Randomised algorithms in NLA

So far, all algorithms have been deterministic (always same output)

- Direct methods (LU for $Ax = b$, QRalg for $Ax = \lambda x$ or $A = U\Sigma V^T$):
  - Incredibly reliable, backward stable
  - Works like magic if $n \lesssim 10000$
  - But not beyond; cubic complexity $O(n^3)$ or $O(mn^2)$
- Iterative methods (GMRES, CG, Arnoldi, Lanczos)
  - Very fast when it works (nice spectrum etc)
  - Otherwise, not so much; need for preconditioning

Randomised algorithms in NLA

So far, all algorithms have been deterministic (always same output)

- Direct methods (LU for $Ax = b$, QRalg for $Ax = \lambda x$ or $A = U\Sigma V^T$):
  - Incredibly reliable, backward stable
  - Works like magic if $n \lesssim 10000$
  - But not beyond; cubic complexity $O(n^3)$ or $O(mn^2)$
- Iterative methods (GMRES, CG, Arnoldi, Lanczos)
  - Very fast when it works (nice spectrum etc)
  - Otherwise, not so much; need for preconditioning
- Randomised algorithms
  - Output differs at every run
  - Ideally succeed with enormous probability, e.g. $1 - \exp(-cn)$
  - Often by far the fastest&only feasible approach
  - Not for all problems—active field of research

We'll cover two NLA topics where randomisation very successful: **low-rank approximation (randomised SVD)**, and overdetermined **least-squares problems**

SVD: the most important matrix decomposition

- **Symmetric eigenvalue decomposition**: $A = V\Lambda V^T$
  for symmetric $A \in \mathbb{R}^{n\times n}$, where $V^T V = I_n$, $\Lambda = \mathrm{diag}(\lambda_1, \dots, \lambda_n)$.
- **Singular Value Decomposition (SVD)**: $A = U\Sigma V^T$
  for any $A \in \mathbb{R}^{m\times n}$, $m \geq n$. Here $U^T U = V^T V = I_n$, $\Sigma = \mathrm{diag}(\sigma_1, \dots, \sigma_n)$,
  $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n \geq 0$.

SVD proof: Take Gram matrix $A^T A$ and its eigendecomposition $A^T A = V\Lambda V^T$. Λ is nonnegative, and $(AV)^T(AV)$ is diagonal, so $AV = U\Sigma$ for some orthonormal U. Right-multiply V^T.

SVD useful for

- Finding column space, row space, null space, rank, ...
- Matrix analysis, polar decomposition, ...
- **Low-rank approximation**

(Most) important result in Numerical Linear Algebra

Given $A \in \mathbb{R}^{m\times n}$ $(m \geq n)$, find low-rank (rank r) approximation

$$A \approx \hat{U} \hat{\Sigma} \hat{V}^T, \qquad \hat{\Sigma} \in \mathbb{R}^{r\times r}$$

- Optimal solution $A_r = U_r \Sigma_r V_r^T$ via truncated SVD
  $U_r = U(:, 1:r), \Sigma_r = \Sigma(1:r, 1:r), V_r = V(:, 1:r)$, giving

  $$\|A - A_r\| = \|\text{diag}(\sigma_{r+1}, \dots, \sigma_n)\|$$

  in any unitarily invariant norm [Horn-Johnson 1985]
- But that costs $O(mn^2)$ (bidiagonalisation+QR); look for cheaper approximation

Randomised SVD by HMT

[Halko-Martinsson-Tropp, SIAM Review 2011]

1. Form a random matrix $X \in \mathbb{R}^{n\times r}$, usually $r \ll n$.
2. Compute AX.
3. QR factorisation $AX = QR$.
4. $A \approx Q \; Q^T A$ $(= (QU_0)\Sigma_0 V_0^T)$ is rank-r approximation.

- $O(mnr)$ cost for dense A
- Near-optimal approximation guarantee: for any $\hat{r} < r$,

$$\mathbb{E}\|A - \hat{A}\|_F \leq \left(1 + \frac{r}{r - \hat{r} - 1}\right) \|A - A_{\hat{r}}\|_F$$

where $A_{\hat{r}}$ is the rank $\hat{r}$-truncated SVD (expectation w.r.t. random matrix X)

Goal: understand this, or at least why $\mathbb{E}\|A - \hat{A}\| = O(1)\|A - A_{\hat{r}}\|$

Pseudoinverse and projectors

Given $M \in \mathbb{R}^{m\times n}$ with economical SVD $M = U_r \Sigma_r V_r^T$ ($U_r \in \mathbb{R}^{m\times r}, \Sigma_r \in \mathbb{R}^{r\times r}, V_r \in \mathbb{R}^{n\times r}$ where $r = \mathrm{rank}(M)$ so that $\Sigma_r \succ 0$), the **pseudoinverse** $M^\dagger$ is

$$M^\dagger = V_r \Sigma_r^{-1} U_r^T \in \mathbb{R}^{n\times m}$$

- satisfies $MM^\dagger M = M$, $M^\dagger M M^\dagger = M^\dagger$, $AA^\dagger = (AA^\dagger)^T$, $A^\dagger A = (A^\dagger A)^T$ (which are often taken to be the definition—above is much simpler IMO)
- $M^\dagger = M^{-1}$ if M nonsingular

A square matrix $P \in \mathbb{R}^{n\times n}$ is called a **projector** if $P^2 = P$

- P diagonalisable and all eigenvalues 1 or 0
- $\|P\|_2 \geq 1$ and $\|P\|_2 = 1$ iff $P = P^T$; in this case P is called orthogonal projector
- $I - P$ is another projector, and unless $P = 0$ or $P = I$, $\|I - P\|_2 = \|P\|_2$: Schur form $QPQ^* = \begin{bmatrix} I & B \\ 0 & 0 \end{bmatrix}$, $Q(I-P)Q^* = \begin{bmatrix} 0 & -B \\ 0 & I \end{bmatrix}$; see [Szyld 2006]

HMT approximant: analysis (down from 70 pages!)

$\hat{A} = QQ^TA$, where $AX = QR$. Goal: $\|A - \hat{A}\| = \|(I_m - QQ^T)A\| = O(\|A - A_{\hat{r}}\|)$.

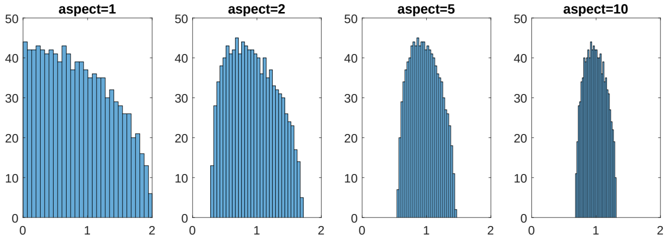
1. $QQ^TAX = AX$ (QQ^T is **orthogonal projector** onto span(AX)). Hence $(I_m - QQ^T)AX = 0$, so $A - \hat{A} = (I_m - QQ^T)A(I_n - XM^T)$ for any $M \in \mathbb{R}^{n\times r}$.
2. Set $M^T = (V^TX)^\dagger V^T$ where $V = [v_1, \ldots, v_{\hat{r}}] \in \mathbb{R}^{n\times\hat{r}}$ top sing vecs of A ($\hat{r} \le r$).
3. $VV^T(I - XM^T) = VV^T(I - X(V^TX)^\dagger V^T) = 0$ if V^TX full row-rank (generic assumption), so $A - \hat{A} = (I_m - QQ^T)A(I - VV^T)(I_n - XM^T)$.
4. Taking norms, $\|A - \hat{A}\|_2 = \|(I_m - QQ^T)A(I - VV^T)(I_n - XM^T)\|_2 = \|(I_m - QQ^T)U_2\Sigma_2V_2^T(I_n - XM^T)\|_2$ where $[V, V_2]$ is orthogonal, so

$$\|A - \hat{A}\|_2 \le \|\Sigma_2\|_2\|(I_n - XM^T)\|_2 = \underbrace{\|\Sigma_2\|_2}_{\text{optimal rank-}\hat{r}} \quad \|XM^T\|_2$$

To see why $\|XM^T\|_2 = O(1)$ (with high probability), we need random matrix theory

Tool from RMT: Rectangular random matrices are well conditioned

Singvals of random matrix $X \in \mathbb{R}^{m\times n}$ $(m \geq n)$ with iid X_{ij} (mean 0, variance 1) follow Marchenko-Pastur (M-P) distribution (proof nonexaminable)



density $\sim \frac{1}{x}\sqrt{((1+\sqrt{\frac{m}{n}})-x)(x-(1-\sqrt{\frac{m}{n}}))}$, support $[\sqrt{m}-\sqrt{n}, \sqrt{m}+\sqrt{n}]$

$\sigma_{\max}(X) \approx \sqrt{m}+\sqrt{n}$, $\sigma_{\min}(X) \approx \sqrt{m}-\sqrt{n}$, hence $\kappa_2(X) \approx \frac{1+\sqrt{m/n}}{1-\sqrt{m/n}} = O(1)$,

Key fact in many breakthroughs in computational maths!

- Randomised SVD, Blendenpik (randomised least-squares)
- (nonexaminable:) Compressed sensing (RIP) [Donoho 06, Candes-Tao 06], Matrix concentration inequalities [Tropp 11], Function approx. by least-squares [Cohen-Davenport-Leviatan 13]

$\|XM^T\|_2 = O(1)$

Recall we've shown for $M^T = (V^T X)^\dagger V^T \; X \in \mathbb{R}^{n\times r}$

$$\|A - \hat{A}\|_2 \le \|\Sigma_2\|_2 \|(I_n - XM^T)\|_2 = \underbrace{\|\Sigma_2\|_2}_{\text{optimal rank-}\hat{r}} \quad \|XM^T\|_2$$

Now $\|XM^T\|_2 = \|X(V^T X)^\dagger V^T\|_2 = \|X(V^T X)^\dagger\|_2 \le \|X\|_2 \|(V^T X)^\dagger\|_2$.
Assume X is random Gaussian $X_{ij} \sim \mathcal{N}(0,1)$. Then

- $V^T X$ is a Gaussian matrix (orthogonal$\times$Gaussian$=$Gaussian; exercise), hence $\|(V^T X)^\dagger\| = 1/\sigma_{\min}(V^T X) \lesssim 1/(\sqrt{r} - \sqrt{\hat{r}})$ by M-P
- $\|X\|_2 \lesssim \sqrt{m} + \sqrt{r}$ by M-P

Together we get $\|XM^T\|_2 \lesssim \frac{\sqrt{m}+\sqrt{r}}{\sqrt{r}-\sqrt{\hat{r}}} =$ "$O(1)$"

- When X non-Gaussian random matrix, perform similarly, harder to analyze

Precise analysis for HMT (nonexaminable)

Theorem (Reproduces HMT 2011 Thm.10.5)

If X Gaussian, for any $\hat{r} < r$, $\mathbb{E}\|E_{\mathrm{HMT}}\|_F \leq \sqrt{\mathbb{E}\|E_{\mathrm{HMT}}\|_F^2} = \sqrt{1 + \frac{r}{r-\hat{r}-1}}\|A - A_{\hat{r}}\|_F.$

PROOF. First ineq: Cauchy-Schwarz. $\|E_{\mathrm{HMT}}\|_F^2$ is

$$\begin{aligned}&\|A(I - VV^T)(I - \mathcal{P}_{X,V})\|_F^2 = \|A(I - VV^T)\|_F^2 + \|A(I - VV^T)\mathcal{P}_{X,V}\|_F^2\\ &= \|\Sigma_2\|_F^2 + \|\Sigma_2\mathcal{P}_{X,V}\|_F^2 = \|\Sigma_2\|_F^2 + \|\Sigma_2(V_\perp^T X)(V^T X)^\dagger V^T\|_F^2.\end{aligned}$$

Now if X is Gaussian then $V_\perp^T X \in \mathbb{R}^{(n-\hat{r})\times r}$ and $V^T X \in \mathbb{R}^{\hat{r}\times r}$ are independent Gaussian. Hence by [HMT Prop. 10.1] $\mathbb{E}\|\Sigma_2(V_\perp^T X)(V^T X)^\dagger\|_F^2 = \frac{r}{r-\hat{r}-1}\|\Sigma_2\|_F^2$, so

$$\mathbb{E}\|E_{\mathrm{HMT}}\|_F^2 = \left(1 + \frac{r}{r-\hat{r}-1}\right)\|\Sigma_2\|_F^2.$$

Generalized Nyström

$X \in \mathbb{R}^{n\times r}$ as before; set $Y \in \mathbb{R}^{n\times(r+\ell)}$, and [N. arXiv 2020]

$$\hat{A} = (AX(Y^TAX)^\dagger Y^T)A = \mathcal{P}_{AX,Y}A$$

Then $A - \hat{A} = (I - \mathcal{P}_{AX,Y})A = (I - \mathcal{P}_{AX,Y})A(I - XM^T)$; choose M s.t. $XM^T = X(V^TX)^\dagger V^T = \mathcal{P}_{X,V}$. Then $\mathcal{P}_{AX,Y}, \mathcal{P}_{X,V}$ projections, and

$$\begin{aligned}\|A - \hat{A}\| &= \|(I - \mathcal{P}_{AX,Y})A(I - \mathcal{P}_{X,V})\| \\ &\le \|(I - \mathcal{P}_{AX,Y})A(I - VV^T)(I - \mathcal{P}_{X,V})\| \\ &\le \|A(I - VV^T)(I - \mathcal{P}_{X,V})\| + \|\mathcal{P}_{AX,Y}A(I - VV^T)(I - \mathcal{P}_{X,V})\|.\end{aligned}$$

- Note $\|A(I - VV^T)(I - \mathcal{P}_{X,V})\|$ exact same as HMT error
- Extra term $\|\mathcal{P}_{AX,Y}\|_2 = O(1)$ as before if $c > 1$ in $Y \in \mathbb{R}^{m\times cr}$
- Overall, about $(1 + \|\mathcal{P}_{AX,Y}\|_2) \approx (1 + \frac{\sqrt{n}+\sqrt{r+\ell}}{\sqrt{r+\ell}-\sqrt{r}})$ times bigger expected error than HMT, **still near-optimal** and **much faster** $O(mn\log n + r^3)$