Regularized Weighted Low Rank Approximation

Frank Ban
UC Berkeley / Google
fban@berkeley.edu

David Woodruff
Carnegie Mellon University
dwoodruf@cs.cmu.edu

Qiuyi (Richard) Zhang
UC Berkeley / Google
qiuyi@berkeley.edu

Abstract

The classical low rank approximation problem is to find a rank k matrix UV (where U has k columns and V has k rows) that minimizes the Frobenius norm of $A - UV$. Although this problem can be solved efficiently, we study an NP-hard variant of this problem that involves weights and regularization. A previous paper of [Razenshteyn et al. '16] derived a polynomial time algorithm for weighted low rank approximation with constant rank. We derive provably sharper guarantees for the regularized version by obtaining parameterized complexity bounds in terms of the statistical dimension rather than the rank, allowing for a rank-independent runtime that can be significantly faster. Our improvement comes from applying sharper matrix concentration bounds, using a novel conditioning technique, and proving structural theorems for regularized low rank problems.

1 Introduction

In the weighted low rank approximation problem, one is given a matrix $M \in \mathbb{R}^{n\times d}$, a weight matrix $W \in \mathbb{R}^{n\times d}$, and an integer parameter k, and would like to find factors $U \in \mathbb{R}^{n\times k}$ and $V \in \mathbb{R}^{k\times d}$ so as to minimize

$$\|W \circ (M - U \cdot V)\|_F^2 = \sum_{i=1}^{n}\sum_{j=1}^{d} W_{i,j}^2 (M_{i,j} - \langle U_{i,*}, V_{*,j}\rangle)^2,$$

where $U_{i,*}$ denotes the i-th row of U and $V_{*,j}$ denotes the j-th column of V. W.l.o.g., we assume $n \geq d$. This is a weighted version of the classical low rank approximation problem, which is a special case when $W_{i,j} = 1$ for all i and j. One often considers the approximate version of this problem, for which one is given an approximation parameter $\varepsilon \in (0, 1)$ and would like to find $U \in \mathbb{R}^{n\times k}$ and $V \in \mathbb{R}^{k\times d}$ so that

$$\|W \circ (M - U \cdot V)\|_F^2 \leq (1 + \varepsilon) \min_{U' \in \mathbb{R}^{n\times k}, V' \in \mathbb{R}^{k\times d}} \|W \circ (M - U' \cdot V')\|_F^2. \tag{1}$$

Weighted low rank approximation extends the classical low rank approximation problem in many ways. While in principal component analysis, one typically first subtracts off the mean to make the matrix M have mean 0, this does not fix the problem of differing variances. Indeed, imagine one of the columns of M has much larger variance than the others. Then in classical low rank approximation with $k = 1$, it could suffice to simply fit this single high variance column and ignore the remaining entries of M. Weighted low rank approximation, on the other hand, can correct for this by re-weighting each of the entries of M to give them similar variance; this allows for the low rank approximation $U \cdot V$ to capture more of the remaining data. This technique is often used in gene expression analysis and co-occurrence matrices; we refer the reader to [SJ03] and the Wikipedia entry on weighted low rank approximation[1]. The well-studied problem of *matrix completion* is

[1] https://en.wikipedia.org/wiki/Low-rank_approximation#Weighted_low-rank_approximation_problems

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

also a special case of weighted low rank approximation, where the entries $W_{i,j}$ are binary, and has numerous applications in recommendation systems and other settings with missing data.

Unlike its classical variant, weighted low rank approximation is known to be NP-hard [GG11]. Classical low rank approximation can be solved quickly via the singular value decomposition, which is often sped up with sketching techniques [Woo14, PW15, TYUC17]. However, in the weighted setting, under a standard complexity-theoretic assumption known as the Random Exponential Time Hypothesis (see, e.g., Assumption 1.3 in [RSW16] for a discussion), there is a fixed constant $\varepsilon_0 \in (0, 1)$ for which any algorithm achieving (1) with constant probability and for $\varepsilon = \varepsilon_0$, and even for $k = 1$, requires $2^{\Omega(r)}$ time, where r is the number of distinct columns of the weight matrix W. Furthermore, as shown in Theorem 1.4 of [RSW16], this holds even if W has both at most r distinct rows and r distinct columns.

Despite the above hardness, in a number of applications the parameter r may be small. Indeed, in a matrix in which the rows correspond to users and the columns correspond to ratings of a movie, such as in the Netflix matrix, one may have a small number of categories of movies. In this case, one may want to use the same column in W for each movie in the same category. It may thus make sense to renormalize user ratings based on the category of movies being watched. Note that any number of distinct rows of W is possible here, as different users may have completely different ratings, but there is just one distinct column of W per category of movie. In some settings one may simultaneously have a small number of distinct rows and a small number of distinct columns. This may occur if say, the users are also categorized into a small number of groups. For example, the users may be grouped by age and one may want to weight ratings of different categories of movies based on age. That is, maybe cartoon ratings of younger users should be given higher weight, while historical films rated by older users should be given higher weight.

Motivated by such applications when r is small, [RSW16] propose several *parameterized complexity algorithms*. They show that in the case that W has at most r distinct rows and r distinct columns, there is an algorithm solving (1) in $2^{O(k^2 r/\varepsilon)}\mathrm{poly}(n)$ time. If W has at most r distinct columns but any number of distinct rows, there is an algorithm achieving (1) in $2^{O(k^2 r^2/\varepsilon)}\mathrm{poly}(n)$ time. Note that these bounds imply that for constant k and ε, even if r is as large as $\Theta(\log n)$ in the first case, and $\Theta(\sqrt{\log n})$ in the second case, the corresponding algorithm is polynomial time.

In [RSW16], the authors also consider the case when the rank of the weight matrix W is at most r, which includes the r distinct rows and columns, as well as the r distinct column settings above, as special cases. In this case the authors achieve an $n^{O(k^2 r/\varepsilon)}$ time algorithm. Note that this is only polynomial time if k, r, and ε are each fixed constants, and unlike the algorithms for the other two settings, this algorithm is not fixed parameter tractable, meaning its running time cannot be written as $f(k, r, 1/\varepsilon) \cdot \mathrm{poly}(nd)$, where f is a function that is independent of n and d.

There are also other algorithms for weighted low rank approximation, though they do not have provable guarantees, or require strong assumptions on the input. There are gradient-based approaches of Shpak [Shp90] and alternating minimization approaches of Lu et al. [LPW97, LA03], which were refined and used in practice by Srebro and Jaakkola [SJ03]. However, neither of these has provable gurantees. While there is some work that has provable guarantees, it makes incoherence assumptions on the low rank factors of M, as well as assumptions that the weight matrix W is spectrally close to the all ones matrix [LLR16] and that there are no 0 weights.

1.1 Our Contributions

The only algorithms with provable guarantees that do not make assumptions on the inputs are slow, and inherently so given the above hardness results. Motivated by this and the widespread use of regularization in machine learning, we propose to look at *provable guarantees for regularized weighted low rank approximation*. In one version of this problem, where the parameter r corresponds to the rank of the weight matrix W, we are given a matrix $M \in \mathbb{R}^{n\times d}$, a weight matrix $W \in \mathbb{R}^{n\times d}$ with rank r, and a target integer $k > 0$, and we consider the problem

$$\min_{U\in\mathbb{R}^{n\times k}, V\in\mathbb{R}^{k\times d}} \|W \circ (UV - M)\|_F^2 + \lambda\|U\|_F^2 + \lambda\|V\|_F^2$$

Let U^*, V^* minimize $\|W \circ (UV - M)\|_F^2 + \lambda\|U\|_F^2 + \lambda\|V\|_F^2$ and OPT be the minimum value.

Regularization is a common technique to avoid overfitting and to solve an ill-posed problem. It has been applied in the context of weighted low rank approximation [DN11], though so far the only such results known for weighted low rank approximation with regularization are heuristic. In this paper we give the first provable bounds, without any assumptions on the input, on regularized weighted low rank approximation.

Importantly, we show that regularization improves our running times for weighted low rank approximation, as specified below. Intuitively, the complexity of regularized problems depends on the "statistical dimension" or "effective dimension" of the underlying problem, which is often significantly smaller than the number of parameters in the regularized setting.

Let U^* and V^* denote the optimal low-rank matrix approximation factors, $D_{W_{i,:}}$ denote the diagonal matrix with the i-th row of W along the diagonal, and $D_{W_{:,j}}$ denote the diagonal matrix with the j-th column of W along the diagonal.

Improving the Exponent: We first show how to improve the $n^{O(k^2 r/\varepsilon)}$ time algorithm of [RSW16] to a running time of $n^{O((s+\log(1/\varepsilon))rk/\varepsilon)}$. Here s is defined to be the maximum statistical dimension of $V^* D_{W_{i,:}}$ and $D_{W_{:,j}} U^*$, over all $i = 1, \ldots, n$, and $j = 1, \ldots, d$, where the statistical dimension of a matrix M is:

Definition 1. *Let $\mathtt{sd}_\lambda(M) = \sum_i 1/(1 + \lambda/\sigma_i^2)$ denote the statistical dimension of M with regularizing weight λ (here σ_i are the singular values of M).*

Note that this maximum value s is always at most k and for any $s \geq \log(1/\varepsilon)$, our bound directly improves upon the previous time bound. Our improvement requires us to sketch matrices with k columns down to s/ε rows where s/ε is potentially smaller than k. This is particularly interesting since most previous provable sketching results for low rank approximation cannot have sketch sizes that are smaller than the rank, as following the standard analysis would lead to solving a regression problem on a matrix with fewer rows than columns.

Thus, we introduce the notion of an upper and lower distortion factor (K_S and κ_S below) and show that the lower distortion factor will satisfy tail bounds only on a smaller-rank subspace of size s/ε, which can be smaller than k. Directly following the analysis of [RSW16] will cause the lower distortion factor to be infinite. The upper distortion factor also satisfies tail bounds via a more powerful matrix concentration result not used previously. Furthermore, we apply a novel conditioning technique that conditions on the product of the upper and lower distortion factors on separate subspaces, whereas previous work only conditions on the condition number of a specific subspace.

We next considerably strengthen the above result by showing an $n^{O(r^2(s+\log(1/\varepsilon))^2/\varepsilon^2))}$ time algorithm. This shows that the rank k need not be in the exponent of the algorithm at all! We do this via a novel projection argument in the objective (sketching on the right), which was not done in [RSW16] and also improves a previous result for the classical setting in [ACW17]. To gain some perspective on this result, suppose ε is a large constant, close to 1, and r is a small constant. Then our algorithm runs in $n^{O(s^2)}$ time as opposed to the algorithm of [RSW16] which runs in $n^{O(k^2)}$ time. We stress in a number of applications, the effective dimension s may be a very small constant, close to 1, even though the rank parameter k can be considerably larger. This occurs, for example, if there is a single dominant singular value, or if the singular values are geometrically decaying. Concretely, it is realistic that k could be $\Theta(\log n)$, while $s = \Theta(1)$, in which case our algorithm is the first polynomial time algorithm for this setting.

Improving the Base: We can further optimize by removing our dependence on n in the base. The *non-negative rank* of a $n \times d$ matrix A is defined to be the least r such that there exist factors $U \in \mathbb{R}^{n \times r}$ and $V \in \mathbb{R}^{r \times d}$ where $A = U \cdot V$ and both U and V have non-negative entries. By applying a novel rounding procedure, if in addition the non-negative rank of W is at most r', then we can obtain a fixed-parameter tractable algorithm running in time $2^{r' r^2(s+\log(1/\varepsilon))^2/\varepsilon^2)}\mathrm{poly}(n)$. Note that $r \leq r'$, where r is the rank of W. Note also that if W has at most r distinct rows or columns, then its non-negative rank is also at most r since we can replace the entries of W with their absolute values without changing the objective function, while still preserving the property of at most r distinct rows and/or columns. Consequently, we significantly improve the algorithms for a small number of distinct rows and/or columns of [RSW16], as our exponent is *independent* of k.

Thus, even if $k = \Theta(n)$ but the statistical dimension $s = O(\sqrt{\log n})$, for constant r' and ε our algorithm is polynomial time, while the best previous algorithm would be exponential time.

We also give ways, other than non-negative rank, for improving the running time. Supposing that the rank of W is r again, we apply iterative techniques in linear system solving like Richardson's Iteration and preconditioning to further improve the running time. We are able to show that instead of an $n^{\mathrm{poly}(rs/\varepsilon)}$ time algorithm, we are able to obtain algorithms that have running time roughly $(\sigma^2/\lambda)^{\mathrm{poly}(rs/\varepsilon)}\mathrm{poly}(n)$ or $(u_W/l_W)^{\mathrm{poly}(rs/\varepsilon)}\mathrm{poly}(n)$, where σ^2 is defined to be the maximum singular value of $V^* D_{W_{i,:}}$ and $D_{W_{:,j}} U^*$, over all $i = 1, \ldots, n$, and $j = 1, \ldots, d$, while u_W is defined to be the maximum absolute value of an entry of W and l_W the minimum absolute value of an entry. In a number of settings one may have $\sigma^2/\lambda = O(1)$ or $u_W/l_W = O(1)$ in which case we again obtain fixed parameter tractable algorithms.

Empirical Evaluation: Finally, we give an empirical evaluation of our results. While the main goal of our work is to obtain the first algorithms with provable guarantees for regularized weighted low rank approximation, we can also use them to guide heuristics in practice. In particular, alternating minimization is a common heuristic for weighted low rank approximation. We consider a sketched version of alternating minimization to speed up each iteration. We show that in the regularized case, the dimension of the sketch can be significantly smaller if the statistical dimension is small, which is consistent with our theoretical results.

2 Preliminaries

We let $\|\cdot\|_F$ denote the Frobenius norm of a matrix and let $\circ$ be the elementwise matrix multiplication operator. We denote $x \in [a, b]\, y$ if $ay \le x \le by$. For a matrix M, let $M_{i,;}$ denote its ith row and let $M_{;,j}$ denote its jth column. For $v \in \mathbb{R}^n$, let D_v denote the $n \times n$ diagonal matrix with its i-th diagonal entry equal to v_i. For a matrix M with non-negative M_{ij}, let $\mathtt{nnr}(M)$ denote the non-negative rank of M. Let $\mathtt{sr}(M) = \|M\|_F^2/\|M\|^2$ denote the stable rank of M. Let $\mathcal{D}$ denote a distribution over $r \times n$ matrices; in our setting, there are matrices with entries that are Gaussian random variables with mean 0 and variance $1/r$ (or $r \times n$ CountSketch matrices [Woo14]).

Definition 2. *For S sampled from a distribution of matrices $\mathcal{D}$ and a matrix M with n rows, let $c_S(M) \ge 1$ denote the smallest (possibly infinite) number such that $\|SMv\|^2 \in [c_S(M)^{-1}, c_S(M)]\|Mv\|^2$ for all v.*

Definition 3. *For S sampled from a distribution of matrices $\mathcal{D}$ and a matrix M, let $K_S(M) \ge 1$ denote the smallest number such that $\|SMv\|^2 \le K_S(M)\|Mv\|^2$ for all v.*

Definition 4. *For S sampled from a distribution of matrices $\mathcal{D}$ and a matrix M, let $\kappa_S(M) \le 1$ denote the largest number such that $\|SMv\|^2 \ge \kappa_S(M)\|Mv\|^2$ for all v.*

Note that by definition, $c_s(M)$ equals the max of $K_S(M)$ and $\frac{1}{\kappa_S(M)}$. We define the condition number of a matrix A to be $c_A = K_A(I)/\kappa_A(I)$.

2.1 Previous Techniques

Building upon the initial framework established in [RSW16], we apply a polynomial system solver to solve weighted regularized LRA to high accuracy. By applying standard sketching guarantees, v can be made a polynomial function of $k, 1/\varepsilon, r$ that is independent of n.

Theorem 1 ([Ren92a][Ren92b][BPR96])**.** *Given a real polynomial system $P(x_1, x_2, ..., x_v)$ having v variables and m polynomial constraints $f_i(x_1, ..., x_v)\Delta_i 0$, where $\Delta_i \in \{\ge, =, \le\}$, d is the maximum degree of all polynomials, and H is the maximum bitsize of the coefficients of the polynomials, one can determine if there exists a solution to P in $(md)^{O(v)}poly(H)$ time.*

Intuitively, the addition of regularization requires us to only preserve directions with high spectral weight in order to preserve our low rank approximation well enough. This dimension of the subspace spanned by these important directions is exactly the statistical dimension of the problem, allowing us to sketch to a size less than k that could provably preserve our low rank approximation well enough. In line with this intuition, we use an important lemma from [CNW16]

Lemma 2.1. *Let A, B be matrices with n rows and let S, sampled from $\mathcal{D}$, have $\ell = \Omega(\frac{1}{\gamma^2}(K + \log(1/\varepsilon)))$ rows and n columns. Then*

$$\mathbf{Pr}\left[\|A^T S^T S B - A^T B\| > \gamma \cdot \sqrt{(\|A\|^2 + \|A\|_F^2/K)\,(\|B\|^2 + \|B\|_F^2/K)}\right] < \varepsilon$$

In particular, if we choose $K > \Omega(\mathit{sr}(A) + \mathit{sr}(B))$, then we have for some small constant γ',

$$\mathbf{Pr}\left[\|A^T S^T S B - A^T B\| > \gamma' \|A\| \|B\|\right] < \varepsilon$$

3 Multiple Regression Sketches

In this section, we prove our main structural theorem which allows us to sketch regression matrices to the size of the statistical dimension of the matrices while maintaining provable guarantees. Specifically, to approximately solve a sum of regression problems, we are able to reduce the dimension of the problem to the maximum statistical dimension of the regression matrices.

Theorem 2. *Let $M^{(1)}, \ldots, M^{(d)} \in \mathbb{R}^{n\times k}$ and $b^{(1)}, \ldots, b^{(d)} \in \mathbb{R}^n$ be column vectors. Let $S \in \mathbb{R}^{\ell\times n}$ be sampled from $\mathcal{D}$ with $\ell = \Theta(\frac{1}{\varepsilon}(s + \log(1/\varepsilon)))$ and $s = \max_i\{\mathit{sd}_\lambda(M^{(i)})\}$.*

Define $x^{(i)} = \operatorname{argmin}_x \|M^{(i)}x - b^{(i)}\|^2 + \lambda\|x\|^2$ and $y^{(i)} = \operatorname{argmin}_y \|S(M^{(i)}y - b^{(i)})\|^2 + \lambda\|y\|^2$. Then, with constant probability,

$$\sum_{i=1}^{d} \|M^{(i)}y^{(i)} - b^{(i)}\|^2 + \lambda\|y^{(i)}\|^2 \le (1+\varepsilon)\cdot\left(\sum_{i=1}^{d} \|M^{(i)}x^{(i)} - b^{(i)}\|^2 + \lambda\|x^{(i)}\|^2\right)$$

We note that a simple union bound would incur a dependence of a factor of $\log(d)$ in the sketching dimension l. While this might seem mild at first, the algorithms we consider are exponential in l, implying that we would be unable to derive polynomial time algorithms for solving weighted low rank approximation even when the input and weight matrix are both of constant rank. Therefore, we need an average case version of sketching guarantees to hold; however, this is not always the case since l is small and applying Lemma 2.1 naïvely only gives a probability bound. Ultimately, we must condition on the event of a combination of sketching guarantees holding and carefully analyzing the expectation in separate cases.

4 Algorithms

In this section, we present a fast algorithm for solving regularized weighted low rank approximation. Our algorithm exploits the structure of low-rank approximation as a sum of regression problems and applies the main structural theorem of our previous section to significantly reduce the number of variables in the optimization process. Note that we can write

$$\|W \circ (UV - A)\|_F^2 = \sum_{i=1}^{n} \|U_{i,;}VD_{W_{i,;}} - A_{i,;}D_{W_{i,;}}\|^2 = \sum_{j=1}^{d} \|D_{W_{;,j}}UV_{;,j} - D_{W_{;,j}}A_{;,j}\|^2$$

4.1 Using the Polynomial Solver with Sketching

Now we sample Gaussian sketch matrices S' from $\mathbb{R}^{d\times\Theta(\frac{s}{\varepsilon})\log(1/\varepsilon)}$ and S'' from $\mathbb{R}^{\Theta(\frac{s}{\varepsilon})\log(1/\varepsilon)\times n}$. We let $P^{(i)}$ denote $VD_{W_{i,;}}S'$ and $Q^{(j)}$ denote $S''D_{W_{;,j}}U$.

The matrices $P^{(i)}$ and $Q^{(j)}$ can be encoded using $\Theta(\frac{s+\log(1/\varepsilon)}{\varepsilon})kr$ variables. For fixed $P^{(i)}$ and $Q^{(j)}$ we can define

$$\tilde{U} = \underset{U\in\mathbb{R}^{n\times k}}{\operatorname{argmin}} \sum_{i=1}^{n} \|U_{i,;}P^{(i)} - A_{i,;}D_{W_{i,;}}S'\|^2 + \lambda\|U_{i,;}\|^2$$

and

$$\tilde{V} = \underset{V\in\mathbb{R}^{k\times n}}{\operatorname{argmin}} \sum_{j=1}^{d} \|Q^{(j)}V_{;,j} - S''D_{W_{;,j}}A_{;,j}\|^2 + \lambda\|V_{;,j}\|^2$$

Algorithm 1 Regularized Weighted Low Rank Approximation

```
public : procedure RegWeightedLowRank($A, W, \lambda, s, k, \varepsilon$)
    Sample Gaussian sketch $S' \in \mathbb{R}^{d \times \Theta(\frac{s}{\varepsilon}) \log(1/\varepsilon)}$ from $\mathcal{D}$
    Sample Gaussian sketch $S'' \in \mathbb{R}^{\Theta(\frac{s}{\varepsilon}) \log(1/\varepsilon) \times n}$ from $\mathcal{D}$.
    Create matrix variables $P^{(i)} \in \mathbb{R}^{k \times \Theta(\frac{s}{\varepsilon}) \log(1/\varepsilon)}$, $Q^{(j)} \in \mathbb{R}^{k \times \Theta(\frac{s}{\varepsilon}) \log(1/\varepsilon)}$ for $i, j$ from 1 to $r$
                                                        ▷ Variables used in polynomial system solver

    Use Cramer's Rule to express $\tilde{U}_{i,;} = A_{i,;} D_{W_{i,;}} S' (P^{(i)})^T (P^{(i)} (P^{(i)})^T + \lambda I_k)^{-1}$ as a rational
    function of variables $P^{(i)}$; similarly, $\tilde{V}_{;,j} = ((Q^{(j)})^T Q^{(j)} + \lambda I_k)^{-1} (Q^{(j)})^T S'' D_{W_{;,j}} A_{;,j}$
                                                        ▷ $\tilde{U}, \tilde{V}$ are now rational function of variables $P, Q$

    Solve $\min_{\tilde{U}, \tilde{V}} \|W \circ (\tilde{U}\tilde{V} - A)\|_F^2 + \lambda \|\tilde{U}\|_F^2 + \lambda \|\tilde{V}\|_F^2$ and apply binary search to find $\tilde{U}, \tilde{V}$
                                                        ▷ Optimization with polynomial solver of Theorem 1 in variables $P, Q$
return $\tilde{U}, \tilde{V}$
```

to get

$$\tilde{U}_{i,;} = A_{i,;} D_{W_{i,;}} S' (P^{(i)})^T (P^{(i)} (P^{(i)})^T + \lambda I_k)^{-1}$$

and

$$\tilde{V}_{;,j} = ((Q^{(j)})^T Q^{(j)} + \lambda I_k)^{-1} (Q^{(j)})^T S'' D_{W_{;,j}} A_{;,j}$$

so $\tilde{U}$ and $\tilde{V}$ can be encoded as rational functions over $\Theta(\frac{(s+\log(1/\varepsilon))kr}{\varepsilon})$ variables by Cramer's Rule.

By Theorem 2, we can argue that $\min_{\tilde{U},\tilde{V}} \|W \circ (\tilde{U}\tilde{V} - A)\|_F^2 + \lambda \|\tilde{U}\|_F^2 + \lambda \|\tilde{V}\|_F^2$ is a good approximation for $\|W \circ (U^* V^* - A)\|_F^2 + \lambda \|U^*\|_F^2 + \lambda \|V^*\|_F^2$ with constant probability, and so in particular such a good approximation exists. By using the polynomial system feasibility checkers described in Theorem 1 and following similar procedures and doing binary search, we get an polynomial system with $O(nk)$-degree and $O(\frac{s+\log(1/\varepsilon)}{\varepsilon} kr)$ variables after simplifying and so our polynomial solver runs in time $n^{O((s+\log(1/\varepsilon))kr/\varepsilon)} \log^{O(1)}(\Delta/\delta)$.

Theorem 3. *Given matrices $A, W \in \mathbb{R}^{n \times d}$ and $\varepsilon < 0.1$ such that*

1. *rank(W) = r*
2. *non-zero entries of A, W are multiples of $\delta > 0$*
3. *all entries of A, W are at most Δ in absolute value*
4. *$s = \max_{i,j}\{\mathtt{sd}_\lambda(V^* D_{W_{i,;}}), \mathtt{sd}_\lambda(D_{W_{;,j}} U^*)\} < k$*

there is an algorithm to find $U \in \mathbb{R}^{n \times k}, V \in \mathbb{R}^{k \times d}$ in time $n^{O((s+\log(1/\varepsilon))kr/\varepsilon)} \log^{O(1)}(\Delta/\delta)$ such that $\|W \circ (UV - A)\|_F^2 + \lambda \|U\|_F^2 + \lambda \|V\|_F^2 \leq (1+\varepsilon)\mathrm{OPT}$.

4.2 Removing Rank Dependence

Note that the running time of our algorithm still depends on k, the dimension that we are reducing to. To remove this, we prove a structural theorem about low rank approximation of low statistical dimension matrices.

Theorem 4. *Given matrices A, W in $\mathbb{R}^{n \times d}$ and $\varepsilon < 0.1$ such that rank(W) is r, and letting s equal $\max_{i,j}\{\mathtt{sd}_\lambda(V^* D_{W_{i,;}}), \mathtt{sd}_\lambda(D_{W_{;,j}} U^*)\} < k$, if we let* OPT*(k) denote*

$$\min_{U \in \mathbb{R}^{n \times k}, V \in \mathbb{R}^{k \times d}} \|W \circ (UV - A)\|_F^2 + \lambda \|U\|_F^2 + \lambda \|V\|_F^2$$

then $\mathrm{OPT}(O(r(s + \log(1/\varepsilon))/\varepsilon)) \leq (1+\varepsilon)\mathrm{OPT}(k)$

Combining Theorem 3 and Theorem 4, we have our final theorem. We note that this also improves running time bounds of un-weighted regularized low rank approximation in Section 3 of [ACW17].

Theorem 5. *Given matrices $A, W \in \mathbb{R}^{n \times d}$ and $\varepsilon < 0.1$ and the conditions of Theorem 3, there is an algorithm to find $U \in \mathbb{R}^{n \times k}, V \in \mathbb{R}^{k \times d}$ in time $n^{O(r^2(s+\log(1/\varepsilon))^2/\varepsilon^2)} \log^{O(1)}(\Delta/\delta)$ such that $\|W \circ (UV - A)\|_F^2 + \lambda\|U\|_F^2 + \lambda\|V\|_F^2 \leq (1+\varepsilon)\mathrm{OPT}$.*

5 Reducing the Degree of the Solver

5.1 Non-negative Weight Matrix and Non-Negative Rank

Under the case where W is rank r with only r distinct columns (up to scaling), we are able to improve the running time to $poly(n)2^{r^3(s+\log(1/\varepsilon))^2/\varepsilon^2}$ by showing that the degree of the solver is $O(rk)$ as opposed to $O(nk)$. Specifically, the $O(nk)$ degree comes from clearing the denominator of the rational expressions that come from naïvely using and analyzing Cramer's Rule; in this section, we demonstrate different techniques to avoid the dependence on n. We also show the same running time bound under a more relaxed assumption of non-negative rank, which is always less than or equal to the number of distinct columns.

Theorem 6. *Given matrices $A, W \in \mathbb{R}^{n \times d}$ and $\varepsilon < 0.1$ and suppose the conditions of Theorem 3 hold. Furthermore, we are given $Y, Z \geq 0$ such that $W = YZ$ and Y, Z^T has $\mathfrak{nnr}(W) = r'$ columns.*

Then, there is an algorithm to find $U \in \mathbb{R}^{n \times k}$, $V \in \mathbb{R}^{k \times d}$ in time $poly(n) \cdot 2^{O(r'r^2(s+\log(\frac{1}{\varepsilon}))^2\frac{1}{\varepsilon^2})} \cdot \log^{O(1)}\left(\frac{\Delta}{\delta}\right)$ such that $\|W \circ (UV - A)\|_F^2 + \lambda\|U\|_F^2 + \lambda\|V\|_F^2 \leq (1+\varepsilon)\mathrm{OPT}$.

5.2 Richardson's Iteration

Note that the current polynomial solver uses Cramer's rule to solve

$$\tilde{U} = \operatorname*{argmin}_{U \in \mathbb{R}^{n \times k}} \sum_{i=1}^{n} \|U_{i,:}P^{(i)} - A_{i,:}D_{W_{i,:}}S'\|^2 + \lambda\|U_{i,:}\|^2$$

giving

$$\tilde{U}_{i,:} = A_{i,:}D_{W_{i,:}}S'(P^{(i)})^T(P^{(i)}(P^{(i)})^T + \lambda I_k)^{-1}.$$

We want to use Richardson's iteration instead to avoid rational expressions and the dependence on n in the degree that comes from clearing the denominator.

Theorem 7 (Preconditioned Richardson [CKP+17])**.** *Let A, B be symmetric PSD matrices such that $ker(A) = ker(B)$ and $\eta A \preceq B \preceq A$. Then, for any b, if $x_0 = 0$ and $x_{i+1} = x_i - \eta B^{-1}(Ax_i - b)$,*

$$\|x_t - A^{-1}b\| \leq \varepsilon\|A^{-1}b\|$$

for $t = \Omega(\log(c_B/\varepsilon)/\eta)$. Furthermore, we may express x_t as a polynomial of degree $O(t)$ in terms of the entries of B^{-1} and A.

Theorem 8. *Given matrices $A, W \in \mathbb{R}^{n \times d}$ and $\varepsilon < 0.1$ and suppose the conditions of Theorem 3 hold. Furthermore, let $\sigma = \max_{i,j}\{\sigma_1(V^* D_{W_{i,:}}), \sigma_1(D_{W_{:,j}}U^*)\}$.*

There is an algorithm to find $U \in \mathbb{R}^{n \times k}$, $V \in \mathbb{R}^{k \times d}$ in time $poly(n)\left(\frac{\sigma^2}{\lambda} \cdot \log\left(\frac{\Delta(\sigma^2+\lambda)n}{\lambda\tau}\right)\right)^l \cdot \log^{O(1)}\left(\frac{\Delta}{\delta}\right)$, where $l = O((s+\log(\frac{1}{\varepsilon}))^2\frac{r^2}{\varepsilon^2})$, such that $\|W \circ (UV - A)\|_F^2 + \lambda\|U\|_F^2 + \lambda\|V\|_F^2 \leq (1+\varepsilon)\mathrm{OPT} + \tau$.

5.3 Preconditioned GD

Instead of directly using Richardson's iteration, we may use a preconditioner first instead. The right preconditioner can also be guessed at a cost of increasing the number of variables. Note that multiple preconditioners may be used, but for now, we demonstrate the power of a single preconditioner.

Theorem 9. *Given matrices $A, W \in \mathbb{R}^{n \times d}$ and $\varepsilon < 0.1$ and suppose the conditions of Theorem 8 hold. Furthermore, $0 < l_W \leq |W| \leq u_W$. Then, there is an algorithm to find $U \in \mathbb{R}^{n \times k}, V \in \mathbb{R}^{k \times d}$ in time $poly(n) \cdot \left(\frac{u_W}{l_W} \cdot \log\left(\frac{\Delta(\sigma^2+\lambda)n}{\lambda\tau}\right)\right)^l \cdot \log^{O(1)}\left(\frac{\Delta}{\delta}\right)$, where $l = O((s+\log(\frac{1}{\varepsilon}))^2\frac{r^2}{\varepsilon^2})$, such that $\|W \circ (UV - A)\|_F^2 + \lambda\|U\|_F^2 + \lambda\|V\|_F^2 \leq (1+\varepsilon)\mathrm{OPT} + \tau$.*

6 Experiments

The goal of our experiments was to show that sketching down to the statistical dimension can be applied to regularized weighted low rank approximation without sacrificing overall accuracy in the objective function, as our theory predicts. We combine sketching with a common practical alternating minimization heuristic for solving regularized weighted low rank approximation, rather than implementing a polynomial system solver. At each step in the algorithm, we have a candidate U and V and we perform a "best-response" where we either update U to give the best regularized weighted low rank approximation cost for V or we update V to give the best regularized weighted low rank approximation cost for U. We used a synthetic dataset and several real datasets (connectus, NIPS, landmark, and language) [DH11, PJST17]. All our experiments ran on a MacBook Pro 2012 with 8GB RAM and a 2.5GHz Intel Core i5 processor.

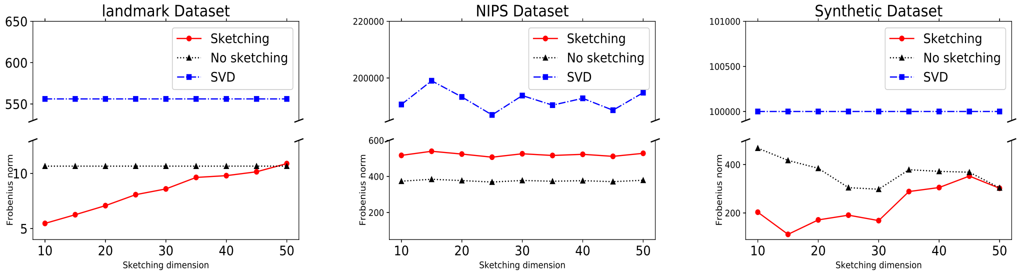

Figure 1: Regularized weighted low-rank approximations with $\lambda = 0.556$ for landmark, $\lambda = 314$ for NIPS, and $\lambda = 1$ for the synthetic dataset.

For all datasets, the task was to find a rank $k = 50$ decomposition of a given matrix A. For the experiments of Figure 1 and Figure 2, we generated dense weight matrices W with the same shape as A and with each entry being a 1 with probability 0.8, a 0.1 with probability 0.15, and a 0.01 with probability 0.05. For the experiments of Figure 3, we generated binary weight matrices where each entry was 1 with probability 0.9. Note that this setting corresponds to a regularized form of matrix completion. We set the regularization parameter λ to a variety of values (described in the Figure captions) to illustrate the performance of the algorithm in different settings.

For the synthetic dataset, we generated matrices A with dimensions 10000×1000 by picking random orthogonal vectors as its singular vectors and having one singular value equal to 10000 and making the rest small enough so that the statistical dimension of A would be approximately 2.

For the real datasets, we chose the connectus, landmark, and language datasets [DH11] and the NIPS dataset [PJST17]. We sampled 1000 rows from each adjacency or word count matrix to form a matrix B and then let A be the radial basis function kernel of B. We performed three algorithms on each dataset: Singular Value Decomposition, Alternating Minimization without Sketching, and Alternating Minimization with Sketching. We parameterized the experiments by t, the sketch size, which took values in $\{10, 15, 20, 25, 30, 35, 40, 45, 50\}$. For each value of t we generated a weight matrix and either generated a synthetic dataset or sampled a real dataset as described in the above paragraphs, then tested our three algorithms.

For the SVD, we just took the best rank k approximation to A as given by the top k singular vectors. We used the built-in `svd` function in numpy's linear algebra package.

For Alternating Minimization without Sketching, we initialized the low rank matrix factors U and V to be random subsets of the rows and columns of A respectively, then performed $n = 25$ steps of alternating minimization.

For Alternating Minimization with Sketching, we initialized U and V the same way, but performed $n = 25$ best response updates in the sketched space, as in Theorem 3. The sketch S was chosen to be a CountSketch matrix with t. Based on Theorem 5, we calculated a rank $t < k$ approximation of A whenever we used a sketch of size t. We plotted the objective value of the low rank approximation for the connectus, NIPS, and synthetic datasets (the other datasets as well as a different family of weight matrices are discussed in the supplementary material) for each value of t and each algorithm in Figure 1. The experiment with the landmark dataset in Figure 1 used a regularization parameter value of $\lambda = 0.556$, while the experiments with the NIPS and synthetic datasets used a value of $\lambda = 1$. Objective values were given in 1000's in the Frobenius norm.

Both forms of alternating minimization greatly outperform the low rank approximation given by the SVD. Alternating minimization with sketching comes within a factor of 1.5 approximation to alternating minimization without sketching and can sometimes slightly outperform alternating minimization without sketching[2], showing that performing CountSketch at each best response step does not result in a critically suboptimal objective value. The runtime of alternating minimization with sketching varies from being around 2 times as fast as alternating minimization without sketching (when the sketch size $t = 10$) to being around 1.4 times as fast (when the sketch size $t = 50$). Table 1 shows the runtimes for the non-synthetic experiments of Figure 1.

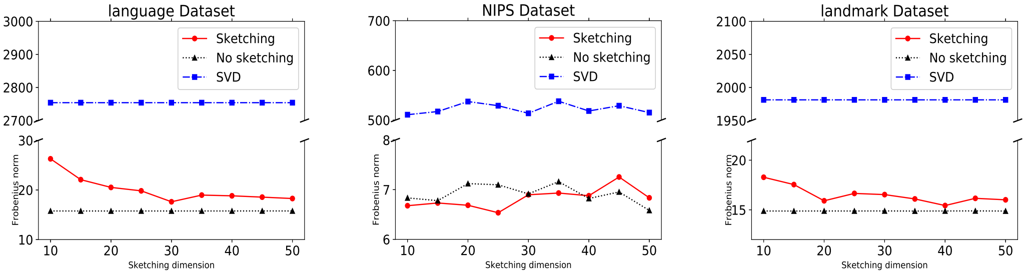

Figure 2: Regularized weighted low-rank approximations with $\lambda = 2.754$ for language, $\lambda = 1$ for NIPS, and $\lambda = 1.982$ for landmark.

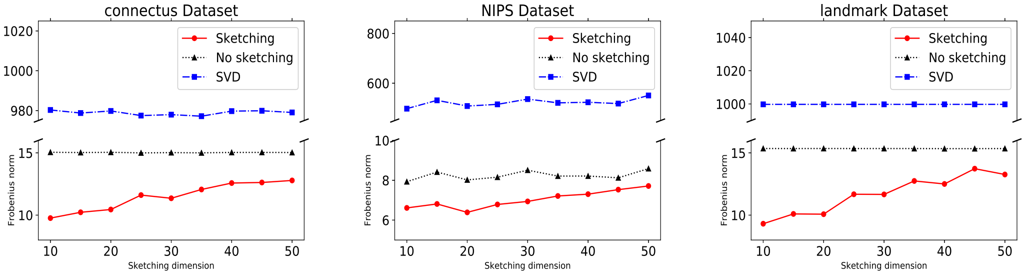

Figure 3: Regularized weighted low-rank approximations with binary weights and $\lambda = 1$.

Runtimes w/ sketching		
t	landmark	NIPS
10	54.31	49.1
15	53.58	50.33
20	57.65	51.8
25	65.53	56.43
30	68.68	57.34
35	72.22	62.66
40	79.94	63.48
45	81.22	67.73
50	72.77	73.11

Runtimes wo/ sketching		
t	landmark	NIPS
10	126.22	104.5
15	113.8	105.75
20	119.17	104.28
25	121.69	104.35
30	123.51	105.42
35	129.87	100.5
40	123.65	101.75
45	109.02	104.93
50	100.61	101.77

Table 1: Runtimes in seconds for alternating minimization with and without sketching.

Acknowledgements: Part of this work was done while D. Woodruff was visiting Google Mountain View as well as the Simons Institute for the Theory of Computing, and was supported in part by an Office of Naval Research (ONR) grant N00014-18-1-2562.

[2] See the supplementary material for additional discussion.

References

[ACW17] Haim Avron, Kenneth L. Clarkson, and David P. Woodruff. Sharper bounds for regularized data fitting. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques, APPROX/RANDOM 2017, August 16-18, 2017, Berkeley, CA, USA*, pages 27:1–27:22, 2017. 1.1, 4.2, A

[BPR96] Saugata Basu, Richard Pollack, and Marie-Françoise Roy. On the combinatorial and algebraic complexity of quantifier elimination. *Journal of the ACM (JACM)*, 43(6):1002–1045, 1996. 1

[CD08] Zizhong Chen and Jack J. Dongarra. Condition numbers of gaussian random matrices. *CoRR*, abs/0810.0800, 2008. A

[CKP$^+$17] Michael B Cohen, Jonathan Kelner, John Peebles, Richard Peng, Anup B Rao, Aaron Sidford, and Adrian Vladu. Almost-linear-time algorithms for markov chains and new spectral primitives for directed graphs. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 410–419. ACM, 2017. 7

[CNW16] Michael B. Cohen, Jelani Nelson, and David P. Woodruff. Optimal Approximate Matrix Product in Terms of Stable Rank. In Ioannis Chatzigiannakis, Michael Mitzenmacher, Yuval Rabani, and Davide Sangiorgi, editors, *43rd International Colloquium on Automata, Languages, and Programming (ICALP 2016)*, volume 55 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 11:1–11:14, Dagstuhl, Germany, 2016. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik. 2.1

[DH11] Timothy A. Davis and Yifan Hu. The university of florida sparse matrix collection. *ACM Trans. Math. Softw.*, 38(1):1:1–1:25, December 2011. 6, 6

[DN11] Saptarshi Das and Arnold Neumaier. Regularized low rank approximation of weighted data sets. *Preprint*, 2011. 1.1

[GG11] Nicolas Gillis and Francois Glineur. Low-rank matrix approximation with weights or missing data is np-hard. *SIAM Journal on Matrix Analysis and Applications*, 32(4):1149–1165, 2011. 1

[LA03] Wu-Sheng Lu and Andreas Antoniou. New method for weighted low-rank approximation of complex-valued matrices and its application for the design of 2-d digital filters. In *ISCAS (3)*, pages 694–697, 2003. 1

[LLR16] Yuanzhi Li, Yingyu Liang, and Andrej Risteski. Recovery guarantee of weighted low-rank approximation via alternating minimization. In *International Conference on Machine Learning*, pages 2358–2367, 2016. 1

[LPW97] W.-S Lu, S.-C Pei, and P.-H Wang. Weighted low-rank approximation of general complex matrices and its application in the design of 2-d digital filters. In *IEEE Transactions on Circuits and Systems*, volume 44, pages 650–655, 1997. 1

[PJST17] Valerio Perrone, Paul A. Jenkins, Dario Spanò, and Yee Whye Teh. Poisson random fields for dynamic feature models. *Journal of Machine Learning Research*, 18:127:1–127:45, 2017. 6, 6

[PW15] Mert Pilanci and Martin J Wainwright. Randomized sketches of convex programs with sharp guarantees. *IEEE Transactions on Information Theory*, 61(9):5096–5115, 2015. 1

[Ren92a] James Renegar. On the computational complexity and geometry of the first-order theory of the reals. part i: Introduction. preliminaries. the geometry of semi-algebraic sets. the decision problem for the existential theory of the reals. *Journal of symbolic computation*, 13(3):255–299, 1992. 1

[Ren92b] James Renegar. On the computational complexity and geometry of the first-order theory of the reals. part ii: The general decision problem. preliminaries for quantifier elimination. *Journal of Symbolic Computation*, 13(3):301–327, 1992. 1

[RSW16] Ilya P. Razenshteyn, Zhao Song, and David P. Woodruff. Weighted low rank approximations with provable guarantees. In *Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2016, Cambridge, MA, USA, June 18-21, 2016*, pages 250–263, 2016. 1, 1.1, 1.1, 2.1, A

[Shp90] D. Shpak. A weighted-least-squares matrix decomposition method with applications to the design of two-dimensional digital filters. In *IEEE Thirty Third Midwest Symposium on Circuits and Systems*, 1990. 1

[SJ03] Nathan Srebro and Tommi S. Jaakkola. Weighted low-rank approximations. In *Machine Learning, Proceedings of the Twentieth International Conference (ICML 2003), August 21-24, 2003, Washington, DC, USA*, pages 720–727, 2003. 1

[TYUC17] Joel A Tropp, Alp Yurtsever, Madeleine Udell, and Volkan Cevher. Practical sketching algorithms for low-rank matrix approximation. *SIAM Journal on Matrix Analysis and Applications*, 38(4):1454–1485, 2017. 1

[Woo14] David P. Woodruff. Sketching as a tool for numerical linear algebra. *Foundations and Trends in Theoretical Computer Science*, 10(1-2):1–157, 2014. 1, 2

A Proof of Theorem 2

Proof. Let $\widehat{S} = \begin{bmatrix} S & 0 \\ 0 & I_k \end{bmatrix} \in \mathbb{R}^{(\ell+k)\times(n+k)}$, $\widehat{M}^{(i)} = \begin{bmatrix} M^{(i)} \\ \sqrt{\lambda} I_k \end{bmatrix} \in \mathbb{R}^{(n+k)\times k}$, and $\widehat{b}^{(i)} = \begin{bmatrix} b^{(i)} \\ 0 \end{bmatrix} \in \mathbb{R}^{n+k}$. Observe that $\|M^{(i)}x - b^{(i)}\|^2 + \lambda\|x\|^2 = \|\widehat{M}^{(i)}x - \widehat{b}^{(i)}\|^2$ and $\|S(M^{(i)}y - b^{(i)})\|^2 + \lambda\|y\|^2 = \|\widehat{S}(\widehat{M}^{(i)}y - \widehat{b}^{(i)})\|^2$.

It suffices to prove that, for $1 \le i \le d$,

$$\mathop{\mathbf{E}}_{S}\left[\|\widehat{M}^{(i)}y^{(i)} - \widehat{b}^{(i)}\|^2 - \|\widehat{M}^{(i)}x^{(i)} - \widehat{b}^{(i)}\|^2\right] \le O(\varepsilon)\cdot\|\widehat{M}^{(i)}x^{(i)} - \widehat{b}^{(i)}\|^2$$

because we can sum over all i and apply Markov's inequality to complete the argument.

Fix some i and set $M = M^{(i)}, b = b^{(i)}, \widehat{M} = \widehat{M}^{(i)}, \widehat{b} = \widehat{b}^{(i)}$, $x = x^{(i)}$, $y = y^{(i)}$. Let $\widehat{U}_b$ be an orthogonal matrix whose columns form an orthonormal basis for the columns of $[\widehat{M}\,\widehat{b}]$. Now define $\Gamma := \|\widehat{U}_b^T\widehat{S}^T\widehat{S}\widehat{U}_b - I_{k+1}\|$. We let U_1 form its first n rows and U_2 form the rest.

Since $s < k$, we can have an unbounded condition number if we just look at $c_{\widehat{S}}(\widehat{M})$ so we need to have a more subtle analysis than in [RSW16]. Instead of simply conditioning on Γ, let us define $M = M_h + M_t$, where M_h is the component of M corresponding to the span of singular vectors with values that are $\sigma_i^2 \ge \lambda$. Then, M_t is the orthogonal component to M_h and is a subspace of the span of singular vectors corresponding to values that are $\sigma_i^2 < \lambda$. Since $2 \ge 1 + \frac{\lambda}{\sigma_i^2}$, for σ_i corresponding to M_h, then the rank of M_h is bounded by $r_h = O(\mathtt{sd}_\lambda(M))$. Since S has at least $\Omega(\mathtt{sd}_\lambda(M)/\varepsilon)$ rows, with probability 1, the condition number $c_h = c_S([M_h, b])$ will be finite with probability 1 (note that b can be assumed to be orthogonal to the image of M).

Let $\alpha = c_h(1+\Gamma)$ be a product of condition numbers. If α is close to 1, then we are in a good regime so we will condition on α.It follows that

$$\begin{aligned}\mathop{\mathbf{E}}_{S}\left[\|\widehat{M}y - \widehat{b}\|^2 - \|\widehat{M}x - \widehat{b}\|^2\right] &= \mathbf{Pr}\left[\alpha > 1.1\right]\cdot\mathop{\mathbf{E}}_{S}\left[\|\widehat{M}y - \widehat{b}\|^2 - \|\widehat{M}x - \widehat{b}\|^2 \middle| \alpha > 1.1\right] \\ &\quad + \mathbf{Pr}\left[\alpha \le 1.1\right]\cdot\mathop{\mathbf{E}}_{S}\left[\|\widehat{M}y - \widehat{b}\|^2 - \|\widehat{M}x - \widehat{b}\|^2 \middle| \alpha \le 1.1\right]\end{aligned}$$

We will now bound the two terms in the sum and our final result follows by combining Claim A.1 and Claim A.4.

Claim A.1.

$$\mathbf{Pr}\left[\alpha > 1.1\right]\mathop{\mathbf{E}}_{S}\left[\|\widehat{M}y - \widehat{b}\|^2 - \|\widehat{M}x - \widehat{b}\|^2 \middle| \alpha > 1.1\right] \le O(\varepsilon)\cdot\|\widehat{M}x - \widehat{b}\|^2$$

Proof. To bound our expression, note that

$$\|\widehat{M}y - \widehat{b}\|^2 \le \frac{1}{\kappa_{\widehat{S}}([\widehat{M}\,\widehat{b}])}\|\widehat{S}(\widehat{M}y - \widehat{b})\|^2 \le \frac{1}{\kappa_{\widehat{S}}([\widehat{M}\,\widehat{b}])}\|\widehat{S}(\widehat{M}x - \widehat{b})\|^2 \le \frac{K_{\widehat{S}}([\widehat{M}\,\widehat{b}])}{\kappa_{\widehat{S}}([\widehat{M}\,\widehat{b}])}\|\widehat{M}x - \widehat{b}\|^2$$

By Claim A.2, $\|\widehat{M}y - \widehat{b}\|^2 \le 10\alpha^2\|\widehat{M}x - \widehat{b}\|^2$. By Claim A.3,

$$\mathop{\mathbf{E}}_{S}\left[\|\widehat{M}y - \widehat{b}\|^2 - \|\widehat{M}x - \widehat{b}\|^2 \middle| \alpha > 1.1\right] \le 10\|\widehat{M}x - \widehat{b}\|^2 \mathop{\mathbf{E}}_{S}\left[\alpha^2 \middle| \alpha > 1.1\right] \le O(\varepsilon)\|\widehat{M}x - \widehat{b}\|^2$$

□

Claim A.2. *$K_{\widehat{S}}([\widehat{M}\,\widehat{b}]) \le \alpha$ and $\frac{1}{\kappa_{\widehat{S}}([\widehat{M}\,\widehat{b}])} \le 10\alpha$*

Proof. First, we bound $K_{\widehat{S}}([\widehat{M}\,\widehat{b}])$. By definition of Γ, $\|S(Mx - b)\|^2 + \lambda\|x\|^2 \le (1+\Gamma)(\|Mx - b\|^2 + \lambda\|x\|^2)$ so $K_{\widehat{S}}([\widehat{M}\,\widehat{b}]) \le 1 + \Gamma \le \alpha$.

More importantly, we want to bound $\frac{1}{\kappa_{\widehat{S}}([\widehat{M}\,\widehat{b}])}$. First, we claim that $\|SM_tx\| \le \sqrt{\lambda(1+2\Gamma)}\|x\|$. We may assume x lies entirely in the column space of M_t and by definition of M_t, we know that

$\|Mx\|^2 = \|M_t x\|^2 \le \lambda\|x\|^2$. Now, by the definition of Γ, $\|SM_t x\|^2 = \|SMx\|^2$ which is at most $(1+\Gamma)\|Mx\|^2 + \Gamma\lambda\|x\|^2 \le (1+2\Gamma)\lambda\|x\|^2$.

We now consider two cases: one where $\|S(M_h x - b)\|$ is at least $2\sqrt{\lambda(1+2\Gamma)}\|x\|$ and one where $\|S(M_h x - b)\| < 2\sqrt{\lambda(1+2\Gamma)}\|x\|$.

For all x such that $\|S(M_h x - b)\| \ge 2\sqrt{\lambda(1+2\Gamma)}\|x\|$, we rewrite $\|S(Mx - b)\|^2 = \|S(M_h x - b) + SM_t x\|^2$. Then, by Cauchy-Schwarz,

$$\begin{aligned}
&\|S(Mx-b)\|^2 \ge \|S(M_h x - b)\|^2 - 2|\langle S(M_h x - b), SM_t x\rangle| \\
&\ge \|S(M_h x - b)\|^2 - 2\|S(M_h x - b)\|\|SM_t x\| = \|S(M_h x - b)\|(\|S(M_h x - b)\| - \|SM_t x\|) \\
&\ge 0.5\|S(M_h x - b)\|^2 \ge (0.5/c_h)\|M_h x - b\|^2,
\end{aligned}$$

where the fourth line follows since $\|S(M_h x - b)\| \ge 2\sqrt{\lambda(1+2\Gamma)}\|x\| \ge 2\|SM_t x\|$ and the fifth line follows from definition of c_h.

Finally, this implies

$$\begin{aligned}
&\|S(Mx-b)\|^2 + \lambda\|x\|^2 \ge (0.5/c_h)(\|M_h x - b\|^2 + 2\lambda\|x\|^2) \\
&\ge (0.5/c_h)(\|M_h x - b\|^2 + \|M_t x\|^2 + \lambda\|x\|^2) \ge (1/2c_h)(\|Mx - b\|^2 + \lambda\|x\|^2)
\end{aligned}$$

where the first line follows since $c_h > 1$, the second line follows from $\|M_t x\|^2 < \lambda\|x\|^2$ and the last line from orthogonality.

Now consider all x such that $\|S(M_h x - b)\|$ is less than $2\sqrt{\lambda K_S(M)}\|x\|$. This means $\|M_h x - b\|$ is less than $2\sqrt{c_h \lambda K_S(M)}\|x\|$. Then,

$$\begin{aligned}
&\|Mx - b\|^2 + \lambda\|x\|^2 \le 4c_h(1+2\Gamma)\lambda\|x\|^2 + \lambda\|x\|^2 \\
&\le \frac{1}{4c_h(1+2\Gamma)+1} \cdot \lambda\|x\|^2 \le \frac{1}{5c_h(1+2\Gamma)} \cdot (\|S(Mx-b)\|^2 + \lambda\|x\|^2)
\end{aligned}$$

Together, we conclude that $\frac{1}{\kappa_{\widehat{S}}([\widehat{M}\ \widehat{b}])} \le \max(5c_h(1+2\Gamma), 2c_h) \le 10\alpha$, where $\alpha = c_h(1+\Gamma)$.

□

Claim A.3.

$$\mathbf{Pr}\left[\alpha > 1.1\right] \mathop{\mathbf{E}}_{S}\left[\alpha^2 \middle| \alpha > 1.1\right] = O(\varepsilon)$$

Proof. Note that for $t > 1$, we have

$$\begin{aligned}
\mathbf{Pr}\left(\alpha > t\right) &\le \mathbf{Pr}\left(1+\Gamma > \sqrt{t}\right) + \mathbf{Pr}\left(1+\Gamma \le \sqrt{t} \text{ and } c_h > \sqrt{t}\right) \\
&\le \mathbf{Pr}\left(1+\Gamma > \sqrt{t}\right) + \mathbf{Pr}\left(c_h > \sqrt{t}\right)
\end{aligned}$$

By Lemma 12 of [ACW17], note that $\|U_1\|_F^2$ is at most $\mathtt{sd}_\lambda(M) + 1$ and $\|U_1\| < 1$. Now, we can express $\Gamma = \|\widehat{U}_b^T \widehat{S}^T \widehat{S} \widehat{U}_b - I_{k+1}\|$ which is equal to $\|U_1^T S^T S U_1 - U_1^T U_1\|$. By Lemma 2.1 with $A = B = U_1$ and $\gamma = \frac{\sqrt{t}}{\|U_1\|^2}$, then for any $t > 1.1$, we have

$$\mathbf{Pr}[1+\Gamma > \sqrt{t}] < \varepsilon t^{-\Omega(1)}$$

since ℓ is larger than $\Omega(\frac{1}{\gamma^2}(\|U_1\|_F^2/\|U_1\|^2 + \log(t/\varepsilon))) = \Omega(\mathtt{sd}_\lambda(M) + \log(1/\varepsilon))$.

Then, by [CD08], since M_h only has less than $O(\mathtt{sd}_\lambda(M))$ columns, then since $\ell > \mathtt{sd}_\lambda(M)/\varepsilon$, we have for $t > 1.1$, $\mathbf{Pr}[c_S(M) > \sqrt{t}] = \Theta(t^{-1/\varepsilon}) < \varepsilon t^{-\Omega(1)}$. Thus, we conclude that

$$\mathbf{Pr}\left[\alpha > 1.1\right] \mathop{\mathbf{E}}_{S}\left[\alpha^2 \middle| \alpha > 1.1\right] \le O(1)\,\mathbf{Pr}[\alpha > 1.1] + \int_{1.1}^{\infty} t\,\mathbf{Pr}[\alpha > t]\,dt = O(\varepsilon)$$

□

Claim A.4. $\mathbf{E}_S\left[\|\widehat{M}y - \widehat{b}\|^2 - \|\widehat{M}x - \widehat{b}\|^2 \middle| \alpha \le 1.1\right] \le O(\varepsilon) \cdot \|\widehat{M}x - \widehat{b}\|^2$

Proof. The normal equations for x tell us that $\widehat{M}^T(\widehat{M}x - \widehat{b})$ is 0. Thus, by the Pythagorean Theorem,

$$\|\widehat{M}y - \widehat{b}\|^2 - \|\widehat{M}x - \widehat{b}\|^2 = \|\widehat{M}(y-x)\|^2 = \|\tilde{y} - \tilde{x}\|^2$$

where $\widehat{U}\tilde{y} = \widehat{M}y$ and $\widehat{U}\tilde{x} = \widehat{M}x$.

We have $\|\tilde{y} - \tilde{x}\| \leq \|(\widehat{U}^T\widehat{S}^T\widehat{S}\widehat{U} - I_k)(\tilde{y} - \tilde{x})\| + \|\widehat{U}^T\widehat{S}^T\widehat{S}\widehat{U}(\tilde{y} - \tilde{x})\|$, so since we are conditioning on $\alpha \leq 1.1$, we know that $\Gamma \leq 0.1$, which implies that $\|\tilde{y} - \tilde{x}\| \leq O(1)\|\widehat{U}^T\widehat{S}^T\widehat{S}\widehat{U}(\tilde{y} - \tilde{x})\|$.

Since $\mathbf{Pr}[\alpha \leq 1.1] \geq 1 - O(\varepsilon)$, then

$$\mathop{\mathbf{E}}_{S}\left[\|\widehat{M}y - \widehat{b}\|^2 - \|\widehat{M}x - \widehat{b}\|^2 \middle| \alpha \leq 1.1\right] \leq O(1) \cdot \mathop{\mathbf{E}}_{S}\left[\|\widehat{U}^T\widehat{S}^T\widehat{S}\widehat{U}(\tilde{y} - \tilde{x})\|^2\right].$$

and the normal equations for $\tilde{y}$ tell us that $\widehat{U}^T\widehat{S}^T\widehat{S}(\widehat{U}\tilde{y} - \widehat{b})$ is 0. Thus,

$$\mathop{\mathbf{E}}_{S}\left[\|\widehat{U}^T\widehat{S}^T\widehat{S}\widehat{U}(\tilde{y} - \tilde{x})\|^2\right] = \mathop{\mathbf{E}}_{S}\left[\|\widehat{U}^T\widehat{S}^T\widehat{S}(\widehat{U}\tilde{x} - \widehat{b})\|^2\right].$$

Let t be a natural number. Note that S has $\Omega(\frac{1}{\varepsilon}(s + \log(1/\varepsilon)) = \Omega(\frac{1}{t\varepsilon}(s + \log((1/\varepsilon)^t))$ rows. Note that $\widehat{S}$ only sketches U_1 but leaves U_2 un-sketched. By Lemma 2.1 with $A = U_1$, $B = U_1\tilde{x} - b$, $\gamma = \sqrt{t\varepsilon}/\|U_1\|$ we have

$$\mathbf{Pr}\left[\|\widehat{U}^T\widehat{S}^T\widehat{S}(\widehat{U}\tilde{x} - \widehat{b})\| > \sqrt{t\varepsilon}\|\widehat{U}\tilde{x} - \widehat{b}\|\right] < O(\varepsilon^t) \tag{2}$$

Let E_t denote the event that $\|\widehat{U}^T\widehat{S}^T\widehat{S}(\widehat{U}\tilde{x} - \widehat{b})\|$ is between $\sqrt{(t-1)\varepsilon}\|\widehat{M}x - \widehat{b}\|$ and $\sqrt{t\varepsilon}\|\widehat{M}x - \widehat{b}$. By inequality (2) we have

$$\mathop{\mathbf{E}}_{S}\left[\|\widehat{U}^T\widehat{S}^T\widehat{S}(\widehat{U}\tilde{x} - \widehat{b})\|^2\right] \leq \|\widehat{M}x - \widehat{b}\|^2 \sum_{t=1}^{\infty} t\varepsilon \cdot \mathbf{Pr}\left[E_t\|\right]$$

$$\leq \varepsilon \cdot \|\widehat{M}x - \widehat{b}\|^2 \sum_{t=1}^{\infty} O(\varepsilon^t) \cdot t \leq O(\varepsilon) \cdot \|\widehat{M}x - \widehat{b}\|^2$$

and we are done. □

□

B Proof of Theorem 4

Proof. Observe that our algorithm in Theorem 3,

$$\tilde{V} = \mathop{\mathrm{argmin}}_{V \in \mathbb{R}^{k \times n}} \sum_{j=1}^{d} \|Q^{(j)}V_{:,j} - S'' D_{W_{:,j}} A_{:,j}\|^2 + \lambda\|V_{:,j}\|^2$$

where $Q^{(j)} = S'' D_{W_{:,j}} U$ which equals $R^{(j)}U$ which is a $O((s + \log(1/\varepsilon))/\varepsilon)$ by k matrix and $R^{(j)} = S'' D_{W_{:,j}}$. Note that $R^{(j)}$ can be written as a linear combination of r matrices with rank at most $O((s + \log(1/\varepsilon))/\varepsilon)$. Therefore, by letting P be the projection matrix on the span of the total

$$O((s + \log(1/\varepsilon))r/\varepsilon)$$

right singular vectors of these r matrices, we see that $\tilde{V}$ equals

$$\mathop{\mathrm{argmin}}_{V \in \mathbb{R}^{k \times n}} \sum_{j=1}^{d} \|R^{(j)}PUV_{:,j} - S'' D_{W_{:,j}} A_{:,j}\|^2 + \lambda\|V_{:,j}\|^2$$

Since this holds for any U, we see that

$$\begin{aligned} &\|W \circ (PU^*\tilde{V} - A)\|_F^2 + \lambda\|PU^*\|_F^2 + \lambda\|\tilde{V}\|_F^2 \\ &\leq \|W \circ (U^*\tilde{V} - A)\|_F^2 + \lambda\|U^*\|_F^2 + \lambda\|\tilde{V}\|_F^2 \end{aligned}$$

Since $PU^*\tilde{V}$ has rank at most the rank of P, we conclude by noting that by the guarantees of Theorem 3,

$$\|W \circ (U^*\tilde{V} - A)\|_F^2 + \lambda\|U^*\|_F^2 + \lambda\|\tilde{V}\|_F^2 \leq (1+\varepsilon)\mathrm{OPT}(k)$$

□

C Proof of Theorem 6

Proof. Since we know $W = YZ$, where $Y \in R^{n\times r'}$ and $Z \in \mathbb{R}^{r'\times n}$ are non-negative, then we claim that we have a rounding procedure to create W' with only $(\log n/\varepsilon)^{r'}$ distinct columns and rows that produces an ε-close solution. The procedure is as expected: round all values in Y, Z to the nearest power of $(1+\varepsilon)$ and call $W' = Y'Z'$ our new matrix. Note that there are at most $(\log n/\varepsilon)^{r'}$ distinct rows, since each row in A takes on only $(\log n/\varepsilon)^{r'}$ possible values. Symmetrically, the number of columns is bounded by the same value. Now, we claim that:

Claim C.1. $(1-\varepsilon)^2 W \leq W' \leq (1+\varepsilon)^2 W$

Proof. It suffices to show that the intermediary matrix $\widehat{W} = Y'Z$ satisfies this bound. Consider each row of $\widehat{W}$, so WLOG, let $\widehat{W}_1$ be the first row of $\widehat{W}$ which can be expressed as $\widehat{W}_1 = \sum_{i=1}^{r'} (Y')_{1i} Z_i$, where Z_i is the i-th row of Z. Note that $W_1 = \sum_{i=1}^{r'} Y_{1i} Z_i$. Finally, since $(Y')_{1i} \in (1-\varepsilon, 1+\varepsilon)Y_{1i}$ by our rounding procedure and all values in Y_{1i}, Z_i are non-negative, we deduce that $(1-\varepsilon)\widehat{W}_1 \leq W_1 \leq (1+\varepsilon)\widehat{W}_1$. □

Since we only have $(\log n/\varepsilon)^{r'}$ distinct columns and rows, when we call the polynomial solver, the degree of our system is at most $(\log n/\varepsilon)^{r'}$ and our bound follows from polynomial system solver guarantees. □

D Proof of Theorem 8

Proof. Fix some i. By Theorem 7, we may express a k-order approximation of $\tilde{U}_{i,;}$ as

$$\tilde{U}^k_{i,;} = A_{i,;} D_{W_{i,;}} S' (P^{(i)})^T p_k (P^{(i)} (P^{(i)})^T + \lambda I_k)$$

where p_k is a degree $O(k)$ polynomial that approximate the inverse. Furthermore, we claim $\lambda I_k \preceq P^{(i)}(P^{(i)})^T + \lambda I_k \preceq \log(n)(1+\sigma/\lambda) I_k$, where $P^{(i)} = V^* D_{w_{i,;}} S$, with high probability.

By the same arguments as in the proof of Theorem 2 in Claim A.3, let $M^T = V^* D_{w_{i,;}}$ and $\widehat{M}$ defined analogously, along with $\widehat{U}, \widehat{S}$. Now define $\Gamma := \|\widehat{U}^T \widehat{S}^T \widehat{S} \widehat{U} - I_k\|$. Again, let U_1 form its first n rows and U_2 form the rest. Note that $\|U_1\|_F^2 \leq \mathtt{sd}_\lambda(M)$ and $\|U_1\| < 1$. By Lemma 2.1 with $A = B = U_1$ and $\gamma = \log(n)/\|U_1\|^2$, then we have

$$\mathbf{Pr}[1 + \Gamma > \log(n)] < n^{-\Omega(1)}$$

since $\ell > \Omega(\frac{1}{\gamma^2}(\|U_1\|_F^2/\|U_1\|^2 + \log(n))) = \Omega(\mathtt{sd}_\lambda(M)) = \Omega(s)$.

This implies that with high probability

$$\|SMx\|^2 + \lambda\|x\|^2 \leq \log(n)(\|Mx\|^2 + \lambda\|x\|^2).$$

Specifically, we have $\sigma_1(P^{(i)})^2 \leq \log(n)\sigma_1(V^* D_{w_{i,;}})^2 + \lambda \log(n) \leq \log(n)\sigma^2 + \lambda\log(n)$.

Therefore, by the guarantees of Theorem 7, we have

$$\tilde{U}^k_{i,;} P^{(i)} - A_{i,;} D_{W_{i,;}} S'\|^2 + \lambda \|\tilde{U}^k_{i,;}\|^2 \leq \|\tilde{U}_{i,;} P^{(i)} - A_{i,;} D_{W_{i,;}} S'\|^2 + \lambda\|\tilde{U}_{i,;}\|^2 + \tau/n$$

holds if $k > \Omega((\sigma^2/\lambda)\log(\Delta(\sigma^2+\lambda)/\lambda n/\tau))$.

Summing up for all i and applying a union bound over failure probabilities, we see that with constant probability, we have

$$\sum_{i=1}^{n} \|\tilde{U}^k_{i,;} P^{(i)} - A_{i,;} D_{W_{i,;}} S'\|^2 + \lambda\|\tilde{U}^k_{i,;}\|^2 \leq (1+\varepsilon)\mathrm{OPT} + \tau.$$

Finally, since the degree of the polynomial system using $\tilde{U}^k$ is simply $O(k)$, our theorem follows. □

E Proof of Theorem 9

Proof. For all i, we want to show there exists some matrix B such that $\eta(P^{(i)}(P^{(i)})^T + \lambda I_k) \preceq B \preceq P^{(i)}(P^{(i)})^T + \lambda I_k$ with constant probability. Then, we may simply guess B^{-1} with only an additional $O((s + \log(1/\varepsilon))^2)$ variables and apply Theorem 7, express a k-order approximation of $\tilde{U}_{i,;}$ as

$$\tilde{U}_{i,;}^{k} = A_{i,;}D_{W_{i,;}}S'(P^{(i)})^T p_k(P^{(i)}(P^{(i)})^T + \lambda I_k, B^{-1})$$

where p_k is a degree $O(k)$ polynomial that approximate the inverse and apply the same analysis as Theorem 8 to see that $k = O(\eta^{-1}\log(c_B/\tau))$ suffices.

In fact, we will explicitly construct B. Let $D \in \mathbb{R}^{n\times n}$ be the diagonal matrix with diagonal entries l_W. Let $P = V^*DS = RS$, where $R = V^*D_l$. Also, we define $P^{(i)} = V^*D_{w_{i,;}}S = R^{(i)}S$. Then, we see that by our bounds on W,

$$\frac{l_W}{u_W}R^{(i)}(R^{(i)})^T \preceq RR^T \preceq R^{(i)}(R^{(i)})^T$$

By using similar arguments for condition number bounds as in Claim A.3 in Theorem 2, we see that

$$\frac{l_W}{u_W\log(n)}R^{(i)}SS^T(R^{(i)})^T \preceq RSS^TR^T \preceq \log(n)R^{(i)}SS^T(R^{(i)})^T$$

with high probability. So, we set $B = (1/\log(n))RSS^TR$ and that implies that $\eta = 1/\log(n)^2(u_W/l_W)$. Then, using Theorem 7, we conclude with the same analysis as in Theorem 8. □

F A Note on the Experiments

It may surprise the reader to see that the objective values when using sketching slightly outperform the objective values without using sketching and that the objective value improves as the sketching dimension decreases. In theory, this should not happen because in low rank approximation problems, it never hurts the objective value to increase the number of columns of U and number of rows of V since one can simply add 0's.

This phenomenon arises due to the use of the alternating minimization heuristic. Although an ideal low rank approximation algorithm would recognize that if U and V have more columns / rows, then one only needs to add 0's, we found in our experiments that alternating minimization tended to add mass to those extra columns / rows. This extra mass resulted in a higher contribution from the regularization terms. Thus, by sketching onto fewer dimensions, the alternating minimization heuristic was improved because it couldn't add mass in the form of extraneous columns for U or rows for V.