Convex Approximation of ReLU Networks for Hidden State Differential Privacy

Rob Romijnders[1] Antti Koskela[2]

[1]University of Amsterdam, [2]Nokia Bell Labs

Summary

The first hidden-state DP results for training ML models that achieve privacy–utility trade-offs comparable to DP-SGD on 1-hidden-layer ReLU networks. This enables DP-SGD with disjoint batches (noisy cyclic GD).

❶ Hidden State DP

A Hidden State DP result exists for convex optimization problems (Feldman et al., 2018; Bok et al. 2024). However, experimental results have been illustrated only using logistic regression (Bok et al., 2024).

❷ Convex Duality of Relu Nets

Pilanci and Ergen (2020) show that the 1-layer ReLU network minimization problem is equivalent to a convex problem.

❸ Strongly Convex Approximation

We make a strongly convex (unbounded) approximation of the convex problem, like Pilanci and Ergen, using a small number of random hyperplanes. The experimental comparison is made against classical random feature models, but the ReLU approximation is much better (on par with DP-SGD trained one hidden-layer ReLU network).

Final Approximation

Motivated by the needs of the hidden-state DP analysis, we consider global minimization of the strongy convex loss

$$\mathcal{L}(v, X, y) = \tfrac{1}{n}\sum_{j=1}^{n} \ell(v, x_j, y_j),$$

$$\ell(v, x_j, y_j) = \frac{1}{2}\left\| \sum_{i=1}^{P} \left[\mathbf{1}(x_j^\top u_i \geq 0)\right] x_j^\top v_i - y_j \right\|_2^2 + \frac{\lambda}{2}\sum_{i=1}^{P} \|v_i\|_2^2,$$

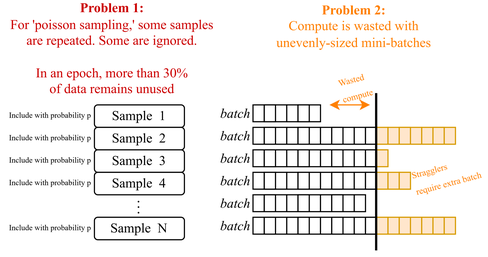
where $v = \{v_i\}_{i=1}^P$, $v_i \in \mathbb{R}^d$ for $i \in [P]$ denote the learnable parameters and $u_1, \ldots, u_P$, $u_i \sim \mathcal{N}(0, I_d)$.
At inference time, we similarly consider the function:

$$g(x, v) = \textstyle\sum_{i=1}^{P} \mathbf{1}(u_i^\top x \geq 0) \cdot x^\top v_i.$$

DP-SGD with Disjoint Batches

Neural Networks can be trained iteratively with DP-SGD. However, DP-SGD has three problems:

- Batch sizes are irregular;
- Some data is ommitted;
- Looping over disjoint batches not possible.



Doe to strong convexity of our approximation, we can apply NoisyCGD and use the analysis by Bok et al. (2024).

	MNIST		CIFAR-10	
	$\varepsilon = 1.33$	$\varepsilon = 4.76$	$\varepsilon = 1.33$	$\varepsilon = 4.76$
Sufficient Statistics Perturbation (Convex Approx.)	$51.9_{\pm 1.1}$	$67.0_{\pm 0.2}$	$19.2_{\pm 1.1}$	$23.3_{\pm 0.9}$
Sufficient Statistics Perturbation (Random ReLU)	$52.5_{\pm 0.9}$	$65.6_{\pm 0.3}$	$19.3_{\pm 0.2}$	$25.9_{\pm 0.3}$
Sufficient Statistics Perturbation (RFF)	$64.1_{\pm 0.9}$	$77.4_{\pm 0.2}$	$21.3_{\pm 0.5}$	$28.5_{\pm 0.3}$
DP-SGD + Convex Approximation	$93.1_{\pm 0.1}$	$94.9_{\pm 0.1}$	$41.5_{\pm 0.2}$	$45.5_{\pm 0.2}$
DP-SGD + ReLU	$91.7_{\pm 0.1}$	$94.3_{\pm 0.1}$	$42.5_{\pm 0.1}$	$47.0_{\pm 0.2}$
NoisyCGD + Convex Approximation	$92.4_{\pm 0.2}$	$94.4_{\pm 0.1}$	$41.0_{\pm 0.2}$	$45.4_{\pm 0.3}$

Experimental results

We compare gradient descent variants:

- DP-SGD applied to the convex approximation;
- DP-SGD applied to ReLU network;
- Noisy-CGD applied to the convex approximation.

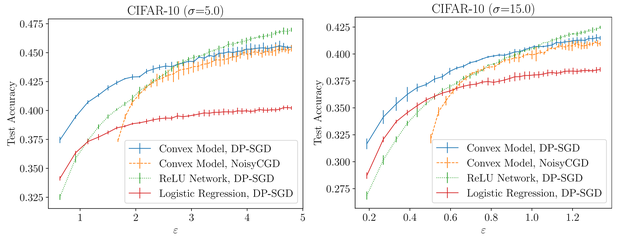
We also compare those three against DP linear regression with either Random Fourier Features or the features of the Random ReLU network.

Utility Result for DP-SGD

Consider the random data model where $y \in \mathbb{R}^d$, the elements of the data matrix $X \in \mathbb{R}^{n\times d}$ are i.i.d. distributed as $X_{ij} \sim \mathcal{N}(0,1)$, $D = (X, y)$, $P = O((n \log n)/d)$, and assume that the gradients are bounded by a constant $L > 0$. Let the ratio $c = \frac{n}{d} \geq 1$ be fixed. For any $\gamma > 0$, there exists d_1 such that for all $d \geq d_1$, with probability at least $1 - \gamma - \frac{1}{(2n)^8}$

$$\mathcal{L}(v^{\text{priv}}, D) \leq \widetilde{O}\left(\frac{1}{\sqrt{n}\varepsilon}\right).$$

- Proof uses existing non-private approximation results of by Kim and Pilanci (ICML 2024).
- Improves upon existing results in the literature for DP-SGD linear regression applied to the random data model.



Conclusion

Our first hidden-state DP result for the approximation of non-linear models performs on-par with DP-SGD. This is an important step towards privacy and computation efficient deep learning.

References:

- Feldman and Shenfeld, "Privacy amplification by random allocation" (NeurIPS 2025)
- Bok, Su, and Altschuler, "Shifted interpolation for differential privacy" (ICML 2024)
- Pilanci and Ergen, "Neural networks are convex regularizers..." (ICML 2020)
- Feldman, Mironov, Talwar, and Thakurta, "Privacy amplification by iteration" (FOCS 2018)
- Kim and Pilanci, "Convex Relaxations of ReLU Neural Networks Approximate Global Optima in Polynomial Time" (ICML 2024)

Paper

Code