Randomized Algorithms for Least Squares Problems

Ilse C.F. Ipsen

Joint work with Jocelyn T. Chi and Thomas Wentworth

North Carolina State University
Raleigh, NC, USA

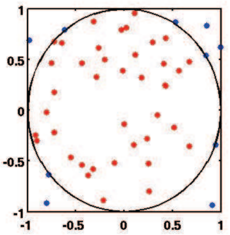
Research supported by DARPA XData, NSF CCF and DMS

Randomized Algorithms

Solve a deterministic problem by statistical sampling

- Monte Carlo Methods

  Von Neumann & Ulam, Los Alamos, 1946



circle area $\approx 4 \, \frac{\#\text{hits}}{\#\text{darts}}$

- Simulated Annealing: global optimization

This Talk: The Ideas behind Randomized Least Squares Solvers

- Deterministic Least Squares Solvers
- Kaczmarz: An Iterative Coordinate Descent Method
- Effect of Sampling on Statistical Model Uncertainty
- How to Do Randomized Sampling
- An Overview of Randomized Least Squares/Regression
- Randomized Row-wise Compression for Dense Matrices
- A Randomized Right Preconditioner for Sparse Matrices
- Probabilistic Bound for Deviation from Orthonormality
- A few Take Aways, and Bibliography

Deterministic Least Squares Solvers

Statistics: Linear Regression

Gaussian linear model

$$b = Ax_0 + \epsilon, \qquad \epsilon \sim \mathcal{N}(0, \sigma^2 I_m)$$

Given: Design matrix $A \in \mathbb{R}^{m \times n}$
Observation vector $b \in \mathbb{R}^m$
Unknown: Parameter vector $x_0 \in \mathbb{R}^n$
Noise vector: ϵ has multivariate normal distribution

Minimize Residual Sum of Squares

$$\mathrm{RSS}(x) = (b - Ax)^T (b - Ax) \qquad \{\text{superscript T is transpose}\}$$

Minimizer x_* is maximum likelihood estimator of x_0

Computational Mathematics: Least Squares

This talk: Well-posed least squares problems

Given: $A \in \mathbb{R}^{m \times n}$ with $\mathrm{rank}(A) = n \leq m$, $\; b \in \mathbb{R}^m$

{tall and skinny A with linearly independent columns}

Solve: $\min_x \|Ax - b\|_2$ {two norm}

Unique solution (in exact arithmetic): $x_* = A^\dagger b$

Moore-Penrose inverse: $A^\dagger \equiv (A^T A)^{-1} A^T$

Hat matrix: $AA^\dagger = A(A^T A)^{-1} A^T$

orthogonal projector onto $\mathrm{range}(A)$

Least squares residual: $b - Ax_* = (I - AA^\dagger)b$

orthogonal projection of b onto $\mathrm{range}(A)^\perp$

Least Squares Solvers for Dense Matrices

Idea: Basis transformation $A = QR$

- Q has orthonormal columns: $Q^T Q = I_n$
  $\{$Orthonormal basis for $\mathrm{range}(A)\}$
- R is triangular nonsingular
  $\{$Easy-to-compute relation between old and new bases$\}$
- Left inverse simplifies: $A^\dagger = (A^T A)^{-1} A^T = R^{-1} Q^T$

Direct method:

1. Thin QR factorization $A = QR$
2. Triangular system solve $R\, x_* = Q^T b$

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

Least Squares Solvers for Sparse Matrices

LSQR [Paige & Saunders 1982]

Krylov space method for solving system with $\begin{pmatrix} I & A \\ A^T & 0 \end{pmatrix}$

Matrix vector products with A and A^T

Conceptually:
Solution of $A^T A x = A^T b$ with approximations at iteration k

$$x_k \in \operatorname{span}\left\{A^T b,\ (A^T A)\, A^T b,\ \ldots,\ (A^T A)^k\, A^T b\right\}$$

Residuals decrease {in exact arithmetic}

$$\|b - A x_k\|_2 \ \leq\ \|b - A x_{k-1}\|_2$$

Fast convergence if condition number $\kappa(A) \equiv \|A\|_2 \|A^\dagger\|_2$ small

$$\|A(x_* - x_k)\|_2^2 \ \leq\ 2\left(\frac{\kappa(A) - 1}{\kappa(A) + 1}\right)^k \|A(x_* - x_0)\|_2^2$$

Summary: Deterministic Least Squares Solvers

Given: $A \in \mathbb{R}^{m\times n}$ with $\mathrm{rank}(A) = n$
Want: Unique solution x_* of $\min_x \|Ax - b\|_2$

- Dense matrix A

  $A = QR$ requires $\mathcal{O}(mn^2)$ flops
  Too expensive when A is large or sparse
  QR produces fill-in

- Sparse matrix A

  Matrix vector products with A and A^T
  Convergence of LSQR depends on $\kappa(A)$
  Need convergence acceleration (preconditioner)
  with low cost per iteration

Kaczmarz:
An Iterative Coordinate Descent Method

Idea Behind Kaczmarz Methods

Each iteration projects on a particular equation

$$A = \begin{pmatrix} a_1^T \\ \vdots \\ a_m^T \end{pmatrix} \in \mathbb{R}^{m \times n} \qquad b = \begin{pmatrix} b_1 \\ \vdots \\ b_m \end{pmatrix} \in \mathbb{R}^m$$

Given iterate $x^{(k-1)}$, compute next iterate $x^{(k)} = x^{(k-1)} + z$
so that $x^{(k)}$ solves equation i

$$z = e_i^T \left(b - Ax^{(k-1)}\right) \frac{a_i}{a_i^T a_i} = \frac{b_i - a_i^T x^{(k-1)}}{\|a_i\|_2^2}\, a_i$$

Then $a_i^T x^{(k)} = b_i$

Kaczmarz Methods for Linear Systems

Input: $A \in \mathbb{R}^{m\times n}$ with $\mathrm{rank}(A) = n$, $b \in \mathbb{R}^m$, $x^{(0)} \in \mathbb{R}^n$
Output: Approximate solution to $Ax_* = b$

for $k = 1, 2, \ldots$ do
  Choose equation i
  $x^{(k)} = x^{(k-1)} + \frac{b_i - a_i^T x^{(k-1)}}{\|a_i\|_2^2}\, a_i$
end for

How to choose equation i?

- Deterministic [Kaczmarz 1937]
  Cycle through the equations: $i = k \bmod m + 1$
- Randomized: Uniform Sampling [Natterer 1986]
  Sample i from $\{1, \ldots, m\}$ with probability $1/m$, independently and with replacement

Randomized Kaczmarz with Non-Uniform Sampling

Let $A \in \mathbb{R}^{m \times n}$ with $\mathrm{rank}(A) = n$
Scaled condition number: $\kappa_{F,2}(A) = \|A\|_F \|A^\dagger\|_2$

Sample rows with large norms

Sample i from $\{1, \ldots, m\}$ with probability $\|a_i\|_2^2 / \|A\|_F^2$ independently and with replacement

Convergence in expectation

- Linear systems $Ax_* = b$ [Strohmer, Vershynin 2009]

$$\mathbb{E}\left[\|x^{(k)} - x_*\|_2^2\right] \leq \left(1 - \frac{1}{(\kappa_{F,2}(A))^2}\right)^k \|x^{(0)} - x_*\|_2^2$$

- Least squares $\min_x \|Ax - b\|_2$ [Needell 2010]

$$\mathbb{E}\left[\|x^{(k)} - x_*\|_2^2\right] \leq \left(1 - \frac{1}{(\kappa_{F,2}(A))^2}\right)^k \|x^{(0)} - x_*\|_2^2 + (\kappa_{F,2}(A))^2 \|b - Ax_*\|_\infty^2$$

Connections, and Related Work: A Very Small Selection

- Sampling rows according to row norms: Diagonal scaling for optimal condition numbers [Van der Sluis 1969]
- Kaczmarz with relaxation factors for least squares [Hanke, Niethammer 1990, 1995]
- Greedy Kaczmarz-Motzkin algorithms [Haddock, Ma 2021]
- Randomized Gauss-Seidel for least squares [Niu, Zheng, 2021]
- Direct projection methods for linear systems [Benzi, Meyer 1995]
- Kaczmarz for detection of corrupted matrix elements [Haddock, Needell 2019]
- Application to medical imaging, computer tomography [Natterer 2001]

Effect of Sampling on Statistical Model Uncertainty

Example: Effect of Sampling on Model Uncertainty

Gaussian linear model

$$b = Ax_0 + \epsilon \qquad A = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 0 \end{pmatrix} \qquad \epsilon \sim \mathcal{N}(0, \sigma^2 I_4)$$

Least squares problem $\min_x \|Ax - b\|_2$ has solution

$$x_* = A^\dagger b \qquad A^\dagger = (A^T A)^{-1} A^T = \begin{pmatrix} \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

Solution is unbiased estimator

$$\mathbb{E}_\epsilon[x_*] = A^\dagger \, \mathbb{E}_\epsilon[b] = A^\dagger A x_0 = x_0$$

with nonsingular variance $\mathbb{V}\mathrm{ar}_\epsilon[x_*] = \sigma^2 (A^T A)^{-1} = \sigma^2 \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & 1 \end{pmatrix}$

Example: Sampling Preserves Rank

Fixed sampling matrix S with $\operatorname{rank}(SA) = \operatorname{rank}(A)$
$\min_x \|S(Ax - b)\|_2$ has unique solution $\tilde{x} = (SA)^\dagger Sb$

- Sampled matrix has full column-rank

$$SA = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = (SA)^\dagger$$

- Unbiased estimator $\mathbb{E}_\epsilon[\tilde{x}] = (SA)^\dagger S\, \mathbb{E}_\epsilon[b] = x_0$
- Increase in variance

$$\mathbb{V}\mathrm{ar}_\epsilon[\tilde{x}] = \sigma^2 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \succcurlyeq \sigma^2 \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & 1 \end{pmatrix} = \mathbb{V}\mathrm{ar}_\epsilon[x_*]$$

Example: Sampling Fails to Preserve Rank

Fixed sampling matrix S with $\mathrm{rank}(SA) < \mathrm{rank}(A)$
$\min_x \|S(Ax-b)\|_2$ has minimal-norm solution $\tilde{x} = (SA)^\dagger Sb$

- Sampled matrix is rank-deficient

$$SA = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} = (SA)^\dagger$$

- Biased estimator $\mathbb{E}_\epsilon[\tilde{x}] = (SA)^\dagger (SA) x_0 = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} x_0 \neq x_0$
- Singular variance

$$\mathbb{V}\mathrm{ar}_\epsilon[\tilde{x}] = \sigma^2 \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \neq \sigma^2 \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & 1 \end{pmatrix} = \mathbb{V}\mathrm{ar}_\epsilon[x_*]$$

Summary: Effect of Sampling on Model Uncertainty

$\min_x \|S(Ax-b)\|_2$ has minimal-norm solution $\tilde{x} = (SA)^\dagger(Sb)$
with expectation $\mathbb{E}_\epsilon[\tilde{x}] = (SA)^\dagger(SA)x_0$

- If S preserves rank: $\mathrm{rank}(SA) = \mathrm{rank}(A)$
  $(SA)^\dagger$ is left inverse: $(SA)^\dagger(SA) = I$
  $\tilde{x}$ is unbiased estimator: $\mathbb{E}_\epsilon[\tilde{x}] = x_0$
- If S loses rank: $\mathrm{rank}(SA) < \mathrm{rank}(A)$
  No left inverse: $(SA)^\dagger(SA) \neq I$
  $\tilde{x}$ is biased estimator: $\mathbb{E}_\epsilon[\tilde{x}] \neq x_0$
  Variance $\mathbb{V}\mathrm{ar}_\epsilon[\tilde{x}]$ is singular

This was a best case analysis: A fixed sampling matrix S.
We did not incorporate the uncertainty due to randomization

How to do Randomized Sampling

How to Sample [Devroye 1986]

Sample t from $\{1, \ldots, m\}$ with probability p_t

- Uniform sampling: $p_i = 1/m, \ \ 1 \le i \le m$

$$
\begin{aligned}
&\upsilon = \texttt{rand} \qquad \{\text{uniform } [0,1] \text{ random variable}\} \\
&t = \lfloor 1 + m\,\upsilon \rfloor
\end{aligned}
$$

- Non-uniform sampling:

$$
\begin{aligned}
&\upsilon = \texttt{rand},\ t = 1, \ \ F = p_1 \\
&\text{while } \upsilon > F \\
&t = t + 1, \ \ F = F + p_t
\end{aligned}
$$

Inversion by sequential search: $F(i) \equiv \sum_{j=1}^{i} p_j$ so that $p_i = F(i) - F(i-1)$

t defined by $F(t-1) < \upsilon \le F(t)$

Matlab: `randi`, `datasample`

R: `sample`

Different Sampling Methods

Want: Sampling matrix S with $\mathbb{E}[S^T S] = I_m$

1. Uniform sampling with replacement
   Sample k_t from $\{1, \ldots, m\}$ with probability $\frac{1}{m}$, $1 \leq t \leq c$
   $S = \sqrt{\frac{m}{c}} \begin{pmatrix} e_{k_1} & \ldots & e_{k_c} \end{pmatrix}^T$

2. Uniform sampling without replacement
   Let $k_1, \ldots, k_m$ be a permutation of $1, \ldots, m$
   $S = \sqrt{\frac{m}{c}} \begin{pmatrix} e_{k_1} & \ldots & e_{k_c} \end{pmatrix}^T$

3. Bernoulli sampling

$$S(t,:) = \sqrt{\frac{m}{c}} \begin{cases} e_t^T & \text{with probability } \frac{c}{m} \\ 0_{1\times m} & \text{with probability } 1 - \frac{c}{m} \end{cases} \qquad 1 \leq t \leq m$$

   Alternative simulation:
   Sample $\tilde{c}$ from $\{1, \ldots, m\}$ with $\mathbb{P}[\tilde{c} = k] = \binom{m}{k}(\frac{c}{m})^k(1 - \frac{c}{m})^{m-k}$
   Sample $k_1, \ldots, k_{\tilde{c}}$ without replacement

Comparison of Different Sampling Methods

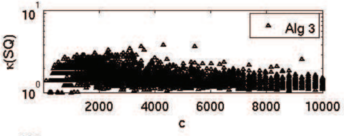
Sampling rows from matrices with orthonormal columns
$10^4 \times 5$ matrices Q with $Q^T Q = I$

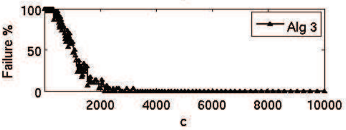
Plots for $5 \leq c \leq 10^4$

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

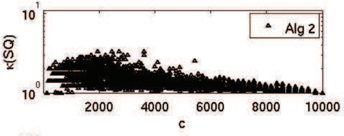
Comparison of Sampling Methods

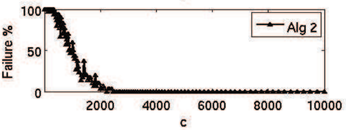
Sampling with replacement



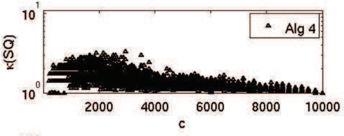


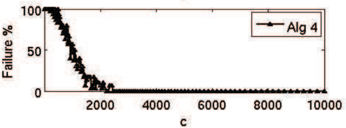
Sampling without replacement





Bernoulli sampling





Summary: Comparison of Different Sampling Methods

Three different sampling methods:

- Uniform sampling with replacement
- Uniform sampling without replacement
- Bernoulli sampling

Conclusion:
Little difference among sampling methods for small amounts of sampling

From now on:
Use sampling with replacement

An Overview of
Randomized Least Squares/Regression

Randomized Least Squares/Regression

(Solvers mostly not ready for production yet)

$\min_{x\in\mathbb{R}^n} \|Ax-b\|_2$ r for $A \in \mathbb{R}^{m\times n}$ with $m \geq n$
Direct methods require $\mathcal{O}(mn^2)$ flops

Classification [Thanei, Heinze, Meinshausen 2017]

- Row-wise compression: $\min_{x\in\mathbb{R}^n} \|S(Ax-b)\|_2$
  $S \in \mathbb{R}^{c\times m}$ with $c \leq m$
  Solver requires $\mathcal{O}(cn^2)$ flops after compression

- Column-wise compression: $\min_{y\in\mathbb{R}^c} \|ASy-b\|_2$
  $S \in \mathbb{R}^{n\times c}$ with $c \leq n$
  Solver requires $\mathcal{O}(m\,c^2)$ flops after compression

  Special case: $S \in \mathbb{R}^{n\times n}$ nonsingular
  Right preconditioning to accelerate iterative methods

Existing Work

Row-wise compression

Bartels, Hennig (2016); Becker, Jawas, Patrick, Ramamurthy (2017)
Boutsidis, Drineas (2009); Dhillon, Lu, Foster, Ungar (2013)
Drineas, Mahoney, Muthukrishnan (2006)
Drineas, Mahoney, Muthukrishnan, Sarlós (2011)
Ipsen, Wentworth (2014)
McWilliams, Krummenacher, Lučić, Buhmann (2014)
Meng, Saunders, Mahoney (2014); Wang, Zhu, Ma (2018)
Zhou, Lafferty, Wasserman (2007)

Column-wise compression

Kabán (2014); Mallard, Munos (2009)
Meng, Saunders, Mahoney (2014)
Thanei, Heinze, Meinshausen (2017)

Right preconditioning

Avron, Maymounkov, Toledo (2010)
Ipsen, Wentworth (2014); Rokhlin, Tygert (2008)

Statistical properties

Ahfock, Astle, Richardson (2017); Chi, Ipsen (2020)
Lopes, Wang, Mahoney (2018); Ma, Mahoney, Yu (2014, 2015)
Raskutti, Mahoney (2016; Thanei, Heinze, Meinshausen (2017)

Randomized Row-Wise Compression for Dense Matrices

Uniform Sampling with Replacement

[Drineas, Kannan & Mahoney 2006]

$S \in \mathbb{R}^{c \times m}$ samples c rows from identity $I_m = \begin{pmatrix} e_1^T \\ \vdots \\ e_m^T \end{pmatrix}$

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

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

- Expected value $\mathbb{E}\left[S^T S\right] = I_m$
- S can sample a row more than once

Example: Uniform Sampling with Replacement

Sample 2 out of 4 rows: $m = 4$, $c = 2$, $\sqrt{\frac{m}{c}} = \sqrt{2}$

$$A = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 0 \end{pmatrix}, \qquad S^{(ij)} = \sqrt{2} \begin{pmatrix} e_i^T \\ e_j^T \end{pmatrix}, \quad 1 \le i, j \le 4$$

Examples of sampled matrices

$$S^{(11)} A = \sqrt{2} \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 0 \end{pmatrix} = \sqrt{2} \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$$

$$S^{(42)} A = \sqrt{2} \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 0 \end{pmatrix} = \sqrt{2} \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$$

Sampling matrices are unbiased estimators of identity

$$\mathbb{E}[S^T S] = \sum_{i=1}^{4} \sum_{j=1}^{4} \tfrac{1}{16} \left(S^{(ij)}\right)^T S^{(ij)} = I_4$$

Row Sampling Algorithm for $\min_{x\in\mathbb{R}^n} \|Ax-b\|_2$

Special case of [Drineas, Mahoney, Muthukrishnan, Sarlós, 2011]

Input: $A \in \mathbb{R}^{m\times n}$ with $\mathrm{rank}(A) = n$, $b \in \mathbb{R}^m$
$c \geq 1$ $\quad$ {sampling amount}

$S = 0_{c\times m}$ $\quad$ {initialize sampling matrix}
for $t = 1 : c$ do
  Sample k_t from $\{1, \ldots, m\}$ with probability $1/m$
  independently and with replacement
  $S(t,:) = \sqrt{\frac{m}{c}}\, e_{k_t}^T$ $\quad$ {row t of sampling matrix}
end for

Output: Minimal norm solution $\tilde{x}$ of $\min_x \|S(Ax-b)\|_2$

Error due to Randomization

Derivation in two steps

1. Structural bound:
   Treat sampling matrix SA as fixed perturbation
   Carry deterministic analysis as far as possible

2. Probabilistic bound:
   Treat sampled matrix SA as random matrix
   Use matrix concentration inequalities

Structural Bound: Absolute Error

- Exact solution $x_* = A^\dagger b$
- Randomized solution $\tilde{x} = (SA)^\dagger Sb$
  Assume: $\mathrm{rank}(SA) = \mathrm{rank}(A)$
- Change of basis: $A = QR$
- Geometric interpretation of error

$$\tilde{x} - x_* = (SA)^\dagger Sb - A^\dagger b = A^\dagger\, Q(SQ)^\dagger S\,(b - Ax_*)$$

  $Q(SQ)^\dagger S$ is oblique projector onto $\mathrm{range}(A)$
  $b - Ax_*$ is exact least squares residual

- If $\|S(b - Ax_*)\|_2 \leq (1+\epsilon)\|b - Ax_*\|_2$ then

$$\|\tilde{x} - x_*\|_2 \leq (1+\epsilon)\|A^\dagger\|_2 \|(SQ)^\dagger\|_2 \|b - Ax_*\|_2$$

Structural Bound: Relative Error

[Drineas, Mahoney, Muthukrishnan, Sarlós, 2011]

If $\mathrm{rank}(SA) = n$ and $\|S(b - Ax_*)\|_2 \leq (1+\epsilon)\|b - Ax_*\|_2$ then

$$\frac{\|\tilde{x} - x_*\|_2}{\|x_*\|_2} \leq (1+\epsilon)\, \|(SQ)^\dagger\|_2\, \kappa(A)\, \underbrace{\frac{\|b - Ax_*\|_2}{\|A\|_2 \|x_*\|_2}}_{\substack{\text{normalized} \\ \text{LS residual}}}$$

$\kappa(A) = \|A\|_2 \|A^\dagger\|_2$ condition of A w.r.t. left inversion

- Relative error depends only on $\kappa(A)$ but not $[\kappa(A)]^2$
- Sensitivity to multiplicative perturbations from randomization is lower than sensitivity to deterministic additive perturbations
- Probabilistic bound for $\|(SQ)^\dagger\|_2$
  Has to take care of $\mathrm{rank}(SA) = n$, and quantify ϵ

Towards a Probabilistic Bound

Given $A \in \mathbb{R}^{m\times n}$ with $\mathrm{rank}(A) = n$

$$\frac{\|\tilde{x} - x_*\|_2}{\|x_*\|_2} \leq (1+\epsilon)\, \|(SQ)^\dagger\|_2\, \kappa(A)\, \frac{\|b - Ax_*\|_2}{\|A\|_2 \|x_*\|_2}$$

- For the analysis (but not computed): $A = QR$
  where $Q \in \mathbb{R}^{m\times n}$ with $Q^T Q = I$
- Idea: $SA = (SQ)\, R$
  Sampling rows from A amounts to sampling rows from Q
- Simplify the analysis to SQ:
  Sampling rows from matrices Q with orthonormal columns

Before doing the analysis:
Look at a randomized solver for sparse matrices, which faces the same situation

A Randomized Right Preconditioner for Sparse Matrices

Right Preconditioning LSQR

Convergence acceleration for LSQR applied to $\min_x \|Ax - b\|_2$

Right preconditioning = change of variables

$$\min_y \|A\,P^{-1}\,\underbrace{(Px)}_{y} - b\|_2$$

1. $\min_y \|A\,P^{-1}y - b\|_2$ {Solve preconditioned problem}
2. Solve $Px_* = y$ {Retrieve solution to original problem}

Requirements for preconditioner P

Fast convergence: $\kappa(A\,P^{-1}) \approx 1$

Linear systems with P are cheap to solve

The Ideal Right Preconditioner

- QR factorization $A = QR$ $\quad Q^T Q = I_n, \; R$ is $\triangle$
- Use R as preconditioner
- Preconditioned matrix $AR^{-1} = Q$

  Orthonormal columns

  Perfect condition number $\kappa(Q) = 1$
- LSQR solves pre-conditioned system in 1 iteration

But:

This is what we are trying to avoid in the first place

Construction of preconditioner is way too expensive

A Randomized Preconditioner

Idea: QR factorization from a few rows of $m \times n$ matrix A

1. Sample $c \geq n$ rows of A: SA
2. QR factorization of sampled matrix
$$SA = Q_s R_s \qquad Q_s^T Q_s = I_n, \; R_s \text{ is } \triangle$$
3. Randomized preconditioner R_s^{-1}

Operation count: $\mathcal{O}(cn^2)$ {independent of large dimension m}

QR Factorization from a Few Rows

Blendenpik [Avron, Maymounkov & Toledo 2010]

Input: $m \times n$ matrix with $\mathrm{rank}(A) = n$, $m \times 1$ vector b
Sampling amount $c \geq n$

Output: Solution x_* to $\min_x \|Ax - b\|_2$

{Construct preconditioner}

Sample c 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 $R_s x_* = y$ {$\triangle$ system}

We 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]

Two QR factorizations

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

Idea

1. Sampling rows of A $\triangleq$ Sampling rows of Q

$$\operatorname{rank}(SA) = \operatorname{rank}(SQ)$$

2. Condition number of preconditioned matrix (2-norm)

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

Simpler problem

Sample from matrices with orthonormal columns

Sampling from Matrices with Orthonormal Columns
What To Expect

Given: $Q \in \mathbb{R}^{8\times 2}$ with $Q^TQ = I$
Want: Sampled matrix SQ with $\mathrm{rank}(SQ) = 2$
Which one is easier?

$$Q = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{pmatrix} \qquad \text{versus} \qquad Q = \frac{1}{\sqrt{8}} \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & -1 \\ 1 & -1 \\ 1 & -1 \\ 1 & -1 \end{pmatrix}$$

Sampling from Matrices with Orthonormal Columns
What To Expect

Given: $Q \in \mathbb{R}^{8\times 2}$ with $Q^TQ = I$
Want: Sampled matrix SQ with $\operatorname{rank}(SQ) = 2$
Which one is easier?

$$Q = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{pmatrix} \qquad \text{versus} \qquad Q = \frac{1}{\sqrt{8}} \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & -1 \\ 1 & -1 \\ 1 & -1 \\ 1 & -1 \end{pmatrix}$$

Row norms (squared)

$\|e_1^TQ\|_2^2 = \|e_2^TQ\|_2^2 = 1$
$\|e_j^TQ\|_2^2 = 0 \quad \text{for } j \geq 3$

$\|e_j^TQ\|_2^2 = \frac{2}{8} = \frac{1}{4} \quad \text{for all } j$

Sampling from Matrices with Orthonormal Columns

$Q \in \mathbb{R}^{8\times 2}$ with $Q^TQ = I$

$$Q = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{pmatrix} \qquad Q = \frac{1}{\sqrt{8}} \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & -1 \\ 1 & -1 \\ 1 & -1 \\ 1 & -1 \end{pmatrix}$$

$\max_j \|e_j^T Q\|_2^2 = 1$ $\qquad$ $\max\limits_j \|e_j^T Q\|_2^2 = \frac{1}{4}$

Sampling is hard $\qquad$ Sampling is easy

Largest row norm distinguishes matrices with orthonormal columns

Use it to quantify difficulty of sampling

Probabilistic Bound for Deviation from Orthonormality

Deviation of SQ from Orthonormality

Given $0 \le \epsilon < 1$, want sampling amount $c \ge n$ so that

$$\|(SQ)^T(SQ) - I\|_2 \le \epsilon$$

This implies for the singular values of $SQ \in \mathbb{R}^{c \times n}$

$$1 - \epsilon \le \sigma_j(SQ)^2 \le 1 + \epsilon, \qquad 1 \le j \le n$$

Therefore

- SQ has full column-rank: $\min_j \sigma_j(SQ) \ge \sqrt{1-\epsilon} > 0$
- Left inverse exists and is bounded

$$\|(SQ)^\dagger\|_2 = \frac{1}{\min_j \sigma_j(SQ)} \le \frac{1}{\sqrt{1-\epsilon}}$$

- Condition number is bounded

$$\kappa_2(SQ) = \|SQ\|_2 \|(SQ)^\dagger\|_2 = \frac{\max_j \sigma_j(SQ)}{\min_j \sigma_j(SQ)} \le \sqrt{\frac{1+\epsilon}{1-\epsilon}}$$

Matrix Bernstein Concentration Inequality [Recht 2011]

Assume

- Zero-mean: Independent random $n \times n$ matrices Y_t with $\mathbb{E}[Y_t] = 0_{n\times n}$
- Boundedness: $\| Y_t \|_2 \leq \tau$ almost surely
- Variance: $\rho_t \equiv \max\{\| \mathbb{E}[Y_t Y_t^T] \|_2, \| \mathbb{E}[Y_t^T Y_t] \|_2\}$
- Desired error tolerance: $0 < \epsilon < 1$
- Failure probability: $\delta = 2n \exp\left(-\frac{3}{2} \frac{\epsilon^2}{3 \sum_t \rho_t + \tau \epsilon}\right)$

Then with probability at least $1 - \delta$

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

Apply the Concentration Inequality

Sampled matrix

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

Zero-mean version

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

Check assumptions

- Zero mean: $\mathbb{E}[Y_t] = 0$ {by construction}
- Boundedness: $\|Y_t\|_2 \leq \frac{m}{c}\,\mu$
- Variance: $\|\mathbb{E}[Y_t^2]\|_2 \leq \frac{m}{c^2}\,\mu$

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

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

Condition Number Bound [Ipsen & Wentworth 2014]

Assume

- $m \times n$ matrix Q with $Q^T Q = I_n$ {orthonormal columns}
- Largest row norm squared: $\mu = \max_{1 \leq j \leq m} \|e_j^T Q\|_2^2$
- Number of sampled rows: $c \geq n$
- Desired error tolerance: $0 < \epsilon < 1$
- Failure probability

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

Then with probability at least $1 - \delta$

Condition number of sampled matrix $\kappa(SQ) \leq \sqrt{\frac{1+\epsilon}{1-\epsilon}}$

Tightness of Condition Number Bound

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

Investigate: $c \times n$ matrix SQ {sampling with replacement}

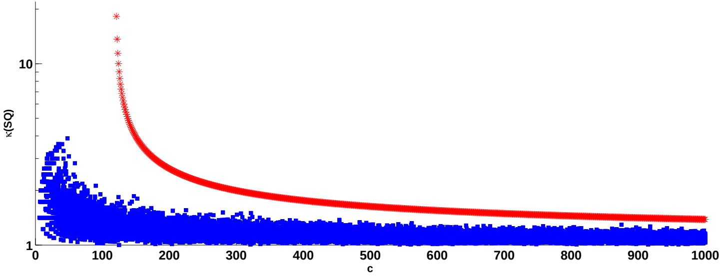
Little sampling: $n \leq c \leq 1000$
A lot of sampling: $1000 \leq c \leq m$

Plots:

1. Exact condition number $\kappa(SQ)$
2. Bound $\kappa(SQ) \leq \sqrt{\frac{1+\epsilon}{1-\epsilon}}$ with probability $1-\delta \equiv .99$

$$\begin{aligned} \epsilon &\equiv \frac{1}{2c}\left(\ell + \sqrt{12c\ell + \ell^2}\right) \\ \ell &\equiv \frac{2}{3}\,(m\,\mu - 1)\,\ln(2n/\,\delta) = \Omega\,(m\,\mu\,\ln n) \end{aligned}$$

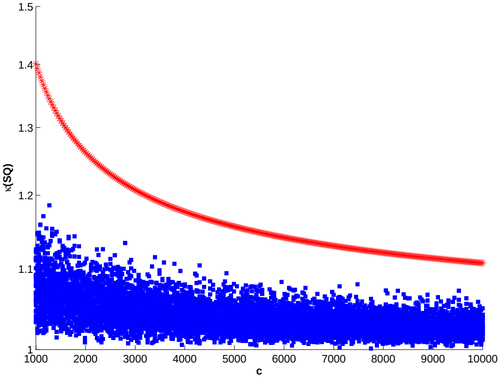
Little sampling ($n \leq c \leq 1000$)



Exact condition numbers $\kappa(SQ)$

Bound holds starting from $c \geq 93 \; \approx \; 3\ell \; = \; \Omega\,(m\,\mu\,\ln n)$

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



Bound predicts correct magnitude of condition numbers

Conclusions for Condition Number Bound

Given: $m \times n$ matrix Q with $Q^T Q = I_n$ {orthonormal columns}
Sampling: $c \times n$ matrix SQ

Bound on condition number $\kappa(SQ)$ of sampled matrix:

- Correct magnitude
- Informative even for small matrix dimensions
  and stringent success probabilities
- Implies lower bound on number of sampled rows

$$c = \Omega\left(m\,\mu \ln n\right)$$

- Depends on coherence of Q: $\mu = \max_{1 \le j \le m} \|e_j^T Q\|_2^2$

  Largest squared row norm of Q
  Reveals distribution of mass in Q

Coherence

Properties of Coherence

Coherence of $m \times n$ matrix Q with $Q^T Q = I_n$ {orthonormal columns}

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

- $n/m \leq \mu(Q) \leq 1$
- Maximal coherence: $\mu(Q) = 1$
  At least one column of Q is column of identity
- Minimal coherence: $\mu(Q) = n/m$
  Columns of Q are columns of Hadamard matrix

Definition can be extended to: general matrices, subspaces

Properties of Coherence

Coherence of $m \times n$ matrix Q with $Q^T Q = I_n$ {orthonormal columns}

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

- $n/m \le \mu(Q) \le 1$
- Maximal coherence: $\mu(Q) = 1$
  At least one column of Q is column of identity
- Minimal coherence: $\mu(Q) = n/m$
  Columns of Q are columns of Hadamard matrix

Coherence

- Measures correlation with standard basis
- Reflects difficulty of recovering the matrix from sampling

The Origins of Coherence

- 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

Effect of Coherence on Sampling

Input: $m \times n$ matrix Q with $Q^T Q = I_n$ {orthonormal columns}
$m = 10^4, \ \ n = 5$

Investigate: $c \times n$ matrix SQ {sampling with replacement}

Question: How does coherence of Q affect sampling?

Two types of matrices Q

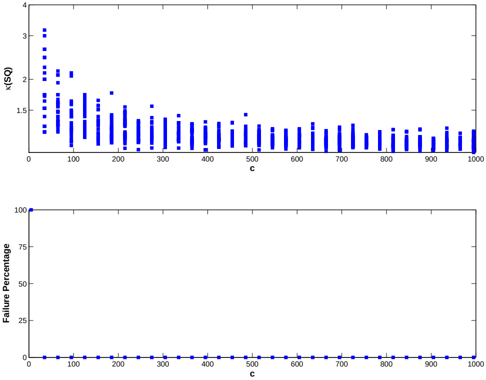
1. Low coherence: $\mu = 7.5 \cdot 10^{-4} = 1.5\, n/m$
2. Higher coherence: $\mu = 7.5 \cdot 10^{-2} = 150\, n/m$

Plots for $n \leq c \leq 1000$

1. Percentage of numerically rank-deficient SQ $\{\kappa(SQ) \geq 10^{16}\}$
2. Condition number of full column-rank SQ

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

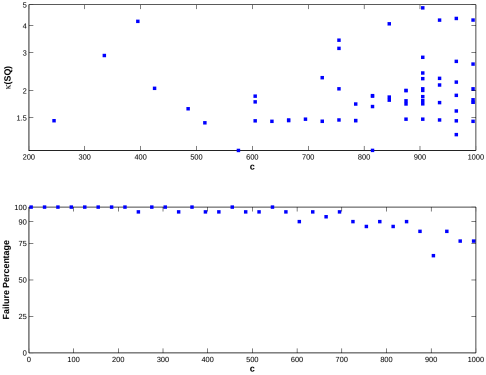
Sampling Rows from Q with Low Coherence



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

Full-rank matrices SQ perfectly conditioned: $\kappa(SQ) < 4$

Sampling Rows from Q with Higher Coherence



Sampling up to 10% of rows:

- Most matrices SQ are rank-deficient
- Full-rank matrices SQ perfectly conditioned: $\kappa(SQ) \leq 5$

Effect of Coherence on Sampling: Conclusions

Given: $m \times n$ matrix Q with $Q^T Q = I_n$ {orthonormal columns}
Investigate: $c \times n$ sampled matrix SQ

Q has low coherence $\mu \approx n/m$

- Most SQ full-rank and perfectly conditioned {even for small c}
- Mass of Q uniformly distributed {it does not matter what you pick}
- Sampling is easy

Q has higher coherence $\mu \approx 100n/m$

- Most SQ rank-deficient {even for larger c}
- Mass of Q concentrated in a few spots {you have to be lucky}
- Sampling is hard

A Few Take Aways for Randomized Least Squares Solvers

$$\min_x \|Ax - b\|_2$$

- Sampling is effective if A has good coherence ('uniformity')
- Powerful matrix concentration inequalities are important
- Not discussed: Improving coherence with fast multiplication by random matrix
- The 'safe' randomized LS solver: *Blendenpik*
  Randomization confined to preconditioner

Research questions

- Numerical behavior in floating point arithmetic
- Effect of sampling on statistical model uncertainty
- Flexible preconditioners that can change in every iteration
- Regularization for ill-posed problems

Resources: Surveys and Books

- L. Devroye: Nonuniform Random Variate Generation
  Springer-Verlag (1986)
- M. Mitzenmacher and E. Upfal:
  Probability and Computing: Randomized Algorithms and Probabilistic Analysis, Cambridge University Press (2005)
- R. Vershynin:
  High-Dimensional Probability: An Introduction with Applications in Data Science, Cambridge University Press (2018)
- J. A. Tropp: An Introduction to Matrix Concentration Inequalities
  Found. Trends Mach. Learning, vol. 8, no. 1-2, pp 1-230 (2015)
- N. Halko, P.G. Martinsson and J.A. Tropp:
  Finding Structure with Randomness: Probabilistic Algorithms for Constructing Approximate Matrix Decompositions
  SIAM Rev., vol. 53, no. 2, pp 217–288 (2011)
- M.W. Mahoney: Randomized Algorithms for Matrices and Data
  Found. Trends Mach. Learn., vol. 3, pp 123–224 (2011)

Resources: Papers Discussed in this Talk

- H. Avron, P. Maymounkov, and S. Toledo
  Blendenpik: Supercharging Lapack's Least-Squares Solver
  SIAM J. Sci. Comput., vol. 32, no. 3, pp 1217–1236 (2010)
- P. Drineas, M.W. Mahoney, S. Muthukrishnan, and T. Sarlós
  Faster Least Squares Approximation
  Numer. Math., vol. 117, no. 2, pp 219–249 (2011)
- I.C.F. Ipsen and T. Wentworth
  The Effect of Coherence on Sampling from Matrices with Orthonormal Columns, and Preconditioned Least Squares Problems
  SIAM J. Matrix Anal. Appl., vol. 35, no. 4, pp 1490–1520 (2014)
- J.T. Chi and I.C.F. Ipsen
  Multiplicative Perturbation Bounds for Multivariate Multiple Linear Regression in Schatten p-Norms
  Linear Algebra Appl. (to appear)
- J.T. Chi and I.C.F. Ipsen
  A Projector-Based Approach to Quantifying Total and Excess Uncertainties for Sketched Linear Regression
  arXiv:1808.0594