DSSGD: A Novel Dual-Stream Stochastic Gradient Descent for Efficient High-Dimensional and Incomplete Matrix Factorization on GPUs Supplementary File

Qicong Hu, *Graduate Student Member, IEEE,* Wen Qin, Hao Wu, *Member, IEEE*, Xin Luo, *Fellow, IEEE*

I. INTRODUCTION

THIS is the supplementary file for paper entitled *"DSSGD: A Novel Dual-Stream Stochastic Gradient Descent for Efficient High-Dimensional and Incomplete Matrix Factorization on GPUs"*. It provides the convergence proof of the DSSGD algorithm and the memory usage comparison.

II. CONVERGENCE PROOF

A. Proof for Lemma 1

Considering two LF vectors $\mathbf{m}_{u1}, \mathbf{m}_{u2} \in \mathbb{R}^f$, we apply the second-order Taylor approximation of $\mathcal{L}_{uv}(\mathbf{m}_{u1})$ around $\mathbf{m}_{u2}$:

$$\begin{aligned}&\mathcal{L}_{uv}\left(\mathbf{m}_{u1}\right)=\mathcal{L}_{uv}\left(\mathbf{m}_{u2}\right)+\nabla\mathcal{L}_{uv}\left(\mathbf{m}_{u2}\right)\left(\mathbf{m}_{u1}-\mathbf{m}_{u2}\right)^T+\frac{1}{2}\left(\mathbf{m}_{u1}-\mathbf{m}_{u2}\right)\nabla^2\mathcal{L}_{uv}\left(\mathbf{m}_{u2}\right)\left(\mathbf{m}_{u1}-\mathbf{m}_{u2}\right)^T\\ \Rightarrow\ &\mathcal{L}_{uv}\left(\mathbf{m}_{u1}\right)-\mathcal{L}_{uv}\left(\mathbf{m}_{u2}\right)=\nabla\mathcal{L}_{uv}\left(\mathbf{m}_{u2}\right)\left(\mathbf{m}_{u1}-\mathbf{m}_{u2}\right)^T+\frac{1}{2}\left(\mathbf{m}_{u1}-\mathbf{m}_{u2}\right)\nabla^2\mathcal{L}_{uv}\left(\mathbf{m}_{u2}\right)\left(\mathbf{m}_{u1}-\mathbf{m}_{u2}\right)^T.\end{aligned}\tag{S1}$$

From **Definition 3**, if $\mathcal{L}_{uv}$ is strongly convex, we obtain the inequality:

$$\mathcal{L}_{uv}\left(\mathbf{m}_{u1}\right)-\mathcal{L}_{uv}\left(\mathbf{m}_{u2}\right)\geq\nabla\mathcal{L}_{uv}\left(\mathbf{m}_{u2}\right)\left(\mathbf{m}_{u1}-\mathbf{m}_{u2}\right)^T+\frac{\mu}{2}\left\|\mathbf{m}_{u1}-\mathbf{m}_{u2}\right\|^2.\tag{S2}$$

Substituting equation (S1) into (S2) yields:

$$\left(\mathbf{m}_{u1}-\mathbf{m}_{u2}\right)\nabla^2\mathcal{L}_{uv}\left(\mathbf{m}_{u2}\right)\left(\mathbf{m}_{u1}-\mathbf{m}_{u2}\right)^T\geq\mu\left\|\mathbf{m}_{u1}-\mathbf{m}_{u2}\right\|^2.\tag{S3}$$

Given that $\nabla^2\mathcal{L}_{uv}(\mathbf{m}_{u2})=\mathbf{n}_v^T\mathbf{n}_v+\lambda I_f$, it follows that:

$$\begin{aligned}&\left(\mathbf{m}_{u1}-\mathbf{m}_{u2}\right)\left(\mathbf{n}_v^T\mathbf{n}_v+\lambda I_f\right)\left(\mathbf{m}_{u1}-\mathbf{m}_{u2}\right)^T\geq\mu\left\|\mathbf{m}_{u1}-\mathbf{m}_{u2}\right\|^2\\ \Rightarrow\ &\left(\mathbf{m}_{u1}-\mathbf{m}_{u2}\right)\left(\mathbf{n}_v^T\mathbf{n}_v+\lambda I_f-\mu I_f\right)\left(\mathbf{m}_{u1}-\mathbf{m}_{u2}\right)^T\geq 0.\end{aligned}\tag{S4}$$

Inequality (S4) is guaranteed provided that $(\mathbf{n}_v^T\mathbf{n}_v+\lambda I_f-\mu I_f)$ remains positive semi-definite. This property is ensured when μ equals the smallest eigenvalue of $(\mathbf{n}_v^T\mathbf{n}_v+\lambda I_f)$. Consequently, this completes the *proof* of **Lemma 1**.

B. Proof of Lemma 2

Considering two LF vectors $\mathbf{m}_{u1}$ and $\mathbf{m}_{u2}$, we have:

$$\begin{aligned}&\nabla\mathcal{L}_{uv}\left(\mathbf{m}_{u1}\right)-\nabla\mathcal{L}_{uv}\left(\mathbf{m}_{u2}\right)=\left(\mathbf{m}_{u1}-\mathbf{m}_{u2}\right)\left(\mathbf{n}_v^T\mathbf{n}_v+\lambda I_f\right)\\ \Rightarrow\ &\left\|\nabla\mathcal{L}_{uv}\left(\mathbf{m}_{u1}\right)-\nabla\mathcal{L}_{uv}\left(\mathbf{m}_{u2}\right)\right\|^2=\left\|\left(\mathbf{m}_{u1}-\mathbf{m}_{u2}\right)\left(\mathbf{n}_v^T\mathbf{n}_v+\lambda I_f\right)\right\|^2\\ &\qquad\leq\left\|\left(\mathbf{m}_{u1}-\mathbf{m}_{u2}\right)\right\|^2\left\|\left(\mathbf{n}_v^T\mathbf{n}_v+\lambda I_f\right)\right\|^2,\end{aligned}\tag{S5}$$

where $\left\|\mathbf{n}_v^T\mathbf{n}_v+\lambda I_f\right\|^2$ represents the greatest eigenvalue of $(\mathbf{n}_v^T\mathbf{n}_v+\lambda I_f)$. Let $L=\left\|\mathbf{n}_v^T\mathbf{n}_v+\lambda I_f\right\|^2$, **Lemma 2** follows.

C. *Proof of Theorem 1*

Based on the M-oriented learning scheme in (6), we obtain:

$$\begin{aligned}\left\|\mathbf{m}_u^{(t)}-\mathbf{m}_u^*\right\|^2&=\left\|\mathbf{m}_u^{(t-1)}-\eta\nabla\mathcal{L}_{uv}\left(\mathbf{m}_u^{(t-1)}\right)-\mathbf{m}_u^*\right\|^2\\&=\left\|\mathbf{m}_u^{(t-1)}-\mathbf{m}_u^*\right\|^2-2\eta\nabla\mathcal{L}_{uv}\left(\mathbf{m}_u^{(t-1)}\right)\left(\mathbf{m}_u^{(t-1)}-\mathbf{m}_u^*\right)^T+\eta^2\left\|\nabla\mathcal{L}_{uv}\left(\mathbf{m}_u^{(t-1)}\right)\right\|^2.\end{aligned}\tag{S6}$$

Through expectation calculation on both sides of (S6):

$$\begin{aligned}\mathbb{E}\left[\left\|\mathbf{m}_u^{(t)}-\mathbf{m}_u^*\right\|^2\right]=\mathbb{E}&\left[\left\|\mathbf{m}_u^{(t-1)}-\mathbf{m}_u^*\right\|^2\right]\\&-2\eta\mathbb{E}\left[\nabla\mathcal{L}_{uv}\left(\mathbf{m}_u^{(t-1)}\right)\left(\mathbf{m}_u^{(t-1)}-\mathbf{m}_u^*\right)^T\right]\\&+\eta^2\mathbb{E}\left[\left\|\nabla\mathcal{L}_{uv}\left(\mathbf{m}_u^{(t-1)}\right)\right\|^2\right].\end{aligned}\tag{S7}$$

According to **Definition 3** and **Lemma 1**, we have:

$$\begin{cases}\mathcal{L}_{uv}\left(\mathbf{m}_u^*\right)\geq\mathcal{L}_{uv}\left(\mathbf{m}_u^{t-1}\right)+\nabla\mathcal{L}_{uv}\left(\mathbf{m}_u^{t-1}\right)\left(\mathbf{m}_u^*-\mathbf{m}_u^{t-1}\right)^T+\frac{\mu}{2}\left\|\mathbf{m}_u^{t-1}-\mathbf{m}_u^*\right\|^2;\\\mathcal{L}_{uv}\left(\mathbf{m}_u^{t-1}\right)\geq\mathcal{L}_{uv}\left(\mathbf{m}_u^*\right)+\nabla\mathcal{L}_{uv}\left(\mathbf{m}_u^*\right)\left(\mathbf{m}_u^{t-1}-\mathbf{m}_u^*\right)^T+\frac{\mu}{2}\left\|\mathbf{m}_u^{t-1}-\mathbf{m}_u^*\right\|^2.\end{cases}\tag{S8}$$

Joining these inequalities leads to:

$$\left[\nabla\mathcal{L}_{uv}\left(\mathbf{m}_u^{t-1}\right)-\nabla\mathcal{L}_{uv}\left(\mathbf{m}_u^*\right)\right]\left(\mathbf{m}_u^{t-1}-\mathbf{m}_u^*\right)^T\geq\mu\left\|\mathbf{m}_u^{t-1}-\mathbf{m}_u^*\right\|^2.\tag{S9}$$

Given that $\mathbf{m}_u^*$ represents the optimal value of $\mathbf{m}_u$, we have:

$$\nabla\mathcal{L}_{uv}\left(\mathbf{m}_u^*\right)=0.\tag{S10}$$

Therefore, (S9) can be rewritten as:

$$\nabla\mathcal{L}_{uv}\left(\mathbf{m}_u^{t-1}\right)\left(\mathbf{m}_u^{t-1}-\mathbf{m}_u^*\right)^T\geq\mu\left\|\mathbf{m}_u^{t-1}-\mathbf{m}_u^*\right\|^2.\tag{S11}$$

Substituting (S11) into (S8) gives:

$$\begin{aligned}\mathbb{E}\left[\left\|\mathbf{m}_u^{(t)}-\mathbf{m}_u^*\right\|^2\right]\leq\mathbb{E}&\left[\left\|\mathbf{m}_u^{(t-1)}-\mathbf{m}_u^*\right\|^2\right]\\&-2\eta\mu\mathbb{E}\left[\left\|\mathbf{m}_u^{t-1}-\mathbf{m}_u^*\right\|^2\right]\\&+\eta^2\mathbb{E}\left[\left\|\nabla\mathcal{L}_{uv}\left(\mathbf{m}_u^{(t-1)}\right)\right\|^2\right].\end{aligned}\tag{S12}$$

Based on **Lemmas 1** and **Lemmas 2**,the subsequent corollary [27] follows:

$$\mathbb{E}\left[\left\|\nabla\mathcal{L}_{uv}\left(\mathbf{m}_u^{(t-1)}\right)\right\|^2\right]\leq G^2.\tag{S13}$$

Substituting (S13) into (S12) yields:

$$\mathbb{E}\left[\left\|\mathbf{m}_u^{(t)}-\mathbf{m}_u^*\right\|^2\right]\leq\left(1-2\eta\mu\right)\mathbb{E}\left[\left\|\mathbf{m}_u^{t-1}-\mathbf{m}_u^*\right\|^2\right]+\eta^2G^2.\tag{S14}$$

Using inductive reasoning over t, we find:

$$\mathbb{E}\left[\left\|\mathbf{m}_u^{(t)}-\mathbf{m}_u^*\right\|^2\right]\leq\left(1-2\eta\mu\right)^{t-1}\mathbb{E}\left[\left\|\mathbf{m}_u^{1}-\mathbf{m}_u^*\right\|^2\right]+\sum_{j=0}^{t-2}\left(1-2\eta\mu\right)^j\left(\eta G\right)^2.\tag{S15}$$

Under **Theorem 1**'s constraint $0<\eta<1/(2\mu)$, it follows that:

$$\sum_{j=0}^{t-2}\left(1-2\eta\mu\right)^j<\sum_{j=0}^{\infty}\left(1-2\eta\mu\right)^j=\frac{1}{2\eta\mu}.\tag{S16}$$

Substituting (S16) into (S15) yields:

$$\mathbb{E}\left[\left\|\mathbf{m}_u^{(t)}-\mathbf{m}_u^*\right\|^2\right] \leq (1-2\eta\mu)^{t-1}\mathbb{E}\left[\left\|\mathbf{m}_u^1-\mathbf{m}_u^*\right\|^2\right]+\frac{\eta G^2}{2\mu}. \tag{S17}$$

Based on the quadratic upper bound property of the L-smoothness function [27], we have:

$$\begin{aligned}\mathcal{L}_{uv}\left(\mathbf{m}_u^t\right)-\mathcal{L}_{uv}\left(\mathbf{m}_u^*\right) &\leq \nabla\mathcal{L}_{uv}\left(\mathbf{m}_u^*\right)\left(\mathbf{m}_u^t-\mathbf{m}_u^*\right)^T+\frac{L}{2}\left\|\mathbf{m}_u^t-\mathbf{m}_u^*\right\|^2\\ &\leq \frac{L}{2}\left\|\mathbf{m}_u^t-\mathbf{m}_u^*\right\|^2.\end{aligned} \tag{S18}$$

By joining (S18) and (S17), it follows that:

$$\begin{aligned}\mathbb{E}\left[\mathcal{L}_{uv}\left(\mathbf{m}_u^t\right)-\mathcal{L}_{uv}\left(\mathbf{m}_u^*\right)\right] &\leq \frac{L}{2}\mathbb{E}\left[\left\|\mathbf{m}_u^t-\mathbf{m}_u^*\right\|^2\right]\\ &\leq \frac{L}{2}\left[(1-2\eta\mu)^{t-1}\mathbb{E}\left[\left\|\mathbf{m}_u^1-\mathbf{m}_u^*\right\|^2\right]+\frac{\eta G^2}{2\mu}\right].\end{aligned} \tag{S19}$$

When the iteration count t is sufficiently large (i.e., as $t\to\infty$):

$$\mathbb{E}\left[\mathcal{L}_{uv}\left(\mathbf{m}_u^t\right)-\mathcal{L}_{uv}\left(\mathbf{m}_u^*\right)\right] \leq \frac{L\eta G^2}{4\mu}. \tag{S20}$$

Let $\Omega(u)$ represents the set of known entries related to node u, then (S20) can be reformulated as:

$$\mathbb{E}\left[\sum_{r_{uv}\in\Lambda(u)}\mathcal{L}_{uv}\left(\mathbf{m}_u^t\right)-\mathcal{L}_{uv}\left(\mathbf{m}_u^*\right)\right] \leq \frac{|\Lambda(u)|\,L\eta G^2}{4\mu}. \tag{S21}$$

For all $u\in U$, equation (S21) can be extended to:

$$\mathbb{E}\left[\sum_{u\in U}\sum_{r_{uv}\in\Lambda(u)}\mathcal{L}_{uv}\left(\mathbf{m}_u^t\right)-\mathcal{L}_{uv}\left(\mathbf{m}_u^*\right)\right] \leq \frac{|\Lambda|\,L\eta G^2}{4\mu}. \tag{S22}$$

This completes the *proof* of **Theorem 1**.

III. Memory Usage Comparison

TABLE S1
GPU Memory Usage Across Datasets

Algorithm	ML25M	ML32M	Yahoo!Music	Netflix
FAMPSGD	1092MB	1282MB	4226MB	2988MB
MSGD	900MB	1038MB	3274MB	2324MB
CUMF_CCD	974MB	1140MB	3450MB	2436MB
DSSGD-C	900MB	1038MB	3274MB	2324MB
DSSGD-D	788MB	888MB	2564MB	1870MB

This table reports the GPU memory usage of DSSGD-C, DSSGD-D, and other GPU-based baselines across various datasets. For DSSGD, we set the batch sizes to 5,000,000 (ML25M), 6,000,000 (ML32M), 30,000,000 (Yahoo!Music), and 20,000,000 (Netflix), such that the number of batches is consistently 4. The results show that DSSGD-C consumes at most 3274MB of GPU memory, which is well within the device's capacity. In this case, introducing batching is unnecessary, as the cost of data transfer may exceed the compute time of each batch, preventing effective overlap. Furthermore, both DSSGD-D and DSSGD-C exhibit the first and second lowest memory usage among all methods, respectively, indicating that DSSGD-D's batch-wise mechanism is effective in reducing memory footprint.