Randomized Numerical Linear Algebra

Review: classical numerical linear algebra

$*$: conjugate transpose

$\mathbb{H}_n = \mathbb{H}_n(\mathbb{F}_n) = \{A \in \mathbb{F}^{n\times n} : A^* = A\}$ $(\mathbb{F} = \mathbb{R}; \mathbb{C})$

Moore-Penrose pseudoinverse: A^+

$A = U\Sigma_r V^*$ $\quad \Sigma_r = \begin{pmatrix} \sigma_1 & & & 0 \\ & \ddots & & \\ & & \sigma_r & \\ 0 & & & 0 \end{pmatrix}$ $\quad r = \text{rank } A$

$\sigma_i = \sqrt{\lambda_i(A^*A)}$

$A^+ = V\begin{pmatrix} \sigma_1^{-1} & & & 0 \\ & \ddots & & \\ & & \sigma_r^{-1} & \\ 0 & & & 0 \end{pmatrix} U^*$ (see the eigenvalues/singular values below)

Properties of A^+:
- $AA^+A = A$, $(AA^+)^* = AA^+ = (AA^+)^2$

  $A^+AA^+ = A^+$, $(A^+A)^* = A^+A = (A^+A)^2$
- If A has full column rank, then $A^+ = (A^*A)^{-1}A^*$
- If A attains an inverse, then $A^+ = A^{-1}$

PSD (positive semidefinite) $= \{A \in \mathbb{H}_n : x^TAx \geqslant 0$ for $x \neq 0\}$

PD (positive definite) $= \{A \in \mathbb{H}_n : x^TAx > 0$ for $x \neq 0\}$

$\preccurlyeq$: semidefinite order : $(A \preccurlyeq B \Leftrightarrow 0 \preccurlyeq B - A \Leftrightarrow B - A \in \text{PSD})$

$\prec$: definite order : $(A \prec B \Leftrightarrow 0 \prec B - A \Leftrightarrow B - A \in \text{PD})$

eigenvalues of $A \in \mathbb{H}_n$, $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_n$, $A = \sum_{i=1}^{n} \lambda_i u_i u_i^*$

$\mathbb{R}$

singular values of $A \in \mathbb{F}^{m\times n}$, $\sigma_1 \geqslant \sigma_2 \geqslant \cdots \geqslant \sigma_r \geqslant 0$, $A = \sum_{i=1}^{r} \sigma_i u_i v_i^*$

functional calculus $f: \mathbb{H}_n \to \mathbb{H}_n$, $f(A) := \sum_{i=1}^{n} f(\lambda_i) u_i u_i^*$

$f: \mathbb{R} \to \mathbb{R} \rightsquigarrow$

$A = \sum_{i=1}^{n} \lambda_i u_i u_i^*$

$\langle a, b\rangle = \sum_{i=1}^{n} (a)_i^* (b)_i$, $\quad \|a\|^2 := \langle a, a\rangle$. $\quad S^{n-1} = S^{n-1}(\mathbb{F}) = \{a \in \mathbb{F}^n : \langle a, a\rangle = 1\}$

$\mathrm{Tr}(A) = \sum_{i=1}^{n} (A)_{ii}$

Equip $\mathbb{F}^{m\times n}$ with the standard trace inner product and Frobenius norm:

$$A, B \in \mathbb{F}^{m\times n}, \quad \langle A, B\rangle := \mathrm{Tr}(A^* B)$$

$$\|A\|_F^2 := \langle A, A\rangle$$

matrix norms:

- operator norms induced by $\|\cdot\|_\alpha$ $\mathbb{F}^n$, $\|\cdot\|_\beta$ $\mathbb{F}^m$

$$\|\cdot\|_{\alpha,\beta} : \mathbb{F}^{m\times n} \to \mathbb{R}, \quad A \mapsto \sup_{\substack{x \in \mathbb{F}^n \\ \|x\|_\alpha \neq 0}} \frac{\|Ax\|_\beta}{\|x\|_\alpha}$$

- alternatively, we may define any function $\|\cdot\| : \mathbb{F}^{m\times n} \to \mathbb{R}$ that fulfills the three axioms $\begin{cases} \text{positivity} \\ \text{homogeneity} \\ \text{triangle inequality} \end{cases}$

- $\|A\|_{\ell^2} =: \|A\|$ (operator norm induced by ℓ^2) (spectral norm) $= \sigma_1 = \sqrt{\lambda_{max}(A^*A)}$

- $\|A\|_* = \sum_{k=1}^{\min(m,n)} \sigma_k$ (nuclear/trace/trace-1 norm)

- $\|A\|_F^2 = \mathrm{Tr}(A^*A) = \sum_{k=1}^{\min(m,n)} \sigma_k^2 = \sum_{i=1}^{m}\sum_{j=1}^{n} |(A)_{ij}|^2$

- $\|A\|_p = \left(\sum_{k=1}^{\min(m,n)} \sigma_k^p\right)^{1/p}$ (trace-p norm or Schatten-p norm)

- (less used) Ky-Fan p-norm $(p \le \min(m,n))$

$$\|A\|_{K,p} = \sum_{k=1}^{p} \sigma_k \quad \begin{pmatrix} \|\cdot\|_* = \|\cdot\|_1 = \|\cdot\|_{K,\min(m,n)} \\ \|\cdot\|_F = \|\cdot\|_2 \\ \|\cdot\| = \|\cdot\|_{K,1} = \|\cdot\|_\infty \end{pmatrix}$$

Intrinsic dimension $A \in PSD$, $\operatorname{intdim}(A) := \frac{\operatorname{Tr} A}{\|A\|} = \frac{\sigma_1 + \cdots + \sigma_r}{\sigma_1}$.

- If $\sigma_1 = \cdots = \sigma_r$, then $\operatorname{intdim}(A) = r$
- If $\sigma_1 > 0$, $\sigma_2, \ldots, \sigma_r \ll \sigma_1$, then $\operatorname{intdim}(A) \approx 1$ $(A \approx \sigma_1 \underbrace{u_1 u_1^*}_{\text{rank-}1})$
- a "continuous" measure of rank.
- In general, $1 \leq \operatorname{intdim}(A) \leq r = \operatorname{rank}(A)$
- The upper-bound is saturated if A is an orthogonal projector $(A = A^2 = (A^*))$

Stable rank $B \in \mathbb{F}^{m \times n}$. $\operatorname{srank} B := \operatorname{intdim}(\underset{PSD}{B^* B}) = \frac{\operatorname{Tr} B^* B}{\|B^* B\|}$

$$= \frac{\sigma_1^2 + \cdots + \sigma_r^2}{\sigma_1^2} = \frac{\|B\|_F^2}{\|B\|^2}$$

Schur complement formula $M = \begin{pmatrix} A & B \\ C & D \end{pmatrix} \in \mathbb{F}^{m \times m}$, $A \in \mathbb{F}^{n \times n}$ $(m > n)$

- Schur complement of D in M: $M/D = A - BD^{-1}C$ (If D invertible)

Then we have $M = \begin{pmatrix} I & BD^{-1} \\ 0 & I \end{pmatrix} \begin{pmatrix} M/D & 0 \\ C & D \end{pmatrix}$. M invertible $\Leftrightarrow (M/D)$ invertible

$$M^{-1} = \begin{pmatrix} (M/D)^{-1} & -(M/D)^{-1} B D^{-1} \\ -D^{-1}C(M/D)^{-1} & D^{-1} + D^{-1}C(M/D)^{-1}BD^{-1} \end{pmatrix}$$

- Schur complement of A in M: $M/A = D - CA^{-1}B$. (If A invertible)

Then we have $M = \begin{pmatrix} I & 0 \\ CA^{-1} & I \end{pmatrix} \begin{pmatrix} A & B \\ 0 & M/A \end{pmatrix}$. M invertible $\Leftrightarrow (M/A)$ invertible

$$M^{-1} = \begin{pmatrix} A^{-1} + A^{-1}B(M/A)^{-1}CA^{-1} & -A^{-1}B(M/A)^{-1} \\ -(M/A)^{-1}CA^{-1} & (M/A)^{-1} \end{pmatrix}$$

- If D or A are singular or not square:

$M \in \mathbb{F}^{m\times n}$. $\underline{\alpha} = (\alpha_1, \dots \alpha_k) \subseteq [m]$, $\underline{\alpha}^c = [m] \setminus \underline{\alpha}$

$\underline{\beta} = (\beta_1, \dots \beta_l) \subseteq [n]$, $\underline{\beta}^c = [~~m~~n] \setminus \underline{\beta}$

The Schur complement of $M[\underline{\alpha}, \underline{\beta}]$ in M

$$M / M[\underline{\alpha}, \underline{\beta}] = M[\underline{\alpha}^c, \underline{\beta}^c] - M[\underline{\alpha}^c \underline{\beta}] (M[\underline{\alpha}, \underline{\beta}])^{+} M[\underline{\alpha}, \underline{\beta}^c]$$

<u>low-rank approximation in $\|\cdot\|$</u>

$A \in \mathbb{F}^{m\times n}$, $\hat{A} \in \mathbb{F}^{m\times n}$ such that $\|A - \hat{A}\| \leq \varepsilon$

(e.g. $\hat{A} = \sum_{i: \sigma_i \geq \varepsilon} \sigma_i u_i v_i^*$)

then we have

- $\forall F \in \mathbb{F}^{m\times n}$, $|\langle F, A\rangle - \langle F, \hat{A}\rangle| \leq \|F\|_* \varepsilon$

  $\hookrightarrow$ nuclear norm
- $|\sigma_j(A) - \sigma_j(\hat{A})| \leq \varepsilon \ (\forall j)$

<u>Rmk</u>, read: Frobenius-norm approximation (Martinsson & Tropp, Page 8)

<u>Preliminaries in probability theories</u> (independent and identically distributed $=$ i.i.d.)

- linearity of expectation $\mathbb{E}(\underset{\text{deterministic}}{A} X) = A\, \mathbb{E}(\underset{\text{random}}{X})$
- isotropic random vector $\mathbb{E}(xx^*) = I$
- centered random vector $\mathbb{E}(x) = 0$

  } isotropic + centered = standardized
- Scalar Rademacher $X \sim \text{unif}\{\pm 1\}$
- $\mu \in \mathbb{F}^n$, $X \sim$ normal(μ, C), $C \in \mathbb{H}_n(\mathbb{F})$ ($C \in$ PSD)

  C: covariance matrix

$$C = \mathbb{E}((X-\mu)(X-\mu)^*)$$
$$= \mathbb{E}(XX^T) - \mu\mu^*$$

- concentration inequalities describe the closeness of a random matrix to its expectation (bound on the probability)

- Scalar case

Markov X nonnegative R.V., $a>0$, then $\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}(X)}{a}$

Chebyshev X R.V., $\sigma^2 = Var(X) < \infty$, $\mu = \mathbb{E}(X)$, then $\mathbb{P}(|X-\mu| \geq k\sigma) \leq \frac{1}{k^2}$ $(\forall k>0)$

or $\mathbb{P}(|X-\mu| \geq k) \leq \frac{\sigma^2}{k^2}$

E.g. for $\{X_i\}_{i=1}^k$ i.i.d.,

$$\mathbb{P}\left(\left|\frac{\sum_{i=1}^k X_i}{k} - \mu\right| \geq t\right) \underset{\text{chebyshev}}{\leq} \frac{\sigma^2}{kt^2} \quad \left(\begin{aligned} &\mathbb{E}\Big(\big(\textstyle\sum_{i=1}^k X_i\big)/k\Big) = \mu k/k = \mu \\ &Var\Big(\textstyle\sum_{i=1}^k X_i/k\Big) = \frac{1}{k^2}(k\sigma^2) = \frac{\sigma^2}{k} \end{aligned}\right)$$

(want to estimate μ by sampling $X_i \overset{i.i.d.}{\sim}$ some distribution with expectation μ & variance σ^2
how many samples do we need to achieve the accuracy ε with failure probability δ? $\delta = \frac{\sigma^2}{kt^2} \Longleftrightarrow k = \frac{\sigma^2}{\delta\varepsilon^2}$ $(\varepsilon = t)$)

Chernoff $\mathbb{P}(X \geq a) = \mathbb{P}(e^{\lambda X} \geq e^{\lambda a}) \leq \frac{\mathbb{E}(e^{\lambda X})}{e^{\lambda a}}$ $\rightsquigarrow$ MGF (moment generating function)

E.g. (Hoeffding's method)
for $\{X_i\}_{i=1}^k$ i.i.d., $a \leq X_i \leq b$ a.s. (bounded R.V.'s), then

$$\mathbb{P}\left(\left|\frac{\sum_{i=1}^k X_i}{k} - \mu\right| \geq t\right) \leq 2\exp\left(-\frac{2kt^2}{(b-a)^2}\right)$$

$$\mathbb{P}\left(\frac{\sum_{i=1}^k X_i}{k} - \mu \geq t\right) \leq \exp\left(-\frac{2kt^2}{(b-a)^2}\right)$$

(This is obtained by the MGF bound

$$\mathbb{E}\left(e^{\lambda(X_i-\mu)}\right) \leq \exp\left(\frac{\lambda^2(b-a)^2}{8}\right)$$

using the convexity of $t \mapsto \exp(t)$)

E.g. (Bernstein's method) $\{X_i\}_{i=1}^k$ i.i.d. s.t. $|X_i - \mathbb{E}(X_i)| \le B$ a.s.

then $\mathbb{P}\left(\left|\frac{\sum_{i=1}^k X_i}{k} - \mu\right| \geq t\right) \le 2\exp\left(-\frac{\frac{1}{2}kt^2}{\sigma^2 + \frac{1}{3}Bt}\right) \underset{\text{large } t}{\lesssim} 2\exp\left(-\frac{3kt}{2B}\right)$

(For small t, Bernstein tighter than Hoeffding

For large t, Hoeffding tighter than Bernstein)

(Proof. exer. Hint: $\forall |x| \le M, |\lambda| \le \frac{1}{M}$, $e^{\lambda x} \le 1 + \lambda x + \frac{\lambda^2 x}{2(1-\lambda M/3)}$)

More concentrated inequalities .... Applications in e.g. trace estimator

- matrix version of concentration inequalities (Tropp 2015)

- Gaussian random matrix

$G \in \mathbb{R}^{m \times n}$, $G_{ij} \overset{\text{i.i.d.}}{\sim} \text{normal}(0,1)$ (i.i.d. Gaussian or "Ginibre")

$\|G\| = \sup_{\|x\|=\|y\|=1} \langle y, Gx \rangle =: \sup_{\|x\|=\|y\|=1} X_{(x,y)}$

$\mathbb{E} X_{(x,y)} X_{(x',y')} = \langle x, x' \rangle \langle y, y' \rangle$

- Slepian inequality X_t, Y_t two centered Gaussian processes, if

$\mathbb{E}(X_s - X_t)^2 \le \mathbb{E}(Y_s - Y_t)^2, \forall s,t$, then

$\mathbb{E}\sup_t X_t \le \mathbb{E}\sup_t Y_t$

(Vershynin 2018) "High-dimensional probability")

- Chevet inequality $\mathbb{E}\sup_{\substack{x\in T\\ y \in S}} \underbrace{\langle y, Gx \rangle}_{= X_{(x,y)}} \lesssim \text{width}(T) + \text{width}(S)$ $\left(\text{width}(T) := \mathbb{E}\sup_{x\in T} \langle g, x \rangle\right)$

Taking $T = S =$ unit ball, $\mathbb{E}\|G\| \lesssim \sqrt{n} + \sqrt{m}$

- Gordon inequality $\mathbb{E}\inf_{x \in T}\sup_{y\in S} X_{(x,y)} \geq \mathbb{E}\|g\| \inf_{x\in T}\|x\| - \mathbb{E}\|h\| \sup_{y\in S}\|y\|$

Taking $T = S =$ unit ball, $\sigma_{\min}(G) \gtrsim \sqrt{m} - \sqrt{n}$ w.h.p.

$(\Longrightarrow \mathbb{E}\,\sigma_{\min}(G) \gtrsim \sqrt{m} - \sqrt{n})$