8
Matrix Completion

In earlier chapters, we saw the power of sparsity. It's possible to recover a sparse vector from many fewer measurements than its dimension. And if we don't know the basis where our vectors are sparse, with enough examples we can learn it. But sparsity is just the beginning. There are many other ways to make the objects we are working with be low-complexity. In this chapter, we will study the matrix completion problem, where the goal is to reconstruct a matrix even when we observe just a few of its entries. Without any assumptions on the matrix, this is impossible, because there are just too many degrees of freedom. But when the matrix is low-rank and incoherent, it turns out that there are simple convex programs that work. You can take these ideas much further and study all sorts of structured recovery problems via convex programs, such as decomposing a matrix into the sum of a sparse matrix and a low-rank matrix. We won't get to these here, but will give pointers to the literature.

8.1 Introduction

In 2006, Netflix issued a grand challenge to the machine learning community: Beat our prediction algorithms for recommending movies to users by more than 10 percent, and we'll give you a million dollars. It took a few years, but eventually the challenge was won and Netflix paid out. During that time, we all learned a lot about how to build good recommendation systems. In this chapter, we will cover one of the main ingredients, which is called the matrix completion problem.

The starting point is to model our problem of predicting movie ratings as a problem of predicting the unobserved entries of a matrix from the ones we do observe. More precisely, if user i rates movie j (from one to five stars), we set $M_{i,j}$ to be the numerical score. Our goal is to use the entries $M_{i,j}$ that

https://doi.org/10.1017/9781316882177.009 Published online by Cambridge University Press

we observe to predict the ones that we don't know. If we could predict these accurately, it would give us a way to suggest movies to users in a way that we are suggesting movies that we think they might like. A priori, there's no reason to believe you can do this. If we think about the entire matrix M that we would get by coercing every user to rate every movie (and in the Netflix dataset there are $480,189$ users and $17,770$ movies), then in principle the entries $M_{i,j}$ that we observe might tell us nothing about the unobserved entries.

We're in the same conundrum we were in when we talked about compressed sensing. A priori, there is no reason to believe you can take fewer linear measurements of a vector x than its dimension and reconstruct x. What we need is some assumption about the structure. In compressed sensing, we assumed that x is sparse or approximately sparse. In matrix completion, we will assume that M is low-rank or approximately low-rank. It's important to think about where this assumption comes from. If M were low-rank, we could write it as

$$M = u^{(1)}(v^{(1)})^T + u^{(2)}(v^{(2)})^T \dots u^{(r)}(v^{(r)})^T.$$

The hope is that each of these rank-one terms represents some category of movies. For example, the first term might represent the category *drama*, and the entries in $u^{(1)}$ might represent for every user, to what extent does he or she like drama movies? Then each entry in $v^{(1)}$ would represent for every movie, to what extent would it appeal to someone who likes drama? This is where the low-rank assumption comes from. What we're hoping is that there are some categories underlying our data that make it possible to fill in missing entries. When I have a user's ratings for movies in each of the categories, I could then recommend other movies in the category that he or she likes by leveraging the data I have from other users.

The Model and Main Results

Now let's be formal. Suppose there are n users and m movies so that M is an $n \times m$ matrix. Let $\Omega \subseteq [n] \times [m]$ be the indices where we observe the value $M_{i,j}$. Our goal is, under the assumption that M is low-rank or approximately low-rank, to fill in the missing entries. The trouble is that in this level of generality, finding the matrix M of lowest rank that agrees with our observations is *NP*-hard. However, there are some by now standard assumptions under which we will be able to give efficient algorithms for recovering M exactly:

(a) The entries we observe are chosen uniformly at random from $[n] \times [m]$.
(b) M has rank r.
(c) The singular vectors of M are uncorrelated with the standard basis (such a matrix is called *incoherent* and we define this later).

https://doi.org/10.1017/9781316882177.009 Published online by Cambridge University Press

In this chapter, our main result is that there are efficient algorithms for recovering M exactly if $m \approx mr \log m$ where $m \geq n$ and $\text{rank}(M) \leq r$. This is similar to compressed sensing, where we were able to recover a k-sparse signal x from $O(k \log n/k)$ linear measurements, which is much smaller than the dimension of x. Here too we can recover a low-rank matrix M from a number of observations that is much smaller than the dimension of M.

Let us examine the assumptions above. The assumption that should give us pause is that Ω is uniformly random. This is somewhat unnatural, since it would be more believable if the probability that we observe $M_{i,j}$ depended on the value itself. Alternatively, a user should be more likely to rate a movie *if he or she actually liked it*.

We already discussed the second assumption. In order to understand the third assumption, suppose our observations are indeed uniformly random. Consider

$$M = \Pi \left[\begin{array}{c|c} I_r & 0 \\ \hline 0 & 0 \end{array}\right] \Pi^T$$

where Π is a uniformly random permutation matrix. M is low-rank, but unless we observe all of the ones along the diagonal, we will not be able to recover M uniquely. Indeed, the top singular vectors of M are standard basis vectors. But if we were to assume that the singular vectors of M are incoherent with respect to the standard basis, we would avoid this snag, because the vectors in our low-rank decomposition of M are spread out over many rows and columns.

Definition 8.1.1 *The coherence μ of a subspace $U \subseteq \mathbb{R}^n$ of dimension $dim(u) = r$ is*

$$\frac{n}{r} \max_i \|P_U e_i\|^2$$

where P_U denotes the orthogonal projection onto U and e_i is the standard basis element.

It is easy to see that if we choose U uniformly at random, then $\mu(U) = \widetilde{O}(1)$. Also we have that $1 \leq \mu(U) \leq n/r$ and the upper bound is attained if U contains any e_i. We can now see that if we set U to be the top singular vectors of the above example, then U has high coherence. We will need the following conditions on M:

(a) Let $M = U\Sigma V^T$, then $\mu(U), \mu(V) \leq \mu_0$.
(b) $\|UV^T\|_\infty \leq \frac{\mu_1\sqrt{r}}{\sqrt{nm}}$, where $||\cdot||_\infty$ denotes the maximum absolute value of any entry.

The main result of this chapter is:

https://doi.org/10.1017/9781316882177.009 Published online by Cambridge University Press

Theorem 8.1.2 *Suppose Ω is chosen uniformly at random. Then there is a polynomial time algorithm to recover M exactly that succeeds with high probability if*

$$|\Omega| \geq C \max(\mu_1^2, \mu_0) r(n+m) \log^2(n+m).$$

The algorithm in the theorem above is based on a convex relaxation for the rank of a matrix called the *nuclear norm*. We will introduce this in the next section and establish some of its properties, but one can think of it as an analogue to the ℓ_1 minimization approach that we used in compressed sensing. This approach was first introduced in Fazel's thesis [70], and Recht, Fazel, and Parrilo [124] proved that this approach exactly recovers M in the setting of *matrix sensing*, which is related to the problem we consider here.

In a landmark paper, Candes and Recht [41] proved that the relaxation based on nuclear norm also succeeds for matrix completion and introduced the assumptions above in order to prove that their algorithm works. There has since been a long line of work improving the requirements on m, and the theorem above and our exposition will follow a recent paper of Recht [123] that greatly simplifies the analysis by making use of matrix analogues of the Bernstein bound and using these in a procedure now called *quantum golfing* that was first introduced by Gross [80].

Remark 8.1.3 *We will restrict to $M \in \mathbb{R}^{n \times n}$ and assume $\mu_0, \mu_1 = \widetilde{O}(1)$ in our analysis, which will reduce the number of parameters we need to keep track of.*

8.2 Nuclear Norm

Here we introduce the nuclear norm, which will be the basis for our algorithms for matrix completion. We will follow an outline parallel to that of compressed sensing. In particular, a natural starting point is the optimization problem:

$$(P_0) \qquad \min \operatorname{rank}(X) \text{ s.t. } X_{i,j} = M_{i,j} \text{ for all } (i,j) \in \Omega$$

This optimization problem is *NP*-hard. If $\sigma(X)$ is the vector of singular values of X, then we can think of the rank of X equivalently as the sparsity of $\sigma(X)$. Recall, in compressed sensing we faced a similar obstacle: finding the sparsest solution to a system of linear equations is also *NP*-hard. But instead we considered the ℓ_1 relaxation and proved that under various conditions, this optimization problem recovers the sparsest solution. Similarly, it is natural to consider the ℓ_1-norm of $\sigma(X)$, which is called the nuclear norm:

Definition 8.2.1 *The nuclear norm of X denoted by $\|X\|_*$ is $\|\sigma(X)\|_1$.*

https://doi.org/10.1017/9781316882177.009 Published online by Cambridge University Press

We will instead solve the convex program:

$$(P_1) \qquad \min \|X\|_* \text{ s.t. } X_{i,j} = M_{i,j} \text{ for all } (i,j) \in \Omega$$

and our goal is to prove conditions under which the solution to (P_1) is exactly M. Note that this is a convex program because $\|X\|_*$ is a norm, and there are a variety of efficient algorithms to solve the above program.

In fact, for our purposes, a crucial notion is that of a *dual norm*. We will not need this concept in full generality, so we state it for the specific case of the nuclear norm. This concept gives us a method to lower-bound the nuclear norm of a matrix:

Definition 8.2.2 *Let* $\langle X, B\rangle = \sum_{i,j} X_{i,j}B_{i,j} = trace(X^T B)$ *denote the matrix inner product.*

Lemma 8.2.3 $\|X\|_* = \max_{\|B\|\leq 1} \langle X, B\rangle$

To get a feel for this, consider the special case where we restrict X and B to be diagonal. Moreover, let $X = \text{diag}(x)$ and $B = \text{diag}(b)$. Then $\|X\|_* = \|x\|_1$ and the constraint $\|B\| \leq 1$ (the spectral norm of B is at most one) is equivalent to $\|b\|_\infty \leq 1$. So we can recover a more familiar characterization of vector norms in the special case of diagonal matrices:

$$\|x\|_1 = \max_{\|b\|_\infty \leq 1} b^T x$$

Proof: We will only prove one direction of the above lemma. What B should we use to certify the nuclear norm of X? Let $X = U_X \Sigma_X V_X^T$, then we will choose $B = U_X V_X^T$. Then

$$\begin{aligned}\langle X, B\rangle &= \text{trace}(B^T X) = \text{trace}(V_X U_X^T U_X \Sigma_X V_X^T) \\ &= \text{trace}(V_X \Sigma_X V_X^T) = \text{trace}(\Sigma_X) = \|X\|_*\end{aligned}$$

where we have used the basic fact that $\text{trace}(ABC) = \text{trace}(BCA)$. Hence this proves $\|X\|_* \leq \max_{\|B\|\leq 1} \langle X, B\rangle$, and the other direction is not much more difficult (see, e.g., [88]). ■

How can we show that the solution to (P_1) is M? Our basic approach will be a proof by contradiction. Suppose not; then the solution is $M + Z$ for some Z that is supported in $\overline{\Omega}$. Our goal will be to construct a matrix B of spectral norm at most one for which

$$\|M + Z\|_* \geq \langle M + Z, B\rangle > \|M\|_*.$$

Hence $M+Z$ would not be the optimal solution to (P_1). This strategy is similar to the one in compressed sensing, where we hypothesized some other solution

https://doi.org/10.1017/9781316882177.009 Published online by Cambridge University Press

w that differs from x by a vector y in the kernel of the sensing matrix A. There, our strategy was to use geometric properties of $\ker(A)$ to prove that w has strictly larger ℓ_1 norm than x. The proof here will be in the same spirit, but considerably more technical and involved.

Let us introduce some basic projection operators that will be crucial in our proof. Recall, $M = U\Sigma V^T$, let $u_1, \ldots, u_r$ be columns of U, and let $v_1, \ldots, v_r$ be columns of V. Choose $u_{r+1}, \ldots, u_n$ so that $u_1, \ldots, u_n$ form an orthonormal basis for all of $\mathbb{R}^n$; i.e., $u_{r+1}, \ldots, u_n$ is an arbitrary orthonormal basis of $U^\perp$. Similarly, choose $v_{r+1}, \ldots, v_n$ so that $v_1, \ldots, v_n$ form an orthonormal basis for all of $\mathbb{R}^n$. We will be interested in the following linear spaces over matrices:

Definition 8.2.4 $T = span\{u_i v_j^T \mid 1 \le i \le r \text{ or } 1 \le j \le r \text{ or both}\}$

Then $T^\perp = \text{span}\{u_i v_j^T \text{ s.t. } r+1 \le i,j \le n\}$. We have $\dim(T) = r^2 + 2(n-r)r$ and $\dim(T^\perp) = (n-r)^2$. Moreover, we can define the linear operators that project into T and $T^\perp$, respectively:

$$P_{T^\perp}[Z] = \sum_{i=r+1}^{n} \sum_{j=r+1}^{n} \langle Z, u_i v_j^T \rangle \cdot u_i v_j^T = P_{U^\perp} Z P_{V^\perp}.$$

And similarly,

$$P_T[Z] = \sum_{(i,j) \in [n]\times[n] - [r+1,n]\times[r+1,n]} \langle Z, u_i v_j^T \rangle \cdot u_i v_j^T = P_U Z + Z P_V - P_U Z P_V.$$

We are now ready to describe the outline of the proof of Theorem 8.1.2. The proof will be based on the following:

(a) We will assume that a certain helper matrix Y exists, and show that this is enough to imply $\|M + Z\|_* > \|M\|_*$ for any Z supported in Ω.
(b) We will construct such a Y using quantum golfing [80].

Conditions for Exact Recovery

Here we will state the conditions we need on the helper matrix Y and prove that if such a Y exists, then M is the solution to (P_1). We require that Y is supported in Ω and

(a) $\|P_T(Y) - UV^T\|_F \le \sqrt{r/8n}$
(b) $\|P_{T^\perp}(Y)\| \le 1/2$.

We want to prove that for any Z supported in $\overline{\Omega}$, $\|M+Z\|_* > \|M\|_*$. Recall, we want to find a matrix B of spectral norm at most one so that $\langle M + Z, B\rangle > \|M\|_*$. Let $U_\perp$ and $V_\perp$ be singular vectors of $P_{T^\perp}[Z]$. Then consider

https://doi.org/10.1017/9781316882177.009 Published online by Cambridge University Press

$$B = \begin{bmatrix} U & U_\perp \end{bmatrix} \cdot \begin{bmatrix} V^T \\ V_\perp^T \end{bmatrix} = UV^T + U_\perp V_\perp^T.$$

Claim 8.2.5 $\|B\| \leq 1$

Proof: By construction, $U^T U_\perp = 0$ and $V^T V_\perp = 0$, and hence the above expression for B is its singular value decomposition, and the claim now follows. ■

Hence we can plug in our choice for B and simplify:

$$\begin{aligned} \|M + Z\|_* &\geq \langle M + Z, B \rangle \\ &= \langle M + Z, UV^T + U_\perp V_\perp^T \rangle \\ &= \underbrace{\langle M, UV^T \rangle}_{\|M\|_*} + \langle Z, UV^T + U_\perp V_\perp^T \rangle \end{aligned}$$

where in the last line we use the fact that M is orthogonal to $U_\perp V_\perp^T$. Now, using the fact that Y and Z have disjoint supports, we can conclude:

$$\|M + Z\|_* \geq \|M\|_* + \langle Z, UV^T + U_\perp V_\perp^T - Y \rangle$$

Therefore, in order to prove the main result in this section, it suffices to prove that $\langle Z, UV^T + U_\perp V_\perp^T - Y \rangle > 0$. We can expand this quantity in terms of its projection onto T and $T^\perp$ and simplify as follows:

$$\begin{aligned} \|M + Z\|_* - \|M\|_* &\geq \langle P_T(Z), P_T(UV^T + U_\perp V_\perp^T - Y) \rangle \\ &\quad + \langle P_{T^\perp}(Z), P_{T^\perp}(UV^T + U_\perp V_\perp^T - Y) \rangle \\ &\geq \langle P_T(Z), UV^T - P_T(Y) \rangle + \langle P_{T^\perp}(Z), U_\perp V_\perp^T - P_{T^\perp}(Y) \rangle \\ &\geq \langle P_T(Z), UV^T - P_T(Y) \rangle + \|P_{T^\perp}(Z)\|_* - \langle P_{T^\perp}(Z), P_{T^\perp}(Y) \rangle \end{aligned}$$

where in the last line we used the fact that $U_\perp$ and $V_\perp$ are the singular vectors of $P_{T^\perp}[Z]$, and hence $\langle U_\perp V_\perp^T, P_{T^\perp}[Z] \rangle = \|P_{T^\perp}[Z]\|_*$.

Now we can invoke the properties of Y that we have assumed in this section, to prove a lower bound on the right-hand side. By property (a) of Y, we have that $\|P_T(Y) - UV^T\|_F \leq \sqrt{\frac{r}{2n}}$. Therefore, we know that the first term $\langle P_T(Z), UV^T - P_T(Y) \rangle \geq -\sqrt{\frac{r}{8n}} \|P_T(Z)\|_F$. By property (b) of Y, we know the operator norm of $P_T^\perp(Y)$ is at most 1/2. Therefore, the third term $\langle P_{T^\perp}(Z), P_{T^\perp}(Y) \rangle$ is at most $\frac{1}{2}\|P_{T^\perp}(Z)\|_*$. Hence

$$\|M + Z\|_* - \|M\|_* \geq -\sqrt{\frac{r}{8n}} \|P_T(Z)\|_F + \frac{1}{2}\|P_{T^\perp}(Z)\|_* \stackrel{?}{>} 0.$$

We will show that with high probability over the choice of Ω, the inequality does indeed hold. We defer the proof of this last fact, since it and the

https://doi.org/10.1017/9781316882177.009 Published online by Cambridge University Press

construction of the helper matrix Y will both make use of the matrix Bernstein inequality, which we present in the next section.

8.3 Quantum Golfing

What remains is to construct a helper matrix Y and prove that with high probability over Ω, for any matrix Z supported in $\overline{\Omega}$, $\|P_{T^\perp}(Z)\|_* > \sqrt{\frac{r}{2n}}\|P_T(Z)\|_F$, to complete the proof we started in the previous section. We will make use of an approach introduced by Gross [80] and follow the proof of Recht in [123], where the strategy is to construct Y iteratively. In each phase, we will invoke concentration results for matrix-valued random variables to prove that the error part of Y decreases geometrically and we make rapid progress in constructing a good helper matrix.

First we will introduce the key concentration result that we will apply in several settings. The following matrix-valued Bernstein inequality first appeared in the work of Ahlswede and Winter related to quantum information theory [6]:

Theorem 8.3.1 *[Noncommutative Bernstein Inequality] Let $X_1 \dots X_l$ be independent mean 0 matrices of size $d \times d$. Let $\rho_k^2 = \max\{\|\mathbb{E}[X_k X_k^T]\|, \|\mathbb{E}[X_k^T X_k]\|\}$ and suppose $\|X_k\| \leq M$ almost surely. Then for $\tau > 0$,*

$$\Pr\left[\left\|\sum_{k=1}^{l} X_k\right\| > \tau\right] \leq 2d \exp\left\{\frac{-\tau^2/2}{\sum_k \rho_k^2 + M\tau/3}\right\}.$$

If $d = 1$, this is the standard Bernstein inequality. If $d > 1$ and the matrices X_k are diagonal, then this inequality can be obtained from the union bound and the standard Bernstein inequality again. However, to build intuition, consider the following toy problem. Let u_k be a random unit vector in $\mathbb{R}^d$ and let $X_k = u_k u_k^T$. Then it is easy to see that $\rho_k^2 = 1/d$. How many trials do we need so that $\sum_k X_k$ is close to the identity (after scaling)? We should expect to need $\Theta(d \log d)$ trials; this is true even if u_k is drawn uniformly at random from the standard basis vectors $\{e_1 \dots e_d\}$ due to the coupon collector problem. Indeed, the above bound corroborates our intuition that $\Theta(d \log d)$ is necessary and sufficient.

Now we will apply the above inequality to build up the tools we will need to finish the proof.

Definition 8.3.2 *Let R_Ω be the operator that zeros out all the entries of a matrix except those in Ω.*

https://doi.org/10.1017/9781316882177.009 Published online by Cambridge University Press

Lemma 8.3.3 *If Ω is chosen uniformly at random and $m \geq nr \log n$, then with high probability*

$$\frac{n^2}{m} \left\| P_T R_\Omega P_T - \frac{m}{n^2} P_T \right\| < \frac{1}{2}.$$

Remark 8.3.4 *Here we are interested in bounding the operator norm of a linear operator on matrices. Let T be such an operator, then $\|T\|$ is defined as*

$$\max_{\|Z\|_F \leq 1} \|T(Z)\|_F.$$

We will explain how this bound fits into the framework of the matrix Bernstein inequality, but for a full proof, see [123]. Note that $\mathbb{E}[P_T R_\Omega P_T] = P_T \mathbb{E}[R_\Omega] P_T = \frac{m}{n^2} P_T$, and so we just need to show that $P_T R_\Omega P_T$ does not deviate too far from its expectation. Let $e_1, e_2, \ldots, e_d$ be the standard basis vectors. Then we can expand:

$$\begin{aligned} P_T(Z) &= \sum_{a,b} \left\langle P_T(Z), e_a e_b^T \right\rangle e_a e_b^T \\ &= \sum_{a,b} \left\langle Z, P_T(e_a e_b^T) \right\rangle e_a e_b^T \end{aligned}$$

Hence $R_\Omega P_T(Z) = \sum_{(a,b)\in\Omega} \left\langle Z, P_T(e_a e_b^T) \right\rangle e_a e_b^T$, and finally we conclude that

$$P_T R_\Omega P_T(Z) = \sum_{(a,b)\in\Omega} \left\langle Z, P_T(e_a e_b^T) \right\rangle P_T(e_a e_b^T).$$

We can think of $P_T R_\Omega P_T$ as the sum of random operators of the form $\tau_{a,b} : Z \to \left\langle Z, P_T(e_a e_b^T) \right\rangle P_T(e_a e_b^T)$, and the lemma follows by applying the matrix Bernstein inequality to the random operator $\sum_{(a,b)\in\Omega} \tau_{a,b}$.

We can now complete the deferred proof of part (a):

Lemma 8.3.5 *If Ω is chosen uniformly at random and $m \geq nr \log n$, then with high probability for any Z supported in $\overline{\Omega}$ we have*

$$\|P_{T^\perp}(Z)\|_* > \sqrt{\frac{r}{2n}} \|P_T(Z)\|_F.$$

Proof: Using Lemma 8.3.3 and the definition of the operator norm (see the remark), we have

$$\left\langle Z, P_T R_\Omega P_T Z - \frac{m}{n^2} P_T Z \right\rangle \geq -\frac{m}{2n^2} \|Z\|_F^2.$$

Furthermore, we can upper-bound the left-hand side as

$$\begin{aligned} \langle Z, P_T R_\Omega P_T Z \rangle &= \langle Z, P_T R_\Omega^2 P_T Z \rangle = \|R_\Omega(Z - P_{T^\perp}(Z))\|_F^2 \\ &= \|R_\Omega(P_{T^\perp}(Z))\|_F^2 \leq \|P_{T^\perp}(Z)\|_F^2 \end{aligned}$$

https://doi.org/10.1017/9781316882177.009 Published online by Cambridge University Press

where in the last line we used that Z is supported in $\overline{\Omega}$, and so $R_\Omega(Z) = 0$. Hence we have that

$$\|P_{T^\perp}(Z)\|_F^2 \geq \frac{m}{n^2}\|P_T(Z)\|_F^2 - \frac{m}{2n^2}\|Z\|_F^2.$$

We can use the fact that $\|Z\|_F^2 = \|P_{T^\perp}(Z)\|_F^2 + \|P_T(Z)\|_F^2$ and conclude that $\|P_{T^\perp}(Z)\|_F^2 \geq \frac{m}{4n^2}\|P_T(Z)\|_F^2$. Now

$$\begin{aligned}\|P_{T^\perp}(Z)\|_*^2 \geq \|P_{T^\perp}(Z)\|_F^2 &\geq \frac{m}{4n^2}\|P_T(Z)\|_F^2 \\ &> \frac{r}{2n}\|P_T(Z)\|_F^2\end{aligned}$$

which completes the proof of the lemma. ■

All that remains is to prove that the helper matrix Y that we made use of actually does exist (with high probability). Recall that we require that Y is supported in Ω and $\|P_T(Y) - UV^T\|_F \leq \sqrt{r/8n}$ and $\|P_{T^\perp}(Y)\| \leq 1/2$. The basic idea is to break up Ω into disjoint sets $\Omega_1, \Omega_2, \ldots \Omega_p$, where $p = \log n$, and use each set of observations to make progress on the remaining $P_T(Y) - UV^T$. More precisely, initialize $Y_0 = 0$, in which case the remainder is $W_0 = UV^T$. Then set

$$Y_{i+1} = Y_i + \frac{n^2}{m}R_{\Omega_{i+1}}(W_i)$$

and update $W_{i+1} = UV^T - P_T(Y_{i+1})$. It is easy to see that $\mathbb{E}[\frac{n^2}{m}R_{\Omega_{i+1}}] = I$. Intuitively, this means that at each step $Y_{i+1} - Y_i$ is an unbiased estimator for W_i, and so we should expect the remainder to decrease quickly (here we will rely on the concentration bounds we derived from the noncommutative Bernstein inequality). Now we can explain the nomenclature *quantum golfing*: at each step, we hit our golf ball in the direction of the hole, but here our target is to approximate the matrix UV^T, which for various reasons is the type of question that arises in quantum mechanics.

It is easy to see that $Y = \sum_i Y_i$ is supported in Ω and that $P_T(W_i) = W_i$ for all i. Hence we can compute

$$\begin{aligned}\|P_T(Y_i) - UV^T\|_F &= \left\| P_T\frac{n^2}{m}R_{\Omega_i}W_{i-1} - W_{i-1}\right\|_F \\ &= \left\| P_T\frac{n^2}{m}R_{\Omega_i}P_TW_{i-1} - P_TW_{i-1}\right\|_F \\ &= \frac{n^2}{m}\left\|P_TR_\Omega P_T - \frac{m}{n^2}P_T\right\| \leq \frac{1}{2}\|W_{i-1}\|_F\end{aligned}$$

https://doi.org/10.1017/9781316882177.009 Published online by Cambridge University Press

where the last inequality follows from Lemma 8.3.3. Therefore, the Frobenius norm of the remainder decreases geometrically, and it is easy to guarantee that Y satisfies condition (a).

The more technically involved part is showing that Y also satisfies condition (b). However, the intuition is that $\|P_{T^\perp}(Y_1)\|$ is itself not too large, and since the norm of the remainder W_i decreases geometrically, we should expect that $\|P_{T^\perp}(Y_i)\|$ does too, and so most of the contribution to

$$\|P_{T^\perp}(Y)\| \leq \sum_i \|P_{T^\perp}(Y_i)\|$$

comes from the first term. For full details, see [123]. This completes the proof that computing the solution to the convex program indeed finds M *exactly*, provided that M is incoherent and $|\Omega| \geq \max(\mu_1^2, \mu_0) r(n+m) \log^2(n+m)$.

Further Remarks

There are many other approaches to matrix completion. What makes the above argument so technically involved is that we wanted to solve exact matrix completion. When our goal is to recover an approximation to M, it becomes much easier to show bounds on the performance of (P_1). Srebro and Shraibman [132] used Rademacher complexity and matrix concentration bounds to show that (P_1) recovers a solution that is close to M. Moreover, their argument extends straightforwardly to the arguably more practically relevant case when M is only entrywise close to being low-rank. Jain et al. [93] and Hardt [83] gave provable guarantees for alternating minimization. These guarantees are worse in terms of their dependence on the coherence, rank, and condition number of M, but alternating minimization has much better running time and space complexity and is the most popular approach in practice. Barak and Moitra [26] studied noisy tensor completion and showed that it is possible to complete tensors better than naively flattening them into matrices, and showed lower bounds based on the hardness of refuting random constraint satisfaction problems.

Following the work on matrix completion, convex programs have proven to be useful in many other related problems, such as separating a matrix into the sum of a low-rank and a sparse part [44]. Chandrasekaran et al. [46] gave a general framework for analyzing convex programs for linear inverse problems and applied it in many settings. An interesting direction is to use reductions and convex programming hierarchies as a framework for exploring computational versus statistical trade-offs [29, 45, 24].

https://doi.org/10.1017/9781316882177.009 Published online by Cambridge University Press