A tutorial on randomized algorithms in numerical linear algebra

Mark Tygert
Program in Applied Mathematics
Yale University

IPAM (at UCLA)
September, 2007

Three problems

1. Verifying that the spectral norm of a matrix is not meaningfully larger than desired

2. Linear least-squares regression (solving overdetermined systems of linear-algebraic equations in the least-squares sense)

3. Low-rank approximation of a matrix (including computing several of the greatest singular values and corresponding singular vectors)

We will survey points 1 and 2 during the present tutorial talk, start discussing point 3 with the next (self-contained) tutorial talk, and finish with a (self-contained) talk during Workshop II.

Recent work on randomized solutions

- J. D. Dixon (1983)
- J. Kuczyński and H. Woźniakowski (1992)
- C. H. Papadimitriou, P. Raghavan, H. Tamaki, and S. Vempala (2000)
- A. Frieze, R. Kannan, and S. Vempala (1999, 2004)
- D. Achlioptas and F. McSherry (2001)
- P. Drineas, R. Kannan, M. W. Mahoney, and S. Muthukrishnan (2006a, 2006b, 2006c, 2006d)
- S. Har-Peled (2006)
- A. Deshpande and S. Vempala (2006)
- S. Friedland, M. Kaveh, A. Niknejad, and H. Zare (2006)
- T. Sarlós (2006a, 2006b, 2006c)
- E. Liberty, P.-G. Martinsson, V. Rokhlin, F. Woolfe, and the speaker (2006, 2007)

Estimating the spectral norm, 1

Suppose we construct an approximation A to a given matrix B.

How can we determine efficiently if the approximation is good, particularly if we cannot afford to access all of the individual entries of $D = A - B$, but can only afford to apply D and D^* to a few vectors?

We need an algorithm for ensuring that the spectral norm of $D = A - B$ is not too large.

We will see that applying $(D^*\, D)^k$ to a normalized random vector, and taking the Euclidean norm of the result, yields a decent estimate to $\|D\|^{2k}$, with probability approaching 1 exponentially fast as k increases.

Estimating the spectral norm, 2

Via remarkably simple calculations concerning the surface area of a band around the equator of a hypersphere, Dixon (1983) showed the following:

Define $\tilde{\omega} = \omega/\|\omega\|$, where ω is an $n \times 1$ vector whose entries are i.i.d. centered Gaussian random variables.

Given any $n \times n$ matrix C, we have $\mathbf{P}\,\{\|C\,\tilde{\omega}\| < \mu\,\|C\|\} \;<\; 0.8\,\mu\,\sqrt{n}$ for any positive real number $\mu < 1$.

Given any $m \times n$ matrix D and any positive integer k, substituting $C = (D^*\,D)^k$ yields $\mathbf{P}\left\{\left\|(D^*\,D)^k\,\tilde{\omega}\right\|^{1/(2k)} < \mu\,\|D\|\right\} \;<\; 0.8\,\mu^{2k}\,\sqrt{n}$ for any positive real number $\mu < 1$.

Estimating the spectral norm, 3

Therefore, $\left\|(D^* D)^k \tilde{\omega}\right\|^{1/(2k)}$ is not much less than $\|D\|$, with probability approaching 1 exponentially fast as k increases.

Of course, $\left\|(D^* D)^k \tilde{\omega}\right\|^{1/(2k)}$ is always less than or equal to $\|D\|$.

Thus, $\left\|(D^* D)^k \tilde{\omega}\right\|^{1/(2k)}$ gives a decent estimate of $\|D\|$ with very high probability, if k is large enough. The probability depends only on k, not on gaps between the singular values of D.

In other words, the (modified) power method with a random start provides a decent estimate of $\|D\|$ with very high probability.

Kuczyński & Woźniakowski (1992) derived somewhat stronger estimates for the classical power and Lanczos methods; their derivations are beyond the scope of this talk.

Properties of certain random matrices, 1

Suppose that k, l, and m are integers such that $0 < k < l < m$.

Suppose further that R is an $l \times m$ random matrix, either [1] consisting of i.i.d. $N\left(0, \frac{1}{l}\right)$ entries, or [2] consisting of uniformly randomly selected rows of the product of an appropriately normalized discrete Fourier transform and a diagonal matrix whose diagonal entries are i.i.d., drawn uniformly from the unit circle.

Then,

$$\|R\| \lesssim \sqrt{m}. \tag{1}$$

Properties of certain random matrices, 2

Moreover, for any $m \times k$ matrix U whose columns are orthonormal, and real number ε such that $0 < \varepsilon < 1$,

$$\sigma_1(RU) \lesssim \sqrt{1+\varepsilon} \tag{2}$$

$$1/\sigma_k(RU) \lesssim \sqrt{1+\varepsilon} \tag{3}$$

with high probability, given that l is sufficiently greater than k (meaning that, for example, $l = 10(k/\varepsilon)^2$ for a random matrix R of type [2]).

Randomized regression utilizes (2) and (3).

Randomized reduced-rank approximation utilizes (1) and (3). Also, (2) is helpful for certain theoretically interesting bounds.

Linear least-squares regression, 1

Suppose that A is an $m \times n$ matrix and b is an $m \times 1$ vector, with $m \gg n$. We would like to find an $n \times 1$ vector x such that $\|A\,x - b\|$ is close to its minimal value.

Standard methods for computing such an x cost $\mathcal{O}(n^2\,m)$.

Sarlós (2006) pointed out that x minimizing $\|R\,A\,x - R\,b\|$ works, where R is the random matrix of type [2] discussed previously. Because R consists of l rows of the product of an $m \times m$ discrete Fourier transform and a diagonal $m \times m$ matrix, computing $R\,A$ and $R\,b$ costs $\mathcal{O}(nm\,\ln(l))$. In all, then, constructing $R\,A$ and $R\,b$, and computing x minimizing $\|R\,A\,x - R\,b\|$ costs $\mathcal{O}(nm\,\ln(l) + n^2\,l)$.

$\mathcal{O}(nm\,\ln(l) + n^2\,l)$ can be less than $\mathcal{O}(n^2\,m)$ when $m \gg n$; current theory requires $l > n^2$, while $l = n + 8$ works well empirically.

Linear least-squares regression, 2

To show that x minimizing $\|R\,A\,x - R\,b\|$ nearly minimizes $\|A\,x - b\|$, we define U to be a matrix whose columns are an orthonormal basis of the subspace spanned by b and the columns of A, and

$$T = U\,((R\,U)^*\,(R\,U))^{-1}\,(R\,U)^*. \tag{4}$$

Combining (4) and the definition of U yields that

$$T\,R\,U = U; \quad T\,R\,A = A, \quad T\,R\,b = b. \tag{5}$$

Recall that, if the number l of rows in R is sufficiently large, then

$$1/\sigma_k(R\,U) \lesssim \sqrt{1+\varepsilon}, \tag{6}$$

where ε is a real number such that $0 < \varepsilon < 1$. Combining (4), (6), and the fact that the columns of U are orthonormal yields

$$\|T\| \lesssim \sqrt{1+\varepsilon}. \tag{7}$$

Linear least-squares regression, 3

We define x to be the $n \times 1$ vector minimizing $\|R\,A\,x - R\,b\|$, and y to be the $n \times 1$ vector minimizing $\|A\,y - b\|$. (5) and (7) yield

$$\begin{aligned} \|A\,x - b\| &= \|T\,R\,A\,x - T\,R\,b\| && (8)\\ &\le \|T\|\,\|R\,A\,x - R\,b\| && (9)\\ &\le \|T\|\,\|R\,A\,y - R\,b\| && (10)\\ &\lesssim \sqrt{1+\varepsilon}\,\|R\,A\,y - R\,b\|. && (11) \end{aligned}$$

Find vectors z and c such that

$$A\,y = U\,z, \quad b = U\,c. \tag{12}$$

(12) and the fact that the columns of U are orthonormal yield

$$\begin{aligned} \|R\,A\,y - R\,b\| &= \|R\,U\,z - R\,U\,c\| && (13)\\ &\le \|R\,U\|\,\|z - c\| && (14)\\ &= \|R\,U\|\,\|U\,z - U\,c\| && (15)\\ &= \|R\,U\|\,\|A\,y - b\|. && (16) \end{aligned}$$

Linear least-squares regression, 4

Recall that, if the number l of rows in R is sufficiently large, then

$$\|R\,U\| \lesssim \sqrt{1+\varepsilon}, \tag{17}$$

where ε is a real number such that $0 < \varepsilon < 1$.
Combining (11), (16), and (17) yields that

$$\|A\,x - b\| \lesssim (1+\varepsilon)\,\|A\,y - b\|. \tag{18}$$

Thus, $\|A\,x - b\|$ is nearly as small as possible with high probability, provided that the number l of rows in R is sufficiently large (where x minimizes $\|R\,A\,x - R\,b\|$).

Again, constructing $R\,A$ and $R\,b$, and computing x minimizing $\|R\,A\,x - R\,b\|$ costs $\mathcal{O}(nm\ln(l) + n^2\,l)$, less than the $\mathcal{O}(n^2\,m)$ needed to compute y minimizing $\|A\,y - b\|$, when $m \gg n$; current theory requires $l > n^2$, while $l = n + 8$ works well empirically.

Empirical performance, 1

We define the $m \times n$ matrix A via the formula

$$A = \sum_{k=1}^{n} 10^{-12(k-1)/(n-1)} \cdot u^{(k)} \cdot (v^{(k)})^*, \tag{19}$$

and the $m \times 1$ vector b via the formula

$$b_{m\times 1} = 10^{-9} \cdot u^{(n+1)} + \sum_{k=1}^{n} 10^{-12(k-1)/(n-1)} \cdot u^{(k)}, \tag{20}$$

with $u^{(1)}$, $u^{(2)}$, ..., $u^{(n)}$, $u^{(n+1)}$ being a set of orthonormal $m \times 1$ vectors, and $v^{(1)}$, $v^{(2)}$, ..., $v^{(n-1)}$, $v^{(n)}$ being an independent set of orthonormal $n \times 1$ vectors, both obtained by applying the Gram-Schmidt process to vectors whose entries are drawn i.i.d. from a pseudorandom number generator with a complex Gaussian distribution of zero mean and unit variance.

Empirical performance, 2

The vector x minimizing $\|A\,x - b\|$ is $x = \sum_{k=1}^{n} v^{(k)}$; for this x, $A\,x - b = -10^{-9} \cdot u^{(n+1)}$, and so $\|A\,x - b\| = 10^{-9}$.

We note that $\|A\| = 1$.

We will compare the randomized algorithm with the classical pivoted "$Q\,R$" decomposition algorithm based on plane (Householder) reflections.

We define x_{fast} to be the vector minimizing $\|R\,A\,x_{\text{fast}} - R\,b\|$.

We report the maximum value δ_{fast} of $\|A\,x_{\text{fast}} - b\|$ encountered during 300 randomized trials, and the average value t_{fast} over 300 randomized trials of the time in seconds taken to compute x_{fast} from A and b.

Empirical performance, 3

m	n	l	δ_{direct}	δ_{fast}
1024	8	16	.100E−08	.218E−08
2048	16	24	.100E−08	.295E−08
4096	32	40	.100E−08	.389E−08
8192	64	72	.100E−08	.476E−08
16384	128	136	.100E−08	.759E−08
32768	256	264	.100E−08	.107E−07

m	n	l	t_{direct}	t_{fast}	$t_{\text{direct}}/t_{\text{fast}}$
1024	8	16	.83E−03	.63E−03	1.3
2048	16	24	.52E−02	.24E−02	2.2
4096	32	40	.40E−01	.13E−01	3.1
8192	64	72	.34E−00	.56E−01	6.1
16384	128	136	.27E+01	.26E−00	10
32768	256	264	.22E+02	.15E+01	15

Empirical performance, 4

We performed all calculations in IEEE standard double-precision variables, whose mantissas have about one bit of precision less than 16 digits (so that the relative precision of the variables is approximately .2E–15).

We compiled the Fortran 77 code using the Lahey/Fujitsu Express v6.2 compiler, with the optimization flag `--o2` enabled.

We ran all computations on one core of a 1.86 GHz Intel Centrino Core Duo processor with 2 MB of L2 cache and 1 GB of RAM.

We employed a double-precision version of P. N. Swarztrauber's FFTPACK library for the fast Fourier transforms required to apply the matrix R.

Proof of the properties for type [2], 1

For a random matrix R of type [2], we have the following

Lemma. Suppose that $\delta, \varepsilon \in \mathbf{R}$, and $k, l, m \in \mathbf{Z}$, such that $\delta, \varepsilon < 1$, $\delta, \varepsilon, k, l, m > 0$, and $m > l > (1 + 1/\varepsilon)^2 \, k^2/\delta$. Suppose also that U is an $m \times k$ matrix whose columns are orthonormal.

Then,

$$\sigma_1(RU) \leq \sqrt{1 + \varepsilon} \tag{21}$$

$$1/\sigma_k(RU) \leq \sqrt{1 + \varepsilon} \tag{22}$$

with probability at least $1 - \delta$.

Hence, when acting from the left on the column space of $U_{m \times k}$, $R_{l \times m}$ preserves norms up to a distortion factor of $\sqrt{1 + \varepsilon}$.

Proof of the properties for type [2], 2

The lemma follows from the straightforward computation that

$$(RU)^* (RU) = \mathbf{1} + E, \tag{23}$$

where the expectation of the sum of the squares of the entries

of the $k \times k$ matrix E is at most $k^2/l < \delta\, \varepsilon^2/(1+\varepsilon)^2$; therefore,

$$\|E\| < \varepsilon/(1+\varepsilon) \tag{24}$$

with probability at least $1 - \delta$. Combining (23) and (24) yields

$$\|RU\|^2 = \|\mathbf{1} + E\| \le 1 + \|E\| \le 1 + \varepsilon \tag{25}$$

$$\left\| \Big((RU)^* (RU) \Big)^{-1} \right\| = \left\| (\mathbf{1} + E)^{-1} \right\| \le \sum_{j=0}^{\infty} \|E\|^j \le 1 + \varepsilon \tag{26}$$

with probability at least $1 - \delta$.

Three problems

1. Verifying that the spectral norm of a matrix is not meaningfully larger than desired

2. Linear least-squares regression (solving overdetermined systems of linear-algebraic equations in the least-squares sense)

3. Low-rank approximation of a matrix (including computing several of the greatest singular values and corresponding singular vectors)

We surveyed points 1 and 2 during the present tutorial talk, will start discussing point 3 with the next (self-contained) tutorial talk, and finish with a (self-contained) talk during Workshop II.