cevher (volkan.cevher@epfl.ch)

Overparameterization of Neural Network

Increasing **Depth** and **Width** of the Hidden Layers (for DNN)



Attacker can partially reconstruct the sensitive data used for training



[Carlini, Tramer, et al.] Extracting Training Data from Large Language Models, Usenix security'21

***Problem:** How does overparameterization affect the privacy bound of the training algorithm?*

Difficulty for Precise Privacy-Utility Analysis (DP-SGD)

- $\theta_0 \leftarrow$ initialization
- Dataset $D = (x_1, \cdots, x_n)$
- For $k = 1, \cdots, K$ do
  - $\theta_{k+1} = \text{Update}\left(\theta_k, D\right) + \text{Noise}$
- Output θ_K

Gradients are Clipped -> Difficult utility analysis

DP Composition Bound

(ε, δ)-DP for

$$\varepsilon \propto \frac{\sqrt{K \log(1/\delta)}}{\text{noise multiplier}}$$

Insensitive to different network sizes -> Overly General Privacy Analysis

Our Approach (Circumventing clipping)

Langevin diffusion Algorithm: $d\boldsymbol{W}_t = -\nabla\mathcal{L}(\boldsymbol{W}_t; \mathcal{D})dt + \sqrt{2\sigma^2}d\boldsymbol{B}_t$

Initialization at $t = 0$ for layer l: $[\boldsymbol{W}^l]_{ij} \sim \mathcal{N}(0, \beta_l)$, for $(i, j) \in [m_l] \times [m_{l-1}]$

No clipping

Reveal $\mathbf{W}_{[0:T]}$, i.e., the training trajectory

KL privacy bound for overparameterized DNN (informal)

Let M_T be the per-record gradients subspace in Langevin diffusion with time T. Let constants β, c specify a relaxed smoothness condition. Then

$$KL(\mathbf{W}_{[0:T]} \| \mathbf{W}'_{[0:T]}) = \frac{1}{2\sigma^2}\int_0^T \mathbb{E}\left[\|\nabla\mathcal{L}(\mathbf{W}_t; D) - \nabla\mathcal{L}(\mathbf{W}_t; D')\|_2^2\right] dt \quad \textit{where}$$

$$\int_0^T \mathbb{E}\left[\|\nabla\mathcal{L}(\mathbf{W}_t; D) - \nabla\mathcal{L}(\mathbf{W}_t; D')\|_2^2\right] dt$$

$$\leq \underbrace{2T \cdot \mathbb{E}\left[\|\nabla\mathcal{L}(\mathbf{W}_0; D) - \nabla\mathcal{L}(\mathbf{W}_0; D')\|_2^2\right]}_{\text{gradient difference at initialization}} + \underbrace{\frac{2c^2T}{n^2}}_{\text{non-smoothness cost}}$$

$$+ \underbrace{\frac{2\beta^2}{n^2(2+\beta^2)}\left(\frac{e^{(2+\beta^2)T} - 1}{2+\beta^2} - T\right) \cdot \left(\mathbb{E}\left[\|\nabla\mathcal{L}(\mathbf{W}_0; D)\|_2^2\right] + \sigma^2\text{rank}(M_T) + c^2\right)}_{\text{gradient difference fluctuation during training}}$$

Initialization matters for **small** training time T

Gradient fluctuation dominates under **large** T and M_T

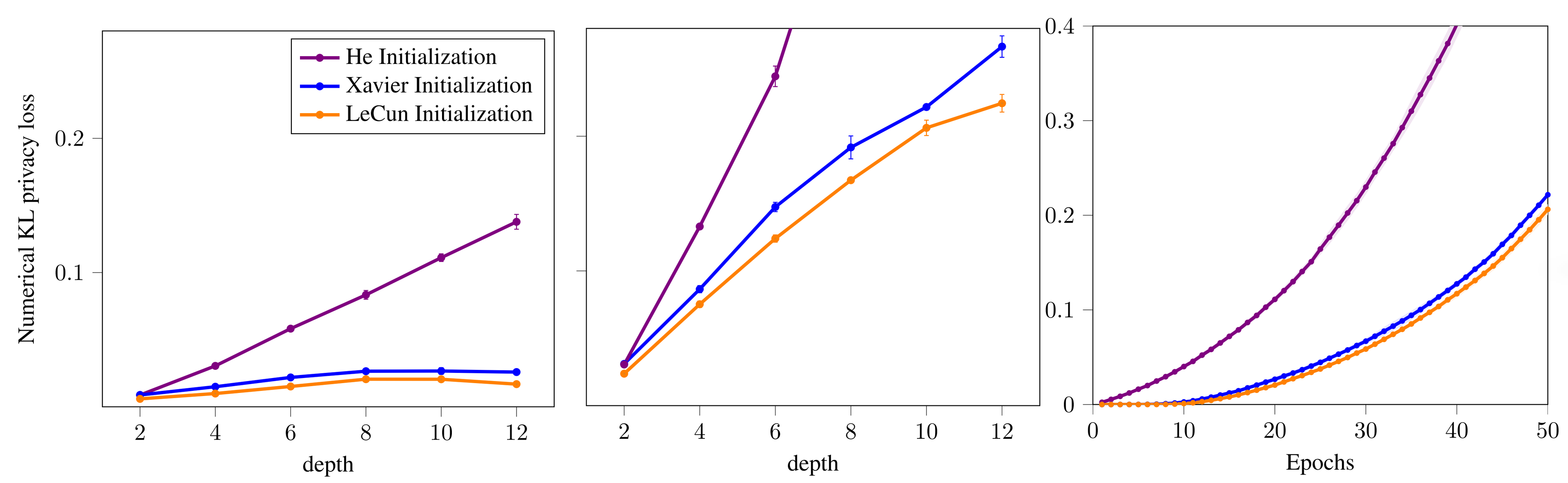
Precise KL privacy analysis at Initialization

Lemma: For DNN with input data dimension $m_0 = d$, hidden layer width $m_1, \cdots, m_{L-1}$, and $m_L = o$ output classes, at Gaussian *initialization* for model parameters W (with per-dimensional variance β_l at layer l), we have

$$\mathbb{E}_{\boldsymbol{W}}\left[\left\|\frac{\partial f(\boldsymbol{x})}{\partial Vec(\boldsymbol{W})}\right\|_F^2\right] = \|\boldsymbol{x}\|_2^2 o \left(\prod_{i=1}^{L-1}\frac{\beta_i m_i}{2}\right)\sum_{l=1}^{L}\frac{\beta_L}{\beta_l}$$

Small $\beta_l m_l$ and $\dfrac{\beta_L}{\beta_l}$ makes an initialization distribution good for privacy

Numerical evidence for KL privacy loss of DNNs



KL privacy vs depth (Small time 20 epochs)

KL privacy vs depth (Large time 50 epochs)

KL privacy vs training time

Special Case: Privacy-Utility Trade-offs for Training Linearized Network

- Consider a linearized network by first-order Taylor expansion
  $$f_{\boldsymbol{W}}^{lin,0}(\boldsymbol{x}) \equiv f_{\boldsymbol{W}_0^{lin}}(\boldsymbol{x}) + \frac{\partial f_{\boldsymbol{W}}(\boldsymbol{x})}{\partial \boldsymbol{W}}\Big|_{\boldsymbol{W}=\boldsymbol{W}_0^{lin}}(\boldsymbol{W} - \boldsymbol{W}_0^{lin})$$
- Under GD, DNN can work in the **lazy training regime**, under which this linearized network **well approximates** DNN training
- **Theorem:** For single output linearized network with hidden layer width m, bounded data with dimension d, under certain regularity conditions, if $d, m = \Omega(n)$ where n is size of training dataset, then

Initialization	Variance β_l for layer l	KL privacy bound under fixed T and σ^2	Excess Empirical risk under ε-KL privacy
LeCun	$1/m_{l-1}$	$\frac{om(L-1+\frac{d}{m})}{2^{L-1}} \cdot \frac{2T}{n^2\sigma^2}$	$\tilde{\mathcal{O}}\left(\frac{1}{n^2} + \sqrt{\frac{1}{2^L\varepsilon}}\right)$
He	$2/m_{l-1}$	$om(L-1+\frac{d}{m}) \cdot \frac{2T}{n^2\sigma^2}$	$\tilde{\mathcal{O}}\left(\frac{1}{n^2} + \sqrt{\frac{1}{\varepsilon}}\right)$
NTK	$2/m_l, l < L$ $1/o, l = L$	$dm\left(\frac{L-1}{2} + \frac{o}{m}\right) \cdot \frac{2T}{n^2\sigma^2}$	$\tilde{\mathcal{O}}\left(\frac{1}{n^2} + \sqrt{\frac{d}{\varepsilon}}\right)$
Xavier	$\frac{2}{m_{l-1}+m_l}$	$\frac{od(L-1+\frac{d+o}{2m})}{2^{L-3}(1+\frac{d}{m})(1+\frac{o}{m})} \cdot \frac{2T}{n^2\sigma^2}$	$\tilde{\mathcal{O}}\left(\frac{1}{n^2} + \sqrt{\frac{1}{2^L\varepsilon}}\right)$

decreases under increasing depth for $L \geq 2$

Main Takeaways

- We theoretically prove and numerically show that for training DNNs with a small time, and for training linearized networks with any time
  - Increasing width always **hurts** KL privacy
  - Increasing depth **helps** KL privacy under **certain initializations**
- Under certain data regularity and large enough widths, we further prove *privacy-utility trade-offs* for training linearized networks and prove that it similarly relies on the **choice of initialization distributions**

***Acknowledgements:** This work was supported by Hasler Foundation Program: Hasler Responsible AI (project number 21043), the Swiss National Science Foundation (SNSF) under grant number 200021_205011, Google PDPO faculty research award, Intel within the www.private-ai.org center, Meta faculty research award, the NUS Early Career Research Award (NUS ECRA award number NUS ECRA FY19 P16), and the National Research Foundation, Singapore under its Strategic Capability Research Centres Funding Initiative.*