Randomly Sampling Rows from Orthonormal Matrices with Application to Least Squares Problems

Ilse Ipsen

Joint work with Thomas Wentworth

North Carolina State University
Raleigh, NC, USA

Research supported in part by NSF CISE CCF

This Talk

Given:

- Real $m \times n$ matrix Q with orthonormal columns, $Q^T Q = I_n$
- Real $c \times m$ "sampling" matrix S with $c \ll m$

Want: Probability that

1. SQ has full column rank $\quad (\mathrm{rank}(SQ) = n)$
2. Condition number: Given η

$$\kappa(SQ) = \|SQ\|_2 \, \|(SQ)^\dagger\|_2 \leq 1 + \eta$$

Analysis:

Probabilistic bound for eigenvalues of $(SQ)^T(SQ)$

Outline

1. Motivation: *Blendenpik*
   A randomized preconditioned least squares solver
2. Sampling rows from matrices with orthonormal columns
   - *Three different strategies*
   - *Numerical comparison*
3. Probabilistic condition number bounds
4. The important property: Coherence
5. Generating matrices with user-specified coherence
6. Improving on coherence: Leverage scores
7. Summary

Motivation: *Blendenpik*
A Randomized Preconditioned Least Squares Solver

Existing Work

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

- Apply QR to sampled rows from (preprocessed) A

  Drineas, Mahoney & Muthukrishnan 2006
  Drineas, Mahoney, Muthukrishnan & Sarlós 2006
  Boutsidis & Drineas 2009

- Preconditioned iterative methods

  Rokhlin & Tygert 2008
  Blendenpik: Avron, Maymounkov & Toledo 2010
  LSRN: Meng, Saunders & Mahoney 2011

- Survey papers

  Halko, Martinsson & Tropp 2011
  Mahoney 2011

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 $A \rightarrow SA$

Thin QR decomposition $SA = Q_s R_s$

{Solve preconditioned problem}

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

Solve $R_s\, z = y$

- Idea: AR_s^{-1} is almost orthonormal
- LSQR converges fast if $\kappa(AR_s^{-1}) \approx 1$

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: $A = QR$

  Sampling rows of A $\equiv$ *Sampling rows of Q*

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

- Preconditioned matrix $A\,R_s^{-1} = QR\,R_s^{-1} \quad \rightarrow RR_s^{-1}$
- Sampled orthonormal matrix

$$\begin{aligned} S\,Q &= S\,AR^{-1} \\ &= SA\,R^{-1} = Q_s R_s\,R^{-1} \quad \rightarrow R_s R^{-1} \end{aligned}$$

From Preconditioned to Orthonormal Matrices

[Avron, Maymounkov & Toledo 2010]

- Blendenpik computes:
  QR decomposition of sampled matrix $SA = Q_s R_s$
- Conceptual aid:
  QR decomposition of whole matrix $A = QR$
- Condition number of preconditioned matrix:

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

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

Sampling Rows from Matrices with Orthonormal Columns

Different Sampling Procedures

Want to uniformly sample c rows from m rows

1. Sampling without replacement

   Each row is sampled at most once
   Number of sampled rows is equal to c

2. Sampling with replacement (Exactly(c))

   A row may be sampled more than once
   Number of sampled rows is equal to c

3. Bernoulli sampling

   Each row is sampled at most once
   Number of sampled rows not known in advance
   Expected value of number of sampled rows equals c

Uniform Sampling without Replacement

[Gittens & Tropp 2011, Gross & Nesme 2010]

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

Sampling matrix $S = \begin{pmatrix} e_{k_1}^T \\ \vdots \\ e_{k_c}^T \end{pmatrix}$

- S is $c \times m$, and samples *exactly* c rows
- Expected value $\mathbf{E}(S^T S) = \frac{c}{m} I_m$

Uniform Sampling with Replacement (Exactly(c))

[Drineas, Kannan & Mahoney 2006]

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

Sampling matrix $\quad S = \sqrt{\frac{m}{c}} \begin{pmatrix} e_{k_1}^T \\ \vdots \\ e_{k_c}^T \end{pmatrix}$

- S is $c \times m$, and samples *exactly* c rows
- Expected value $\mathbf{E}(S^T S) = I_m$
- S can sample a row more than once

Bernoulli Sampling

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

$S = 0_{m \times m}$
for $j = 1 : m$ **do**

$$S_{jj} = \begin{cases} 1 & \text{with probability } \frac{c}{m} \\ 0 & \text{with probability } 1 - \frac{c}{m} \end{cases}$$

end for

- S is $m \times m$, and samples each row at most once
- Expected value $\mathbf{E}(S^T S) = \frac{c}{m}\, I_m$
- Expected number of sampled (non zero) rows: c

Comparison of Sampling Strategies

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

Sampled matrices SQ from three strategies:

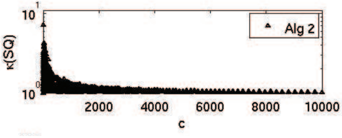
Sampling without replacement
Sampling with replacement (Exactly(c))
Bernoulli sampling

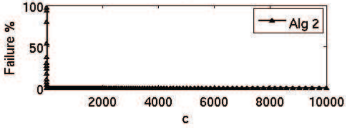
Plots:

1. Two-norm 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

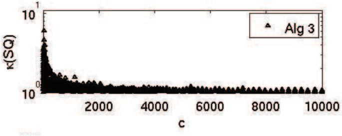
First Comparison

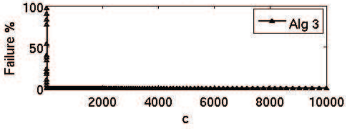
Sampling without replacement



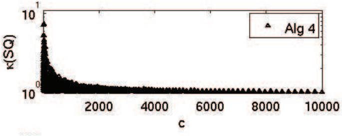


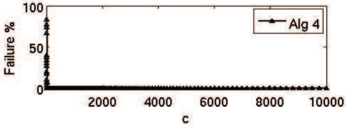
Sampling with replacement (Exactly(c))





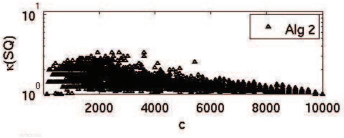
Bernoulli sampling

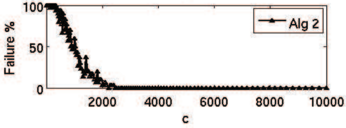




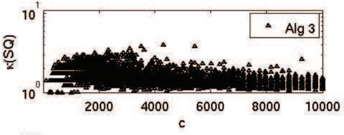
Second Comparison

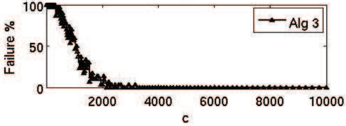
Sampling without replacement



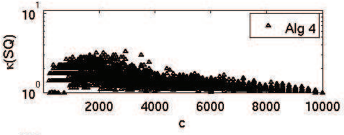


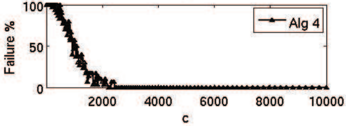
Sampling with replacement (Exactly(c))





Bernoulli sampling





Comparison of Sampling Strategies

Sampled matrices SQ from three strategies:

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

Summary

Little difference among the sampling strategies
If SQ has full rank then $\kappa(SQ) \leq 10$

Rest of the talk: Sampling with replacement

Fast: need to generate/inspect only c values
Easy to implement
Replacement does not affect accuracy
(for small amounts of sampling)

Probabilistic Condition Number Bounds

Sampling with Replacement (Exactly(c))

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

$$S = \sqrt{\frac{m}{c}} \begin{pmatrix} e_{k_1}^T \\ \vdots \\ e_{k_c}^T \end{pmatrix}$$

- Unbiased estimator: $\mathbf{E}\left[Q^T S^T S Q\right] = Q^T Q = I_n$
- Sum of c random matrices: $Q^T S^T S Q = X_1 + \cdots + X_c$

$$X_t = \tfrac{m}{c}\, Q^T e_{k_t} e_{k_t}^T Q, \qquad 1 \leq t \leq c$$

Bernstein-Type Concentration Inequality [Recht 2011]

- Y_t independent random $n \times n$ matrices with $\mathbf{E}[Y_t] = 0$
- $\|Y_t\|_2 \leq \tau$ almost surely
- $\rho_t \equiv \max\{\|\mathbf{E}[Y_t Y_t^T]\|_2, \|\mathbf{E}[Y_t^T Y_t]\|_2\}$
- Desired error $0 < \epsilon < 1$
- Failure probability $\quad \delta = 2n \exp\left(-\frac{3}{2} \frac{\epsilon^2}{3\sum_t \rho_t + \tau\, \epsilon}\right)$

With probability at least $1 - \delta$

$$\left\| \sum_t Y_t \right\|_2 \leq \epsilon \qquad \{\text{Deviation from mean}\}$$

Applying the Concentration Inequality

- Sampled matrix:

$$Q^T S^T S Q = X_1 + \cdots + X_c, \quad X_t = \tfrac{m}{c} Q^T e_{k_t} e_{k_t}^T Q$$

- Zero mean version:

$$Q^T S^T S Q - I_n = Y_1 + \cdots + Y_c, \quad Y_t = X_t - \tfrac{1}{c} I_n$$

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

$$\|Y_t\|_2 \leq \tfrac{m}{c}\,\mu, \qquad \mathbf{E}[Y_t^2] \leq \tfrac{m}{c^2}\,\mu$$

Largest row norm squared: $\mu = \max_{1 \leq j \leq m} \|e_j^T Q\|_2^2$

With probability at least $1 - \delta$, $\|(SQ)^T (SQ) - I_n\|_2 \leq \epsilon$

Condition Number Bound

- $m \times n$ matrix Q with orthonormal columns
- Largest row norm squared: $\mu = \max_{1 \le j \le m} \|e_j^T Q\|_2^2$
- Number of rows to be sampled: $c \ge n$
- $0 < \epsilon < 1$

Failure probability

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

With probability at least $1 - \delta$:

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

Tightness of Condition Number Bound

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

1. Exact condition number from sampling with replacement

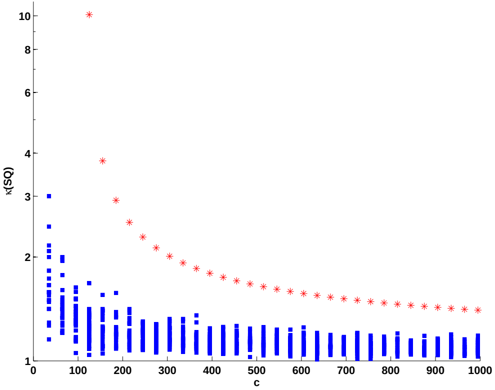
   Little sampling: $n \leq c \leq 1000$
   A lot of sampling: $1000 \leq c \leq m$

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

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

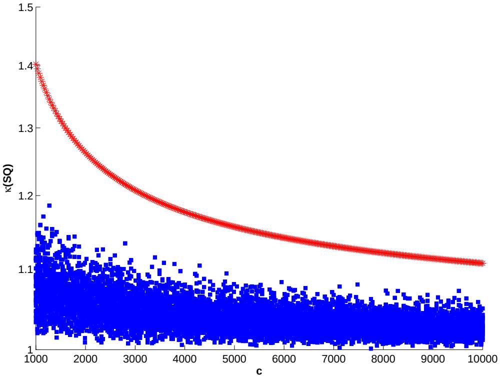
$$\epsilon \equiv \frac{1}{2c} \left(\ell + \sqrt{12c\ell + \ell^2} \right) \qquad \ell \equiv \tfrac{2}{3} \, (m\,\mu - 1) \, \ln(2n/\,\delta)$$

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 squared: $\mu = \max_{1 \le j \le m} \|e_j^T Q\|_2^2$
- Number of rows to be sampled: $c \ge n$
- $0 < \epsilon < 1$
- Failure probability

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

With probability at least $1 - \delta$:

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

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

Conclusions from the Bound

Input: $m \times n$ matrix Q with $\mu = \max_{1 \leq j \leq m} \|e_j^T Q\|_2^2$

- Correct magnitude for condition number of sampled matrix, even for small matrix dimensions
- Required number of samples $c = \mathcal{O}\left(m\,\mu \ln n\right)$
- Slightly tighter bound for failure probability

$$\delta \equiv n \left\{ \left(e^{-\epsilon} (1-\epsilon)^{-(1-\epsilon)} \right)^{c/(m\,\mu)} + \left(e^{\epsilon} (1+\epsilon)^{-(1+\epsilon)} \right)^{c/(m\,\mu)} \right\}$$

use [Tropp 2011]

- Similar bounds for

  Sampling without replacement
  Bernoulli sampling

- Important ingredient $\mu = \max_{1 \leq j \leq m} \|e_j^T Q\|_2^2$

The Important Property: Coherence

Coherence = Largest Row Norm^2

Q is $m \times n$ with orthonormal columns: $\mu = \max_{1 \le j \le m} \|e_j^T Q\|_2^2$

$$Q = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{pmatrix}$$

high coherence: $\mu = 1$

$$Q = \begin{pmatrix} 1/\sqrt{6} & -1/\sqrt{6} \\ 1/\sqrt{6} & 1/\sqrt{6} \\ 1/\sqrt{6} & -1/\sqrt{6} \\ 1/\sqrt{6} & 1/\sqrt{6} \\ 1/\sqrt{6} & -1/\sqrt{6} \\ 1/\sqrt{6} & 1/\sqrt{6} \end{pmatrix}$$

low coherence $\mu = \frac{1}{3}$

Properties of Coherence

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

$$\mu = \max_{1 \leq j \leq m} \|e_j^T Q\|_2^2$$

- $n/m \leq \mu(Q) \leq 1$
- 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"

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 $\mathcal{Q}$

$$\mu_0(\mathcal{Q}) = \frac{m}{n} \max_{1 \leq j \leq m} \|e_j^T P\|_2^2 \qquad (1 \leq \mu_0 \leq \tfrac{m}{n})$$

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

$$\mu(A) = \max_{1 \leq j \leq m} \|e_j^T Q\|_2^2 \qquad (\tfrac{n}{m} \leq \mu \leq 1)$$

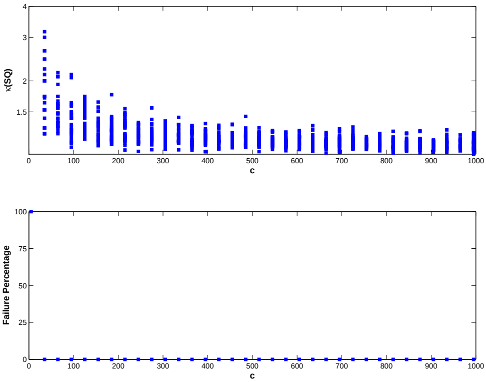
- Reflects difficulty of recovering the matrix from sampling

Effect of Coherence on Sampling

Input: $m \times n$ *matrix* Q *with* $Q^T Q = I_n$
$m = 10^4$, $n = 5$
Sampling with replacement

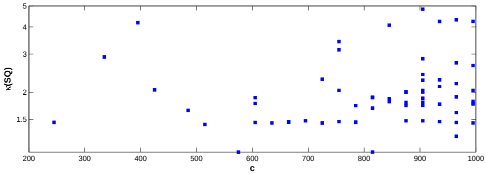
1. Low coherence: $\mu = 7.5 \cdot 10^{-4} = 1.5\, n/m$
2. Higher coherence: $\mu = 7.5 \cdot 10^{-2} = 150\, n/m$

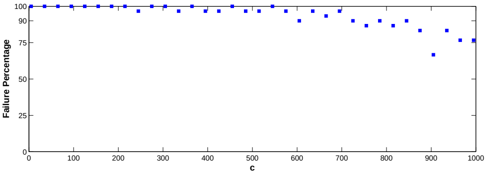
Low Coherence



Only a single failure (for $c = 5$)

Higher Coherence





Very high failure rate when sampling at most 10% of rows

Coherence Isn't Everything

$$G_{ood} = \begin{pmatrix} 1/2 & 0 \\ 1/2 & 0 \\ 1/2 & 0 \\ 1/2 & 0 \\ 0 & -1/2 \\ 0 & -1/2 \\ 0 & 1/\sqrt{2} \end{pmatrix} \qquad B_{ad} = \begin{pmatrix} 1/\sqrt{2} & 0 \\ -1/\sqrt{2} & 0 \\ 0 & 1/\sqrt{2} \\ 0 & -1/\sqrt{2} \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{pmatrix}$$

- Same coherence: $\mu(G_{ood}) = \mu(B_{ad}) = 1/2$
- Sampling with replacement: $c = 3$

$$\textbf{Prob}[SG_{ood} \textit{ has full column rank}] \geq 73\%$$
$$\textbf{Prob}[SB_{ad} \textit{ has full column rank}] < 35\%$$

- Sampled bad matrices more likely to be rank deficient

Generating Matrices With User-Specified Coherence

Good Matrices with Specified Coherence

Algorithm for generating Hermitian matrices with prescribed diagonal elements and eigenvalues [Dhillon, Heath, Sustik & Tropp 2005]

Input: Dimensions m and n with $m \geq n$
Desired row norms^2 ℓ_j, $1 \leq j \leq m$

Output: $m \times n$ matrix Q with orthonormal columns
Row norms^2 $\|e_j^T Q\|_2^2 = \ell_j$
Coherence $\mu = \max_{1 \leq j \leq m} \ell_j$

Initialize $Q_0 = \begin{pmatrix} I_n \\ 0 \end{pmatrix}$
Rotate rows of Q_0 until row norms ℓ_j achieved

Bad Matrices with Specified Coherence

Idea

Lower bound for coherence: $\mu \geq n/m$
Given n *and* μ*, minimal number rows is* $m_0 = \lceil n/\mu \rceil$

Algorithm

Initialize $m_0 = \lceil n/\mu \rceil$
Generate $m_0 \times n$ matrix Q_0 with coherence μ
Set $Q = \begin{pmatrix} Q_0 \\ 0_{(m-m_0)\times n} \end{pmatrix}$

Q has coherence μ and maximal number of zero rows

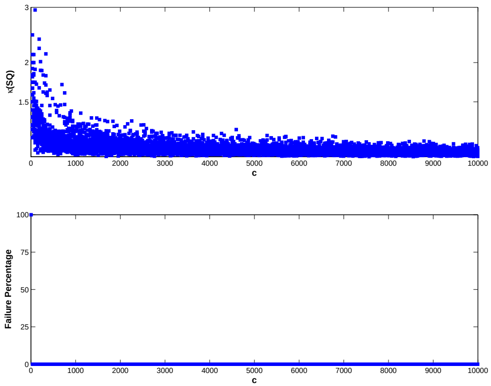
Difference between Good and Bad Matrices

Input: $m \times n$ *matrices* Q *with* $Q^T Q = I_n$
$m = 10^4$, $n = 5$, $\mu = .05$
Sampling with replacement

Two matrices with same coherence

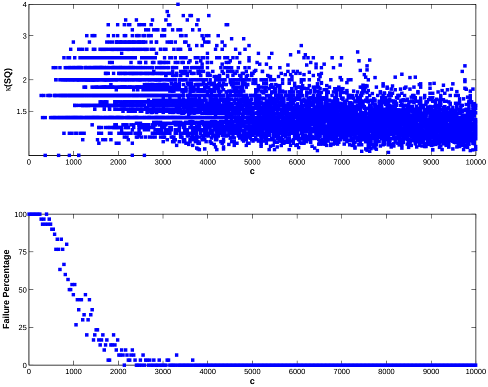
1. Good matrices: No zero rows
2. Bad matrices: 9900 zero rows

Good Matrices



Only a single failure (for $c = 5$)

Bad Matrices



High failure percentage when sampling at most 20% of rows

Improving on Coherence: Leverage Scores

Distinguishing Good and Bad Matrices with Same Coherence

Idea: Use all row norms

- Q is $m \times n$ with orthonormal columns
- Leverage scores = row norms^2

$$\ell_k = \|e_k^T Q\|_2^2, \qquad 1 \leq k \leq m$$

- 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 $\mathcal{R}(A)$

$$\ell_k(A) = \|e_k^T Q\|_2^2, \qquad 1 \leq k \leq m$$

Statistical Leverage Scores

Hoaglin & Welsch 1978
Chatterjee & Hadi 1986

- Identify potential outliers in $\min_x \|Ax - b\|_2$
- 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: $A = QR$

$$H_{kk} = \|e_k^T Q\|_2^2 = \ell_k(A)$$

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

Leverage Score Bound

- $m \times n$ matrix Q with orthonormal columns
- Leverage scores $\ell_j = \|e_j^T Q\|_2^2, \quad \mu = \max_{1 \le j \le m} \ell_j$

$$L = \mathrm{diag}\begin{pmatrix} \ell_1 & \dots & \ell_m \end{pmatrix}$$

- Sampling with replacement
- $0 < \epsilon < 1$

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 $1 - \delta$: $\quad \kappa(SQ) \le \sqrt{\frac{1+\epsilon}{1-\epsilon}}$

Leverage Scores vs. Coherence

- Failure probability

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

- Bounds in terms of coherence:

$$\mu^2 \leq \|Q^T LQ\|_2 \leq \mu$$

- Estimation in terms of largest leverage scores
  If $k = 1/\mu$ is an integer then

$$\|Q^T LQ\|_2 \leq \mu \sum_{j=1}^{k} \ell_{[j]}$$

where $\ell_{[1]} \geq \cdots \geq \ell_{[m]}$

Summary

- Motivation: Randomized preconditioner for least squares
- Preconditioned matrix $\sim$ sampled orthonormal matrix
- Three different sampling strategies:
  Essentially the same for small amounts of sampling

- Bounds for condition number of sampled orthonormal matrices
  Explicit and non-asymptotic
  Predictive even for small matrix dimensions

- Coherence: Largest row norm^2
- Algorithms to generate matrices with user-specified coherence
- Leverage scores: row norms^2
- Tighter bounds: Replace coherence by leverage scores