CS 761 : Randomized Algorithms , Spring 2023, Waterloo

Lecture 11 : Numerical Linear Algebra

We study the subspace embedding technique to design fast approximation algorithms for linear regression, and see three constructions of subspace embedding.

We will also briefly see a fast algorithm for Laplacian solver.

Fast Linear Regression

Random sampling and dimension reduction are widely used in designing fast algorithms for numerical linear algebra problems.

We illustrate these ideas in a basic problem, the least square problem.

In the least square problem, we are given an $n \times d$ matrix $A \in \mathbb{R}^{n \times d}$ and $b \in \mathbb{R}^n$, and the objective is to find an $x \in \mathbb{R}^d$ to minimize $\|Ax - b\|_2$.

We think of $n \gg d$, so the problem is over-constrained.

To solve it exactly, the runtime is $\Omega(n\,\mathrm{poly}(d))$, which is too slow for large n.

We would like to find an approximation algorithm with $\|Ax' - b\|_2 \leq (1+\varepsilon) \min_x \|Ax - b\|_2$ in $\tilde{O}(nd + \mathrm{poly}(d/\varepsilon))$ time, which is near linear when $n \gg d$.

The idea is to use a near-linear time algorithm to compress the matrix A into a $k \times d$ matrix A' with $k = \mathrm{poly}(d/\varepsilon)$, and then solve $\|A'x - b\|_2$ exactly as our approximate solution.

<u>Definition</u> (Subspace embedding) A $(1 \pm \varepsilon)$ ℓ_2-subspace embedding for the column space of an $n \times d$ matrix A is a matrix S for which $(1-\varepsilon)\|Ax\|_2^2 \leq \|SAx\|_2^2 \leq (1+\varepsilon)\|Ax\|_2^2 \quad \forall x \in \mathbb{R}^d$.

Suppose we have such a matrix $k \times n$ matrix S with $k = \mathrm{poly}(d/\varepsilon)$.

Then we just solve $\min_x \|SAx - Sb\|_2$ instead in $\mathrm{poly}(d/\varepsilon)$ time and use this solution as our approximate solution, as $\|Ax - b\|_2^2 \leq (1+\varepsilon)\|S(Ax-b)\|_2^2 \leq (1+\varepsilon)\|S(Ax^* - b)\|_2^2 \leq \frac{1+\varepsilon}{1-\varepsilon}\|Ax^* - b\|_2^2$ where x and x^* are the minimizers for the compressed and the original problem respectively.

Subspace Embedding

There are two approaches to do subspace embedding, oblivious embedding and row sampling, both of which we have seen.

Oblivious Embedding

As you may imagine, the Johnson-Lindenstrauss theorem will be useful here, i.e. $S = \frac{1}{\sqrt{k}} G$ where each entry is a standard normal random variable.

One technical detail is that in the Johnson-Lindenstrauss transform, for $k = O(\frac{1}{\varepsilon^2}\log(\frac{1}{\delta}))$, it works for one specific vector with probability $\geq 1-\delta$.

The subspace embedding requires that it works for all vectors in $\mathbb{R}^d$, and there are infinitely many.

The analysis is to show that if it works for an ε-net for some constant ε (i.e. a discretization of the unit sphere), then it works for all vectors $\mathbb{R}^d$.

We have seen in L05 that an ε-net is of size c^n for some constant c.

Therefore, the Johnson-Lindenstrauss transform will work if $k = \Theta(d/\varepsilon^2)$, by setting $\delta = \frac{1}{c^d}$ and by union bound.

We omit the details of the proof as it is quite similar to that for compressed sensing in L05, although the way to use the ε-net and the argument using inner product is different.

See the survey by Woodruff for details.

Fast Oblivious Embedding

Besides the number of rows of the matrix S, another important point is to compute SA and Sb efficiently, as matrix multiplication is slow and so compression may already take too much time.

There are much research in fast dimension reduction, and it is possible to do the compression in $O(nd\log n)$ time, which is near linear when $A \in \mathbb{R}^{n\times d}$ is a dense matrix (see Woodruff's survey).

A surprising result of Clarkson and Woodruff proves that a very sparse matrix S works:

Set $k = \Theta(\frac{d^2}{\varepsilon^2}\,\mathrm{polylog}(\frac{d}{\varepsilon}))$, for each column, choose a random location, set it to be $+1$ with probability $\frac{1}{2}$ and -1 with probability $\frac{1}{2}$.

So, each column has only one non-zero entry, and the compression can be done very efficiently.

We will present a simple proof by Nelson and Nguyen (also Meng and Mahoney) using a spectral analysis.

Spectral Analysis

Recall that the goal is to find a matrix $S \in \mathbb{R}^{m\times n}$ (with m small and S sparse) so that $\|SAx\|_2^2 = (1\pm\varepsilon)\|Ax\|_2^2 \;\; \forall x \in \mathbb{R}^d$

Let $U \in \mathbb{R}^{n\times d}$ be an orthonormal basis of the column space of A so that $\{Ax \mid x \in \mathbb{R}^d\} = \{Uy \mid y \in \mathbb{R}^d\}$.

Then the goal is equivalent to finding $S \in \mathbb{R}^{m\times n}$ so that $(1-\varepsilon)\|Uy\|_2^2 \leq \|SUy\|_2^2 \leq (1+\varepsilon)\|Uy\|_2^2 \;\; \forall y \in \mathbb{R}^d$.

As $U^TU = I_d$, this can be rewritten as $-\varepsilon\|y\|_2^2 \leq y^T(U^TS^TSU - I_d)y \leq \varepsilon\|y\|_2^2$, which is equivalent to $\|U^TS^TSU - I_d\|_{op} \leq \varepsilon$, where $\|M\|_{op}$ is the maximum absolute value of an eigenvalue for symmetric M.

Let $\Pi = U^TS^TSU$. To bound the probability that $\|\Pi - I\|_{op} > \varepsilon$, the idea is simply to use Markov's inequality

$$\Pr[\|\Pi - I\|_{op} > \varepsilon] = \Pr[\|\Pi - I\|_{op}^2 > \varepsilon^2] \le \mathbb{E}[\|\Pi - I\|_{op}^2]/\varepsilon^2 \le \mathbb{E}[\|\Pi - I\|_F^2]/\varepsilon^2,$$

where $\|M\|_F^2 = \sum_{i,j} M_{i,j}^2$ is the Frobenius norm, and the last inequality is because

$$\|M\|_F^2 = tr(M^T M) = \sum_i \lambda_i(M^T M) = \sum_i \lambda_i^2(M) \quad \text{and so} \quad \|M\|_F^2 \ge \|M\|_{op}^2.$$

Construction $\quad S$ is an $m \times n$ matrix.

Let $h : [n] \to [m]$ be a hash function, and $\sigma \in \{-1, +1\}^n$ be random signs.

Set $S_{h(i),i} = \sigma(i)$ for $1 \le i \le n$. This is the construction.

We also write $S_{i,j} = \delta_{i,j}\sigma_{i,j}$ where $\delta_{i,j}$ is the indicator random variable whether $S_{i,j} \ne 0$ and $\sigma_{i,j}$ is random sign.

Theorem $\quad$ If h is pairwise independent and σ is 4-wise independent, then $\Pi = U^T S^T S U$ is a $(1 \pm \varepsilon)$-subspace embedding for A with probability at least $1 - \delta$ as long as $m \gtrsim d^2/\delta\varepsilon^2$.

Proof $\quad$ The plan is to bound $\mathbb{E}[\|\Pi - I\|_F^2]$ and then apply Markov's inequality as stated above.

To do so, we compute the entries of Π.

Note that $(SU)_{r,k} = \sum_{i=1}^n \delta_{r,i}\sigma_{r,i} u_i^k$ where u^k is the k-th column of U.

$$\text{So,} \quad \Pi_{k,k'} = \sum_{r=1}^m (SU)_{r,k}(SU)_{r,k'} = \sum_{r=1}^m \Big(\sum_{i=1}^n \delta_{r,i}\sigma_{r,i}u_i^k\Big)\Big(\sum_{i=1}^n \delta_{r,i}\sigma_{r,i}u_i^{k'}\Big)$$

$$= \sum_{r=1}^m \delta_{r,i} \sum_{i=1}^n u_i^k u_i^{k'} + \sum_{r=1}^m \sum_{i \ne j} \delta_{r,i}\delta_{r,j}\sigma_{r,i}\sigma_{r,j} u_i^k u_j^{k'}$$

$$= \langle u^k, u^{k'} \rangle + \sum_{r=1}^m \sum_{i \ne j} \delta_{r,i}\delta_{r,j}\sigma_{r,i}\sigma_{r,j} u_i^k u_j^{k'} \qquad \text{as } \sum_{r=1}^m \delta_{r,i} = 1.$$

As the columns of U are orthonormal, it follows that $(\Pi - I)_{k,k'} = \sum_{r=1}^m \sum_{i \ne j} \delta_{r,i}\delta_{r,j}\sigma_{r,i}\sigma_{r,j} u_i^k u_j^{k'}$.

Now, we just bound the square of each entry of $\Pi - I$.

For the diagonal entries, $\mathbb{E}[(\Pi - I)_{k,k}^2] = \sum_{r=1}^m \sum_{i \ne j} \frac{2}{m^2} (u_i^k)^2 (u_j^k)^2 \quad$ // only when $(i,j) = (i',j')$ as $\mathbb{E}[\sigma_{r,i}] = 0$

$$\le \frac{2}{m} \sum_{i \ne j} (u_i^k)^2 (u_j^k)^2 \le \frac{2}{m}\|u^k\|^4 \le \frac{2}{m}.$$

For off-diagonal entries, $\mathbb{E}[(\Pi - I)_{k,k'}^2] = \frac{1}{m^2}\sum_{r=1}^m \sum_{i \ne j} \Big((u_i^k)^2 (u_j^{k'})^2 + u_i^k u_j^k u_i^{k'} u_j^{k'}\Big) \quad$ // only for $(i,j),(i,j)$ and $(i,j),(j,i)$

$$= \frac{1}{m}\sum_{i \ne j}\Big((u_i^k)^2 (u_j^{k'})^2 + u_i^k u_j^k u_i^{k'} u_j^{k'}\Big)$$

$$\le \frac{1}{m}\sum_{i \ne j} (u_i^k)^2 (u_j^{k'})^2 \le \frac{1}{m}\|u^k\|^2 \cdot \|u^{k'}\|^2 = \frac{1}{m},$$

where the first inequality is because $0 = \langle u^k, u^{k'} \rangle^2 = \sum_{i=1}^n (u_i^k)^2 (u_i^{k'})^2 + \sum_{i \ne j} u_i^k u_j^k u_i^{k'} u_j^{k'} \Rightarrow \sum_{i \ne j} u_i^k u_j^k u_i^{k'} u_j^{k'} \le 0$.

Therefore, $\mathbb{E}[\|\Pi - I\|_F^2] \le \frac{2d}{m} + \frac{d(d-1)}{m} = \frac{d^2 + d}{m}$.

We conclude that $\Pr[\|\Pi - I\|_{op} > \varepsilon] \le \mathbb{E}[\|\Pi - I\|_F^2]/\varepsilon^2 \le \frac{d^2 + d}{\varepsilon^2 m} \le \delta$ by our choice of m.

Note that we only need pairwise independence of h (and thus $\delta_{r,i}$ and $\delta_{r,j}$) and 4-wise independence of σ. $\square$

Leverage Score Sampling

Another approach of subspace embedding is similar to what we have seen in spectral sparsification.

Given $A \in \mathbb{R}^{n \times d}$ and $b \in \mathbb{R}^n$, we first reduce the problem to the case when the columns of A are orthonormal.

This is reminiscent to the reduction to the identity matrix in spectral sparsification, so that $A^T A = I_d$, or equivalently $\sum_{i=1}^{n} a_i a_i^T = I_d$ where a_i is the i-th row of A.

Then, we construct a matrix B by sampling and rescaling each row proportional to its length, so that $\sum_{i=1}^{n} s_i a_i a_i^T \approx I_d$ with only $O(d \log d / \varepsilon^2)$ nonzeros, i.e. B has $O(d \log d / \varepsilon^2)$ rows. where each row of B is $\sqrt{s_i}\, a_i$ so that $(1-\varepsilon) A^T A \preceq B^T B \preceq (1+\varepsilon) A^T A$.

It is a good subspace embedding as $\|Ax\|_2^2 \approx \|Bx\|_2^2$ because $x^T A^T A x \approx x^T B^T B x$.

All the technical details are very similar to those in spectral sparsification, e.g. matrix Chernoff bound.

The sampling probability is called the leverage score of a row, a generalization of effective resistance.

These ideas are very useful in numerical linear algebra, which have further applications in other areas such as optimization (e.g. fast interior point algorithms for linear programming crucially used many of these tricks).

Laplacian Solver

I will tell some history about Laplacian solvers, and briefly describe the simple and elegant algorithm by Kyng and Sachdeva in class.

References

- Woodruff, Sketching as a tool for numerical linear algebra, 2014.
- Nelson, Nguyen. Faster numerical linear algebra algorithms via sparser subspace embeddings, 2014.
- Kyng, Sachdeva, Approximate Gaussian elimination for Laplacians: fast, sparse, and simple, 2016.