Recovery guarantee of weighted low-rank approximation via alternating minimization

Yuanzhi Li YUANZHIL@CS.PRINCETON.EDU
Yingyu Liang YINGYUL@CS.PRINCETON.EDU
Andrej Risteski RISTESKI@CS.PRINCETON.EDU
Princeton University, 35 Olden St, Princeton, NJ 08540 USA

Abstract

Many applications require recovering a ground truth low-rank matrix from noisy observations of the entries, which in practice is typically formulated as a weighted low-rank approximation problem and solved by non-convex optimization heuristics such as alternating minimization. In this paper, we provide provable recovery guarantee of weighted low-rank via a simple alternating minimization algorithm. In particular, for a natural class of matrices and weights and without any assumption on the noise, we bound the spectral norm of the difference between the recovered matrix and the ground truth, by the spectral norm of the weighted noise plus an additive error term that decreases exponentially with the number of rounds of alternating minimization, from either initialization by SVD or, more importantly, random initialization. These provide the first theoretical results for weighted low-rank approximation via alternating minimization with non-binary deterministic weights, significantly generalizing those for matrix completion, the special case with binary weights, since our assumptions are similar or weaker than those made in existing works. Furthermore, this is achieved by a very simple algorithm that improves the vanilla alternating minimization with a simple clipping step.

1. Introduction

Recovery of low-rank matrices has been a recurring theme in recent years in machine learning, signal processing, and numerical linear algebra, since in many applications, the data is a noisy observation of a low-rank ground truth matrix. Typically, the noise on different entries is not identically distributed, which naturally leads to a weighted low-rank approximation problem: given the noisy observation $\mathbf{M}$, one tries to recover the ground truth by finding $\widetilde{\mathbf{M}}$ that minimizes $\|\mathbf{M}-\widetilde{\mathbf{M}}\|_{\mathbf{W}}^2 = \sum_{ij} \mathbf{W}_{i,j}(\mathbf{M}_{i,j}-\widetilde{\mathbf{M}}_{i,j})^2$ where the weight matrix $\mathbf{W}$ is chosen according to prior knowledge about the noise. For example, the co-occurrence matrix for words in natural language processing applications (Pennington et al., 2014; Arora et al., 2016) is such that the noise is larger when the co-occurrence of two words is rarer. When doing low-rank approximation on the co-occurrence matrix to get word embeddings, it has been observed empirically that a simple weighting can lead to much better performance than the unweighted formulation (see, e.g., (Levy & Goldberg, 2014)). In biology applications, it is often the case that the variance of the noise is different for each entry of a data matrix, due to various reasons such as different properties of different measuring devices. A natural approach to recover the ground truth matrix is to solve a weighted low-rank approximation problem where the weights are inversely proportional to the variance in the entries (Gadian, 1982; Wentzell et al., 1997). Even for collaborative filtering, which is typically modeled as a matrix completion problem that assigns weight 1 on sampled entries and 0 on non-sampled entries, one can achieve better results when allowing non-binary weights (Srebro & Jaakkola, 2003).

In practice, the weighted low-rank approximation is typically solved by non-convex optimization heuristics. One of the most frequently used is alternating minimization, which sets $\widetilde{\mathbf{M}}$ to be the product of two low-rank matrices and alternates between updating the two matrices. Although it is a natural heuristic to employ and also an interesting theoretical question to study, to the best of our knowledge there is no guarantee for alternating minimization for weighted low-rank approximation. Moreover, general weighted low-rank approximation is NP-hard, even when the ground truth is a rank-1 matrix (Gillis & Glineur, 2011).

Proceedings of the 33^{rd} International Conference on Machine Learning, New York, NY, USA, 2016. JMLR: W&CP volume 48.

A special case of weighted low-rank approximation is matrix completion, where the weights are binary. Most methods proposed for solving this problem rely on the assumptions that the observed entries are sampled uniformly at random, and additionally often the observations need to be re-sampled across different iterations of the algorithm. This is inherently infeasible for the more general weighted low-rank approximation, and thus their analysis is not portable to the more general problem. The few exceptions that work with deterministic weights are (Heiman et al., 2014; Lee & Shraibman, 2013; Bhojanapalli & Jain, 2014). In this line of work the state-of-the-art is (Bhojanapalli & Jain, 2014), who proved recovery guarantees under the assumptions that the ground truth has a strong version of incoherence and the weight matrix has a sufficiently large spectral gap. However, their results still only work for binary weights, use a nuclear norm convex relaxation and do not consider noise on the observed entries.

In this paper, we provide the first theoretical guarantee for weighted low-rank approximation via alternating minimization, under assumptions generalizing those in (Bhojanapalli & Jain, 2014). In particular, assuming that the ground truth has a strong version of incoherence and the weight matrix has a sufficiently large spectral gap, we show that the spectral norm of the difference between the recovered matrix and the ground truth matrix is bounded by the spectral norm of the weighted noise plus an additive error term that decreases exponentially with the number of rounds of alternating minimization, from either initialization by SVD or, more importantly, random initialization. We emphasize that the bounds hold without any assumption on the noise, which is particularly important for handling complicated noise models. Since uniform sampling can satisfy our assumptions, our guarantee naturally generalizes those in previous works on matrix completion. See Section 4.1 for a detailed comparison.

The guarantee is proved by showing that the distance between the intermediate solution and the ground truth is improved at each iteration, which in spirit is similar to the framework in previous works. However, the lack of randomness in the weights and the exclusion of re-sampling (i.e., using independent samples at each iteration) lead to several technical obstacles that need to be addressed. Our proof of the improvement is then significantly different (and more general) from previous ones. In particular, showing improvement after each step is only possible when the intermediate solution has some additional special properties in terms of incoherence and spectrum. Prior works ensure such properties by using re-sampling (and sometimes assumptions about the noise), which are not available in our setting. We address this by showing that the spectral property only needs to hold in an average sense, which can be achieved by a simple clipping step. This results in a very simple algorithm that almost matches the practical heuristics, and thus provides explanation for them and also suggests potential improvement of the heuristics.

Further results The above results build on the insight that the spectral property only need to hold in an average sense. However, we can even make sure that the spectral property holds at each step strictly by a whitening step. More precisely, the clipping step is replaced by a whitening step using SDP and Rademacher rounding, which ensures that the intermediate solutions are incoherent and have the desired spectral property (the smallest eigenvalues of some related matrices are bounded). The technique of maintaining the smallest eigenvalues may be applicable to some other non-convex problems, and thus is of independent interest. The details are presented in the full version of this paper on arXiv (Li et al., 2016).

Furthermore, combining our insight that the spectral property only need to hold in an average sense with the framework in (Sun & Luo, 2015), we are able to show provable guarantees for the family of algorithms analyzed there, including stochastic gradient descent. We demonstrate this by including the proof details for stochastic gradient descent in the full version (Li et al., 2016).

2. Related work

Being a common practical problem (e.g., (Lu et al., 1997; Srebro & Jaakkola, 2003; Li et al., 2010; Eriksson & van den Hengel, 2012)), multiple heuristics for non-convex optimization such as alternating minimization have been developed, but they come with no guarantees. On the other hand, weighted low-rank approximation is NP-hard in the worst case, even when the ground truth is a rank-1 matrix (Gillis & Glineur, 2011).

On the theoretical side, the only result we know of is (Razenshteyn et al., 2016), who provide a fixed-parameter tractability result when *additionally* the weight matrix is low-rank. Namely, when the weight matrix has rank r, they provide an algorithm for outputting a matrix $\widetilde{\mathbf{M}}$ which approximates the optimization objective up to a $1+\epsilon$ multiplicative factor, and runs in time $n^{O(k^r/\epsilon)}$.

A special case of weighted low rank approximation is matrix completion, where the goal is to recover a low-rank matrix from a subset of the matrix entries and corresponds to the case when the weights are in $\{0,1\}$. For this special case much more is known theoretically. It is known that matrix completion is NP-hard in the case when the $k = 3$ (Peeters, 1996). Assuming that the matrix is incoherent and the observed entries are chosen uniformly at random, Candès & Recht (2009) showed that nuclear norm convex relation can recover an $n \times n$ rank-k matrix us-

ing $m = O(n^{1.2}k\log(n))$ entries. The sample size is improved to $O(nk\log(n))$ in subsequent papers (Candès & Tao, 2010; Recht, 2011; Gross, 2011). Candes & Plan (2010) relaxed the assumption to tolerate noise and showed the nuclear norm convex relaxation can lead to a solution such that the Frobenius norm of the error matrix is bounded by $O(\sqrt{n^3/m})$ times that of the noise matrix. However, all these results are for the restricted case with uniformly random binary weight matrices.

The only relaxations to random sampling to the best of our knowledge are in (Heiman et al., 2014; Lee & Shraibman, 2013; Bhojanapalli & Jain, 2014). In this line the state-of-the-art is (Bhojanapalli & Jain, 2014), where the support of the observation is a d-regular expander such that the weight matrix has a sufficiently large spectral gap. However, it only works for binary weights, and is for a nuclear norm convex relaxation and does not incorporate noise.

Recently, there is an increasing interest in analyzing non-convex optimization techniques for matrix completion. In two seminal papers (Jain et al., 2013; Hardt, 2014), it was shown that with an appropriate SVD-based initialization, the alternating minimization algorithm (with a few modifications) recovers the ground-truth. These results are for random binary weight matrix and crucially rely on re-sampling (i.e., using independent samples at each iteration), which is inherently not possible for the setting studied in this paper. More recently, Sun & Luo (2015) proved recovery guarantees for a family of algorithms including alternating minimization on matrix completion without re-sampling. However, the result is still for random binary weights and has not considered noise. More detailed comparison of our result with prior work can be found in Section 4, and comments on whether their arguments can be applied in our setting can be found in Section 5.

We also mention (Negahban & Wainwright, 2012) who consider random sampling, but one that is not uniformly random across the entries. In particular, their sampling produces a rank-1 matrix. (Additionally, they require the ground truth matrix to have nice properties such as low-rankness and spikiness.) The rank-1 assumption on the weight matrix is typically not true for many applications that introduce the weights to battle the different noise across the different entries of the matrix.

Finally, two related works are (Bhojanapalli et al., 2015a;b). The former implements faster SVD decomposition via weighted low rank approximation. However, here the weights in the weighted low rank problem come from leverage scores, so have a very specific structure, specially designed for performing SVD decompositions. The latter concerns optimization of strongly convex functions $f(\mathbf{V})$ when $\mathbf{V}$ is in the set of positive-definite matrices. It does this in a non-convex manner, by setting $\mathbf{V} = \mathbf{U}\mathbf{U}^\top$ and using the entries of $\mathbf{U}$ as variables. Our work focus on the recovery of the ground truth under the generative model, rather than on the optimization.

3. Problem definition and assumptions

For a matrix $\mathbf{A}$, let $\mathbf{A}_i$ denote its i-th column, $\mathbf{A}^j$ denote its j-th row, and $\mathbf{A}_{i,j}$ denote the element in i-th row and j-th column. Let $\odot$ denote the Hadamard product, i.e., $\mathbf{C} = \mathbf{A} \odot \mathbf{B}$ means $\mathbf{C}_{i,j} = \mathbf{A}_{i,j}\mathbf{B}_{i,j}$.

Let $\mathbf{M}^* \in \mathbb{R}^{n\times n}$ be a rank-k matrix. Given the observation $\mathbf{M} = \mathbf{M}^* + \mathbf{N}$ where $\mathbf{N}$ is a noise matrix, we want to recover the ground truth $\mathbf{M}^*$ by solving the weighted low-rank approximation problem for $\mathbf{M}$ and a non-negative weight matrix $\mathbf{W}$:

$$\min_{\widetilde{\mathbf{M}}\in\mathcal{R}_k} \left\|\widetilde{\mathbf{M}} - \mathbf{M}\right\|_{\mathbf{W}}^2$$

where $\mathcal{R}_k$ is the set of rank-k n by n matrices, and $\|\mathbf{A}\|_{\mathbf{W}}^2 = \sum_{i,j} \mathbf{W}_{i,j}\mathbf{A}_{i,j}^2$ is the weighted Frobenius norm. Our goal is to specify conditions about $\mathbf{M}^*$ and $\mathbf{W}$, under which $\mathbf{M}^*$ can be recovered up to small error by alternating minimization, i.e., set $\widetilde{\mathbf{M}} = \mathbf{X}\mathbf{Y}^\top$ where $\mathbf{X}$ and $\mathbf{Y}$ are n by k matrices, and then alternate between updating the two matrices. Ideally, the recovery error should be bounded by $\|\mathbf{W}\odot\mathbf{N}\|_2$, since this allows selecting weights according to the noise to make the error bound small.

As mentioned before, the problem is NP-hard in general, so we will need to impose some conditions. We summarize our assumptions as follows, and then discuss their necessity and the connections to existing ones.

- $(\mathbf{A1})$ *Ground truth is incoherent*: $\mathbf{M}^*$ has SVD $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$, where $\max_{i=1}^n\{||\mathbf{U}^i||_2^2, ||\mathbf{V}^i||_2^2\} \le \frac{\mu k}{n}$. Additionally, assume $\sigma_{\max}(\mathbf{\Sigma}) = \Theta(1)$. (See discussion below.) Denote its condition number as $\tau = \sigma_{\max}(\mathbf{\Sigma})/\sigma_{\min}(\mathbf{\Sigma})$.
- $(\mathbf{A2})$ *Weight matrix has a spectral gap*: $||\mathbf{W} - \mathbf{E}||_2 \le \gamma n$, where $\gamma < 1$ and $\mathbf{E}$ is the all-one matrix.
- $(\mathbf{A3})$ *Weight is not degenerate*: Let $\mathbf{D}_i = \mathsf{Diag}(\mathbf{W}^i)$, i.e., $\mathbf{D}_i$ is a diagonal matrix whose diagonal entries are the i-th row of $\mathbf{W}$. Then there are $0 < \underline{\lambda} \le 1 \le \overline{\lambda}$:

$$\underline{\lambda}\mathbf{I} \preceq \mathbf{U}^\top\mathbf{D}_i\mathbf{U} \preceq \overline{\lambda}\mathbf{I}, \text{and } \underline{\lambda}\mathbf{I} \preceq \mathbf{V}^\top\mathbf{D}_i\mathbf{V} \preceq \overline{\lambda}\mathbf{I}(\forall i \in [n]).$$

The incoherence assumption on the ground truth matrix is standard in the context of matrix completion. It is known that this is necessarily required for recovering the ground truth matrix. The assumption that $\sigma_{\max}(\mathbf{\Sigma}) = \Theta(1)$ is without loss of generality: one can estimate $\sigma_{\max}(\mathbf{\Sigma})$ up to a constant factor, scale the data and apply our results. The full details are included in the appendix.

The spectrum assumption on the weight matrix is a natural generalization of the randomness assumption typically made in matrix completion scenario (e.g., (Candes & Plan, 2010; Jain et al., 2013; Hardt, 2014)). In that case, $\mathbf{W}$ is a matrix with $d = \Omega(\log n)$-nonzeros in each row chosen uniformly at random, which corresponds to $\gamma = O\left(\frac{1}{\sqrt{d}}\right)$ in $(\mathbf{A2})$. Our assumption is also a generalization of the one in (Bhojanapalli & Jain, 2014), which requires $\mathbf{W}$ to be d-regular expander-like (i.e., to have a spectral gap) but is concerned only with matrix completion where the entries of $\mathbf{W}$ can be 0 or 1 only.

The final assumption $(\mathbf{A3})$ is a generalization of the assumption A2 in (Bhojanapalli & Jain, 2014) that, intuitively, requires the singular vectors to satisfy RIP (restricted isometry property). This is because when the weights are binary, $\mathbf{U}^\top \mathbf{D}_i \mathbf{U} = \sum_{j \in S} (\mathbf{U}^k)(\mathbf{U}^k)^\top$ where S is the support of $\mathbf{W}^i$, so after proper scaling the assumption is a strict weakening of theirs. They viewed it as a stronger version of incoherence, discussed the necessity and showed that it is implied by the strong incoherence property assumed in (Candès & Tao, 2010). In the context of more general weights, the necessity of $(\mathbf{A3})$ is even more clear, as elaborated below.

Note that since $(\mathbf{A2})$ does not require $\mathbf{W}$ to be random or d-regular, it does not a-priori exclude the degenerate case that $\mathbf{W}$ has one all-zero column. In that case, clearly one cannot hope to recover the corresponding column of $\mathbf{M}^*$. So, we need to make a third, non-degeneracy assumption about $\mathbf{W}$, saying that it is "correlated" with $\mathbf{M}^*$. The assumption is actually quite weak in the sense that when $\mathbf{W}$ is chosen uniformly at random, this assumption is true automatically: in those cases, $\mathbb{E}[\mathbf{D}_i] = \mathbf{I}$ and thus $\mathbb{E}[\mathbf{U}^\top \mathbf{D}_i \mathbf{U}] = \mathbf{I}$ since $\mathbf{U}$ is orthogonal. A standard matrix concentration bound can then show that our assumption $(\mathbf{A3})$ holds with high probability. Therefore, it is only needed when considering a deterministic $\mathbf{W}$. Intuitively, this means that the weights should cover the singular vectors of $\mathbf{M}^*$. This prevents the aforementioned degenerate case when $\mathbf{W}_i = 0$ for some i, and also some other degenerate cases. For example, consider the case when $\mathbf{N} = 0$, all rows of $\mathbf{M}^*$ are the same vector with first $\Theta(\log n)$ entries being zero and the rest being one, and in one row of $\mathbf{M}^*$ the non-zeros entries all have zero weight. In this case, there is also no hope to recover $\mathbf{M}^*$, which should be excluded by our assumption.

4. Algorithm and results

We prove guarantees for the vanilla alternating minimization with a simple clipping step, from either SVD initialization or random initialization. The algorithm is specified in Algorithm 1. Overall, it follows the usual alternating minimization framework: it keeps two working matrices $\mathbf{X}$ and $\mathbf{Y}$, and alternates between updating them. In an $\mathbf{X}$ update step, it first updates $\mathbf{X}$ to be the minimizer of the weighted low rank objective while fixing $\mathbf{Y}$, which can be done efficiently since now the optimization is convex. Then it performs a "clipping" step which zeros out rows of the matrix with too large norm,[1] and then make it orthogonal by QR-factorization.[2] At the end, the algorithm computes a final solution $\widetilde{\mathbf{M}}$ from the two iterates.

The two iterates can be initialized by performing SVD on the weighted observation (Algorithm 3), which is a weighted version of SVD initialization typically used in matrix completion. Moreover, we show that the algorithm works with random initialization (Algorithm 4), which is a simple and widely used heuristic in practice but rarely understood well.

We are now ready to state our main results. Theorem 1

Algorithm 1 Main Algorithm (ALT)

Input: Noisy observation $\mathbf{M}$, weight matrix $\mathbf{W}$, number of iterations T

Initialize $\mathbf{Y}_1$ using either $\mathbf{Y}_1 = \text{SVDINITIAL}(\mathbf{M}, \mathbf{W})$ or $\mathbf{Y}_1 = \text{RANDINITIAL}$
for $t = 1, 2, ..., T$ **do**
  $\widetilde{\mathbf{X}}_{t+1} = \operatorname{argmin}_{\mathbf{X} \in \mathbb{R}^{n \times k}} \left\| \mathbf{M} - \mathbf{X}\mathbf{Y}_t^\top \right\|_{\mathbf{W}}$
  $\overline{\mathbf{X}}_{t+1} = \text{CLIP}(\widetilde{\mathbf{X}}_{t+1})$
  $\mathbf{X}_{t+1} = \text{QR}(\overline{\mathbf{X}}_{t+1})$
  $\widetilde{\mathbf{Y}}_{t+1} = \operatorname{argmin}_{\mathbf{Y} \in \mathbb{R}^{n \times k}} \left\| \mathbf{M} - \mathbf{X}_{t+1}\mathbf{Y}^\top \right\|_{\mathbf{W}}$
  $\overline{\mathbf{Y}}_{t+1} = \text{CLIP}(\widetilde{\mathbf{Y}}_{t+1})$
  $\mathbf{Y}_{t+1} = \text{QR}(\overline{\mathbf{Y}}_{t+1})$
end for

Output: $\widetilde{\mathbf{M}} = \overline{\mathbf{X}}_{T+1}\mathbf{Y}_T$

Algorithm 2 Clipping (CLIP)

Input: matrix $\widetilde{\mathbf{X}}$

Output: matrix $\overline{\mathbf{X}}$ with

$$\overline{\mathbf{X}}^i = \begin{cases} \widetilde{\mathbf{X}}^i & \text{if } \|\widetilde{\mathbf{X}}^i\|_2^2 \le \xi := \frac{2\mu k}{n} \\ 0 & \text{otherwise.} \end{cases}$$

[1] The clipping step zeros out rows with square ℓ_2 norm twice larger than the upper bound $\mu k/n$ imposed by our incoherence assumption $(\mathbf{A1})$. One can choose the threshold to be $c\mu k/n$ where $c \ge 2$ is a constant and can choose to shrink the row to have norm no greater than $\mu k/n$, and our analysis still holds. The current choices are only for ease of presentation.

[2] The QR-factorization step is not necessary for our analysis. But since it is widely used in practice for numerical stability, we prefer to analyze the algorithm with QR.

Algorithm 3 SVD Initialization (SVDINITIAL)

Input: observation $\mathbf{M}$, weight $\mathbf{W}$

1: $(\widetilde{\mathbf{X}}, \widetilde{\mathbf{\Sigma}}, \widetilde{\mathbf{Y}}) =$ rank-k SVD$(\mathbf{W} \odot \mathbf{M})$, i.e., the columns of $\widetilde{\mathbf{Y}}$ are the top k right singular vectors of $\mathbf{W} \odot \mathbf{M}$

2: $\overline{\mathbf{Y}} =$ CLIP$(\widetilde{\mathbf{Y}})$, $\mathbf{Y} =$ QR$(\overline{\mathbf{Y}})$

Output: $\mathbf{Y}$

Algorithm 4 Random Initialization (RANDINITIAL)

1: Let $\mathbf{Y} \in \mathbb{R}^{n \times k}$ generated as $\mathbf{Y}_{i,j} = b_{i,j} \frac{1}{\sqrt{n}}$, where $b_{i,j}$'s are independent uniform from $\{-1, 1\}$

Output: $\mathbf{Y}$

describes our guarantee for the algorithm with SVD initialization, and Theorem 3 is for random initialization.

Theorem 1 (Main, SVD initialization)**.** *Suppose* $\mathbf{M}^*, \mathbf{W}$ *satisfy assumptions* ***(A1)-(A3)*** *with*

$$\gamma = O\left(\min\left\{\sqrt{\frac{n}{D_1}}\frac{\underline{\lambda}}{\tau\mu^{3/2}k^2}, \frac{\underline{\lambda}}{\tau^{3/2}\mu k^2}\right\}\right),$$

where $D_1 = \max_{i\in[n]} \|\mathbf{W}^i\|_1$*. Then after* $O(\log(1/\epsilon))$ *rounds of Algorithm 1 with initialization from Algorithm 3 outputs a matrix* $\widetilde{\mathbf{M}}$ *that satisfies*

$$\|\widetilde{\mathbf{M}} - \mathbf{M}^*\|_2 \le O\left(\frac{k\tau}{\underline{\lambda}}\right)\|\mathbf{W} \odot \mathbf{N}\|_2 + \epsilon.$$

The running time is polynomial in n *and* $\log(1/\epsilon)$*.*

The theorem is stated in its full generality. To emphasize the dependence on the matrix size n, the rank k and the incoherence μ, we can consider a specific range of parameter values where the other parameters (the spectral bounds, condition number, D_1/n) are constants. Also, these parameter values are typical in matrix completion, which facilitates our comparison in the next subsection.

Corollary 2. *Suppose* $\underline{\lambda}, \overline{\lambda}$ *and* τ *are all constants,* $D_1 = \Theta(n)$*, and* $T = O(\log(1/\epsilon))$*. Furthermore,*

$$\gamma = O\left(\frac{1}{\mu^{3/2}k^2}\right).$$

Then Algorithm 1 with initialization from Algorithm 3 outputs a matrix $\widetilde{\mathbf{M}}$ *that satisfies*

$$\|\widetilde{\mathbf{M}} - \mathbf{M}^*\|_2 \le O(k)\|\mathbf{W} \odot \mathbf{N}\|_2 + \epsilon.$$

Remarks The theorem bounds the spectral norm of the error matrix by the spectral norm of the weighted noise plus an additive error term that decreases exponentially with the number of rounds of alternating minimization. We emphasize that our guarantee holds for any $\mathbf{M}^*, \mathbf{W}$ satisfying our deterministic assumptions; the high success probability is with respect to the execution of the algorithm, not to the input. This ensures the freedom in choosing the weights to battle the noise. We also emphasize that the bounds hold *without any assumption* on the noise, which is particularly important here since weighted low rank is typically applied to complicated noise models.

Bounding the error by $\|\mathbf{W} \odot \mathbf{N}\|_2$ is particularly useful when the noise is not uniform across the entries: prior knowledge about the noise (e.g., the different variances of noise on different entries) can be taken into account by setting up a reasonable weight matrix[3], such that $\|\mathbf{W} \odot \mathbf{N}\|_2$ can be significantly smaller than $\|\mathbf{N}\|_2$. Also, in recovering the ground truth, a spectral norm bound is more preferred than a Frobenius norm bound, since typically the Frobenius norm is $\sqrt{n}$ larger than the spectral norm.

Furthermore, when $\|\mathbf{W} \odot \mathbf{N}\|_2 = 0$ (as in matrix completion without noise), the ground truth is recovered in a geometric rate.

Finally, in matrix completion with uniform random sampled observations, the term D_1 concentrates around n, so $\frac{D_1}{n}$ disappears in this case.

Theorem 3 (Main, random initialization)**.** *Suppose* $\mathbf{M}^*, \mathbf{W}$ *satisfy assumptions* ***(A1)-(A3)*** *with*

$$\gamma = O\left(\min\left\{\sqrt{\frac{n}{D_1}}\frac{\underline{\lambda}}{\tau\mu^2k^{5/2}}, \frac{\underline{\lambda}}{\tau^{3/2}\mu^{3/2}k^{5/2}}\right\}\right),$$

$$\|\mathbf{W}\|_\infty = O\left(\frac{\underline{\lambda}n}{k^2\mu\log^2 n}\right),$$

where $D_1 = \max_{i\in[n]} \|\mathbf{W}^i\|_1$*. Then after* $O(\log(1/\epsilon))$ *rounds Algorithm 1 with initialization from Algorithm 3 outputs a matrix* $\widetilde{\mathbf{M}}$ *that with probability at least* $1 - \frac{1}{n^2}$ *satisfies*

$$\|\widetilde{\mathbf{M}} - \mathbf{M}^*\|_2 \le O\left(\frac{k\tau}{\underline{\lambda}}\right)\|\mathbf{W} \odot \mathbf{N}\|_2 + \epsilon.$$

The running time is polynomial in n *and* $\log(1/\epsilon)$*.*

Remarks Compared to SVD initialization, we need slightly stronger assumptions for random initialization to work. There is an extra $1/(\mu^{1/2}k^{1/2})$ in the requirement of the spectral parameter γ. We note that the same error bound is obtained when using random initialization. Roughly speaking, this is because our analysis shows that the updates can make improvement under rather weak requirements that random initialization can satisfy, and after the first step the rest updates make the same progress as in the case using SVD initialization.

[3]Note that $\mathbf{W}$ cannot be made arbitrarily small since it should satisfy our assumptions. Roughly speaking, $\mathbf{W}$ has spectral norm n and is flexible to take into account the prior knowledge about the noise. In particular, it can be set to the all one matrix, reducing to the unweighted case.

	weight values	determin. weights	tolerate noise	alter. min.	order of γ (spectral gap)	Bound on $\Delta = \widetilde{\mathbf{M}} - \mathbf{M}^*$
(1)	0-1	no	yes	no	$\frac{1}{\mu k^{1/2}\mathsf{poly}(\log n)}$	$\|\Delta\|_F = O(\sqrt{\frac{n^3}{m}}\|\mathbf{N}_\Omega\|_F)$
(2)	0-1	no	yes	no	$\sqrt{\frac{1}{\mu k \log n}}$	$\|\Delta\|_F = O(\frac{n^2\sqrt{k}}{m}\|\mathbf{N}_\Omega\|_2)$
(3)	0-1	yes	no	no	$\frac{1}{\mu k}$	exact recovery
(4)	0-1	no	yes	yes	$\frac{1}{k\epsilon\sqrt{\mu \log n}}$	$\|\Delta\|_F \leq \epsilon\|\mathbf{M}^* + \mathbf{N}\|_F$
(5)	0-1	no	no	yes	$\frac{1}{\max\{\sqrt{k\mu \log n}, \mu k^{3.5}\}}$	exact recovery
ours (SVD init)	real	yes	yes	yes	$\frac{1}{\mu^{3/2}k^2}$	$\|\Delta\|_2 = O(k)\|\mathbf{W} \odot \mathbf{N}\|_2 + \epsilon$
ours (random init)	real	yes	yes	yes	$\frac{1}{\mu^2 k^{5/2}}$	$\|\Delta\|_2 = O(k)\|\mathbf{W} \odot \mathbf{N}\|_2 + \epsilon$

Table 1. Comparison with related work on matrix completion: (1) (Candes & Plan, 2010); (2) (Keshavan et al., 2009); (3) (Bhojanapalli & Jain, 2014); (4) (Hardt, 2014). (5) (Sun & Luo, 2015). Technical details are ignored. Especially, parameters other than the matrix size n, the rank k and the incoherence μ are regarded as constants.

4.1. Comparison with prior work

For the sake of completeness, we will give a more detailed comparison with representative prior work on matrix completion from Section 2, emphasizing the dependence on n, k and μ and regarding the other parameters as constants. We first note that when the m observed entries are sampled uniformly at random from an n by n matrix, the corresponding binary weight matrix will have a spectral gap $\gamma = O(\sqrt{\frac{n}{m}})$ (see, e.g., (Feige & Ofek, 2005)). Converting the sample bounds in the prior work to the spectral gap, we see that in general our result has worse dependence on parameters like the rank than those by convex relaxations, but has slightly better dependence than those by alternating minimization. The comparison is summarized in Table 1.

The seminal paper (Candès & Recht, 2009) showed that a nuclear norm convex relaxation approach can recover the ground truth matrix using $m = O(n^{1.2}k\log^2 n)$ entries chosen uniformly at random and without noise. The sample size was improved to $O(nk\log^6 n)$ in (Candès & Tao, 2010) and then $O(nk\log n)$ in subsequent papers. Candes & Plan (2010) generalized the result to the case with noise: the same convex program using $m = O(nk\log^6 n)$ entries recovers a matrix $\widetilde{\mathbf{M}}$ s.t. $\|\widetilde{\mathbf{M}} - \mathbf{M}^*\|_F \leq (2 + 4\sqrt{(2+p)n/p})\|\mathbf{N}_\Omega\|_F$ where $p = m/n^2$ and $\mathbf{N}_\Omega$ is the noise projected on the observed entries.

Keshavan et al. (2009) showed that with $m = O(n\mu k\log n)$, one can recover a matrix $\widetilde{\mathbf{M}}$ such that $\left\|\mathbf{M}^* - \widetilde{\mathbf{M}}\right\|_F = O\left(\frac{n^2\sqrt{k}}{m}\|\mathbf{N}_\Omega\|_2\right)$ by an optimization over a Grassmanian manifold.

Bhojanapalli & Jain (2014) relaxed the assumption that the entries are randomly sampled. They showed that the nuclear norm relaxation recovers the ground truth, assuming that the support Ω of the observed matrix forms a d-regular expander graph (or alike), i.e., $|\Omega| = dn$, $\sigma_1(\Omega) = d$ and $\sigma_2(\Omega) \leq c\sqrt{d}$ and $d \geq c^2\mu^2k^2$. This would correspond to a parameter $\gamma = O(\frac{1}{\mu k})$ for us. They did not consider the robustness to noise.

Hardt (2014) showed that with an appropriate initialization alternating minimization recovers the ground truth approximately. Precisely, they assumed $\mathbf{N}$ satisfies: (1). $\mu(\mathbf{N}) \lesssim \sigma_{\min}(\mathbf{M}^*)^2$;(2). $\|\mathbf{N}\|_\infty \leq \frac{\mu}{n}\|\mathbf{M}^*\|_F$. Then, he shows that $\log(\frac{n}{\epsilon}\log n)$ alternating minimization steps recover a matrix $\widetilde{\mathbf{M}}$ such that $\|\widetilde{\mathbf{M}} - \mathbf{M}^*\|_F \leq \epsilon\|\mathbf{M}\|_F$ provided that $pn \geq k(k + \log(n/\epsilon))\mu \times \left(\frac{\|\mathbf{M}^*\|_F + \|\mathbf{N}\|_F/\epsilon}{\sigma_k}\right)^2\left(1 - \frac{\sigma_{k+1}}{\sigma_k}\right)^5$ where σ_k is the k-th singular value of the ground-truth matrix. The parameter γ corresponding to the case considered there would be roughly $O(\frac{1}{k\sqrt{\mu \log n}})$. While their algorithm has a good tolerance to noise, $\mathbf{N}$ is assumed to have special structure for him that we do not assume in our setting.

Sun & Luo (2015) proved recovery guarantees for a family of algorithms including alternating minimization on matrix completion. They showed that by using $m = O(nk\max\{\mu\log n, \mu^2k^6\})$ randomly sampled entries without noise, the ground truth can be recovered in a geometric rate. This corresponds to a spectral gap of $O\left(\frac{1}{\max\{\sqrt{k\mu \log n}, \mu k^{3.5}\}}\right)$. Our result is more general and also handles noise. When specialized to their setting, we also have a geometric rate with a slightly better dependence on the rank k but a slightly worse dependence on the incoherence μ.

5. Proof sketch

Before going into our analysis, we first discuss whether arguments in prior work can be applied. Most of the work on matrix completion uses convex optimization and thus their analysis is not applicable in our setting. There indeed exists some other work that analyzes non-convex optimization for matrix completion, and it is tempting to adopt their

arguments. However, there exist fundamental difficulties in porting their arguments. All of them crucially rely on the randomness in sampling the observed entries. Keshavan et al. (2009) analyzed optimization over a Grassmanian manifold, which uses the fact that $\mathbb{E}[\mathbf{W} \odot \mathbf{S}] = \mathbf{S}$ for any matrix $\mathbf{S}$. In (Jain et al., 2013; Hardt, 2014), re-sampling of new observed entries in different iterations was used to get around the dependency of the iterates on the sample set, a common difficulty in analyzing alternating minimization. The subtlety and the drawback of re-sampling were discussed in detail in (Bhojanapalli & Jain, 2014; Candes et al., 2015; Sun & Luo, 2015). We note that (Sun & Luo, 2015) only needs sampling before the algorithm starts and does not need re-sampling in different iterations, but still relies on the randomness in the sampled entries. In particular, in all the aforementioned work, the randomness guarantees that the iterates $\mathbf{X}, \mathbf{Y}$ stay incoherent and have good spectrum properties. Given these, alternating minimization can make progress towards the ground truth in each iteration. Nevertheless, since we focus on deterministic weights, such randomness is inherently infeasible in our setting. In this case, after just one iteration, it is unclear if the iterates can have incoherence and good spectrum properties required to progress towards the ground truth, even under our current assumptions. The whole algorithm thus breaks down. To address this, we show that it is sufficient to ensure the spectral property in an average sense and then introduce our clipping step to achieve that, arriving at our current algorithm.

Here for simplicity, we drop the subscription t in all iterates, and we only focus on important factors, dropping other factors and the big-O notation. We only consider the case when $\mathbf{W} \odot \mathbf{N} = 0$, so as to emphasize the main technical challenges.

On a high level, our analysis of the algorithm maintains potential functions $\mathsf{dist}_c(\mathbf{X}, \mathbf{U})$ and $\mathsf{dist}_c(\mathbf{Y}, \mathbf{V})$ between our working matrices $\mathbf{X}, \mathbf{Y}$ and the ground truth $\mathbf{U}, \mathbf{V}$ (recall that $\mathbf{M}^* = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$):

$$\mathsf{dist}_c(\mathbf{X}, \mathbf{U}) = \min_{\mathbf{Q} \in \mathsf{O}_{k\times k}} \|\mathbf{X}\mathbf{Q} - \mathbf{U}\|_2$$

and

$$\mathsf{dist}_c(\mathbf{Y}, \mathbf{V}) = \min_{\mathbf{Q} \in \mathsf{O}_{k\times k}} \|\mathbf{Y}\mathbf{Q} - \mathbf{V}\|_2,$$

where $\mathsf{O}_{k\times k}$ are the set of $k \times k$ rotation matrices. The key is to show that they decrease after each update step, so $\mathbf{X}$ and $\mathbf{Y}$ get closer to the ground truth.[4] The strategy of maintaining certain potential function measuring the distance between the iterates and the ground truth is also used in prior work (Bhojanapalli & Jain, 2014; Candes et al., 2015; Sun & Luo, 2015). We will point out below the key technical difficulties that are not encountered in prior work and make our analysis substantially different. The complete proofs are provided in the appendix due to space limitation.

[4]Note that we also need a good initialization, which can be done by SVD. Since our analysis requires rather weak warm start, we are able to show that simple random initialization is also sufficient (at the cost of slightly worse bounds).

5.1. Update

We would like to show that after an $\mathbf{X}$ update, the new matrix $\widetilde{\mathbf{X}}$ satisfies $\mathsf{dist}_c(\mathbf{X}, \mathbf{U}) \le \mathsf{dist}_c(\mathbf{Y}, \mathbf{V})/2 + c$ for some small c (similarly for a $\mathbf{Y}$ update).

Consider the update step

$$\widetilde{\mathbf{X}} \leftarrow \mathrm{argmin}_{\mathbf{A} \in \mathbb{R}^{n\times k}} \left\|\mathbf{M} - \mathbf{A}\mathbf{Y}^\top\right\|_{\mathbf{W}}.$$

By setting the gradient to 0 and with some algebraic manipulation, we have $\widetilde{\mathbf{X}} - \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top\mathbf{Y} = \mathbf{G}$ where

$$\mathbf{G}^i := \mathbf{U}^i\mathbf{\Sigma}\mathbf{V}^\top\mathbf{Y}_\perp\mathbf{Y}_\perp^\top\mathbf{D}_i\mathbf{Y}(\mathbf{Y}^\top\mathbf{D}_i\mathbf{Y})^{-1}.$$

where $\mathbf{D}_i = \mathsf{Diag}(\mathbf{W}^i)$. Since $\widetilde{\mathbf{X}}$ is the value prior to performing QR decomposition, we want to show that $\widetilde{\mathbf{X}}$ is close to $\mathbf{U}^i\mathbf{\Sigma}\mathbf{V}^\top\mathbf{Y}$, i.e., the error term $\mathbf{G}$ on right hand side is small. In the ideal case when the error term is 0, then $\widetilde{\mathbf{X}} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top\mathbf{Y}$ and thus $\mathsf{dist}_c(\mathbf{X}, \mathbf{U}) = 0$, meaning that with one update $\widetilde{\mathbf{X}}$ already hits into the correct subspace. So we would like to show that it is small so that the iterate still makes progress. Let

$$\mathbf{P}_i = \mathbf{V}^\top\mathbf{Y}_\perp\mathbf{Y}_\perp^\top\mathbf{D}_i\mathbf{Y} \text{ and } \mathbf{O}_i = (\mathbf{Y}^\top\mathbf{D}_i\mathbf{Y})^{-1},$$

so that $\mathbf{G}^i = \mathbf{U}^i\mathbf{\Sigma}\mathbf{P}_i\mathbf{O}_i$. Now the two challenges are to bound $\mathbf{P}_i$ and $\mathbf{O}_i$.

Let us first consider the simpler case of matrix completion, where the entries of the matrix are randomly sampled by probability p. Then $\mathbf{D}_i$ is a random diagonal matrix with $\mathbb{E}[\mathbf{D}_i] = \mathbf{I}$ and $\mathbb{E}[\mathbf{D}_i^2] = \frac{1}{p}\mathbf{I}$. Furthermore, for $n \times k$ orthogonal matrices $\mathbf{Y}$, $\mathbf{O}_i = (\mathbf{Y}^\top\mathbf{D}_i\mathbf{Y})^{-1}$ concentrates around $\mathbf{I}$. Then in expectation, $||\mathbf{P}_i||$ is about $||\mathbf{V}^\top\mathbf{Y}_\perp||/\sqrt{p}$ and $\|\mathbf{O}_i\|$ is about 1, so $\|\mathbf{G}^i\|$ is as small as $\mu k||\mathbf{V}^\top\mathbf{Y}_\perp||/(\sqrt{p}n) = \mu k \sin\theta(\mathbf{V}, \mathbf{Y})/(\sqrt{p}n)$. High probability can then be established by the trick of re-sampling.

However, in our setting, we have to deal with two major technical obstacles due to deterministic weights.

1. There is no expectation for $\mathbf{D}_i$. Since $\|\mathbf{D}_i\|_\infty^2$ can be as large as $\frac{n^2}{\mathsf{poly}(\log n)}$, $\|\mathbf{P}_i\|$ can potentially be as large as $\sin\theta(\mathbf{Y}, \mathbf{V})\frac{n}{\mathsf{poly}(\log n)}$, which is almost a factor n larger than the bound for random $\mathbf{D}_i$. This is clearly insufficient to show the progress.
2. A priori the norm of $\mathbf{O}_i = (\mathbf{Y}^\top\mathbf{D}_i\mathbf{Y})^{-1}$ may be large. Especially, in the algorithm $\mathbf{Y}$ is given by the

alternating minimization steps and giving an upper bound on $\|\mathbf{O}_i\|$ at all steps seems hard.

The first issue For this, we exploit the incoherence of $\mathbf{Y}$ and the spectral property of the weight matrix. If $\mathbf{D}_i$ is the identity matrix, then $\mathbf{P}_i = 0$ which, intuitively, means that there are cancellations between negative part and positive parts. When $\mathbf{W}$ is expander-like, it will put roughly equal weights on the negative part and the positive part. If furthermore we have that $\mathbf{Y}$ is incoherent (i.e., the negative and positive parts are spread out), then $\mathbf{W}$ can mix the terms and lead to a cancellation similar to that when $\mathbf{D}_i = \mathbf{I}$. More precisely, consider the (j, j')-th element in $\mathbf{P}_i$. Define a new vector $x \in \mathbb{R}^n$ such that

$$x_i = (\widetilde{\mathbf{V}}_j)_i (\mathbf{Y}_{j'})_i, \text{ where } \widetilde{\mathbf{V}} = \mathbf{V}^\top \mathbf{Y}_\perp \mathbf{Y}_\perp^\top.$$

Then we have the cancellation in the form of $\sum_i x_i = 0$. When $\mathbf{D}_i = \mathbf{I}$, we simply get $(\mathbf{P}_i)_{j,j'} = \sum_i x_i = 0$. When $\mathbf{D}_i \neq \mathbf{I}$, we have $(\mathbf{P}_i)_{j,j'} = \sum_{s\in[n]} (\mathbf{D}_i)_s x_s^{j,j'}$. Now mix over all i, we have

$$\begin{aligned}\sum_{i\in[n]} ((\mathbf{P}_i)_{j,j'})^2 &= \left(\sum_{s\in[n]} (\mathbf{D}_i)_s x_s\right)^2 = \|\mathbf{W}x\|^2 \\ &= \|(\mathbf{W}-\mathbf{E})x\|^2 \qquad (\text{since } \mathbf{E}x = 0) \\ &\leq \gamma^2 n^2 \|x\|^2\end{aligned}$$

where in the last step we use the expander-like property of $\mathbf{W}$ (Assumption $(\mathbf{A2})$) to gain the cancellation. Furthermore, if $\|\mathbf{Y}_{j'}\|_\infty$ is small, by definition $\|x\|^2$ is also small, so we can get an upper bound on $\sum_{i\in[n]} \|\mathbf{P}_i\|_F^2$.

Then the problem reduces to maintaining the incoherence of $\mathbf{Y}$. This is taken care of by our clipping step (Algorithm 2), which sets to 0 the rows of $\mathbf{Y}$ that are too large. Of course, we have to show that this will not increase the distance of the clipped $\mathbf{Y}$ and $\mathbf{V}$. The intuition is that we clip only when $\|\mathbf{Y}^i\| \geq 2\mu k/n$. But $\|\mathbf{V}^i\| \leq \mu k/n$, so after clipping, $\mathbf{Y}^i$ only gets closer to $\mathbf{V}^i$.

The second issue This is the more difficult technical obstacle, i.e., $\|\mathbf{O}_i\| = \|(\mathbf{Y}^\top \mathbf{D}_i \mathbf{Y})^{-1}\|$ can be large. Our key idea is that although individual $\|\mathbf{O}_i\|$ can indeed be large, this cannot be the case on average. We show that there can just be a few i's such that $\|\mathbf{O}_i\|$ is large, and they will not contribute much to $\|\mathbf{G}\|$, so the update can make progress.

To be more formal, we wish to bound the number of indices i such that $\sigma_{\min}\left(\mathbf{Y}^\top \mathbf{D}_i \mathbf{Y}\right) \leq \frac{\lambda}{4}$. Consider an arbitrary unit vector a. Then,

$$a\mathbf{Y}^\top \mathbf{D}_i \mathbf{Y} a = \sum_j a\mathbf{Y}^\top (\mathbf{D}_i)_j \mathbf{Y} a = \sum_j (\mathbf{D}_i)_j \langle a, \mathbf{Y}^j\rangle^2.$$

We know that $\mathbf{Y}$ is close to $\mathbf{V}$, so we rewrite the above using some algebraic manipulation as

$$\begin{aligned}&\sum_j (\mathbf{D}_i)_j \langle a, (\mathbf{Y}^j - \mathbf{V}^j) + \mathbf{V}^j\rangle^2 \\ \geq\ & \frac{1}{4}\sum_j (\mathbf{D}_i)_j \langle a, \mathbf{V}^j\rangle^2 - \frac{1}{3}\sum_j (\mathbf{D}_i)_j \langle a, \mathbf{Y}^j - \mathbf{V}^j\rangle^2\end{aligned}$$

For j's such that $\mathbf{Y}^j$ is close to $\mathbf{V}^j$ (denote these j's as $\mathcal{S}_g$), then the terms can be easily bounded since $\mathbf{V}^\top \mathbf{D}_i \mathbf{V} \geq \underline{\lambda} I$ by assumption. So we only need to consider j's such that $\mathbf{Y}^j$ is far from $\mathbf{V}^j$. Since we have incoherence, we know that $\|\mathbf{Y}^j - \mathbf{V}^j\|$ is still bounded in the order of $\mu k/n$. So $a\mathbf{Y}^\top \mathbf{D}_i \mathbf{Y} a$ can be small only when $\sum_{j\notin\mathcal{S}_g} (\mathbf{D}_i)_j$ is large.

Let $\mathcal{S}$ denote those bad i's. Let $u_{\mathcal{S}}$ be the indicator vector for $\mathcal{S}$ and u_g be the indicator vector for $[n - \mathcal{S}_g]$.

$$\begin{aligned}\sum_{i\in\mathcal{S}}\sum_{j\notin\mathcal{S}_g} (\mathbf{D}_i)_j &= u_{\mathcal{S}}^\top \mathbf{W} u_g \\ &\leq |\mathcal{S}|(n - |\mathcal{S}_g|) + \gamma n\sqrt{|\mathcal{S}|(n-|\mathcal{S}_g|)}\end{aligned}$$

where the last step is due to the spectral property of $\mathbf{W}$. Therefore, there can be only a few i's with large $\sum_{j\notin\mathcal{S}_g} (\mathbf{D}_i)_j$.

5.2. Proofs of main results

We only need to show that we can get an initialization close enough to the ground truth so that we can apply the above analysis for the update. For SVD initialization,

$$[\mathbf{X}, \mathbf{\Sigma}, \mathbf{Y}] = \text{rank-}k\ \text{SVD}(\mathbf{W}\odot\mathbf{M}^* + \mathbf{W}\odot\mathbf{N}).$$

Since $||\mathbf{W}\odot\mathbf{N}||_2 \leq \delta$ can be regarded as small, the idea is to show that $\mathbf{W}\odot\mathbf{M}^*$ is close to $\mathbf{M}^*$ in spectral norm and then apply Wedin's theorem (Wedin, 1972). We show this by the spectral gap property of $\mathbf{W}$ and the incoherence property of $\mathbf{U}, \mathbf{V}$.

For random initialization, the proof is only a slight modification of that for SVD initialization, because the update requires rather mild conditions on the initialization such that even the random initialization is sufficient (with a slightly worse parameters).

6. Conclusion

In this paper we presented the first recovery guarantee of weighted low-rank matrix approximation via alternating minimization. Our work generalized prior work on matrix completion, and revealed technical obstacles in analyzing alternating minimization, i.e., the incoherence and spectral properties of the intermediate iterates need to be preserved. We addressed the obstacles by a simple clipping step, which resulted in a very simple algorithm that almost matches the practical heuristics.

Acknowledgements

This work was supported in part by NSF grants CCF-1527371, DMS-1317308, Simons Investigator Award, Simons Collaboration Grant, and ONR-N00014-16-1-2329.

References

Arora, Sanjeev, Li, Yuanzhi, Liang, Yingyu, Ma, Tengyu, and Risteski, Andrej. A latent variable model approach to pmi-based word embeddings. *To appear in Transactions of the Association for Computational Linguistics*, 2016.

Bhojanapalli, Srinadh and Jain, Prateek. Universal matrix completion. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pp. 1881–1889, 2014.

Bhojanapalli, Srinadh, Jain, Prateek, and Sanghavi, Sujay. Tighter low-rank approximation via sampling the leveraged element. In *Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 902–920. SIAM, 2015a.

Bhojanapalli, Srinadh, Kyrillidis, Anastasios, and Sanghavi, Sujay. Dropping convexity for faster semi-definite optimization. *arXiv preprint arXiv:1509.03917*, 2015b.

Buck, Christian, Heafield, Kenneth, and van Ooyen, Bas. N-gram counts and language models from the common crawl. In *Proceedings of the Language Resources and Evaluation Conference*, Reykjavk, Icelandik, Iceland, May 2014.

Candes, Emmanuel J and Plan, Yaniv. Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925–936, 2010.

Candès, Emmanuel J and Recht, Benjamin. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717–772, 2009.

Candès, Emmanuel J and Tao, Terence. The power of convex relaxation: Near-optimal matrix completion. *Information Theory, IEEE Transactions on*, 56(5):2053–2080, 2010.

Candes, Emmanuel J, Li, Xiaodong, and Soltanolkotabi, Mahdi. Phase retrieval via wirtinger flow: Theory and algorithms. *Information Theory, IEEE Transactions on*, 61(4):1985–2007, 2015.

Eriksson, Anders and van den Hengel, Anton. Efficient computation of robust weighted low-rank matrix approximations using the l_1 norm. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(9):1681–1690, 2012.

Feige, Uriel and Ofek, Eran. Spectral techniques applied to sparse random graphs. *Random Structures & Algorithms*, 27(2):251–275, 2005.

Gadian, David G. *Nuclear magnetic resonance and its applications to living systems*. Clarendon Press; Oxford University Press, 1982.

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

Gross, David. Recovering low-rank matrices from few coefficients in any basis. *Information Theory, IEEE Transactions on*, 57(3):1548–1566, 2011.

Hardt, Marcus. Understanding alternating minimization for matrix completion. In *Foundations of Computer Science (FOCS), 2014 IEEE 55th Annual Symposium on*, pp. 651–660. IEEE, 2014.

Heiman, Eyal, Schechtman, Gideon, and Shraibman, Adi. Deterministic algorithms for matrix completion. *Random Structures & Algorithms*, 45(2):306–317, 2014.

Jain, Prateek, Netrapalli, Praneeth, and Sanghavi, Sujay. Low-rank matrix completion using alternating minimization. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pp. 665–674. ACM, 2013.

Keshavan, Raghunandan, Montanari, Andrea, and Oh, Sewoong. Matrix completion from noisy entries. In *Advances in Neural Information Processing Systems*, pp. 952–960, 2009.

Lee, Troy and Shraibman, Adi. Matrix completion from any given set of observations. In *Advances in Neural Information Processing Systems*, pp. 1781–1787, 2013.

Levy, Omer and Goldberg, Yoav. Neural word embedding as implicit matrix factorization. In *Advances in Neural Information Processing Systems*, pp. 2177–2185, 2014.

Li, Yanen, Hu, Jia, Zhai, ChengXiang, and Chen, Ye. Improving one-class collaborative filtering by incorporating rich user information. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pp. 959–968. ACM, 2010.

Li, Yuanzhi, Liang, Yingyu, and Risteski, Andrej. Recovery guarantee of weighted low-rank approximation via alternating minimization. *CoRR*, abs/1602.02262, 2016. URL `http://arxiv.org/abs/1602.02262`.

Lu, W-S, Pei, S-C, and Wang, P-H. Weighted low-rank approximation of general complex matrices and its application in the design of 2-d digital filters. *Circuits and*

Systems I: Fundamental Theory and Applications, IEEE Transactions on, 44(7):650–655, 1997.

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

Peeters, René. Orthogonal representations over finite fields and the chromatic number of graphs. *Combinatorica*, 16 (3):417–431, 1996.

Pennington, Jeffrey, Socher, Richard, and Manning, Christopher D. Glove: Global vectors for word representation. *Proceedings of the Empiricial Methods in Natural Language Processing (EMNLP 2014)*, 12:1532–1543, 2014.

Razenshteyn, Ilya, Song, Zhao, and Woodruff, David. Weighted low rank approximations with provable guarantees. In *Proceedings of the 48th Annual Symposium on the Theory of Computing*, 2016.

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

Srebro, Nathan and Jaakkola, Tommi. Weighted low-rank approximations. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, pp. 720–727, 2003.

Sun, Ruoyu and Luo, Zhi-Quan. Guaranteed matrix completion via nonconvex factorization. In *IEEE 56th Annual Symposium on Foundations of Computer Science*, pp. 270–289, 2015.

Wedin, Per-Åke. Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics*, 12(1):99–111, 1972.

Wentzell, Peter D, Andrews, Darren T, and Kowalski, Bruce R. Maximum likelihood multivariate calibration. *Analytical chemistry*, 69(13):2299–2311, 1997.

Wikimedia. English Wikipedia dump. http://dumps.wikimedia.org/enwiki/latest/enwiki-latest-pages-articles.xml.bz2, 2012. Accessed Mar-2015.