Collaborative Filtering with Graph Information: Consistency and Scalable Methods

Nikhil Rao **Hsiang-Fu Yu** **Pradeep Ravikumar** **Inderjit S. Dhillon**

{nikhilr, rofuyu, paradeepr, inderjit}@cs.utexas.edu
Department of Computer Science
University of Texas at Austin

Abstract

Low rank matrix completion plays a fundamental role in collaborative filtering applications, the key idea being that the variables lie in a smaller subspace than the ambient space. Often, additional information about the variables is known, and it is reasonable to assume that incorporating this information will lead to better predictions. We tackle the problem of matrix completion when pairwise relationships among variables are known, via a graph. We formulate and derive a highly efficient, conjugate gradient based alternating minimization scheme that solves optimizations with over 55 million observations up to 2 orders of magnitude faster than state-of-the-art (stochastic) gradient-descent based methods. On the theoretical front, we show that such methods generalize weighted nuclear norm formulations, and derive statistical consistency guarantees. We validate our results on both real and synthetic datasets.

1 Introduction

Low rank matrix completion approaches are among the most widely used collaborative filtering methods, where a partially observed matrix is available to the practitioner, who needs to impute the missing entries. Specifically, suppose there exists a ratings matrix $Y \in \mathbb{R}^{m \times n}$, and we only observe a subset of the entries $Y_{ij}, \forall (i,j) \in \Omega, |\Omega| = N \ll mn$. The goal is to estimate $Y_{i,j}, \forall (i,j) \notin \Omega$. To this end, one typically looks to solve one of the following (equivalent) programs:

$$\hat{Z} = \arg\min_{Z} \frac{1}{2} \|\mathcal{P}_\Omega(Y - Z)\|_F^2 + \lambda_z \|Z\|_* \tag{1}$$

$$\hat{W}, \hat{H} = \arg\min_{W,H} \frac{1}{2} \|\mathcal{P}_\Omega(Y - WH^T)\|_F^2 + \frac{\lambda_w}{2} \|W\|_F^2 + \frac{\lambda_h}{2} \|H\|_F^2 \tag{2}$$

where the nuclear norm $\|Z\|_*$, given by the sum of singular values, is a tight convex relaxation of the non convex rank penalty, and is equivalent to the regularizer in (2). $\mathcal{P}_\Omega(\cdot)$ is the projection operator that only retains those entries of the matrix that lie in the set Ω.

In many cases however, one not only has the partially observed ratings matrix, but also has access to additional information about the relationships between the variables involved. For example, one might have access to a social network of users. Similarly, one might have access to attributes of items, movies, etc. The nature of the attributes can be fairly arbitrary, but it is reasonable to assume that "similar" users/items share "similar" attributes. A natural question to ask then, is if one can take advantage of this additional information to make better predictions. In this paper, we assume that the row and column variables lie on graphs. The graphs may naturally be part of the data (social networks, product co-purchasing graphs) or they can be constructed from available features. The idea then is to incorporate this additional structural information into the matrix completion setting.

We not only require the resulting optimization program to enforce additional constraints on Z, but we also require it to admit efficient optimization algorithms. We show in the sections that follow that this in fact is indeed the case. We also perform a theoretical analysis of our problem when the observed entries of Y are corrupted by additive white Gaussian noise. To summarize, the contributions of our paper are as follows:

- We provide a *scalable* algorithm for matrix completion graph with structural information. Our method relies on efficient Hessian-vector multiplication schemes, and is orders of magnitude faster than (stochastic) gradient descent based approaches.
- We make connections with other structured matrix factorization frameworks. Notably, we show that our method generalizes the weighted nuclear norm [20], and methods based on Gaussian generative models [26].
- We derive consistency guarantees for graph regularized matrix completion, and empirically show that our bound is smaller than that of traditional matrix completion, where graph information is ignored.
- We empirically validate our claims, and show that our method achieves comparable error rates to other methods, while being significantly more scalable.

Related Work and Key Differences

For convex methods for matrix factorization, Haeffele et al. [9] provided a framework to use regularizers with norms other than the Euclidean norm in (2). Abernethy et al. [1] considered a kernel based embedding of the data, and showed that the resulting problem can be expressed as a norm minimization scheme. Srebro and Salakhutdinov [20] introduced a weighted nuclear norm, and showed that the method enjoys superior performance as compared to standard matrix completion under a non-uniform sampling scheme. We show that the graph based framework considered in this paper is in fact a generalization of the weighted nuclear norm problem, with non-diagonal weight matrices.

In the context of matrix factorization with graph structural information, [5] considered a graph regularized nonnegative matrix factorization framework and proposed a gradient descent based method to solve the problem. In the context of recommendation systems in social networks, Ma et al. [14] modeled the weight of a graph edge[1] explicitly in a re-weighted regularization framework. Li and Yeung [13] considered a similar setting to ours, but a key point of difference between all the aforementioned methods and our paper is that we consider the partially observed ratings case. There are some works developing algorithms for the situation with partially observations [12, 25, 26]; however, none of them provides statistical guarantees. Weighted norm minimization has been considered before ([15, 20]) in the context of low rank matrix completion. The thrust of these methods has been to show that despite suboptimal conditions (correlated data, non-uniform sampling), the sample complexity does not change. None of these methods use graph information. We are interested in a complementary question: *Given variables conforming to graph information, can we obtain better guarantees under uniform sampling to those achieved by traditional methods?*

2 Graph-Structured Matrix Factorization

Assume that the "true" target matrix can be factorized as $Z^\star = W^\star (H^\star)^T$, and there exist a graph (V^w, E^w) whose adjacency matrix encodes the relationships between the m rows of $W^\star$ and a graph (V^h, E^h) for n rows of $H^\star$. In particular, two rows (or columns) connected by an edge in the graph are "close" to each other in the Euclidean distance. In the context of graph-based embedding, [3, 4] proposed a smoothing term of the form

$$\frac{1}{2}\sum_{i,j} E^w_{ij}(\boldsymbol{w}_i - \boldsymbol{w}_j)^2 = \mathrm{tr}(W^T \mathbf{Lap}(E^w) W) \tag{3}$$

where $\mathbf{Lap}(E^w) := D^w - E^w$ is the graph Laplacian for (V^w, E^w), where D^w is the diagonal matrix with $D^w_{ii} = \sum_{j\sim i} E^w_{ij}$. Adding (3) into the minimization problem (2) encourages solutions where $\boldsymbol{w}_i \approx \boldsymbol{w}_j$ when E^w_{ij} is large. A similar argument holds for $H^\star$ and the associated graph Laplacian $\mathbf{Lap}(E^h)$.

[1] The authors call this the "trust" between links in a social network

We would thus not only want the target matrix to be low rank, but also want the variables W, H to be faithful to the underlying graph structure. To this end, we consider the following problem:

$$\min_{W,H} \ \frac{1}{2}\|\mathcal{P}_\Omega\big(Y - WH^T\big)\|_F^2 + \frac{\lambda_L}{2}\{\mathrm{tr}(W^T\,\mathbf{Lap}(E^w)W) + \mathrm{tr}(H^T\,\mathbf{Lap}(E^h)H)\} + \frac{\lambda_w}{2}\|W\|_F^2 + \frac{\lambda_h}{2}\|H\|_F^2 \tag{4}$$

$$\equiv \min_{W,H} \ \frac{1}{2}\|\mathcal{P}_\Omega\big(Y - WH^T\big)\|_F^2 + \frac{1}{2}\big\{\mathrm{tr}(W^T L_w W) + \mathrm{tr}(H^T L_h H)\big\} \tag{5}$$

where $L_w := \lambda_L\,\mathbf{Lap}(E^w) + \lambda_w I_m$, and L_h is defined similarly. Note that we subsume the regularization parameters in the definition of L_w, L_h. Note that $\|W\|_F^2 = \mathrm{tr}(W^T I_m W)$.

The regularizer in (5) encourages solutions that are smooth with respect to the corresponding graphs. However, the Laplacian matrix can be replaced by other (positive, semi-definite) matrices that encourage structure by different means. Indeed, a very general class of Laplacian based regularizers was considered in [19], where one can replace L_w by a function:

$$\langle \boldsymbol{x}, \tau(\mathbf{Lap}(E))\boldsymbol{x}\rangle \quad \text{where} \quad \tau(\mathbf{Lap}(E)) \equiv \sum_{i=1}^{|V|} \tau(\lambda_i)\boldsymbol{q}_i\boldsymbol{q}_i^T,$$

where $\{(\lambda_i, \boldsymbol{q}_i)\}$ constitute the eigen-system of $\mathbf{Lap}(E)$ and $\tau(\lambda_i)$ is a scalar function of the eigenvalues. Our case corresponds to $\tau(\cdot)$ being the identity function. We briefly summarize other schemes that fit neatly into (5), apart from the graph regularizer we consider:

Covariance matrices for variables: [26] proposed a kernelized probabilistic matrix factorization (KPMF), which is a generative model to incorporate covariance information of the variables into matrix factorization. They assumed that each row of $W^\star$, $H^\star$ is generated according to a multivariate Gaussian, and solving the corresponding MAP estimation procedure yields exactly (5), with $L_w = C_w^{-1}$ and $L_h = C_h^{-1}$, where C_w, C_h are the associated covariance matrices.

Feature matrices for variables: Assume that there is a feature matrix $X \in \mathbb{R}^{m\times d}$ for objects associated rows. For such X, one can construct a graph (and hence a Laplacian) using various methods such as k-nearest neighbors, ϵ-nearest neighbors etc. Moreover, one can assume that there exists a kernel $k(\boldsymbol{x}_i, \boldsymbol{x}_j)$ that encodes pairwise relations, and we can use the Kernel Gram matrix as a Laplacian.

We can thus see that problem (5) is a very general scheme, and can incorporate information available in many different forms. In the sequel, we assume the matrices L_w, L_h are given. In the theoretical analysis in Section 5, for ease of exposition, we further assume that the minimum eigenvalues of L_w, L_h are unity. A general (nonzero) minimum eigenvalue will merely introduce multiplicative constants in our bounds.

3 GRALS: Graph Regularized Alternating Least Squares

In this section, we propose efficient algorithms for (5), which is convex with respect to W or H separately. This allows us to employ alternating minimization methods [24] to solve the problem. When Y is fully observed, Li and Yeung [13] propose an alternating minimization scheme using block steepest descent. We deal with the partially observed setting, and propose to apply conjugate gradient (CG), which is known to converge faster than steepest descent, to solve each subproblem. We propose a very efficient Hessian-vector multiplication routine that results in the algorithm being highly scalable, compared to the (stochastic) gradient descent approaches in [14, 26].

We assume that $Y \in \mathbb{R}^{m\times n}$, $W \in \mathbb{R}^{m\times k}$ and $H \in \mathbb{R}^{n\times k}$. When optimizing H with W fixed, we obtain the following sub-problem.

$$\min_H \ f(H) = \frac{1}{2}\|\mathcal{P}_\Omega\big(Y - WH^T\big)\|_F^2 + \frac{1}{2}\,\mathrm{tr}(H^T L_h H). \tag{6}$$

Optimizing W while H fixed is similar, and thus we only show the details for solving (6). Since L_h is nonsingular, (6) is strongly convex.[2] We first present our algorithm for the fully observed case, since it sets the groundwork for the partially observed setting.

[2] In fact, a nonsingular L_h can be handled using proximal updates, and our algorithm will still apply

Algorithm 1 Hv-Multiplication for $g(\boldsymbol{s})$

- **Given:** Matrices L_h, W
- **Initialization:** $\boldsymbol{G} = W^T W$
- **Multiplication:** $\nabla^2 g(\boldsymbol{s}_0)\boldsymbol{s}$:
  1 **Input:** $S \in \mathbb{R}^{n\times k}$ s.t. $\boldsymbol{s} = \text{vec}(S)$
  2 $A \leftarrow S\boldsymbol{G} + L_h S$
  3 **Return**: $\text{vec}(A)$

Algorithm 2 Hv-Multiplication for $g_\Omega(\boldsymbol{s})$

- **Given:** Matrices L_h, W, Ω
- **Multiplication:** $\nabla^2 g(\boldsymbol{s}_0)\boldsymbol{s}$:
  1 **Input:** $S \in \mathbb{R}^{k\times n}$ s.t. $\boldsymbol{s} = \text{vec}(S)$
  2 Compute $K = [\boldsymbol{k}_1, \ldots, \boldsymbol{k}_n]$ s.t. $\boldsymbol{k}_j \leftarrow \sum_{i\in\Omega_j} (\boldsymbol{w}_i^T \boldsymbol{s}_j)\boldsymbol{w}_i$
  3 $A \leftarrow K + SL_h$
  4 **Return**: $\text{vec}(A)$

3.1 Fully Observed Case

As in [5, 13] among others, there may be scenarios where Y is completely observed, and the goal is to find the row/column embeddings that conform to the corresponding graphs. In this case, the loss term in (6) is simply $\|Y - WH^T\|_F^2$. Thus, setting $\nabla f(H) = 0$ is equivalent to solving the following Sylvester equation for an $n \times k$ matrix H:

$$HW^T W + L_h H = Y^T W. \tag{7}$$

(7) admits a closed form solution. However the standard Bartels-Stewart algorithm for the Sylvester equation requires transforming both $W^T W$ and L_h into Schur form (diagonal in our case where $W^T W$ and L_h are symmetric) by the QR algorithm, which is time consuming for a large L_h. Thus, we consider applying conjugate gradient (CG) to minimize $f(H)$ directly. We define the following quadratic function:

$$g(\boldsymbol{s}) := \frac{1}{2}\boldsymbol{s}^T M\boldsymbol{s} - \text{vec}\big(Y^T W\big)^T \boldsymbol{s}, \;\; \boldsymbol{s} \in \mathbb{R}^{nk}, \;\; M = I_k \otimes L_h + (W^T W) \otimes I_n$$

It is not hard to show that $f(H) = g(\text{vec}(H))$ and so we apply CG to minimize $g(\boldsymbol{s})$.

The most crucial step in CG is the Hessian-vector multiplication. Using the identity $(B^T \otimes A)\,\text{vec}(X) = \text{vec}(AXB)$, it follows that

$$(I_k \otimes L_h)\boldsymbol{s} = \text{vec}(L_h S), \quad \text{and} \quad \big((W^T W) \otimes I_n\big)\boldsymbol{s} = \text{vec}\big(SW^T W\big),$$

where $\text{vec}(S) = \boldsymbol{s}$. Thus the Hessian-vector multiplication can be implemented by a series of matrix multiplications as follows.

$$M\boldsymbol{s} = \text{vec}\big(L_h S + S(W^T W)\big),$$

where $W^T W$ can be pre-computed and stored in $O(k^2)$ space. The details are presented in Algorithm 1. The time complexity for a single CG iteration is $O(nnz(L_h)k + nk^2)$, where $nnz(\cdot)$ is the number of non zeros. Since in most practical applications k is generally small, the complexity is essentially $O(nnz(L_h)k)$ as long as $nk \leq nnz(L_h)$.

3.2 Partially Observed Case

In this case, the loss term of (6) becomes $\sum_{(i,j)\in\Omega}(Y_{ij} - \boldsymbol{w}_i^T \boldsymbol{h}_j)^2$, where $\boldsymbol{w}_i^T$ is the i-th row of W and $\boldsymbol{h}_j$ is the j-th column of H^T. Similar to the fully observed case, we can define:

$$g_\Omega(\boldsymbol{s}) := \frac{1}{2}\boldsymbol{s}^T M_\Omega \boldsymbol{s} - \text{vec}\big(W^T Y\big)^T \boldsymbol{s},$$

where $M_\Omega = \bar{B} + L_h \otimes I_k$, $\bar{B} \in \mathbb{R}^{nk\times nk}$ is a block diagonal matrix with n diagonal blocks $B_j \in \mathbb{R}^{k\times k}$. $B_j = \sum_{i\in\Omega_j} \boldsymbol{w}_i \boldsymbol{w}_i^T$, where $\Omega_j = \{i : (i,j) \in \Omega\}$. Again, we can see $f(H) = g_\Omega(\text{vec}\big(H^T\big))$. Note that the transpose H^T is used here instead of H, which is used in the fully observed case.

For a given $\boldsymbol{s}$, let $S = [\boldsymbol{s}_1, \ldots \boldsymbol{s}_j, \ldots \boldsymbol{s}_n]$ be a matrix such that $\text{vec}(S) = \boldsymbol{s}$ and $K = [\boldsymbol{k}_1, \ldots, \boldsymbol{k}_j, \ldots, \boldsymbol{k}_n]$ with $\boldsymbol{k}_j = B_j \boldsymbol{s}_j$. Then $\bar{B}\boldsymbol{s} = \text{vec}(K)$. Note that since n can be very large in practice, it may not be feasible to compute and store all B_j in the beginning. Alternatively, $B_j \boldsymbol{s}_j$ can be computed in $O(|\Omega_j|k)$ time as follows.

$$B_j \boldsymbol{s}_j = \sum_{i\in\Omega_j} (\boldsymbol{w}_i^T \boldsymbol{s}_j)\boldsymbol{w}_i.$$

Thus $\bar{B}\boldsymbol{s}$ can be computed in $O(|\Omega|k)$ time, and the Hessian-vector multiplication $M_\Omega \boldsymbol{s}$ can be done in $O(|\Omega|k + nnz(L_h)k)$ time. See Algorithm 2 for a detailed procedure. As a result, each CG iteration for minimizing $g_\Omega(\boldsymbol{s})$ is also very cheap.

Remark on Convergence. In [2], it is shown that any local minimizer of (5) is a global minimizer of (5) if k is larger than the true rank of the underlying matrix.[3] From [24], the alternating minimization procedure is guaranteed to globally converge to a block coordinate-wise minimum[4] of (5). The converged point might not be a local minimizer, but it still yields good performance in practice. Most importantly, since the updates are cheap to perform, our algorithm scales well to large datasets.

4 Convex Connection via Generalized Weighted Nuclear Norm

We now show that the regularizer in (5) can be cast as a generalized version of the weighted nuclear norm. The weights in our case will correspond to the scaling factors introduced on the matrices W, H due to the eigenvalues of the shifted graph Laplacians L_w, L_h respectively.

4.1 A weighted atomic norm:

From [7], we know that the nuclear norm is the gauge function induced by the atomic set: $\mathscr{A}_* = \{\boldsymbol{w}_i \boldsymbol{h}_i^T \ : \ \|\boldsymbol{w}_i\| = \|\boldsymbol{h}_i\| = 1\}$. Note that all rank one matrices in $\mathscr{A}_*$ have unit Frobenius norm. Now, assume $P = [\boldsymbol{p}_1, \ldots, \boldsymbol{p}_m] \in \mathbb{R}^{m\times m}$ is a basis of $\mathbb{R}^m$ and $S_p^{-1/2}$ is a diagonal matrix with $(S_p^{-1/2})_{ii} \geq 0$ encoding the "preference" over the space spanned by $\boldsymbol{p}_i$. The more the preference, the larger the value. Similarly, consider the basis Q and the preference $S_q^{-1/2}$ for $\mathbb{R}^n$. Let $A = PS_p^{-1/2}$ and $B = QS_q^{-1/2}$, and consider the following "preferential" atomic set:

$$\mathscr{A} := \{\mathbb{a}_i = \boldsymbol{w}_i \boldsymbol{h}_i^T : \boldsymbol{w}_i = A\boldsymbol{u}_i, \boldsymbol{h}_i = B\boldsymbol{v}_i, \|\boldsymbol{u}_i\| = \|\boldsymbol{v}_i\| = 1\}. \tag{8}$$

Clearly, each atom $\mathbb{a}$ in $\mathscr{A}$ has non-unit Frobenius norm. This atomic set allows for biasing of the solutions towards certain atoms. We then define a corresponding atomic norm:

$$\|Z\|_{\mathscr{A}} = \inf \sum_{\mathbb{a}_i \in \mathscr{A}} |c_i| \quad \text{s.t. } Z = \sum_{\mathbb{a}_i \in \mathscr{A}} c_i \mathbb{a}_i. \tag{9}$$

It is not hard to verify that $\|Z\|_{\mathscr{A}}$ is a norm and $\{Z : \|Z\|_{\mathscr{A}} \leq \tau\}$ is closed and convex.

4.2 Equivalence to Graph Regularization

The graph regularization (5) can be shown to be a special case of the atomic norm (9), as a consequence of the following result:

Theorem 1. *For any $A = PS_p^{-1/2}$, $B = QS_q^{-1/2}$, and corresponding weighted atomic set $\mathscr{A}$,*

$$\|Z\|_{\mathscr{A}} = \inf_{W,H} \quad \frac{1}{2}\{\|A^{-1}W\|_F^2 + \|B^{-1}H\|_F^2\} \quad \textit{s.t.} \quad Z = WH^T.$$

We prove this result in Appendix A. Theorem 1 immediately leads us to the following equivalence result:

Corollary 1. *Let $L_w = U_w S_w U_w^T$ and $L_h = U_h S_h U_h^T$ be the eigen decomposition for L_w and L_h. We have*

$$\mathrm{Tr}\big(W^T L_w W\big) = \|A^{-1}W\|_F^2, \qquad \textit{and} \quad \mathrm{Tr}\big(H^T L_h H\big) = \|B^{-1}H\|_F^2,$$

where $A = U_w S_w^{-1/2}$ and $B = U_h S_h^{-1/2}$. As a result, $\|M\|_{\mathscr{A}}$ with the preference pair $(U_w, S_w^{-1/2})$ for the column space and the preference pair $(U_h, S_h^{-1/2})$ for row space is a weighted atomic norm equivalent for the graph regularization using L_w and L_h.

The results above allow us to obtain the dual weighted atomic norm for a matrix Z

$$\|Z\|_{\mathscr{A}}^* = \|A^T Z B\ \| = \|S_w^{-\frac{1}{2}} U_w^T Z U_h S_h^{-\frac{1}{2}}\| \tag{10}$$

[3] The authors actually show this for a more general class of regularizers.

[4] Nash equilibrium is used in [24].

which is a weighted spectral norm. An elementary proof of this result can be found in Appendix B. Note that we can then write

$$\|Z\|_{\mathscr{A}} = \|A^{-1}ZB^{-T}\|_* = \|S_w^{\frac{1}{2}}U_w^{-1}ZU_h^{-T}S_h^{\frac{1}{2}}\|_* \tag{11}$$

In [20], the authors consider a norm similar to (11), but with A, B being diagonal matrices. In the spirit of their nomenclature, we refer to the norm in (11) as the *generalized* weighted nuclear norm.

5 Statistical Consistency in the Presence of Noisy Measurements

In this section, we derive theoretical guarantees for the graph regularized low rank matrix estimators. We first introduce some additional notation. We assume that there is a $m \times n$ matrix $Z^\star$ of rank k with $\|Z^\star\|_F = 1$, and $N = |\Omega|$ entries of $Z^\star$ are uniformly sampled[5] and revealed to us (i.e., $Y = \mathcal{P}_\Omega(Z^\star)$). We further assume an one-to-one mapping between the set of observed indices Ω and $\{1, 2, \ldots, N\}$ so that the t^{th} measurement is given by

$$y_t = Y_{i(t),j(t)} = \langle \boldsymbol{e}_{i(t)}\boldsymbol{e}_{j(t)}^T, Z^\star\rangle + \frac{\sigma}{\sqrt{mn}}\eta_t \qquad \eta_t \sim \mathcal{N}(0,1). \tag{12}$$

where $\langle\cdot,\cdot\rangle$ denotes the matrix trace inner product, and $i(t), j(t)$ is a randomly selected coordinate pair from $[m]\times[n]$. Let A, B are corresponding matrices defined in Corollary 1 for the given L_w, L_h. W.L.O.G, we assume that the minimum singular value of both L_w and L_h is 1. We then define the following graph based complexity measures:

$$\alpha_g(Z) := \sqrt{mn}\frac{\|A^{-1}ZB^{-T}\|_\infty}{\|A^{-1}ZB^{-T}\|_F}, \qquad \beta_g(Z) := \frac{\|A^{-1}ZB^{-T}\|_*}{\|A^{-1}ZB^{-T}\|_F} \tag{13}$$

where $\|\cdot\|_\infty$ is the element-wise ℓ_∞ norm. Finally, we assume that the true matrix $Z^\star$ can be expressed as a linear combination of atoms from (8) (we define $\alpha^\star := \alpha_g(Z^\star)$):

$$Z^\star = AU^\star(V^\star)^TB^T, \;\; U^\star \in \mathbb{R}^{m\times k}, V^\star \in \mathbb{R}^{n\times k}, \tag{14}$$

Our goal in this section will be to characterize the solution to the following *convex* program, where the constraint set precludes selection of overly complex matrices in the sense of (13):

$$\hat{Z} = \arg\min_{Z\in\mathcal{C}} \frac{1}{N}\|\mathcal{P}_\Omega(Y-Z)\|_F^2 + \lambda\|Z\|_{\mathscr{A}} \text{ where } \mathcal{C} := \left\{Z : \alpha_g(Z)\beta_g(Z) \le \bar{c}_0\sqrt{\frac{N}{\log(m+n)}}\right\}, \tag{15}$$

where $\bar{c}_0$ is a constant depending on $\alpha^\star$.

A quick note on solving (15): since $\|\cdot\|_{\mathscr{A}}$ is a weighted nuclear norm, one can resort to proximal point methods [6], or greedy methods developed specifically for atomic norm constrained minimization [17, 21]. The latter are particularly attractive, since the greedy step reduces to computing the maximum singular vectors which can be efficiently computed using power methods. However, such methods will first involve computing the eigen decompositions of the graph Laplacians, and then storing the large, dense matrices A, B. We refrain from resorting to such methods in Section 6, and instead use the efficient framework derived in Section 3. We now state our main theoretical result:

Theorem 2. *Suppose we observe N entries of the form* (12) *from a matrix $Z^\star \in \mathbb{R}^{m\times n}$, with $\alpha^\star := \alpha_g(Z^\star)$ and which can be represented using at most k atoms from* (8)*. Let $\hat{Z}$ be the minimizer of the convex problem* (15) *with $\lambda \ge C_1\sqrt{\frac{(m+n)\log(m+n)}{N}}$. Then, with high probability, we have*

$$\|\hat{Z} - Z^\star\|_F^2 \le C\alpha^{\star 2}\max\{1,\sigma^2\}\frac{k(m+n)\log(m+n)}{N} + O\left(\frac{\alpha^{\star 2}}{N}\right)$$

where C, C_1 are positive constants.

See Appendix C for the detailed proof. A proof sketch is as follows:

[5] Our results can be generalized to non uniform sampling schemes as well.

Proof Sketch: There are three major portions of the proof:

- Using the fact that $Z^\star$ has unit Frobenius norm and can be expressed as a combination of at most k atoms, we can show $\|Z^\star\|_{\mathscr{A}} \leq \sqrt{k}$ (Appendix C.1)
- Using (10), we can derive a bound for the dual norm of the gradient of the loss $\mathcal{L}(Z)$, given by $\|\nabla\mathcal{L}(Z)\|^*_{\mathscr{A}} = \|S_w^{-\frac{1}{2}} U_w^T \nabla\mathcal{L}(Z) U_h S_h^{-\frac{1}{2}}\|$. (Appendix C.2)
- Finally, using (13), we define a notion of restricted strong convexity (RSC) that the "error" matrices $Z^\star - \hat{Z}$ lie in. The proof follows closely along the lines of the equivalent result in [15], with appropriate modifications to accommodate our generalized weighted nuclear norm. (Appendix C.3).

5.1 Comparison to Standard Matrix Completion:

It is instructive to consider our result in the context of noisy matrix completion with uniform samples. In this case, one would replace L_w, L_h by identity matrices, effectively ignoring graph information available. Specifically, the "standard" notion of spikiness ($\alpha_n := \sqrt{mn}\frac{\|Z\|_\infty}{\|Z\|_F}$) defined in [15] will apply, and the corresponding error bound (Theorem 2) will have $\alpha^\star$ replaced by $\alpha_n(Z^\star)$. In general, it is hard to quantify the relationship between α_g and α_n, and a detailed comparison is an interesting topic for future work. However, we show below using simulations for various scenarios that the former is much smaller than the latter. We generate $m \times m$ matrices of rank $k = 10$, $M = U\Sigma V^T$ with U, V being random orthonormal matrices and Σ having diagonal elements picked from a uniform$[0, 1]$ distribution. We generate graphs at random using the schemes discussed below, and set $Z = AMB^T$, with A, B as defined in Corollary 1. We then compute α_n, α_g for various m.

Comparing α_g to α_n: Most real world graphs exhibit a power law degree distribution. We generated graphs with the i^{th} node having degree $(m \times i^p)$ with varying negative p values. Figure 1(a) shows that as $p \to 0$ from below, the gains received from using our norm is clear compared to the standard nuclear norm. We also observe that in general the weighted formulation is never worse then unweighted (The dotted magenta line is $\alpha_n/\alpha_g = 1$). The same applies for random graphs, where there is an edge between each (i, j) with varying probability p (Figure 1(b)).

(a) Power Law (b) Random (c) Sample Complexity

Figure 1: (a), (b): Ratio of spikiness measures for traditional matrix completion and our formulation. (c): Sample complexity for the nuclear norm (NN) and generalized weighted nuclear norm (GWNN)

Sample Complexity: We tested the sample complexity needed to recover a $m = n = 200$, $k = 20$ matrix, generated from a power law distributed graph with $p = -0.5$. Figure 1(c) again outlines that the atomic formulation requires fewer examples to get an accurate recovery. We average the results over 10 independent runs, and we used [17] to solve the atomic norm constrained problem.

6 Experiments on Real Datasets

Comparison to Related Formulations: We compare GRALS to other methods that incorporate side information for matrix completion: the ADMM method of [12] that regularizes the entire target matrix; using known features (IMC) [10, 23]; and standard matrix completion (MC). We use the MOVIELENS $100k$ dataset,[6] that has user/movie features along with the ratings matrix. The dataset contains user features (such as age (numeric), gender (binary), and occupation), which we map

[6] http://grouplens.org/datasets/movielens/

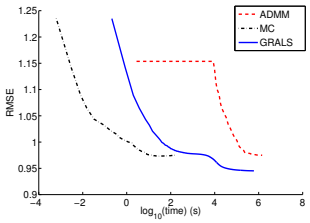


Figure 2: Time comparison of different methods on MOVIELENS 100k

Method	RMSE
IMC	1.653
Global mean	1.154
User mean	1.063
Movie mean	1.033
ADMM	0.996
MC	0.973
GRALS	**0.945**

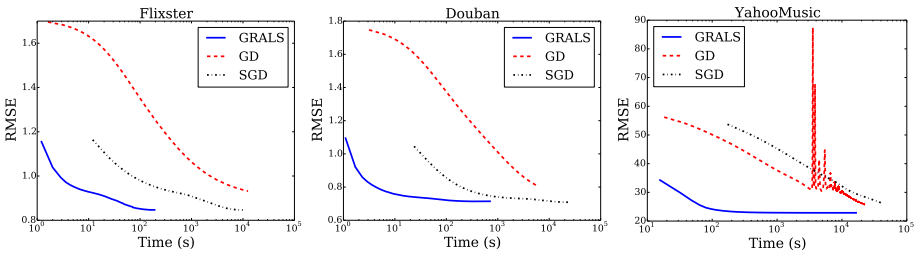
Table 1: RMSE on the MOVIELENS dataset

Table 2: Data statistics.

Dataset	# users	# items	# ratings	# links	rank used
Flixster ([11])	147,612	48,794	8,196,077	2,538,746	10
Douban ([14])	129,490	58,541	16,830,839	1,711,802	10
YahooMusic ([8])	249,012	296,111	55,749,965	57,248,136	20

into a 22 dimensional feature vector per user. We then construct a 10-nearest neighbor graph using the euclidean distance metric. We do the same for the movies, except in this case we have an 18 dimensional feature vector per movie. For IMC, we use the feature vectors directly. We trained a model of rank 10, and chose optimal parameters by cross validation. Table 1 shows the RMSE obtained for the methods considered. Figure 2 shows that the ADMM method, while obtaining a reasonable RMSE does not scale well, since one has to compute an SVD at each iteration.

Scalability of GRALS: We now demonstrate that the proposed GRALS method is more efficient than other state-of-the-art methods for solving the graph-regularized matrix factorization problem (5). We compare GRALS to the SGD method in [26], and GD: ALS with simple gradient descent. We consider three large-scale real-world collaborate filtering datasets with graph information: see Table 2 for details.[7] We randomly select 90% of ratings as the training set and use the remaining 10% as the test set. All the experiments are performed on an Intel machine with Xeon CPU E5-2680 v2 Ivy Bridge and enough RAM. Figure 3 shows orders of magnitude improvement in time compared to SGD. More experimental results are provided in the supplementary material.



(a) Flixster (b) Douban (c) YahooMusic

Figure 3: Comparison of GRALS, GD, and SGD. The x-axis is the computation time in log-scale.

7 Discussion

In this paper, we have considered the problem of collaborative filtering with graph information for users and/or items, and showed that it can be cast as a generalized weighted nuclear norm problem. We derived statistical consistency guarantees for our method, and developed a highly scalable alternating minimization method. Experiments on large real world datasets show that our method achieves ~ 2 orders of magnitude speedups over competing approaches.

Acknowledgments

This research was supported by NSF grant CCF-1320746 and the gift from Adobe. H.-F. Yu acknowledges support from an Intel PhD fellowship. NR was supported by an ICES fellowship.

[7] See more details in Appendix D.

References

[1] Jacob Abernethy, Francis Bach, Theodoros Evgeniou, and Jean-Philippe Vert. Low-rank matrix factorization with attributes. *arXiv preprint cs/0611124*, 2006.

[2] Francis Bach, Julien Mairal, and Jean Ponce. Convex sparse matrix factorizations. *CoRR*, abs/0812.1869, 2008.

[3] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *NIPS*, volume 14, pages 585–591, 2001.

[4] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 15(6):1373–1396, 2003.

[5] Deng Cai, Xiaofei He, Jiawei Han, and Thomas S Huang. Graph regularized nonnegative matrix factorization for data representation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(8): 1548–1560, 2011.

[6] Jian-Feng Cai, Emmanuel J Candès, and Zuowei Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4):1956–1982, 2010.

[7] Venkat Chandrasekaran, Benjamin Recht, Pablo A Parrilo, and Alan S Willsky. The convex geometry of linear inverse problems. *Foundations of Computational Mathematics*, 12(6):805–849, 2012.

[8] Gideon Dror, Noam Koenigstein, Yehuda Koren, and Markus Weimer. The yahoo! music dataset and kdd-cup'11. In *KDD Cup*, pages 8–18, 2012.

[9] Benjamin Haeffele, Eric Young, and Rene Vidal. Structured low-rank matrix factorization: Optimality, algorithm, and applications to image processing. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 2007–2015, 2014.

[10] Prateek Jain and Inderjit S Dhillon. Provable inductive matrix completion. *arXiv preprint arXiv:1306.0626*, 2013.

[11] Mohsen Jamali and Martin Ester. A matrix factorization technique with trust propagation for recommendation in social networks. In *Proceedings of the Fourth ACM Conference on Recommender Systems*, RecSys '10, pages 135–142, 2010.

[12] Vassilis Kalofolias, Xavier Bresson, Michael Bronstein, and Pierre Vandergheynst. Matrix completion on graphs. In *Neural Information Processing Systems 2014, Workshop" Out of the Box: Robustness in High Dimension"*, 2014.

[13] Wu-Jun Li and Dit-Yan Yeung. Relation regularized matrix factorization. In *21st International Joint Conference on Artificial Intelligence*, 2009.

[14] Hao Ma, Dengyong Zhou, Chao Liu, Michael R. Lyu, and Irwin King. Recommender systems with social regularization. In *Proceedings of the fourth ACM international conference on Web search and data mining*, WSDM '11, pages 287–296, Hong Kong, China, 2011.

[15] Sahand Negahban and Martin J Wainwright. Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *The Journal of Machine Learning Research*, 13(1):1665–1697, 2012.

[16] Sahand N Negahban, Pradeep Ravikumar, Martin J Wainwright, and Bin Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science*, 27(4): 538–557, 2012.

[17] Nikhil Rao, Parikshit Shah, and Stephen Wright. Conditional gradient with enhancement and truncation for atomic-norm regularization. In *NIPS Workshop on Greedy Algorithms*, 2013.

[18] Benjamin Recht. A simpler approach to matrix completion. *The Journal of Machine Learning Research*, 12:3413–3430, 2011.

[19] Alexander J Smola and Risi Kondor. Kernels and regularization on graphs. In *Learning theory and kernel machines*, pages 144–158. Springer, 2003.

[20] Nathan Srebro and Ruslan R Salakhutdinov. Collaborative filtering in a non-uniform world: Learning with the weighted trace norm. In *Advances in Neural Information Processing Systems*, pages 2056–2064, 2010.

[21] Ambuj Tewari, Pradeep K Ravikumar, and Inderjit S Dhillon. Greedy algorithms for structurally constrained high dimensional problems. In *Advances in Neural Information Processing Systems*, pages 882–890, 2011.

[22] Roman Vershynin. A note on sums of independent random matrices after ahlswede-winter. *Lecture notes*, 2009.

[23] Miao Xu, Rong Jin, and Zhi-Hua Zhou. Speedup matrix completion with side information: Application to multi-label learning. In *Advances in Neural Information Processing Systems*, pages 2301–2309, 2013.

[24] Yangyang. Xu and Wotao Yin. A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion. *SIAM Journal on Imaging Sciences*, 6(3):1758–1789, 2013.

[25] Zhou Zhao, Lijun Zhang, Xiaofei He, and Wilfred Ng. Expert finding for question answering via graph regularized matrix completion. *Knowledge and Data Engineering, IEEE Transactions on*, PP(99), 2014.

[26] Tinghui Zhou, Hanhuai Shan, Arindam Banerjee, and Guillermo Sapiro. Kernelized probabilistic matrix factorization: Exploiting graphs and side information. In *SDM*, volume 12, pages 403–414. SIAM, 2012.

A Proof of Theorem 1

Proof For all $Z = \sum_i c_i \mathbb{o}_i$ with $\|A\|_{\mathscr{A}} = \sum_i |c_i|$, where $\mathbb{o}_i = A\boldsymbol{u}_i \boldsymbol{v}_i^T B^T$, we can construct the i-th column of W and H as

$$\boldsymbol{w}_i = \sqrt{|c_i|} A\boldsymbol{u}_i \quad \text{and} \quad \boldsymbol{h}_i = \sqrt{|c_i|} B\boldsymbol{v}_i.$$

Clearly, we have $Z = WH^T$ and

$$\|A^{-1}W\|_F^2 = \|B^{-1}H\|_F^2 = \sum_i |c_i|$$

Thus, if follows that LHS $\geq$ RHS. Oh the other hand, for a matrix $Z = WH^T$, we can construct

$$\boldsymbol{u}_i = \frac{A^{-1}\boldsymbol{w}_i}{\|A^{-1}\boldsymbol{w}_i\|} \quad \text{and} \quad \boldsymbol{v}_i = \frac{B^{-1}\boldsymbol{h}_i}{\|B^{-1}\boldsymbol{h}_i\|},$$

and $c_i = \|A^{-1}\boldsymbol{w}_i\|\|B^{-1}\boldsymbol{h}_i\|$. Clearly, we have $\boldsymbol{w}_i\boldsymbol{h}_i^T = c_i A\boldsymbol{u}_i\boldsymbol{v}_i B^T$ and $Z = \sum_i c_i A\boldsymbol{u}_i\boldsymbol{v}_i^T B^T$. We also have

$$|c_i| = \|A^{-1}\boldsymbol{w}_i\|\|B^{-1}\boldsymbol{h}_i\| \leq \frac{1}{2}\big(\|A^{-1}\boldsymbol{w}_i\|^2 + \|B^{-1}\boldsymbol{h}_i\|^2\big)$$

by AM-GM inequality. Thus, we have LHS $\leq$ RHS. ■

B Dual Weighted Generalized Nuclear Norm

Recall the definition of the weighted atomic set:

$$\mathscr{A} := \{\mathbb{o}_i = \boldsymbol{w}_i\boldsymbol{h}_i^T : \boldsymbol{w}_i = A\boldsymbol{u}_i, \boldsymbol{h}_i = B\boldsymbol{v}_i, \|\boldsymbol{u}_i\| = \|\boldsymbol{v}_i\| = 1\}.$$

We derive the dual norm as follows.

$$\begin{aligned}
&\|Z\|_{\mathscr{A}}^* = \sup_{\mathbb{o}\in\mathscr{A}} \langle \mathbb{o}, Z\rangle \\
=&\sup_{\boldsymbol{u},\boldsymbol{v}} \langle A\boldsymbol{u}\boldsymbol{v}^T B^T, Z\rangle, \quad \text{s.t. } \|\boldsymbol{u}\| = \|\boldsymbol{v}\| = 1 \\
=&\sup_{\boldsymbol{u},\boldsymbol{v}} \operatorname{Tr}\big(B\boldsymbol{v}\boldsymbol{u}^T A^T Z\big), \quad \text{s.t. } \|\boldsymbol{u}\| = \|\boldsymbol{v}\| = 1 \\
=&\sup_{\boldsymbol{u},\boldsymbol{v}} \boldsymbol{u}^T A^T Z B\boldsymbol{v}, \quad \text{s.t. } \|\boldsymbol{u}\| = \|\boldsymbol{v}\| = 1 \\
=&\|A^T Z B\|
\end{aligned}$$

C Proof of Theorem 2

The proof of our main Theorem 2 follows the similar steps used in [15]. The main idea is to use Theorem 3 [16] to obtain the consistency guarantee. Our proof steps (and indeed that of [15]) are a consequence of carefully bounding the various quantities needed to make Theorem 3 hold:

Theorem 3 (Theorem 1 of [16])**.** *For the convex optimization problem of the following form:*

$$\hat{Z} = \arg\min_{Z\in\mathbb{R}^{m\times n}} \mathcal{L}(Z; X_1, \ldots, X_N) + \lambda\mathcal{R}(Z),$$

where

- *(a) the regularizer $\mathcal{R}$ is a norm and is* **decomposable** *with respect to the subspace pair $(\mathcal{M}, \bar{\mathcal{M}}^\perp)$, where $\mathcal{M} \subseteq \bar{\mathcal{M}}$ is a subspace.*
- *(b) the loss function $\mathcal{L}$ is convex and differentiable, and satisfies* **restricted strong convexity** *with curvature κ and tolerance τ*

with a strictly positive regularization constant $\lambda \geq 2\mathcal{R}^*(\nabla\mathcal{L}(Z^\star))$, any optimal solution $\hat{Z}$ satisfies the bound

$$\|\hat{Z} - Z^\star\|^2 \leq 9\frac{\lambda^2}{\kappa^2}\Psi(\mathcal{M})^2 + \frac{\lambda}{\kappa}\big\{2\tau^2(Z^\star) + 4\mathcal{R}(Z^\star_{M^\perp})\big\}, \tag{C.1}$$

where $\Psi(\mathcal{M}) := \sup_{Z\in\mathcal{M}\setminus\{0\}} \frac{\mathcal{R}(Z)}{\|Z\|_F}$. Furthermore, if $Z^\star \in \mathcal{M}$, then the bound becomes

$$\|\hat{Z} - Z^\star\|^2 \leq 9\frac{\lambda^2}{\kappa^2}\Psi(\mathcal{M})^2. \tag{C.2}$$

See [16] for the detailed definitions of **decomposable norms** and **restricted strong convexity**.

To apply Theorem 3 to analyze the consistency of (15), we make the following remarks:

- $\mathcal{R}(Z) = \|Z\|_{\mathscr{A}}$: the weighted atomic norm defined in (9).
- $\mathcal{R}^*(Z) = \|Z\|^*_{\mathscr{A}}$: the dual norm of the weighted atomic norm.
- $\mathcal{M} = \{Z = AMB^T : \mathrm{rank}(Z) = k\}$: the subspace we are interested in.
- $\mathcal{L}(Z; X_1, \ldots, Z_n) = \frac{1}{N}\sum_{i=1}^N (y_t - \langle X_t, Z\rangle)^2$: where $X_t = \boldsymbol{e}_{i(t)}\boldsymbol{e}^T_{j(t)}$ (See the corresponding measurement model in (12)). Because the squared-L_2 loss is used in our setting, the restricted strong convexity parameter κ is related to the minimum singular value of the Hessian of $\mathcal{L}(Z; X_1, \ldots, X_N)$. Thus, from (C.1) and (C.2) we can see that the bounds remain the same when we scale $\{X_t\}$ and $\{y_t\}$ by the same constant as both κ and the lower bound of λ (which is $\mathcal{R}^*(\nabla\mathcal{L}(Z^\star))$) are scaled with the same constant. Thus, in the following proof, we consider the following equivalent statistical measurement model:

  $$y_t = \langle\sqrt{mn}\epsilon_t\boldsymbol{e}_{i(t)}\boldsymbol{e}^T_{j(t)}, Z^\star\rangle + \sigma\epsilon_t\eta_t \tag{C.3}$$

  where ϵ_t are i.i.d. Rademacher random variables [15]. Let's re-define

  $$\begin{aligned} X_t &:= \sqrt{mn}\epsilon_t\boldsymbol{e}_{i(t)}\boldsymbol{e}^T_{j(t)}, \\ y_t &:= \langle X_t, Z\rangle + \sigma\epsilon_t\eta_t. \end{aligned} \tag{C.4}$$

  In addition, we also define $\mathscr{X}(Z) \in \mathbb{R}^N$ be the vector such that $\mathscr{X}(Z)_t = \langle X_t, Z\rangle$.
- The exact restricted strong convexity condition we need for (15) is as follows:

  $$\frac{1}{\sqrt{N}}\|\mathscr{X}(Z)\| \geq \frac{1}{8}\|Z\|_F\left\{1 - \hat{c}_0\frac{\alpha_g(Z)}{\sqrt{N}}\right\} \ \forall Z \in \mathcal{C}, \tag{C.5}$$

  where $\mathcal{C}$ is defined in (15) and $\hat{c}_0$ is a constant (similar to [15, Eq. 28]).

In the following subsections, we prove bounds for the quantities needed for establishing Theorem 2 via the following steps:

- In Section C.1, we derive an upper bound for $\Psi(\mathcal{M})$.
- In Section C.2, we derive an upper bound for $\mathcal{R}^*(\nabla\mathcal{L}(Z^*))$.
- In Section C.3, we prove that the restricted strong convexity (C.5) holds $\mathcal{L}$ with exponentially high probability.

Note that for the sake of proving our results, we assume that the target matrix $Z^\star$ is exactly low rank, and the minimum singular values of $A,\ B$ are 1. Our results can be extended in a straightforward manner when $Z^\star$ does not exactly lie in $\mathcal{M}$ (it is approximately low rank).

C.1 Bounding the Atomic Norm

Based on the definition of $\Psi(\mathcal{M})$, we can derive its upper bound on the atomic norm of $Z \in \mathcal{M}$ with $\|Z\|_F = 1$.

Lemma 1. *Let $Z \in \mathbb{R}^{m\times n},\ Z = AUV^TB^T,\ \mathrm{rank}(Z) = k$ be a linear combination of atoms in $\mathscr{A}$. Then, with the assumption $\|Z\|_F = 1$ we have*

$$\|Z\|_{\mathscr{A}} \leq \sqrt{k}$$

Proof

$$\|Z\|_{\mathscr{A}} = \|UV^T\|_* \leq \sqrt{k}\|UV^T\|_F \leq \sqrt{k}\|A^{-1}AUV^TB^{-T}\|_F,$$

where the first inequality follows from Cauchy Schwartz, and the second inequality follows from noting that $\|A^{-1}\| \leq 1$ and likewise for $\|B^{-1}\|$, since we assumed that the minimum singular value of both L_w, L_h is unity. ■

C.2 Bounding the Dual Norm of the Gradient of Loss Function

A key ingredient for our main result will be a bound on the dual norm of the gradient of the loss function, which we will use to bound the regularizer λ. From Eq. (11), and our problem set up in Eq. (16), we have the following set of inequalities:

$$\begin{aligned}\|\nabla\mathcal{L}\|_{\mathscr{A}}^* &= \|S_w^{-\frac{1}{2}}U_w^T\nabla\mathcal{L}U_hS_h^{-\frac{1}{2}}\| \overset{(i)}{\leq} \|S_w^{-\frac{1}{2}}U_w^T\|\|U_hS_h^{-\frac{1}{2}}\|\|(\nabla\mathcal{L})\| \\ &= \frac{\|\nabla\mathcal{L}\|}{\sigma_{min}(L_w^{\frac{1}{2}})\sigma_{min}(L_h^{\frac{1}{2}})} \overset{(ii)}{\leq} C\sigma\sqrt{\frac{(m+n)\log(m+n)}{N}}, \end{aligned} \tag{C.6}$$

with probability at least $1 - \exp\Big(c\sqrt{N\log(m+n)}\Big)$. (i) appeals to submultiplicativity, and we prove (ii) below. From our assumption about unit minimum singular values, we can ignore the denominator.

Here we develop a bound on the spectral norm of the gradient of the loss function, specifically step (ii) in (C.6). Our proof follows that of the corresponding result in [15], which we show here for completeness.

Recall the definition of $X_t := \sqrt{mn}\epsilon_t\boldsymbol{e}_{i(t)}\boldsymbol{e}_{j(t)}^T$ in (C.4), we have the gradient of the loss function given by

$$\nabla\mathcal{L} = \frac{\sigma}{N}\left\|\sum_{t=1}^{N}\eta_tX_t\right\| \tag{C.7}$$

For ease of exposition, assume $m = n$. We now show that with high probability, the quantity in (C.7) is bounded above by $C\sigma\sqrt{\frac{m\log(m)}{N}}$. For $m \neq n$, our bound can be made necessarily better since the result we prove can be seen as holding for $\max\{m, n\}$. To prove our result, we make use of the matrix noncommutative Bernstein inequality (Theorem 3.2 in [18]):

Lemma 2. *Let $X_1, \cdots, X_N$ be independent, zero mean random matrices of size $m \times n$. Suppose $\rho_t^2 := \max\{\|\mathbb{E}[X_tX_t^T]\|, \|\mathbb{E}[X_t^TX_t]\|\}$, and suppose $\|X_t\| \leq \bar{M}$ almost surely $\forall t$. Then for any $\tau > 0$*

$$\mathbb{P}\left[\left\|\sum_{t=1}^{N}X_t\right\| > \tau\right] \leq (m+n)\exp\left(\frac{-\frac{\tau^2}{2}}{\sum_{t=1}^{N}\rho_t^2 + \frac{M\tau}{3}}\right)$$

The above result holds even for sub-exponential random variables [22] and the Orlicz norm instead of the spectral norm being bounded above by a constant $\bar{M}$.

To use Lemma 2, we first derive bounds on the relevant quantities. First, note that for all t, X_t has a single non zero entry of magnitude m. Noting that η_t is a standard Gaussian random variable, we can bound the Orlicz norm $\|\eta_tX_t\|_{\psi 1} \leq m$. Also

$$\mathbb{E}\big[\eta_t^2X_t^TX_t\big] = \mathbb{E}\Big[m^2\boldsymbol{e}_{j(t)}\boldsymbol{e}_{i(t)}^T\boldsymbol{e}_{i(t)}\boldsymbol{e}_{j(t)}^T\Big] = m^2\mathbb{E}\Big[\boldsymbol{e}_{j(t)}\boldsymbol{e}_{j(t)}^T\Big]$$

The matrix inside the expectation has a 1 in the $j(t), j(t)$ location. Since $j(t)$ is chosen uniformly at random, the expected value of the non zero entry is $1/m$. This means

$$\|\mathbb{E}\big[\eta_t^2X_t^TX_t\big]\| = m = \|\mathbb{E}\big[\eta_t^2X_tX_t^T\big]\|$$

This gives $M = \rho_t^2 = m$. Setting $\tau = N\delta$, and from Lemma 2, we get

$$\mathbb{P}\left[\frac{\sigma}{N}\left\|\sum_{t=1}^{N}\eta_tX_t\right\| > \sigma\delta\right] \leq 2m\exp\left(-\frac{CN\delta}{m}\right)$$

Our result then follows by setting $\delta = c\sqrt{\frac{m\log(m)}{N}}$.

C.3 Restricted Strong Convexity for Generalized Weighted Nuclear Norm

The proof of this result mirrors the corresponding proof in [15]. Hence, to keep things simple, we skip the steps that are common between our method and [15], and only pause to highlight the differences.

First, note that since we assume uniformly weighted samples, we need not concern ourselves with the "weight" matrices that are considered in [15]. Also, define $\mathscr{X}(Z)_t = \langle X_t, Z\rangle$, where X_t is defined as in Appendix C.2. Then, the RSC condition requires us to prove (C.5), which we re-state it again as follows.

$$\frac{1}{\sqrt{N}}\|\mathscr{X}(Z)\| \geq \frac{1}{8}\|Z\|_F\left\{1 - \hat{c}_0\frac{\alpha_g(Z)}{\sqrt{N}}\right\} \ \forall Z \in \mathcal{C},$$

where $\mathcal{C}$ is defined in (15) and $\hat{c}_0$ is a constant In other words, we wish to prove that the following event holds with high probability:

$$\mathcal{E}_1 := \left\{\forall Z \in \mathcal{C} : \frac{1}{\sqrt{N}}\|\mathscr{X}(Z)\| \geq \frac{1}{8}\|Z\|_F - \frac{\hat{c}_0 m}{8\sqrt{N}}\|Z\|_F\frac{\|M\|_\infty}{\|M\|_F}\right\}, \tag{C.8}$$

where $M := A^{-1}ZB^{-T}$ and α_g is the spikiness defined in (13). Subtracting $\|Z\|_F$ from both sides of the inequality in (C.8), we get

$$\frac{1}{\sqrt{N}}\|\mathscr{X}(Z)\| - \|Z\|_F \geq -\frac{7}{8}\|Z\|_F - \frac{\hat{c}_0 m}{8\sqrt{N}}\|Z\|_F\frac{\|M\|_\infty}{\|M\|_F},$$

and hence we can define a "bad" event as

$$\mathcal{E}_2 := \left\{\exists Z \in \mathcal{C} : \left|\frac{1}{\sqrt{N}}\|\mathscr{X}(Z)\| - \|Z\|_F\right| > \frac{7}{8}\|Z\|_F + \frac{\hat{c}_0 m}{8\sqrt{N}}\|Z\|_F\frac{\|M\|_\infty}{\|M\|_F}\right\} \tag{C.9}$$

Now, due to the definition of $\mathcal{C}$, event $\mathcal{E}_2$ is invariant under rescaling of Z (so as $M := A^{-1}ZB^{-T}$). Thus, without loss of generality, we may assume that $\|M\|_\infty = 1/m$. Then, the remaining degrees of freedom in the set $\mathcal{C}$ can be parameterized in terms of the quantities $D = \|M\|_F$ and $\rho = \|M\|_*$. For any $Z = AMB^T \in \mathcal{C}$ with $\|M\|_\infty = 1/d$ and $\|M\|_F \leq D$, we have $\|M\|_* \leq \rho(D)$, where

$$\rho(D) := \bar{c}_0 D^2\left(\frac{N}{m\log(m)}\right)^{\frac{1}{2}}.$$

For each radius $D > 0$, consider the set

$$\mathscr{B}(D) := \left\{Z = AMB^T : \ \|M\|_\infty = 1/m, \ \|M\|_F \leq D, \ \|M\|_* \leq \rho(D)\right\}, \tag{C.10}$$

and consider the event

$$\mathcal{E}_3(D) := \left\{\exists Z \in \mathscr{B}(D) : \left|\frac{1}{\sqrt{N}}\|\mathscr{X}(Z)\| - \|Z\|_F\right| > \frac{3}{4}D + \frac{\hat{c}_0 m}{8\sqrt{N}}\|Z\|_F\frac{\|M\|_\infty}{\|M\|_F}\right\} \tag{C.11}$$

Now, note that the RHS of inequality in the above event satisfies, for $Z \in \mathscr{B}(D)$

$$\begin{aligned}\frac{3}{4}D + \frac{\hat{c}_0 m}{8\sqrt{N}}\|Z\|_F\frac{\|M\|_\infty}{\|M\|_F} &= \frac{3}{4}D + \frac{\hat{c}_0}{8\sqrt{N}}\frac{\|Z\|_F}{\|M\|_F} \\ &\geq \frac{3}{4}D + \frac{\hat{c}_0}{8\sqrt{N}},\end{aligned}$$

where the first equality follows since $Z = AMB^T \in \mathcal{C} \Rightarrow \|M\|_\infty = 1/m$, and the last inequality follows since $\|Z\|_F = \|AMB^T\|_F \geq \sigma_{min}(A)\sigma_{min}(B)\|M\|_F$, and noting that the minimum singular values of A, B are unity. Finally, we define the event

$$\mathcal{E}_4(D) := \left\{\exists Z \in \mathscr{B}(D) : \left|\frac{1}{\sqrt{N}}\|\mathscr{X}(Z)\| - \|Z\|_F\right| \geq \frac{3}{4}D + \frac{\hat{c}_0}{8\sqrt{N}}\right\} \tag{C.12}$$

Let $\mathcal{S}_1$ be the set of Z that satisfy event $\mathcal{E}_1$, and similarly define sets $\mathcal{S}_2, \mathcal{S}_3(D), \mathcal{S}_4(D)$. The following statement will be used to prove our results: for each fixed $D > 0$,

$$\mathcal{S}_4(D) \supset \mathcal{S}_3(D) \supset \mathcal{S}_2 \supset \mathcal{S}_1^c \tag{C.13}$$

Meaning that if we can show that $\mathcal{E}_4$ holds with very low probability for a fixed D, then it follows from (C.13) that $\mathcal{E}_1$ holds with high probability. The remainder of the proof will focus on doing so.

First, note that the event $\mathcal{E}_4$ defined in (C.12) is exactly the same as the event defined in [15, Eq. 29]. Hence, we can use the exact same argument as described in [15, Section 5.2] to obtain

$$\mathbb{P}(\mathcal{E}_4(D)) \leq c_1 \exp\big(-c_2 N D^2\big).$$

Now, we have the following result:

Lemma 3. *Suppose there are constants c_1, c_2 so that, for each fixed $D > 0$,*

$$\mathbb{P}(\mathcal{E}_4(D)) \leq c_1 \exp\big(-c_2 N D^2\big)$$

then $\exists$ a universal constant c_2' so that

$$\mathbb{P}(\mathcal{E}_2) \leq c_1 \frac{\exp(-c_2' m \log(m))}{1 - \exp(-c_2' m \log(m))}.$$

The statement is the same as [15, Lemma 3], but we have to slightly modify the proof to adapt it to our setting. We do this in Appendix C.4.

Lemma 3 allows us to shows that if $\mathcal{E}_4$ holds with low probability, then $\mathcal{E}_2$ holds with low probability as well. Since by construction, $\mathcal{E}_1^c \subset \mathcal{E}_2$, the RSC result follows.

Since the results derived here are for the statistical model defined by (C.3), we go from this model to the initial model that we consider in (12). To this end, one needs to make the following two transformations, as explained in the remarks following Theorem 3:

- Scale the magnitude of X_t, and consequently λ by $1/m$
- Scale the noise variance σ by m.

The rates we obtain in Theorem 2 remain unchanged as a result of this scaling.

C.4 Proof of Lemma 3

The proof is similar to [15, Lemma 3], we include it with our notation for completeness. For any $Z = AMB^T \in \mathcal{C}$, with $\|M\|_\infty = 1/m$, based on the definition of $\mathcal{C}$ in (15), we have

$$\|M\|_F^2 \geq \bar{c}_0^{-1} \|M\|_* \left(\frac{m \log(m)}{N}\right)^{\frac{1}{2}} \geq \bar{c}_0^{-1} \|M\|_F \left(\frac{m \log(m)}{N}\right)^{\frac{1}{2}},$$

which gives us $\|M\|_F \geq \bar{c}_0^{-1} \left(\frac{m \log(m)}{N}\right)^{\frac{1}{2}}$. Hence, we only need to focus on sets $\mathscr{B}(D)$ where $D > \mu := \bar{c}_0^{-1} \left(\frac{m \log(m)}{N}\right)^{\frac{1}{2}}$. For $l = 1, 2, \ldots$ and $a = \frac{7}{6}$ define

$$S_l := \big\{ Z = AMB^T \in \mathcal{C} : \ \|M\|_\infty = 1/m, \ \ a^{l-1}\mu \leq \|M\|_F \leq a^l \mu, \ \ \|M\|_* \leq \rho(a^l \mu) \big\}$$

From the definition of (C.9), we have $S_l \subset \mathscr{B}(a^l \mu)$. Now, if $\mathcal{E}_2$ holds for some Z, then Z must belong to S_l for some l. When $Z \in S_l$, we know $\exists Z \in \mathscr{B}(a^l \mu)$ such that

$$\begin{aligned}
\left| \frac{\|\mathscr{X}(Z)\|}{\sqrt{N}} - \|Z\|_F \right| &\geq \frac{7}{8} \|Z\|_F + \frac{\hat{c}_0 m}{8\sqrt{N}} \|Z\|_F \frac{\|M\|_\infty}{\|M\|_F} \\
&\geq \frac{7}{8} \|Z\|_F + \frac{\hat{c}_0}{8\sqrt{N}} \\
&\geq \frac{7}{8} a^{l-1} \mu + \frac{\hat{c}_0}{8\sqrt{N}} \\
&= \frac{3}{4} a^l \mu + \frac{\hat{c}_0}{8\sqrt{N}} \qquad \text{since a = } 7/6.
\end{aligned}$$

Thus, we have shown that when this $Z \in S_l$, then $\mathcal{E}_4(a^l\mu)$ must hold. Because any Z which make the event $\mathcal{E}_2$ hold must fall into some set S_l, the union bound implies that

$$\begin{aligned}
\mathbb{P}[\![\mathcal{E}_2]\!] &\leq \sum_{l=1}^{\infty} \mathbb{P}\left[\![\mathcal{E}(a^l\mu)]\!\right] \\
&\leq c_1 \sum_{l=1}^{\infty} \exp\left(-c_2 N a^{2l}\mu^2\right) \\
&\leq c_1 \sum_{l=1}^{\infty} \exp\left(-2c_2 \log(a) l N\mu^2\right) \\
&\leq c_1 \frac{\exp(-\bar{c}_2 N\mu^2)}{1-\exp(-\bar{c}_2 N\mu^2)} \\
&= c_1 \frac{\exp(-c_2' m\log m)}{1-\exp(-c_2' m\log m)},
\end{aligned}$$

where the last equality follows as $N\mu^2 = \bar{c}_0^{-1}(m\log m)$.

D Additional Details for Experimental Results

Experimental environment and Implementation. All the experiments are generated on an Intel machine with 2 Xeon CPU E5-2680 v2 Ivy Bridge and 256 GB ram. GRALS is implemented using a MEX routine written in C++. For SGD and GD, we optimize the code from [26] in several ways: vectorization of for-loops and parallel residual computation using a MEX routine using C++. All the implementations employ embarrassing parallelization for BLAS operations whenever applicable (either through parallel BLAS library in Matlab, or simple OpenMP parallel-for loop).

Parameters. In Section 6, we show the results in Figure 3 to demonstrate the superiority of the proposed algorithm GRALS over the existing approaches: SGD and GD [26]. In Table Supp-1, we list the parameters used to generate the results. Note that in all the datasets we used, there is only one set of variables which comes with graph information (say W). Thus, the regularization consists of three terms as follows:

$$\lambda_L \operatorname{Tr}\left(W^T \mathbf{Lap}(E^w) W\right) + \lambda_w \|W\|_F^2 + \lambda_h \|H\|_F^2.$$

In addition to the regularization parameters, there are algorithmic parameters for each approach:

- GRALS: the number of CG iterations to solve each sub-problem
- SGD: the learning rate, η_{sgd}
- GD: the learning rate, η_{gd}

In Table Supp-1, we also report the best algorithm-specific parameters for each method.

Table Supp-1: Parameters used in the experiments for Figure 3

	λ_L	λ_w	λ_h	GRALS CG-iters	SGD η_{sgd}	GD η_{gd}
Flixster	0.01	0.01	0.02	3	10^{-4}	10^{-6}
Douban	1	0.01	1.01	5	10^{-4}	10^{-6}
YahooMusic	100	100	200	20	10^{-6}	10^{-6}

Graph Information in Datasets. For Flixster and Douban, the datasets come with the graph information among users. For YahooMusic, we use the Yahoo Music Track 2 dataset from KDDCup 2011 [8] for the purpose of showing that GRALS scales much better than other approaches. As most of entries in the test split of the Track 2 dataset are marked as -1 (for the classification purpose in that track), we only use the training set in our experiments. The original training set is randomly partitioned into a $90-10$ training-test split. There is no explicit graph information in YahooMusic. Thus, we use the provided "album", "artist", and "genre" attributes for each item (or music track)

to construct a binary indicator vector and construct a 10-NN graph graph using the inner product distance over all the items.

RMSE Performance. Because the aim of this paper is to develop scalable algorithms and consistency results for graph regularized matrix factorization (4), we did not include the performance comparison table (similar to Table 1) for other large datasets for want of space. Here, we report the results in Table Supp-2. Note that there are other approaches to incorporate graph information into collaborative filtering, which might lead to different RMSEs. A detailed comparison to all such methods is beyond the scope of this paper. However, whenever there is a means to incorporate pairwise relationships between user-user variables or item-item variables, we can use GRALS to achieve the same results as other approaches, but at a much faster rate. Note that the Yahoo Music dataset has ratings in the range $[0, 100]$ and hence the larger RMSE values. A fairer comparison can be obtained by dividing the results by 20, to correspond to ratings in the range $[0, 5]$.

Table Supp-2: RMSE of various methods on the datasets considered in Figure 3. PMF : Our method with graph Laplacians replaced by identity matrices.

DATASET	PMF	GRALS	Global Mean	User Mean	Item Mean
Flixster	0.923	0.845	1.092	0.979	1.088
Douban	0.719	0.714	0.907	0.848	0.790
YahooMusic	23.823	22.872	37.941	43.308	38.042