Randomized Algorithms for Matrix Computations

Ilse Ipsen

Students: John Holodnak, Kevin Penner, Thomas Wentworth

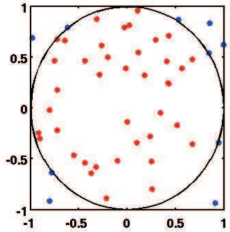
Research supported in part by NSF CISE CCF, DARPA XData

Randomized Algorithms

Solve a deterministic problem by statistical sampling

- Monte Carlo Methods

  Von Neumann & Ulam, Los Alamos, 1946



- Simulated Annealing: global optimization

Application of Randomized Algorithms

- Astronomy: Tamás Budavári (Johns Hopkins)
  Classification of galaxies
  Singular Value Decomposition/PCA, subset selection

- Nuclear Engineering: Hany Abdel-Khalik (NCSU)
  Model reduction in nuclear fission
  Low-rank approximation

- Population Genomics: Petros Drineas (RPI)
  Find SNPs that predict population association
  SVD/PCA, subset selection

- Software for Machine Learning Applications:
  Matrix multiplication, linear systems, least squares,
  SVD/PCA, low rank approximation, matrix functions

  Michael Mahoney (Stanford), Ken Clarkson and David Woodruf (IBM Almaden)

  Haim Avron and Christos Boutsidis (IBM Watson), Costas Bekas (IBM Zürich)

This Talk

Least squares/regression problems

- Traditional, deterministic algorithms
- A randomized algorithm

  Krylov space method
  Randomized preconditioner

- Random sampling of rows

  To construct preconditioned matrix
  Need: Sampled matrices with low condition numbers

- A probabilistic bound for condition numbers
- Coherence
- Leverage scores

Least Squares/Regression Problems: Traditional, Deterministic Algorithms

Least Squares for Dense Matrices

- Given:

  Real $m \times n$ *matrix* A *with* $\text{rank}(A) = n$
  Real $m \times 1$ *vector* b

- Want: $\min_x \|Ax - b\|_2$ (two-norm)

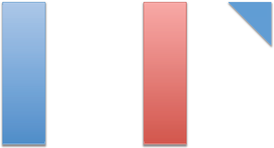
  Unique solution $x = A^\dagger b$

- Direct method:

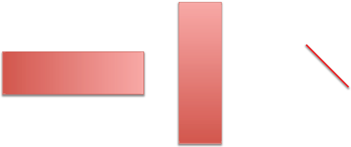
  QR factorization $A = QR$ $Q^T Q = I$, R *is* $\triangle$
  Solve $Rx = Q^T b$

  Operation count: $\mathcal{O}(mn^2)$

QR Factorization of Tall & Skinny Matrix



$A \quad = \quad Q \quad R$



$Q^T \quad Q \quad = \quad \mathbf{I}$

Least Squares for Dense Matrices

- Given:

  Real $m \times n$ *matrix* A *with* $\mathrm{rank}(A) = n$
  Real $m \times 1$ *vector* b

- Want: $\min_x \|Ax - b\|_2$ (two-norm)

  Unique solution $x = A^\dagger b$

- Direct method:

  QR factorization $A = QR$ $Q^T Q = I$, R *is* $\triangle$
  Solve $Rx = Q^T b$

  Operation count: $\mathcal{O}(mn^2)$

- But: Too expensive when A is large & sparse

  QR factorization produces fill-in

Least Squares for Sparse Matrices: LSQR [Paige & Saunders 1982]

Krylov space method

Iterates x_k computed with matrix vector multiplications

Residuals decrease (in exact arithmetic)

$$\|b - Ax_k\|_2 \leq \|b - Ax_{k-1}\|_2$$

Convergence fast if condition number $\kappa(A)$ small

$$\kappa(A) = \|A\|_2 \|A^\dagger\|_2 \geq 1$$

Accelerating Convergence

Find preconditioner P with

$\kappa(A\,P)$ small
Linear system $Px = rhs$ is easy to solve

Preconditioned least squares problem

Solve $\min_y \|A\,P\,y - b\|_2$ *by LSQR* $\{$*fast convergence*$\}$
Retrieve solution to original problem: $x = Py$

The ideal preconditioner

QR factorization $A = QR$ $\quad Q^T Q = I$, R is $\triangle$
Preconditioner $P = R^{-1} \quad \Rightarrow AR^{-1} = Q$
$\kappa(Q) = 1:$ *LSQR converges in 1 iteration!!!*

But: Construction of preconditioner much too expensive!

A Randomized Preconditioner for Least Squares Problems

A Cheaper, Randomized Preconditioner

QR factorization from only a few rows of A

- Sample $c \geq n$ rows of A: SA
- "Small" QR factorization

$$SA = Q_s R_s \qquad Q_s^T Q_s = I, \quad R_s \text{ is } \triangle$$

- Randomized preconditioner R_s^{-1}

QR Factorizations of Original and Sampled Matrices

Blendenpik [Avron, Maymounkov & Toledo 2010]

- Solve $\min_z \|Az - b\|_2$
- A is $m \times n$, $\mathrm{rank}(A) = n$ and $m \gg n$

{Construct preconditioner}

Sample $c \geq n$ rows of $A \rightarrow SA$ {fewer rows}

QR factorization $SA = Q_s R_s$

{Solve preconditioned problem}

Solve $\min_y \|AR_s^{-1}\, y - b\|_2$ with LSQR

Solve $R_s\, z = y$ {$\triangle$ system}

Hope:

AR_s^{-1} has almost orthonormal columns

Condition number almost perfect: $\kappa(AR_s^{-1}) \approx 1$

From Sampling to Condition Numbers

[Avron, Maymounkov & Toledo 2010]

- *Computed*
  QR decomposition of sampled matrix: $SA = Q_s R_s$
- *Conceptual*
  QR decomposition of full matrix: $A = QR$

The Idea

1. Sampling rows of A $\equiv$ Sampling rows of Q
2. Condition number of preconditioned matrix:

$$\kappa(AR_s^{-1}) = \kappa(SQ)$$

We analyze $\kappa(SQ)$
Sampled matrices with orthonormal columns

Random Sampling of Rows

Random Sampling of Rows

Given: Matrix Q with m rows

1. Sampling strategies pick indices:
   Sample c indices from $\{1, \ldots, m\}$
2. Construct sampled matrix SQ
3. Determine expected value of SQ

Three sampling strategies:

Sampling without replacement
Sampling with replacement (Exactly(c))
Bernoulli sampling

Uniform Sampling without Replacement

[Gittens & Tropp 2011, Gross & Nesme 2010]

Choose random permutation $k_1, \ldots, k_m$ of $1, \ldots, m$
Pick $k_1 \ldots, k_c$

Properties:

- Picks exactly c indices $k_1, \ldots, k_c$
- Each index is picked at most once
- The picks depend on each other

- Random permutation of m indices

  A permutation that is equally likely to occur among all $m!$ permutations

Uniform Sampling with Replacement (Exactly(c))

[Drineas, Kannan & Mahoney 2006]

for $t = 1 : c$ **do**
  Pick k_t from $\{1, \ldots, m\}$ with probability $1/m$
    independently and with replacement
end for

Properties:

- Picks exactly c indices $k_1, \ldots, k_c$
- An index can be picked more than once
- All picks are independent

Bernoulli Sampling

[Avron, Maymounkov & Toledo 2010, Gittens & Tropp 2011]

for $j = 1 : m$ **do**
  Pick j with probability c/m
end for

Properties:

- Samples each index at most once
- Expected number of sampled indices: c
- All picks are independent

Construction of Sampled Matrix

- Sampling with and without replacement:

$$SQ = \sqrt{\frac{m}{c}} \begin{pmatrix} \text{row } k_1 \text{ of } Q \\ \vdots \\ \text{row } k_c \text{ of } Q \end{pmatrix}$$

  Sampled matrix has c rows from Q

- Bernoulli sampling:

$$\text{row } j \text{ of } SQ = \sqrt{\frac{m}{c}} \begin{cases} \text{row } j \text{ of } Q & \text{with probability } \frac{c}{m} \\ 0 & \text{with probability } 1 - \frac{c}{m} \end{cases}$$

  Sampled matrix has m rows
  Expected number of rows from Q is c

Approach towards Expected Value of SQ

- Two-norm condition number:

$$\kappa(SQ)^2 = \kappa\left((SQ)^T(SQ)\right)$$

  $(SQ)^T(SQ)$ is a symmetric matrix

- Matrix product $(SQ)^T(SQ)$: Sum of outer products
- Expected value of one sampled outer product
- Expected value of $(SQ)^T(SQ)$

Matrix Product = Sum of Outer Products

$$Q = \begin{pmatrix} x & x \\ x & x \\ x & x \end{pmatrix}$$

$$\begin{aligned} \text{Product} \quad Q^T Q \quad &= \begin{pmatrix} x & x & x \\ x & x & x \end{pmatrix} \begin{pmatrix} x & x \\ x & x \\ x & x \end{pmatrix} \\ &= \begin{pmatrix} x \\ x \end{pmatrix} \begin{pmatrix} x & x \end{pmatrix} + \begin{pmatrix} x \\ x \end{pmatrix} \begin{pmatrix} x & x \end{pmatrix} + \begin{pmatrix} x \\ x \end{pmatrix} \begin{pmatrix} x & x \end{pmatrix} \\ &= \text{Sum of outer products} \end{aligned}$$

Outer Product Representations

Q has m rows

$$Q = \begin{pmatrix} r_1 \\ \vdots \\ r_m \end{pmatrix} \quad \text{with} \quad r_1^T r_1 + \cdots + r_m^T r_m = Q^T Q$$

Sampled matrix SQ has c rows

$$SQ = \sqrt{\frac{m}{c}} \begin{pmatrix} r_{k_1} \\ \vdots \\ r_{k_c} \end{pmatrix}$$

with

$$\frac{m}{c} \left(r_{k_1}^T r_{k_1} + \cdots + r_{k_c}^T r_{k_c} \right) = (SQ)^T (SQ)$$

Expected Value of Sampled Matrix

Given: Row r_{k_1} of Q from sampling with replacement

Expected value of matrix random variable

$$\begin{aligned} \mathbf{E}[r_{k_1}^T r_{k_1}] &= \tfrac{1}{m} r_1^T r_1 + \cdots + \tfrac{1}{m} r_m^T r_m \\ &= \tfrac{1}{m} \left(r_1^T r_1 + \cdots + r_m^T r_m \right) \\ &= \tfrac{1}{m}\, Q^T Q \end{aligned}$$

Linearity of expected value

$$\begin{aligned} \mathbf{E}\left[(SQ)^T (SQ) \right] &= \tfrac{m}{c} \left(\mathbf{E}[r_{k_1}^T r_{k_1}] + \cdots + \mathbf{E}[r_{k_c}^T r_{k_c}] \right) \\ &= \tfrac{m}{c} \left(\tfrac{1}{m} Q^T Q + \cdots + \tfrac{1}{m} Q^T Q \right) = Q^T Q \end{aligned}$$

$(SQ)^T (SQ)$ is unbiased estimator of $Q^T Q$

Expected Value when Sampling from Matrices with Orthonormal Columns

- Given: Matrix Q with orthonormal columns, $Q^TQ = I$
- Matrix SQ produced by

  Sampling with replacement
  Sampling without replacement
  Bernoulli sampling

- Expected value $\mathbf{E}\left[(SQ)^T(SQ)\right] = Q^TQ = I$

  Cross product of sampled matrix is
  Unbiased estimator of identity matrix

- "In expectation": $\kappa(SQ) = 1$
  Expected value of sampled matrix perfectly conditioned

Where We Are Now in the Talk

- Solution of least squares problems $\min_x \|Ax - b\|_2$
- Krylov space method + randomized preconditioner
- Condition number of preconditioned matrix:

$$\kappa(SQ) = \|SQ\|_2 \, \|(SQ)^\dagger\|_2$$

  Q has orthonormal columns, $Q^T Q = I$

- Sampled matrix SQ
  Three different strategies to sample rows from Q
- Expected value of sampled matrix

$$\mathbf{E}\left[(SQ)^T(SQ)\right] = Q^T Q = I$$

Issues

- How to implement sampling?
- Does sampling with replacement really make sense here?
- How large are the condition numbers of the sampled matrices?
- How many sampled matrices are rank deficient?
- Can we get bounds for the condition numbers of the sampled matrices?

  How does the expected value differ
  from the actually sampled value?

How to Implement Uniform Sampling With Replacement

Sample k_t from $\{1, \ldots, m\}$ with probability $1/m$ independently and with replacement

- [Devroye 1986]

$$\eta = rand \qquad \{\textit{uniform } [0,1] \textit{ random variable}\}$$
$$k_t = \lfloor 1 + m\,\eta \rfloor$$

- Matlab: $k_t = \texttt{randi(m)}$

Condition Numbers and Rank Deficiency

Sampling c rows from $m \times n$ matrix Q with $Q^T Q = I$
$m = 10^4$, $n = 5$ (30 runs for each value of c)

Sampled matrices SQ from three strategies:

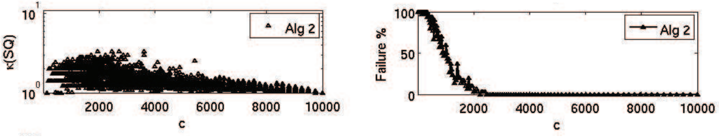
Sampling without replacement
Sampling with replacement
Bernoulli sampling

Plots:

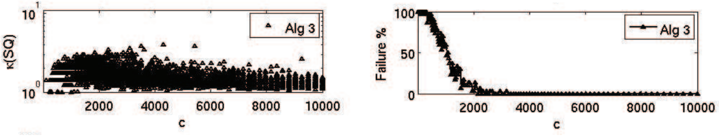
1. Condition number of SQ
   $\kappa(SQ) = \|SQ\|_2 \, \|(SQ)^\dagger\|_2$ (if SQ has full column rank)
2. Percentage of matrices SQ that are rank deficient
   ($\kappa(SQ) \geq 10^{16}$)

Comparison of Sampling Strategies

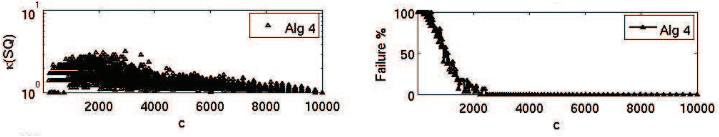
Sampling without replacement



Sampling with replacement (Exactly(c))



Bernoulli sampling



Comparison of Sampling Strategies

Sampled matrices SQ from three strategies:

Little difference among the sampling strategies

Same behavior for small c

If SQ has full rank, then very well conditioned

$$\kappa(SQ) \leq 10$$

Advantages of sampling with replacement

Fast: Need to generate/inspect only c values

Easy to implement, embarrassingly parallel

Replacement does not affect accuracy

(for small amounts of sampling)

A Probabilistic Bound for
The Condition Number of the Sampled Matrices

Setup for the Bound

- Given: $m \times n$ matrix Q with orthonormal columns
- Sampling c rows from Q

$$Q = \begin{pmatrix} r_1 \\ \vdots \\ r_m \end{pmatrix} \qquad SQ = \sqrt{\frac{m}{c}} \begin{pmatrix} r_{k_1} \\ \vdots \\ r_{k_c} \end{pmatrix}$$

- Unbiased estimator: $\mathbf{E}\left[(SQ)^T(SQ)\right] = Q^TQ = I$
- Sum of c random matrices: $(SQ)^T SQ = X_1 + \cdots + X_c$

$$X_t = \frac{m}{c}\, r_{k_t} r_{k_t}^T \qquad 1 \leq t \leq c$$

Matrix Bernstein Concentration Inequality [Recht 2011]

- Y_t independent random $n \times n$ matrices with $\mathbf{E}[Y_t] = 0$
- $\|Y_t\| \leq \tau$ almost surely (two-norm)
- $\rho_t \equiv \max\{\|\mathbf{E}[Y_t Y_t^T]\|, \|\mathbf{E}[Y_t^T Y_t]\|\}$
- Desired error $0 < \epsilon < 1$
- Failure probability $\delta = 2n \exp\left(-\frac{3}{2}\,\frac{\epsilon^2}{3\sum_t \rho_t + \tau\,\epsilon}\right)$

With probability at least $1 - \delta$

$$\left\|\sum_t Y_t\right\| \leq \epsilon \qquad \{\text{Deviation from mean}\}$$

Applying the Concentration Inequality

- Sampled matrix:

$$(SQ)^T(SQ) = X_1 + \cdots + X_c, \quad X_t = \tfrac{m}{c}\, r_{k_t} r_{k_t}^T$$

- Zero mean version:

$$(SQ)^T(SQ) - I = Y_1 + \cdots + Y_c, \quad Y_t = X_t - \tfrac{1}{c} I$$

- By construction: $\mathbf{E}[Y_t] = 0$

$$\|Y_t\| \leq \tfrac{m}{c}\, \mu, \qquad \mathbf{E}[Y_t^2] \leq \tfrac{m}{c^2}\, \mu$$

Largest row norm squared: $\mu = \max_{1 \leq j \leq m} \|r_j\|^2$

With probability at least $1 - \delta$, $\quad \|(SQ)^T(SQ) - I\| \leq \epsilon$

Condition Number Bound

- $m \times n$ matrix Q with orthonormal columns
- Largest row norm of Q squared: $\mu = \max_{1 \leq j \leq m} \|r_j\|^2$
- Number of rows to be sampled: $c \geq n$
- $0 < \epsilon < 1$

Failure probability

$$\delta = 2n \exp\left(-\frac{c}{m\,\mu}\,\frac{\epsilon^2}{3+\epsilon}\right)$$

With probability at least $1 - \delta$:

$$\kappa(SQ) \leq \sqrt{\frac{1+\epsilon}{1-\epsilon}}$$

Tightness of Condition Number Bound

Input: $m \times n$ *matrix* Q *with* $Q^T Q = I$ *with*
$m = 10^4$, $n = 5$, $\mu = 1.5\, n/m$

1. Exact condition number from sampling with replacement

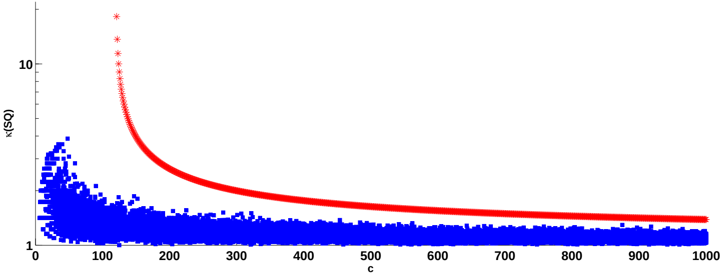
   Little sampling: $n \leq c \leq 1000$
   A lot of sampling: $1000 \leq c \leq m$

2. Condition number bound $\sqrt{\frac{1+\epsilon}{1-\epsilon}}$

   where success probability $1 - \delta \equiv .99$

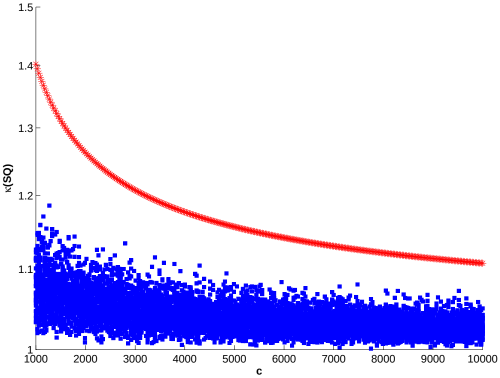
$$\epsilon \equiv \frac{1}{2c}\left(\ell + \sqrt{12c\ell + \ell^2}\right) \qquad \ell \equiv \tfrac{2}{3}\,(m\,\mu - 1)\,\ln(2n/\delta)$$

Little sampling ($n \leq c \leq 1000$)



Bound holds for $c \geq 93 \approx 2(m\mu - 1)\ \ln(2n/\delta)/\epsilon^2$

A lot of sampling ($1000 \leq c \leq m$)



Bound predicts correct magnitude of condition number

Condition Number Bound

- $m \times n$ matrix Q with orthonormal columns
- Largest row norm of Q squared: $\mu = \max_{1 \leq j \leq m} \|r_j\|^2$
- Number of rows to be sampled: $c \geq n$
- $0 < \epsilon < 1$
- Failure probability

$$\delta = 2n \,\exp\left(-\frac{c}{m\,\mu}\,\frac{\epsilon^2}{3+\epsilon}\right)$$

With probability at least $1-\delta$:

$$\kappa(SQ) \leq \sqrt{\frac{1+\epsilon}{1-\epsilon}}$$

The only distinction among different $m \times n$ matrices Q with orthonormal columns is μ

Conclusions from the Bound

- Correct magnitude for condition number of sampled matrix, even for small matrix dimensions

- Lower bound on number of sampled rows

$$c = \mathcal{O}\left(m\,\mu \ln n\right)$$

- Important ingredient: Largest row norm of Q squared

$$\mu = \max_{1 \le j \le m} \|\text{row } j \text{ of } Q\|^2$$

Coherence

$$\mu = \max_{1 \le j \le m} \|\text{row } j \text{ of } Q\|^2$$

Largest squared row norm of
matrix Q with orthonormal columns

Properties of Coherence

Coherence of $m \times n$ matrix Q with $Q^T Q = I$

$$\mu = \max_{1 \le j \le m} \|\text{row } j \text{ of } Q\|^2$$

- $n/m \le \mu(Q) \le 1$
- Maximal coherence: $\mu(Q) = 1$
  At least one column of Q is a canonical vector
- Minimal coherence: $\mu(Q) = n/m$
  Columns of Q are columns of a Hadamard matrix
- Coherence measures "correlation with standard basis"
- Indicator for how well the "mass" of the matrix is distributed

Coherence in General

- Donoho & Huo 2001
  Mutual coherence of two bases
- Candés, Romberg & Tao 2006
- Candés & Recht 2009
  Matrix completion: Recovering a low-rank matrix by sampling its entries
- Mori & Talwalkar 2010, 2011
  Estimation of coherence
- Avron, Maymounkov & Toledo 2010
  Randomized preconditioners for least squares
- Drineas, Magdon-Ismail, Mahoney & Woodruff 2011
  Fast approximation of coherence

Different Definitions

- Coherence of subspace
  $\mathcal{Q}$ is subspace of $\mathbb{R}^m$ of dimension n
  P orthogonal projector onto $\mathcal{Q}$

$$\mu_0(\mathcal{Q}) = \frac{m}{n} \max_{1\leq j\leq m} \|\text{row } j \text{ of } P\|^2 \qquad (1 \leq \mu_0 \leq \tfrac{m}{n})$$

- Coherence of full rank matrix
  A is $m \times n$ with $\mathrm{rank}(A) = n$
  Columns of Q are orthonormal basis for $\mathcal{R}(A)$

$$\mu(A) = \max_{1\leq j\leq m} \|\text{row } j \text{ of } Q\|^2 \qquad (\tfrac{n}{m} \leq \mu \leq 1)$$

- Reflects difficulty of recovering the matrix from sampling

Effect of Coherence on Sampling

Input: $m \times n$ *matrix* Q *with* $Q^TQ = I$

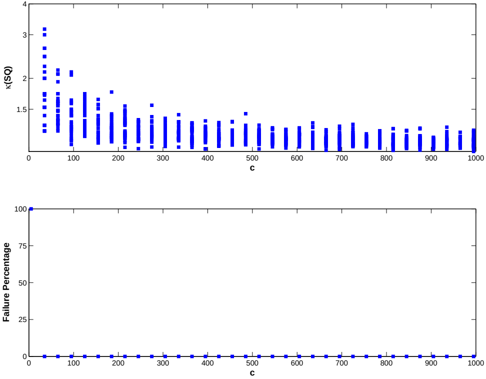
Coherence $\mu = \max_{1 \le j \le m} \|$*row* j *of* $Q\|^2$

$m = 10^4$, $n = 5$

Sampling with replacement

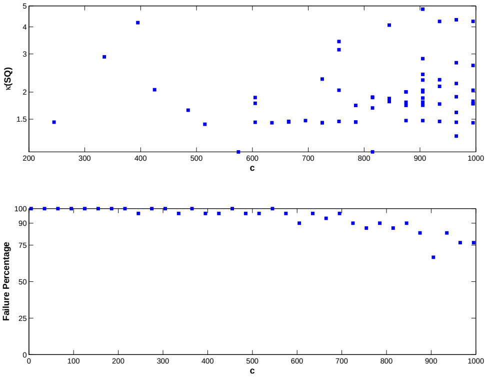
1. Low coherence: $\mu = 7.5 \cdot 10^{-4} = 1.5\, n/m$
2. Higher coherence: $\mu = 7.5 \cdot 10^{-2} = 150\, n/m$

Low Coherence



Only a single rank deficient matrix (for $c = 5$)

Higher Coherence



Most matrices rank deficient, when sampling at most 10% of rows

Improving on Coherence:
Leverage Scores

Leverage Scores

Idea: Use all row norms

- Q is $m \times n$ with orthonormal columns
- Leverage scores = squared row norms

$$\ell_k = \|\text{row } k \text{ of } Q\|^2 \qquad 1 \leq k \leq m$$

- Coherence $\mu = \max_k \ell_k$
- Low coherence $\approx$ uniform leverage scores

- Leverage scores of full column rank matrix A:
  Columns of Q are orthonormal basis for $\mathcal{R}(A)$

$$\ell_k(A) = \|\text{row } k \text{ of } Q\|^2 \qquad 1 \leq k \leq m$$

Statistical Leverage Scores

Hoaglin & Welsch 1978
Chatterjee & Hadi 1986

- Identify potential outliers in $\min_x \|Ax - b\|$
- Hb: Projection of b onto $\mathcal{R}(A)$ where $H = A(A^T A)^{-1} A^T$
- Leverage score: $H_{kk} \sim$ influence of kth data point on LS fit

- QR decomposition: $A = QR$

$$H_{kk} = \|\text{row } k \text{ of } Q\|^2 = \ell_k(A)$$

Application to randomized algorithms: Mahoney & al. 2006–2012

Leverage Score Bound

- $m \times n$ matrix Q with orthonormal columns
- Leverage scores

$$\ell_j = \|\text{row } j \text{ of } Q\|^2 \qquad L = \operatorname{diag}\begin{pmatrix}\ell_1 & \dots & \ell_m\end{pmatrix}$$

- Coherence $\mu = \max_{1 \le j \le m} \ell_j$
- $0 < \epsilon < 1$

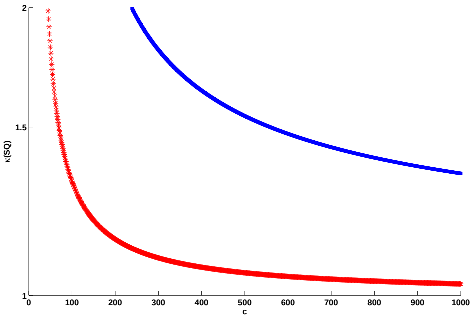
Failure probability

$$\delta = 2n \exp\left(-\frac{3}{2}\, \frac{c\, \epsilon^2}{m\,(3\, \|Q^T L Q\|_2 + \mu\, \epsilon)}\right)$$

With probability at least $1 - \delta$: $\kappa(SQ) \le \sqrt{\frac{1+\epsilon}{1-\epsilon}}$

Improvement with Leverage Score Bound

Low coherence: $\mu = 1.5n/m$, small amounts of sampling



Leverage score bound vs. Coherence bound

($m = 10^4$, $n = 4$, $\delta = .01$)

Summary

Least squares problems $\min_z \|Az - b\|_2$
A *is* $m \times n$, $\operatorname{rank}(A) = n$ *and* $m \gg n$

- Solution by iterative Krylov space method LSQR
- Randomized preconditioner
- Condition number of preconditioned matrix $= \kappa(SQ)$

  Q has orthonormal columns, SQ is sampled matrix
  Three strategies to sample rows from Q

- Bounds for condition number of sampled matrix SQ

  Explicit, non-asymptotic, predictive even for small matrix dimensions

- Sampling accuracy depends on coherence
- Tighter bounds: Replace coherence by leverage scores

Existing Work

Randomized algorithms for least squares

Drineas, Mahoney & Muthukrishnan 2006
Drineas, Mahoney, Muthukrishnan & Sarlós 2006
Rokhlin & Tygert 2008
Boutsidis & Drineas 2009
Blendenpik: Avron, Maymounkov & Toledo 2010
LSRN: Meng, Saunders & Mahoney 2011

Survey papers for randomized algorithms

Halko, Martinsson & Tropp 2011
Mahoney 2011

Graph theory

Preconditioning for graph Laplacians, graph sparsification

Spielman & Teng 2006, Koutis, Miller & Peng 2012, ...

Effective resistance = leverage scores

Drineas & Mahoney 2010