1 Overview

- Low-rank approximation,
- Compression Sensing

2 Last Time

We looked at three different regression methods. The first was based on $\varepsilon$-subspace embeddings. The second was an iterative approach, building a well-conditioned matrix good for stochastic gradient descent. The third was formulated as follows:

For the least square problem $\min_x \| Sx - b \|_2^2$, which has optimal solution $x^* = S^+ b$, and approximate solution $\tilde{x}^* = \arg\min_x \| Sx - \Pi x \|_2^2$, we let $x^* = U\alpha$, $w = Sx^* - b$, $U\beta = S\tilde{x}^* - Sx^*$ where $S = U\Sigma V^T$. We proved that $(\Pi U)^T (\Pi U) \beta = (\Pi U)^T \Pi w$ last time.

These results from regression will appear in our work for low-rank approximation.

3 Low-rank approximation

The basic idea is a huge matrix $A \in \mathbb{R}^{n \times d}$ with $n, d$ both very large - say, $n$ users rating $d$ movies. We might believe that the users are linear combinations of a few ($k$) basic types. We want to discover this low-rank structure. More formally:

Given a matrix $A \in \mathbb{R}^{n \times d}$, we want to compute $A_k := \arg\min_{\text{rank}(B) \leq k} \| A - B \|_F$.

Some now argue that we should look for a non-negative matrix factorization; nevertheless, this version is still used.

**Theorem 1** (Eckart-Young). Let $A = U\Sigma V^T$ be a singular-value decomposition of $A$ where $\text{rank}(A) = r$ and $\Sigma$ is diagonal with entries $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r > 0$, then under $\| \cdot \|_X = \| \cdot \|_F$, $A_k = U_k \Sigma_k V_k^T$ is the minimizer where $U_k$ and $V_k$ are the first $k$ columns of $U$ and $V$ and $\Sigma_k = \text{diag}(\sigma_1, \ldots, \sigma_k)$.

Our output is then $U_k, \Sigma_k, V_k$. We can calculate $A_k$ in $O(nd^2)$ time, by calculating the SVD of $A$. We would like to do better. First, a few definitions:

**Definition 2.** Proj$_A B$ is the projection of the columns of $B$ onto the colspace($A$).

**Definition 3.** Let $A = U\Sigma V^T$ be a singular decomposition. $A^+ = V\Sigma^{-1}U^T$ is called Moore-Penrose pseudoinverse of $A$. 

3.1 Algorithm

Today we are going to use a sketch which is used both in subspace embedding and approximate matrix multiplication to compute $\tilde{A}_k$ with rank at most $k$ such that $\|A - \tilde{A}_k\|_F \leq (1+\epsilon)\|A - A_k\|_F$, following Sarlós’ approach [8]. The first works which got some decent error (like $\epsilon \|A\|_F$) was due to Papadimitriou [7] and Frieze, Kannan and Vempala [5].

**Theorem 4.** Define $\tilde{A}_k = \text{Proj}_{A\Pi^T k}(A)$. As long as $\Pi \in \mathbb{R}^{m \times n}$ is an $1/2$ subspace embedding for a certain $k$-dimensional subspace $V_k$ and satisfies approximate matrix multiplication with error $\sqrt{\epsilon/k}$, then

$$\|A - \tilde{A}_k\|_F \leq (1 + O(\epsilon))\|A - A_k\|_F,$$

where $\text{Proj}_{V,k}(A)$ is the best rank $k$ approximation to $\text{Proj}_V(A)$, i.e., projecting the columns of $A$ to $V$.

Before we prove this theorem, let us first convince ourselves that this algorithm is fast, and that we can compute $\text{Proj}_{A\Pi^T k}(A)$ quickly. To satisfy the conditions in the above theorem, we know that $\Pi \in \mathbb{R}^{m \times d}$ can be chosen with $m = O(k/\epsilon)$ e.g. using a random sign matrix (or slightly larger $m$ using a faster subspace embedding). We need to multiply $A\Pi^T$. We can use a fast subspace embedding to compute $A\Pi^T$ quickly, then we can compute the SVD of $A\Pi^T = U^\prime S^\prime V'^T$ in $O(mn^2)$ time. Let $[\cdot]_k$ denote the best rank-$k$ approximation under Frobenius norm. We then want to compute $[U^\prime U'^T A]_k = U'[U'^T A]_k$. Computing $U'^T A$ takes $O(mnd)$ time, then computing the SVD of $U'^T A$ takes $O(dm^2)$ time. Note that this is already better than the $O(nd^2)$ time to compute the SVD of $A$, but we can do better if we approximate. In particular, by using the right combination of subspace embeddings, for constant $\epsilon$ the scheme described here can be made to take $O(nnz(A)) + O(ndk)$ time (where $\tilde{O}$ hides log $n$ factors). We will shoot instead for $O(nnz(A)) + O(nk^2)$.

Consider that:

- We want to compute $\tilde{A}_k = \text{argmin}_{X, \text{rank}(X) \leq k} \|U'X - A\|_F^2$. If $X^+$ is the argmin without the rank constraint, then the argmin with the rank constraint is $[U^\prime X^+ ]_k = U'[X^+]_k$, where $[\cdot]_k$ denotes the best rank-$k$ approximation under Frobenius error.

- Rather than find $X^+$, we use approximate regression to find an approximately optimal $\tilde{X}$. That is, we compute $\tilde{X} = \text{argmin}_X \|\Pi' U'X - \Pi' A\|_F^2$, where $\Pi'$ is an $\alpha$-subspace embedding for the column space of $U'$ (note $U'$ has rank $m$). Then our final output is $U'[\tilde{X}]_k$.

Why does the above work? (Thanks to Michael Cohen for describing the following simple argument.) First note

$$\left(\frac{1 + \alpha}{1 - \alpha}\right) \cdot \|U'X^+ - A\|_F^2 \geq \|U'\tilde{X} - A\|_F^2,$$

$$= \|(U'X^+ - A) + U'(\tilde{X} - X^+)\|_F^2$$

$$= \|U'X^+ - A\|_F^2 + \|U'(\tilde{X} - X^+)\|_F^2$$

$$= \|U'X^+ - A\|_F^2 + \|\tilde{X} - X^+\|_F^2.$$
and thus $\|\tilde{X} - X^+\|_F^2 \leq O(\alpha) \cdot \|U'X^+ - A\|_F^2$. The second equality above holds since the matrix $U'$ preserves Frobenius norms, and the first equality since $U'X^+ - A$ has a column space orthogonal to the column space of $U'$. Next, suppose $f, \tilde{f}$ are two functions mapping the same domain to $\mathbb{R}$ such that $|f(x) - \tilde{f}(x)| \leq \eta$ for all $x$ in the domain. Then clearly $f(\text{argmin}_x \tilde{f}(x)) \leq \min_x f(x) + 2\eta$.

Now, let the domain be the set of all rank-$k$ matrices, and let $f(Z) = \|U'X^+ - Z\|_F$ and $\tilde{f}(Z) = \|U'\tilde{X} - Z\|_F$. Then $\eta = \|U'X^+ - U'\tilde{X}\|_F = \|X^+ - \tilde{X}\|_F$. Thus

$$\|U'[\tilde{X}]_k - A\|_F^2 = \|U'[\tilde{X}]_k - U'X^+\|_F^2 + \|(I - U'U^T)A\|_F^2 \leq (\|U'[X]^+\|_k - U'X^+\|_F + 2 \cdot X^+ - \tilde{X}\|_F^2 + \|(I - U'U^T)A\|_F^2.$$

where (1) used that $U'X^+ - A$ has rank $k$-subspace embedding for the column space of $A$ and let $\Pi_1$ be the best Frobenius approximation to $U'X^+$. Also, (2) used that $U'X^+ + U'X^+ - A\|_F^2 = U'[X]^+ - A\|_F^2 + \|U'[X]^+ - U'X^+\|_F^2$. since $U'X^+ - A$ has columns orthogonal to the column space of $U'$. Finally, (3) again used $\|U'X^+ - A\|_F^2 \leq \|U'[X]^+\|_A - A\|_F^2$, and also used the triangle inequality

$$\|U'[X]^+\|_k - U'X^+\|_F \leq \|U'[X]^+\|_k - A\|_F + \|U'X^+ - A\|_F \leq 2 \cdot \|U'[X]^+\|_k - A\|_F.$$

Thus we have established the following theorem, which follows from the above calculations and Theorem 4.

**Theorem 5.** Let $\Pi_1 \in \mathbb{R}^{m_1 \times n}$ be a 1/2 subspace embedding for a certain $k$-dimensional subspace $V_k$, and suppose $\Pi_1$ also satisfies approximate matrix multiplication with error $\sqrt{\varepsilon/k}$. Let $\Pi_2 \in \mathbb{R}^{m_2 \times n}$ be an $\alpha$-subspace embedding for the column space of $U'$, where $A\Pi_1^T = U'S\Pi_1$ is the SVD (and hence $U'$ has rank at most $m_1$). Let $\tilde{A}_k' = U'[\tilde{X}]_k$ where

$$\tilde{X} = \text{argmin}_X \|\Pi_2 U'X - \Pi_2 A\|_F^2.$$

Then $\tilde{A}_k'$ has rank $k$ and

$$\|A - \tilde{A}_k'\|_F \leq (1 + O(\varepsilon) + O(\sqrt{\alpha})) \|A - A_k\|_F.$$

In particular, the error is $(1 + O(\varepsilon)) \|A - A_k\|_F$ for $\alpha = \varepsilon$. 

3
In the remaining part of these lecture notes, we show that \( \text{Proj}_{A^T, k}(A) \) actually is a good rank-\( k \) approximation to \( A \) (i.e. we prove Theorem 4). In the following proof, we will denote the first \( k \) columns of \( U \) and \( V \) as \( U_k \) and \( V_k \) and the remaining columns by \( U_{k}^\perp \) and \( V_{k}^\perp \).

**Proof.** Let \( Y \) be the column span of \( \text{Proj}_{A^T}(A_k) \) and the orthogonal projection operator onto \( Y \) as \( P \). Then,

\[
\|A - \text{Proj}_{A^T,k}(A)\|_F^2 \leq \|A - PA\|_F^2 = \|A_k - PA_k\|_F^2 + \|A_k - PA_k\|_F^2.
\]

Then we can bound the second term in that sum:

\[
\|A_k\| = \|(I - P)A_k\|_F \leq \|A_k\|_F^2.
\]

Now we just need to show that \( \|A_k - PA_k\|_F^2 \leq \varepsilon \|A_k\|_F^2 \):

\[
\|A - PA\|_F^2 = \|A_k - (A^T)(A^T)^+ A_k\|_F^2 \leq \|A_k - (A^T)(A^T)^+ A_k\|_F^2
\]

\[
= \|A_k^T - A_k^T (A^T)^+ (A^T)^+ A_k\|_F^2
\]

\[
= \sum_{i=1}^n \|A_k^T(i) - A_k^T (A^T)^+ (A^T)^+ (A^T)^+ (A^T)^+ A_k(i)\|_F^2
\]

Here superscript \((i)\) means the \( i \)th column. Now we have a bunch of different approximate regression problems which have the following form:

\[
\min_x \| A_k^T x - \Pi (A^T)^+ (A^T)^+ (A^T)^+ A_k(i) \|_2,
\]

which has optimal value \( \hat{x}^* = (\Pi A^T)^+ (A^T)^+ (A^T)^+ A_k(i) \). Consider the problem \( \min_x \| A_k^T x - (A^T)^+ (A^T)^+ A_k(i) \|_2 \) as original regression problem. In this case optimal \( x^* \) gives \( A_k^T x^* = \text{Proj}_{A^T}(A^T)^+ (A^T)^+ A_k(i) \). Now we can use the analysis on the approximate least square from last week.

In our problem, we have a bunch of \( w_i, \beta_i, \alpha_i \) with \( S = A_k^T = V_k \Sigma_k U_k^T \) and \( b_i = (A^T)^+ A_k(i) \). Here, \( \|w_i\|^2 = \|Sx^* - b\|^2 = \|(A_k^T)^+ (A^T)^+ (A^T)^+ A_k(i)\|^2 \). Hence \( \sum_i \|w_i\|^2 = \|A - A_k\|_F^2 \). On the other hand, \( \sum_i \|\beta_i\|^2 = \|A_k^T - A_k^T (A^T)^+ (A^T)^+ A_k(i)\|_F^2 \). Since \( (IV_k)^T (IV_k) \beta_i = (IV_k)^T \Pi w_i \), if all singular values of \( IV_k \) are at least \( 1/2^{1/4} \), we have

\[
\frac{\sum_i \|\beta_i\|^2}{2} \leq \sum_i \|(IV_k)^T (IV_k) \beta_i\|^2 = \sum_i \|(IV_k)^T \Pi w_i\|^2 = \|(IV_k)^T \Pi W\|^T
\]

where \( W \) has \( w_i \) as \( i \)th column. What does it look like? \( (IV_k)^T \Pi W \) exactly look like approximate matrix multiplication of \( V_k \) and \( W \). Since columns of \( W \) and \( V_k \) are orthogonal, we have \( V_k^T W = 0 \), hence if \( \Pi \) is a sketch for approximate matrix multiplication of error \( \varepsilon' = \sqrt{\varepsilon/k} \), then

\[
\mathbb{P}(\|(IV_k)^T (IV_k)\|_F^2 \geq \varepsilon \|W\|_F^2) < \delta
\]

since \( \|V_k\|_F^2 = k \). Clearly \( \|W\|_F^2 = \sum_i \|w_i\|^2 = \|A - A_k\|_F^2 \), we get the desired result. \( \square \)
3.2 Further results

What we just talked about gives a good low-rank approximation but every column of $\tilde{A}_k$ is a linear combination of potentially all columns of $A$. In applications (e.g. information retrieval), we want a few number of columns be spanning our low dimensional subspace. There has been work on finding fewer columns of $A$ (call them $C$) such that $\|A - (CC^+A)k\|_F^2$ is small, but we will not talk about it deeply.

- Boutsidis et al. [1] showed that we can take $C$ with $\approx 2k/\varepsilon$ columns and error $\leq \varepsilon\|A - A_k\|_F$.
- Guruswami and Sinop got $C$ with $\leq k\varepsilon + k - 1$ columns such that $\|A - CC^+A\|_F \leq (1 + \varepsilon)\|A - A_k\|_F$.

3.3 K-Means as a Low-Rank Approximation Problem

The k-means problem, which was stated on the problem set, involved a set of points $(x_1, ..., x_n) \in \mathbb{R}^d$. Let $A$ be the matrix with the $i$th row equal to $x_i^T$. Given a partition $P(P_1, ..., P_k)$ of points into $k$ clusters, then the best centroids are averages of the clusters. Define the matrix $X_p \in \mathbb{R}^{n \times k}$ such that:

$$(X_p)_{i,j} = \begin{cases} \frac{1}{\sqrt{|P_j|}} & \text{if } i \in P_j \\ 0 & \text{otherwise} \end{cases}$$

Note that $X_p^TX_p = I$. It can be shown that the $i$th row of $X_pX_p^TA$ is the centroid of the cluster that $X_i$ belongs to. Thus, solving k-means is equivalent to finding some $P^* = \arg\min\|A - X_pX_p^TA\|$ - this is a constrained rank-k approximation problem. Cohen et. al[3] show that $\Pi$ can have $m = O(k/\varepsilon^2)$ for a $(1 + \varepsilon)$ approximation, or a $m = O(\log k/\varepsilon^2)$ for a $(9 + \varepsilon)$ approximation (the second bound is specifically for the k-means problem). It is an open problem whether this second bound can get a better approximation.

4 Compressed Sensing

4.1 Basic Idea

Consider $x \in \mathbb{R}^n$. If $x$ is a $k$ sparse vector, we could represent it in a far more compressed manner. Thus, we define a measure of how ”compressible” a vector is as a measure of how close it is to being $k$ sparse.

**Definition 6.** Let $x_{\text{head}(k)}$ be the $k$ elements of largest magnitude in $x$. Let $x_{\text{tail}(k)}$ be the rest of $x$.

Therefore, we call $x$ compressible if $\|x_{\text{tail}(k)}\|$ is small.

The goal here is to approximately recover $x$ from few linear measurements. Consider we have a matrix $\Pi x$ such that each the $i$th row is equal to $\langle \alpha_i, x \rangle$ for some $\alpha_1, ..., \alpha_m \in \mathbb{R}^n$. We want to recover a $\tilde{x}$ from $\Pi x$ such that $\|x - \tilde{x}\|_p \leq C_{\varepsilon, p,q} \|x_{\text{tail}(k)}\|_q$, where $C_{\varepsilon, p,q}$ is some constant dependent on $\varepsilon$, $p$ and $q$. Depending on the problem formulation, I may or may not get to choose this matrix $\Pi$. 
4.2 Approximate Sparsity

There are many practical applications in which approximately sparse vectors appear. Pixelated images, for example, are usually approximately sparse in some basis $U$. For example, consider an $n$ by $n$ image $x \in \mathbb{R}^{n^2}$, then $x = Uy$ for some basis $U$, and $y$ is approximately sparse. Thus we can get measurements from $\Pi Uy$.

Images are typically sparse in the wavelet basis. We will describe how to transform to the Haar wavelet basis here. Assume that $n$ is a power of two. Then:

1. Break the image $x$ into squares of size four pixels.
2. Initialize a new image, with four regions $R_1, R_2, R_3, R_4$.
3. Each block of four pixels, $b$, in $x$ has a corresponding single pixel in each of $R_{1b}, R_{2b}, R_{3b},$ and $R_{4b}$ based on its location. For each block of four $b$:
   - Let the $b$ have pixel values $p_1, p_2, p_3,$ and $p_4$.
   - $R_{1b} \leftarrow \frac{1}{4}(p_1 + p_2 + p_3 + p_4)$
   - $R_{2b} \leftarrow \frac{1}{4}(p_1 - p_2 + p_3 - p_4)$
   - $R_{3b} \leftarrow \frac{1}{4}(p_1 - p_2 - p_3 + p_4)$
   - $R_{4b} \leftarrow \frac{1}{4}(p_1 - p_2 + p_3 - p_4)$
4. Recurse on $R_1, R_2, R_3,$ and $R_4$.

The general idea is this: usually, pixels are relatively constant in certain regions. Thus, the values in all regions except for the first are usually relatively small. If you view images after this transform, the upper left hand regions will often be closer to white, while the rest will be relatively sparse.

**Theorem 7** (Candès, Romberg, Tao [2], Donoho [4]). There exists a $\Pi \in \mathbb{R}^{m \times n}$ with $m = O(k \log(n/k))$ and a poly-time algorithm $\text{Alg}$ s.t. if $\tilde{x} = \text{Alg}(\Pi x)$ then $\|x - \tilde{x}\|_2 \leq O(k^{-1/2})\|x_{\text{tail}(k)}\|_1$

If $x$ is actually k-spares, $2k$ measurements are necessary and sufficient. We will see this by examining Prony’s method in one of our problem sets, and investigate compressed sensing further next class.

References


