An Efficient, Sparsity-Preserving, Online Algorithm for Low-Rank Approximation

Dave Anderson and Ming Gu

Department of Mathematics, University of California, Berkeley

Background

Goal: low-rank approximations (LRAs) that are

- high-quality
- efficient
- can be continued
- spectrum-preserving
- sparsity-preserving
- can be updated
- highlights features

Approach: use the LU decomposition

$$\Pi_1 \mathbf{A} \Pi_2^T = \begin{matrix} \\ {\scriptstyle m-k} \end{matrix} \overset{\begin{matrix} {\scriptstyle k} & {\scriptstyle m-k} \end{matrix}}{\begin{pmatrix} \mathbf{L}_{11} & \\ \mathbf{L}_{21} & \mathbf{I}_{m-k} \end{pmatrix}} \overset{\begin{matrix} {\scriptstyle k} & {\scriptstyle n-k} \end{matrix}}{\begin{pmatrix} \mathbf{U}_{11} & \mathbf{U}_{12} \\ & \mathbf{S} \end{pmatrix}} \approx \begin{pmatrix} \mathbf{L}_{11} \\ \mathbf{L}_{21} \end{pmatrix} \begin{pmatrix} \mathbf{U}_{11} & \mathbf{U}_{12} \end{pmatrix} \overset{\text{def}}{=} \widehat{\mathbf{L}}\widehat{\mathbf{U}}$$

Problem: the LU decomposition is unstable and both Π_1 and Π_2^T are needed. Computing both is expensive

Solution: use randomization to approximate complete pivoting by cheaply calculating Π_1 and Π_2^T

Previous Work

Deterministic Algorithms

- Truncated SVD, QR, ID, Constrained Nuclear Norm, others

Randomized Algorithms for Faster LRA:

- Black Box Algorithms [2, 6]

$$\mathbf{QR} = \mathbf{A}\Omega$$
$$\mathbf{A} \approx \mathbf{Q}^T\mathbf{QA}$$

- Sampling Algorithms [4]

$$\mathbf{A} \approx \mathbf{C}\left(\mathbf{C}^\dagger \mathbf{A} \mathbf{R}^\dagger\right)\mathbf{R}$$

- Complexity:

$$O\left(\text{nnz}\left(\mathbf{A}\right)\right) + n \cdot \text{poly}\left(\frac{k}{\epsilon}\right)$$

- Error bound:

$$\left\|\mathbf{A} - \widehat{\mathbf{A}}\right\| \leq (1+\epsilon)\left\|\mathbf{A} - \mathbf{A}_k\right\|$$

...but $\epsilon \not\to 0$ because of the complexity

Truncated LU Theory

General Truncated LU Bounds

Theorem

Let $(\cdot)_s$ denote the rank-s truncated SVD for $s \leq k \ll m, n$. Then for any truncated LU factorization with Schur complement $\mathbf{S}$:

$$\left\|\Pi_1 \mathbf{A} \Pi_2^T - \widehat{\mathbf{L}}\widehat{\mathbf{U}}\right\| = \|\mathbf{S}\|$$

for any norm, and

$$\left\|\Pi_1 \mathbf{A} \Pi_2^T - \left(\widehat{\mathbf{L}}\widehat{\mathbf{U}}\right)_s\right\|_2 \leq 2\|\mathbf{S}\|_2 + \sigma_{s+1}\left(\mathbf{A}\right).$$

Theorem

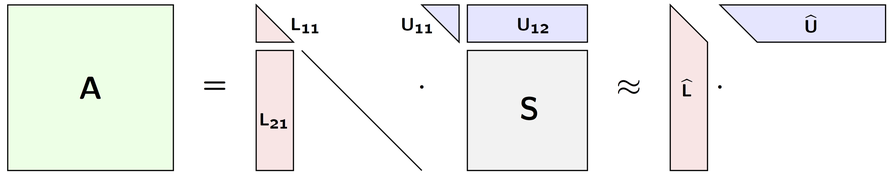
For a general rank-k truncated LU decomposition, we have for all $1 \leq j \leq k$,

$$\sigma_j\left(\mathbf{A}\right) \leq \sigma_j\left(\widehat{\mathbf{L}}\widehat{\mathbf{U}}\right)\left(1 + \left(1 + \frac{\|\mathbf{S}\|_2}{\sigma_k\left(\widehat{\mathbf{L}}\widehat{\mathbf{U}}\right)}\right)\frac{\|\mathbf{S}\|_2}{\sigma_j\left(\mathbf{A}\right)}\right).$$

Theorem

$$\left\|\Pi_1 \mathbf{A} \Pi_2^T - \widehat{\mathbf{L}}\mathbf{M}\widehat{\mathbf{U}}\right\|_2 \leq 2\|\mathbf{S}\|_2$$
$$\left\|\Pi_1 \mathbf{A} \Pi_2^T - \widehat{\mathbf{L}}\mathbf{M}\widehat{\mathbf{U}}\right\|_F \leq \|\mathbf{S}\|_F.$$

The Truncated LU Factorization



Spectrum-Revealing LU (SRLU)

SRLU = TRLUCP + SRP

Algorithm 1 Truncated Randomized LU with Complete Pivoting (TRLUCP)

Inputs: Data matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$, target rank k, block size b, oversampling parameter $p \geq b$, random Gaussian matrix $\Omega \in \mathbb{R}^{p \times m}$, $\widehat{\mathbf{L}}$ and $\widehat{\mathbf{U}}$ are initially 0 matrices

Calculate random projection $\mathbf{R} = \Omega\mathbf{A}$

for $j = 0, b, 2b, \cdots, k-b$ **do**

Perform column selection algorithm on $\mathbf{R}$ and swap columns of $\mathbf{A}$

Update block column of $\widehat{\mathbf{L}}$

Perform block LU with partial row pivoting and swap rows of $\mathbf{A}$

Update block row of $\widehat{\mathbf{U}}$

Update $\mathbf{R}$

end for

How did we do?

Let α be the largest element in the Schur complement:

$$\Pi_1 \mathbf{A} \Pi_2^T = \widehat{\mathbf{L}}\widehat{\mathbf{U}} + \begin{pmatrix} 0 & 0 \\ 0 & \mathbf{S} \end{pmatrix} = \begin{pmatrix} \mathbf{L}_{11} \\ \ell^T \\ \mathbf{L}_{31} \end{pmatrix} \begin{pmatrix} \mathbf{U}_{11} & u & \mathbf{U}_{13} \end{pmatrix} + \begin{pmatrix} 0 & 0 & \\ 0 & \alpha & s_{12}^T \\ & s_{21} & \mathbf{S}_{22} \end{pmatrix}$$

Compare what we have to what we missed:

Define: $\overline{\mathbf{A}}_{11} \overset{\text{def}}{=} \begin{pmatrix} \mathbf{L}_{11} \\ \ell^T & 1 \end{pmatrix} \begin{pmatrix} \mathbf{U}_{11} & u \\ & \alpha \end{pmatrix}$ and test: $\left\|\left(\overline{\mathbf{A}}_{11}\right)^{-1}\right\|_{\text{MAX}} \overset{?}{\leq} \frac{f}{|\alpha|}$

Rewrite as an algorithm:

Algorithm 2 Spectrum-Revealing Pivoting (SRP)

Input: Truncated LU factorization $\mathbf{A} \approx \widehat{\mathbf{L}}\widehat{\mathbf{U}}$, tolerance $f > 1$

while $\|\overline{\mathbf{A}}_{11}^{-1}\|_{\max} > \frac{f}{|\alpha|}$ **do**

Set α to be the largest element in $\mathbf{S}$ (or find an approximate α using $\mathbf{R}$)

Swap row and column containing α with row and column of largest element in $\overline{\mathbf{A}}_{11}^{-1}$

Update truncated LU factorization

end while

Improving Quality

Use a CUR technique:

$$\Pi_1 \mathbf{A} \Pi_2^T \approx \widehat{\mathbf{L}} \underbrace{\left(\widehat{\mathbf{L}}^\dagger \Pi_1 \mathbf{A} \Pi_2^T \widehat{\mathbf{U}}^\dagger\right)}_{\mathbf{M}} \widehat{\mathbf{U}}$$

$\mathbf{M}$ is dense, but small

Online Updating

Begin with an SRLU factorization:

$$\Pi_1 \mathbf{A} \Pi_2^T = \begin{pmatrix} \mathbf{L}_{11} \\ \mathbf{L}_{21} & \mathbf{I} \end{pmatrix} \begin{pmatrix} \mathbf{U}_{11} & \mathbf{U}_{12} \\ & \mathbf{S} \end{pmatrix}$$

Receive new data $\mathbf{B}\Pi_2^T = \begin{pmatrix} \mathbf{B}_1 & \mathbf{B}_2 \end{pmatrix}$ and combine:

$$\begin{pmatrix} \Pi_1 \mathbf{A} \Pi_2^T \\ \mathbf{B}\Pi_2^T \end{pmatrix} = \begin{pmatrix} \mathbf{L}_{11} & & \\ \mathbf{L}_{21} & \mathbf{I} & \\ \mathbf{B}_1\mathbf{U}_{11}^{-1} & & \mathbf{I} \end{pmatrix} \begin{pmatrix} \mathbf{U}_{11} & \mathbf{U}_{12} \\ & \mathbf{S} \\ & \mathbf{B}_2 - \mathbf{B}_1\mathbf{U}_{11}^{-1}\mathbf{U}_{12} \end{pmatrix}$$

Perform Spectrum-Revealing Pivoting

Theoretical Results

SRLU Bounds

Theorem

For $j \leq k$ and $\gamma = O\left(fk\sqrt{mn}\right)$, SRP produces a rank-$k$ SRLU factorization with

$$\|\Pi_1 \mathbf{A} \Pi_2^T - \widehat{\mathbf{L}}\widehat{\mathbf{U}}\|_2 \leq \gamma\sigma_{k+1}\left(\mathbf{A}\right),$$
$$\|\Pi_1 \mathbf{A} \Pi_2^T - \left(\widehat{\mathbf{L}}\widehat{\mathbf{U}}\right)_j\|_2 \leq \sigma_{j+1}\left(\mathbf{A}\right)\left(1 + 2\gamma\frac{\sigma_{k+1}(\mathbf{A})}{\sigma_{j+1}(\mathbf{A})}\right)$$

Theorem

For $1 \leq j \leq k$, SRP produces a rank-k SRLU factorization with

$$\frac{\sigma_j\left(\mathbf{A}\right)}{1 + \tau\frac{\sigma_{k+1}(\mathbf{A})}{\sigma_j(\mathbf{A})}} \leq \sigma_j\left(\widehat{\mathbf{L}}\widehat{\mathbf{U}}\right) \leq \sigma_j\left(\mathbf{A}\right)\left(1 + \tau\frac{\sigma_{k+1}\left(\mathbf{A}\right)}{\sigma_j\left(\mathbf{A}\right)}\right)$$

for $\tau \leq O\left(mnk^2 f^3\right)$.

Theorem

$$\left\|\Pi_1 \mathbf{A} \Pi_2^T - \widehat{\mathbf{L}}\mathbf{M}\widehat{\mathbf{U}}\right\|_2 \leq 2\gamma\sigma_{k+1}\left(\mathbf{A}\right)$$
$$\left\|\Pi_1 \mathbf{A} \Pi_2^T - \widehat{\mathbf{L}}\mathbf{M}\widehat{\mathbf{U}}\right\|_F \leq \omega\sigma_{k+1}\left(\mathbf{A}\right),$$

where $\gamma = O\left(fk\sqrt{mn}\right)$ is the same as in Theorem 4, and $\omega = O\left(fkmn\right)$.

Theorem

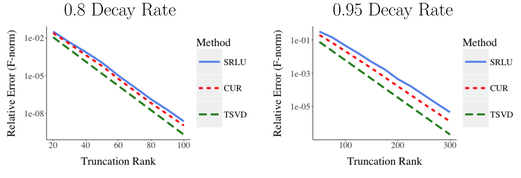
If $\sigma_j^2\left(\mathbf{A}\right) > 2\|\mathbf{S}\|_2^2$ then

$$\sigma_j\left(\mathbf{A}\right) \geq \sigma_j\left(\widehat{\mathbf{L}}\mathbf{M}\widehat{\mathbf{U}}\right) \geq \sigma_j\left(\mathbf{A}\right)\sqrt{1 - 2\gamma\left(\frac{\sigma_{k+1}\left(\mathbf{A}\right)}{\sigma_j\left(\mathbf{A}\right)}\right)^2}$$

for $\gamma = O\left(mnk^2 f^2\right)$ and f is an input parameter controlling a tradeoff of quality vs. speed as before.

Experiments

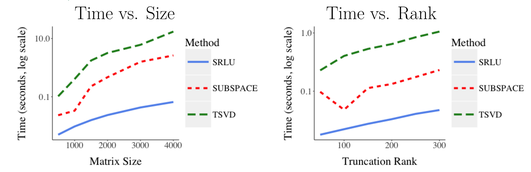
Accuracy:



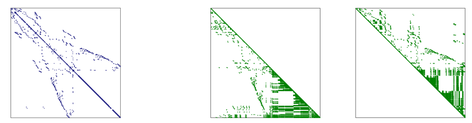
Roughly a constant amount of oversampling is needed to match the accuracy of the Truncated SVD

Experiments (Continued)

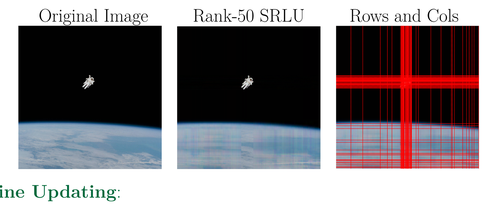
Efficiency:



Sparsity Preservation:



Feature Selection:



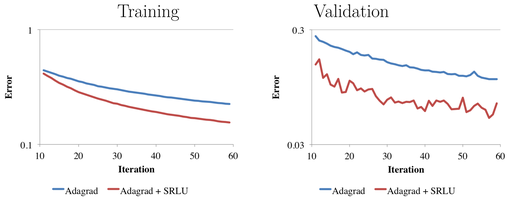
Online Updating:

SRLU finds top words in the Enron email corpus when processed in blocks

Going Forward

Efficient Optimizers: Approximation of Full Adagrad

- Based on recent work by Krummenacher et. al. [3]
- Use SRLU now (column pivots applied to rows too)
- Symmetric implementation
- Neural net with 3 Convolutional layers, approximately 33,000 parameters



References

[1] Clarkson K. L. and Woodruff, D. P., Low Rank Approximation and Regression in Input Sparsity Time, CoRR, abs/1207.6365, 2012.

[2] Gu, M. Subspace Iteration Randomization and Singular Value Problems, SIAM J. Scientific Computing, 37(3), 2015.

[3] Krummenacher, G. and McWilliams, B., Kilcher, Y., Buhmann, J. M., and Meinshausen, N, Scalable Adaptive Stochastic Optimization Using Random Projections, NIPS, pp. 1750-1758, 2016.

[4] Mahoney, M. W. and Drineas, P., CUR Matrix Decompositions for Improved Data Analysis, PNAS, 106(3), pp. 697-702, 2009.

[5] Melgaard, C. and Gu, M., Gaussian Elimination with Randomized Complete Pivoting, CoRR, abs/1511.08528, 2015.

[6] Sarlós, T., Improved Approximation Algorithms for Large Matrices Via Random Projections, FOCS, pp. 143-152, 2006.