Reweighted Solutions for Weighted Low Rank Approximation

David P. Woodruff [1] **Taisuke Yasuda** [1]

Abstract

Weighted low rank approximation (WLRA) is an important yet computationally challenging primitive with applications ranging from statistical analysis, model compression, and signal processing. To cope with the NP-hardness of this problem, prior work considers heuristics, bicriteria, or fixed parameter tractable algorithms to solve this problem. In this work, we introduce a new relaxed solution to WLRA which outputs a matrix that is not necessarily low rank, but can be stored using very few parameters and gives provable approximation guarantees when the weight matrix has low rank. Our central idea is to use the weight matrix itself to reweight a low rank solution, which gives an extremely simple algorithm with remarkable empirical performance in applications to model compression and on synthetic datasets. Our algorithm also gives nearly optimal communication complexity bounds for a natural distributed problem associated with this problem, for which we show matching communication lower bounds. Together, our communication complexity bounds show that the rank of the weight matrix provably parameterizes the communication complexity of WLRA. We also obtain the first relative error guarantees for feature selection with a weighted objective.

1. Introduction

The approximation of matrices by matrices of lower rank has been, and continues to be, one of the most intensely studied and applied computational problems in statistics, machine learning, and signal processing. The classical approach to this problem is to approximate a matrix $\mathbf{A} \in \mathbb{R}^{n\times d}$ by a rank k matrix $\tilde{\mathbf{A}} \in \mathbb{R}^{n\times d}$ that minimizes the Frobenius norm error

$$\|\mathbf{A}-\tilde{\mathbf{A}}\|_F^2 \coloneqq \sum_{i=1}^{n}\sum_{j=1}^{d}|\mathbf{A}_{i,j}-\tilde{\mathbf{A}}_{i,j}|^2,$$

where $\mathrm{rank}(\tilde{\mathbf{A}}) \le k$. This problem can be solved exactly by the singular value decomposition (SVD), which can be computed in polynomial time. We will write $\mathbf{A}_k$ to denote the optimal rank k approximation to $\mathbf{A}$ in the Frobenius norm, and we will write $\mathbf{A}_{-k} \coloneqq \mathbf{A}-\mathbf{A}_k$ to denote the residual error of this approximation.

While this simple choice often gives satisfactory results, this loss function treats all entries of the matrix uniformly when trying to fit $\tilde{\mathbf{A}}$, which may not exactly align with the practitioner's desires if some of the entries are more crucial to fit than others. If one additionally has such information available in the form of non-negative weights $\mathbf{W}_{i,j} \ge 0$ that reflect some measure of importance of each of the entries (i,j), then this can be encoded in the loss function by incorporating weights as

$$\|\mathbf{A}-\tilde{\mathbf{A}}\|_{\mathbf{W},F}^2 \coloneqq \sum_{i=1}^{n}\sum_{j=1}^{d}\mathbf{W}_{i,j}^2\cdot|\mathbf{A}_{i,j}-\tilde{\mathbf{A}}_{i,j}|^2,$$

where $\mathrm{rank}(\tilde{\mathbf{A}}) \le k$. This problem is known as the *weighted low rank approximation (WLRA)* problem. We write $\mathbf{A}\circ\mathbf{B}$ to denote the entrywise product for two matrices $\mathbf{A}$ and $\mathbf{B}$, so we may also write

$$\|\mathbf{A}-\tilde{\mathbf{A}}\|_{\mathbf{W},F}^2 \coloneqq \|\mathbf{W}\circ(\mathbf{A}-\tilde{\mathbf{A}})\|_F^2 = \|\mathbf{W}\circ\mathbf{A}-\mathbf{W}\circ\tilde{\mathbf{A}}\|_F^2$$

The incorporation of weights into the low rank approximation problem gives this computational problem an incredible versatility for use in a long history of applications starting with its use in factor analysis in the early statistical literature (Young, 1941). A popular special case is the *matrix completion* problem (Rennie & Srebro, 2005; Candès & Tao, 2010; Keshavan et al., 2010), where the weights $\mathbf{W} \in \{0,1\}^{n\times d}$ are binary and encode whether a given entry of $\mathbf{A}$ is observed or not. This primitive has been useful in the design of recommender systems (Koren et al., 2009; Chen et al., 2015; Lee et al., 2016), and has been famously applied in the 2006 Netflix Prize problem. More generally, the weights $\mathbf{W}$ can be used to reflect the variance or number of samples

*Equal contribution [1]School of Computer Science, Carnegie Mellon University, Pittsburgh, PA. Correspondence to: David P. Woodruff <dwoodruf@cs.cmu.edu>, Taisuke Yasuda <taisukey@cs.cmu.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

obtained for each of the entries, so that more "uncertain" entries can influence the objective function less (Anandan & Irani, 2002; Srebro & Jaakkola, 2003). In the past few years, weighted low rank approximation has also been used to improve model compression algorithms, especially those for large scale LLMs, based on low rank approximations of weight matrices by taking into account the importance of parameters (Arora et al., 2016; Hsu et al., 2022; Hua et al., 2022). Given the rapid growth of large scale machine learning models, model compression techniques such as WLRA are expected to bring high value to engineering efforts for these models. Other applications of WLRA include ecology (Robin et al., 2019; Kidzinski et al., 2022), background modeling (Li et al., 2017; Dutta et al., 2018), computational biology (Tuzhilina et al., 2022), and signal processing (Shpak, 1990; Lu et al., 1997).

Approximation algorithms have long been considered for efficient low rank approximation problems, and we formalize the approximation guarantee that we study in Definition 1.1.

Definition 1.1 (Approximate weighted low rank approximation)**.** Let $\mathbf{W} \in \mathbb{R}^{n\times d}$ be non-negative, let $\mathbf{A} \in \mathbb{R}^{n\times d}$, and let $k \in \mathbb{N}$. Then in the *κ-approximate rank k weighted low rank approximation* problem, we seek to output a matrix $\tilde{\mathbf{A}} \in \mathbb{R}^{n\times d}$ such that

$$\|\mathbf{A} - \tilde{\mathbf{A}}\|_{\mathbf{W},F} \leq \kappa \min_{\mathrm{rank}(\mathbf{A}')\leq k} \|\mathbf{A} - \mathbf{A}'\|_{\mathbf{W},F}.$$

In Definition 1.1, we have purposefully under-specified requirements on $\tilde{\mathbf{A}}$. Part of this is to cope with the computational difficulty of WLRA. Indeed, while we ideally would like $\tilde{\mathbf{A}}$ to have rank at most k, solving for even an approximate such solution (with $\kappa = (1 + 1/\operatorname{poly}(n))$) is an NP-hard problem (Gillis & Glineur, 2011). On the other hand, allowing for additional flexibility in the choice of $\tilde{\mathbf{A}}$ may still be useful as long as $\tilde{\mathbf{A}}$ satisfies some sort of "parameter reduction" guarantee. A common choice is to allow $\tilde{\mathbf{A}}$ to have rank $k' \geq k$ slightly larger than k, which is known as a *bicriteria* guarantee. In this work, we will show a new relaxation of the constraints on $\tilde{\mathbf{A}}$ that allows us to achieve new approximation guarantees for WLRA.

1.1. Our results

We present our main contribution in Theorem 1.2, which gives a simple approach to WLRA, under the assumption that the weight matrix $\mathbf{W}$ has low rank $\mathrm{rank}(\mathbf{W}) \leq r$. We note that this assumption is very natural and captures natural cases, for example when $\mathbf{W}$ has block structure, and has been motivated and studied in prior work (Razenshteyn et al., 2016; Ban et al., 2019). We also empirically verify this assumption in our experiments. We defer a further discussion of the low rank $\mathbf{W}$ assumption to Section 1.1.3 as well as prior works (Razenshteyn et al., 2016; Ban et al., 2019).

The algorithm (shown in Algorithm 1) that we propose is extremely simple: compute a rank rk approximation of $\mathbf{W} \circ \mathbf{A}$, and then divide the result entrywise by $\mathbf{W}$. Note that if we exactly compute the low rank approximation step by an SVD, then the optimal rank rk approximation $(\mathbf{W} \circ \mathbf{A})_{rk}$ given by the SVD requires only $(n + d)rk$ parameters to store, and $\mathbf{W}$ also only requires nr parameters to store. Thus, denoting the entrywise inverse of $\mathbf{W}$ by $\mathbf{W}^{\circ -1}$, the solution $\mathbf{W}^{\circ -1} \circ (\mathbf{W} \circ \mathbf{A})_{rk}$ can be stored in a total of $O((n + d)rk)$ parameters, which is nearly optimal for constant rank $r = O(1)$.[1]

Algorithm 1 Weighted low rank approximation

input: input matrix $\mathbf{A} \in \mathbb{R}^{n\times d}$, non-negative weights $\mathbf{W} \in \mathbb{R}^{n\times d}$ with rank r, rank parameter k.
output: approximate solution $\tilde{\mathbf{A}}$.

1: Compute a rank rk approximation $\tilde{\mathbf{A}}_{\mathbf{W}}$ of $\mathbf{W} \circ \mathbf{A}$
2: Return $\tilde{\mathbf{A}} := \mathbf{W}^{\circ -1} \circ \tilde{\mathbf{A}}_{\mathbf{W}}$

While our discussion thus far has simply used the SVD to compute the rank rk approximation $(\mathbf{W} \circ \mathbf{A})_{rk}$, we obtain other useful guarantees by allowing for approximate solutions $\tilde{\mathbf{A}}_{\mathbf{W}}$ that only approximately minimize $\|\mathbf{W} \circ \mathbf{A} - \tilde{\mathbf{A}}_{\mathbf{W}}\|_F$. For example, by computing the rank rk approximation $\tilde{\mathbf{A}}_{\mathbf{W}}$ using faster randomized approximation algorithms for the SVD (Clarkson & Woodruff, 2013; Musco & Musco, 2015; Avron et al., 2017), we obtain algorithms for WLRA with similar running time. In general, we prove the following theorem:

Theorem 1.2. *Let $\mathbf{W} \in \mathbb{R}^{n\times d}$ be a non-negative weight matrix with rank r. Let $\mathbf{A} \in \mathbb{R}^{n\times d}$ and let $k \in \mathbb{N}$. Suppose that $\tilde{\mathbf{A}}_{\mathbf{W}} \in \mathbb{R}^{n\times d}$ satisfies*

$$\begin{aligned}\|\mathbf{W} \circ \mathbf{A} - \tilde{\mathbf{A}}_{\mathbf{W}}\|_F^2 &\leq \kappa \min_{\mathrm{rank}(\mathbf{A}')\leq rk} \|\mathbf{W} \circ \mathbf{A} - \mathbf{A}'\|_F^2 \\ &= \kappa \|(\mathbf{W} \circ \mathbf{A})_{-rk}\|_F^2\end{aligned}$$

and let $\tilde{\mathbf{A}} := \mathbf{W}^{\circ -1} \circ \tilde{\mathbf{A}}_{\mathbf{W}}$, where $\mathbf{W}^{\circ -1} \in \mathbb{R}^{n\times d}$ denotes the entrywise inverse of $\mathbf{W}$. Then,

$$\|\mathbf{A} - \tilde{\mathbf{A}}\|_{\mathbf{W},F}^2 \leq \kappa \min_{\mathrm{rank}(\mathbf{A}')\leq k} \|\mathbf{A} - \mathbf{A}'\|_{\mathbf{W},F}^2$$

In particular, we obtain a solution with $\kappa = (1 + \varepsilon)$ in running time $O(\mathsf{nnz}(\mathbf{A})) + \tilde{O}(n(rk)^2/\varepsilon + \operatorname{poly}(rk/\varepsilon))$ by using randomized low rank approximation algorithms of (Avron et al., 2017).

We prove Theorem 1.2 in Section 2. In the special case of binary weight matrices, our result shows that "zero-filling"

[1] If $\mathbf{W}_{i,j} = 0$, we take $\mathbf{W}_{i,j}^{\circ -1} = \infty$. Note that this entry is ignored by the cost, i.e., $(\mathbf{W} \circ \mathbf{W}^{\circ -1})_{i,j} = 0$ by convention.

the missing entries leads to relative error guarantees in this natural setting, which is perhaps surprising due to a number of works studying this algorithm that suggest otherwise (Balzano et al., 2010; Wang & Singh, 2017).

Note that as stated, the approximation given by Algorithm 1 may not always be desirable, since in general, $\mathbf{W}^{\circ -1}$ cannot be computed without multiplying out the low rank factors of $\mathbf{W}$. However, we show in Lemma 1.3 that for a broad family of structured matrices formed by the sum of support-disjoint rank 1 matrices and a sparse matrix, $\mathbf{W}^{\circ -1}$ can in fact be stored and applied in the same time as $\mathbf{W}$. These capture a large number of commonly used weight matrix patterns in practice, such as Low-Rank Plus Sparse, Low-Rank Plus Diagonal, Low-Rank Plus Block Diagonal, Monotone Missing Data Pattern, and Low-Rank Plus Banded matrices (Musco et al., 2021) (see Corollary 1.4). These results are proved in Appendix 3.

Lemma 1.3. *Let $\mathbf{W} \in \mathbb{R}^{n\times d}$ be structured as $\mathbf{W} = \mathbf{E} + \sum_{i=1}^{r'} \mathbf{S}_i$, where $\mathbf{S}_i$ are rank 1 matrices with disjoint rectangular supports $S_i \times T_i$ for $S_i \subseteq [n]$ and $T_i \subseteq [d]$, and $\mathbf{E}$ is a sparse matrix with $\mathsf{nnz}(\mathbf{E})$ non-zero entries. Let $\mathbf{A} = \mathbf{U}\mathbf{V}$ for $\mathbf{U} \in \mathbb{R}^{n\times k}$ and $\mathbf{V} \in \mathbb{R}^{k\times d}$ be a rank k matrix. Then, $\mathbf{W}^{\circ -1} \circ \mathbf{A}$ can be stored in $O(\mathsf{nnz}(\mathbf{E}) + \sum_{i=1}^{r'} |S_i| + |T_i|)$ space and can be applied to a vector $\mathbf{x} \in \mathbb{R}^d$ in $O(\mathsf{nnz}(\mathbf{E}) + \sum_{i=1}^{r'} |S_i| + |T_i|)$ time. Furthermore, $\mathbf{W}$ has rank at most $r = \mathsf{nnz}(\mathbf{E}) + r'$.*

Corollary 1.4. *Let $\mathbf{W} \in \mathbb{R}^{n\times d}$. The following hold:*

- ***Low-Rank Plus Sparse**: Suppose that $\mathbf{W}$ has at most t non-zeros per row. Then, $\mathbf{W}^{\circ -1}$ can be stored and applied in $O(nt)$ time and space.*
- ***Low-Rank Plus Diagonal**: Suppose that $\mathbf{W}$ is all ones except for zeros along the diagonal. Then, $\mathbf{W}^{\circ -1}$ can be applied and stored in $O(n)$ time and space.*
- ***Low-Rank Plus Block Diagonal**: Suppose that $\mathbf{W}$ is all ones except for r block diagonal matrices that are zeros. Then, $\mathbf{W}^{\circ -1}$ can be applied and stored in $O(nr)$ time and space.*
- ***Monotone Missing Data Pattern**: Suppose that $\mathbf{W}$ is a rank r matrix where each row is a prefix of all ones followed by a suffix of all zeros. Then, $\mathbf{W}^{\circ -1}$ can be applied and stored in $O(nr)$ time and space.*
- ***Low-Rank Plus Banded**: Suppose that $\mathbf{W}$ is all ones except for zeros on "band" entries, i.e., $\mathbf{W}_{i,j} = 0$ for $|i-j| \le p$. Then, $\mathbf{W}^{\circ -1}$ can be applied and stored in $O(np)$ time and space.*

Thus, our results yield efficient algorithms with provable approximation guarantees for a wide class of structured weight matrices encountered in practice. Furthermore, for general weight matrices, our results can be applied by first computing a low rank approximation of the weight matrices or an approximation by one of the above structured classes of weight matrices for further improvements in efficiency.

1.1.1. Column subset selection for weighted low rank approximation

Another advantage of allowing for approximation algorithms for computing low rank approximations to $\mathbf{W} \circ \mathbf{A}$ is that we can employ *column subset selection* approaches to low rank approximation (Frieze et al., 2004; Deshpande & Vempala, 2006; Drineas et al., 2006; 2008; Boutsidis et al., 2016; Altschuler et al., 2016). That is, it is known that the Frobenius norm low rank approximation problem admits $(1+\varepsilon)$-approximate low rank approximations whose left factor is formed by a subset of at most $O(k/\varepsilon)$ columns of the input matrix. In particular, these results show the existence of approximate solutions to the low rank approximation problem that preserve the sparsity of the input matrix, and thus can lead to a reduced solution size when the input matrix has sparse columns. Furthermore, column subset selection solutions to low rank approximation give a natural approach for *unsupervised feature selection*. Thus, as a corollary of Theorem 1.2, we obtain the first relative error guarantee for unsupervised feature selection with a weighted Frobenius norm objective. Weaker additive error guarantees were previously studied by (Dai, 2023; Axiotis & Yasuda, 2023)[2].

Corollary 1.5 (Column subset selection for weighted low rank approximation)**.** *There is an algorithm that computes a subset $S \subseteq [d]$ of $|S| = O(rk/\varepsilon)$ columns and $\mathbf{X} \in \mathbb{R}^{|S|\times d}$ such that*

$$\begin{aligned} &\left\|\mathbf{A} - \mathbf{W}^{\circ -1} \circ ((\mathbf{W}\circ\mathbf{A})|^S \mathbf{X})\right\|_{\mathbf{W},F}^2 \\ \le\ &(1+\varepsilon) \min_{\mathrm{rank}(\mathbf{A}')\le k} \|\mathbf{A} - \mathbf{A}'\|_{\mathbf{W},F}^2 \end{aligned}$$

where for a matrix $\mathbf{B} \in \mathbb{R}^{n\times d}$, $\mathbf{B}|^S$ denotes the matrix formed by the columns of $\mathbf{B}$ indexed by S.

Proof. This follows from Theorem 1.2 by computing the rank rk approximation $\tilde{\mathbf{A}}_{\mathbf{W}}$ to $\mathbf{W} \circ \mathbf{A}$ via column subset selection algorithms given by, e.g., (Boutsidis et al., 2016). □

Note that in Corollary 1.5, the approximation $\mathbf{W}^{\circ -1} \circ ((\mathbf{W} \circ \mathbf{A})|^S \mathbf{X})$ only depends on $\mathbf{A}$ through the columns $\mathbf{A}|^S$, and thus giving an approach to column subset selection with a weighted objective.

[2]The result of (Dai, 2023) contained an error, which we correct, tighten, and simplify in Appendix C.

1.1.2. Nearly optimal communication complexity bounds

As a consequence of Corollary 1.5, we obtain another important result for WLRA in the setting of *communication complexity*. Here, we obtain nearly optimal communication complexity bounds for constant factor approximations (i.e., $\kappa = O(1)$) to distributed WLRA for a wide range of parameters. While many works have studied distributed LRA in depth (Sarlós, 2006; Clarkson & Woodruff, 2009; 2013; Macua et al., 2010; Kannan et al., 2014; Ghashami et al., 2016; Boutsidis et al., 2016), we are surprisingly the first to initiate a study of this problem for WLRA.

The communication setting we consider is as follows. We have two players, Alice and Bob, where Alice has an input matrix $\mathbf{A}$ and would like to communicate an approximate WLRA solution to Bob. Communication complexity is of great interest in modern computing, where exchanging bits can be a critical bottleneck in large scale computation. While we consider two players in this discussion for simplicity, our algorithms also apply to a distributed computing setting, where the columns of the input matrix are partitioned among m servers as m matrices $\mathbf{A}^{(1)}, \mathbf{A}^{(2)}, \ldots, \mathbf{A}^{(m)}$, and some central coordinator outputs a WLRA of the concatenation $\mathbf{A} = [\mathbf{A}^{(1)}, \mathbf{A}^{(2)}, \ldots, \mathbf{A}^{(m)}]$ of these columns.

Definition 1.6 (WLRA: communication game)**.** Let Alice and Bob be two players. Let $\mathbf{W} \in \mathbb{Z}^{n\times d}$ be non-negative, let $\mathbf{A} \in \mathbb{Z}^{n\times d}$, and let $k \in \mathbb{N}$. Furthermore, let $\mathbf{W}$ and $\mathbf{A}$ have entries at most $\mathbf{W}_{i,j}, |\mathbf{A}_{i,j}| \leq \text{poly}(nd)$. We let both Alice and Bob receive the weight matrix $\mathbf{W}$ as input, and we give only Alice the input matrix $\mathbf{A}$. We say that Alice and Bob solve the *κ-approximate rank k weighted low rank approximation communication game* using B bits of communication if Alice sends at most B bits to Bob, and Bob outputs *any* matrix $\tilde{\mathbf{A}} \in \mathbb{R}^{n\times d}$ satisfying

$$\|\mathbf{A} - \tilde{\mathbf{A}}\|_{\mathbf{W},F} \leq \kappa \min_{\text{rank}(\mathbf{A}')\leq k} \|\mathbf{A} - \mathbf{A}'\|_{\mathbf{W},F}.$$

Suppose that $\mathbf{A}$ has columns which each have at most s non-zero entries. Then, the solution given by Corollary 1.5 can be communicated to Bob using just $O(srk/\varepsilon + rkd)$ numbers ($O(srk/\varepsilon)$ for the $O(rk/\varepsilon)$ columns of $\mathbf{A}$ and $O(rkd)$ for $\mathbf{X}$), or $O((srk/\varepsilon + rkd)\log(nd))$ bits under our bit complexity assumptions. Thus, when the number of columns d is at most the column sparsity s, then we obtain an algorithm which uses only $O((srk/\varepsilon)\log(nd))$ bits of communication. More generally, if the columns of $\mathbf{A}$ are distributed among m servers, then a solution can be computed using $O((msrk/\varepsilon)\log(nd))$ bits of communication by using work of (Boutsidis et al., 2016).

In fact, we show a nearly matching communication lower bound. In particular, we show that $\Omega(srk)$ bits of communication is required to output *any* matrix (not necessarily structured) that achieves a weighted Frobenius norm loss that is *any* finite factor within the optimal solution. Our lower bound is information-theoretic, and also immediately implies an $\Omega(msrk)$ bit lower bound in the distributed setting of m servers if each server must output a solution, as considered by (Boutsidis et al., 2016).

Theorem 1.7. *Let $\mathbf{W}$ be a binary block diagonal mask (Definition 4.2) and let $k \in \mathbb{N}$. Suppose that a randomized algorithm solves, for every $\mathbf{C} \in \mathbb{Z}^{n\times n}$ with at most s non-zero entries in each column, the κ-approximate weighted low rank approximation problem on input $\mathbf{C}$ using B bits of communication with probability at least $2/3$, for any $1 \leq \kappa < \infty$. If $s, k \leq n/r$, then $B = \Omega(srk)$.*

By proving a nearly tight communication complexity bound of $\tilde{\Theta}(rsk)$ for computing constant factor WLRAs, we arrive at the following qualitative observation: *the rank r of the weight matrix $\mathbf{W}$ parameterizes the communication complexity of WLRA*. A similar conclusion was drawn for the *computational* complexity of WLRA in the work of (Razenshteyn et al., 2016), where it was shown that WLRA is fixed parameter tractable in the parameter r, and also must have running time exponential in r under natural complexity theoretic assumptions. Thus, an important contribution of our work is to provide further evidence, both empirical and theoretical, that the rank r of the weight matrix $\mathbf{W}$ is a natural parameter to consider when studying WLRA.

1.1.3. Experiments

We demonstrate the empirical performance of our WLRA algorithms through experiments for model compression tasks. This application of WLRA was suggested by (Hsu et al., 2022; Hua et al., 2022), which we find to be a particularly relevant application of weighted low rank approximation due to the trend of extremely large models. In the model compression setting, we wish to approximate the hidden layer weight matrices of neural networks by much smaller matrices. A classical way to do this is to use low rank approximation (Sainath et al., 2013; Kim et al., 2016; Chen et al., 2018). While this often gives reasonable results, the works of (Hsu et al., 2022; Hua et al., 2022) show that significant improvements can be obtained by taking into account the importance of each of the parameters in the LRA problem. We thus conduct our experiments in this setting.

We first show in Section 5.1 that the importance matrices arising this application are indeed very low rank. We may interpret this phenomenon intuitively: we hypothesize that the importance score of some parameter $\mathbf{A}_{i,j}$ is essentially the product of the importance of the corresponding input i and the importance of the corresponding output j. This observation may be of independent interest, and also empirically justifies the low rank weight matrix assumption that we make in this work, as well as works of (Razenshteyn

et al., 2016; Ban et al., 2019). While WLRA with a rank 1 weight matrix is known to be solvable efficiently via the SVD, our result shows that general low rank weight matrices also yield efficient algorithms via the SVD.

Next in Section 5.2, we conduct experiments which demonstrate the superiority of our methods in practice. Of the algorithms that we compare to, an expectation-minimization (EM) approach of (Srebro & Jaakkola, 2003) gives the smallest loss albeit with a very high running time, and our algorithm nearly matches this loss with an order of magnitude lower running time. We also show that this solution can be refined with EM, producing the best trade-off between efficiency and accuracy. One of the baselines we compare is a sampling algorithm of (Dai, 2023), whose analysis contains an error which we correct, simplify, and tighten.

1.2. Related work

We survey a number of related works on approximation algorithms for weighted low rank approximation. One of the earliest algorithms for this problem is a natural EM approach proposed by (Srebro & Jaakkola, 2003). Another related approach is to parameterize the low rank approximation $\tilde{\mathbf{A}}$ as the product $\mathbf{UV}$ of two matrices $\mathbf{U} \in \mathbb{R}^{n\times k}$ and $\mathbf{V} \in \mathbb{R}^{k\times d}$ and alternately minimize the two matrices, known as *alternating least squares*. This algorithm has been studied in a number of works (Hastie et al., 2015; Li et al., 2016; Song et al., 2023). The work of (Bhaskara et al., 2021) proposes an approach to weighted low rank approximation based on a greedy pursuit, where rank one factors are iteratively added based on an SVD of the gradient matrix. Finally, fixed parameter tractable algorithms have been considered in (Razenshteyn et al., 2016; Ban et al., 2019) based on sketching techniques.

2. Approximation algorithms

The following simple observation is the key idea behind Theorem 1.2:

Lemma 2.1. *Let $\mathbf{W}, \mathbf{A}' \in \mathbb{R}^{n\times d}$ with $\mathrm{rank}(\mathbf{W}) \le r$ and $\mathrm{rank}(\mathbf{A}') \le k$. Then, $\mathrm{rank}(\mathbf{W} \circ \mathbf{A}') \le rk$.*

Proof. Since $\mathrm{rank}(\mathbf{W}) \le r$, it can be written as $\mathbf{W} = \sum_{i=1}^{r} \mathbf{u}_i\mathbf{v}_i^\top$ for $\mathbf{u}_i \in \mathbb{R}^n$ and $\mathbf{v}_i \in \mathbb{R}^d$. Then,

$$\mathbf{W} \circ \mathbf{A}' = \sum_{i=1}^{r} (\mathbf{u}_i\mathbf{v}_i^\top) \circ \mathbf{A}' = \sum_{i=1}^{r} \mathrm{diag}(\mathbf{u}_i)\mathbf{A}'\,\mathrm{diag}(\mathbf{v}_i)$$

so $\mathbf{W} \circ \mathbf{A}'$ is the sum of r matrices, each of which is rank k. Thus, $\mathbf{W} \circ \mathbf{A}'$ has rank at most rk. □

Using Lemma 2.1, we obtain the following:

Theorem 1.2. *Let $\mathbf{W} \in \mathbb{R}^{n\times d}$ be a non-negative weight matrix with rank r. Let $\mathbf{A} \in \mathbb{R}^{n\times d}$ and let $k \in \mathbb{N}$. Suppose that $\tilde{\mathbf{A}}_{\mathbf{W}} \in \mathbb{R}^{n\times d}$ satisfies*

$$\begin{aligned}\|\mathbf{W} \circ \mathbf{A} - \tilde{\mathbf{A}}_{\mathbf{W}}\|_F^2 &\le \kappa \min_{\mathrm{rank}(\mathbf{A}')\le rk} \|\mathbf{W} \circ \mathbf{A} - \mathbf{A}'\|_F^2 \\ &= \kappa \|(\mathbf{W} \circ \mathbf{A})_{-rk}\|_F^2\end{aligned}$$

and let $\tilde{\mathbf{A}} := \mathbf{W}^{\circ -1} \circ \tilde{\mathbf{A}}_{\mathbf{W}}$, where $\mathbf{W}^{\circ -1} \in \mathbb{R}^{n\times d}$ denotes the entrywise inverse of $\mathbf{W}$. Then,

$$\|\mathbf{A} - \tilde{\mathbf{A}}\|_{\mathbf{W},F}^2 \le \kappa \min_{\mathrm{rank}(\mathbf{A}')\le k} \|\mathbf{A} - \mathbf{A}'\|_{\mathbf{W},F}^2$$

In particular, we obtain a solution with $\kappa = (1+\varepsilon)$ in running time $O(\mathsf{nnz}(\mathbf{A})) + \tilde{O}(n(rk)^2/\varepsilon + \mathrm{poly}(rk/\varepsilon))$ by using randomized low rank approximation algorithms of (Avron et al., 2017).

Proof. Note that $\|\mathbf{W}^{\circ -1} \circ \tilde{\mathbf{A}}_{\mathbf{W}} - \mathbf{A}\|_{\mathbf{W},F}^2 = \|\tilde{\mathbf{A}}_{\mathbf{W}} - \mathbf{W} \circ \mathbf{A}\|_F^2$, which is at most $\kappa\|(\mathbf{W} \circ \mathbf{A})_{-rk}\|_F^2$ by assumption. On the other hand for any rank k matrix $\mathbf{A}'$, $\|\mathbf{A}' - \mathbf{A}\|_{\mathbf{W},F} = \|\mathbf{W} \circ \mathbf{A}' - \mathbf{W} \circ \mathbf{A}\|_F$ can be lower bounded by $\|(\mathbf{W} \circ \mathbf{A})_{-rk}\|_F$ since $\mathbf{W} \circ \mathbf{A}'$ has rank at most rk by Lemma 2.1. Thus,

$$\begin{aligned}\|\mathbf{W}^{\circ -1} \circ \tilde{\mathbf{A}}_{\mathbf{W}} - \mathbf{A}\|_{\mathbf{W},F}^2 &\le \kappa\|(\mathbf{W} \circ \mathbf{A})_{-rk}\|_F^2 \\ &\le \kappa \min_{\mathrm{rank}(\mathbf{A}')\le k} \|\mathbf{A} - \mathbf{A}'\|_{\mathbf{W},F}^2.\end{aligned}$$

□

3. Matrices with structured entrywise inverses

We present a general lemma which shows how to handle a wide family of structured matrices which often arise in practice as weight matrices for weighted low rank approximation.

Lemma 3.1. *Let $\mathbf{W} \in \mathbb{R}^{n\times d}$ be structured as $\mathbf{W} = \mathbf{E} + \sum_{i=1}^{r'} \mathbf{S}_i$, where $\mathbf{S}_i$ are rank 1 matrices with disjoint rectangular supports $S_i \times T_i$ for $S_i \subseteq [n]$ and $T_i \subseteq [d]$, and $\mathbf{E}$ is a sparse matrix with $\mathsf{nnz}(\mathbf{E})$ non-zero entries. Let $\mathbf{A} = \mathbf{UV}$ for $\mathbf{U} \in \mathbb{R}^{n\times k}$ and $\mathbf{V} \in \mathbb{R}^{k\times d}$ be a rank k matrix. Then, $\mathbf{W}^{\circ -1} \circ \mathbf{A}$ can be stored in $O(\mathsf{nnz}(\mathbf{E}) + \sum_{i=1}^{r'} |S_i| + |T_i|)$ space and can be applied to a vector $\mathbf{x} \in \mathbb{R}^d$ in $O(\mathsf{nnz}(\mathbf{E}) + \sum_{i=1}^{r'} |S_i| + |T_i|)$ time. Furthermore, $\mathbf{W}$ has rank at most $r = \mathsf{nnz}(\mathbf{E}) + r'$.*

Proof. Let $\mathbf{S} = \sum_{i=1}^{r} \mathbf{S}_i$. We can then write $\mathbf{W}^{\circ -1} = \mathbf{E}' + \mathbf{S}^{\circ -1}$, where $\mathbf{E}'$ is a sparse matrix with $\mathsf{nnz}(\mathbf{E}') = \mathsf{nnz}(\mathbf{E})$. Since $\mathbf{S}_i$ have disjoint supports, we can also write $\mathbf{S}^{\circ -1} = \sum_{i=1}^{r} \mathbf{S}_i^{\circ -1}$. Thus,

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

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

Note that $\mathsf{nnz}(\mathbf{E}' \circ \mathbf{A}) \leq \mathsf{nnz}(\mathbf{E})$, so this can be stored and applied in $O(\mathsf{nnz}(\mathbf{E}))$ time and space. For each i, $\mathbf{S}_i^{\circ -1}$ is just a rank 1 matrix supported on $S_i \times T_i$, so this can be stored and applied in $O(|S_i| + |T_i|)$ time and space. □

As a corollary of Lemma 1.3, we show that we can efficiently handle all five families of commonly encountered weight matrices discussed in (Musco et al., 2021). We refer to (Musco et al., 2021) on the large body of work studying these classes of weight matrices.

Corollary 3.2. *Let $\mathbf{W} \in \mathbb{R}^{n \times d}$. The following hold:*

- ***Low-Rank Plus Sparse**: Suppose that $\mathbf{W}$ has at most t non-zeros per row. Then, $\mathbf{W}^{\circ -1}$ can be stored and applied in $O(nt)$ time and space.*
- ***Low-Rank Plus Diagonal**: Suppose that $\mathbf{W}$ is all ones except for zeros along the diagonal. Then, $\mathbf{W}^{\circ -1}$ can be applied and stored in $O(n)$ time and space.*
- ***Low-Rank Plus Block Diagonal**: Suppose that $\mathbf{W}$ is all ones except for r block diagonal matrices that are zeros. Then, $\mathbf{W}^{\circ -1}$ can be applied and stored in $O(nr)$ time and space.*
- ***Monotone Missing Data Pattern**: Suppose that $\mathbf{W}$ is a rank r matrix where each row is a prefix of all ones followed by a suffix of all zeros. Then, $\mathbf{W}^{\circ -1}$ can be applied and stored in $O(nr)$ time and space.*
- ***Low-Rank Plus Banded**: Suppose that $\mathbf{W}$ is all ones except for zeros on "band" entries, i.e., $\mathbf{W}_{i,j} = 0$ for $|i - j| \leq p$. Then, $\mathbf{W}^{\circ -1}$ can be applied and stored in $O(np)$ time and space.*

Proof. For the Low-Rank Plus Sparse weight matrices, $\mathbf{W}$ itself has at most $O(nt)$ non-zero entries and thus can be set to $\mathbf{E}$ in Lemma 1.3. For the Low-Rank Plus Diagonal weight matrices, $\mathbf{W}$ can be written as the sum of the rank 1 matrix of all ones and a sparse matrix supported on the diagonal. For the Low-Rank Plus Block Diagonal weight matrices, we can write the complement of the blocks along their rows as disjoint rank 1 matrices. Since there are at most r blocks, this is the sum of at most $r + 1$ disjoint rank 1 matrices. For the Monotone Missing Data Pattern weight matrices, there can only be at most r different patterns of rows, and these can be written as the sum of r disjoint rank 1 matrices. For the Low-Rank Plus Banded weight matrices, $\mathbf{W}$ can be written as the sum of the rank 1 matrix of all ones and a sparse matrix supported on the diagonal band. □

4. Communication complexity bounds

We show that our approach to weighted low rank approximation in Theorem 1.2 gives nearly optimal bounds for this problem in the setting of communication complexity.

Our first result is an upper bound for the communication game in Definition 1.6.

Theorem 4.1. *Let $\mathbf{W} \in \mathbb{Z}^{n \times d}$ be a non-negative rank k weight matrix and let $\mathbf{A} \in \mathbb{Z}^{n \times d}$ be an input matrix with at most s non-zero entries in each column. There is an algorithm which solves the $(1 + \varepsilon)$-approximate weighted low rank approximation communication game (Definition 1.6) using at most $B = O((srk/\varepsilon + rkd)\log(nd))$ bits of communication.*

Proof. The algorithm is to use the column subset selection-based WLRA algorithm of Corollary 1.5 and then to send the columns of $\mathbf{A}$ indexed by the subset S and $\mathbf{X}$. □

On the other hand, we show a communication complexity lower bound showing that the number of bits B exchanged by Alice and Bob must be at least $\Omega(rsk)$. Our lower bound holds even when the weight matrix $\mathbf{W}$ is the following simple binary matrix.

Definition 4.2 (Block diagonal mask)**.** Let $r \in \mathbb{N}$ and let n be an integer multiple of r. Then, $\mathbf{W} \in \{0, 1\}^{n \times n}$ is the *block diagonal mask* associated with these parameters if $\mathbf{W}$ is the $r \times r$ block diagonal matrix with diagonal blocks given by the $n/r \times n/r$ all ones matrix and off-diagonal blocks given by the $n/r \times n/r$ all zeros matrix.

We give our communication complexity lower bound in the following theorem.

Theorem 1.7. *Let $\mathbf{W}$ be a binary block diagonal mask (Definition 4.2) and let $k \in \mathbb{N}$. Suppose that a randomized algorithm solves, for every $\mathbf{C} \in \mathbb{Z}^{n \times n}$ with at most s non-zero entries in each column, the κ-approximate weighted low rank approximation problem on input $\mathbf{C}$ using B bits of communication with probability at least $2/3$, for any $1 \leq \kappa < \infty$. If $s, k \leq n/r$, then $B = \Omega(srk)$.*

Proof. Let $\mathbf{A}_{\text{dense}} \in \{0, 1\}^{sr \times k}$ be a uniformly random binary matrix, and let $\mathbf{A}_{\text{pad}} \in \{0, 1\}^{n \times n/r}$ be formed by padding the columns of $\mathbf{A}_{\text{dense}}$ with $n/r - k$ zero columns and padding each block of s contiguous rows with $n/r - s$ zero rows. For $j \in [r]$, let $\mathbf{A}_{\text{pad}}^{(j)}$ be the restriction of $\mathbf{A}_{\text{pad}}$ to the jth contiguous block of n/r rows. We then construct $\mathbf{A} \in \mathbb{R}^{n \times n}$ by horizontally concatenating r copies of $\mathbf{A}_{\text{pad}}$.

Note that the optimal rank k approximation achieves 0 loss in the $\mathbf{W}$-weighted Frobenius norm. Indeed, we can take $\mathbf{A}^*$ to be the horizontal concatenation of r copies of $\mathbf{A}_{\text{pad}}$. Since $\mathbf{A}_{\text{pad}}$ has rank k, $\mathbf{A}^*$ also has rank k. Furthermore,

on the j-th nonzero blocks of $\mathbf{W}$, $\mathbf{A}_{\text{pad}}$ has the same entries as $\mathbf{A}_{\text{pad}}^{(j)}$. Thus, an approximation $\tilde{\mathbf{A}}$ that achieves *any* finite approximation factor κ must exactly recover $\mathbf{A}$, restricted to the support of $\mathbf{W}$. In turn, this means that such an approximation $\tilde{\mathbf{A}}$ can also be used to recover $\mathbf{A}_{\text{dense}}$.

It now follows by a standard information-theoretic argument that $B = \Omega(srk)$ (see Appendix A for further details). □

5. Experiments

As discussed in Section 1.1.3, we first conduct experiments for WLRA in the setting of model compression (as proposed by (Hsu et al., 2022; Hua et al., 2022)). In our experiments, we train a basic multilayer perceptron (MLP) on four image datasets, `mnist`, `fashion_mnist`, `smallnorb`, and `colorectal_histology` which were selected from the `tensorflow_datasets` catalogue for simplicity of processing (e.g., fixed feature size, no need for embeddings, etc.). We then compute a matrix of importances of each of the parameters in a hidden layer of the MLP given by the Fisher information matrix. Finally, we compute a WLRA of the hidden layer matrix using the Fisher information matrix as the weights $\mathbf{W}$.

Our experiments are conducted on a 2019 MacBook Pro with a 2.6 GHz 6-Core Intel Core i7 processor. All code used in the experiments are available in the supplementary.

Table 1: Datasets used in experiments

Dataset	Image dim.	Flattened dim.	Neurons	Matrix dim.
`mnist`	$(28, 28, 1)$	784	128	784×128
`fashion_mnist`	$(28, 28, 1)$	784	128	784×128
`smallnorb`	$(96, 96, 1)$	9216	1024	9216×1024
`colorectal_histology`	$(150, 150, 3)$	67500	1024	67500×1024

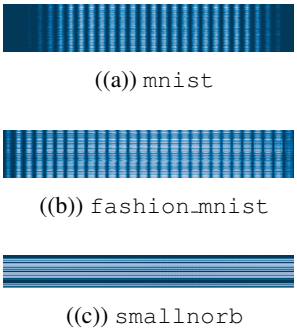
5.1. The low rank weight matrix assumption in practice

We first demonstrate that for the task of model compression, the weight matrix is approximately low rank in practice. The weight matrix $\mathbf{W}$ in this setting is the empirical Fisher information matrix of the hidden layer weights $\mathbf{A}$, where the empirical Fisher information of the (i, j)-th entry $\mathbf{A}_{i,j}$ is given by

$$\mathbf{W}_{i,j} \coloneqq \mathbb{E}_{\mathbf{x}\sim\mathcal{D}}\left[\left(\frac{\partial}{\partial \mathbf{A}_{i,j}}\mathcal{L}(\mathbf{x}; \mathbf{A})\right)^2\right]$$

where $\mathcal{L}(\mathbf{x}; \mathbf{A})$ denotes the loss of the neural network on the data point $\mathbf{x}$ and hidden layer weights $\mathbf{A}$, and $\mathcal{D}$ denotes the empirical distribution (that is, the uniform distribution over the training data).

Plots of the empirical Fisher matrix (Figure 1) reveal low rank structure to the matrices, and the spectrum of the Fisher matrix confirms that the vast majority of the Frobenius norm is contained in the first singular value (Table 2). We also plot the spectrum itself in Figure 6 in the appendix.



((a)) `mnist`

((b)) `fashion_mnist`

((c)) `smallnorb`

Figure 1: Low rank structure of Fisher weight matrices

Table 2: % mass of Fisher matrix in 1st singular value

Dataset	% mass
`mnist`	95.4%
`fashion_mnist`	95.9%
`smallnorb`	99.9%
`colorectal_histology`	99.3%

5.2. Approximation quality and running time

In this section, we compare the performance of our Algorithm 1 (denoted as `svd_w` in the following discussion) with a variety of previously proposed algorithms for WLRA.

We consider the following algorithms: `adam`, `em`, `greedy`, `sample`, and `svd`, which we next explain in detail. We first consider `adam`, in which we simply parameterize the WLRA problem as an optimization problem in the factorized representation $\mathbf{UV}$ for factors $\mathbf{U} \in \mathbb{R}^{n\times k}$ and $\mathbf{V} \in \mathbb{R}^{k\times d}$ (Burer & Monteiro, 2003), and optimize this loss function using the Adam optimizer provided in the `tensorflow` library. Such an approach is well-studied for the standard low rank approximation problem (Li et al., 2018; Ye & Du, 2021), and empirically performs well for weighted low rank approximation as well. This was run for 100 epochs, with an initial learning rate of 1.0 decayed by a factor of 0.7 every 10 steps. The `em` algorithm was proposed by (Srebro & Jaakkola, 2003) for the WLRA problem, and involves iteratively "filling in" missing values and recomputing a low rank approximation. In the experiments, we run 25 iterations. The `greedy` algorithm is a greedy basis pursuit algorithm proposed by (Bhaskara et al., 2021) and iteratively adds new directions to the low rank approximation by taking an SVD of the gradient of the objective. Similar algorithms were also studied in (Shalev-Shwartz et al., 2011; Khanna et al., 2017; Axiotis & Sviridenko, 2021) for general rank-constrained convex optimization problems. The `sample` algorithm is a row norm sampling approach studied by (Dai,

2023). Finally, `svd` simply computes an SVD of the original matrix $\mathbf{W}$, without regard to the weights $\mathbf{W}$.

We compute low rank approximations for ranks 1 through 20 on four datasets, and plot the loss and the running time against the rank in Figures 2 and 3, respectively. The values in the figures are tabulated at ranks 20, 10, and 5 in Tables 3, 4, and 5 in the supplementary. We observe that our `svd_w` algorithm performs among the best in the approximation loss (Figure 2), nearly matching the approximation quality achieved by much more computational expensive algorithms such as `adam` and `em`, while requiring much less computational time (Figure 3).

While in some cases the `em` algorithm may eventually produce a better solution, we note that our `svd_w` may be improved by initializing the `em` algorithm with this solution, which produces an algorithm which quickly produces a superior solution with many fewer iterations (Figure 4).

5.3. Experiments on synthetic datasets

Finally, we perform additional experiments on a synthetic datasets based on a mixture of Gaussians. In this experiment, we consider a uniform mixture of k Gaussians in d dimensions with diagonal covariance matrices, each which has variances that take at most r distinct values. The input matrix $\mathbf{A}$ is taken to be n i.i.d. observations of this distribution, while the weight matrix scales a Gaussian with variance σ^2 by $1/\sigma^2$. It can easily be observed that this weight matrix has rank at most kr, and thus our results apply. The variances for are chosen randomly by taking the 4th powers of random Gaussian variables. The 4th power is taken to make the variances more varied, which makes the WLRA problem more interesting. In Figure 5, we show the results for $n = 1000$, $d = 50$, $k = 5$, and $r = 3$. Our results again show that our algorithm achieves superior losses with a running time that is competitive with a standard SVD.

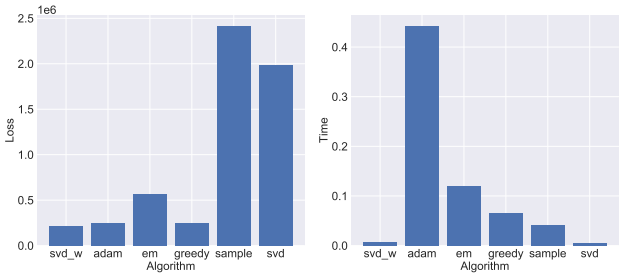


Figure 5: Loss and running time of WLRA on a synthetic dataset based on a mixture of Gaussians. Results are averaged over 5 trials.

6. Conclusion

In this work, we studied new algorithms for the weighted low rank approximation problem, which has countless applications in statistics, machine learning, and signal processing. We propose an approach based on reweighting a low rank matrix, which is a novel class of relaxed solutions to the WLRA problem, and give provable guarantees under the assumption that the weight matrix $\mathbf{W}$ has low rank. Theoretically, this allows us to obtain an algorithm for WLRA with nearly optimal communication complexity, for which we show nearly matching communication complexity lower bounds, which shows that the rank of the weight matrix tightly parameterizes the communication complexity of this problem. We also give the first guarantees for column subset selection for weighted low rank approximation, which gives a notion of feature selection with a weighted objective. Finally, we show that in practice, our approach gives a highly efficient algorithm that outperforms prior algorithms for WLRA, particularly when combined with refinement using expectation-maximization.

Impact statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

Acknowledgements

We thank the anonymous reviewers for useful feedback on improving the presentation of this work. We also thank Yucheng Dai for helpful discussions. David P. Woodruff and Taisuke Yasuda were supported by a Simons Investigator Award.

References

Altschuler, J. M., Bhaskara, A., Fu, G., Mirrokni, V. S., Rostamizadeh, A., and Zadimoghaddam, M. Greedy column subset selection: New bounds and distributed algorithms. In Balcan, M. and Weinberger, K. Q. (eds.), *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, volume 48 of *JMLR Workshop and Conference Proceedings*, pp. 2539–2548. JMLR.org, 2016.

Anandan, P. and Irani, M. Factorization with uncertainty. *Int. J. Comput. Vis.*, 49(2-3):101–116, 2002. doi: 10.1023/A:1020137420717.

Arora, S., Li, Y., Liang, Y., Ma, T., and Risteski, A. A latent variable model approach to pmi-based word embeddings.

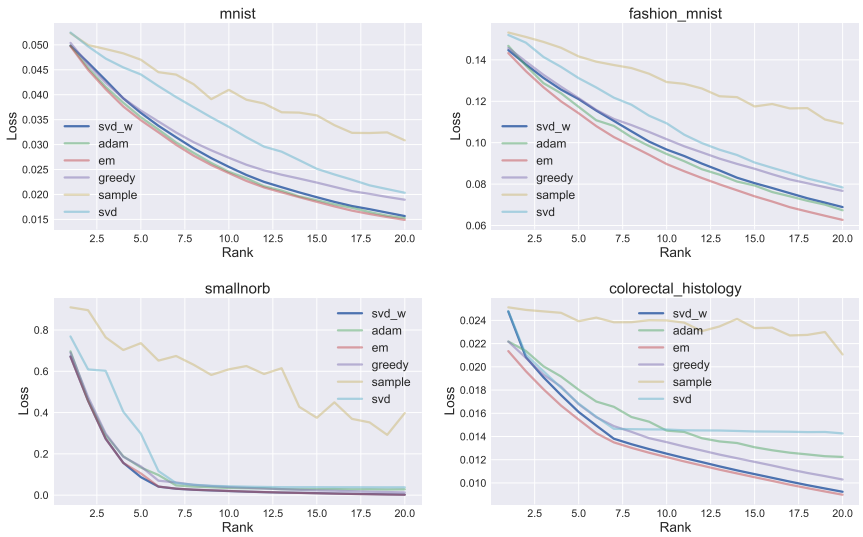


Figure 2: Fisher-weighted low rank approximation loss of weighted low rank approximation algorithms for model compression of four datasets. Results are averaged over 5 trials.

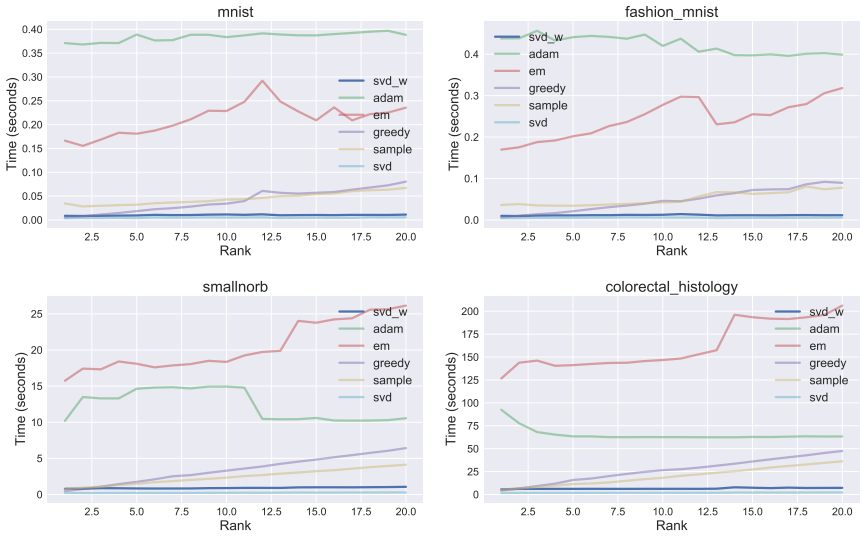


Figure 3: Running time of weighted low rank approximation algorithms for model compression of four datasets. Results are averaged over 5 trials.

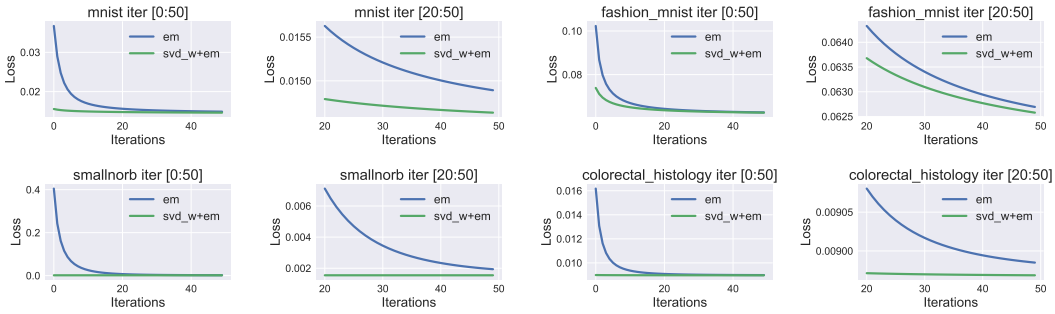


Figure 4: Improving the `svd_w` solution with `em` iterations for a rank 20 approximation.

Trans. Assoc. Comput. Linguistics, 4:385–399, 2016. doi: 10.1162/tacl_a_00106.

Avron, H., Clarkson, K. L., and Woodruff, D. P. Sharper bounds for regularized data fitting. In Jansen, K., Rolim, J. D. P., Williamson, D., and Vempala, S. S. (eds.), *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques, APPROX/RANDOM 2017, August 16-18, 2017, Berkeley, CA, USA*, volume 81 of *LIPIcs*, pp. 27:1–27:22. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2017. doi: 10.4230/LIPICS.APPROX-RANDOM.2017.27.

Axiotis, K. and Sviridenko, M. Local search algorithms for rank-constrained convex optimization. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

Axiotis, K. and Yasuda, T. Performance of ℓ_1 regularization for sparse convex optimization. *CoRR*, abs/2307.07405, 2023. doi: 10.48550/arXiv.2307.07405.

Balzano, L., Recht, B., and Nowak, R. D. High-dimensional matched subspace detection when data are missing. In *IEEE International Symposium on Information Theory, ISIT 2010, June 13-18, 2010, Austin, Texas, USA, Proceedings*, pp. 1638–1642. IEEE, 2010. doi: 10.1109/ISIT.2010.5513344.

Ban, F., Woodruff, D. P., and Zhang, Q. R. Regularized weighted low rank approximation. In Wallach, H. M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. B., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 4061–4071, 2019.

Bhaskara, A., Ruwanpathirana, A. K., and Wijewardena, M. Additive error guarantees for weighted low rank approximation. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 874–883. PMLR, 2021.

Boutsidis, C., Woodruff, D. P., and Zhong, P. Optimal principal component analysis in distributed and streaming models. In Wichs, D. and Mansour, Y. (eds.), *Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2016, Cambridge, MA, USA, June 18-21, 2016*, pp. 236–249. ACM, 2016. doi: 10.1145/2897518.2897646.

Burer, S. and Monteiro, R. D. C. A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization. volume 95, pp. 329–357. 2003. doi: 10.1007/s10107-002-0352-8. Computational semidefinite and second order cone programming: the state of the art.

Candès, E. J. and Tao, T. The power of convex relaxation: near-optimal matrix completion. *IEEE Trans. Inf. Theory*, 56(5):2053–2080, 2010. doi: 10.1109/TIT.2010.2044061.

Chen, C., Li, D., Zhao, Y., Lv, Q., and Shang, L. WEMAREC: accurate and scalable recommendation through weighted and ensemble matrix approximation. In Baeza-Yates, R., Lalmas, M., Moffat, A., and Ribeiro-Neto, B. A. (eds.), *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, Santiago, Chile, August 9-13, 2015*, pp. 303–312. ACM, 2015. doi: 10.1145/2766462.2767718.

Chen, P. H., Si, S., Li, Y., Chelba, C., and Hsieh, C. Groupreduce: Block-wise low-rank approximation for neural language model shrinking. In Bengio, S., Wallach, H. M., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pp. 11011–11021, 2018.

Clarkson, K. L. and Woodruff, D. P. Numerical linear algebra in the streaming model. In Mitzenmacher, M. (ed.), *Proceedings of the 41st Annual ACM Symposium on Theory of Computing, STOC 2009, Bethesda, MD, USA, May 31 - June 2, 2009*, pp. 205–214. ACM, 2009. doi: 10.1145/1536414.1536445.

Clarkson, K. L. and Woodruff, D. P. Low rank approximation and regression in input sparsity time. In Boneh, D., Roughgarden, T., and Feigenbaum, J. (eds.), *Symposium on Theory of Computing Conference, STOC'13, Palo Alto, CA, USA, June 1-4, 2013*, pp. 81–90. ACM, 2013. doi: 10.1145/2488608.2488620.

Cover, T. M. and Thomas, J. A. *Elements of information theory (2. ed.)*. Wiley, 2006. ISBN 978-0-471-24195-9.

Dai, Y. On algorithms for weighted low rank approximation. Master's thesis, Carnegie Mellon University, 2023.

Deshpande, A. and Vempala, S. S. Adaptive sampling and fast low-rank matrix approximation. In Díaz, J., Jansen, K., Rolim, J. D. P., and Zwick, U. (eds.), *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques, 9th International Workshop on Approximation Algorithms for Combinatorial Optimization Problems, APPROX 2006 and 10th International Workshop on Randomization and Computation, RANDOM 2006, Barcelona, Spain, August 28-30 2006, Proceedings*, volume 4110 of *Lecture Notes in Computer Science*, pp. 292–303. Springer, 2006. doi: 10.1007/11830924_28.

Drineas, P., Mahoney, M. W., and Muthukrishnan, S. Subspace sampling and relative-error matrix approximation: Column-row-based methods. In Azar, Y. and Erlebach, T. (eds.), *Algorithms - ESA 2006, 14th Annual European Symposium, Zurich, Switzerland, September 11-13, 2006, Proceedings*, volume 4168 of *Lecture Notes in Computer Science*, pp. 304–314. Springer, 2006. doi: 10.1007/11841036_29.

Drineas, P., Mahoney, M. W., and Muthukrishnan, S. Relative-error CUR matrix decompositions. *SIAM J. Matrix Anal. Appl.*, 30(2):844–881, 2008. doi: 10.1137/07070471X.

Dutta, A., Li, X., and Richtárik, P. Weighted low-rank approximation of matrices and background modeling. *CoRR*, abs/1804.06252, 2018.

Frieze, A. M., Kannan, R., and Vempala, S. S. Fast monte-carlo algorithms for finding low-rank approximations. *J. ACM*, 51(6):1025–1041, 2004. doi: 10.1145/1039488.1039494.

Ghashami, M., Liberty, E., Phillips, J. M., and Woodruff, D. P. Frequent directions: Simple and deterministic matrix sketching. *SIAM J. Comput.*, 45(5):1762–1792, 2016. doi: 10.1137/15M1009718.

Gillis, N. and Glineur, F. Low-rank matrix approximation with weights or missing data is NP-hard. *SIAM J. Matrix Anal. Appl.*, 32(4):1149–1165, 2011. doi: 10.1137/110820361.

Hastie, T., Mazumder, R., Lee, J. D., and Zadeh, R. Matrix completion and low-rank SVD via fast alternating least squares. *J. Mach. Learn. Res.*, 16:3367–3402, 2015. doi: 10.5555/2789272.2912106.

Hsu, Y., Hua, T., Chang, S., Lou, Q., Shen, Y., and Jin, H. Language model compression with weighted low-rank factorization. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022.

Hua, T., Hsu, Y., Wang, F., Lou, Q., Shen, Y., and Jin, H. Numerical optimizations for weighted low-rank estimation on language models. In Goldberg, Y., Kozareva, Z., and Zhang, Y. (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pp. 1404–1416. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.emnlp-main.91.

Kannan, R., Vempala, S. S., and Woodruff, D. P. Principal component analysis and higher correlations for distributed data. In Balcan, M., Feldman, V., and Szepesvári, C. (eds.), *Proceedings of The 27th Conference on Learning Theory, COLT 2014, Barcelona, Spain, June 13-15, 2014*, volume 35 of *JMLR Workshop and Conference Proceedings*, pp. 1040–1057. JMLR.org, 2014.

Keshavan, R. H., Montanari, A., and Oh, S. Matrix completion from a few entries. *IEEE Trans. Inf. Theory*, 56(6):2980–2998, 2010. doi: 10.1109/TIT.2010.2046205.

Khanna, R., Elenberg, E. R., Dimakis, A. G., Ghosh, J., and Negahban, S. N. On approximation guarantees for greedy low rank optimization. In Precup, D. and Teh, Y. W. (eds.), *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1837–1846. PMLR, 2017.

Kidzinski, L., Hui, F. K. C., Warton, D. I., and Hastie, T. J. Generalized matrix factorization: efficient algorithms for fitting generalized linear latent variable models to large data arrays. *J. Mach. Learn. Res.*, 23:291:1–291:29, 2022.

Kim, Y., Park, E., Yoo, S., Choi, T., Yang, L., and Shin, D. Compression of deep convolutional neural networks for fast and low power mobile applications. In Bengio, Y. and LeCun, Y. (eds.), *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016.

Koren, Y., Bell, R. M., and Volinsky, C. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009. doi: 10.1109/MC.2009.263.

Lee, J., Kim, S., Lebanon, G., Singer, Y., and Bengio, S. LLORMA: local low-rank matrix approximation. *J. Mach. Learn. Res.*, 17:15:1–15:24, 2016.

Li, X., Dutta, A., and Richtárik, P. A batch-incremental video background estimation model using weighted low-rank approximation of matrices. In *2017 IEEE International Conference on Computer Vision Workshops, ICCV Workshops 2017, Venice, Italy, October 22-29, 2017*, pp. 1835–1843. IEEE Computer Society, 2017. doi: 10.1109/ICCVW.2017.217.

Li, Y., Liang, Y., and Risteski, A. Recovery guarantee of weighted low-rank approximation via alternating minimization. In Balcan, M. and Weinberger, K. Q. (eds.), *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, volume 48 of *JMLR Workshop and Conference Proceedings*, pp. 2358–2367. JMLR.org, 2016.

Li, Y., Ma, T., and Zhang, H. Algorithmic regularization in over-parameterized matrix sensing and neural networks

with quadratic activations. In Bubeck, S., Perchet, V., and Rigollet, P. (eds.), *Conference On Learning Theory, COLT 2018, Stockholm, Sweden, 6-9 July 2018*, volume 75 of *Proceedings of Machine Learning Research*, pp. 2–47. PMLR, 2018.

Lu, W.-S., Pei, S.-C., and Wang, P.-H. Weighted low-rank approximation of general complex matrices and its application in the design of 2-d digital filters. *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, 44(7):650–655, 1997.

Macua, S. V., Belanovic, P., and Zazo, S. Consensus-based distributed principal component analysis in wireless sensor networks. In *2010 IEEE 11th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, pp. 1–5. IEEE, 2010.

Musco, C. and Musco, C. Randomized block krylov methods for stronger and faster approximate singular value decomposition. In Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pp. 1396–1404, 2015.

Musco, C., Musco, C., and Woodruff, D. P. Simple heuristics yield provable algorithms for masked low-rank approximation. In Lee, J. R. (ed.), *12th Innovations in Theoretical Computer Science Conference, ITCS 2021, January 6-8, 2021, Virtual Conference*, volume 185 of *LIPIcs*, pp. 6:1–6:20. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2021. doi: 10.4230/LIPIcs.ITCS.2021.6.

Razenshteyn, I. P., Song, Z., and Woodruff, D. P. Weighted low rank approximations with provable guarantees. In Wichs, D. and Mansour, Y. (eds.), *Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2016, Cambridge, MA, USA, June 18-21, 2016*, pp. 250–263. ACM, 2016. doi: 10.1145/2897518.2897639.

Rennie, J. D. M. and Srebro, N. Fast maximum margin matrix factorization for collaborative prediction. In Raedt, L. D. and Wrobel, S. (eds.), *Machine Learning, Proceedings of the Twenty-Second International Conference (ICML 2005), Bonn, Germany, August 7-11, 2005*, volume 119 of *ACM International Conference Proceeding Series*, pp. 713–719. ACM, 2005. doi: 10.1145/1102351.1102441.

Robin, G., Josse, J., Moulines, E., and Sardy, S. Low-rank model with covariates for count data with missing values. *J. Multivar. Anal.*, 173:416–434, 2019. doi: 10.1016/j.jmva.2019.04.004.

Sainath, T. N., Kingsbury, B., Sindhwani, V., Arisoy, E., and Ramabhadran, B. Low-rank matrix factorization for deep neural network training with high-dimensional output targets. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2013, Vancouver, BC, Canada, May 26-31, 2013*, pp. 6655–6659. IEEE, 2013. doi: 10.1109/ICASSP.2013.6638949.

Sarlós, T. Improved approximation algorithms for large matrices via random projections. In *47th Annual IEEE Symposium on Foundations of Computer Science (FOCS 2006), 21-24 October 2006, Berkeley, California, USA, Proceedings*, pp. 143–152. IEEE Computer Society, 2006. doi: 10.1109/FOCS.2006.37.

Shalev-Shwartz, S., Gonen, A., and Shamir, O. Large-scale convex minimization with a low-rank constraint. In Getoor, L. and Scheffer, T. (eds.), *Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28 - July 2, 2011*, pp. 329–336. Omnipress, 2011.

Shpak, D. J. A weighted-least-squares matrix decomposition method with application to the design of two-dimensional digital filters. In *Proceedings of the 33rd Midwest Symposium on Circuits and Systems*, pp. 1070–1073. IEEE, 1990.

Song, Z., Ye, M., Yin, J., and Zhang, L. Efficient alternating minimization with applications to weighted low rank approximation. *CoRR*, abs/2306.04169, 2023. doi: 10.48550/arXiv.2306.04169.

Srebro, N. and Jaakkola, T. S. Weighted low-rank approximations. In Fawcett, T. and Mishra, N. (eds.), *Machine Learning, Proceedings of the Twentieth International Conference (ICML 2003), August 21-24, 2003, Washington, DC, USA*, pp. 720–727. AAAI Press, 2003.

Tuzhilina, E., Hastie, T. J., and Segal, M. R. Principal curve approaches for inferring 3D chromatin architecture. *Biostatistics*, 23(2):626–642, 2022. ISSN 1465-4644. doi: 10.1093/biostatistics/kxaa046.

Wang, Y. and Singh, A. Provably correct algorithms for matrix column subset selection with selectively sampled data. *J. Mach. Learn. Res.*, 18:156:1–156:42, 2017.

Ye, T. and Du, S. S. Global convergence of gradient descent for asymmetric low-rank matrix factorization. In Ranzato, M., Beygelzimer, A., Dauphin, Y. N., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 1429–1439, 2021.

Young, G. Maximum likelihood estimation and factor analysis. *Psychometrika*, 6:49–53, 1941. ISSN 0033-3123. doi: 10.1007/BF02288574.

A. Missing proofs from Section 4

We provide the standard information-theoretic argument omitted in the proof of Theorem 1.7 in Section 4.

Let M denote the communication transcript between Alice and Bob, and let $\tilde{\mathbf{A}}_{\text{dense}}$ denote the reconstruction of $\mathbf{A}_{\text{dense}}$ based on the approximate weighted low rank approximation solution $\tilde{\mathbf{A}}$, which can be constructed based on M. Recall that the algorithm succeeds in outputting a correct approximation with probability at least $2/3$, so $\tilde{\mathbf{A}}_{\text{dense}} = \mathbf{A}_{\text{dense}}$ with probability at least $2/3$. Then by Fano's inequality (Theorem 2.10.1 of (Cover & Thomas, 2006)), we have that

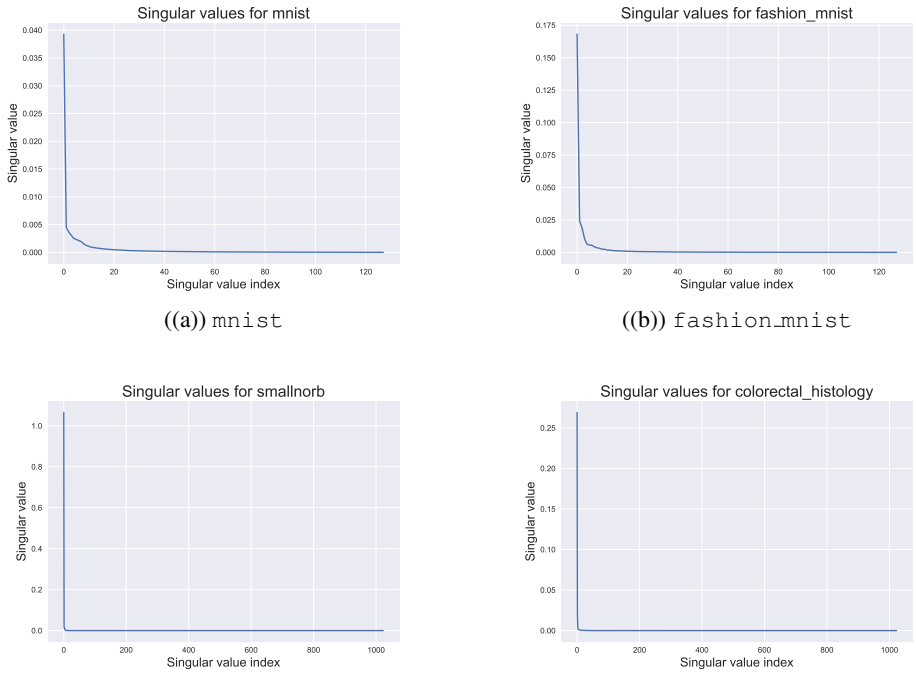
$$H(\mathbf{A}_{\text{dense}} \mid \tilde{\mathbf{A}}_{\text{dense}}) \leq h(1/3) + \frac{1}{3}\log_2|\mathcal{A}| = h(1/3) + \frac{1}{3}srk \tag{1}$$

where $\mathcal{A}$ denotes the support of the random variable $\mathbf{A}_{\text{dense}}$ and h denotes the binary entropy function. It then follows from the data processing inequality (Theorem 2.8.1 of (Cover & Thomas, 2006)) and the previous bound that the message length $B = |M|$ is at least

$$\begin{aligned}
B = |M| &\geq H(M) \geq I(M; \mathbf{A}_{\text{dense}}) \\
&\geq I(\tilde{\mathbf{A}}_{\text{dense}}; \mathbf{A}_{\text{dense}}) && \text{data processing inequality} \\
&= H(\mathbf{A}_{\text{dense}}) - H(\mathbf{A}_{\text{dense}} \mid \tilde{\mathbf{A}}_{\text{dense}}) \\
&\geq rsk - (h(1/3) + srk/3) = \Omega(rsk) && (1).
\end{aligned}$$

B. Additional figures for experiments

In Figure 6, we plot the spectrum of the weight matrices that we consider, showing that the assumption of a low rank weight matrix is a highly practical one.



((a)) `mnist` ((b)) `fashion_mnist`

((c)) `smallnorb` ((d)) `colorectal_histology`

Figure 6: Spectrum of Fisher weight matrices

In Tables 3 through 5, we tabulate the approximation loss and running time of various algorithms for ranks 20, 10, and 5, respectively.

Table 3: Fisher-weighted rank 20 approximation loss and running time

Algorithm	mnist	fashion_mnist	smallnorb	colorectal_histology
svd_w loss	0.0157	0.0689	0.0023	0.0092
adam loss	0.0153	0.0673	0.0281	0.0122
em loss	0.0149	0.0627	0.0019	0.0090
greedy loss	0.0189	0.0767	0.0131	0.0103
sample loss	0.0308	0.1093	0.3978	0.0211
svd loss	0.0203	0.0783	0.0380	0.0143
svd_w time (s)	0.0112	0.0115	1.0586	7.2273
adam time (s)	0.3883	0.3988	10.5479	63.3477
em time (s)	0.2356	0.3183	26.1445	206.1457
greedy time (s)	0.0803	0.0895	6.4177	47.4794
sample time (s)	0.0672	0.0779	4.1263	36.2301
svd time (s)	0.0055	0.0057	0.2831	2.2419

Table 4: Fisher-weighted rank 10 approximation loss and running time

Algorithm	mnist	fashion_mnist	smallnorb	colorectal_histology
svd_w loss	0.0255	0.0967	0.0198	0.0125
adam loss	0.0245	0.0945	0.0348	0.0145
em loss	0.0243	0.0897	0.0202	0.0122
greedy loss	0.0274	0.1017	0.0391	0.0135
sample loss	0.0410	0.1293	0.6094	0.0240
svd loss	0.0335	0.1094	0.0429	0.0146
svd_w time (s)	0.0114	0.0125	0.8855	6.1798
adam time (s)	0.3835	0.4199	14.9386	62.5755
em time (s)	0.2285	0.2776	18.3505	146.7829
greedy time (s)	0.0342	0.0460	3.2980	26.5377
sample time (s)	0.0428	0.0421	2.3308	18.2120
svd time (s)	0.0049	0.0059	0.2339	1.8703

C. Row norm sampling for weighted low rank approximation

We correct an error of a row norm sampling result of (Dai, 2023), and further tighten the result and simplify the proof. The algorithm we study is to repeatedly sample rows according to a *row norm sampling* distribution (Definition C.1).

Definition C.1 (Row norm sampling)**.** Let $\mathbf{A} \in \mathbb{R}^{n\times d}$. Then, a *row norm sample* from this matrix samples the row $\mathbf{e}_i^\top \mathbf{A}$ for $i \in [n]$ with probability

$$p_i = \frac{\|\mathbf{e}_i^\top \mathbf{A}\|_2^2}{\|\mathbf{A}\|_F^2}.$$

We prove the following theorem, which corrects and improves Theorem 3 of (Dai, 2023).

Theorem C.2. *Let $\mathbf{A} \in \mathbb{R}^{n\times d}$ be a rank d matrix and let $\mathbf{W} \in \mathbb{R}^{n\times d}$ be non-negative weights bounded by 1. Let $\mathbf{A}^* \in \mathbb{R}^{n\times d}$ be a rank k matrix satisfying*

$$\|\mathbf{A} - \mathbf{A}^*\|_{\mathbf{W},F}^2 = \min_{\mathrm{rank}(\mathbf{A}')\le k} \|\mathbf{A} - \mathbf{A}'\|_{\mathbf{W},F}^2.$$

Let $T \subseteq [n]$ be a multiset of t indices sampled according to the distribution of Definition C.1. If $t \ge (2\sqrt{10} +$

Table 5: Fisher-weighted rank 5 approximation loss and running time

Algorithm	mnist	fashion_mnist	smallnorb	colorectal_histology
svd_w loss	0.0364	0.1210	0.0872	0.0161
adam loss	0.0355	0.1173	0.1337	0.0180
em loss	0.0349	0.1142	0.1052	0.0155
greedy loss	0.0368	0.1215	0.1393	0.0168
sample loss	0.0470	0.1417	0.7363	0.0239
svd loss	0.0441	0.1312	0.2979	0.0168
svd_w time (s)	0.0094	0.0110	0.8328	6.1427
adam time (s)	0.3890	0.4410	14.6393	63.4094
em time (s)	0.1809	0.2019	18.0931	141.1849
greedy time (s)	0.0185	0.0211	1.7466	15.8465
sample time (s)	0.0320	0.0343	1.4726	11.4331
svd time (s)	0.0039	0.0050	0.2129	1.8317

$1)^2\|\mathbf{A}^*\mathbf{A}^-\|_F^2/\varepsilon^2$*, then with probability at least* $9/10$*, there is a matrix* $\hat{\mathbf{A}} \in \mathbb{R}^{n\times d}$ *with rows spanned by the rows sampled in* T *such that*

$$\|\mathbf{A}-\hat{\mathbf{A}}\|_{\mathbf{W},F}^2 \le \|\mathbf{A}-\mathbf{A}^*\|_{\mathbf{W},F}^2 + \varepsilon\|\mathbf{A}\|_F^2.$$

We make a couple of remarks about this result before showing the proof.

Remark C.1. *If* $\mathbf{W}$ *is all ones and* $\mathbf{A}^* = \mathbf{A}_k$ *is the optimal rank* k *approximation in the standard Frobenius norm given by the SVD, then* $\|\mathbf{A}_k\mathbf{A}^-\|_F^2 = k$ *and thus we recover a result of (Frieze et al., 2004). We empirically estimate the value of* $\|\mathbf{A}_k\mathbf{A}^-\|_F^2$ *for weighted low rank approximation by treating the best solution we find as the "true" solution* $\mathbf{A}^*$*, and we find that this value is* $\le 2k$ *on these datasets.*

Proof of Theorem C.1. Define the random matrix

$$\hat{\mathbf{A}} = \frac{1}{t}\sum_{i\in T}\frac{\mathbf{A}^*\mathbf{A}^-\mathbf{e}_i\mathbf{e}_i^\top\mathbf{A}}{p_i}$$

Note that $\mathbf{E}[\hat{\mathbf{A}}] = \mathbf{A}^*$ and that $\hat{\mathbf{A}}$ is in the row span of the rows of $\mathbf{A}$ sampled by T. Then, the variance of a single sample in T is bounded by

$$\sum_{i=1}^n \frac{1}{p_i}\|\mathbf{A}^*\mathbf{A}^-\mathbf{e}_i\mathbf{e}_i^\top\mathbf{A}\|_F^2 = \sum_{i=1}^n \frac{\|\mathbf{A}\|_F^2}{\|\mathbf{e}_i^\top\mathbf{A}\|_2^2}\|\mathbf{A}^*\mathbf{A}^-\mathbf{e}_i\|_2^2\|\mathbf{e}_i^\top\mathbf{A}\|_2^2 = \|\mathbf{A}^*\mathbf{A}^-\|_F^2\|\mathbf{A}\|_F^2$$

so the variance of $\hat{\mathbf{A}}$ is bounded by

$$\mathbf{Var}(\hat{\mathbf{A}}) = \mathbf{E}\|\mathbf{A}^*-\hat{\mathbf{A}}\|_F^2 \le \frac{1}{t}\|\mathbf{A}^*\mathbf{A}^-\|_F^2\|\mathbf{A}\|_F^2.$$

Thus by Markov's inequality,

$$\|\mathbf{A}^*-\hat{\mathbf{A}}\|_F^2 \le \frac{10}{t}\|\mathbf{A}^*\mathbf{A}^-\|_F^2\|\mathbf{A}\|_F^2$$

with probability at least $9/10$. Then,

$$\begin{aligned}
\|\mathbf{A}-\hat{\mathbf{A}}\|_{\mathbf{W},F} &\le \|\mathbf{A}-\mathbf{A}^*\|_{\mathbf{W},F} + \|\mathbf{A}^*-\hat{\mathbf{A}}\|_{\mathbf{W},F} && \text{triangle inequality}\\
&\le \|\mathbf{A}-\mathbf{A}^*\|_{\mathbf{W},F} + \|\mathbf{A}^*-\hat{\mathbf{A}}\|_F\\
&\le \|\mathbf{A}-\mathbf{A}^*\|_{\mathbf{W},F} + \frac{\sqrt{10}}{\sqrt{t}}\|\mathbf{A}^*\mathbf{A}^-\|_F\|\mathbf{A}\|_F.
\end{aligned}$$

Squaring both sides yields

$$\begin{aligned}
\|\mathbf{A}-\hat{\mathbf{A}}\|_{\mathbf{W},F}^2 &\leq \|\mathbf{A}-\mathbf{A}^*\|_{\mathbf{W},F}^2 + 2\|\mathbf{A}-\mathbf{A}^*\|_{\mathbf{W},F}\frac{\sqrt{10}}{\sqrt{t}}\|\mathbf{A}^*\mathbf{A}^-\|_F\|\mathbf{A}\|_F + \frac{1}{t}\|\mathbf{A}^*\mathbf{A}^-\|_F^2\|\mathbf{A}\|_F^2 \\
&\leq \|\mathbf{A}-\mathbf{A}^*\|_{\mathbf{W},F}^2 + 2\frac{\sqrt{10}}{\sqrt{t}}\|\mathbf{A}^*\mathbf{A}^-\|_F\|\mathbf{A}\|_F^2 + \frac{1}{t}\|\mathbf{A}^*\mathbf{A}^-\|_F^2\|\mathbf{A}\|_F^2 \\
&= \|\mathbf{A}-\mathbf{A}^*\|_{\mathbf{W},F}^2 + \left(2\frac{\sqrt{10}}{\sqrt{t}}\|\mathbf{A}^*\mathbf{A}^-\|_F + \frac{1}{t}\|\mathbf{A}^*\mathbf{A}^-\|_F^2\right)\|\mathbf{A}\|_F^2 \\
&\leq \|\mathbf{A}-\mathbf{A}^*\|_{\mathbf{W},F}^2 + (2\sqrt{10}+1)\frac{\|\mathbf{A}^*\mathbf{A}^-\|_F}{\sqrt{t}}\|\mathbf{A}\|_F^2 \\
&\leq \|\mathbf{A}-\mathbf{A}^*\|_{\mathbf{W},F}^2 + \varepsilon\|\mathbf{A}\|_F^2.
\end{aligned}$$

□