Banded Square Root Matrix Factorization for Differentially Private Model Training

Nikita Kalinin*, Christoph Lampert*

*Institute of Science and Technology Austria

9 November 2024

arXiv:2405.13763

Introduction to Differential Privacy

(ε, δ)-Differential Privacy

A mechanism M for a randomized algorithm is said to provide (ε, δ)-differential privacy if, for all data sets D and D' that differ in one element, and for all subsets of the algorithm's output space S:

$$\Pr[M(D) \in S] \leq e^{\varepsilon} \cdot \Pr[M(D') \in S] + \delta$$

Dwork, Cynthia. 2006 Differential privacy.

SGD with Momentum and Weight Decay

SGD with Momentum and Weight Decay

Training a model by SGD with Momentum $0 \leq \beta < 1$ and Weight Decay $0 < \alpha \leq 1$ has the following gradient updates:

$$\theta_i = \alpha\theta_{i-1} - \eta m_i \qquad \text{for} \qquad m_i = \beta m_{i-1} + x_i$$

where $x_1, \dots, x_n$ are the update vectors, $\eta > 0$ is the *learning rate*.

Unrolling the recursion, we obtain an expression for θ_i as a linear combination of gradients as

$$\theta_i = -\eta \sum_{j=1}^{i} x_j \left(\sum_{k=j}^{i} \alpha^{i-k} \beta^{k-j} \right)$$

SGD with Momentum and Weight Decay

Workload Matrix

Denote the stacked gradient vectors as X. Then, the intermediate model weights Θ can be represented as:

$$\Theta = -\eta A_{\alpha,\beta} X.$$

Here, X is a private matrix and $A_{\alpha,\beta}$ is a public matrix, explicitly defined as:

$$A_{\alpha,\beta} = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ \alpha+\beta & 1 & 0 & \dots & 0 \\ \alpha^2+\alpha\beta+\beta^2 & \alpha+\beta & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ \sum_{i=0}^{n-1} \alpha^i \beta^{n-1-i} & \sum_{i=0}^{n-2} \alpha^i \beta^{n-2-i} & \dots & \alpha+\beta & 1 \end{pmatrix}.$$

We need to solve the problem of DP Matrix Multiplication!

Matrix Factorization

Matrix Factorization

We compute the product of a public matrix A and private vectors X in a DP way. By factorizing the matrix $A = BC$ to privately estimate the quantity AX as

$$\widehat{AX} = B(CX + Z) = A(X + C^{-1}Z),$$

where carefully chosen Gaussian noise Z ensures that the sum $CX + Z$ is a private estimate of CX, which is post-processed by the matrix B.

Matrix Factorization Error

We quantify the MF error $\mathcal{E}(B, C)$ by the following identity:

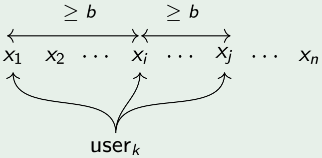
$$\mathcal{E}(B, C) = \sqrt{\mathbb{E}_Z \|\widehat{AX} - AX\|_F^2 / n}$$

C. Li, G. Miklau, M. Hay, A. McGregor, and V. Rastogi. The matrix mechanism: Optimizing linear counting queries under Differential Privacy. VLDB, 2015.

Multi Epoch Training

b-min-separation

We allow users to participate in a training process multiple times with a restriction on the time gap between two consecutive participations:



C. A. Choquette-Choo, A. Ganesh, M. H. B. McKenna, R., J. K. Rush, A. G. Thakurta, and X. Zheng. (Amplified) banded matrix factorization: A unified approach to private training. In Conference on Neural Information Processing Systems (NeurIPS), 2023.

Approximately Optimal Factorization

Approximately Optimal Factorization

For a workload matrix A we solve optimization problem

$$\arg\min_{S \in \mathcal{S}_+^n} tr[A^T A\, S^{-1}] \quad \text{subject to} \quad \mathrm{diag}(S) = 1 \quad \text{and} \quad S_{[i,j]} = 0 \quad \text{for} \quad |i - j| \geq b.$$

Then, $C^\top C = S$ and $B = AC^{-1}$.

C. A. Choquette-Choo, A. Ganesh, M. H. B. McKenna, R., J. K. Rush, A. G. Thakurta, and X. Zheng. (Amplified) banded matrix factorization: A unified approach to private training. In Conference on Neural Information Processing Systems (NeurIPS), 2023.

Banded Square Root

Lemma (Banded Square-Root Decomposition for Regularized SGD with Momentum)

Let $A_{\alpha,\beta} \in \mathbb{R}^{n\times n}$ be the workload matrix. Then $A_{\alpha,\beta} = B^{|p|}_{\alpha,\beta} C^{|p|}_{\alpha,\beta}$ for

$$C^{|p|}_{\alpha,\beta} = \begin{pmatrix} 1 & 0 & \dots & 0 & 0 & \dots & 0 \\ c_1 & 1 & \dots & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{0} & c_{p-1} & \dots & 1 & 0 & \dots & 0 \\ \boldsymbol{0} & \boldsymbol{0} & \dots & c_1 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \dots & \boldsymbol{0} & c_{p-1} & \dots & 1 \end{pmatrix},$$

for $c_k = \sum_{i=0}^{k} \alpha^{k-i} r_{k-i} r_i \beta^i$ with coefficients $r_k = \left|\binom{-1/2}{k}\right|$. Where $A_{\alpha,\beta} = (C^{|n|}_{\alpha,\beta})^2$.

Matrix Factorization Error

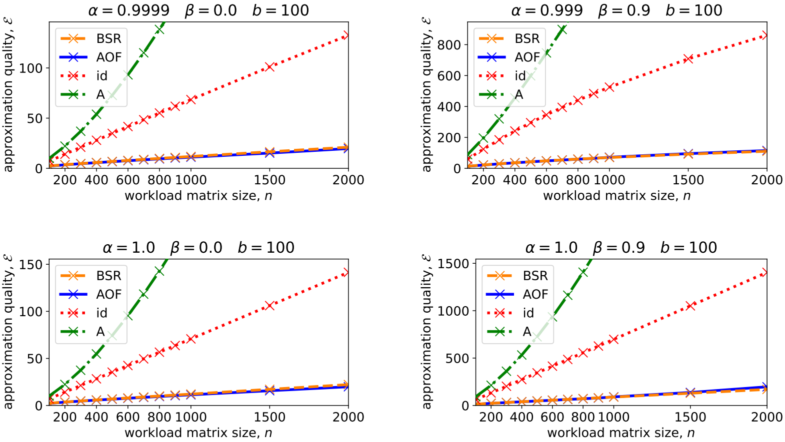
Theorem (Factorization Error in the Setting of Multi Participation)

Setting of Multi Participation] Let $A_{1,\beta} \in \mathbb{R}^{n \times n}$ be the workload matrix of SGD with momentum $0 \leq \beta < 1$. Then, for any $b \in \{1, \ldots, n\}$ it holds that

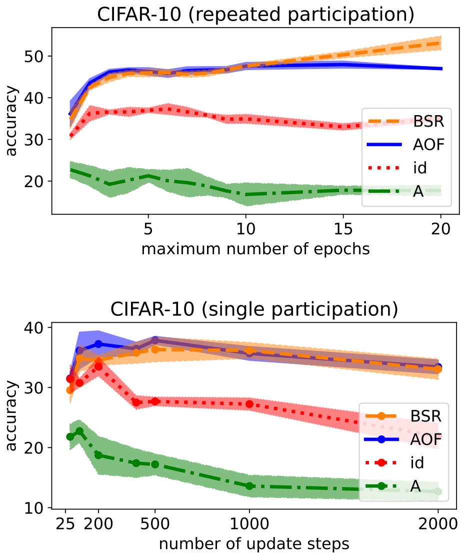
$$\mathcal{E}(B_{1,\beta}^{|p|}, C_{1,\beta}^{|p|}) = O_\beta \left(\sqrt{\frac{kn \log p}{p}} \right) + O_{\beta,p} \left(\sqrt{k} \right)$$

where $k \leq \lceil \frac{n}{b} \rceil$ is the number of participations and $p \leq b$.

Matrix Factorization Numerical Experiments



Matrix Factorization Mechanism for DP Model Training



Summary

Summary

1. Propose Banded Square Root Factorization
2. Derive an explicit and efficient SGDM factorization
3. Analyze sensitivity for decreasing Lower Triangular Toeplitz Matrices
4. Establish upper and lower bounds on matrix factorization error for both multiple and single participation
5. Compare numerically with approximately optimal factorization
6. Train a CIFAR-10 model using the Banded Square Root MF mechanism