Learning Low Rank Matrices from $O(n)$ Entries

Raghunandan H. Keshavan, Andrea Montanari and Sewoong Oh

***Abstract*— How many random entries of an $n \times n\alpha$, rank r matrix are necessary to reconstruct the matrix within an accuracy δ? We address this question in the case of a random matrix with bounded rank, whereby the observed entries are chosen uniformly at random. We prove that, for any $\delta > 0$, $C(r, \delta)n$ observations are sufficient.**

Finally we discuss the question of reconstructing the matrix *efficiently*, and demonstrate through extensive simulations that this task can be accomplished in nPoly$(\log n)$ operations, for small rank.

I. Introduction and main results

A. Problem definition

Let M be an $n \times m$ matrix of rank (at most) r and assume that $n\epsilon$ uniformly random entries of M are revealed. Does this knowledge allow to approximately reconstruct M?

The answer is negative unless the matrix has some specific structure. In this paper we assume that M is a *random rank-r matrix*, i.e. $\mathsf{M} = \mathsf{U} \cdot \mathsf{V}$ where U is a $n \times r$ matrix with iid entries and V an independent $r \times m$ matrix with iid entries. The distributions of the entries of U and V are denoted, respectively as p_0 and q_0.

The metric we shall consider is the root mean square error (RMSE). If $\{\mathsf{M}_{i,a}\}$ are the entries of M, and $\widehat{\mathsf{M}}$ is its estimate based on the observed entries, we have

$$D(\mathsf{M}, \widehat{\mathsf{M}}) \equiv \left\{ \frac{1}{nm} \sum_{i,a} |\mathsf{M}_{i,a} - \widehat{\mathsf{M}}_{i,a}|^2 \right\}^{1/2} . \tag{1}$$

Notice that this coincides, up to a factor, with the distance induced by the Frobenius norm $D(\mathsf{M}, \widehat{\mathsf{M}}) = ||\mathsf{M} - \widehat{\mathsf{M}}||_{\rm F}/\sqrt{nm}$.

In the following we shall denote by $R \ni i, j, k, \dots$ the set of rows of M and by $C \ni a, b, c, \dots$ its set of columns. The subset of revealed entries will be denoted by $E \subseteq R \times C$.

B. Motivation and related work

Low rank matrices have been proposed as statistical models to describe a number of complex data sources. For instance, the matrix of empirical correlations among stock prices in a market is approximately low rank if price fluctuations are driven by a few underlying mechanisms [1]. A completely different application is provided by the matrix of square distances among n sensors in 3 dimension, which has rank $r = 5$ [2].

Low rank matrices have been proposed as a model for collaborative filtering data. As a concrete example we shall focus here on the Netflix Challenge dataset [3]. This dataset concerns a set C of approximately $5 \cdot 10^5$ customers and R of $2 \cdot 10^4$ movies. For about 10^8 customer-movie pairs $(i, a) \in E$, the corresponding rating (an integer between 1 and 5) is provided. The challenge consists in predicting the ratings of 10^6 non-revealed customer-movie pairs within a root mean square error smaller than 0.8563.

One possible approach consists in considering the customer-movie matrix M (or a rescaled version of it) and assuming that it has low rank to predict the requested entries. Indeed, a simple coordinate descent algorithm that minimizes the energy function

$$\sum_{(i,a)\in E} (\mathsf{M}_{i,a} - (\mathsf{UV})_{i,a})^2 + \lambda||\mathsf{U}||_{\rm F}^2 + \lambda||\mathsf{V}||_{\rm F}^2 \tag{2}$$

provides good predictions (within the Netflix competition, it was used by SimonFunk).

In general, the matrix completion problem is not convex, and the descent algorithm is not guaranteed to converge to the original matrix M even if this is the unique rank r matrix consistent with the observations. A possible alternative consists in relaxing the rank constraint, by looking instead for a matrix $\widehat{\mathsf{M}}$ of minimal nuclear norm (recall that the nuclear norm of $\widehat{\mathsf{M}}$ is the sum of the absolute values of its singular values). The problem then becomes convex and indeed reducible to semidefinite programming. In [4] it was shown that this relaxation indeed recovers the original low rank matrix M, given that a sufficient number of random linear combinations of its entries are revealed.

The case in which a random subset of the entries is revealed (which is relevant for collaborative filtering) was treated in [5]. This paper proves that the convex relaxation is tight with high probability[1] if $\epsilon \geq C\, r\, n^{1/5} \log n$. In particular this implies two statements: (i) For $\epsilon \geq C\, r\, n^{1/5} \log n$, $n\epsilon$ random entries uniquely determine a random rank-r matrix. (ii) This matrix is the unique minimum of a semidefinite program.

C. Main results

The results briefly reviewed above leave open several key issues:

1. Why is it necessary to observe $\Theta(n^{6/5})$ entries to reconstruct a rank-r matrix, that has $\Theta(n)$ degrees of freedom?
2. As the Netflix challenge shows, it is not realistic nor necessary to reconstruct M exactly. What is the

Raghunandan H. Keshavan is with the Department of Electrical Engineering, Stanford University, raghuram@stanford.edu. Andrea Montanari is with Departments of Electrical Engineering and Statistics, Stanford University, montanari@stanford.edu. Sewoong Oh is with the Department of Electrical Engineering, Stanford University, swoh@stanford.edu.

[1] Strictly speaking, the matrix model treated in [5] is slightly different from the one considered here. However it should not be hard to prove that the two models are asymptotically equivalent for large n.

trade-off between RMSE distortion and number of observations?

3. In general, semidefinite programming has $\Theta(n^6)$ complexity [6]. This is affordable up to $n \approx 10^2$, but way beyond current capabilities when $n \approx 10^5$ as in modern datasets.

In this paper we address the first two points and show that $O(n)$ observations are sufficient to reconstruct a low rank matrix within any positive distortion.

Theorem I.1. *Let $\mathsf{M} = \mathsf{U} \cdot \mathsf{V}$ be a random rank-r matrix with n rows and $n\alpha$ columns and assume the distributions of $\mathsf{U}_{i,k}$ and $\mathsf{V}_{k,a}$ to have support in $[-1, 1]$. Let E be a random subset of $n\epsilon$ entries in $R \times C$. Then, with high probability, any matrix rank-r matrix $\widehat{\mathsf{M}}$ such that $|\mathsf{M}_{i,a} - \widehat{\mathsf{M}}_{i,a}| \le \Delta$ for all $(i,a) \in E$, and with factors $\mathsf{U}_{i,k}, \mathsf{V}_{k,a} \in [-1, 1]$, also satisfies*

$$D(\mathsf{M}, \widehat{\mathsf{M}}) \le \Delta + 2r\, \widetilde{\epsilon}^{-1/2} \log(10\widetilde{\epsilon})\,, \tag{3}$$

where $\widetilde{\epsilon} \equiv \epsilon/(1+\alpha)r$.

Notice that the term Δ in the above inequality is unavoidable. Since we are looking for matrices that match the observed entries only within precision Δ, we cannot hope for a RMSE smaller than Δ. In the second term, the factor $2r$ corresponds to the maximal distance between matrix entries in the present model, while the ϵ-dependent factor tends to 0 as $\epsilon \to \infty$. Notice that $\widetilde{\epsilon}$ is exactly the number of observations per degree of freedom.

The proof of this statement is given in Section III, which also provides a much more accurate upper bound. The latter is –however– not straightforward to evaluate. While it is clear that small RMSE cannot be achieved with less than $\Theta(n)$ observed matrix elements, Section IV proves a quantitative lower bound of this form.

In Section V we address the question of efficient reconstruction and demonstrate that $O(n \log n)$ operations are sufficient to reconstruct random low rank matrices with rank $r \le 4$, from $O(n)$ entries. Indeed such performances are achieved by a straightforward stochastic local search algorithm that we refer to as WalkRank or by a coordinate descent algorithm. A formal analysis of these algorithms will be presented in a future publication. Finally, in Section VI we use these results to compare random low rank matrices and the Netflix dataset.

Before dwelling into the intricacies of the full problem, the next Section discusses a particularly simple but perhaps instructive case: rank $r = 1$.

II. A WARMUP EXAMPLE

If M has rank 1, most of the questions listed above have a simple answer with a suggestive graph-theoretical interpretation.

Assume to know 3 entries of the matrix M that belong to the same 2×2 minor. Explicitly, for two row indices $i, j \in R$ and two column indices $a, b \in C$, the entries $\mathsf{M}_{i,a}$, $\mathsf{M}_{j,a}$, $\mathsf{M}_{i,b}$ are known. Unless $\mathsf{M}_{i,a} = 0$, the fourth entry of the same minor is then uniquely determined $\mathsf{M}_{j,b} = \mathsf{M}_{j,a}\mathsf{M}_{i,b}/\mathsf{M}_{i,a}$. The case $\mathsf{M}_{i,a} = 0$ can be treated separately but, for the sake of simplicity we shall assume that the distribution p_0, q_0 do not have mass on 0.

This observation suggests a simple matrix completion algorithm: Recursively look for a 2×2 minor wit a unique unknown entry and complete it according to the rule $\mathsf{M}_{j,b} = \mathsf{M}_{j,a}\mathsf{M}_{i,b}/\mathsf{M}_{i,a}$. As anticipated above, this algorithm has a nice graph-theoretic interpretation. Consider the bipartite graph $G = (R, C, E)$ with vertices corresponding to the row and columns of M and edges for the observed entries. If a 2×2 minor has a unique unknown entry, it means that the corresponding vertices $j \in R$, $b \in C$ are connected by a length-3 path in G. Hence the algorithm recursively adds edges to G connecting distance-3 vertices.

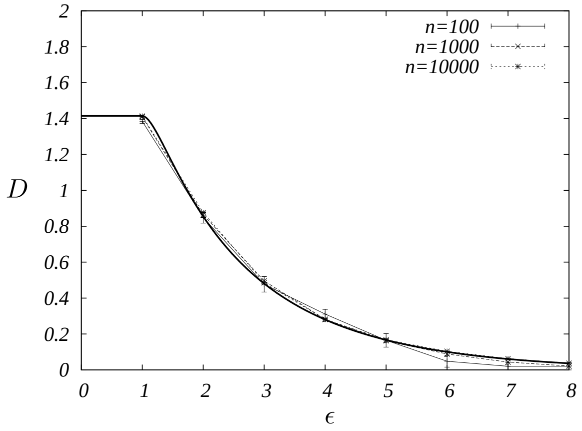

Fig. 1. Learning random rank-1 matrices. The continuous line is the optimal distortion (achieved by the recursive completion algorithm). Data points correspond to a $O(n)$ complexity local search algorithm.

After at most $O(n^2)$ operations the process described halts on a graph that is a disjoint union of cliques, corresponding to the connected components in G. Each edge corresponds to a correctly predicted matrix entry. Clearly, in the large n-limit the only components with $\Theta(n)$ matter (as they have $\Theta(n^2)$ edges). It is a fundamental result in random graph theory that there is no such component for $\epsilon \le 1/\sqrt{\alpha}$. For $\epsilon > 1/\sqrt{\alpha}$ there is one such component involving approximately $n\xi$ in R and $m\zeta$ vertices in C, where (ξ, ζ) is the unique positive solution of

$$\xi = 1 - e^{-\epsilon\alpha\zeta}\,, \qquad \zeta = 1 - e^{-\epsilon\xi}\,. \tag{4}$$

This analysis implies the following result.

Proposition II.1. *Let $\mathsf{M} = \mathsf{U}\cdot\mathsf{V}$ be a random rank 1 matrix, and denote by $\xi(\epsilon)$, $\zeta(\epsilon)$ the largest solution of Eq. (4). Then there exists an algorithm with $O(n^2)$ complexity achieving, with high probability, RMSE*

$$D(\mathsf{M}, \widehat{\mathsf{M}}) = \sqrt{1 - \xi(\epsilon)\zeta(\epsilon)}\, D_0 + O(\sqrt{(\log n)/n})\,. \tag{5}$$

where $D_0 \equiv \sqrt{\mathbb{E}(V_1^2)\mathbb{E}(U_1^2)}$. Further, if the entries U_i, V_a have symmetric distribution, then no algorithm does achieve smaller distortion.

Proof. The mentioned distortion is achieved by the recursive completion algorithm, whereby matrix element corresponding to vertex pairs in distinct components are predicted to vanish. This is optimal if the matrix element distribution is symmetric. Indeed the conditional matrix element distribution remains symmetric even given the observations. □

For massive datasets even $O(n^2)$ complexity is unaffordable. Figure 1 compares the minimal distortion guaranteed by Proposition II.1 with the performances of the WalkRank algorithm described in Section V. Here the factors U_i, V_a where chosen uniformly in $\{+1, -1\}$.

III. Upper Bound and Proof of Theorem I.1

In this section we prove the upper bound on distortion stated in Theorem I.1. The proof proceeds in three steps. First we will consider the case in which the factor entries $\mathsf{U}_{i,k}$, $\mathsf{V}_{k,a}$ are supported on a finite set, and prove a (tighter) upper bound via a counting argument. Then we'll use a quantization argument to generalize this bound to the continuous case. Finally, we simplify our bound to get the pleasing expression in Theorem I.1. Unhappily this simplification entails a worsening of the bound.

A. The discrete case

We start by introducing a couple of new notations. Given a row index $i \in R$, we let $\vec{u}_i^0 = (\mathsf{U}_{i,1}, \dots, \mathsf{U}_{i,r})$ be the i-th row of U. Analogously, for $a \in C$, let $\vec{v}_a^0$ be the a-th column of V. We then have

$$\mathsf{M}_{i,a} = \vec{u}_i^0 \cdot \vec{v}_a^0 \, . \tag{6}$$

We also write $\vec{u}_i^0 = (u_{i,1}^0, \dots, u_{i,r}^0)$ and $\vec{v}_a^0 = (v_{a,1}^0, \dots, v_{a,r}^0)$ for the components of these vectors. These are assumed to be iid's with distributions p_0 (for $\vec{u}$) and q_0 (for $\vec{v}$) supported on a finite set $A_N \subset \mathbb{R}$ with $|A_N| = N$ points. Typical examples are $A_2 = \{-1, +1\}$ or $A_{2M+1} \equiv \{-M\varepsilon, -(M-1)\varepsilon, \dots, (M-1)\varepsilon, M\varepsilon\}$). Our basic counting estimate is stated below.

Proposition III.1. *Let $\Delta \ge 0$ and M be a random rank-r matrix with factors supported in A_N. Then, with high probability any rank-r matrix $\widehat{\mathsf{M}}$ with factors supported in A_N that satisfies $|\mathsf{M}_{ia} - \widehat{\mathsf{M}}_{ia}| \le \Delta$ for all $(i,a) \in E$ also satisfies $D(\mathsf{M}, \widehat{\mathsf{M}}) \le \overline{\delta}(\epsilon, \alpha, \Delta) + o_n(1)$, where*

$$\overline{\delta}(\epsilon, \alpha, \Delta) = \sup_{p \in \mathcal{D}(p_0), q \in \mathcal{D}(q_0)} \{ \, d(p,q) \; : \; \phi_\Delta(p,q) \; \ge 0 \} \; . \tag{7}$$

Here the $\sup$ *over p (over q) is taken over the space of distributions $\mathcal{D}(p_0)$ (respectively $\mathcal{D}(q_0)$) over $(A_N)^r \times (A_N)^r$ such that $\sum_{\vec{u}} p(\vec{u}, \vec{u}^0) = p_0(\vec{u}^0)$ (respectively $\sum_{\vec{v}} q(\vec{v}, \vec{v}^0) = q_0(\vec{v}^0)$). The functionals appearing in Eq. (7) are defined by*

$$d(p,q) \equiv \left\{ \mathbb{E}_{p,q} \, |\vec{u} \cdot \vec{v} - \vec{u}^0 \cdot \vec{v}^0|^2 \right\}^{1/2} \, , \tag{8}$$

and

$$\begin{aligned} \phi_\Delta(p,q) \equiv\ & H(p) - H(p_0) + \alpha[H(q) - H(q_0)] + \\ & + \epsilon \, \mathbb{E}_{p_0,q_0} \log \mathbb{P}_{p,q}\big\{ |\vec{u} \cdot \vec{v} - \vec{u}^0 \cdot \vec{v}^0| \le \Delta \, \big| \, \vec{u}^0, \vec{v}^0 \big\} \, , \end{aligned} \tag{9}$$

Proof. Define $Z_G(\Delta, \delta)$ (G is the bipartite graph with edge set E) as the number of matrices $\widehat{\mathsf{M}}$ of the form (6) such that:

(1) $|\mathsf{M}_{i,a} - \widehat{\mathsf{M}}_{i,a}| \le \Delta$ for all $(i,a) \in E$;
(2) $D(\mathsf{M}, \widehat{\mathsf{M}}) \ge \delta$.

This can be written as

$$Z_G(\Delta, \delta) = \sum_{\{\vec{u}_i, \vec{v}_a\} \in C(\delta)} \; \prod_{(i,a) \in E} \mathbb{I}(|\vec{u}_i \cdot \vec{v}_a - \vec{u}_i^0 \cdot \vec{v}_a^0| \le \Delta) \quad , \tag{10}$$

where $C(\delta)$ is the set of vectors that satisfy condition (2) above. We further define the set of *typical instances* (M, E), $\mathsf{Typ}(\gamma)$ through the following conditions:

(a) Let $\theta_{\mathsf{U}}(\,\cdot\,)$ be the type of factor U, namely $n\theta_{\mathsf{U}}(\vec{u})$ is the number of row indices $i \in R$ such that $\vec{u}_i = \vec{u}$. Then for $(\mathsf{M}, E) \in \mathsf{Typ}(\gamma)$, we have $D(\theta_{\mathsf{U}} || p_0) \le \gamma$.
(b) Analogously, for the type of factor V we require $D(\theta_{\mathsf{V}} || q_0) \le \gamma$.
(c) Finally, let $\theta_E(\,\cdot\,,\,\cdot\,)$ be the edge type, i.e. $n\epsilon\theta_E(\vec{u}, \vec{v})$ is the number of edges $(i,a) \in E$ such that $\vec{u}_i = \vec{u}$ and $\vec{v}_a = \vec{v}$. We then require $D(\theta_{\mathsf{V}} || p_0 \cdot q_0) \le \gamma$ (where $p_0 \cdot q_0$ is the product distribution on $\vec{u}$, $\vec{v}$).

By standard arguments [7] we have $\mathbb{P}\{\mathsf{Typ}(\gamma)\} \to 1$ for any positive γ as $n \to \infty$. We then define

$$\widehat{Z}_G(\Delta, \delta) \equiv Z_G(\Delta, \delta) \, \mathbb{I}\big((\mathsf{M}, E) \in \mathsf{Typ}(\gamma)\big) \, . \tag{11}$$

According to lemma III.2, the expectation of $\widehat{Z}_G(\Delta, \delta)$ vanishes as n tends to infinity for $\delta > \overline{\delta}(\epsilon, \alpha, \Delta)$. Since $\mathbb{P}\{\mathsf{Typ}(\gamma)\} \to 1$ and using Markov inequality, this implies that $\lim_{n \to \infty} \mathbb{P}\{Z_G(\Delta, \delta) > 0\} = 0$. In conclusion, any matrix $\widehat{\mathsf{M}}$ that satisfies $|\mathsf{M}_{ia} - \widehat{\mathsf{M}}_{ia}| \le \Delta$ for all $(i,a) \in E$ results in a distance metric smaller than $\overline{\delta}(\epsilon, \alpha, \Delta)$ with high probability, as n tends to infinity. □

Lemma III.2. *For any $\delta > \overline{\delta}(\epsilon, \alpha, \Delta)$ there exists $\gamma > 0$ such that $\lim_{n \to \infty} \mathbb{E}_{E,\mathsf{M}}\{\widehat{Z}_G(\Delta, \delta)\} = 0$.*

Proof. $Z_G(\Delta, \delta)$ is a random variable where the randomness comes from the matrix elements $\mathsf{M}_{i,a}$ and the choice of the sampling set E. Since E is uniformly random, we can take any realization of $\mathsf{M} = \mathsf{U} \cdot \mathsf{V}$ from the typical set according to iid p_0 and iid q_0. Given one such realization of $\mathsf{U} = (\vec{u}_1^0, \dots, \vec{u}_n^0)$ and $\mathsf{V} = (\vec{v}_1^0, \dots, \vec{v}_m^0)$, go through all the estimations $\widehat{\mathsf{M}} = \widehat{\mathsf{U}} \cdot \widehat{\mathsf{V}}$, where $\widehat{\mathsf{U}} = (\vec{u}_1, \dots, \vec{u}_n)$ and $\widehat{\mathsf{V}} = (\vec{v}_1, \dots, \vec{v}_m)$. Now group the set of assignments $\widehat{\mathsf{U}}$ and $\widehat{\mathsf{V}}$ that have the same empirical distribution, and let $p(\vec{u}, \vec{u}^0)$ and $q(\vec{v}, \vec{v}^0)$ denote the joint distribution. Then, the number of different assignments with same empirical distribution (p,q) is $e^{n\{H(p) - H(p_0)\} + m\{H(q) - H(q_0)\}}$. For each distribution pair (p,q) that satisfy condition (2) above, we fix the factors $\widehat{\mathsf{U}}$ and $\widehat{\mathsf{V}}$ and compute the probability that they satisfies condition (1). Denoting by $\mathbb{E}'_{E,\mathsf{M}}\{\cdots\} = \mathbb{E}_{E,\mathsf{M}}\{\cdots \mathbb{I}((E, \mathsf{M}) \in \mathsf{Typ}(\gamma))\}$ the expectation restricted to

$(E, \mathsf{M}) \in \mathsf{Typ}(\gamma)$, we have

$$\begin{aligned} &\mathbb{E}'_{E,\mathsf{M}}\{Z_G(\Delta, \delta)\} \\ &= \mathbb{E}'_{E,\mathsf{M}}\left\{ \sum_{\{\vec{u}_i,\vec{v}_a\}\in C(\delta)} \prod_{(i,a)\in E} \mathbb{I}(|\vec{u}_i \cdot \vec{v}_a - \vec{u}_i^0 \cdot \vec{v}_a^0| \le \Delta) \right\} \\ &\doteq \sum_{\substack{p\in\mathcal{D}(p_0),q\in\mathcal{D}(q_0)\\ d(p,q)\ge\delta}} e^{nH(p|p_0)+mH(q|q_0)}. \\ &\qquad \mathbb{E}'_E\left\{ \prod_{(i,a)\in E} \mathbb{I}(|\vec{u}_i \cdot \vec{v}_a - \vec{u}_i^0 \cdot \vec{v}_a^0| \le \Delta) \right\} \end{aligned}$$

To compute the expectation in the last inequality, we look at a typical realization of E and partition it into subsets $\{E_{\vec{u}^0,\vec{v}^0}\}$, for $(\vec{u}^0, \vec{v}^0) \in (A_N)^r \times (A_N)^r$, defined as follows. $(i, a) \in E$ is in $E_{\vec{u}^0,\vec{v}^0}$ if $\vec{u}_i^0 = \vec{u}^0$ and $\vec{v}_a^0 = \vec{v}^0$. By definition $|E_{\vec{u}^0,\vec{v}^0}| = n\epsilon\theta_E(\vec{u}_0, \vec{v}_0)$. Further $E_{\vec{u}^0,\vec{v}^0}$ is uniformly random given its size. Within the typical set $\mathsf{Typ}(\gamma)$, $\theta_E(\vec{u}_0, \vec{v}_0)$ is close to $p_0(\vec{u}^0)q_0(\vec{v}^0)$. We thus get

$$\begin{aligned} &\mathbb{E}'_E\left\{ \prod_{(i,a)\in E} \mathbb{I}(|\vec{u}_i \cdot \vec{v}_a - \vec{u}_i^0 \cdot \vec{v}_a^0| \le \Delta) \right\} \\ &\doteq \prod_{\vec{u}^0,\vec{v}^0} \mathbb{E}_{E_{\vec{u}^0,\vec{v}^0}} \left\{ \prod_{(i,a)\in E_{\vec{u}^0,\vec{v}^0}} \mathbb{I}(|\vec{u}_i \cdot \vec{v}_a - \vec{u}_i^0 \cdot \vec{v}_a^0| \le \Delta) \right\} \\ &\overset{\mathrm{d}}{=} \prod_{\vec{u}^0,\vec{v}^0} \mathbb{P}\left\{ |\vec{u}_i \cdot \vec{v}_a - \vec{u}_i^0 \cdot \vec{v}_a^0| \le \Delta \;\middle|\; \vec{u}^0, \vec{v}^0 \right\}^{n\epsilon\theta_E(\vec{u}^0,\vec{v}^0)} . \end{aligned}$$

Finally, we get,

$$\mathbb{E}'_{E,\mathsf{M}}\{Z_G(\Delta, \delta)\} \le e^{n\kappa(\gamma)} \sum_{\substack{p\in\mathcal{D}(p_0),q\in\mathcal{D}(q_0)\\ d(p,q)\ge\delta}} e^{n\phi_\Delta(p,q)} . \tag{12}$$

where $\kappa(\gamma) \to 0$ as $\gamma \to 0$. For (p, q) that satisfies $d(p, q) > \overline{\delta}(\epsilon, \alpha, \Delta)$, we know that $\phi_\Delta(p, q) < 0$ by definition. Hence, for γ small enough, $\delta > \overline{\delta}(\epsilon, \alpha)$ is a sufficient condition for $\lim_{n\to\infty} \mathbb{E}_{E,\mathsf{M}}\{\widehat{Z}_G(\Delta, \delta)\} = 0$. □

B. General distributions via quantization

Above tighter upper bound can be generalized to matrices in theorem I.1 via quantization argument. In this section, we're interested in recovering a continuous real valued matrix M from samples of its entries. First, we estimate it using factors $\widehat{\mathsf{U}}_{i,k}$, $\widehat{\mathsf{V}}_{k,a}$ supported in the continuous alphabet. Then, the distortion is bounded using the upper bound from section III-A via quantization.

Proposition III.3. *Let $\Delta \ge 0$ and M be a random rank-r matrix with factors supported in continuous bounded alphabet A_c. Let A_δ be discrete quantized alphabet of A_c, with maximum quantization error less than $\delta/2$. $\widehat{\mathsf{M}}$ is the rank-r estimation with factors supported in A_c. Then, with high probability, any matrix $\widehat{\mathsf{M}}$ that satisfies $|\mathsf{M}_{ia} - \widehat{\mathsf{M}}_{ia}| \le \Delta$ for all $(i, a) \in E$ also satisfies $D(\mathsf{M}, \widehat{\mathsf{M}}) \le \overline{\delta}(\epsilon, \alpha, \Delta + 2err(\delta)) + 2err(\delta) + o_n(1)$, where $\overline{\delta}(\epsilon, \alpha, \Delta)$ is defined as in Eq. (7) and $err(\delta)$ is the quantization error which only depends on δ.*

Proof. Let M^δ be the quantized version of the original matrix M, which is defined as follows. Define $\vec{u}_i^\delta \in (A_\delta)^r$ and $\vec{v}_a^\delta \in (A_\delta)^r$ to be the quantized version of $\vec{u}_i$ and $\vec{v}_a$ respectively, where $\vec{u}_i$ is the i-th row of U and $\vec{v}_a$ is the a-th column V. Then, M^δ is defined as,

$$\mathsf{M}^\delta_{i,a} = \vec{u}_i^\delta \cdot \vec{v}_a^\delta .$$

Note that $\mathsf{M}^\delta_{i,a}$ satisfies $|\mathsf{M}_{i,a} - \mathsf{M}^\delta_{i,a}| \le err(\delta)$. Analogously, define $\widehat{\mathsf{M}}^\delta$ to be the quantized version of the estimated matrix $\widehat{\mathsf{M}}$. Then, the M^δ and $\widehat{\mathsf{M}}^\delta$ satisfy $|\widehat{\mathsf{M}}^\delta_{i,a} - \mathsf{M}^\delta_{i,a}| \le \Delta + 2err(\delta)$ for all $(i, a) \in E$.

Let $\overline{\delta}(\epsilon, \alpha, \Delta)$ be the upper bound in proposition III.1. Then, the distortion is bounded with high probability by

$$\begin{aligned} D(\mathsf{M}, \widehat{\mathsf{M}}) &\le D(\mathsf{M}, \mathsf{M}^\delta) + D(M^\delta, \widehat{\mathsf{M}}^\delta) + D(\widehat{\mathsf{M}}^\delta, \widehat{\mathsf{M}}) \\ &\le \overline{\delta}(\epsilon, \alpha, \Delta + 2err(\delta)) + 2err(\delta) \ . \end{aligned} \tag{13}$$

Note that twice the quantization error is added to Δ since now we only have $|\widehat{\mathsf{M}}^\delta_{i,a} - \mathsf{M}^\delta_{i,a}| \le \Delta + 2err(\delta)$ for all $(i, a) \in E$. □

C. Simplified bound

The (tighter) upper bound in proposition III.1 is not easily computed. To get a bound that can be analyzed, we relax the constraint $\phi_\Delta \ge 0$ and get a relaxed or simplified upper bound on $\overline{\delta}(\epsilon, \alpha, \Delta)$. Furthermore, this simplified upper bound is used to prove theorem I.1.

Proposition III.4. *For all $\epsilon \ge 0$, $\alpha \ge 0$ and $\Delta \ge 0$, we have*

$$\begin{aligned} &\overline{\delta}(\epsilon, \alpha, \Delta) \le \\ &\left\{ \overline{d}^2 - (\overline{d}^2 - \Delta^2) \exp\left(-\frac{\overline{H}(p|p_0) + \alpha\overline{H}(q|q_0)}{\epsilon} \right) \right\}^{1/2} , \end{aligned}$$

where $\overline{\delta}(\epsilon, \alpha, \Delta)$ is defined as in proposition III.1, $\overline{H}(p|p_0) = \max_{p\in\mathcal{D}(p_0)}\{H(p)\} - H(p_0)$, $\overline{H}(q|q_0) = \max_{q\in\mathcal{D}(q_0)}\{H(q)\} - H(q_0)$, and $\overline{d} = max\{|\vec{u} \cdot \vec{v} - \vec{u}^0 \cdot \vec{v}^0|\}$.

Proof. Define the upper bound $\overline{\delta}^u(\epsilon, \alpha, \Delta)$ as

$$\overline{\delta}^u(\epsilon, \alpha, \Delta) = \sup_{p\in\mathcal{D}(p_0),q\in\mathcal{D}(q_0)} \{\, d(p, q) \ : \ \phi^u_\Delta(p, q) \ \ge 0\} \ , \tag{14}$$

where $\mathcal{D}(p_0)$, $\mathcal{D}(p_0)$ and $d(p, q)$ are defined in Eq. (7). The only difference is the relaxed constraint function ϕ^u_Δ, defined as

$$\phi^u_\Delta(p, q) \equiv \overline{H}(p|p_0) + \alpha\overline{H}(q|q_0) + \epsilon \log\left(\frac{\overline{d}^2 - d(p, q)^2}{\overline{d}^2 - \Delta^2} \right) .$$

By Jensen's and Markov inequality, $\phi^u_\Delta(p, q)$ is larger than $\phi_\Delta(p, q)$. This implies that the supremum in the simplified upper bound is taken over a larger set of distributions than the tighter upper bound, hence we have $\overline{\delta}(\epsilon, \alpha, \Delta) \le \overline{\delta}^u(\epsilon, \alpha, \Delta)$. And after some computation, it's easy to show that $\overline{\delta}^u(\epsilon,\alpha,\Delta) = \left\{\overline{d}^2 - (\overline{d}^2 - \Delta^2)\exp\left(-\frac{1}{\epsilon}\left[\overline{H}(p|p_0) + \alpha\overline{H}(q|q_0)\right]\right)\right\}^{1/2}$, which concludes the proof. □

This simplified upper bound can be generalized, in the same manner, to the continuous support case. The following example illustrates this generalization and introduces bounds necessary in the proof of theorem I.1.

For the original matrix $\mathsf{M} = \mathsf{U}\cdot\mathsf{V}$, assume the distributions of $\mathsf{U}_{i,k}$ and $\mathsf{V}_{k,a}$ to have support in $A_\delta = \{-1, -1+\delta, \ldots, 1-\delta, 1\}$. Also, the factors of the rank-r solution $\widehat{\mathsf{M}}$ are supported on the same discrete set. Then, the simplified upper bound is given by

$$\overline{\delta}^u(\epsilon,\alpha,\Delta) = \left(\Delta^2 + (4r^2 - \Delta^2)\left(1 - \exp\left\{-\frac{\log N}{\widetilde{\epsilon}}\right\}\right)\right)^{1/2},$$

where $N = |A_\delta|$ and $\widetilde{\epsilon} \equiv \epsilon/(1+\alpha)r$. Note that $\lim_{\epsilon\to\infty}\overline{\delta}^u(\epsilon,\alpha,\Delta) = \Delta$, which means that we cannot get RMSE smaller than Δ.

The maximum quantization error associated with $M_{i,a}$ is $r(\delta - \delta^2/4)$, which happens when all the entries of $\vec{u}_i^0$ and $\vec{v}_a^0$ are $1-\delta/2$ and quantized to 1. For simplicity, $err(\delta) = r\delta$ is used. Combined with Eq. (13), we have a simple analytical upper bound on the distortion when the original matrix and the estimation have continuous support $[-1, 1]$.

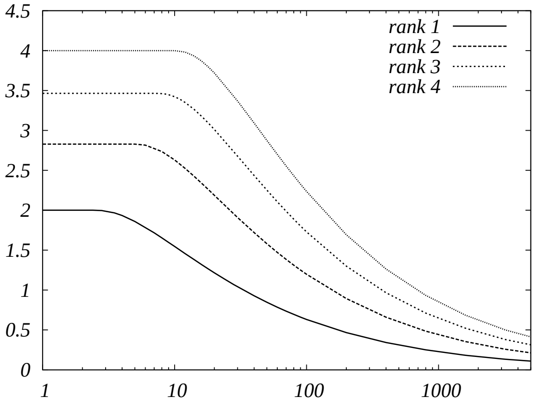

Fig. 2. The upper bound in Eq. (13) with simplified upper bound $\overline{\delta}^u(\epsilon\alpha\Delta)$, for $\alpha = 1$ and $\Delta = 0$ and a few values of the rank r.

Proof of Theorem I.1. From the example above, we can compute the simplified upper bound directly to bound the distortion.

$$\begin{aligned}
&D(\mathsf{M},\widehat{\mathsf{M}}) \\
&\le \left\{4r^2 - (4r^2 - (\Delta+2r\delta)^2)\left(\exp\left(-\frac{\log N}{\widetilde{\epsilon}}\right)\right)\right\}^{1/2} + 2r\delta \\
&\le \left\{(\Delta+2r\delta)^2 + 4r^2\left(1 - \exp\left(-\frac{\log N}{\widetilde{\epsilon}}\right)\right)\right\}^{\frac{1}{2}} + 2r\delta \\
&\le \Delta + 4r\delta + 2r\left(1 - \exp\left(-\frac{\log N}{\widetilde{\epsilon}}\right)\right)^{\frac{1}{2}} \\
&\le \Delta + 4r\delta + 2r\left(\frac{\log N}{\widetilde{\epsilon}}\right)^{\frac{1}{2}} .
\end{aligned}$$

Remember N is defined as the alphabet size $|A_\delta|$, where the discrete alphabet $A_\delta = \{-1, -1+\delta, \cdots, 1-\delta, 1\}$ is used. Fixing $\delta = \frac{2}{N-1}$, we can minimize the right hand side of the last inequality with respect to the alphabet size N. Since the exact minimizer cannot be represented in a closed form, we use instead an approximate minimizer $N = \left\lceil 4\sqrt{\widetilde{\epsilon}}\right\rceil + 1$, which results in

$$\begin{aligned}
&D(\mathsf{M},\widehat{\mathsf{M}}) \\
&\le \Delta + 2r\left\{\frac{4}{\left\lceil 4\sqrt{\widetilde{\epsilon}}\right\rceil} + \left(\frac{\log\left(\left\lceil 4\sqrt{\widetilde{\epsilon}}\right\rceil + 1\right)}{\widetilde{\epsilon}}\right)^{\frac{1}{2}}\right\} \\
&\le \Delta + \frac{2r}{\sqrt{\widetilde{\epsilon}}}\left\{1 + \left(\log\left(\left\lceil 4\sqrt{\widetilde{\epsilon}}\right\rceil + 1\right)\right)^{\frac{1}{2}}\right\} \\
&\le \Delta + \frac{2r}{\sqrt{\widetilde{\epsilon}}}\log(10\widetilde{\epsilon}) \ , \qquad (15)
\end{aligned}$$

where the last inequality in (15) is true for $\widetilde{\epsilon} > 1.5$. This is practical since we are typically interested in the region where $\frac{\log(10\widetilde{\epsilon})}{\sqrt{\widetilde{\epsilon}}} \le 1$.

□

IV. Lower Bound

When the number of observed elements is smaller than $\Theta(n)$, high distortion is inevitable. In this section we derive a quantitative lower bound which supports this observation.

Proposition IV.1. *Let $\mathsf{M} = \mathsf{U}\cdot\mathsf{V}$ be a random rank-r matrix with n rows and $n\alpha$ columns and assume the distributions of $\mathsf{U}_{i,k}$ and $\mathsf{V}_{k,a}$ to have support in $[-1, 1]$, and E a random subset of $n\epsilon$ row-column pairs. Then, with high probability, any rank-r matrix $\widehat{\mathsf{M}}$ such that $|\mathsf{M}_{i,a} - \widehat{\mathsf{M}}_{i,a}| = 0$ for all $(i,a) \in E$, also satisfies*

$$D(\mathsf{M},\widehat{\mathsf{M}}) \ge \tilde{c}\cdot e^{-\epsilon} \ , \qquad (16)$$

where $\tilde{c}$ is a strictly positive constant that only depends on the rank r and the initial distributions p_0 and q_0.

Proof. Think of the following algorithm which has clearly better performance than any other that satisfies the assumptions. Consider the bipartite graph $G = (R, C, E)$ with vertices corresponding to the row and columns of M and edges for the observed entries. For every pair of row and column indices (i,a), $i \in R$ and $a \in C$, that is not connected by an edge, we do the following. If degree of i (a) is less than r, we assume that all the neighbors of node i (a) is known and make MMSE estimation of $\vec{u}_i^0$ $(\vec{v}_a^0)$. If degree of i (a) is greater than $r-1$, we assign the correct value of $\vec{u}_i^0$ $(\vec{v}_a^0)$. With high probability the resulting RMSE is greater than $\underline{\delta}(\epsilon,\alpha)$ as defined below.

$$\underline{\delta}(\epsilon,\alpha) = \sqrt{(1-(1-\xi)(1-\zeta))\tilde{c}}\ , \qquad (17)$$

where $\xi = \mathbb{P}\{degree(i) < r\} = \sum_{k=0}^{r-1} \frac{\epsilon^{-k}}{k!} e^{-\epsilon}$, $\zeta = \mathbb{P}\{degree(a) < r\} = \sum_{k=0}^{r-1} \frac{(\epsilon/\alpha)^{-k}}{k!} e^{-\epsilon/\alpha}$ and $\tilde{c} = \min\{\mathbb{E}\{\vec{u}_i^0 \cdot (\vec{v}_a^0 - \vec{v}'_a)\}, \mathbb{E}\{(\vec{u}_i^0 - \vec{u}'_i) \cdot \vec{v}_a^0\}\}$. Here, $\vec{u}'_i$ and $\vec{v}'_a$ represent the MMSE estimate of $\vec{u}_i^0$ and $\vec{v}_a^0$ respectively, assuming that $r-1$ neighbors and corresponding edges are known.

Without loss of generality, assume $\alpha \geq 1$. Then, we can simplify above bound to get, Eq. (16) □

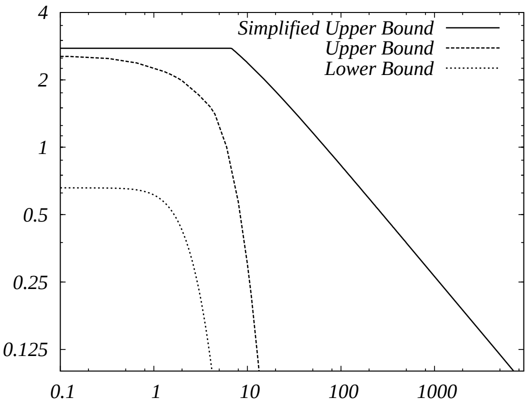

Fig. 3. The upper bound $\overline{\delta}(\epsilon,\alpha,\Delta)$, the simplified upper bound $\overline{\delta}^u(\epsilon,\alpha,\Delta)$ and the lower bound $\underline{\delta}(\epsilon,\alpha)$ for rank $r=2$, $\alpha=1$, $\Delta=0$. Here the factors U_{ik}, V_{ka} take values in $\{-1,0,1\}$.

V. Efficient matrix completion

In the previous sections we proved that $O(n)$ random entries determine a random low rank matrix within an arbitrarily small RMSE. How hard is it to find such a matrix? In this section we present a numerical investigation using a low complexity stochastic local search algorithm that we called WalkRank.

WalkRank is inspired to successful local search algorithms for constraint satisfaction problem, such as WalkSAT [8]. It is particularly suited to low-rank matrices whose factors $\mathsf{U}_{i,k}$, $\mathsf{V}_{k,a}$ take values in a finite set A_N. The algorithm tries to find assignments of the vectors $\{\vec{u}_1,\dots,\vec{u}_n\}$, and $\{\vec{v}_1,\dots,\vec{v}_m\}$ that minimize the cost function

$$\mathcal{C}(\{\vec{u}_i,\vec{v}_a\}) = \sum_{(i,a)\in E} \mathbb{I}(|\vec{u}_i \cdot \vec{u}_a - \mathsf{M}_{ia}| > \Delta)\,, \qquad (18)$$

which counts the number of observations M_{ia} that are not described by the current assignment.

The algorithm initializes the vectors $\{\vec{u}_i\}$, $\{\vec{v}_a\}$ to random iid values and then alternates between two type of moves. The first are greedy moves, described here in the case of U factors.

```
Greedy move, U factors
1: Sample a column index $i \in C$ uniformly;
2: Find $\vec{u}_i^{\text{new}}$ that minimizes $\mathcal{C}(\{\vec{u}_i,\vec{v}_a\})$ over $\vec{u}_i$;
3: Set $\vec{u}_i \leftarrow \vec{u}_i^{\text{new}}$
```

Greedy moves for V factors are defined analogously.

The second type of move potentially increases the cost function.

```
Walk move
1: Sample $(i,a) \in E$ s.t. $|\vec{u}_i \cdot \vec{v}_a - \mathsf{M}_{ia}| > \Delta$;
2: Find $\vec{u}_i^{\text{new}} \cdot \vec{v}_a^{\text{new}}$ such that $|\vec{u}_i^{\text{new}} \cdot \vec{v}_a^{\text{new}} - \mathsf{M}_{ia}| \leq \Delta$
3: Set $\vec{u}_i \leftarrow \vec{u}_i^{\text{new}}$, and $\vec{v}_a \leftarrow \vec{v}_a^{\text{new}}$
```

WalkRank recursively executes one of these moves, choosing a walk move with probability ρ, and a greedy one with probability $1-\rho$. The parameter ρ can be optimized over, and we found $\rho \approx 0.1$ to be a reasonable choice.

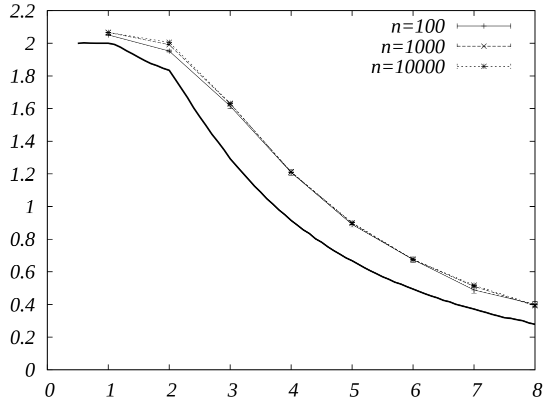

Fig. 4. Performances of the WalkRank algorithm on random rank 2 matrices. The bold line is a lower bound on the distortion obtained by the maximum likelihood algorithm.

In Figures 4 to 6 we present the distortion achieved by the WalkRank algorithm, averaged over 10 instances. We used factors with entries $\mathsf{U}_{i,k}$, $\mathsf{V}_{k,a}$ uniformly distributed in $\{+1,-1\}$. It is clear that the resulting distortion is essentially independent of n over two orders of magnitude and decreases rapidly with ϵ.

We compare these numerical results with an analytical lower bound on the distortion achieved by a maximum likelihood algorithm. The latter fills each unknown position in M with its most likely value. While there exists no practical implementation of the maximum likelihood rule, we can provide a sharp lower bound on its performances using techniques explained in [9]. It appears that, for low values of the rank, WalkRank achieves the same distortion as maximum likelihood, provided it is given one or two more entries per column/row.

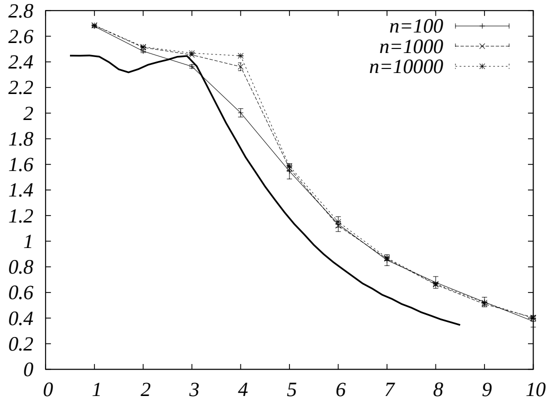

Fig. 5. Performances of the WalkRank algorithm on random rank 3 matrices.

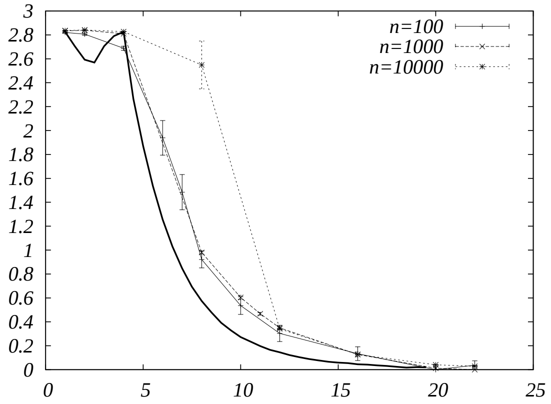

Fig. 6. Performances of the WalkRank algorithm on random rank 4 matrices.

The complexity of one WalkRank step is independent of the matrix size (but grows with the rank). The results in Figures 4 to 6 were obtained with a number of steps slightly superlinear in n. In Fig. 7 we show the evolution of the cost function for averaged over 10 instances for $n = 10^3$ to 10^5, $r = 3$ and $\epsilon = 8$. The number of steps per variable required to reach the asymptotic value increases mildly with n. A reasonable conjecture is that the number of steps scales like $n{\cdot}\text{Poly}(\log n)$.

VI. Back to the Netflix Data

As shown in the last section, local search algorithms efficiently fit low rank matrices of very large dimensions, using few observations. They therefore provide an efficient tool for checking whether a dataset is well described by the random low rank model.

In Figures 8 and 9 we compare the evolution of fit and prediction error for three matrices with $n = m = 5 \cdot 10^3$:

1. A submatrix of the Netflix dataset given by the first $5 \cdot 10^3$ movies and customers.
2. A matrix with the same subset E of revealed entries, each of them chosen uniformly at random in $[-1, +1]$.

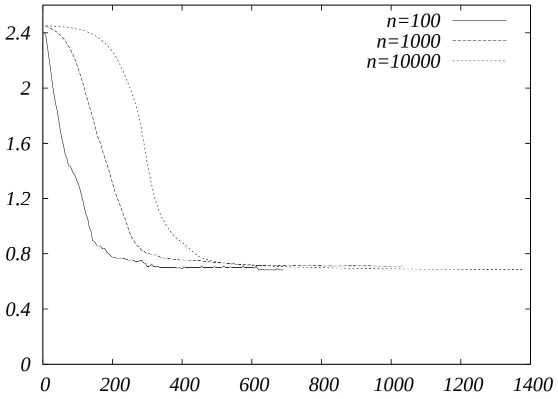

Fig. 7. Typical evolution of the cost function under the WalkRank algorithm. Here the rank is $r = 3$, and $\epsilon = 8$.

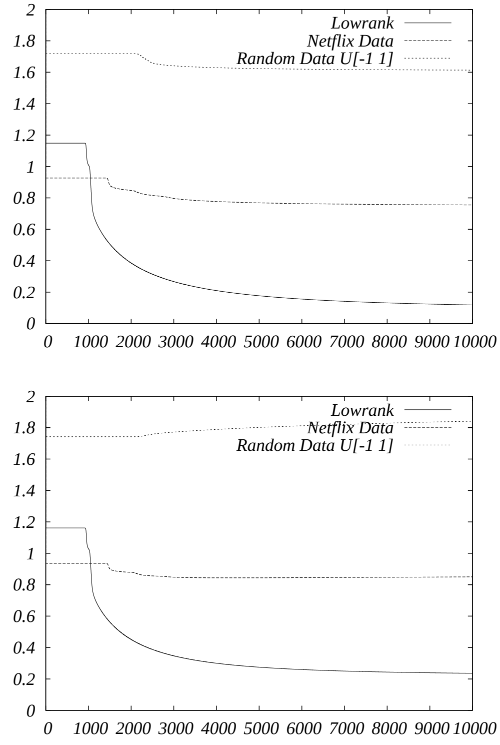

Fig. 8. Evolution of the fit error (top frame) and prediction error (lower frame) for fitting three matrices with a rank 3 model. The curves are obtained using coordinate descent in the factors.

3. A random rank-3 matrix (for Fig. 8) or rank-5 matrix (for Fig. 9), with set of revealed entries as above.

The fit error is defined by restricting the average in Eq. (1) to $(i, a) \in E$. The prediction error is instead obtained by averaging over $(i, a) \notin E$. In the case of the Netflix matrix the latter was estimated by hiding 10^3 entries from the dataset, and averaging over those.

We used a coordinate descent algorithm in the factors $\{\vec{u}_i\}$, $\{\vec{v}_a\}$, with regularized cost function given by Eq. (2). In agreement with the results of previous sections, random low rank matrices are efficiently fitted with small fitting *and*

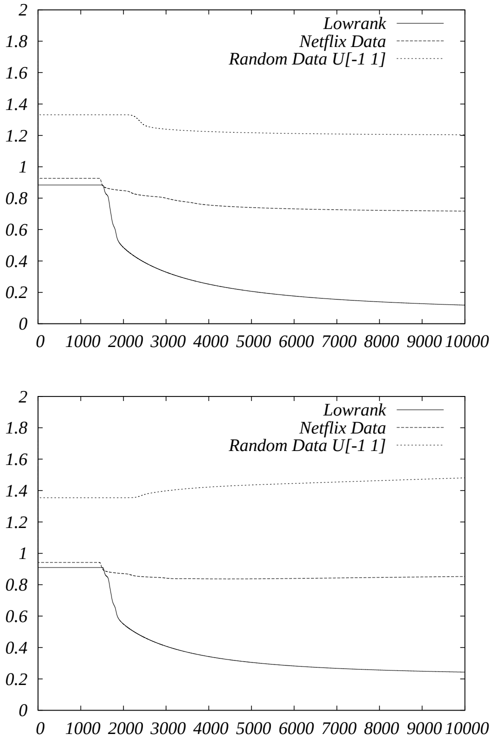

Fig. 9. As in Figure 8, but for a rank 5 model.

prediction error. The difference with iid entries is striking. The fit error decreases only slowly over time, while the prediction error actually increases. As expected, revealed entries do not provide any information on the hidden ones. Netflix data lie somewhat in between: both fit and prediction error decrease over time, albeit not as sharply as for genuine low rank matrices.

Acknowledgements

Andrea Montanari is partially supported by a Terman fellowship and by the NSF CAREER grant 0743978.

References

[1] I. M. Johnstone “High Dimensional Statistical Inference and Random Matrices”, Proc. Intl. Congr. of Math., Madrid, July 2006

[2] A. M.-C. So and Y. Ye, “Theory of semidefinite programming for sensor network localization”, Math. Progr. Series B, 109 (2007), 267-385

[3] “Netflix prize”, `http://www.netflixprize.com/`

[4] B. Recht, M. Fazel, and P. A. Parrilo, “Guaranteed minimum rank solutions to linear matrix equations via nuclear norm minimization”, preprint (2007), submitted to SIAM Review.

[5] E. J. Candés and B. Recht, “Exact Matrix Completion via Convex Optimization”, preprint (2008), available at `http://www.acm.caltech.edu/~emmanuel/papers/MatrixCompletion.pdf`.

[6] Z. Liu and L. Vandenberghe, “Interior-point method for nuclear norm approximation with application to system identification”, preprint (2008), available at `http://www.ee.ucla.edu/~vandenbe/publications/nucnrm.pdf`.

[7] T. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley Interscience, New York, 1991

[8] B. Selman, H. A. Kautz, and B. Cohen “Noise strategies for improving local search,” in *Proc. of AAAI-94*, Seattle, WA.

[9] R. H. Keshavan, A. Montanari and S. Oh, “Learning low-rank matrices from $O(n)$ observations: Algorithms and phase transitions”, *in preparation.*