18.409 The Behavior of Algorithms in Practice 2/14/2

Lecture 3

Lecturer: Dan Spielman *Scribe: Arvind Sankar*

1 Largest singular value

In order to bound the condition number, we need an upper bound on the largest singular value in addition to the lower bound on the smallest that we derived last class. Since the largest singular value of $A+G$ can be bounded by

$$\sigma_n(A+G) = \|A+G\| \leq \|A\| + \|G\|$$

and we can't really do much about $\|A\|$, the important thing to do is bound $\|G\|$. To start off with a weak but easy bound, we use the following simple lemma.

Lemma 1. *If a_i denote the columns of the matrix A, then*

$$\max_i \|a_i\| \leq \|A\| \leq \sqrt{d} \max_i \|a_i\|$$

Proof. If e_i denotes the vector with 1 in the ith component but 0's everywhere else, then

$$Ae_i = a_i$$

Hence the left-hand inequality is clear. For the other inequality, let x be a unit vector and write

$$Ax = A\left(\sum_i x_i e_i\right) = \sum_i x_i a_i$$

Therefore

$$\|Ax\| \leq \sum_i |x_i| \|a_i\|$$

Applying Cauchy-Schwarz and using the fact that $\|x\| = 1$, we get

$$\|Ax\| \leq \|x\| \sqrt{\sum_i \|a_i\|^2} \leq \sqrt{d \max_i \|a_i\|^2}$$

which is what we want. □

If g is a vector of Gaussian random variables with variance 1, then $\|g\|^2$ is distributed according to the χ^2 distribution with d degrees of freedom, which has density function

$$\frac{x^{d/2-1} e^{-x/2}}{\Gamma(d/2) 2^{d/2}}$$

We need the following bound on how large a χ^2 random variable can be.

Lemma 2. *If X is a random variable distributed according to the χ^2 distribution with d degrees of freedom, then*

$$\Pr\{X \geq kd\} \leq k^{d/2-1}e^{-d(k-1)/2}$$

Since $\|G\| \geq kd$ implies $\max_i \|g_i\| \geq k\sqrt{d}$, hence using lemma 2 and the union bound, we get

$$\Pr\{\|G\| \geq kd\} \leq dk^{d-2}e^{-d(k^2-1)/2}$$

2 A sharper bound using nets

The bound above is unsatisfying: for any fixed unit vector x, the vector Gx is a Gaussian random vector, and so its length should be about $\sqrt{d}$ on average. This section will show how to get a bound on $\|G\|$ that uses this idea to get a bound on $\|G\|$ that grows as $\sqrt{d}$ rather than as d.

Let S^{d-1} denote the $(d-1)$-dimensional unit sphere (the boundary of the unit ball in d dimensions).

Definition 1. *A λ-net on S^{d-1} is a collection of points $\{x_1, x_2, \ldots x_n\}$ such that for any $x \in S^{d-1}$,*

$$\min_i \|x - x_i\| \leq \lambda$$

We will use only 1-nets, and the following lemma claims that they need not be too large.

Lemma 3. *For $d \geq 2$, there exists a 1-net with at most $2^d(d-1)$ points.*

Using this lemma, we can prove the following bound on $\|G\|$:

Lemma 4. *If G is a matrix of standard normal variables, then*

$$\Pr\{\|G\| \geq 2k\sqrt{d}\} \leq 2^d(d-1)k^{d-2}e^{-d(k^2-1)/2}$$

(This lemma appears with a slightly different bound as lemma 2.8 on pg. 907 of [Sza90])

Proof. Let N be the 1-net given by lemma 3. Let $G = U\Sigma V^T$ be the singular value decomposition of G, and let u_i and v_i be the columns of U and V respectively. By definition of the net, there exists a vector $x \in N$ such that

$$\|v_n - x\| \leq 1$$

This is equivalent to

$$v_n \cdot x \geq \frac{1}{2}$$

Expanding x in the basis v_i, we obtain

$$x = \sum_i x_i v_i$$

with $x_n \geq 1/2$. Hence

$$\|Gx\| = \|\sum_i x_i G v_i\| = \|\sum_i x_i \sigma_i u_i\| \geq x_n \sigma_n \geq \|G\|/2$$

Hence $\|G\| \geq 2k\sqrt{d}$ implies that there exists $x \in N$ such that

$$\|Gx\| \geq k\sqrt{d}$$

By the union bound and lemma 2, we obtain

$$\Pr\{\|G\| \geq 2k\sqrt{d}\} \leq |N| k^{d-2} e^{-d(k^2-1)/2}$$

which is the stated result. $\square$

3 Gaussian elimination

In the next couple of lectures, we will use the results we have proved to analyze Gaussian elimination. Briefly, Gaussian elimination solves a system

$$Ax = b$$

by performing row and column operations on A to reduce it to an upper triangular matrix, which can then be easily solved.

Theoretically, one can view this process as factoring A into a product of a lower triangular matrix representing the row operations performed (actually, their inverses), and an upper triangular matrix representing the result of these operations. This is called the *LU-factorization* of A.

There are three pivoting strategies one can use while performing this algorithm (pivoting is the process of permuting rows and/or columns before doing the elimination).

1. *No pivoting*: Just what it says. This can be done only if we never run into zeros on the diagonal. This is easy to analyze.

2. *Partial pivoting*: Here only row permutations are permitted. The strategy is to bring the largest entry in the column we are considering onto the diagonal. The LU-factorization now actually has to be written as

$$LU = PA$$

where P is a permutation matrix representing the row permutations performed. Partial pivoting guarantees that no entry in L can exceed 1 in absolute value.

3. *Complete pivoting*: Here both row and column permutations are permitted, and the strategy is to move the largest entry in the part of the matrix that we have not yet processed to the diagonal. The factorization now looks like

$$LU = PAQ$$

where P and Q are permutation matrices.

Wilkinson showed that if $\hat{L}$, $\hat{U}$ and $\hat{x}$ represent the computed values of L, U and x in floating point to an accuracy of ϵ, then

$$\exists \delta A \text{ such that } (A + \delta A)\hat{x} = b$$

with

$$\|\delta A\| \le d\epsilon(3\|A\|_\infty + 5\|L\|_\infty \|U\|_\infty)$$

Matlab uses partial pivoting, and it can be shown that there exist matrices A for which partial pivoting fails, in the sense that $\|U\|_\infty$ becomes exponentially large (in d). This leads to a total loss of precision unless at least d bits are used to store intermediate results.

Wilkinson also showed that for complete pivoting,

$$\frac{\|U\|_\infty}{\|A\|_\infty} \le d^{\frac{1}{2}\lg d}$$

which means that the number of bits required is only $\lg^2 d$ in the worst case. However, complete pivoting is much more expensive in floating point than partial pivoting, which seems to work quite well in practice. One of the goals of this class is to understand why. In the next couple of lectures, we will show in fact that *no* pivoting does well most of the time.

4 Proof of technical lemmas

For completeness, we give the proofs of lemmas 2 and 3.

Proof of lemma 2. We have

$$\begin{aligned} \Pr\{X \ge kd\} &= \int_{kd}^{\infty} \frac{x^{d/2-1} e^{-x/2}}{\Gamma(d/2)2^{d/2}}\, dx \\ &= \int_{d}^{\infty} \frac{(x + (k-1)d)^{d/2-1}\, e^{-(k-1)d/2 - x/2}}{\Gamma(d/2)2^{d/2}}\, dx \end{aligned}$$

Using $x + (k-1)d \le kx$,

$$\begin{aligned} &\le k^{d/2-1} e^{-(k-1)d/2} \int_{d}^{\infty} \frac{x^{d/2-1} e^{-x/2}}{\Gamma(d/2)2^{d/2}}\, dx \\ &\le k^{d/2-1} e^{-(k-1)d/2} \end{aligned}$$

and we are done. □

Proof of lemma 3. Let N be a maximal set of points on the unit sphere such that the great-circle distance between any two points in N is at least $\pi/3$. Then N will be a 1-net, because if u were a unit vector such that no vector in N is within distance 1 of u, then there would be no point of N within great-circle distance $\pi/3$ of u, so u could be added to N.

To see that $|N| \leq (d-1)2^d$, observe that the sets

$$B(x, \pi/6) = \{u \in S^{d-1} : d(u, x) \leq \pi/6\}, \quad x \in N$$

are disjoint. A lower bound on the $(d-1)$-dimensional volume of each $B(x, \pi/6)$ is given by the volume of the $(d-1)$-dimensional ball of radius $\sin(\pi/6) = 1/2$. If S_{d-1} denotes the volume of S^{d-1} and V_d the volume of the unit ball in d dimensions, then

$$V_d = \frac{2\pi^{d/2}}{d\Gamma(d/2)} \quad \text{and} \quad S_{d-1} = \frac{2\pi^{d/2}}{\Gamma(d/2)}$$

Hence

$$\begin{aligned} |N| &\leq 2^{d-1} \frac{S_{d-1}}{V_{d-1}} \\ &= 2^{d-1}(d-1)\sqrt{\pi}\frac{\Gamma((d-1)/2)}{\Gamma(d/2)} \\ &\leq 2^d(d-1) \end{aligned}$$

A somewhat tighter bound can be obtained by using the fact that

$$\lim_{d\to\infty} \frac{\Gamma((d-1)/2)}{\Gamma(d/2)} = \frac{e}{\sqrt{d}}$$

□

References

[Sza90] Stanislaw J. Szarek, *Spaces with large distance to ℓ_∞^n and random matrices*, American Journal of Mathematics **112** (1990), no. 6, 899–942.