Online Matrix Completion Through Nuclear Norm Regularisation

Charanpal Dhanjal, Romaric Gaudel, Stéphan Clémençon

ONLINE MATRIX COMPLETION THROUGH NUCLEAR NORM REGULARISATION

CHARANPAL DHANJAL AND STÉPHAN CLÉMENÇON

Institut Mines-Télécom; Télécom ParisTech: CNRS LTC, 46 rue Barrault, 75634 Paris Cedex 13, France

ROMARIC GAUDEL

Université Lille 3, Domaine Universitaire du Pont de Bois, 59653 Villeneuve d'Ascq Cedex, France

Abstract. It is the main goal of this paper to propose a novel method to perform matrix completion *on-line*. Motivated by a wide variety of applications, ranging from the design of recommender systems to sensor network localization through seismic data reconstruction, we consider the matrix completion problem when entries of the matrix of interest are observed *gradually*. Precisely, we place ourselves in the situation where the predictive rule should be refined incrementally, rather than recomputed from scratch each time the sample of observed entries increases. The extension of existing matrix completion methods to the sequential prediction context is indeed a major issue in the Big Data era, and yet little addressed in the literature. The algorithm promoted in this article builds upon the Soft Impute approach introduced in [17]. The major novelty essentially arises from the use of a randomised technique for both computing and updating the *Singular Value Decomposition (SVD)* involved in the algorithm. Though of disarming simplicity, the method proposed turns out to be very efficient, while requiring reduced computations. Several numerical experiments based on real datasets illustrating its performance are displayed, together with preliminary results giving it a theoretical basis.

1. Introduction

The task of finding unknown entries in a matrix given a sample of observed entries is known as the *matrix completion* problem, see [6]. Stated in such a general manner, this corresponds to a wide variety of problems, including collaborative filtering, dimensionality reduction, image processing or multi-class classification for instance. In particular, it has recently received much attention in the area of *recommender systems*, for e-commerce especially. In this context, users rate a selection of items (*e.g.* books, movies, news) and then an algorithm predicts future still unobserved ratings based on the observed ones. The ratings can be

E-mail addresses: {charanpal.dhanjal, stephan.clemencon}@telecom-paristech.fr, romaric.gaudel@univ-lille3.fr.

Date: January 9, 2014.

represented through a matrix $\mathbf{X} \in \mathbb{R}^{m\times n}$, where the rows correspond to the users and the columns to the items: the ij-th entry is the rating, valued in $\{1, 2, \ldots, 5\}$ say, by user $i \in \{1, \ldots, m\}$ of item $j \in \{1, \ldots, n\}$ and only a sample of the entries of $\mathbf{X}$ are observed. The goal is to predict the rating for an unobserved entry $\mathbf{X}_{ij}$. In general, matrix completion is an impossible task, however under certain assumptions on $\mathbf{X}$ (typically $\mathbf{X}$ has low rank and has a certain number of observed entries) [6, 4, 12], the remaining entries of this matrix can be accurately recovered.

A number of approaches have recently been proposed for matrix completion, the key idea of which is to solve the rank minimisation problem. As this optimisation is NP hard, one can use the *nuclear norm*, or equivalently the sum of the singular values, as a surrogate which leads to a tractable problem, see [6, 4, 12, 3, 5] for example. Most algorithms documented in the literature do not deal with common scenario of having an increasing amount of data, and yet developing methods capable to meet real-time constraints is of huge importance in the beginning Big Data era. Thus it is precisely this scenario which we address in this paper, known as *incremental* or *online learning*. In incremental matrix completion, given a sequence of incomplete matrices $\mathbf{X}_1, \ldots, \mathbf{X}_T$, possibly with different sizes, one wishes to complete each one without full recomputation at each iteration.

Here we build upon the work of the *Soft Impute* algorithm of [17] for nuclear norm regularised matrix completion. This work is simple to implement and analyse, scales to relatively large matrices and achieves competitive errors compared to state-of-the-art algorithms, however a bottleneck in the algorithm is the use of the Singular Value Decomposition (SVD, [8]) of a large matrix at each iteration. We use recent work on the theory and practice of randomised SVDs [9] along with a novel updating method to improve the efficiency of matrix completion both in offline and online cases. In the offline case, we show that the algorithm can be efficiently and accurately implemented using only a single SVD of the input matrix and thereafter inexpensive updates of the SVD can be applied. In the online case, we provide an efficient path to updating the solution matrices under possibly high-rank perturbations of the input matrices. The randomised SVD can introduce errors into the final solution, and we give some theoretical and empirical insight into this error. The resulting randomised online learning algorithm is simple to implement, and readily parallelisable to make full use of modern computer architectures. Computational results on a toy dataset and several large real movie recommendations datasets shows the efficacy of the approach.

In the following section, we present the main results in nuclear norm regularised matrix completion. Following on in Section 3 we introduce the online matrix completion algorithm and provide a preliminary theoretical analysis of it. Section 4 presents an empirical study of the resulting matrix completion approach and finally some concluding remarks are collected in Section 5.

2. Matrix Completion

Consider again the matrix $\mathbf{X} \in \mathbb{R}^{m\times n}$ and a set of indices of the observed entries $\omega \subset \{1, \ldots, m\} \times \{1, \ldots, n\}$. A useful optimisation to consider in terms of the applications outlined above is that of the minimum rank subject to bounded errors on the observed entries,

$$\begin{array}{ll} \min & \operatorname{rank}(\mathbf{Z}) \\ \text{s.t.} & \sum_{i,j\in\omega}(\mathbf{X}_{ij} - \mathbf{Z}_{ij})^2 \leq \delta, \end{array} \tag{1}$$

for user defined $\delta \geq 0$, and some $\mathbf{Z} \in \mathbb{R}^{m \times n}$, however this optimisation is NP hard. A number of ways to tackle this optimisation exist, for example using greedy selection [22, 14]. Another approach is to relax the rank term into a *trace* or nuclear norm, i.e. solve

$$\begin{array}{ll} \min & \|\mathbf{Z}\|_* \\ \text{s.t.} & \sum_{i,j\in\omega}(\mathbf{X}_{ij} - \mathbf{Z}_{ij})^2 \leq \delta, \end{array} \tag{2}$$

which is a convex problem. The nuclear norm $\|\cdot\|_*$ is the sum of the singular values of a matrix $\sum_{i=1}^{r}\sigma_i$, where r is the rank of the matrix. The above can be reformulated in Lagrange form as follows:

$$\min \frac{1}{2}\sum_{i,j\in\omega}(\mathbf{X}_{ij} - \mathbf{Z}_{ij})^2 + \lambda\|\mathbf{Z}\|_*, \tag{3}$$

where λ is a user-defined regularisation parameter. Another way of writing the above is in terms of the projection operator P_ω: $\min \frac{1}{2}\|P_\omega(\mathbf{X}) - P_\omega(\mathbf{Z})\|_F^2 + \lambda\|\mathbf{Z}\|_*$, where $P_\omega(\mathbf{X})$ is the matrix whose (i,j)th entry is $\mathbf{X}_{ij}$ if $i,j \in \omega$ otherwise it is 0, and $\|\mathbf{X}\|_F^2 = \sum_{ij}\mathbf{X}_{ij}^2$ is the Frobenius norm. Similarly, $P_\omega^\perp(\mathbf{X})$ is the matrix whose (i,j)th entry is $\mathbf{X}_{ij}$ if $i,j \notin \omega$ otherwise it is 0. In [4] it is shown that if $\mathbf{X}$ has entries sampled uniformly at random with $|\omega| \geq Cp^{5/4}r\log p$, for a positive constant C and with $p = \max(m,n)$, then one can recover all the entries of $\mathbf{X}$ with no error with high probability using trace norm minimisation.

In [18] the nuclear norm penalised objective is approximated by writing the penalty in terms of the minimum Frobenius norm factorisation, and solving it using a parallel projected incremental gradient method. The resulting problem can be considered as a generalisation of Maximal-Margin Matrix Factorisation (MMMF, [23]). Note that the optimisation of Equation (3) can be considered a generalisation of the ℓ_1 regularised least squares problem which is addressed in [24]. As with ℓ_1 versus ℓ_0 linear regression, minimising the nuclear norm can outperform the rank minimised solution, as supported empirically in [17].

Classical algorithms for solving semi-definite programs such as interior point methods are expensive for large datasets, and this has motivated a number of algorithms with better scaling. In [3] the authors propose a Singular Value Thresholding (SVT) algorithm for the optimisation of Equation (2) in which $\delta = 0$. In [16] the authors use a similar approach based on Bregman iteration, and [25] uses an accelerated proximal gradient algorithm which gives an ϵ-accurate solution in $\mathcal{O}(1/\sqrt{\epsilon})$ steps. In [10] a variant of Equation (2) is solved in which there is an upper bound on the nuclear norm. The authors transform the problem into a convex one on positive semi-definite matrices. A nuclear norm minimisation subject to linear and second order cone constraints is solved in [15], with an application to recommendation on a large movie rating dataset. The soft impute algorithm of [17] is inspired by SVT, however unlike [3, 16, 11] it does not require a step size parameter. Instead, soft impute is controlled using the regularisation parameter λ, and using warm restarts one can compute the complete regularisation path for model selection. The algorithm is shown to be competitive to SVT and MMMF and scalable to relatively large datasets.

2.1. **Soft Impute.** Our novel online matrix completion scheme is based on soft impute, and hence we present the pseudo code of soft impute in Algorithm 1. As input, it takes a partially observed matrix $\mathbf{X}$ and a sequence of regularisation

parameters $\lambda_1 > \lambda_2 > \ldots > \lambda_k$. The core part of the algorithm is the loop at Step 4 which updates the current solution $\mathbf{Z}_{j+1}$ and checks convergence with successive solutions.

At step 5 one computes the SVD of $P_\omega(\mathbf{X}) - P_\omega(\mathbf{Z})$, which is the most expensive part of this algorithm. The SVD of $\mathbf{A} \in \mathbb{R}^{m\times n}$ is the decomposition $\mathbf{A} = \mathbf{P\Sigma Q}^T$, where $\mathbf{P} = [\mathbf{p}_1, \ldots, \mathbf{p}_r]$, $\mathbf{Q} = [\mathbf{q}_1, \ldots, \mathbf{q}_r]$ are respective matrices whose columns are left and right singular vectors, and $\mathbf{\Sigma} = \text{diag}(\sigma_1, \ldots, \sigma_r)$ is a diagonal matrix of singular values $\sigma_1 \geq \sigma_2 \geq, \ldots, \geq \sigma_r$, with $r = \min(m, n)$. One then applies the *matrix shrinkage operator*, $S_\lambda(\mathbf{A}) = \mathbf{P\Sigma}_\lambda\mathbf{Q}^T$, in which $\mathbf{\Sigma}_\lambda = \text{diag}((\sigma_1 - \lambda)_+, \ldots, (\sigma_r - \lambda)_+)$ where $t_+ = \max(t, 0)$. The idea behind this step is to converge to the stationary point of the objective of Equation (3), which is the solution to $\mathbf{Z} = S_\lambda(P_\omega(\mathbf{X}) - P_\omega^\perp(\mathbf{Z}))$, the proof of which is given in [17]. An

Algorithm 1 Pseudo-code for Soft Impute

Require: Matrix $\mathbf{X}$, regularisation parameters $\lambda_1 > \lambda_2 > \ldots > \lambda_k$, error threshold ϵ

1: $\mathbf{Z}^{(1)} = 0$, $j = 1$
2: **for** $i = 1 \to k$ **do**
3: $\quad\gamma = \epsilon + 1$
4: $\quad$**while** $\gamma > \epsilon$ **do**
5: $\quad\quad\mathbf{Z}^{(j+1)} \leftarrow S_{\lambda_i}(P_\omega(\mathbf{X}) + P_\omega^\perp(\mathbf{Z}^{(j)}))$
6: $\quad\quad\gamma = \frac{\|\mathbf{Z}^{(j+1)} - \mathbf{Z}^{(j)}\|_F^2}{\|\mathbf{Z}^{(j)}\|_F^2}$
7: $\quad\quad j \leftarrow j + 1$
8: $\quad$**end while**
9: $\quad\mathbf{Z}_{\lambda_i} \leftarrow \mathbf{Z}^{(j)}$
10: **end for**
11: **return** Solutions $\mathbf{Z}_{\lambda_1}, \ldots, \mathbf{Z}_{\lambda_k}$

important point about soft impute is that it appears to generate a dense matrix in step 5, however the authors note that $\mathbf{Y} = P_\omega(\mathbf{X}) + P_\omega^\perp(\mathbf{Z})$ can be written as the sum of a sparse term $P_\omega(\mathbf{X}) - P_\omega(\mathbf{Z})$ and a low rank term $\mathbf{Z}$. Since the fundamental step in computing an SVD of $\mathbf{Y}$ is matrix-vector multiplications $\mathbf{Yu}$ and $\mathbf{Y}^T\mathbf{v}$, one can use this observation to improve the efficiency of the soft-thresholded SVD. First note that to compute $P_\omega(\mathbf{Z})$ requires $\mathcal{O}(|\omega|\tilde{r})$ operations using the SVD, in which $\mathbf{Z}$ has a rank $\tilde{r} \ll m, n$. Furthermore a matrix-vector multiplication of $\mathbf{Y}$ can be found in order $\mathcal{O}((m + n)\tilde{r} + |\omega|)$ operations, the sum of the low-rank and sparse matrix multiplications. If there are s singular vectors then the total computational cost is $\mathcal{O}((m + n)\tilde{r}s + |\omega|s)$. The final solution of $\mathbf{Z}$ has rank $r \approx \tilde{r}$ and given we want to find approximately r singular vectors this cost can be written as $\mathcal{O}((m + n)r^2t + |\omega|rt)$ assuming that t iterations are required.

3. Online Matrix Completion

One disadvantage of Algorithm 1 is that at each stage one must compute the SVD of a large matrix (in common with other SVT-based algorithms). Normally one uses a Krylov subspace method such Lanczos or Arnoldi (e.g. PROPACK [13]) to compute the rank-k SVD at cost $\mathcal{O}(kT_{mult} + (m + n)k^2)$ where T_{mult} is the cost of a matrix-vector multiplication. In a sparse matrix, T_{mult} is the number

of nonzero elements $|\omega|$ in the matrix. A second disadvantage is that one must compute successive SVDs of a matrix $P_\omega(\mathbf{X}) + P_\omega^\perp(\mathbf{Z})$ ignoring previous computations of this SVD. One attempt to address the former point is given in [28] which decomposes the input into a set of Kronecker products of several smaller matrices in conjunction with the algorithm of [17]. This leads to two convex subproblems on smaller matrices, however a drawback of the approach is that one must specify the size of the decomposition in advance and the best choice is unknown a priori.

Recently, randomised method for computing the SVD have been studied in the literature [9]. These approaches are competitive in terms of computational time with state-of-the-art Krylov methods, robust, well studied theoretically, and benefit from simple implementations and easy parallelisation. Whereas Lanczos and Arnoldi algorithms are numerically unstable, randomised algorithms are stable and come with performance guarantees that do not depend on the spectrum of the input matrix. The key idea of randomised algorithms is to project the rows onto a subspace which captures most of the "action" of the matrix. To illustrate the point, we recount an algorithm from [9] which is used in conjunction with kernel Principal Components Analysis (KPCA, [21]) in [26]. Algorithm 2 provides the associated pseudo-code. The purpose of the first three steps is to find a matrix $\mathbf{V} \in \mathbb{R}^{m\times(k+p)}$ such that the projection of $\mathbf{A}$ onto $\mathbf{V}$ is a good approximation of $\mathbf{Y}$. In other words, we hope to find $\mathbf{V}$ with orthogonal columns such that $\|\mathbf{A}-\mathbf{V}\mathbf{V}^T\mathbf{A}\|_2$ is small, where $\|\mathbf{A}\|_2 = \sigma_1$ is the spectral norm. The above norm is minimised when the columns of $\mathbf{V}$ are made up of the $(k+p)$ largest left singular vectors of $\mathbf{A}$. When $q = 0$, the matrix $\mathbf{Y}$ is one whose columns are random samples from the range of $\mathbf{A}$ under a rank-$(k+p)$ projection $\mathbf{\Omega}$. The columns of $\mathbf{\Omega}$ are likely to be linearly independent as are those of $\mathbf{Y}$ which span much of the range of $\mathbf{A}$ provided its range is not much larger than $(k+p)$. Hence, the resulting projection is orthogonalised to form $\mathbf{V}$ and then one need only find the SVD of the smaller matrix $\mathbf{B} = \mathbf{V}^T\mathbf{A}$. When $q > 0$ the quality of $\mathbf{V}$ is improved when the spectrum of the data matrix decays slowly, as is often the case in matrix completion. Note that to reduce rounding errors one othogonalises the projected matrix before each multiplication with $\mathbf{A}$ or $\mathbf{A}^T$. The complexity of the complete approach is $\mathcal{O}((q+1)(k+p)T_{mult}+(k+p)^2(m+n))$. This

Algorithm 2 Randomised SVD [9]

Require: Matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$, target rank k, oversampling projection vectors p, exponent q

1: Generate a random Gaussian matrix $\mathbf{\Omega} \in \mathbb{R}^{n\times(k+p)}$
2: Create $\mathbf{Y} = (\mathbf{A}\mathbf{A}^T)^q\mathbf{A}\mathbf{\Omega}$ by alternative multiplication with $\mathbf{A}$ and $\mathbf{A}^T$
3: Compute $\mathbf{Y} = \mathbf{V}\mathbf{R}$ using the QR-decomposition
4: Form $\mathbf{B} = \mathbf{V}^T\mathbf{A}$ and compute SVD $\mathbf{B} = \hat{\mathbf{P}}\mathbf{\Sigma}\mathbf{Q}^T$
5: Set $\mathbf{P} = \mathbf{V}\hat{\mathbf{P}}$
6: **return** Approximate SVD $\mathbf{A} \approx \mathbf{P}\mathbf{\Sigma}\mathbf{Q}^T$

then gives us the primary ingredients we require for online matrix completion. At each iteration of soft impute we use the randomised SVD of Algorithm 2 to compute an approximate SVD. For a sequence of matrices $\mathbf{X}_1, \ldots, \mathbf{X}_T$ with corresponding nonzero indices $\omega_1, \ldots, \omega_T$ let the ith solution with regularisation parameter λ be written $\mathbf{Z}_\lambda^{[i]}$. To compute a solution for $\mathbf{X}_{i+1}$ we use the decomposition computed for $\mathbf{Z}_\lambda^{[i]}$ as a seed for the first randomised SVD of soft impute. If $\mathbf{X}_i$ and $\mathbf{X}_{(i+1)}$

have different sizes then we can adjust the initial solution size accordingly, padding with zeros if $\mathbf{X}_{i+1}$ is larger than $\mathbf{X}_i$.

There are two important advantages of this algorithm over the use of traditional Krylov subspace methods in conjunction with soft impute: the first is that the above algorithm can be effective even when there is not a jump in the spectrum in the incomplete matrices. Secondly, one can trivially compute $\mathbf{A}\boldsymbol{\Omega}$ in parallel, which is the most expensive step. Hence, one can make full use of modern multi-core CPUs.

3.1. **SVD of a Perturbed Matrix.** It turns out that we can improve the efficiency of the online matrix completion approach further still by studying perturbations of the SVD. The problem of updating an SVD given a change is a rather important one as many problems such as clustering, denoising, and dimensionality reduction can be solved in part using the SVD. A particular issue in recommendation is that a new user who has rated few items would not get accurate recommendations, known as the "cold start" problem [20]. In the matrix completion context, a special case of this problem has been considered in [19] which studies the online learning of symmetric adjacency matrices. Furthermore, an online approach for matrix completion albeit without trace norm regularisation is presented in [1] which uses gradient descent along the lines of Grassmannian and an incremental approach to compute solutions as columns are added.

Consider the partial SVD given by $\mathbf{A}_k = \mathbf{P}_k\boldsymbol{\Sigma}_k\mathbf{Q}_k^T$, where $\mathbf{P}_k$, $\mathbf{Q}_k$ have as columns the left and right singular vectors, and $\boldsymbol{\Sigma}_k$ has diagonal entries corresponding to the k largest singular values. It is known that $\mathbf{A}_k$ is the best k-rank approximation of $\mathbf{A}$ using the Frobenius norm error. The change we are interested in can be encapsulated in a general sense as $\hat{\mathbf{A}} = \mathbf{A} + \mathbf{U}$ in which $\mathbf{U} \in \mathbb{R}^{m\times n}$. Note that one may also be interested in the addition of rows and columns to $\mathbf{A}$. In this case, it is trivial to phrase these changes in terms of that given above by noting: $\begin{bmatrix} \mathbf{A} \\ \mathbf{0} \end{bmatrix} = \begin{bmatrix} \mathbf{P} \\ \mathbf{0} \end{bmatrix} \boldsymbol{\Sigma}\mathbf{Q}^T$ and $\begin{bmatrix} \mathbf{A} & \mathbf{0} \end{bmatrix} = \mathbf{P}\boldsymbol{\Sigma}\begin{bmatrix} \mathbf{Q} & \mathbf{0} \end{bmatrix}^T$, and then adding an update matrix. Thus we will focus on finding the rank-k approximation of $\hat{\mathbf{A}}$, $\hat{\mathbf{A}}_k$, given the first k singular values and vectors of $\mathbf{A}$ and the perturbation $\mathbf{U}$.

There is a range of work which focuses on the above problem when $\mathbf{U}$ is low rank. One such method [27] uses the SVD of $\mathbf{A}$ to approximate the SVD of $\mathbf{A} + \mathbf{B}\mathbf{C}^T$, with $\mathbf{B} \in \mathbb{R}^{m\times p}$ and $\mathbf{C} \in \mathbb{R}^{n\times p}$, without recomputing the SVD of the new matrix. The total complexity is $\mathcal{O}((m+n)(pk+p^2)+k^3)$ where k is the rank of $\mathbf{U}$, however one requires first the decomposition of $\mathbf{U}$ into $\mathbf{B}\mathbf{C}^T$. An improvement in complexity based on similar principals is provided in [2] which costs $\mathcal{O}(mnk)$ for $k \leq \sqrt{\min(m,n)}$. Unfortunately in our case the update is small in terms of the number of nonzero elements but typically has a large rank.

To address this problem we present a simple randomised method for updating the SVD of $\mathbf{A}$ given a sparse (but not necessarily low rank) update $\mathbf{U}$. Unlike the updating methods mentioned above, we leverage our knowledge of the SVD algorithm to improve the approximation of the perturbed matrix $\hat{\mathbf{A}}$. We use Algorithm 2 with $\boldsymbol{\Omega} = [\mathbf{Q}_k\ \hat{\boldsymbol{\Omega}}]$ in which $\hat{\boldsymbol{\Omega}} \in \mathbb{R}^{n\times p}$ is a random Gaussian matrix. The idea is that one already has a good approximation of the first k right singular vectors of $\mathbf{A}$, which are not changed significantly by adding $\mathbf{U}$ and we improve the projection matrix by adding p random projections, where p is small. Notice also the following:

$$\hat{\mathbf{A}}\mathbf{Q}_k = \mathbf{P}_k\mathbf{\Sigma}_k + \mathbf{U}\mathbf{Q}_k,$$

and hence we need only compute $\mathbf{U}\mathbf{Q}_k$ and add it to precomputed $\mathbf{P}_k\mathbf{\Sigma}_k$. One need not use any power iteration (i.e. $q = 0$) and hence the complexity of the new step 2 is $\mathcal{O}((p + nk)|\alpha| + p|\omega|)$ where α is the set of nonzero entries in $\mathbf{U}$. Of note is that this step requires only a single scan of $\hat{\mathbf{A}}$ versus the $2q + 1$ required in Algorithm 2 and yet a highly accurate solution is obtained, as we shall later see.

3.2. **Analysis.** Here we study the online matrix completion method by looking at the error introduced by using the approximate SVD. Consider again Algorithm 2 whose error is bounded by the following theorem.

Theorem 1. [9] *Define $\boldsymbol{A} \in \mathbb{R}^{m\times n}$. Select an exponent q, a target number of singular vectors k with $2 \leq k \leq 0.5\min(m, n)$ and let $p = k$. Algorithm 2 returns a rank $2k$ factorisation $\boldsymbol{P}\boldsymbol{\Sigma}\boldsymbol{Q}^T$ such that*

$$\mathbb{E}\|\boldsymbol{A} - \boldsymbol{P}\boldsymbol{\Sigma}\boldsymbol{Q}^T\|_2 \leq \left[1 + 4\sqrt{\frac{2\min(m, n)}{k - 1}}\right]^{1/(2q+1)} \sigma_{k+1},$$

where $\mathbb{E}$ is the expectation with respect to the random projection matrix, $\|\cdot\|_2$ is the spectral norm and σ_{k+1} is the $(k + 1)$th largest singular value of $\boldsymbol{A}$.

Of note from this theorem is that the error decreases the term in square brackets exponentially fast as q increases. Furthermore, the expectation is shown to be almost always close to the typical outcome due to measure concentration effects. We can now study the error introduced by the randomised SVD to each iteration of Algorithm 1.

Theorem 2. *Define $f_\lambda(\boldsymbol{Z}) = \frac{1}{2}\|P_\omega(\boldsymbol{X}) - P_\omega(\boldsymbol{Z})\|_F^2 + \lambda\|\boldsymbol{Z}\|_*$ and let $\boldsymbol{Z} = S_\lambda(\boldsymbol{Y})$ for some matrix $\boldsymbol{Y}$. Furthermore, denote by $\hat{S}_\lambda$ the soft thresholding operator using the SVD as computed using Algorithm 2 with $p = k$ and let $\hat{\boldsymbol{Z}} = \hat{S}_\lambda(\boldsymbol{Y})$. Then the following bound holds:*

$$\mathbb{E}|f_\lambda(\boldsymbol{Z}) - f_\lambda(\hat{\boldsymbol{Z}})| \leq \lambda k(1 + \theta)\sigma_{k+1} + \frac{1}{2}\|\boldsymbol{\sigma}_{>k}\|_2^2 + k\left(1 + \frac{k}{k - 1}\right)^{\frac{1}{2q+1}} \|\boldsymbol{\sigma}_{>k}\|_{(2q+1)},$$

where k is the rank of the partial SVDs, $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r$ are the singular values of $\boldsymbol{Y}$, $\boldsymbol{\sigma}_{>k} = [\sigma_{k+1} \ \cdots \ \sigma_r]^T$, and $\theta = \left[1 + 4\sqrt{\frac{2\min(m,n)}{k-1}}\right]^{1/(2q+1)}$.

The proof of this theorem is deferred to the appendix. Notice that the first two terms on the right side of the inequality come from Theorem 1 and constitute the error in the approximation of the trace norm of $\mathbf{Z}$. The final terms represent the bound of the Frobenius norm difference between the approximate and real SVDs, the last of which will tend towards $k\sigma_{k+1}$ as q increases. Naturally, the bound is favourable when the singular values of the residual matrix after k are small. An interesting consequence of the bound in conjunction with soft impute is that successive matrices $\mathbf{Y}$ (or equivalently $\mathbf{Z}^{(j)}$ in Algorithm 1) have decreasing singular values towards the end of the spectrum and hence the error between $f_\lambda(\mathbf{Z})$ and $f_\lambda(\hat{\mathbf{Z}})$ decreases as one iterates. This helps to explain the good convergence

of the online algorithm. We later examine the error in the randomised SVD in practice.

4. Computational Results

In this section we highlight the efficacy of online matrix completion approach on one toy and several real datasets. We compare the online algorithm in conjunction with the randomised SVD approach at each iteration, the SVD updating method of Section 3.1, and PROPACK, denoted `RVSD`, `RSVD+` and `PROPACK` respectively, fixing $\epsilon = 10^{-3}$. Note that comparisons have already been made in [17] with MMMF and SVT on several datasets with competitive results in the offline case. All experimental code is implemented in Python with critical sections implemented in C++. For `RVSD` and `RSVD+` we parallelise the multiplication of a sparse matrix by a projection matrix. We attempted the same parallelisation strategy for `PROPACK` however the algorithm works using matrix-vector multiplications for which the overhead of parallelisation exceeded the computational gains made. Timings are recorded on a 24 core Intel Xeon X5670 CPU with 192GB RAM.

4.1. **Synthetic Datasets.** To begin with we consider a simple dataset generated using the following process. There are 20 matrices in the sequence, of which the first 10 are in $\mathbb{R}^{5000\times1000}$, and from the 10th to the 20th matrix sizes are increased uniformly in both dimensions to $\mathbb{R}^{10000\times1500}$. The fully observed matrix is of rank $r = 50$ with a decomposition of the form $\mathbf{P}\mathbf{\Sigma}\mathbf{Q}^T$ such that $\mathbf{P}, \mathbf{Q}$ are random orthogonal matrices and the singular values $\mathbf{\Sigma}_{ii}$s are uniformly randomly selected, with $\mathbf{\Sigma}_{ij} = 0$, $i \neq j$. For the first 10 matrices in the sequence, elements are observed initially with probability 0.03 and this increases in equal steps to 0.10. Individual elements are normalised to have a standard deviation 1 and we add a noise term $\mathcal{N}(0, 0.01)$. For each matrix in the sequence we split the observations into a training and test set of approximately the same sizes. As a preprocessing step, the nonzero elements of each row of the training matrices are centered according to the mean value of the corresponding row, and the equivalent transformation is applied to the test matrices. Note that we studied the rank of these matrices and found them to be nearly full rank despite being generated by a low rank decomposition. Furthermore the differences between successive matrices was high rank, making the low rank SVD update strategies of Section 3.1 impractical.

Before looking at error rates, we first studied the subspaces generated by soft impute in conjunction with `PROPACK` with the aim of observing how they change as the algorithm iterates on the training matrices. The same dataset of 20 matrices is used however we reset soft impute at each matrix so that the initial solution $\mathbf{Z}^{(1)} = \mathbf{0}$. We use a value of $k = 50$ for the partial SVD for each iteration of the algorithm which corresponds to the rank of the underlying matrices. Instead of directly using the λ parameter, which is sensitive to variations in the size and number of observed entries in a matrix, we use $\rho = \lambda/\sigma_1$ where σ_1 is the largest singular value of $\mathbf{X}_i$ and $\rho = 0.5$ in this case. As well as recording γ we also compute at the ith iteration

$$\theta_{\mathbf{P}_i} = \frac{\|(\mathbf{I} - \mathbf{P}_{(i-1)}\mathbf{P}_{(i-1)}^T)\mathbf{P}_i\|_F^2}{\|\mathbf{P}_i\|_F^2} = \frac{k - \|\mathbf{P}_i^T\mathbf{P}_{i-1}\|_F^2}{k},$$

where k is the dimensionality of $\mathbf{P}_i$ which is the matrix of left singular vectors of $\mathbf{Z}^{(i)}$. The corresponding measure for the right singular vectors of $\mathbf{Z}^{(i)}$ $\mathbf{Q}_i$, is denoted $\theta_{\mathbf{Q}_i}$. We additionally compute the change in the thresholded singular values in a similar fashion, $\phi_{\boldsymbol{\sigma}_i} = \|\boldsymbol{\sigma}_i - \boldsymbol{\sigma}_{i-1}\|^2 / \|\boldsymbol{\sigma}_i\|^2$, where $\boldsymbol{\sigma}_i$ is the soft thresholded singular values of the ith matrix. These measures are computed over all 20 training matrices and averaged. It is clear from Figure 1 that the left and right subspaces of $\mathbf{Z}$ rapidly

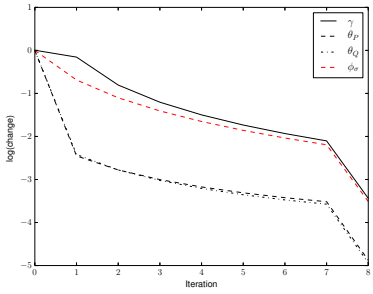

FIGURE 1. Mean differences in successive left and right subspaces of $\mathbf{Z}$ and error measure γ for soft impute.

decrease after the first iteration to approximately 3.7×10^{-3} and continue to fall. The error γ decreases at a slower rate and we see that most of the change occurs with the soft thresholded singular values of $\mathbf{Z}$. The important point to take note of is that the slowly varying subspaces of $\mathbf{Z}$ play into the updating of the randomised SVD presented in Section 3.1.

Next we evaluate the generalisation error of the matrix completion approaches by recording the root mean squared error (RMSE) on the observed entries,

$$\text{RMSE}(\mathbf{X}, \hat{\mathbf{X}}) = \sqrt{\frac{1}{|\omega|} \sum_{(i,j) \in \omega} (\mathbf{X}_{ij} - \hat{\mathbf{X}}_{ij})^2},$$

where $\hat{\mathbf{X}}_{ij}$ are the predicted elements. Furthermore, we postprocess the singular values as suggested in [17] so as to minimise $\|P_\omega(\mathbf{X}) - \sum_i \sigma_i P_\omega(\mathbf{p}_i \mathbf{q}_i^T)\|_F^2$, for $\sigma_i \geq 0$, $i = 0, \ldots, r$, using a maximum of 10^6 nonzero elements. The RMSEs are recorded on the training and test observations. The experiment is repeated using `RVSD`, `RSVD+` and `PROPACK` in conjunction with soft impute. With `RVSD` and `RSVD+` we explore different values of p and q, in particular the following pairs of values are used: $\{(10, 2), (50, 2), (10, 5)\}$. For the $(10, 5)$ case we also compute using cold restarts (i.e $\mathbf{Z}^{[i]} = \mathbf{0}$, $\forall i$) to contrast it to the use of previous solutions. The errors of the test observations are shown Figure 2. The randomised SVD methods are very similar although slightly worse than `PROPACK` for the most of the matrices in the sequence. Interestingly `RSVD+` matches or improves over `RSVD` particular towards the end of the sequence. The effect of cold restarts on the error is mixed: for the initial 10 matrices it provides an improvement in error however for the final 10 it seems to be a disadvantage. The key point however, is that the timing are considerably better for the randomised methods with warm restarts compared to both `PROPACK` and `RSVD` with cold restarts. With `RSVD+` $p = 10, q = 2$ for example, the total time for

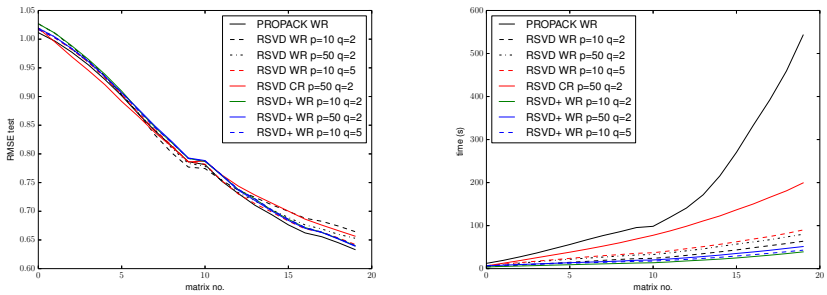

FIGURE 2. Errors and timings on the synthetic dataset, using Warm Restarts (WR) and Cold Restarts (CR).

	m	n	$\|\omega_1\|$	$\|\omega_T\|$
Flixster	50,130	18,369	6,376,264	7,825,955
ML	71,567	10,681	6,569,292	10,000,054
Netflix	480,189	17,770	17,023,860	100,480,507

TABLE 1. Information about the real datasets.

completing the entire sequence of matrices is 38.9s versus 543s for `PROPACK` and yet the error is only 0.007 larger for the final matrix.

4.2. **Real Datasets.** Next we use the online matrix completion algorithms in conjunction with three real datasets. The Netflix dataset contains ratings on a scale of 1 to 5 augmented with the date the rating was made which allows us to explore recommendation accuracy with time. We start with the set of ratings made by the end of 2003 and compute predictions for rating matrices at 30 day intervals, a total of 26 matrices. MovieLens 10M has ratings from 0 to 5 in steps of 0.5. We start at the rating matrix at the end of 2004, incrementing in 30 day intervals until March 2008 giving a total of 40 matrices. Finally, we use the Flixster movie dataset after processing it so that we keep only movies and users with 10 ratings or more, which are given on a scale of 1 to 5 in steps of 0.5. Ratings are used from January 2007 to November 2009 in 30 day increments, resulting in 21 matrices. Table 1 gives some information about these datasets. For each matrix we use a training/test split of 0.8/0.2, preprocessed by centering the rows as described for the synthetic data above. For the initial matrix in the sequence we perform model selection on the training set in order to set the parameters of the matrix completion methods. This is performed on a sample of at most 5×10^6 randomly sampled elements the training matrix using 5-fold cross validation. We select ρ from $\{0.05, 0.1, \ldots, 0.4\}$ and k is chosen from $\{8, 16, 32, 64, 128\}$. We postprocess the singular values as described above and set $p = 50, q = 2$, $p = 50, q = 3$ and $p = 10, q = 3$ for `RSVD` and `RSVD+`. As before, we train on the training observations and record the RMSE on the test observations.

Figures 3 show the resulting errors and timings for the MovieLens dataset. The prediction errors generally improve over time as one has access to more entries of the incomplete matrices. We see that `PROPACK` takes considerably longer to complete

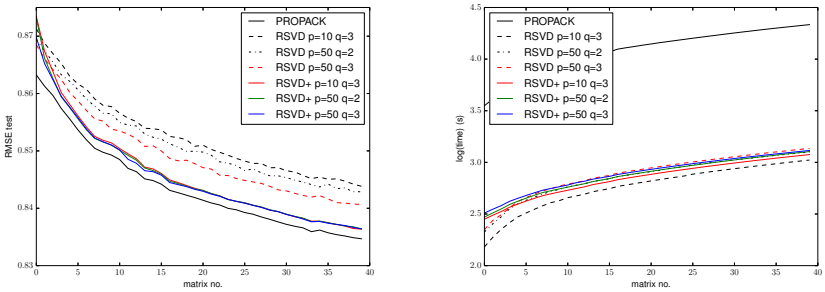

FIGURE 3. Errors and timings on the MovieLens dataset

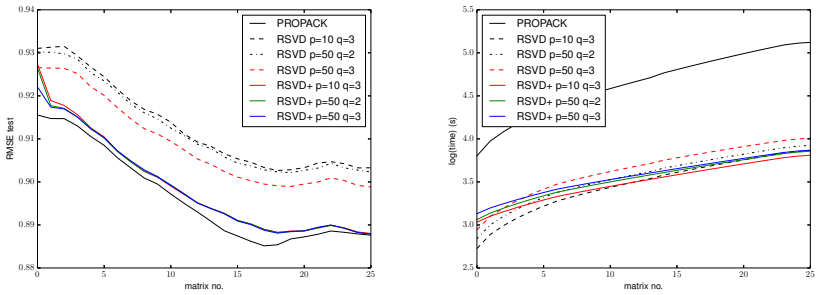

FIGURE 4. Errors and timings on the Netflix dataset

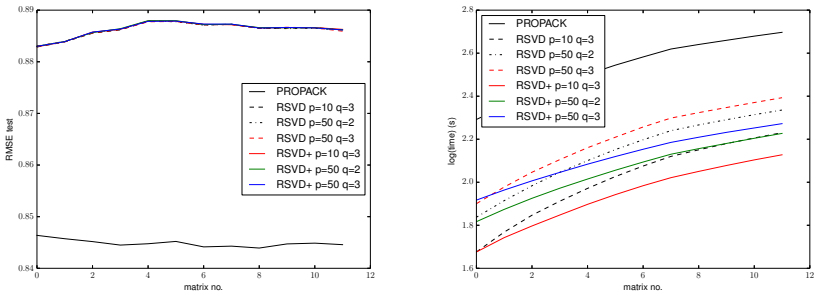

FIGURE 5. Errors and timings on the Flixster dataset

the sequence of matrices, however it does have a slight advantage in error. `PROPACK` requires 21,513 seconds to complete all the matrices whereas `RSVD+` $p = 10$, $q = 3$ took just 1189 seconds, both for rank-128 solutions. The differences in timings between the `RSVD+` methods is small since power iteration is only used for the initial solutions. `RSVD` favours a rank-64 decomposition due to errors in the error grid, and this explains why the corresponding RMSEs are worse than the other

methods and timings are better with $p = 10$, $q = 3$ relative to `RSVD+` for example. The final difference in RMSE between `RSVD+` $p = 10$ $q = 3$ and `PROPACK` is 0.0016.

On Netflix, Figure 4, we see that `PROPACK` requires 132,354 seconds to complete all matrices versus 6463 seconds for `RSVD+` $p = 10$, $q = 3$, an improvement factor of approximately 20. The difference in RMSE at the final matrix between these methods is negligible at 0.0004. Note that the `RSVD` methods choose $k = 64$ versus $k = 128$ for `RSVD+`. If we compare `RSVD` and `RSVD+` for $p = 50$, $q = 3$, the latter improves the error of the former at the final matrix by 0.01 and timings are 10187 and 7394 respectively. As with MovieLens we see that under difference values of p and q the `RSVD+` results converge to similar errors. If the spaces of the top k singular vectors of the input matrices change only slight then it follows that repeated sampling in the manner used in `RSVD+` will produce increasingly accurate results, a key strength of the method in this iterative context.

Finally, we come to Flixster in Figure 5 and on this dataset the methods all chose $k = 8$ $\rho = 0.05$ during model selection and the randomised methods have very similar errors. Of note from the plots is that the matrices do not become easier to complete over time, and `PROPACK` improves the error of the randomised methods by approximately 0.04. We believe that this dataset is particularly sensitive to round-off errors in the random SVD procedure. The timings shown in Figure 5 show that there is a clear advantage in the `RSVD+` methods over `RSVD`, for example observe the respective cumulative timings for $p = 50$, $q = 3$, 187s and 247s with negligible difference in error.

5. Conclusions

We addressed a critical issue in practical applications of matrix completion, namely online learning, based on soft impute which uses a trace norm penalty. The principal bottleneck of this algorithm is the computation of the SVD using PROPACK, which is not readily parallelisable and we motivated the randomised SVD for efficiently evaluating the SVD. Additionally, we showed how matrix completion can be conducted in an online setting by using previous solutions and a method to update the SVD under a potentially high-rank change. The resulting algorithm is simple to implement, and easily parallelisable for effective use of modern multi-core computer architectures. The expectation of the error of the algorithm in terms of the objective function is bounded theoretically, and empirical evidence is provided on its efficacy. In particular, on the large MovieLens and Netflix datasets, RSVD+ significantly reduces computational time upon PROPACK for a small penalty in error, and improves the efficiency of the randomised SVD.

Our novel SVD updating method opens up work in other algorithms which require the computation of SVDs or eigen-decompositions under high rank perturbations. We plan to study theoretically in more detail the error and convergence properties of our online matrix completion approach.

Appendix A. Proof of Theorem 2

This appendix section details the proof of the main theorem in the article. We start with a result which is analogous to Theorem 1 and bounds the Frobenius norm error of a random projection.

Theorem 3. [9] *Suppose that* $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ *with singular values* $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r$. *Choose a target rank* k *and an oversampling parameter* $p \geq 2$ *where* $k + p \leq$

$\min(m,n)$. *Draw an* $n \times (k+p)$ *standard Gaussian matrix* $\boldsymbol{\Omega}$*, and construct* $\boldsymbol{Y} = \boldsymbol{A\Omega}$*. Then the expected approximation error is bounded by*

$$\mathbb{E}\|(\boldsymbol{I} - \boldsymbol{P_Y})\boldsymbol{A}\|_F \leq \left(1 + \frac{k}{p-1}\right)^{1/2} \left(\sum_{j>k} \sigma_j^2\right)^{1/2},$$

where $\boldsymbol{P_Y}$ *is the projection matrix constructed from the orthogonalisation of* $\boldsymbol{Y}$*.*

Note that this error bound is the same as that between $\mathbf{A}$ and its randomised SVD, $\mathbf{P\Sigma Q}^T$, since from Algorithm 2 $\mathbf{P\Sigma Q}^T = \mathbf{VV}^T\mathbf{A}$. We now introduce another result which characterises the error using the power scheme

Theorem 4. [9] *Suppose that* $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ *and let* $\boldsymbol{\Omega}$ *be an* $n \times \ell$ *matrix. For some nonnegative integer* q*, let* $\boldsymbol{B} = (\boldsymbol{AA}^T)^q\boldsymbol{A}$ *and compute* $\boldsymbol{Y} = \boldsymbol{B\Omega}$*, then*

$$\|(\boldsymbol{I} - \boldsymbol{P_Y})\boldsymbol{A}\|_2 \leq \|(\boldsymbol{I} - \boldsymbol{P_Y})\boldsymbol{B}\|_2^{1/(2q+1)},$$

and it follows that

$$\|(\boldsymbol{I} - \boldsymbol{P_Y})\boldsymbol{A}\|_F \leq \sqrt{\ell}\|(\boldsymbol{I} - \boldsymbol{P_Y})\boldsymbol{B}\|_2^{1/(2q+1)}.$$

This gives us the necessary ingredients to derive a bound on the expected Frobenius norm error.

Theorem 5. *Define rank-r matrix* $\boldsymbol{A} \in \mathbb{R}^{m\times n}$*, select an exponent* q*, a target number of singular vectors* k *and a nonnegative oversampling parameter* p*. Let* $\boldsymbol{B} = (\boldsymbol{AA}^T)^q\boldsymbol{A}$ *and compute* $\boldsymbol{Y} = \boldsymbol{B\Omega}$ *in which* Ω *is an* $n \times (k+p)$ *standard Gaussian matrix. Then the expected approximation error is bounded by*

$$\mathbb{E}\|(\boldsymbol{I} - \boldsymbol{P_Y})\boldsymbol{A}\|_F^2 \leq (k+p)\left(1 + \frac{k}{p-1}\right)^{\frac{1}{2q+1}} \|\boldsymbol{\sigma}_{>k}\|_{(2q+1)},$$

where $\boldsymbol{\sigma}_{>k}$ *is a vector of singular values* $\sigma_{k+1}, \ldots, \sigma_r$ *and* $\|\cdot\|_p$ *is the p-norm of the input vector.*

Proof. We begin by using Hölder's inequality to bound $\mathbb{E}\|(\mathbf{I} - \mathbf{P_Y})\mathbf{A}\|_F^2$

$$\begin{aligned}
&\leq \left(\mathbb{E}\|(\mathbf{I} - \mathbf{P_Y})\mathbf{A}\|_F^{2q+1}\right)^{\frac{2}{2q+1}} \\
&\leq \left((k+p)^{\frac{2q+1}{2}}\mathbb{E}\|(\mathbf{I} - \mathbf{P_Y})\mathbf{B}\|_2\right)^{\frac{2}{2q+1}} \\
&\leq (k+p)\left(\left(1 + \frac{k}{p-1}\right)\sum_{j>k}\sigma_j^{(2q+1)}\right)^{\frac{1}{2q+1}},
\end{aligned}$$

where the second step makes use of Theorem 4. The final line uses Theorem 3 noting both that $\|\mathbf{X}\|_2 \leq \|\mathbf{X}\|_F$ for any matrix $\mathbf{X}$, and the singular values of $\mathbf{B}$ are given by $\sigma_1^{2q+1}, \ldots, \sigma_r^{2q+1}$. □

Before introducing the main theorem we present a lemma pertaining to the norm of thresholded SVDs.

Lemma 6. [17, 16] *The shrinkage operator $S_\lambda(\cdot)$ satisfies the following for any $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ with matching dimensions:*

$$\|S_\lambda(\boldsymbol{W}_1) - S_\lambda(\boldsymbol{W}_2)\|_F^2 \leq \|\boldsymbol{W}_1 - \boldsymbol{W}_2\|_F^2,$$

which implies $S_\lambda(\boldsymbol{W})$ is a continuous map in $\boldsymbol{W}$.

We can now study the error introduced by the randomised SVD to each iteration of Algorithm 2.

Theorem 7. *Define $f_\lambda(\boldsymbol{Z}) = \frac{1}{2}\|P_\omega(\boldsymbol{X}) - P_\omega(\boldsymbol{Z})\|_F^2 + \lambda\|\boldsymbol{Z}\|_*$ and let $\boldsymbol{Z} = S_\lambda(\boldsymbol{Y})$ for some matrix $\boldsymbol{Y}$. Furthermore, denote by $\hat{S}_\lambda$ the soft thresholding operator using the SVD as computed using Algorithm 2 with $p = k$ and let $\hat{\boldsymbol{Z}} = \hat{S}_\lambda(\boldsymbol{Y})$. Then the following bound holds:*

$$\begin{aligned}\mathbb{E}|f_\lambda(\boldsymbol{Z}) - f_\lambda(\hat{\boldsymbol{Z}})| \leq \lambda k(1+\theta)\sigma_{k+1} + \\ \frac{1}{2}\|\boldsymbol{\sigma}_{>k}\|_2^2 + k\left(1 + \frac{k}{k-1}\right)^{\frac{1}{2q+1}}\|\boldsymbol{\sigma}_{>k}\|_{(2q+1)},\end{aligned}$$

where k is the rank of the partial SVDs, $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r$ are the singular values of $\boldsymbol{Y}$, $\boldsymbol{\sigma}_{>k} = [\sigma_{k+1} \ \cdots \ \sigma_r]^T$, and $\theta = \left[1 + 4\sqrt{\frac{2\min(m,n)}{k-1}}\right]^{1/(2q+1)}$.

Proof. Define $T_k(\cdot)$ and $\hat{T}_k(\cdot)$ respectively as the k-rank SVD and randomised SVD of the input, then $\mathbb{E}\|T_k(\mathbf{Y}) - \hat{T}_k(\mathbf{Y})\|_2$ is

$$\begin{aligned} &= \mathbb{E}\|T_k(\mathbf{Y}) - \mathbf{Y} + \mathbf{Y} - \hat{T}_k(\mathbf{Y})\|_2 \\ &\leq \mathbb{E}(\|T_k(\mathbf{Y}) - \mathbf{Y}\|_2 + \|\mathbf{Y} - \hat{T}_k(\mathbf{Y})\|_2) \\ &\leq (1+\theta)\sigma_{k+1},\end{aligned}$$

where the 2nd line follows from the triangle inequality and the last uses Theorem 1 and the result from [7] that the best k rank approximation of a matrix is given by its k largest singular values and vectors with spectral residual σ_{k+1}. With similar remarks

$$\begin{aligned}&\mathbb{E}\|T_k(\mathbf{Y}) - \hat{T}_k(\mathbf{Y})\|_F^2 \\ &\quad= \mathbb{E}\|T_k(\mathbf{Y}) - \mathbf{Y} + \mathbf{Y} - \hat{T}_k(\mathbf{Y})\|_F^2 \\ &\quad\leq \mathbb{E}(\|T_k(\mathbf{Y}) - \mathbf{Y}\|_F^2 + \|\mathbf{Y} - \hat{T}_k(\mathbf{Y})\|_F^2) \\ &\quad\leq \|\boldsymbol{\sigma}_{>k}\|_2^2 + 2k\left(1 + \frac{k}{k-1}\right)^{\frac{1}{2q+1}}\|\boldsymbol{\sigma}_{>k}\|_{(2q+1)},\end{aligned}$$

where the final line uses Theorem 5.

Now we can write $|f_\lambda(\mathbf{Z}) - f_\lambda(\hat{\mathbf{Z}})|$ as follows:

$$\begin{aligned}
&\leq \frac{1}{2}\|P_\omega(\mathbf{Z}) - P_\omega(\hat{\mathbf{Z}})\|_F^2 + \lambda\left|(\|\mathbf{Z}\|_* - \|\hat{\mathbf{Z}}\|_*)\right| \\
&\leq \frac{1}{2}\|P_\omega(\mathbf{Z}) - P_\omega(\hat{\mathbf{Z}})\|_F^2 + \lambda\|\mathbf{Z} - \hat{\mathbf{Z}}\|_* \\
&\leq \frac{1}{2}\|\mathbf{Z} - \hat{\mathbf{Z}}\|_F^2 + \lambda\sqrt{k}\|\mathbf{Z} - \hat{\mathbf{Z}}\|_F \\
&= \frac{1}{2}\|S_\lambda(\mathbf{Y}) - \hat{S}_\lambda(\mathbf{Y})\|_F^2 + \lambda\sqrt{k}\|S_\lambda(\mathbf{Y}) - \hat{S}_\lambda(\mathbf{Y})\|_F \\
&\leq \frac{1}{2}\|T_k(\mathbf{Y}) - \hat{T}_k(\mathbf{Y})\|_F^2 + \lambda\sqrt{k}\|T_k(\mathbf{Y}) - \hat{T}_k(\mathbf{Y})\|_F \\
&\leq \frac{1}{2}\|T_k(\mathbf{Y}) - \hat{T}_k(\mathbf{Y})\|_F^2 + \lambda k\|T_k(\mathbf{Y}) - \hat{T}_k(\mathbf{Y})\|_2,
\end{aligned}$$

where the 5th line uses Lemma 6 and we assume all singular values after k are zero in the soft thresholded SVDs. If we take expectations of the above then

$$\begin{aligned}
\mathbb{E}|f_\lambda(\mathbf{Z}) - f_\lambda(\hat{\mathbf{Z}})| \leq \lambda k(1+\theta)\sigma_{k+1} + \\
\frac{1}{2}\|\boldsymbol{\sigma}_{>k}\|_2^2 + k\left(1 + \frac{k}{k-1}\right)^{\frac{1}{2q+1}}\|\boldsymbol{\sigma}_{>k}\|_{(2q+1)},
\end{aligned}$$

which is the required result. □

References

[1] L. Balzano, R. Nowak, and B. Recht. Online identification and tracking of subspaces from highly incomplete information. In *Communication, Control, and Computing (Allerton), 2010 48th Annual Allerton Conference on*, pages 704–711. IEEE, 2010.

[2] M. Brand. Fast low-rank modifications of the thin singular value decomposition. *Linear algebra and its applications*, 415(1):20–30, 2006.

[3] J.-F. Cai, E. J. Candès, and Z. Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4):1956–1982, 2010.

[4] E. Candès and T. Tao. The power of convex relaxation: Near-optimal matrix completion. *Information Theory, IEEE Transactions on*, 56(5):2053–2080, 2010.

[5] E. J. Candes and Y. Plan. Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925–936, 2010.

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

[7] C. Eckart and G. Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.

[8] G. H. Golub and C. F. Van Loan. *Matrix computations*, volume 3. JHUP, 2012.

[9] N. Halko, P.-G. Martinsson, and J. A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, 53(2):217–288, 2011.

[10] M. Jaggi and M. Sulovskỳ. A simple algorithm for nuclear norm regularized problems. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 471–478, 2010.

[11] S. Ji and J. Ye. An accelerated gradient method for trace norm minimization. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 457–464. ACM, 2009.

[12] R. H. Keshavan, A. Montanari, and S. Oh. Matrix completion from a few entries. *Information Theory, IEEE Transactions on*, 56(6):2980–2998, 2010.

[13] R. M. Larsen. Lanczos bidiagonalization with partial reorthogonalization. *DAIMI Report Series*, 27(537), 1998.

[14] K. Lee and Y. Bresler. Admira: Atomic decomposition for minimum rank approximation. *Information Theory, IEEE Transactions on*, 56(9):4402–4416, 2010.

[15] Y.-J. Liu, D. Sun, and K.-C. Toh. An implementable proximal point algorithmic framework for nuclear norm minimization. *Mathematical Programming*, 133(1-2):399–436, 2012.

[16] S. Ma, D. Goldfarb, and L. Chen. Fixed point and bregman iterative methods for matrix rank minimization. *Mathematical Programming*, 128(1-2):321–353, 2011.

[17] R. Mazumder, T. Hastie, and R. Tibshirani. Spectral regularization algorithms for learning large incomplete matrices. *The Journal of Machine Learning Research*, 11:2287–2322, 2010.

[18] B. Recht, C. Ré, and S. Wright. Parallel stochastic gradient algorithms for large-scale matrix completion. *Optimization Online*, 2011.

[19] E. Richard, N. Baskiotis, T. Evgeniou, and N. Vayatis. Link discovery using graph feature tracking. In *Proceedings of Neural Information Processing Systems (NIPS)*, 2010.

[20] A. Schein, A. Popescul, L. Ungar, and D. Pennock. Methods and metrics for cold-start recommendations. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 253–260. ACM, 2002.

[21] B. Schölkopf, A. J. Smola, and K.-R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5):1299–1319, 1998.

[22] S. Shalev-Shwartz, A. Gonen, and O. Shamir. Large-scale convex minimization with a low-rank constraint. *arXiv preprint arXiv:1106.1622*, 2011.

[23] N. Srebro, J. D. Rennie, and T. Jaakkola. Maximum-margin matrix factorization. *Advances in neural information processing systems*, 17(5):1329–1336, 2005.

[24] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[25] K.-C. Toh and S. Yun. An accelerated proximal gradient algorithm for nuclear norm regularized linear least squares problems. *Pacific Journal of Optimization*, 6(615-640):15, 2010.

[26] J.-M. Yun and S. Choi. Nyström approximations for scalable face recognition: A comparative study. In *Neural Information Processing*, pages 325–334. Springer, 2011.

[27] H. Zha and H. Simon. On updating problems in latent semantic indexing. *SIAM Journal on Scientific Computing*, 21:782, 1999.

[28] H. Zhao, J. Han, N. Wang, C. Xu, and Z. Zhang. A fast spectral relaxation approach to matrix completion via kronecker products. In *Twenty-Fifth AAAI Conference on Artificial Intelligence*, 2011.