Numerical Issues in Randomized Algorithms: Effect of Sampling on Condition Numbers

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, $Q^T Q = I$
- Real $c \times m$ "sampling" matrix S with $c \ll m$
- Desired error $0 < \epsilon < 1$

Want: Probability that

$$\|(SQ)^T(SQ) - I\|_2 \le \epsilon$$
$$\kappa(SQ) = \|SQ\|_2 \, \|(SQ)^\dagger\|_2 \le \sqrt{\tfrac{1+\epsilon}{1-\epsilon}}$$

Motivation: Randomized preconditioned LS solver *Blendenpik*
[Avron, Maymounkov & Toledo 2010]

$\kappa(SQ) =$ Condition number of preconditioned matrix

Outline

1. Exactly(c) sampling
2. Sampling rows from matrices with orthonormal columns
3. Important property: Coherence
4. Probabilistic condition number bound for sampled matrices
5. Improving on coherence: Leverage scores
6. Summary

Exactly(c) Sampling [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 $S = \sqrt{\frac{m}{c}} \begin{pmatrix} e_{k_1}^T \\ \vdots \\ e_{k_c}^T \end{pmatrix}$

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

Sampling from Matrices with Orthonormal Columns

Example: $m = 6,\ n = 2,\ c = 3$

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

Prob$[SQ$ has full rank$] \approx 11\%$

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

Prob$[SQ$ has full rank$] = 50\%$

Coherence = Largest Row Norm Squared

Q is $m \times n$ with orthonormal columns: $\mu = \max_{1 \leq k \leq m} \|e_k^T Q\|_2^2$

$$Q = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{pmatrix} \qquad \text{high coherence: } \mu = 1$$

$$Q = \begin{pmatrix} 1/\sqrt{6} & -1/\sqrt{6} \\ 1/\sqrt{6} & 1/\sqrt{6} \\ 1/\sqrt{6} & -1/\sqrt{6} \\ 1/\sqrt{6} & 1/\sqrt{6} \\ 1/\sqrt{6} & -1/\sqrt{6} \\ 1/\sqrt{6} & 1/\sqrt{6} \end{pmatrix} \qquad \text{low coherence } \mu = \frac{1}{3}$$

Properties of Coherence

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

$$\mu = \max_{1 \le k \le m} \|e_k^T Q\|_2^2$$

- $n/m \le \mu(Q) \le 1$
- 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 canonical 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

Different Definitions

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

$$\mu_0(\mathcal{Q}) = \frac{m}{n} \max_{1 \le k \le m} \|e_k^T P\|_2^2 \qquad \left(1 \le \mu_0(\mathcal{Q}) \le \tfrac{m}{n}\right)$$

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

$$\mu(A) = \max_{1 \le k \le m} \|e_k^T Q\|_2^2 \qquad \left(\tfrac{n}{m} \le \mu(A) \le 1\right)$$

- Reflects difficulty of recovering the matrix from sampling

Sampling from Matrices with Orthonormal Columns

- 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: $\mathbf{E}\left[Q^T S^T S Q\right] = Q^T Q = I$
- Sum of c random matrices:

$$Q^T S^T S Q = \frac{m}{c}\, Q^T e_{k_1} e_{k_1}^T Q + \cdots + \frac{m}{c}\, Q^T e_{k_c} e_{k_c}^T Q$$

Matrix Bernstein Inequality [Recht 2011]

- X_t independent random $n \times n$ matrices
- Expected value: $\mathbf{E}[X_t] = 0$
- Uniform boundedness: $\|X_t\|_2 \leq \tau$ almost surely
- Variance: $\rho_t \equiv \max\{\|\mathbf{E}[X_t X_t^T]\|_2, \|\mathbf{E}[X_t^T X_t]\|_2\}$
- 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 $1 - \delta$

$$\left\|\sum_t X_t\right\|_2 \leq \epsilon$$

Assumptions for Our Problem

- $m \times n$ matrix Q with orthonormal columns
- Coherence $\mu = \max_{1 \le k \le m} \|e_k^T Q\|_2^2$
- Sum of c matrices

$$(SQ)^T(SQ) - I = \sum_{t=1}^{c} X_t \qquad X_t = \frac{m}{c} Q^T e_{k_t} e_{k_t}^T Q - \frac{1}{c} I$$

- Expected value: $\mathbf{E}[X_t] = 0$
- Uniform boundedness: $\|X_t\|_2 \le m\,\mu/c$
- Variance: $\mathbf{E}[X_t^2] \le m\,\mu/c^2$

Condition Number Bound

- Desired error $0 < \epsilon < 1$
- Failure probability

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

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

This implies

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

Implications of Bound

- Coherence must be sufficiently low

$$\mu < \tfrac{3}{8} \frac{c}{m \ln(2n/\delta)}$$

{Follows from $\epsilon < 1$}

- Amount of sampling must be sufficiently large

$$c \geq \tfrac{8}{3} \frac{m\,\mu}{\epsilon^2} \ln(2n/\,\delta)$$

Minimal coherence $\mu = n/m$:

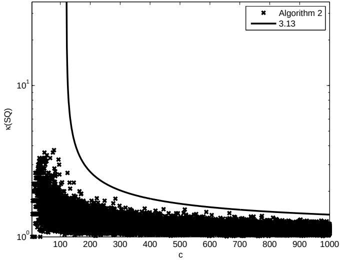
$$c \gtrsim (n \ln n)/\epsilon^2$$

Tightness of Condition Number Bound

Q is $m \times n$ with orthonormal columns, $m = 10^4$, $n = 5$

Coherence $\mu = 1.5n/m$, success probability $1 - \delta = .99$

Little sampling: $n \leq c \leq 1000$



Bound holds for $c \geq 144 \approx \frac{8}{3}\,\frac{m\mu}{\epsilon^2}\ln(2n/\delta)$

Predictive for $c \geq 200$

Coherence is not Enough

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

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

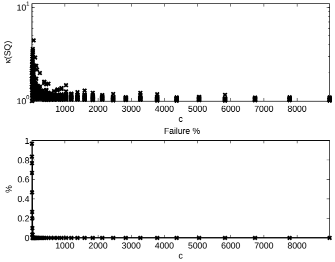
$$\textbf{Prob}[SG_{ood} \text{ has full rank}] \geq 73\%$$
$$\textbf{Prob}[SB_{ad} \text{ has full rank}] < 35\%$$

- Sampled B_{ad} matrices more likely to be rank deficient

Good Matrices

Q is $m \times n$ with orthonormal columns, $m = 10^4$, $n = 5$

Coherence $\mu = .05 = 100n/m$



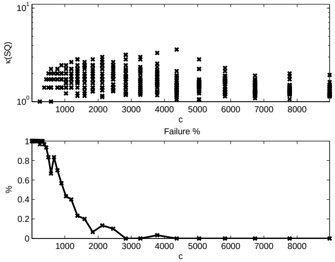
If SQ has full rank, then $\kappa(SQ) \ll 10$

Low percentage of rank deficient SQ for $c \geq n$

Bad Matrices

Q is $m \times n$ with orthonormal columns, $m = 10^4$, $n = 5$

Coherence $\mu = .05 = 100n/m$



High percentage of rank deficient SQ for $c \leq 2000 = m/n$

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 squared

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

- 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 $\min_x \|Ax - b\|_2$
- Orthogonal projector onto $\mathcal{R}(A)$: $H = A(A^T A)^{-1} A^T$
- Leverage score H_{kk}: Influence of kth data point on LS fit

Statistical Leverage Scores

Hoaglin & Welsch 1978
Chatterjee & Hadi 1986

- Identify potential outliers in $\min_x \|Ax - b\|_2$
- Orthogonal projector onto $\mathcal{R}(A)$: $\quad H = A(A^T A)^{-1} A^T$
- Leverage score H_{kk}: 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:
Drineas, Mahoney & al. 2006–2012

Assumptions for Our Problem

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

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

- Sum of c matrices

$$(SQ)^T(SQ) - I = \sum_{t=1}^{c} X_t \qquad X_t = \frac{m}{c} Q^T e_{k_t} e_{k_t}^T Q - \frac{1}{c} I$$

- Expected value: $\mathbf{E}[X_t] = 0$
- Uniform boundedness: $\|X_t\|_2 \le m\,\mu/c$
- Variance: $\mathbf{E}[X_t^2] \le m\,\|Q^T L Q\|_2/c^2$

Condition Number Bound with Leverage Scores

- Desired error $0 < \epsilon < 1$
- 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 $1-\delta$: $\quad \|(SQ)^T(SQ) - I\|_2 \leq \epsilon$

This implies

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

Leverage Scores vs. Coherence

Failure probability

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

- Bounds in terms of coherence:

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

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

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

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

Summary

- Randomized sampling of rows from matrices with orthonormal columns
- Sampling strategy: Exactly(c)
  {Bernoulli sampling is very similar}
- Coherence: *Largest* row norm squared
- Bounds for condition number of sampled matrices

  Explicit and non-asymptotic

  Realistic even for small matrix dimensions
- Leverage scores: Row norms squared
- Tighter bounds: Replace coherence by leverage scores

How much tighter???