# Introduction to Randomized Matrix Algorithms

Ilse Ipsen

Students: John Holodnak, Thomas Wentworth

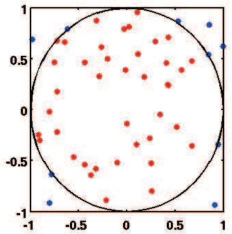
Research supported by NSF CISE CCF, NSF DMS, DARPA XData

# Randomized algorithms

Solve a deterministic problem by statistical sampling

- Monte Carlo methods

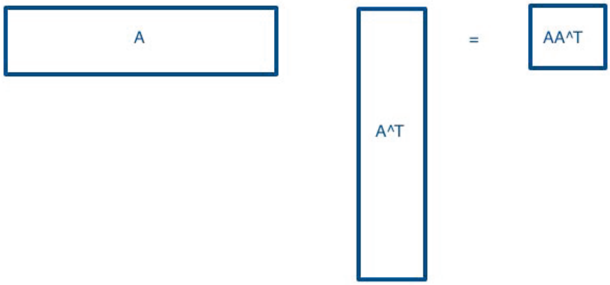
  Von Neumann & Ulam, Los Alamos, 1946




- Simulated annealing: global optimization

# This talk

Given: Real matrix $A$ with more columns than rows
Want: Monte Carlo algorithm for matrix product $AA^T$




Why is this important?

- Monte Carlo algorithm produces approximation $X = BB^T$

# Overview

- Deterministic conditions for exact representation

  When is $BB^T = AA^T$ possible?

- Monte Carlo algorithm

  Samples $B$ so that $\mathbb{E}[BB^T] = AA^T$

- Probabilistic bounds

  Error $BB^T - AA^T$, and number of columns in $B$

- Matrices with orthonormal rows, and singular values

  How close is $B$ to having orthonormal rows?

- Coherence

  Quantifying the difficulty of sampling: For which $A$ can we get a good $B$?

- Leverage scores

  Improving on coherence

- Condition numbers with respect to inversion

  Departure of a basis from orthonormality

# Deterministic conditions for exact representation

# Gram product: $AA^T$

Real matrix $A = \begin{pmatrix} A_1 & \dots & A_n \end{pmatrix}$ with $n$ columns

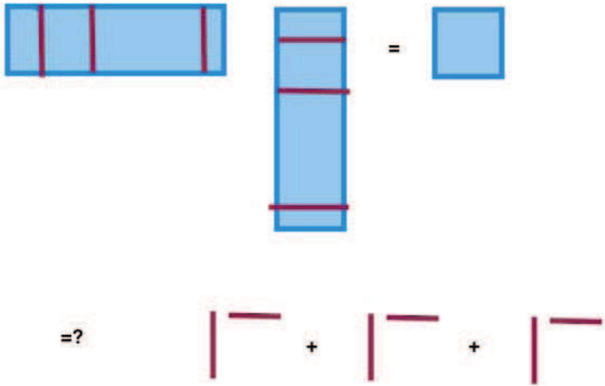
- Exact computation

$$AA^T = A_1 A_1^T + \cdots + A_n A_n^T$$

- Monte Carlo algorithm [Drineas, Kannan & Mahoney]

  Sample $c$ columns

$$X = w_1 \, A_{t_1} A_{t_1}^T + \cdots + w_c \, A_{t_c} A_{t_c}^T$$



# Gram product: $AA^T$

Real matrix $A = \begin{pmatrix} A_1 & \dots & A_n \end{pmatrix}$ with $n$ columns

- Exact computation

$$AA^T = A_1 A_1^T + \cdots + A_n A_n^T$$

- Monte Carlo algorithm [Drineas, Kannan & Mahoney]

  Sample $c$ columns

$$X = w_1 \, A_{t_1} A_{t_1}^T + \cdots + w_c \, A_{t_c} A_{t_c}^T$$

  Weights $w_j \geq 0$ chosen so that $X$ is *unbiased estimator*

$$\mathbb{E}[X] = AA^T$$

# Existing work

## Randomized matrix multiplication

*Cohen & Lewis 1997, 1999*
*Rudelson 1999, Drineas & Kannan 2001*
*Frieze, Kannan & Vempala 2004*
*Drineas, Kannan & Mahoney 2006, Sarlós 2006*
*Rudelson & Vershynin 2007*
*Belabbas & Wolfe 2008*
*Magdon-Ismail 2010, Drineas & Zouzias 2010, Magen & Zouzias 2010*
*Pagh 2011*
*Hsu, Kakade & Zhang 2012, Li, Miller & Peng 2012*
*Liberty 2013*

## Connections to

*Matrix concentration (Minsker, Tropp, ...)*
*Low-rank approximations, subset selection (Boutsidis, ...)*
*Nyström approximations (Gittens, ...)*
*Graph sparsification (Spielman, Srivastava, ...)*
*Compressed sensing (Donoho, Candés, ...)*
*Matrix completion (Recht, ...)*

# Why is this a good idea?

Want:

$$AA^T = A_1 A_1^T + \cdots + A_n A_n^T$$

Monte Carlo algorithm:

$$X = w_1\, A_{t_1} A_{t_1}^T + \cdots + w_c\, A_{t_c} A_{t_c}^T$$

- Why should $c$ columns produce a good approximation?
- How to determine the columns and weights?

Use the SVD

# Singular Value Decomposition (SVD)

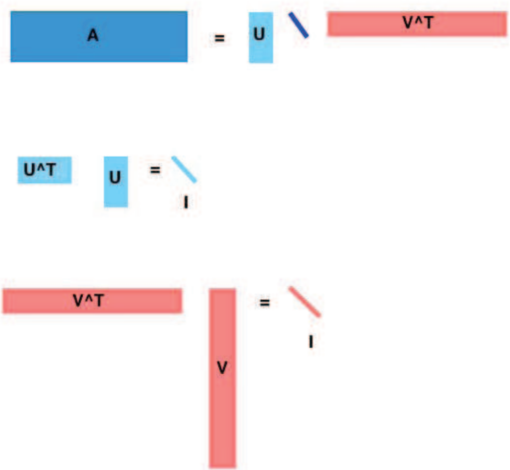
Real $m \times n$ matrix $A$ with $\text{rank}(A) = r$

$$A = U \, \Sigma \, V^T$$

- Left singular vector matrix
  $U$ is $m \times r$ with orthonormal columns: $U^T U = I_r$
- Right singular vector matrix
  $V$ is $n \times r$ with orthonormal columns: $V^T V = I_r$
- Singular values

$$\Sigma = \begin{pmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_r \end{pmatrix} \qquad \sigma_1 \geq \cdots \geq \sigma_r > 0$$

# SVD of a short & fat matrix



# Deterministic conditions for exact representation

**[Holodnak & Ii 2013]**

Given: Real matrix $A$ and $c \geq \text{rank}(A)$

There exist indices $t_1 \leq \cdots \leq t_c$ and weights $w_j \geq 0$ so that

$$w_1 \, A_{t_1} A_{t_1}^T + \cdots + w_c \, A_{t_c} A_{t_c}^T = AA^T$$

if and only if

$$\begin{pmatrix} \sqrt{w_1} \, e_{t_1} & \ldots & \sqrt{w_c} \, e_{t_c} \end{pmatrix}^T V$$

has orthonormal columns

*Exact representation depends on right singular vectors*
*Indices not necessarily distinct*
*Columns of $A$ can occur repeatedly*

# Proof of principle

## Exact representation

$$w_1 \, A_{t_1} A_{t_1}^T + \cdots + w_c \, A_{t_c} A_{t_c}^T = A A^T$$

- Necessary & sufficient conditions for existence
- Conditions depend on right singular vector matrix $V$
- There are matrices that do not satisfy these conditions
- Connections to rank-constrained matrix approximation [Friedland & Torokhti 2007]

# Monte Carlo algorithm

# Monte Carlo algorithm [Drineas et al. 2006, 2010]

**Input:** Real matrix $A$ with $n$ columns
Sampling amount $c \geq 1$
Probabilities $p_j \geq 0$ with $\sum_{j=1}^{n} p_j = 1$

**for** $j = 1$ to $c$ **do**
Sample $t_j$ from $\{1,\ldots,n\}$ with probability $p_{t_j}$
independently and with replacement
$w_j \equiv 1/(c p_{t_j})$
**end for**

**Output:** $X = w_1 \, A_{t_1} A_{t_1}^T + \cdots + w_c \, A_{t_c} A_{t_c}^T$

# How to sample

Given: Probabilities $0 \leq p_1 \leq \cdots \leq p_n$ with $\sum_{j=1}^{n} p_j = 1$

Want: Sample index $t = j$ from {1,. . . ,n} with probability $p_j$

Inversion by sequential search [Devroye 1986]

1. Determine partial sums

$$S_k \equiv \sum_{i=1}^{k} p_i \qquad 1 \leq k \leq n$$

2. Pick uniform $[0, 1]$ random variable $U$
3. Determine integer $j$ with $S_{j-1} < U \leq S_j$
4. Sampled index: $t = j$ with probability $p_j = S_j - S_{j-1}$

# Expected value (mean)

$$X = \frac{1}{c\, p_{t_1}} A_{t_1} A_{t_1}^T + \cdots + \frac{1}{c\, p_{t_c}} A_{t_c} A_{t_c}^T$$

Expected value of a single sample

$$\mathbb{E}\left[\frac{1}{c\, p_{t_j}} A_{t_j} A_{t_j}^T\right] = \sum_{k=1}^{n} p_k \frac{1}{c\, p_k} A_k A_k^T = \frac{1}{c} \sum_{k=1}^{n} A_k A_k^T = \frac{1}{c} AA^T$$

Sampling independently & with replacement:

$$\begin{aligned} \mathbb{E}[X] &= \mathbb{E}\left[\frac{1}{c\, p_{t_1}} A_{t_1} A_{t_1}^T\right] + \cdots + \mathbb{E}\left[\frac{1}{c\, p_{t_c}} A_{t_c} A_{t_c}^T\right] = c\mathbb{E}\left[\frac{1}{c\, p_{t_j}} A_{t_j} A_{t_j}^T\right] \\ &= AA^T \end{aligned}$$

Unbiased estimator: $\mathbb{E}[X] = AA^T$

# Concentration around the mean

$$X = \frac{1}{c\, p_{t_1}}\, A_{t_1} A_{t_1}^T + \cdots + \frac{1}{c\, p_{t_c}}\, A_{t_c} A_{t_c}^T$$

- Unbiased estimator: $\mathbb{E}[X] = AA^T$
- Column norm probabilities [Drineas, Kannan & Mahoney 2006]

$$p_j = \|A_j\|_2^2 / \|A\|_F^2 \qquad 1 \leq j \leq n$$

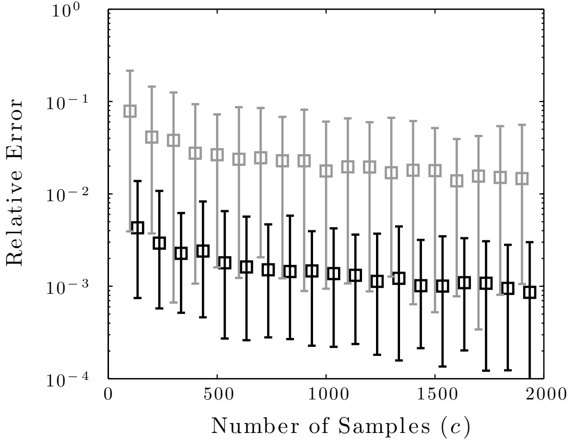
minimize $\mathbb{E}\left[\|X - AA^T\|_F^2\right]$

- We want: For any $\delta > 0$ with probability at least $1 - \delta$

$$\frac{\|X - AA^T\|_2}{\|AA^T\|_2} \leq f(\delta, c, \ldots)$$

- Idea: $X$ is sum of $c$ matrix-valued random variables

# $8 \times 4177$ **Abalone matrix** [Bache & Lichman 2013]




Monte Carlo algorithm has low relative accuracy

# Probabilistic bounds

# Matrix Bernstein concentration inequality [Tropp 2011]

- Independent random real symmetric $m \times m$ matrices $X_j$
- $\mathbb{E}[X_j] = 0$ {zero mean}
- $\|X_j\|_2 \leq \tau$ {bounded}
- $\left\|\sum_j \mathbb{E}[X_j^2]\right\|_2 \leq \rho$ {"variance"}

For any $\epsilon > 0$

$$\mathbb{P}\left[\left\|\sum_j X_j\right\|_2 \geq \epsilon\right] \leq m \exp\left(-\frac{\epsilon^2/2}{\rho + \tau\,\epsilon/3}\right)$$

{deviation from the mean}

# Relative error due to randomization [Holodnak & II]

Given: Real matrix $A = \begin{pmatrix} A_1 & \dots & A_n \end{pmatrix}$

Stable rank: $\mathrm{sr}(A) \equiv \|A\|_F^2 / \|A\|_2^2$

Monte Carlo algorithm (with probabilities $p_j = \|A_j\|_2^2 / \|A\|_F^2$)

$$X = \tfrac{1}{c\,p_{t_1}}\, A_{t_1} A_{t_1}^T + \cdots + \tfrac{1}{c\,p_{t_c}}\, A_{t_c} A_{t_c}^T$$

For any $\delta > 0$, with probability at least $1 - \delta$

$$\frac{\|X - AA^T\|_2}{\|AA^T\|_2} \leq \gamma + \sqrt{\gamma\,(6 + \gamma)}$$

where

$$\gamma \equiv \frac{\ln\left(\mathrm{rank}(A)/\delta\right)}{3\,c}\, \mathrm{sr}(A)$$

# Lower bound on number of samples [Holodnak & II]

Given: Real matrix $A = \begin{pmatrix} A_1 & \dots & A_n \end{pmatrix}$

Monte Carlo algorithm (with probabilities $p_j = \|A_j\|_2^2 / \|A\|_F^2$)

$$X = \tfrac{1}{c\, p_{t_1}}\, A_{t_1} A_{t_1}^T + \cdots + \tfrac{1}{c\, p_{t_c}}\, A_{t_c} A_{t_c}^T$$

If $0 < \epsilon < 1$, $0 < \delta < 1$ and

$$c \geq \tfrac{8}{3}\, \frac{\ln\left(\mathrm{rank}(A)/\delta\right)}{\epsilon^2}\, \mathrm{sr}(A)$$

then with probability at least $1 - \delta$

$$\frac{\|X - AA^T\|_2}{\|AA^T\|_2} \leq \epsilon$$

# Summary of probabilistic bounds

Upper bound on 2-norm relative error due to randomization
Lower bound on number of samples

Bounds

- depend on the rank and stable rank
- do not depend on matrix dimensions
- informative even for small matrix dimensions and stringent success probabilities (99 percent)

Not discussed

- Sampling with replacement, Bernoulli sampling
- Probabilities based on leverage scores
- Tightness of bounds

# Special case:
# Matrices with orthonormal rows

# From matrix multiplication to singular values

Given: Real $m \times n$ matrix $Q$ with $QQ^T = I_m$
Singular values: $\sigma_j(Q) = 1, \ 1 \leq j \leq m$

Monte Carlo algorithm: $X = \tilde{Q}\tilde{Q}^T$ where $\tilde{Q}$ has $c \geq m$ columns

$$\|\tilde{Q}\tilde{Q}^T - I\|_2 \leq \epsilon$$

Matrix multiplication bounds imply singular value bounds

- Singular values of $\tilde{Q}$

$$\sqrt{1-\epsilon} \leq \sigma_j(\tilde{Q}) \leq \sqrt{1+\epsilon} \qquad 1 \leq j \leq m$$

- Condition number of $\tilde{Q}$ with respect to inversion

$$\|\tilde{Q}\|_2 \|\tilde{Q}^\dagger\|_2 = \frac{\sigma_1(\tilde{Q})}{\sigma_m(\tilde{Q})} = \sqrt{\frac{1+\epsilon}{1-\epsilon}}$$

# Singular value bounds [Holodnak & II]

Given: Real matrix $Q = \begin{pmatrix} Q_1 & \dots & Q_n \end{pmatrix}$ with $QQ^T = I_m$

Monte Carlo algorithm (with probabilities $p_j = \|Q_j\|_2^2/m$)

$$X = \tilde{Q}\tilde{Q}^T \qquad \tilde{Q} \equiv \begin{pmatrix} \sqrt{\frac{1}{c\, p_{t_1}}}\, Q_{t_1} & \dots & \sqrt{\frac{1}{c\, p_{t_c}}}\, Q_{t_c} \end{pmatrix}$$

If $0 < \epsilon < 1$, $0 < \delta < 1$ and

$$c \geq 2(1 + \tfrac{\epsilon}{3})\, m\, \frac{\ln(m/\delta)}{\epsilon^2}$$

then with probability at least $1 - \delta$

$$\sqrt{1-\epsilon} \leq \sigma_j(\tilde{Q}) \leq \sqrt{1+\epsilon} \qquad 1 \leq j \leq m$$

# Uniform sampling [Holodnak & II]

Given: Real matrix $Q = \begin{pmatrix} Q_1 & \dots & Q_n \end{pmatrix}$ with $QQ^T = I_m$

Largest column norm $\mu \equiv \max_{1 \le j \le n} \|Q_j\|_2^2$

Monte Carlo algorithm (with probabilities $p_j = 1/n$)

$$X = \tilde{Q}\tilde{Q}^T \qquad \tilde{Q} \equiv \begin{pmatrix} \sqrt{\frac{1}{c\, p_{t_1}}}\, Q_{t_1} & \dots & \sqrt{\frac{1}{c\, p_{t_c}}}\, Q_{t_c} \end{pmatrix}$$

If $0 < \epsilon < 1$, $0 < \delta < 1$ and

$$c \ge 2(1 + \tfrac{\epsilon}{3})\, n\, \mu\, \frac{\ln(m/\delta)}{\epsilon^2}$$

then with probability at least $1 - \delta$

$$\sqrt{1-\epsilon} \le \sigma_j(\tilde{Q}) \le \sqrt{1+\epsilon} \qquad 1 \le j \le m$$

# Summary: Matrices with orthonormal rows

Probabilistic singular value bounds

$$\sqrt{1-\epsilon} \le \sigma_j(\tilde{Q}) \le \sqrt{1+\epsilon} \qquad 1 \le j \le m$$

- Column norm probabilities $p_j = \|Q_j\|_2^2/m$
  Number of samples $c = \Omega\left(m \ \ln m/\epsilon^2\right)$

- Uniform probabilities $p_j = 1/n$
  Number of samples $c = \Omega\left(n\mu \ \ln m/\epsilon^2\right) \qquad \mu \equiv \max_j \|Q_j\|_2^2$

Connections to

- Coupon collector's problem (Halko, Martinsson & Tropp)
- Compressed sensing (Donoho, Candés, ...)

# Coherence

# Properties of Coherence

Real matrix $Q = \begin{pmatrix} Q_1 & \dots & Q_n \end{pmatrix}$ with $QQ^T = I_m$

Coherence $\mu \equiv \max_{1 \le j \le n} \|Q_j\|_2^2$

- $m/n \le \mu \le 1$
- Maximal coherence: $\mu = 1$
  At least one row of $Q$ is a canonical vector
- Minimal coherence: $\mu = m/n$
  Rows of $Q$ are rows of a Hadamard matrix
- Coherence measures "correlation with standard basis"
- Quantifies difficulty of recovering matrix from sampling

# 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
  Meng, Saunders & Mahoney 2011

  *Randomized preconditioners for least squares*

- Drineas, Magdon-Ismail, Mahoney & Woodruff 2011

  *Fast approximation of coherence*

# Leverage scores

# Leverage scores

$Q = \begin{pmatrix} Q_1 & \dots & Q_n \end{pmatrix}$ with $QQ^T = I_m$

Idea: Use all column norms

- Leverage scores = squared column norms of $Q$

$$\ell_j = \|Q_j\|_2^2 \qquad\qquad 1 \leq j \leq n$$

- Coherence = largest leverage score

$$\mu = \max_{1 \leq j \leq n} \ell_j$$

- Low coherence $\Longleftrightarrow$ uniform leverage scores

Leverage scores: Importance sampling in randomized algorithms

[Drineas & Mahoney 2006, ...]

# Leverage scores are ubiquitous

- Statistics

  [Hoaglin & Welsch 1978,Velleman & Welsch 1981, Chatterjee & Hadi 1986]

  Leverage scores: Outliers in regression problems

- Astronomy

  [Yip, Mahoney, Szalay, Csabai, Budavári, Wyse & Dobos 2013]

  Leverage scores: Important wave lengths in galaxy evolution

- Electronic structure calculations

  [Bekas, Kokiopoulou & Saad 2008]

  Leverage scores: Charge densities

- Graph Theory

  [Drineas & Mahoney 2010]

  Leverage scores: Effective resistance of edges

# Condition Number Bound [Il & Wentworth]

- $m \times n$ matrix $Q$ with orthonormal rows
- Leverage scores $\ell_j = \|Q_j\|_2^2$

$$L = \operatorname{diag}\begin{pmatrix}\ell_1 & \dots & \ell_n\end{pmatrix}$$

- Coherence $\mu = \|L\|_2 = \max_{1 \le j \le n} \ell_j$
- Uniform sampling, number of sampled columns $c \geq 1$
- Error tolerance $0 < \epsilon < 1$

Failure probability

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

With probability at least $1-\delta$: $\|\tilde{Q}\|_2 \|\tilde{Q}^\dagger\|_2 \leq \sqrt{\frac{1+\epsilon}{1-\epsilon}}$

# What to do about $\|QLQ^T\|_2$

Failure probability

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

where

$$\mu^2 \leq \|QLQ^T\|_2 \leq \mu$$

- Want: Simple accurate approximation of $\|QLQ^T\|_2$
- How: Derive bound for general scaled matrices
- Connections to

  *Majorization, lattice superadditive maps*

  *Inverse eigenvalue problems* [Dhillon et al. 2005]

## General scaled matrices [Wentworth & II]

- $m \times n$ matrix $Z$ with $\text{rank}(Z) = m$
- Largest squared column norm $\mu_z \equiv \max_{1 \le j \le n} \|Z_j\|_2^2$
- Diagonal matrix $D = \text{diag}\begin{pmatrix} d_1 & \cdots & d_n \end{pmatrix}$

$$d_{[1]} \ge \cdots \ge d_{[n]}$$

*Bound $\|Z\,D\|_2$ in terms of $\mu_z$ and largest elements of $D$*

If $t = \left\lfloor 1/(\|Z^\dagger\|_2^2\,\mu_z) \right\rfloor$ then

$$\|Z\,D\|_2^2 \le \mu_z \sum_{j=1}^{t} d_{[j]}^2 + \left(\|Z\|_2^2 - t\,\mu_z\right)\; d_{[k]}^2$$

where $k = 1$ or $t + 1$

# Bound for $\|QLQ^T\|_2$

- $m \times n$ matrix $Q$ with $QQ^T = I_m$
- Coherence $\mu \equiv \max_{1\leq j\leq n} \|Q_j\|_2^2$
- Leverage scores $\ell_{[1]} \geq \cdots \geq \ell_{[n]}$

If $t = \lfloor 1/\mu \rfloor$ then

$$\|Q\, L\, Q^T\|_2 = \|Q\, L^{1/2}\|_2^2 \leq \mu \sum_{j=1}^{t} \ell_{[j]} + (1 - t\,\mu)\ \ell_{[t+1]}$$

If $t = 1/\mu$ is an integer then

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

Bound for $\|QLQ^T\|_2$ tighter than coherence $\mu$

# Simpler probabilistic bound [Wentworth & II]

- $m \times n$ matrix $Q$ with $QQ^T = I_m$
- Leverage scores $\mu \equiv \ell_{[1]} \geq \cdots \geq \ell_{[n]}$
- Uniform sampling of columns
- Approximation to $\|QLQ^T\|_2$

$$\tau \equiv \mu \sum_{j=1}^{t} \ell_{[j]} + (1 - t\,\mu)\,\ell_{[t+1]} \qquad t = \lfloor 1/\mu \rfloor$$

If

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

then with probability at least $1 - \delta$

$$\|\tilde{Q}\|_2 \|\tilde{Q}^\dagger\|_2 \leq \sqrt{\frac{1+\epsilon}{1-\epsilon}}$$

# Summary

Monte Carlo algorithm for Gram product $AA^T$

- Deterministic conditions for exact representation

  Depend on right singular vector matrix

- Probabilistic bounds for 2-norm relative error, number of sampled columns

  Depend on rank and stable rank of $A$, but not dimension

- Probabilistic singular value bounds

  Matrices with orthonormal rows

  Uniform sampling: Bounds depend on coherence

- Probabilistic condition number bounds

  Matrices with orthonormal rows

  Uniform sampling: Tighter bounds in terms of leverage scores

- Bound for 2-norm of scaled matrices

  In terms of largest column norm, and elements of diagonal matrix

# Why randomized algorithms?

- Reduction of massive data sets, for low-accuracy requirements

  Least squares/regression, SVD/PCA, subspace approximation, model reduction

- Advantages

  "Easy" to analyze, forgiving, probabilistic bounds more optimistic

- Applications

  Machine learning, population genomics, astronomy, nuclear engineering

- Survey papers

  *Halko, Martinsson & Tropp 2011*
  *Mahoney 2011*