Randomized Matrix Computations Lecture 5

Daniel Kressner
Chair for Numerical Algorithms and HPC
Institute of Mathematics, EPFL
daniel.kressner@epfl.ch

Random Embeddings

- Overdetermined least-squares
- JL and subspace embeddings
- Gaussian embeddings
- Structured random embeddings
- Overdetermined least-squares revisited
- Sketched Gram-Schmidt
- Kaczmarz

Literature:

Tropp'2020 Joel A. Tropp. *Randomized Algorithms for Matrix Computations*, Lecture notes, Caltech, 2020.

Vershynin'2012 Roman Vershynin. *Introduction to the non-asymptotic analysis of random matrices*. In "Compressed Sensing, Theory and Applications". CUP'2012.

Vershynin'2018 Roman Vershynin's HDP.

pdf available on Moodle

Overdetermined least-squares problems

Consider overdetermined least-squares problem

$$\min\{\|Ax - b\|_2 : x \in \mathbb{R}^m\}, \quad A \in \mathbb{R}^{d\times m}, \quad b \in \mathbb{R}^d.$$

Assumptions: d (number of observations) $\gg$ m (number of variables). A has full rank $m \rightsquigarrow$ solution x uniquely determined

Classical approach:

- Compute QR decomposition $A = QR$ s.t. $R \in \mathbb{R}^{m\times m}$ is upper triangular and $Q \in \mathbb{R}^{d\times n}$ has orthonormal columns.
- Solve reduced problem $Rx = Q^\top b$.

Reinterpretation: The application of $Q^\top$ produces a *sketch* $Q^\top A \approx Q^\top b$ of $Ax \approx b$ and we obtain the solution of the original problem from solving the sketched problem.

Overdetermined least-squares problems

Idea of sketching: Replace $Q^\top \in \mathbb{R}^{m\times d}$ by more general sketching matrix $S \in \mathbb{R}^{n\times d}$ with $m \le n \ll d$ and solve sketched problem

$$\min\{\|SA\tilde{x} - Sb\|_2 : \tilde{x} \in \mathbb{R}^m\}. \tag{1}$$

A good sketching matrix S should approximately capture the range of A and (optionally) b. A common way to quantify this is the subspace embedding property. Given subspace $\mathcal{U} \subset \mathbb{R}^d$, S is called an ϵ-subspace embedding for $0 < \epsilon < 1$ if

$$(1-\epsilon)\|u\|_2^2 \le \|Su\|_2^2 \le (1+\epsilon)\|u\|_2^2 \quad \forall u \in \mathcal{U}.$$

Lemma (sketch and solve). Suppose that $\tilde{x}$ solves the sketched least-squares problem (1) with an ϵ-subspace embedding S of the subspace span$([A, b])$. Then

$$\|A\tilde{x} - b\|_2^2 \le \frac{1+\epsilon}{1-\epsilon}\|Ax - b\|_2^2.$$

Overdetermined least-squares problems

Proof of Lemma. Let x solve $\min \|Ax - b\|_2$. Then

$$\begin{aligned} \|A\tilde{x} - b\|_2^2 &\leq \frac{1}{1-\epsilon}\|SA\tilde{x} - Sb\|_2^2 \leq \frac{1}{1-\epsilon}\|SAx - Sb\|_2^2 \\ &\leq \frac{1+\epsilon}{1-\epsilon}\|Ax - b\|_2^2, \end{aligned}$$

where we used that $Ax - b, A\tilde{x} - b \in \text{span}([A, b])$. $\diamond$

A trivial subspace embedding (with $\epsilon = 0$) is to take $S = U^\top$, where U is an orthonormal basis of $\text{span}([A, b])$. But this is not what the lemma is aiming for. We aim for cheap constructions that use little or even no information about A, b.

OSE

A subspace embedding is called oblivious if it works for an arbitrary subspace.

Definition A (k, ϵ, δ)-Oblivious Subspace Embedding (OSE) is a random matrix S such that, for a *fixed but arbitrary* k-dimensional subspace $\mathcal{U} \subset \mathbb{R}^d$, S is an ϵ-subspace embedding with probability at least $1 - \delta$, that is,

$$(1 - \epsilon)\|u\|_2^2 \leq \|Su\|_2^2 \leq (1 + \epsilon)\|u\|_2^2 \quad \forall u \in \mathcal{U} \tag{2}$$

holds.[1]

EFY. Let $U \in \mathbb{R}^{n \times k}$ be an orthonormal basis for $\mathcal{U}$. Show that each of the following properties is equivalent to OSE:

1. $\max\{|\|Su\|_2^2 - 1| : u \in \mathcal{U}, \|u\|_2 = 1\} \leq \epsilon$
2. $\sigma_{\min}(SU)^2 \geq 1 - \epsilon$, $\sigma_{\max}(SU)^2 \leq 1 + \epsilon$
3. $\lambda_{\min}(U^\top S^\top SU) \geq 1 - \epsilon$, $\lambda_{\max}(U^\top S^\top SU) \leq 1 + \epsilon$
4. $\|I - U^\top S^\top SU\|_2 \leq \epsilon$

[1] One sometimes finds this definition without the squares. This only makes a marginal difference.

OSE and a second encounter with JL

For $k = 1$, OSE reduces to JL.

Definition A random matrix S satisfies the (ϵ, δ)-Johnson-Lindenstrauss (JL) property if for a *fixed but arbitrary* vector u the inequalities

$$(1 - \epsilon)\|u\|_2^2 \le \|Su\|_2^2 \le (1 + \epsilon)\|u\|_2^2 \tag{3}$$

hold with probability at least $1 - \delta$.

We have already seen that $S = \Omega/\sqrt{n}$ for $n \times d$ Gaussian random matrix satisfies the (ϵ, δ)-JL property if

$$n \ge 8\epsilon^{-2} \log 2/\delta.$$

Bad dependence on ϵ; great dependence on δ!

From JL to OSE

When extending JL to OSE for k-dimensional subspaces $\mathcal{U} \subset \mathbb{R}^d$, the union bound suffers from the obvious problem that $\mathcal{U}$ contains infinitely many vectors. Popular techniques in stochastic analysis to overcome such problems: epsilon nets and chaining.

Idea of ϵ-nets: Given ONB $U \in \mathbb{R}^{d\times k}$, every (normalized) vector in $\mathcal{U}$ takes the form Ux, $x \in S^{k-1}$ (unit sphere in $\mathbb{R}^k$). Cover the sphere with vectors up to a distance $\epsilon = O(1)$ and use union bound. MANY vectors will be needed, so $O(|\log\delta|)$ dependence of JL is needed to save us!

Lemma 5.2 in Vershynin'2012. S^{k-1} has an epsilon net $\mathcal{N}_{\epsilon_{\text{net}}} \subset S^{k-1}$ of cardinality at most $(1+2/\epsilon_{\text{net}})^k$, that is, for every $x \in S^{k-1}$ there is $y \in N_{\epsilon_{\text{net}}}$ such that

$$\|x-y\|_2 \le \epsilon_{\text{net}}.$$

Think of ϵ_{net} not too small, like $\epsilon_{\text{net}} = 1/2$ or $\epsilon_{\text{net}} = 1/4$.

From JL to OSE

For *symmetric* $k \times k$ matrix C: $\|C\|_2 = \max\{|x^\top Cx| : x \in S^{k-1}\}$.

Corollary. Let $N_{\epsilon_{\text{net}}}$ be epsilon net from Lemma. Then

$$\|C\|_2 \leq (1 - 2\epsilon_{\text{net}}^2)^{-1} \max\{|y^\top Cy| : y \in N_{\epsilon_{\text{net}}}\}.$$

Proof. Let $x \in S^{k-1}$ s.t. $Cx = \lambda_1 x$ and $\|C\|_2 = |\lambda_1|$. Let $y \in N_{\epsilon_{\text{net}}}$ s.t. $\|x - y\|_2 \leq \epsilon_{\text{net}}$. Then

$$\left|y^\top Cy - x^\top Cx\right| = \left|(x-y)^\top (C - \lambda_1 I)(x-y)\right| \leq 2\|C\|_2 \|x-y\|_2^2 \leq 2\epsilon_{\text{net}}^2 \|C\|_2.$$

This implies $|y^\top Cy| \geq |x^\top Cx| - |x^\top Cx - y^\top Cy| \geq (1 - 2\epsilon_{\text{net}}^2)\|C\|_2$. ◇

From JL to OSE

Theorem. Any random matrix S satisfying the $(\epsilon/2, \delta/5^k)$-JL property also satisfies the (k, ϵ, δ)-OSE property.

Proof. Choose $\epsilon = 1/2$. By JL we have

$$\left|y^\top \left(I - U^\top S^\top SU\right) y\right| \leq \epsilon/2 \quad \text{with probability} \geq 1 - \delta/5^k.$$

for arbitrary $y \in S^{k-1}$. By the union bound this shows JL holds for all vectors in $\mathcal{N}_{1/2}$ with prob $\geq 1 - \delta$. By the corollary,

$$\|I - U^\top S^\top SU\|_2 \leq 2 \max\{|y^\top (I - U^\top S^\top SU) y| : y \in N_{1/2}\} \leq \epsilon,$$

which shows OSE. $\diamond$

By JL for Gaussian random matrices, this establishes OSE if

$$n \geq 32\epsilon^{-2}(k \log 5 + \log 2/\delta) = \mathcal{O}\left(\epsilon^{-2}(k + \log \delta^{-1})\right).$$

This general construction gets the asymptotics right, but the constants are slightly too large.

Tighter bounds on Gaussian embeddings

Let Ω be an $n \times k$ Gaussian random matrix. Then Section 7.3 of [Versyhnin'2018] shows the following properties:

$$\mathbb{E}\|A\|_2 \leq \sqrt{n} + \sqrt{k}, \quad \mathbb{E}\sigma_{\min}(A) \geq \sqrt{n} - \sqrt{k}.$$

Now let S be an $n \times d$ Gaussian random matrix scaled by $1/\sqrt{n}$. Then, by invariance of Gaussian random vectors under rotations, $\tilde{\Omega} = SU$ is $n \times k$ Gaussian random matrix scaled by $1/\sqrt{n}$. Section 8.5 in [Martinsson/Tropp] shows

$$\mathbb{P}\left\{\sigma_{\min}(\tilde{\Omega}) \leq 1 - \frac{\sqrt{k}+1}{\sqrt{n}} - t\right\} \leq e^{-nt^2/2}$$

$$\mathbb{P}\left\{\|\tilde{\Omega}\|_2 \geq 1 + \frac{\sqrt{k}}{\sqrt{n}} + t\right\} \leq e^{-nt^2/2}$$

EFY: Show that these bounds imply that S is (k, ϵ, δ)-OSE if

$$n \geq 4\epsilon^{-2}(1 + k + \log 2/\delta).$$

OSE beyond Gaussians

Results essentially extend to matrices with sub-Gaussian iid entries (e.g., Rademacher).

Sketching arbitrary $d \times m$ matrix A with $n \times d$ (sub-)Gaussian matrix S requires require $O(ndm) = O(kdm)$ operations. Can be reduced by imposing structure on S.

Most popular choices for structured random embeddings:

- Coordinate sampling
- Sparse sign matrices
- Subsampled unitary transforms
- Khatri-Rao products

Coordinate sampling

An immediate and cheap choice:

$$S = \begin{bmatrix} s_1^\top \\ s_2^\top \\ \vdots \\ s_n^\top \end{bmatrix}, \quad s_i \text{ are iid with } \mathbb{P}\{s_i = e_j/\sqrt{p_j n}\} = p_j,$$

for unit vectors $e_1, \dots, e_d \in \mathbb{R}^d$ and prescribed sampling probabilities $p_1, \dots, p_n$.

- Computing SA requires nm operations. Looks like the LARGE dimension d disappeared![2]
- OSE characterization 4 (see Slide 6) links to matrix Monte Carlo:

$$\|I - U^\top S^\top SU\|_F \le \epsilon.$$

(k, ϵ, δ)-OSE = Approximate matmul from Lecture 4 applied to $U^\top U = I$ returns error ϵ with probability $\geq 1 - \delta$.

[2] well, well... this is not exactly true as we will see on the next slides

Coordinate sampling: Uniform

Uniform sampling: $p_1 = \cdots = p_d = 1/d$. Already know that performance depends on coherence

$$\mu(U) = d \cdot \max_{i=1,\ldots,d} \|U(i,:)\|_2^2$$

- $\mu(U)$ is independent of choice of ONB U for $\mathcal{U}$
- $k \leq \mu(U) \leq d$. Smaller $\mu(U)$ is better.
  EFY: Prove lower bound. Can you find a matrix U that nearly attains lower bound?

Apply Matrix Monte Carlo Theorem (L4S16) with $\|X\|_2$, $\|\mathbb{E}[XX^\top]\|_2$, $\|\mathbb{E}[X^\top X]\|_2$ bounded by $\mu(U)$:

$$\mathbb{P}\{\|I - U^\top S^\top SU\|_F \geq \epsilon\} \leq 2k \exp\Big(-\frac{n\epsilon^2}{\mu(U)(1+2\epsilon/3))}\Big).$$

Hence, given U, S is ϵ-subspace embedding for $0 < \epsilon \leq 1$ with probability $\geq 1 - \delta$ when

$$n \geq 2\mu(U)\epsilon^{-2}\big(\log(2k) + \log \delta^{-1}\big).$$

$\rightsquigarrow$ In the best case $n = O(k \log k)$. In the worst case $n = O(d \log k)$ (completely useless).

Coordinate sampling: Leverage scores

We now set

$$p_i = \frac{1}{k} \| U(i, :) \|_2^2, \quad i = 1, \dots, d. \tag{4}$$

By the discussion from L4S20, we can apply Matrix Monte Carlo Theorem with

$$\|X\|_2 \leq k, \quad \|\mathbb{E}[XX^\top]\|_2 \leq 2k, \quad \|\mathbb{E}[X^\top X]\|_2 \leq 2k.$$

Hence, S is (k, ϵ, δ)-subspace embedding for U with $0 < \epsilon \leq 1$ when

$$n \geq 3k\epsilon^{-2}\big(\log(2k) + \log \delta^{-1}\big).$$

This looks good, but subspace embedding is not oblivious.

Sparse sign matrices

Sparse sign matrices come in two flavors:

1. Fixed sparsity sign matrices:

   Each *column* of S has exactly s (scaled) ± 1 entries at random locations.

   Extreme case $s = 1$: OSE holds for $n = O(k^2\epsilon^{-2}\delta^{-1})$ [Nelson/Nguyen'2013].

   Reasonable choice $s = O(\epsilon^{-1}(\log k + \log \delta^{-1}))$:
   OSE holds for $n = O(\epsilon^{-2}k(\log k + \log \delta^{-1}))$ [Cohen'2016]

2. iid sparsity sign matrices: Consider random sparse sign matrix

$$S = \frac{1}{\sqrt{pn}} \begin{bmatrix} s_{11} & \cdots & s_{1d} \\ \vdots & & \vdots \\ s_{n1} & \cdots & s_{nd} \end{bmatrix} \in \mathbb{R}^{n\times d}, \text{with iid } s_{ij} \text{ s.t. } \begin{array}{l} \mathbb{P}(s_{ij} = +1) = p/2, \\ \mathbb{P}(s_{ij} = -1) = p/2, \\ \mathbb{P}(s_{ij} = 0) = 1 - p. \end{array}$$

   Scaling ensures that $\mathbb{E}[S^\top S] = I_d$.

Sparse sign matrices

Focus on iid sparsity sign matrices in the following.

EFY: Show that $\mathbb{E}\|Sx\|_2^2 = \|x\|_2^2$ for all $x \in \mathbb{R}^d$. An embedding satisfying this property is called isotropic.

Decompose

$$Y = U^\top S^\top SU = \frac{1}{pn}\sum_{j=1}^{n} U^\top s_j s_j^\top U = \sum_{j=1}^{n} X_j, \quad X_j := \frac{1}{pn} U^\top s_j s_j^\top U,$$

where $s_j^\top$ is jth row of S. Then $\mathbb{E}Y = I$ and

$$\mathbb{E}\|X_j\|_2 = \frac{1}{pn}\mathbb{E}\|U^\top s_j\|_2^2 = \frac{k}{n}, \quad \text{but} \quad \|X_j\|_2 = \frac{1}{pn}\|U^\top s_j\|_2^2 \le \frac{d}{pn}.$$

Unfortunate dependence on LARGE d in upper bound!

Sparse sign matrices

Cohen'2016 / Tropp'2020: Truncation of distribution to cap large $\|X_j\|$ + Chernoff.

Intuition of the argument: Need sufficiently large p to have concentration of norm:

$$p \gtrsim \epsilon^{-2} n^{-1} (\log k + \log \delta^{-1})$$

Chernoff gives (k, ϵ, δ)-OSE for

$$n \gtrsim \epsilon^{-2} k (\log k + \log \delta^{-1})$$

More refined analysis in [Tropp'2016] based on Matrix Rosenthal inequalities.

For fixed ϵ^{-2}, δ, there are $O(d \log k)$ nonzero entries in S
$\rightsquigarrow$ Sketching $d \times m$ matrix requires $O(md \log k)$ operations.

However, reduced dimension increases to $k \log k$. Idea: Apply another $O(k \times k \log k)$ Gaussian sketch to bring dimension down to $O(k)$. Cheap when $d \gg k^2$.

Subsampled trigonometric transforms

Idea: First apply (random) orthogonal transformation to assure incoherence. Then use uniform sampling.

Theorem [Avron et al.2010] Let $U \in \mathbb{R}^{d \times k}$ be ONB. Let F be $d \times d$ orthogonal matrix and D diagonal with iid Rademacher diagonal entries. Then

$$\mu(FDU) \leq Ckd\eta \log d, \quad \text{with} \quad \eta = \max_{ij} |f_{ij}|^2.$$

holds with probability at least 0.95 for some constant C.

Proof. W.l.o.g., may assume $\eta = 1$. Let x_{ij} denote (i, j) entry of FDU with $i = 1, \ldots, d$, $j = 1, \ldots, k$. EFY: Show that x_{ij} is sub-Gaussian(1).

Hoeffding's inequality shows that

$$\mathbb{P}\{|x_{ij}| \geq t\} \leq 2\exp(-t^2/2).$$

By the union bound, this implies

$$\mathbb{P}\{|x_{ij}| \geq t\} \leq 2dk\exp(-t^2/2), \quad \forall i, j.$$

Subsampled trigonometric transforms

For the squared row norms, this implies that

$$\mathbb{P}\left\{|x_{i1}|^2 + \cdots + |x_{ik}|^2 \geq kt^2\right\} \leq 2dk \exp(-t^2/2), \quad \forall i.$$

Setting $t = \sqrt{2\log(40dk)}$ gives

$$\mathbb{P}\left\{|x_{i1}|^2 + \cdots + |x_{ik}|^2 \geq 2k\log(40dk)\right\} \leq 0.05, \quad \forall i.$$

Using that $k \leq d$, this shows the desired result by the definition of μ. $\diamond$

Subsampled trigonometric transforms

The result of the theorem is *not* optimal! We can avoid taking the union bound wrt j by using refined results on functions of Rademacher vectors.

Lemma [Ledoux'1996] Suppose that $f : \mathbb{R}^d \to \mathbb{R}$ is convex and Lipschitz continuous with Lipschitz constant L_f. Let $Z \in \mathbb{R}^d$ be a Rademacher random vector. Then for all $t \geq 0$,

$$\mathbb{P}\Big\{ f(Z) \geq \mathbb{E} f(Z) + L_f t \Big\} \leq e^{-t^2/8}.$$

We apply this result to the norm of the ith row of FDU:

$$f(z) = \| e_i^\top F \mathrm{diag}(z) U \|_2 = \| z^\top E U \|_2, \quad E = \mathrm{diag}(f_{i1}, \ldots, f_{id}).$$

f is clearly convex. EFY: Show that f is Lipschitz with $L_f = 1$, assuming that $\eta = 1$. Show that $\mathbb{E} f(Z) = \sqrt{k}$. Ledoux tells us that

$$\mathbb{P}\Big\{ \| e_i^\top F \mathrm{diag}(z) U \|_2 \geq \sqrt{k} + t \Big\} \leq e^{-t^2/8}.$$

Subsampled trigonometric transforms

By the union bound, $\mathbb{P}\left\{\|e_i^\top F\mathrm{diag}(z)U\|_2 \geq \sqrt{k}+t\right\} \leq de^{-t^2/8}, \quad \forall i.$
Following the steps above, this shows:

Theorem [Tropp'2011] Let $U \in \mathbb{R}^{d\times k}$ be ONB. Let F be $d\times d$ orthogonal matrix and D diagonal with iid Rademacher diagonal entries. Then

$$\mu(FDU) \leq C\eta d(k+\log d), \quad \text{with} \quad \eta = \max_{ij}|f_{ij}|^2.$$

holds with probability at least 0.95 for some constant C.

Note the result also holds for unitary matrix F.

- Best we can hope for is $\eta = 1/d$.
- Let R be $n\times d$ coordinate sampling matrix, that is, each row is $e_j^\top/\sqrt{n}$ with probability $1/d$.
- Then $S = RFD$ is (k,ϵ,δ)-OSE for $0<\epsilon\leq 1$ when

$$n \sim \mu(S)\epsilon^{-2}\log k \sim \epsilon^{-2}(k+\log d)\log k,$$

  for fixed δ. EFY: Prove the second relation rigorously and work out the asymptotic dependence on δ.

Subsampled trigonometric transforms

Most popular choices for F:

- SRFT (Subsampled Randomized Fourier Transform): F is the discrete Fourier transform

$$f_{ij} = \frac{1}{\sqrt{d}}\left(e^{-2\pi \mathrm{i}/d}\right)^{(i-1)(j-1)}, \quad \eta = 1/d.$$

  Applying subsampled Fourier transform RF to a vector can be carried out in $O(d \log n)$ ops. With $n \sim (k + \log d)\log k$, need

$$O\big(d(\log(k + \log d) + \log\log k)\big) \approx O(d \log k)$$

  ops to apply RFD to a vector.
- Subsampled Randomized Hartley Transform / Subsampled Randomized Cosine Transform = real variants of SRFT.
- SRHT (Subsampled randomized Hadamard transform) $\Omega = RHD$, where $H = \frac{1}{\sqrt{n}}\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \otimes \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \otimes \cdots \otimes \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$
  (zero padding if n is not a power of 2)
  OSE holds for $n = \mathcal{O}\big(k \log(1/\delta) \log(d/\delta)\big)$ [Boutsidis/Gittens'2013]

Overdetermined least-squares revisited

$$\min\{\|Ax - b\|_2 : x \in \mathbb{R}^m\}, \quad A \in \mathbb{R}^{d\times m}, \quad b \in \mathbb{R}^d.$$

Recall result of Lemma (sketch and solve):

$$\|A\tilde{x} - b\|_2^2 \leq \frac{1+\epsilon}{1-\epsilon}\|Ax - b\|_2^2.$$

Two issues:

- Quality of sketch directly impacts quality of the LSQ solution.
- VERY expensive to get high accuracy! ($n \sim \epsilon^{-2}$)

Idea: Use sketching as preconditioner in iterative solver $\rightsquigarrow$ BLENDENPIK.

Iterative solvers for LSQ problems

Consider $m \times m$ SPD linear system $Cx = d$. The method of conjugate gradients (CG) only requires j matrix-vector products with C and $O(m)$ extra storage to produce approximation x_j satisfying

$$\|x - x_j\|_C \leq 2\|x - x_0\|_C \left(\frac{\sqrt{\kappa(C)} - 1}{\sqrt{\kappa(C)} + 1} \right)^j,$$

with condition number

$$\kappa(C) = \frac{\sigma_{\max}(C)}{\sigma_{\min}(C)} = \frac{\lambda_{\max}(C)}{\lambda_{\min}(C)},$$

see, e.g., [Golub/Van Loan'2013].

LSQR [Paige/Saunders] for solving LSQ problem is *mathematically* equivalent to applying CG to normal equations $A^\top Ax = A^\top b$. Convergence:

$$\|b - Ax_j\|_2 \leq 2\|b - Ax_0\|_2 \left(\frac{\kappa(A) - 1}{\kappa(A) + 1} \right)^j,$$

Iterative solvers for LSQ problems

Use sketching as preconditioner: Compute QR decomposition

$$SA = \hat{Q}\hat{R}$$

and precondition least-squares problem

$$\min \|Ax - b\|_2 = \min \|A\hat{R}^{-1} \underbrace{\hat{R}x}_{:=\tilde{x}} - b\|_2$$

Let S be ϵ-embedding of range(A) and consider QR decomposition $A = QR$. Then

$$\kappa(A\hat{R}^{-1}) = \kappa(QR\hat{R}^{-1}) = \kappa(R\hat{R}^{-1}) = \kappa(\hat{Q}\hat{R}R^{-1}) = \kappa(SQ) \le \sqrt{\frac{1+\epsilon}{1-\epsilon}}.$$

Choose $\epsilon = 1/2$. Then LSQR applied to $\min \|AR^{-1}\tilde{x} - b\|_2$ converges at rate ≈ 0.27.

To attain accuracy $\varepsilon \rightsquigarrow \sim |\log \varepsilon|$ iterations needed. Complexity depends on $|\log \varepsilon|$ instead of ε^{-2}! More detailed discussion of complexity in Sec 8.3 of [Tropp'2020].

Sketched Gram-Schmidt

An ubiquitous (and often expensive) task in scientific computing: Given a set of (linearly independent) vectors $x_1, \ldots, x_k \in \mathbb{R}^d$, determine orthonormal basis $q_1, \ldots, q_k$ of span$(x_1, \ldots, x_k)$.

Gram-Schmidt process. For $j = 1, \ldots, k$:

$$q_j' = x_j - \sum_{i=1}^{j-1} \langle q_i, x_j \rangle q_i = x_j - Q_{j-1} Q_{j-1}^{\top} x_j, \quad q_j = q_j' / \| q_j' \|_2, \tag{5}$$

where $Q_{j-1} = [q_1, \ldots, q_{j-1}]$. On (distributed memory) massive parallel computers, $\langle q_i, x_j \rangle$ requires global communication and is expensive.

Idea: Instead of orthonormality, attain sketch-orthonormality, that is, only $\hat{Q}_k = SQ_k \in \mathbb{R}^{n \times k}$ is orthonormal.

Sketched Gram-Schmidt process:

$$\hat{q}_j' = Sx_j - \hat{Q}_{j-1} \hat{Q}_{j-1}^{\top} (Sx_j), \quad \hat{q}_j = \hat{q}_j' / \| \hat{q}_j' \|_2.$$

Sketched Gram-Schmidt

Set $r_j = \hat{Q}_{j-1}^{\top}(Sx_j)$ and $r_{jj} = \|\hat{q}_j'\|_2$. Then sketched Gram-Schmidt becomes

$$\hat{q}_j' = Sx_j - \hat{Q}_{j-1} r_j, \quad \hat{q}_j = \hat{q}_j' / r_{jj}.$$

At the same time, we compute

$$q_j' = x_j - Q_{j-1} r_j, \quad q_j = q_j' / r_{jj}.$$

By induction $\rightsquigarrow$ Relation $\hat{Q}_j = SQ_j$ is maintained.

Randomized Gram-Schmidt

Input: $x_1, \ldots, x_k \in \mathbb{R}^d$, sketching matrix $S \in \mathbb{R}^{n \times d}$

Output: Sketch-orthonormal basis Q_k of span$\{x_1, \ldots, x_k\}$.

1: $Q_0 = [\,]$, $\hat{Q}_0 = [\,]$.

2: **for** $j = 1, 2, \ldots, k$ **do**

3: Sketch vector $s_j = Sx_j$

4: Sketched GS: $r_j = \hat{Q}_{j-1}^{\top}(Sx_j)$, $\hat{q}_j' = s_j - \hat{Q}_{j-1} r_j$, $r_{jj} = \|\hat{q}_j'\|_2$, $\hat{Q}_j = [\hat{Q}_{j-1}, \hat{q}_j'/r_{jj}]$

5: Update in $\mathbb{R}^d$: $q_j' = x_j - Q_{j-1} r_j$, $Q_j = [Q_{j-1}, q_j'/r_{jj}]$

6: **end for**

Sketched Gram-Schmidt

Analysis of Randomized Gram-Schmidt: By construction, $\hat{Q}_k$, Q_k satisfy the decompositions

$$S[x_1 \dots, x_k] = \hat{Q}_k \hat{R}_k, \quad [x_1 \dots, x_k] = Q_k \hat{R}_k,$$

where the upper triangular matrix $R_k \in \mathbb{R}^{k \times k}$ contains the coefficients from the sketched GS process.

By a QR decomposition $[x_1 \dots, x_k] = UR_U$, we compute an ONB U of $[x_1 \dots, x_k]$. This yields $SU = \hat{Q}_k R$ for $R = \hat{R}_k R_U^{-1}$. If S is random and has the (k, ϵ, δ)-OSE property we have

$$\kappa(SU)^2 = \kappa(R)^2 \leq \frac{1+\epsilon}{1-\epsilon} \quad \text{with probability} \geq 1-\delta.$$

OTOH, $U = Q_k R$ and hence $Q_k = UR^{-1}$, which implies

$$\kappa(Q_k)^2 = \kappa(R)^2 \leq \frac{1+\epsilon}{1-\epsilon} \quad \text{with probability} \geq 1-\delta.$$

For $\epsilon = 1/2 \rightsquigarrow$ reasonably well-conditioned Q_k.

Sketched Gram-Schmidt

Additional remarks:

- A proper analysis also needs to take roundoff error into account [Balabanov/Grigori'2022].
- Krylov subspaces + Gram-Schmidt = Arnoldi [CLA] $\rightsquigarrow$

  Krylov subspaces + randomized Gram-Schmidt
  = randomized Arnoldi.

  Basis of randomized iterative solvers for linear systems, eigenvalue problems, matrix functions. Development and analysis of such solver under active development.

  [Balabanov/Grigori'2022], [Burke/Güttel'2023], [Cortinovis/DK/Nakatsukasa'2024], [de Damas/Grigori'2024], [Güttel/Schweitzer'2024], [Nakatsukasa/Tropp'2024], [Palitta/Schweitzer/Simoncini'2023], [Timsit/Grigori/Balabanov'2023], and many more.

Kaczmarz

Disadvantage of BLENDENPIK: Need to assemble and apply whole matrix A (even when doing coordinate sampling).

Idea of (randomized) Kaczmarz: Merge coordinate sampling with simple iterative refinement.

Suppose one has an approximation x_{t-1} of the minimizer for $\|Ax - b\|$. To determine next iterate $x_t = x_{t-1} + c$, the optimal correction c solves

$$\min \|A(x_{t-1} + c) - b\|_2.$$

Sketching this correction equation $\rightsquigarrow \min \|SA(x_{t-1} + c) - Sb\|_2$. Kaczmarz takes an extreme choice for S. Sample *one* coordinate j:

$$\min \|e_{j(t)}^\top A(x_{t-1} + c) - e_{j(t)}^\top b\|_2 = \min \|\langle a_j, x_{t-1}\rangle + \langle a_j, c\rangle - b_j\|_2 = 0,$$

where $a_j^\top$ denotes jth row of A.

Many choices of c possible. The solution of smallest 2-norm is given by

$$c = a_j \frac{b_j - \langle a_j, x_{t-1}\rangle}{\|a_j\|_2^2}.$$

Randomized Kaczmarz

Randomized Kaczmarz chooses j randomly and independently in each iteration from discrete pdf $p_1, \ldots, p_d$. Canonical choices: Uniform sampling and Leverage scores / Importance sampling:

$$\mathbb{P}\{J(t) = i\} = p_i = \frac{\|a_i\|_2^2}{\|A\|_F^2}, \quad i = 1, \ldots, d.$$

Input: $A \in \mathbb{R}^{d \times m}$, $b \in \mathbb{R}^n$, initial iterate x_0, #iterations T.
Output: Approximation x_T of LSQ problem $\min \|Ax - b\|_2$.

Set $p_i = \|a_i\|_2^2 / \|A\|_F^2$, $i = 1, \ldots, d$
for $t = 1, 2, \ldots, T$ **do**
  Sample $j(t) \in \{1, \ldots, d\}$ according to pdf $(p_1, \ldots, p_d)$.
  $x_t = x_{t-1} - \frac{\langle a_{j(t)}, x_{t-1} \rangle - b_{j(t)}}{\|a_{j(t)}\|_2^2} \cdot a_{j(t)}$
end for

See also [Kireeva/Tropp'2024, arXiv:2402.17873] for a nice intro to Kaczmarz.

Analysis of randomized Kaczmarz

Simplifying assumption:[3] LSQ problem is consistent, that is, $b \in \mathrm{range}(A)$. A has full column rank $\rightsquigarrow \exists! x_*$ s.t. $Ax_* = b$.

Theorem [Strohmer/Vershynin'2009]. The iterates of randomized Kaczmarz satisfy

$$\mathbb{E}\|x_t - x_*\|_2^2 \le \big(1 - \kappa_{\mathrm{Demmel}}^{-2}\big)^t \cdot \|x_0 - x_*\|_2^2,$$

with $\kappa_{\mathrm{Demmel}} = \|A\|_F / \sigma_{\min}(A)$.

Proof. The expectation is to be understood with respect to the randomness in the choice of row indices in every step $1, \dots, t$. Let $J(1), \dots, J(t)$ denote corresponding r.v. Law of total expectation helps us to reduce the analysis to a single step:

$$\begin{aligned}
\mathbb{E}\|x_t - x_*\|_2^2 &= \mathbb{E}_{J(1),\dots,J(t)}\|x_t - x_*\|_2^2 \\
&= \mathbb{E}_{J(1),\dots,J(t-1)}\big[\mathbb{E}_{J(t)}\|x_t - x_*\|_2^2 \mid J(1), \dots, J(t-1)\big].
\end{aligned}$$

To simplify notation, we will simply write $\mathbb{E}_{J(t)}\|x_t - x_*\|_2^2$.

[3] Work by Zouzias/Freris'2013 lifts this assumption.

Analysis of randomized Kaczmarz

For fixed $j \equiv j(t)$, we can write

$$\begin{aligned} e_t &:= x_t - x_* = (x_{t-1} - x_*) - \frac{\langle a_j, x_{t-1} \rangle - b_j}{\|a_j\|_2^2} \cdot a_j \\ &= e_{t-1} - \frac{a_j a_j^\top}{\|a_j\|_2^2} \cdot e_{t-1} = (I - P_j) e_{t-1}, \end{aligned}$$

where we used $\langle a_j, x_* \rangle = b_j$ in the 2nd equality and set $P_j := a_j a_j^\top / \|a_j\|_2^2$. Using that P_j is an orthogonal projector, one obtains

$$\|e_t\|_2^2 = e_{t-1}^\top (I - P_j)(I - P_j) e_{t-1} = e_{t-1}^\top (I - P_j) e_{t-1}.$$

Now, consider random choice $J(t)$ for j. Then

$$\mathbb{E}_{J(t)} P_{J(t)} = \sum_{j=1}^{d} \mathbb{P}\{J(t) = j\} \cdot P_j = \sum_{j=1}^{d} \frac{\|a_j\|_2^2}{\|A\|_F^2} \cdot \frac{a_j a_j^\top}{\|a_j\|_2^2} = \frac{1}{\|A\|_F^2} A^\top A.$$

Analysis of randomized Kaczmarz

Hence,

$$\begin{aligned}
\mathbb{E}_{J(t)}\|e_t\|_2^2 &= e_{t-1}^\top(I-\mathbb{E}_{J(t)}P_j)e_{t-1} = e_{t-1}^\top(I-\|A\|_F^{-2}A^\top A)e_{t-1} \\
&\le \lambda_{\max}\left(I-\|A\|_F^{-2}A^\top A\right)\|e_{t-1}\|_2^2 = \left(1-\|A\|_F^{-2}\sigma_{\min}(A)^2\right)\|e_{t-1}\|_2^2 \\
&= \left(1-\kappa_{\text{Demmel}}^{-2}\right)\|e_{t-1}\|_2^2.
\end{aligned}$$

In summary, we obtain

$$\begin{aligned}
\mathbb{E}\|x_t-x_*\|_2^2 &\le \left(1-\kappa_{\text{Demmel}}^{-2}\right)\mathbb{E}\|x_{t-1}-x_*\|_2^2 \\
&\le \left(1-\kappa_{\text{Demmel}}^{-2}\right)^2\mathbb{E}\|x_{t-2}-x_*\|_2^2 \\
&\vdots \\
&\le \left(1-\kappa_{\text{Demmel}}^{-2}\right)^t\mathbb{E}\|x_0-x_*\|_2^2,
\end{aligned}$$

which concludes the proof. $\diamond$

EFY: Using the Borel-Cantelli lemma, conclude from the theorem that Kaczmarz converges almost surely with a rate arbitrarily close to $1-\kappa_{\text{Demmel}}^{-2}$.

Kaczmarz is SGD

Stochastic gradient descent (SGD) applies to differentiable objective function of the form

$$\varphi(x) = \varphi_1(x) + \varphi_2(x) + \cdots + \varphi_d(x).$$

Each step of SGD updates

$$x_t = x_{t-1} - \eta \nabla \varphi_j(x_{t-1})$$

with randomly chosen index j and learning rate $\eta > 0$.

For $\varphi(x) := \|Ax - b\|_2^2$, we have the decomposition

$$\|Ax - b\|_2^2 = (\langle a_1, x\rangle - b_1)^2 + (\langle a_2, x\rangle - b_2)^2 + \cdots + (\langle a_d, x\rangle - b_d)^2.$$

Because of $\nabla(\langle a_j, x\rangle - b_j)^2 = 2(\langle a_j, x\rangle - b_j)a_j$, one step of SGD becomes

$$x_t = x_{t-1} - 2\eta(\langle a_j, x_{t-1}\rangle - b_j)a_j.$$

With adaptive learning rate $\eta = 1/(2\|a_j\|_2^2)$, this is Kaczmarz!