Fast Computation of Low Rank Matrix Approximations

Dimitris Achlioptas*
optas@microsoft.com

Frank M$^{\text{c}}$Sherry†
mcsherry@cs.washington.edu

Abstract

In many practical applications, given an $m \times n$ matrix A it is of interest to find an approximation to A that has low rank. We introduce a technique that exploits spectral structure in A to accelerate Orthogonal Iteration and Lanczos Iteration, the two most common methods for computing such approximations. Our technique amounts to independently sampling and/or quantizing the entries of the input matrix A, thus speeding up computation by reducing the number of non-zero entries and/or the length of their representation. Our analysis s based on observing that both sampling and quantization can be viewed as adding a random matrix E to A, where the entries of E are independent, zero-mean random variables of bounded variance. Such random matrices posses no significant linear structure, and we can thus prove that the effect of sampling and quantization nearly vanishes when a low rank approximation to A is computed. In fact, the more prominent the linear structure in A is, the more data we can afford to discard and, ultimately, the faster we can discover it. We give bounds on the quality of our approximation both in the L2 and in the Frobenius norm.

*Author's address: Microsoft Research, One Microsoft Way, Redmond, WA 98052.
†Author's address: Department of Computer Science, University of Washington, Seattle, WA 98105. Research performed while visiting Microsoft Research and supported in part by the Microsoft Corporation.

1 Introduction

Many aspects of machine learning and data mining are affected by what has become known as "the curse of dimensionality". In order to find more sophisticated trends in data, potential correlations between larger and larger groups of variables must be considered. Unfortunately, the number of such correlations generally increases exponentially with the number of input variables and, as a result, brute force approaches become infeasible. What often makes this all the more frustrating is the realization that human beings have no difficulty identifying all salient features of, say, a high resolution image — a very high dimensional object from the machine learning point of view.

How is it that humans are able to process such images almost instantly? This discrepancy is generally attributed to the observation that *meaningful* images do not exploit their full allotment of dimensionality. There are hundreds of thousands of pixels in an image of a face, yet a face has only so many muscles and there are only so many understood expressions. Humans are not so concerned with the nature of individual pixels, but rather with the implications of a particular trend, the upturning of the mouth to form a smile. This structure inherent in the domain of facial gestures suggests that there are just a few important "axes" for understanding a facial expression: Is the mouth open? What emotion does the face exhibit? Is the person confused? Each of these questions represents a meaningful dimension for the data. Understanding an image could be as simple as discovering the right questions to ask; the proper dimensions to consider.

1.1 Low Rank Approximations

A natural goal for machine learning is to attempt to identify and isolate these characteristic dimensions. We would like to simplify the data sufficiently so that we can apply traditional machine learning techniques, yet we do not wish to oversimplify and leave out information crucial to understanding. A method that is widely used for this purpose is to first cast the data as a matrix A (indexed by $\langle instance, attribute \rangle$) and then compute a matrix of low rank, D, that approximates A. The idea is that the rank of a matrix corresponds roughly to the degrees of freedom of its entries. By constraining the rank of D we aim to capture only the most pertinent characteristics of the data in A, leaving behind dimensions in which the data appears "random".

Such low rank approximations are most often derived by computing the Singular Value Decomposition (SVD) of A and taking the rank k matrix, A_k, that corresponds to the k largest singular values. Recall that for an arbitrary matrix A its Frobenius norm, $|A|_F$, is given by

$$|A|_F^2 = \sum_{i,j} A(i,j)^2 \ .$$

Perhaps the best-known property of A_k is that for any rank k matrix D,

$$|A - D|_F \geq |A - A_k|_F \ . \tag{1}$$

Considering the low rank approximations suggested by the SVD has proven empirically successful in a number of different areas, including Information Retrieval, with Latent Semantic Analysis (LSA) [2, 3], and Face Recognition [11]. In particular, empirical evidence suggests that indeed meaningful dimensions of the data are captured by this approach. At the same time, a theoretical understanding of this fact remains elusive. One way to put the current state of affairs is that we don't really know "To which data analysis question is A_k the answer?" In particular, the characterization of A_k offered by equation (1) does not seem to lend an insight into how A_k captures *meaningful* dimensions of the data in A.

Our starting point in this paper is another, perhaps less well known, property of the SVD. Namely, that the approximations it offers to A are also optimal with respect to the L2 norm,

$$|A| \;=\; |A|_2 \;=\; \max_{|x|=1} |Ax| \;\; .$$

We will see that the L2 norm, also known as spectral norm, measures "linear structure" in data, i.e., the tendency of instances (rows, columns) to be linear combinations of a small number of other instances. Thus, minimizing the L2 norm of $A - D$ appears naturally related to extracting dimensions which are significant for the data in A. We further posit that considering the L2 norm enables a principled approach to determining "how many" dimensions we should retain (what value of k). The key fact in this vein is the stability of spectral structure with respect to perturbations caused by random noise. This will suggest that one should measure $|A - A_k|_2$ vis-à-vis the L2 norm of a random "A-like" matrix (to be defined precisely).

Our main technical contribution lies in exploiting the stability mentioned above in order to inject a special form of "noise" into the data, namely, independent sampling and quantization of individual entries. Each one of these "matrix simplifications" greatly accelerates algorithms that compute low rank approximations, such as Orthogonal Iteration and Lanczos Iteration. Further, we will show how to combine sampling and quantization to yield a dramatic reduction in the amount of data needed to compute a good low rank approximation to a matrix A. Thus, besides speeding up computation, our technique also benefits data storage and transmission.

1.2 Our Results

Both results we present have the ultimate goal of accelerating the computation of low rank matrix approximations. The two most commonly used algorithms for this task are Orthogonal Iteration and Lanczos Iteration. While the two algorithms are quite distinct (for an excellent discussion see [7]), both spend the bulk of their time performing matrix-vector multiplications. Our approach speeds up such multiplications by i) sampling (and thus sparsifying) the input matrix A, and/or ii) quantizing the entries of A so that arithmetic operations can be performed faster.

The approximations to A_k that we compute are of high quality (additive error), as expressed in Theorems 1 and 2. In particular, the bounds we present for our approximation D are of the form

$$|A - D| \;\;\leq\;\; |A - A_k| + \mathit{Error} \;\; ,$$

where *Error* is caused by sampling/quantization. As a yardstick for our approximations, we define ψ_A as the *minimum* amount of linear structure contained in any matrix whose entries are "in the same range" as A. More precisely, for an $m \times n$ matrix A with $\max_{i,j} |A(i,j)| = b$, we define

$$\psi_A = \min_{Q \,\in\, \{-b,+b\}^{m \times n}} |Q| \;\; .$$

Note that ψ_A is bounded by the norm of the random $m \times n$ matrix where each $Q(i,j)$ is $\pm b$ with equal probability; arguably such a matrix contains no meaningful linear structure. Our bounds can afford to measure *Error* in units of ψ_A.

To simplify notation, both Theorems 1 and 2 are stated with probability of failure $1/(m+n)$. As we will see, this probability can be driven to any $1/\mathrm{poly}(m+n)$ factor by a modest increase of the range constants in each theorem (we elaborate on this point in Section 3.1).

Our first result allows us to randomly omit a large fraction of the entries in a matrix without destroying its spectral structure. More precisely, the error incurred is parameterized by the fraction of entries discarded. Hence, the more prominent the linear structure is in A, the more data we can afford to discard and the faster we can discover it.

Theorem 1 *Let A be any $m \times n$ matrix. For $s \geq 1$, define $\widehat{A}$ to be a random $m \times n$ matrix where*

$$\widehat{A}(i,j) \quad = \quad \begin{cases} 0 & w.p.\ 1 - 1/s \\ \\ sA(i,j) & w.p.\ 1/s \ . \end{cases}$$

If $s \leq \frac{m+n}{11^6 \log^6(m+n)}$, then with probability at least $1 - 1/(m+n)$,

$$|A - \widehat{A}_k| \quad < \quad |A - A_k| + 7\,\psi_A \sqrt{s} \ .$$

We note that in practice it is not necessary to multiply entries by s. Omitting s from all $sA(i,j)$ terms would result in exactly $\widehat{A}_k/s$, which could readily be scaled by s afterwards.

Our second result allows us to "round" the entries of A randomly, again without destroying its spectral structure. In Theorem 2 below we are rounding entries to two values, thus requiring only a single bit to store each entry. This represents a 32 to 64 factor of compression over storing floating point numbers. Naturally, one can generalise the rounding process to a larger set of numbers, trading representation length for error.

Theorem 2 *Let A be any $m \times n$ matrix and let $b = \max_{i,j} |A(i,j)|$. Define $\widehat{A}$ to be a random $m \times n$ matrix where*

$$\widehat{A}(i,j) \quad = \quad \begin{cases} +b & w.p.\ \dfrac{1}{2} + \dfrac{A(i,j)}{2b} \\ \\ -b & w.p.\ \dfrac{1}{2} - \dfrac{A(i,j)}{2b} \ . \end{cases}$$

Then with probability at least $1 - 1/(m+n)$,

$$|A - \widehat{A}_k| \quad < \quad |A - A_k| + 7\,\psi_A \ .$$

As with Theorem 1, the use of b is not necessary in practice. Entries could be rounded to ± 1 just as easily, enabling addition in place of multiplication, with a final scaling of the result by b.

We note that there are several occasions when one might seek particularly *weak* linear structure; vectors in the null space of a matrix, for example. Our results are not useful in these domains. In essence, our results are useful whenever the structure we seek in the data is beyond the realm of random noise or, equivalently, whenever adding random noise cannot obscure the structure.

1.3 Related Work

The singular value decomposition has received a lot of attention in information retrieval where its use was pioneered by Deerwester et al. [4] with Latent Semantic Analysis (LSA). Given a collection of vectors, each one capturing the terms appearing in a single document, LSA considers a low rank approximation, D, of their cosine similarity matrix (a symmetric matrix with bounded entries). Considering distances between documents/terms with respect to D goes a long way in addressing both synonymy and polysemy in practice. Recently, certain progress has also been made in providing a theoretical explanation for the empirical success of LSA. This originated with the work of Papadimitriou et al. [10] who proved that LSA works in the context of a simplified probabilistic "corpus-generating" model. A related work, albeit indirectly, is that of Kleinberg [9] on authoritative sources in a hyperlinked environment. Kleinberg's HITS algorithm computes the singular value decomposition for the adjacency matrix of the directed web link graph.

Recently, Azar et al. [1] extended the results of [10], considering a corpus generated by exposing an approximately low-rank matrix A to a particular form of random error E. Our work is inspired by the ideas in [1] and our results can be viewed as turning the random process acting on the data "on its head": rather than viewing the random error matrix E as a foe corrupting the data, we co-opt the random process and "shape" E so that $A + E$ has useful properties.

In terms of computing low rank approximations to large matrices in practice, a relatively common approach is to use "incremental" algorithms. Such algorithms bring as much data as possible into memory, compute the SVD, and then update this SVD in an incremental fashion with the remaining data. To the best of our knowledge, such algorithms come with no guarantees regarding the quality of the approximation produced.

The first mathematically rigorous approach to speeding up the computation of low rank approximations was offered by Frieze, Kannan and Vempala [5]. They showed that by performing a weighted sampling of the columns of A one can in fact afford to keep a matrix $\hat{A}$ whose size depends *only* on k. In particular, they show how to compute a $1 + \epsilon$ approximation with respect to the Frobenius norm by keeping a square submatrix of dimension $10^7 k^4/\epsilon^6$.

While the result of [5] is theoretically intriguing, it suffers from two drawbacks. The first one is pragmatic: the constants involved are far from being practical. (In fairness, it is not clear that practical considerations were the main aim of [5]). The second drawback is more subtle but, we feel, more germane: the approximation offered is good only with respect to the Frobenius norm. As we will see in Section 4, the Frobenius norm is a poor measure of linear structure. In fact, we will see that other than allowing for good "data reconstruction", good approximations with respect to the Frobenius norm do not appear to have other clear implications in the realm of data mining.

To the best of our knowledge, our approach gives the first method which is both practical and armed with strong performance guarantees (both with respect to the Frobenius and the L2 norm). Moreover, we feel that by focusing on the L2 norm we bring forward a potentially *more fruitful* viewpoint for evaluating low rank approximations.

Paper Outline. In Section 2 we discuss linear structure, its measurement by the L2 norm, and its absence from random matrices. In Section 3 we discuss the acceleration of low rank matrix approximations and prove Theorems 1 and 2. In Section 4 we discus the Frobenius norm, why it does not capture linear structure, and give Frobenius norm bounds for our approximations. In section 5 we discuss: i) how to combine sampling and quantization, and ii) how to get further sparsification when there is great variance in the magnitudes of the entries by using non-uniform sampling. We conclude with a summary of our results and some directions for further research.

2 Linear Structure and Randomness

A matrix A can be viewed as a collection of instances (rows) each comprised of values for a common set of attributes (column indices). Linear structure in such a data set is the tendency of the instances towards a particular (consistent) set of ratios between attribute values. That is, fixing any one of the attributes influences other attributes in a linear fashion. We can test for linear structure by multiplying our matrix A with a candidate ratio vector v, normalized so that $|v|=1$. The vector Av then describes the structure captured by v for each instance; entries with large magnitude correspond to rows that have significant projection onto v. Therefore, the strongest linear correlation in our matrix is witnessed by the unit vector x maximizing $|Ax|$, i.e., the witness for the L2 norm of A.

From the above discussion we see that in forming an approximation of A which captures its linear structure, a natural first step is to determine the strongest linear correlation in A (as witnessed by the vector x identified by the L2 norm) and extract it from the matrix. Thus, $D_1 = Axx^T$ becomes our first approximation to A. If all the rows of A are colinear, i.e., $\mathrm{rank}(A) = 1$, then $A = D_1$ and $|A - D_1| = 0$. In general, of course, there is still linear structure left in $A - D_1$. Using the same method, we can now determine a second matrix, D_2, that approximates $A - D_1$. We incorporate the structure that D_2 captures by adding it to D_1. As we repeat this process, we get a more precise approximation to A and an error matrix $A - \sum D_k$ with decreasingly significant linear structure.

What is truly remarkable is that the greedy process described above is in fact optimal. That is, it maintains an optimal approximation to A for all k.

Theorem 3 *Let $A_k = \sum_{i \le k} D_i$. For any rank k matrix D, $|A - D| \ge |A - A_k|$.*

Recall now our motivation for seeking a low rank approximation D to A: if D is to be both low rank and a good approximation to A, then it must capture only the pertinent characteristics of A. Theorem 3 organizes dimensions by pertinence, leaving us with: "What is the right value of k?"

A direct answer to the above question seems hard to come by. Nonetheless, remember that we are seeking dimensions along which the data exhibits structure, i.e., along which it appears "non-random". Therefore, having computed A_k, we can consider the following thought experiment: take the "remaining" matrix $A - A_k$; change the sign of each entry independently and with probability 1/2 to get a matrix P; compute $|P|$. If it turns out that $|A - A_k| \approx |P|$ then it seems fair to say that $A - A_k$ contains no meaningful linear structure. Otherwise, $A - A_k$ still contains linear structure to be extracted.

Fortunately, we don't have to scramble $A - A_k$ in order to decide. Norms of random matrices are well-understood and can be readily compared vis-á-vis $|A - A_k|$.

2.1 The Norms of Random Matrices

At what point is a linear correlation "too strong" to be attributed to chance? Alternatively, how much linear structure is there in a random matrix? Wigner's famous semi-circle law [12] gave a first answer to this question for random symmetric matrices. His result was later refined by Juhász [8] and Füredi and Komlós [6]. We state below a straightforward extension of the Füredi–Komlós bound to non-symmetric matrices.

Theorem 4 *Let E be a random $m \times n$ matrix with $E(i,j) = r_{ij}$ where the $\{r_{ij}\}$ are independent random variables and for all i, j: $r_{ij} \in [-K, K]$, $\mathbf{E}[r_{ij}] = 0$ and $\mathrm{Var}(r_{ij}) \le \sigma^2$. For any $\alpha > 1/2$, if*

$$K \quad < \quad \sigma\sqrt{m+n}\,(7\alpha\log(m+n))^{-3}$$

then

$$\Pr\left[|E| > (7/3)\sigma\sqrt{m+n}\,\right] \quad < \quad (m+n)^{1/2-\alpha} \quad .$$

Proof: We consider the $(m+n) \times (m+n)$ matrix

$$F = \begin{bmatrix} 0 & E^T \\ E & 0 \end{bmatrix} ,$$

use that $|E| = |F|$, and apply Theorem 2 of [6] for random symmetric matrices. (More precisely, we first reparameterize Theorem 2 of [6] to allow for variable probability of success and then consider a straightforward generalization that allows for entries whose variance is *bounded by* σ^2.) $\square$

Remark. Note that although the proof of Theorem 4 appears rather naive, the bound it gives is tight up to a constant factor; if $m > n$ then one should expect $|E| \sim \sigma\sqrt{m}$ since with high probability the rows of E are essentially orthogonal and, as a result, $|E| \approx |E^{(1)}|$, where $E^{(1)}$ is the first row of E.

To put Theorem 4 in perspective let us consider the following two canonical random $m \times n$ matrices. In the first one, N, we have $r_{ij} = N(0,1)$ for all i, j; in the second one, B, for all i, j, $r_{ij} = \pm 1$, each value having probability 1/2. Using that for any matrix A, $|A|_F^2 \leq |A|^2 \times \min\{m, n\}$, it is easy to show that with exponentially high probability

$$|N| > (2/3)\sqrt{m+n} \qquad \text{and} \qquad |B| > (2/3)\sqrt{m+n} \ .$$

At the same time, Theorem 4 implies (immediately for B, almost immediately for N) that with probability, say, $1 - 1/(m+n)$

$$|N| < (7/3)\sqrt{m+n} \qquad \text{and} \qquad |B| < (7/3)\sqrt{m+n} \ .$$

Thus, we see that in both examples the theorem is tight up to a small constant factor.

Finally, let us note that the fact $|A|_F^2 \leq |A|^2 \times \min\{m, n\}$ also implies

$$\min_{Q \in \{-b,+b\}^{m \times n}} |Q| \geq (b/\sqrt{2})\sqrt{m+n} \ ,$$

and hence B above is very close to being a minimizer of the L2 norm.

3 Computing Low Rank Approximations

Lanczos Iteration and Orthogonal Iteration are the most commonly used techniques to compute low rank approximations of matrices. (When one is interested in determining *all* singular values, often the matrix is first brought to a tridiagonal form — a $O((m+n)^3)$ operation.) The particulars of these two algorithms are not really important to our discussion, save for a common feature: to compute a rank k approximation, each algorithm repeatedly multiplies the input matrix A with a (changing) set of k orthogonal vectors. In fact, the bulk of the running time for both algorithms is comprised of these matrix-vector multiplications. (The k vectors are initially arbitrary, random say, but with each iteration their span gets closer to the top k-dimensional invariant subspace of A.) It is worth pointing out that to compute even a single matrix-vector product one needs to read into memory the entire matrix from wherever it is stored. For many practical applications, this last requirement alone makes the computation of a low rank approximation infeasible. Using sampling and quantization can give dramatic gains in each of the following respects.

- **Time**: The number of operations required for a matrix-vector multiplication is proportional to the number of *non-zero* entries in the matrix.
- **Space**: The amount of data that needs to be processed (and hence stored/transferred) for a matrix-vector multiplication is proportional to the *representation length* of each entry.

3.1 L2 Bounds

We present now our main technical theorem, relating $|A - \widehat{A}_k|$ to $|A - A_k|$. Theorems 1 and 2 will follow from Theorem 5 by considering the corresponding random matrix for each case.

Theorem 5 *Let A be any $m \times n$ matrix and let $b = \max_{i,j} |A(i,j)|$. Let $\widehat{A}$ be a random $m \times n$ matrix with $\widehat{A}(i,j) = a_{ij}$ where the $\{a_{ij}\}$ are independent random variables such that for all i, j: $\mathbf{E}[a_{ij}] = A(i,j)$, $\mathrm{Var}(a_{ij}) \leq (\sigma b)^2$, and*

$$|A(i,j) - a_{ij}| \quad \leq \quad \sigma b \sqrt{m+n}\,(7\alpha \log(m+n))^{-3} \quad . \tag{2}$$

Then with probability at least $1 - (m+n)^{1/2-\alpha}$,

$$|A - \widehat{A}_k| \quad \leq \quad |A - A_k| + 7\,\sigma\,\psi_A \quad .$$

Proof: Before we commence the proof we need an observation that allows us to consider the relation between $|A - A_k|$ and $|\widehat{A} - \widehat{A}_k|$ in terms of $|\widehat{A} - A|$. In particular, for all k

$$|\widehat{A} - \widehat{A}_k| \quad \leq \quad |A - A_k| + |A - \widehat{A}| \quad . \tag{3}$$

The above is essentially equivalent to a variational property of singular values, a high dimensional application of the triangle inequality. To make the proof self-contained we will prove (3) below.

Now, starting with $|A - \widehat{A}_k|$, applying the triangle inequality and using (3) we get

$$\begin{aligned} |A - \widehat{A}_k| &\leq |A - \widehat{A}| + |\widehat{A} - \widehat{A}_k| \\ &\leq |A - A_k| + 2|A - \widehat{A}| \quad . \end{aligned}$$

To prove the theorem we observe that the random matrix $A - \widehat{A}$ fits the conditions of Theorem 4. Therefore, with probability at least $1 - (m+n)^{1/2-\alpha}$, it has norm less than $(7/3)\sigma b\sqrt{m+n}$. To conclude the proof we recall that $\psi_A \geq (b/\sqrt{2})\sqrt{m+n}$.

To prove (3) we will use the Minimax characterisation of the singular values of A. In particular, for each k, a player *min* (minimizer) choose a k dimensional subspace Y. Then, in response, a player *max* (maximizer) attempts to chose a vector x orthogonal to Y so as to maximize $|Ax|$. The Minimax characterisation of the singular values of A is

$$|A - A_k| \quad = \quad \min_{|Y|=k} \max_{x \perp Y} |Ax| \quad .$$

(Recall from our construction of A_k that we repeatedly removed dimensions that had large L2 norm; that is, we played the role of the minimizer.) With the Minimax definition in mind, we have

$$\begin{aligned} |\widehat{A} - \widehat{A}_k| &= \min_{|Y|=k} \max_{x \perp Y} |\widehat{A}x| \\ &= \min_{|Y|=k} \max_{x \perp Y} |(A + \widehat{A} - A)x| \\ &\leq \min_{|Y|=k} \max_{x \perp Y} |Ax| + \max_{x} |(\widehat{A} - A)x| \\ &= |A - A_k| + |\widehat{A} - A| \end{aligned}$$

□

Theorems 1 and 2 follow from Theorem 5 almost trivially. In each case, the expectation of $\widehat{A}$ is A while σ is $\sqrt{s}$ and 1, respectively. Taking $\alpha = 3/2$ yields the stated bounds.

Our choice of $\alpha = 3/2$ is somewhat arbitrary; we chose 3/2 because it happens to make the probabilities relatively simple to read. We can drive the probability of failure to any polynomial of $m + n$ by changing the leading constant in (2) in Theorem 5. In particular, for Theorem 1, to achieve probability of success $1-(m+n)^{1/2-\alpha}$, we must require that $s < (m+n)(7\alpha\log(m+n))^{-6}$. In Theorem 2, the probability amplification only changes (by a constant factor) the smallest value of $m + n$ for which the Theorem holds.

4 The Frobenius Norm

We start by discussing the insensitivity of the Frobenius norm to linear structure and the implications of bounding approximation error in terms of that norm. In spite of our negative conclusions, it is clear that there are domains in which the Frobenius norm is relevant. For example, entrywise confidence is important in data reconstruction and is naturally captured by the Frobenius norm. In Section 4.3, we give bounds on the accuracy of the approximations resulting from sampling and quantization with respect to the Frobenius norm.

4.1 The Singular Value Decomposition

Our discussion of the Frobenius norm will require the formal definition of the singular value decomposition: any $m \times n$ matrix A can be written as

$$A = \sum_{1\le i\le n} \sigma_i u_i v_i^T \quad , \tag{4}$$

where the singular values, $\{\sigma_i\}$, are non-increasing, non-negative scalars and the singular vectors, $\{u_i\}$ and $\{v_i\}$, are each orthonormal bases. Note that A_k, the optimal rank k approximation to A, can be written as the partial sum

$$A_k = \sum_{1\le i\le k} \sigma_i u_i v_i^T \quad .$$

4.2 Linear Structure and the Frobenius Norm

Recall that the Frobenius norm $|\cdot|_F$ is given by

$$|A|_F^2 = \sum_{i,j} A(i,j)^2 \quad .$$

Let's start our examination of the Frobenius norm with a somewhat foreboding example. For a matrix A, consider a new matrix B resulting by changing the sign of each entry in A independently and with probability 1/2. It's not hard to see that B fits the criteria of Theorem 4, and thus has a very small amount of linear structure. However, the Frobenius norm of B is identical to that of A.

While this example suggests that something may be amiss with the Frobenius norm, perhaps a more insightful point of view is the following. Observe that for any vector x,

$$
\begin{aligned}
|Ax|^2 &= \left|\sum_{i=1}^{n}(\sigma_i u_i v_i^T)x\right|^2 && \text{(The definition of the SVD)} \\
&= \sum_{i=1}^{n}\left(\sigma_i v_i^T x\right)^2 && \text{(The } u_i \text{ are orthonormal)} \\
&= \sum_{i=1}^{n}\sigma_i^2\left(v_i^T x\right)^2 \quad .
\end{aligned}
$$

For unit x, we see that the length of x after being multiplied by A is a convex combination of the squares of A's singular values. As we've seen, the L2 norm is described by the maximizer of $|Ax|$, namely $x = v_1$, and thus it captures σ_1, the *largest* stretch induced by A. Rather than taking $x = v_1$, let us now choose x uniformly at random from the n-dimensional unit sphere. By spherical symmetry each $v_i^T x$ is identically distributed; since $\sum_i \sigma_i^2 = |A|_F^2$ we get

$$
\begin{aligned}
\mathbf{E}\left[|Ax|^2\right] &= \mathbf{E}\left[(v_i^T x)^2\right]\cdot\sum_{i=1}^{n}\sigma_i^2 \\
&= \frac{1}{n}\cdot|A|_F^2 \quad .
\end{aligned}
$$

Thus, $|A|_F$ measures the *average* stretch induced by A. Such averaging nature is not what we are looking for when we seek structure; as the number of dimensions in A gets large, it is easy for the strong correlations to get lost in the averaging, desensitising the Frobenius norm to linear structure.

To conclude, let us consider a particular example of the insensitivity mentioned above. Consider an $n \times n$ matrix E whose elements are chosen uniformly from $\{-1, +1\}$. This matrix has large (n) Frobenius norm, but relatively small L2 norm. Let $A = \sigma u v^T$ be a matrix with reasonably strong linear structure, i.e., $\sqrt{n} \ll \sigma \ll n$. It is illustrative to observe how the Frobenius norm changes as the linear structure is added and removed from E:

$$|E|_F = n \quad \leftrightarrow \quad |E + A|_F \le n + \sigma$$

$$|E| \le 3\sqrt{n} \quad \leftrightarrow \quad |E + A| \ge \sigma - 3\sqrt{n}$$

For σ in the range noted above, $|E + A|_F = (1 + o(1))|E|_F$. Thus, the precision of *any* Frobenius norm approximation must be very large in order to be at all sensitive to A. For example, note that E_k, which obviously does not capture anything meaningful about $E + A$, is a pretty good approximation to $E + A$ in the Frobenius sense. In particular, it is only $1 + o(1)$ worse than the optimal approximation, namely $(E + A)_k$.

4.3 Frobenius Bounds

As we said earlier, the Frobenius norm does represent a good measure of entry-wise error. In particular, if we are are simply interested in compression, the Frobenius norm indicates how closely we match the original data. Our bounds suggest that as long as the input matrix A has strong linear structure we can still provide a good entrywise approximation to A_k after sampling and/or quantizing the entries of A.

In our Frobenius norm bounds we will not use the matrix $\widehat{A}_k$ to approximate A. Instead, we will compute the space spanned by the top k singular vectors of $\widehat{A}$ and use the projection of A onto that space. This approach greatly simplifies the analysis and parallels that of [5], in that we only compute a "representation" of the the low rank approximation. It is quite possible that $\widehat{A}_k$ itself is also a good approximation to A in the Frobenius norm but we do not consider this possibility here.

Theorem 6 *Given matrices A and $\widehat{A}$, let $E = \widehat{A} - A$. If $\widehat{V}_k$ is the matrix whose columns are the top k right singular vectors of $\widehat{A}$, then*

$$|A - A(\widehat{V}_k \widehat{V}_k^T)|_F^2 \quad \leq \quad |A - A_k|_F^2 + \frac{4|E|}{\sigma_k}|A_k|_F^2 \ .$$

The proof of this theorem can be found in the appendix.

We will apply Theorem 6 to analyze sampling and quantization. Recall that we assume that k is such that the linear structure in A_k is significant, i.e.,

$$L = \sigma_k(A)/\psi_A \gg 1 \ .$$

Corollary 7 *Let A be any $m \times n$ matrix and let $b = \max_{i,j} |A(i,j)|$. Let $L = \sigma_k(A)/\psi_A$. Define $\widehat{A}$ to be a random $m \times n$ matrix where*

$$\widehat{A}(i,j) \quad = \quad \begin{cases} 0 & w.p.\ 1 - 1/s \\ \\ sA(i,j) & w.p.\ 1/s \ . \end{cases}$$

Let $\widehat{V}_k$ be the matrix whose columns are the top k right singular vectors of $\widehat{A}$. If $s \leq \frac{m+n}{11^6 \log^6(m+n)}$, then with probability at least $1 - 1/(m+n)$,

$$|A - A(\widehat{V}_k \widehat{V}_k^T)|_F^2 \quad = \quad |A - A_k|_F^2 + \frac{28\sqrt{s}}{L}|A_k|_F^2 \ .$$

Corollary 8 *Let A be any $m \times n$ matrix and let $b = \max_{i,j} |A(i,j)|$. Let $L = \sigma_k(A)/\psi_A$. Define $\widehat{A}$ to be a random $m \times n$ matrix where*

$$\widehat{A}(i,j) \quad = \quad \begin{cases} +b & w.p.\ \dfrac{1}{2} + \dfrac{A(i,j)}{2b} \\ \\ -b & w.p.\ \dfrac{1}{2} - \dfrac{A(i,j)}{2b} \ . \end{cases}$$

Let $\widehat{V}_k$ be the matrix whose columns are the top k right singular vectors of $\widehat{A}$. With probability at least $1 - 1/(m+n)$,

$$|A - A(\widehat{V}_k \widehat{V}_k^T)|_F^2 \quad = \quad |A - A_k|_F^2 + \frac{28}{L}|A_k|_F^2 \ .$$

5 Implementation Details

While our results are mathematical, we hope that they affect the practice of computing low rank approximations. Hence, we feel that it is appropriate to discuss two particular practical points in some detail. The first such point pertains to combining our two techniques effectively. The second point is a slightly more sophisticated sampling process which will generally result in smaller error than that specified by Theorem 1.

5.1 Combination of Results

Mathematically, the results of Theorems 1 and 2 are entirely orthogonal and can thus be combined readily. In practice, though, we must face the fact that a significant fraction of the representation of a sparse matrix lies in representing is structure, not its entries. A typical representation of a sparse matrix stores a list of triples $\langle row, col, val \rangle$, each representing a non-zero entry in the matrix. Compressing the *val* to one bit results in relatively minor compression.

To address this, we will make use of the fact that the sparsity structure is introduced by a process that *we control*. That is, in practice, by using only the seed of the "random" number generator we can reconstruct the list of non-zero positions that our process introduced. We store this seed along with the list of bits indicating the rounded values in non-zero positions. When we multiply the matrix by a vector, we simultaneously "run" the omission process, generating the list of non-zero entries. For each entry, we read off the bit value and perform the appropriate multiplication. This allows us to reap the combined benefits of sparsification and quantization.

Note that in terms of computation the above scheme introduces the overhead of a random number generator that produces geometrically distributed random variables. To put this in perspective, recall that the generation of these random variables can be performed without memory interaction and is much faster than reading data from memory.

5.2 Non-Uniform Sampling

Recall that the bound we use for the norm of a random matrix, Theorem 4, relies only on the maximum deviation of any entry. Moreover, note that there is no a priori reason to omit each entry in A with the same probability. In particular, we can tailor each omission probability so that all entries in $\widehat{A} - A$ have the same deviation (maximum) and thus omit even more entries. More precisely, if $E = \widehat{A} - A$ then, if we keep entry (i, j) with probability $1/s_{ij}$, we have

$$\mathrm{Var}\,(E(i,j)) \quad = \quad A(i,j)^2(s_{ij} - 1) \ .$$

If we require that $s_{ij} = s$, then the maximum variance is $\sigma^2 = \max_{i,j} A(i,j)^2(s-1)$. Any entry which has magnitude less than the maximum can, thus, have its probability discounted without increasing the maximum deviation. In particular, if we set

$$s_{ij} \quad = \quad 1 + \frac{\sigma^2}{A(i,j)^2}$$

then the variance corresponding to each entry will be exactly σ^2, while $s_{ij} \geq s$ for all i, j. Assuming that not all entries are of equal magnitude, we have decreased the expected number of non-zero entries, without affecting the provable norm of the matrix. Note that care must be taken to ensure that we do not violate the range constraint of Theorem 4. Whenever we increase the s_{ij}, we increase the range of the random variable. We may not be able to fully enlarge s_{ij} for very small $|A(i,j)|$.

6 Conclusions and Future Directions

We examined the computation of low rank approximations for the purpose of extracting structure from data. In particular, we observed that the empirical success of such approximations is related to their minimization of the L2 norm along with the stability of the latter in the presence of random error. We have used these facts to give matrix simplification techniques that accelerate the computation of low rank approximations.

While our theoretical results are, in our eyes, compelling at the same time we feel that ultimately "the proof of the algorithm is in the running." We have conducted a few very preliminary experiments to get some sense of what the appropriate parameters would be in practical applications. For example, we tested a sequence of human faces (the *eigenfaces* domain) for a relatively small corpus ($m \times n = 2000 \times 1000$). We found that we could set $s = 15$, i.e., keep only 7% of the data, without suffering noticeable error. Naturally, these results are preliminary, and further experimentation is necessary. A general trend, though, that we discovered (confirming mathematical intuition) is that the larger the input size is, the *greater* an omission rate, s, we can afford.

Another direction that merits investigation is the application of our techniques to accelerate the computation of *errorless* low rank approximations. The key idea in that direction is the following. Imagine that we have run, say, Orthogonal Iteration on $\widehat{A}$ to the point where we are relatively close to a (perfect) rank k approximation to (the imperfect) $\widehat{A}$. At that point, rather than letting the method run to convergence, we can instead *add* (put back) data from A to $\widehat{A}$, lowering the fraction of entries omitted, and continue running with this new matrix. Clearly, repeating this process until we have put back all the data converges to A_k. Moreover, this scheme fits perfectly into the incremental nature of both Orthogonal and Lanczos Iteration. The computational savings come from the fact that until we get quite close to the k dimensional invariant subspace of A, a "rough approximation" of A is "just as good" in terms of driving each method's convergence. Understanding the behaviour of this scheme, in particular the appropriate rate to reintroduce entries, remains an interesting open problem.

References

[1] Yossi Azar, Amos Fiat, Anna Karlin, Frank McSherry, and Jared Saia, *Data mining through spectral analysis*, submitted to STOC 2001.

[2] Michael W. Berry, Zlatko Drmač, and Elizabeth R. Jessup, *Matrices, vector spaces, and information retrieval*, SIAM Rev. **41** (1999), no. 2, 335–362 (electronic).

[3] Michael W. Berry, Susan T. Dumais, and Gavin W. O'Brien, *Using linear algebra for intelligent information retrieval*, SIAM Rev. **37** (1995), no. 4, 573–595.

[4] Scott Deerwester, Susan T. Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman, *Indexing by latent semantic analysis*, Journal of the Society for Information Science **41** (1990), no. 6, 391–407.

[5] Alan Frieze, Ravi Kannan, and Santosh Vempala, *Fast monte-carlo algorithms for finding low-rank approximations*, 39th Annual Symposium on Foundations of Computer Science (Palo Alto, CA, 1998), 1998, pp. 370–378.

[6] Zoltán Füredi and János Komlós, *The eigenvalues of random symmetric matrices*, Combinatorica **1** (1981), no. 3, 233–241.

[7] Gene H. Golub and Charles F. Van Loan, *Matrix computations*, third ed., Johns Hopkins University Press, Baltimore, MD, 1996.

[8] F. Juhász, *On the spectrum of a random graph*, Algebraic methods in graph theory, Vol. I, II (Szeged, 1978), North-Holland, Amsterdam, 1981, pp. 313–316.

[9] Jon M. Kleinberg, *Authoritative sources in a hyperlinked environment*, J. ACM **46** (1999), no. 5, 604–632.

[10] Christos H. Papadimitriou, Prabhakar Raghavan, Hisao Tamaki, and Santosh Vempala, *Latent semantic indexing: A probabilistic analysis*, 17th Annual Symposium on Principles of Database Systems (Seattle, WA, 1998), 1998, pp. 159–168.

[11] Matthew Turk and Alex Pentland, *Eigenfaces for recognition*, Journal of Cognitive Neuroscience **3** (1991), no. 1, 71–86.

[12] Eugene P. Wigner, *On the distribution of the roots of certain symmetric matrices*, Ann. of Math. (2) **67** (1958), 325–327.

7 Appendix

We note that the proof of Theorem 6 is rather more technical than the other proofs in this paper. In particular, some unmotivated definitions and unproven inequalities are used. All can be found in the excellent text of Golub and Van Loan [7].

Proof: We will need a few facts about the Frobenius norm. Principal among them is the equality

$$|A|_F^2 = \sum_i |Aq_i|^2 \ , \tag{5}$$

valid for any orthonormal basis $\{q_i\}$. We will consider the bases $V = \{v_i\}$ and $\widehat{V} = \{\widehat{v}_i\}$ corresponding to the right singular vectors of A and $\widehat{A}$, respectively. Applying (5) to V and $\widehat{V}$ we get

$$\sum_i |Av_i|^2 = \sum_i |A\widehat{v}_i|^2 \ . \tag{6}$$

Applying (5) to $A - A_k$ with respect to V we further get

$$\begin{aligned} |A - A_k|_F^2 &= \sum_{i\le k} |(A - A_k)v_i|^2 + \sum_{i>k} |(A - A_k)v_i|^2 \\ &= \sum_{i>k} |(A - A_k)v_i|^2 \\ &= \sum_{i>k} |Av_i|^2 \ . \end{aligned} \tag{7}$$

Similarly applying (5) to $A - A(\widehat{V}_k\widehat{V}_k^T)$ with respect to $\widehat{V}$ we get

$$|A - A(\widehat{V}_k\widehat{V}_k^T)|_F^2 = \sum_{i>k} |A\widehat{v}_i|^2 \ . \tag{8}$$

Thus, we can rewrite (6) as

$$\sum_{i\le k} |Av_i|^2 + |A - A_k|_F^2 = \sum_{i\le k} |A\widehat{v}_i|^2 + |A - A(\widehat{V}_k\widehat{V}_k^T)|_F^2 \ . \tag{9}$$

Ultimately, the second and fourth terms in (9) are going to participate in our bound, so we are interested in bounding the remaining terms. Rearranging, we have

$$\begin{aligned} |A - A(\widehat{V}_k \widehat{V}_k^T)|_F^2 &= |A - A_k|_F^2 + \sum_{i \le k} \left(|Av_i|^2 - |A\widehat{v}_i|^2\right) \\ &= |A - A_k|_F^2 + \sum_{i \le k} \left(\sigma_i^2 - |(\widehat{A} - E)\widehat{v}_i|^2\right) \quad . \end{aligned}$$

To conclude the proof we will use a fairly common inequality about singular values, namely that adding a matrix E cannot change the k^{th} singular value by more than $|E|$. In fact, this is the same observation as equation (3). We thus get

$$\begin{aligned} |A - A(\widehat{V}_k \widehat{V}_k^T)|_F^2 &\le |A - A_k|_F^2 + \sum_{i \le k} \left(\sigma_i^2 - (\sigma_i - 2|E|)^2\right) \\ &\le |A - A_k|_F^2 + \sum_{i \le k} 4\sigma_i |E| \\ &\le |A - A_k|_F^2 + \frac{4|E|}{\sigma_k} \sum_{i \le k} \sigma_i^2 \\ &= |A - A_k|_F^2 + \frac{4|E|}{\sigma_k} |A_k|_F^2 \quad . \end{aligned}$$

□