Reweighted Solutions for Weighted Low Rank Approximation

David P. Woodruff and Taisuke Yasuda

Carnegie Mellon University
Computer Science Department

Weighted Low Rank Approximation

- **Low rank approximation (LRA)**: given $\mathbf{A} \in \mathbb{R}^{n \times d}$, solve

$$\min \|\mathbf{A} - \tilde{\mathbf{A}}\|_F^2 = \sum_{i=1}^{n} \sum_{j=1}^{d} |\mathbf{A}_{i,j} - \tilde{\mathbf{A}}_{i,j}|^2 \qquad \text{s.t. } \operatorname{rank}(\tilde{\mathbf{A}}) \leq k$$

- **Weighted LRA (WLRA)**: given $\mathbf{A} \in \mathbb{R}^{n \times d}$ and weights $\mathbf{W} \in \mathbb{R}^{n \times d}$, solve

$$\min \|\mathbf{A} - \tilde{\mathbf{A}}\|_{\mathbf{W},F}^2 = \sum_{i=1}^{n} \sum_{j=1}^{d} \mathbf{W}_{i,j}^2 \cdot |\mathbf{A}_{i,j} - \tilde{\mathbf{A}}_{i,j}|^2 \qquad \text{s.t. } \operatorname{rank}(\tilde{\mathbf{A}}) \leq k$$

Entrywise weights $\mathbf{W}$

- **Advantages of introducing weights**
  - Model missing entries by setting $\mathbf{W}_{i,j} = 0$ (aka matrix completion)
  - Model uncertainty of an entry
- **Disadvantages of introducing weights**
  - Problem becomes NP hard, even to approximate

> **Goal**. Design efficient bicriteria approximation algorithms for WLRA.

- We focus on **low rank weight matrices** (i.e. $\operatorname{rank}(\mathbf{W}) \leq r$)
  - This assumption is necessary in general [RWZ16]
    - There exist input instances with rank r that require $\exp(r)$ time under natural complexity assumptions
  - This seems to be a decent assumption in practice

A Simple New Algorithm

> **Algorithm: WLRA**.
>
> 1. Compute a rank rk LRA $\tilde{\mathbf{A}}_{\mathbf{W}}$ of $\mathbf{W} \circ \mathbf{A}$
> 2. Return $\tilde{\mathbf{A}} := \mathbf{W}^{\circ -1} \circ \tilde{\mathbf{A}}_{\mathbf{W}}$

Entrywise product $\circ$

Reweight by the entrywise inverse $\mathbf{W}^{\circ -1}$

- This solution is a **reweighting** of a rank rk solution, but not low rank itself
- Efficient storage: $\mathbf{W}$ and $\tilde{\mathbf{A}}_{\mathbf{W}}$ are both low rank
- Efficient application: if $\mathbf{W}$ is structured

Proof

> **Lemma**. If $\operatorname{rank}(\mathbf{W}) \leq r$ and $\operatorname{rank}(\tilde{\mathbf{A}}) \leq k$, then $\operatorname{rank}(\mathbf{W} \circ \tilde{\mathbf{A}}) \leq rk$.

- **Proof:** We have

$$\begin{aligned} \mathbf{W} \circ \tilde{\mathbf{A}} &= \left(\sum_{i=1}^{r} \mathbf{u}_i \mathbf{v}_i^\top\right) \circ \left(\sum_{j=1}^{k} \mathbf{b}_j \mathbf{c}_j^\top\right) \\ &= \sum_{i=1}^{r} \sum_{j=1}^{k} \left(\mathbf{u}_i \mathbf{v}_i^\top\right) \circ \left(\mathbf{b}_j \mathbf{c}_j^\top\right) \\ &= \sum_{i=1}^{r} \sum_{j=1}^{k} \left(\mathbf{u}_i \circ \mathbf{b}_i\right) \left(\mathbf{v}_j \circ \mathbf{c}_j\right)^\top \end{aligned}$$

> **Theorem**. The algorithm outputs a solution with cost at most the optimal rank k solution.

- **Proof:** For any rank k matrix $\mathbf{A}'$, we have

$$\begin{aligned} \|\mathbf{A} - \mathbf{W}^{\circ -1} \circ \tilde{\mathbf{A}}_{\mathbf{W}}\|_{\mathbf{W},F}^2 &= \|\mathbf{W} \circ \mathbf{A} - \tilde{\mathbf{A}}_{\mathbf{W}}\|_F^2 \\ &\leq \|\mathbf{W} \circ \mathbf{A} - \mathbf{W} \circ \mathbf{A}'\|_F^2 \\ &= \|\mathbf{A} - \mathbf{A}'\|_{\mathbf{W},F}^2 \end{aligned}$$

Rank at most rk by lemma

Matrices with Structured Entrywise Inverses

- The entrywise inverse is unfavorable…
- This can be fixed for structured matrices!

> **Lemma**. If $\operatorname{rank}(\mathbf{A}) \leq k$ and $\mathbf{W} = \mathbf{E} + \sum_{i=1}^{r'} \mathbf{S}_i$ for a sparse matrix $\mathbf{E}$ and rank 1 matrices $\mathbf{S}_i$ with disjoint support, then $\mathbf{W}^{\circ -1} \circ \mathbf{A}'$ can be applied quickly to a vector.

- **Proof:** We can write $\mathbf{W}^{\circ -1} = \mathbf{E}' + \sum_{i=1}^{r} \mathbf{S}_i^{\circ -1}$, where $\mathbf{E}'$ is a sparse matrix that has the same support as $\mathbf{E}$. Then,

$$(\mathbf{W}^{\circ -1} \circ \mathbf{A}')\mathbf{x} = (\mathbf{E}' \circ \mathbf{A}')\mathbf{x} + \sum_{i=1}^{r} (\mathbf{S}_i^{\circ -1} \circ \mathbf{A}')\mathbf{x}$$

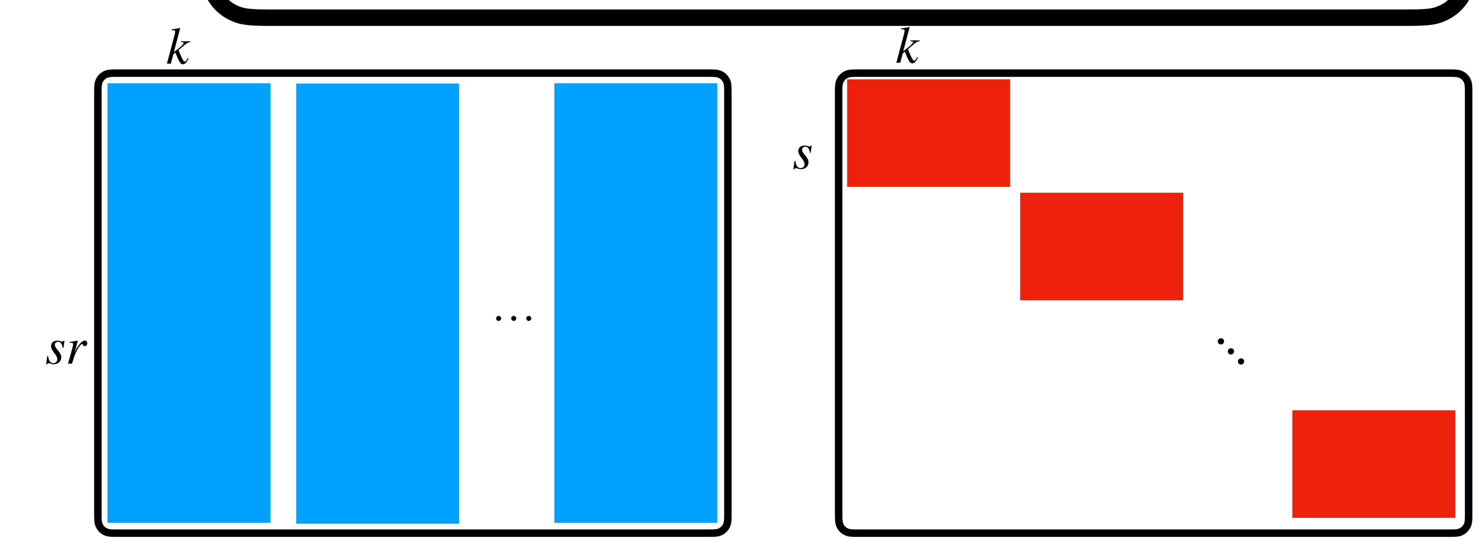
Sparse

Low rank

Communication Complexity

- **Column subset selection**: there is a set of $\tilde{O}(k/\varepsilon)$ columns that span a $(1+\varepsilon)$ relative error LRA
- If the weight matrix $\mathbf{W}$ has s-sparse columns, then by using LRA based on column subset selection, we can represent a WLRA solution in roughly $srk + rkd$ space.
- When $d \ll s \leq \frac{n}{r}$, this is nearly optimal!

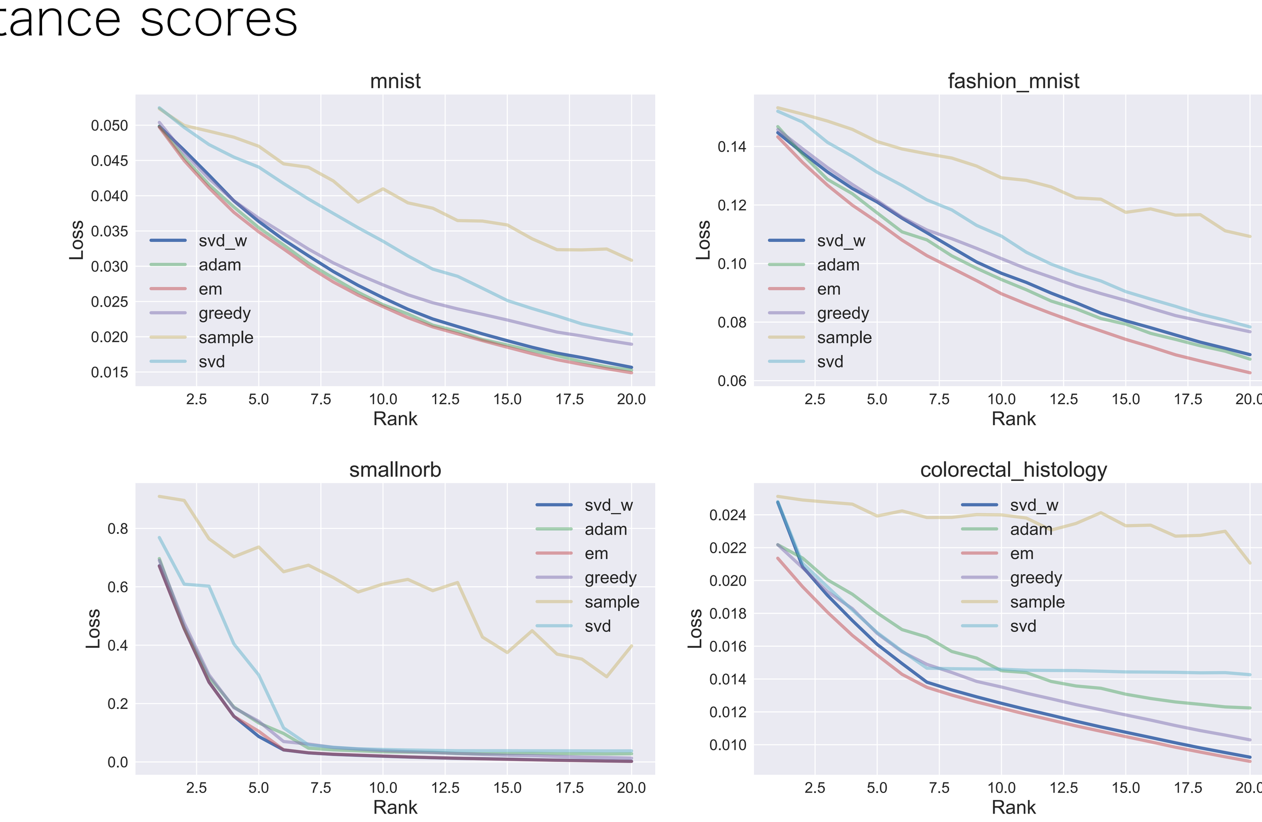
> **Theorem**. Representing a relative error WLRA solution for a matrix $\mathbf{W}$ with s-sparse columns requires $\Omega(srk)$ bits of space.



$\mathbf{A}$ is r copies of a $sr \times k$ matrix $\mathbf{W}$ is r copies of $s \times k$ blocks

Experiments

- Real world dataset: DNN weight matrices, weighted by importance scores



- Synthetic dataset: mixture of Gaussians, weighted by inverse variance