Matrix Norms

- Vector norms are all measures of how "big" the vectors are. Similarly, we want to have measures for how "big" matrices are.
- the definition of a norm of a matrix is the same as that of a vector:
- $f : \mathbb{R}^{m\times n} \to \mathbb{R}$ is a norm if

  (i) $f(\mathbf{A}) \geq 0$ for all $\mathbf{A}$ (positive definiteness)

  (ii) $f(\mathbf{A}) = 0$ if and only if $\mathbf{A} = \mathbf{0}$ (positive definiteness)

  (iv) $f(\alpha\mathbf{A}) = |\alpha| f(\mathbf{A})$ for any $\alpha, \mathbf{A}$ (homogeneity)

  (iii) $f(\mathbf{A}+\mathbf{B}) \leq f(\mathbf{A})+f(\mathbf{B})$ for any $\mathbf{A}, \mathbf{B}$ (sub-additive or triangle inequality)
- we usually use the notation $\|\cdot\|$ to denote a matrix norm
- Since $\mathbf{I}^2 = \mathbf{I}$, from $\|\mathbf{I}\| = \|\mathbf{I}^2\| \leq \|\mathbf{I}\|^2$, we get $\|\mathbf{I}\| \geq 1$, for every matrix norm.

"Elementwise" Norms

- "elementwise" norm: treat $\mathbf{A}$ as a $m \times n$ vector
- in general, for $p,\ q \geq 1$ it is given by

$$f(\mathbf{A}) = \left(\sum_{j=1}^{n} \left(\sum_{i=1}^{m} |a_{ij}|^p \right)^{\frac{q}{p}} \right)^{\frac{1}{q}}$$

- for $p = q = 2$, we have the Frobenius norm $\|\mathbf{A}\|_F = \sqrt{\sum_{i,j} |a_{ij}|^2} = [\mathrm{tr}(\mathbf{A}^T\mathbf{A})]^{\frac{1}{2}}$
  - note Frobenius norm has the orthogonal invariance property, then $\|\mathbf{A}\|_F = \|\mathbf{U}^T\mathbf{A}\mathbf{V}\|_F = \|\mathbf{\Sigma}\|_F = \sqrt{\sigma_1^2 + \ldots + \sigma_r^2}$
- for $p = q = 1$, we have the sum-absolute-value norm $\|\mathbf{A}\|_1 = \sum_{i,j} |a_{ij}|$
- for $p = q = \infty$, we have the max-absolute-value norm $\|\mathbf{A}\|_\infty = \max_{i,j} |a_{ij}|$

Induced Norms

- induced norm or operator norm: for $\mathbf{A} \in \mathbb{R}^{m\times n}$, the function

$$f(\mathbf{A}) = \max_{\mathbf{x}\in\mathbb{R}^n, \mathbf{x}\neq\mathbf{0}} \frac{\|\mathbf{A}\mathbf{x}\|_p}{\|\mathbf{x}\|_q} = \max_{\|\mathbf{x}\|_q \leq 1} \|\mathbf{A}\mathbf{x}\|_p$$

  where $\|\cdot\|_p$, $\|\cdot\|_q$ denote any vector norms; sometimes written as $\|\mathbf{A}\|_{p,q}$
- The matrix norms induced by vector norms measure how much the mapping induced by $\mathbf{A}$ can "stretch" (the "length" of) a vector.
- induced p-norm: matrix norms induced by the vector p-norm ($p \geq 1$)

$$\|\mathbf{A}\|_p = \max_{\|\mathbf{x}\|_p \leq 1} \|\mathbf{A}\mathbf{x}\|_p$$

- it is known that
  - $\|\mathbf{A}\|_1 = \max_{1\leq j\leq n} \sum_{i=1}^m |a_{ij}|$, i.e., the maximum column sum.
  - $\|\mathbf{A}\|_\infty = \max_{1\leq i\leq m} \sum_{j=1}^n |a_{ij}|$, i.e., the maximum row sum. (proof as a Quiz)
- how about $p = 2$?

Induced Norms

- matrix 2-norm or spectral norm:

$$\|\mathbf{A}\|_2 = \sigma_{\max}(\mathbf{A}).$$

- proof:
  - for any $\mathbf{x}$ with $\|\mathbf{x}\|_2 \leq 1$,

$$\begin{aligned}\|\mathbf{A}\mathbf{x}\|_2 = \|\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T\mathbf{x}\|_2 = \|\mathbf{\Sigma}\mathbf{V}^T\mathbf{x}\|_2 \\ \leq \sigma_1\|\mathbf{V}^T\mathbf{x}\|_2 = \sigma_1\|\mathbf{x}\|_2 \leq \sigma_1\end{aligned}$$

  - $\|\mathbf{A}\mathbf{x}\|_2 = \sigma_1$ if we choose $\mathbf{x} = \mathbf{v}_1$

- implication to linear systems: let $\mathbf{y} = \mathbf{A}\mathbf{x}$ be a linear system. Under the input energy constraint $\|\mathbf{x}\|_2 \leq 1$, the system output energy $\|\mathbf{y}\|_2^2$ is maximized when $\mathbf{x}$ is chosen as the 1st right singular vector

- corollary: $\min_{\|\mathbf{x}\|_2=1} \|\mathbf{A}\mathbf{x}\|_2 = \sigma_{\min}(\mathbf{A})$ if $m \geq n$

- corollary: if $\mathbf{A}$ is invertible, $\|\mathbf{A}^{-1}\|_2 = \sigma_{\min}^{-1}(\mathbf{A})$

Induced Norms

Properties for the matrix induced norms:

- $\|\mathbf{Ax}\|_p \leq \|\mathbf{A}\|_{p,q}\|\mathbf{x}\|_q$; $\|\mathbf{Ax}\|_p \leq \|\mathbf{A}\|_p\|\mathbf{x}\|_p$
- $\|\mathbf{AB}\|_{p,q} \leq \|\mathbf{A}\|_{p,q}\|\mathbf{B}\|_q$; $\|\mathbf{AB}\|_p \leq \|\mathbf{A}\|_p\|\mathbf{B}\|_p$ (submultiplicative or consistent)
  - we also have $\|\mathbf{AB}\|_F \leq \|\mathbf{A}\|_F\|\mathbf{B}\|_F$ (proof by yourself)
- $\|\mathbf{QAW}\|_2 = \|\mathbf{A}\|_2$ for any orthogonal $\mathbf{Q}, \mathbf{W}$
  - we also have $\|\mathbf{QAW}\|_F = \|\mathbf{A}\|_F$ for any orthogonal $\mathbf{Q}, \mathbf{W}$
- $\|\mathbf{A}\|_2 \leq \|\mathbf{A}\|_F \leq \sqrt{p}\|\mathbf{A}\|_2$ (here $p = \min\{m, n\}$)
  - proof: $\|\mathbf{A}\|_F = \|\mathbf{\Sigma}\|_F = \sqrt{\sum_{i=1}^{p} \sigma_i^2}$, and $\sigma_1^2 \leq \sum_{i=1}^{p} \sigma_i^2 \leq p\sigma_1^2$

Application: Equivalent Forms of $\|\mathbf{A}\|_2 \leq 1$

- we have

$$\begin{aligned}
\|\mathbf{A}\|_2 \leq 1 &\Longleftrightarrow \|\mathbf{A}\mathbf{x}\|_2 \leq 1, \quad \forall \|\mathbf{x}\|_2 = 1 \\
&\Longleftrightarrow 1 - \|\mathbf{A}\mathbf{x}\|_2^2 \geq 0, \quad \forall \|\mathbf{x}\|_2 = 1 \\
&\Longleftrightarrow \mathbf{x}^T(\mathbf{I} - \mathbf{A}\mathbf{A}^T)\mathbf{x} \geq 0, \quad \forall \|\mathbf{x}\|_2 = 1 \\
&\Longleftrightarrow \mathbf{I} - \mathbf{A}\mathbf{A}^T \succeq \mathbf{0} \Longleftrightarrow \begin{bmatrix} \mathbf{I} & \mathbf{A} \\ \mathbf{A}^T & \mathbf{I} \end{bmatrix} \succeq \mathbf{0}
\end{aligned}$$

- or, equivalently,

$$\begin{aligned}
\|\mathbf{A}\|_2 \leq 1 &\Longleftrightarrow \sigma_{\max}(\mathbf{A}) \leq 1 \\
&\Longleftrightarrow \lambda_{\max}(\mathbf{A}\mathbf{A}^T) \leq 1 \\
&\Longleftrightarrow \mathbf{I} \succeq \mathbf{A}\mathbf{A}^T \Longleftrightarrow \begin{bmatrix} \mathbf{I} & \mathbf{A} \\ \mathbf{A}^T & \mathbf{I} \end{bmatrix} \succeq \mathbf{0}
\end{aligned}$$

Schatten Norms

- applying the p-norm to the vector of singular values of matrix $\mathbf{A}$

$$f(\mathbf{A}) = \left(\sum_{i=1}^{\min\{m,n\}} \sigma_i(\mathbf{A})^p \right)^{\frac{1}{p}}, \qquad p \geq 1,$$

  is known to be a norm and is called the Schatten p-norm
- Frobenius norm when $p = 2$; spectral norm when $p = \infty$
- nuclear norm (or trace norm) when $p = 1$:

$$\|\mathbf{A}\|_* = \|\mathbf{A}\|_{\text{tr}} = \sum_{i=1}^{\min\{m,n\}} \sigma_i(\mathbf{A}) = \text{tr}((\mathbf{A}^T\mathbf{A})^{\frac{1}{2}})$$

  - a special case of the Schatten p-norm
  - a way to prove that the nuclear norm is a norm:
    - show that $f(\mathbf{A}) = \max_{\|\mathbf{B}\|_2 \leq 1} \text{tr}(\mathbf{B}^T\mathbf{A})$ is a norm
    - show that $f(\mathbf{A}) = \sum_{i=1}^{\min\{m,n\}} \sigma_i$
  - finds applications in rank approximation, e.g., for compressive sensing and matrix completion **[Recht-Fazel-Parrilo'10]**

Rank Function

- $\mathrm{rank}(\mathbf{A})$ is nonconvex in $\mathbf{A}$ and is arguably hard to do optimization with it
- Idea: the rank function can be expressed as

$$\mathrm{rank}(\mathbf{A}) = \sum_{i=1}^{\min\{m,n\}} \mathbb{1}\{\sigma_i(\mathbf{A}) \neq 0\},$$

  and why not approximate it by

$$f(\mathbf{A}) = \textstyle\sum_{i=1}^{\min\{m,n\}} \varphi(\sigma_i(\mathbf{A}))$$

  for some friendly function φ?

- nuclear norm

$$\|\mathbf{A}\|_* = \textstyle\sum_{i=1}^{\min\{m,n\}} \sigma_i(\mathbf{A})$$

  - uses $\varphi(z) = z$
  - is convex in $\mathbf{A}$
  - a convex envelope of $\mathrm{rank}(\mathbf{A})$

Nuclear Norm and Low-Rank Matrix Factorization

Fact: Given $\mathbf{A} \in \mathbb{R}^{m \times n}$ and the low-rank factorization $\mathbf{A} = \mathbf{LR}$ with $\mathbf{L} \in \mathbb{R}^{m \times p}$ and $\mathbf{R} \in \mathbb{R}^{p \times n}$ where $p = \min\{m, n\}$, we have

$$\|\mathbf{A}\|_* = \frac{1}{2}(\|\mathbf{L}\|_F^2 + \|\mathbf{R}\|_F^2).$$

Proof:

- Let the thin SVD of $\mathbf{A}$ is $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ with $\mathbf{U} \in \mathbb{R}^{m \times p}$, $\mathbf{\Sigma} \in \mathbb{S}^p$, and $\mathbf{V} \in \mathbb{R}^{n \times p}$. Let $\mathbf{L} = \mathbf{U}\mathbf{\Sigma}^{\frac{1}{2}}$ and $\mathbf{R} = \mathbf{\Sigma}^{\frac{1}{2}}\mathbf{V}^T$.
- $\|\mathbf{L}\|_F^2 = \mathrm{tr}(\mathbf{L}^T\mathbf{L}) = \mathrm{tr}((\mathbf{U}\mathbf{\Sigma}^{\frac{1}{2}})^T\mathbf{U}\mathbf{\Sigma}^{\frac{1}{2}}) = \mathrm{tr}(\mathbf{\Sigma}) = \sum_{i=1}^{p} \sigma_i(\mathbf{A}) = \|\mathbf{A}\|_*$. Similarly, we have $\|\mathbf{R}\|_F^2 = \|\mathbf{A}\|_*$.
- this technique can be used for low-rank matrix learning problems

Dual Norm

For a given norm $\|\cdot\|$ on $\mathbb{R}^{m\times n}$, the dual norm, denoted $\|\cdot\|^{\mathrm{d}}$, is the function from $\mathbb{R}^{m\times n}$ to $\mathbb{R}$ with values

$$\|\mathbf{Y}\|^{\mathrm{d}} = \max_{\|\mathbf{X}\|\leq 1} \langle \mathbf{X}, \mathbf{Y}\rangle,$$

- The above definition indeed corresponds to a norm: it is convex, as it is the pointwise maximum of convex (in fact, linear) functions $\mathbf{X} \to \langle \mathbf{X}, \mathbf{Y}\rangle$; it is homogeneous of degree 1, i.e., $\|\alpha\mathbf{X}\|_* = \alpha\|\mathbf{X}\|_*$ for every $\mathbf{X}$ in $\mathbb{R}^{m\times n}$ and $\alpha \geq 0$.
- By definition of the dual norm,

$$\mathrm{tr}(\mathbf{Y}^T\mathbf{X}) \leq \|\mathbf{X}\| \cdot \|\mathbf{Y}\|^{\mathrm{d}}.$$

- Examples:
  - The norm dual to the Frobenius norm is itself.
  - The norm dual to the elementwise 1-norm is the elementwise ∞-norm.
  - The norm dual to the spectral norm is the nuclear norm.
- The dual of the dual norm is the original norm.

Dual Norm

Fact: The norm dual to the spectral norm is the nuclear norm, i.e.,

$$\|\mathbf{Y}\|_* = \max_{\|\mathbf{X}\|_2 \leq 1} \operatorname{tr}(\mathbf{Y}^T\mathbf{X}),$$

Proof:

- Let $\mathbf{Y} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$. We have

$$\begin{aligned}
\max_{\|\mathbf{X}\|_2 \leq 1} \operatorname{tr}(\mathbf{Y}^T\mathbf{X}) &= \max_{\|\mathbf{X}\|_2 \leq 1} \operatorname{tr}(\mathbf{V}\mathbf{\Sigma}\mathbf{U}^T\mathbf{X}) = \max_{\|\mathbf{X}\|_2 \leq 1} \operatorname{tr}(\mathbf{\Sigma}\mathbf{U}^T\mathbf{X}\mathbf{V}) \\
&= \max_{\|\mathbf{X}\|_2 \leq 1} \sum_{i=1}^{\min(m,n)} \sigma_i(\mathbf{Y}) \left[\mathbf{U}^T\mathbf{X}\mathbf{V}\right]_{ii} = \max_{\|\mathbf{X}\|_2 \leq 1} \sum_i \sigma_i(\mathbf{Y}) \mathbf{u}_i^T\mathbf{X}\mathbf{v}_i \\
&\leq \max_{\|\mathbf{X}\|_2 \leq 1} \sum_i \sigma_i(\mathbf{Y})\sigma_{\max}(\mathbf{X}) = \sum_i \sigma_i(\mathbf{Y}) = \|\mathbf{Y}\|_*
\end{aligned}$$

- Let $\mathbf{X}_1 = \mathbf{U}_1\mathbf{V}_1^T$ which satisfies $\|\mathbf{X}_1\|_2 \leq 1$. Then,

$$\max_{\|\mathbf{X}\|_2 \leq 1} \operatorname{tr}(\mathbf{Y}^T\mathbf{X}) \geq \operatorname{tr}(\mathbf{Y}^T\mathbf{X}_1) = \operatorname{tr}(\mathbf{V}_1\tilde{\mathbf{\Sigma}}\mathbf{U}_1^T\mathbf{U}_1\mathbf{V}_1^T) = \operatorname{tr}(\tilde{\mathbf{\Sigma}}) = \sum_{i=1}^{r} \sigma_i(\mathbf{Y}) = \|\mathbf{Y}\|_*.$$

Linear Systems: Interpretation under SVD

- consider the linear system

$$\mathbf{y} = \mathbf{A}\mathbf{x}$$

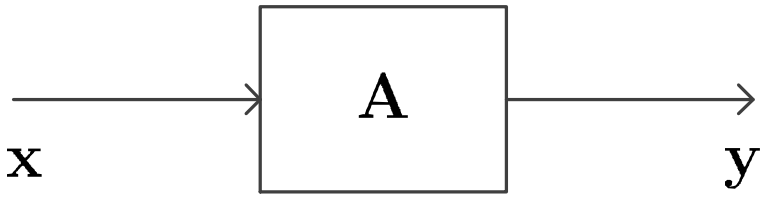
where $\mathbf{A} \in \mathbb{R}^{m \times n}$ is the system matrix; $\mathbf{x} \in \mathbb{R}^n$ is the system input (the domain); $\mathbf{y} \in \mathbb{R}^m$ is the system output (the range)

- by SVD we can write

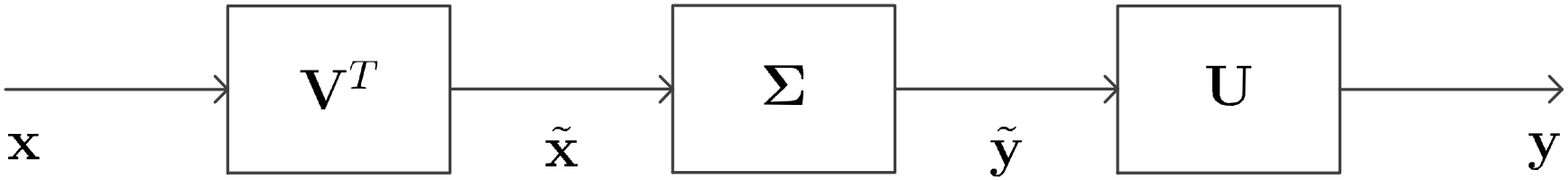
$$\mathbf{y} = \mathbf{U}\tilde{\mathbf{y}}, \qquad \tilde{\mathbf{y}} = \mathbf{\Sigma}\tilde{\mathbf{x}}, \qquad \tilde{\mathbf{x}} = \mathbf{V}^T\mathbf{x}$$

- **Implication:** *every* linear system $\mathbf{A}$ (a mapping from $\mathbb{R}^n$ to $\mathbb{R}^m$) works by performing three processes in cascade, namely,
  - rotate/reflect the system input $\mathbf{x}$ to form an intermediate system input $\tilde{\mathbf{x}}$
  - form an intermediate system output $\tilde{\mathbf{y}}$ by element-wise rescaling $\tilde{\mathbf{x}}$ w.r.t. σ_i's and by either removing some entires of $\tilde{\mathbf{x}}$ or adding some zeros
  - rotate/reflect $\tilde{\mathbf{y}}$ to form the system output $\mathbf{y}$
- **Implication:** *every* linear system $\mathbf{A}$ reduces to the diagonal matrix $\mathbf{\Sigma}$ when the range $\mathbf{y}$ is expressed in the basis of columns of $\mathbf{U}$ and the domain $\mathbf{x}$ is expressed in the basis of columns of $\mathbf{V}$

Linear Systems: Interpretation under SVD



(a) linear system

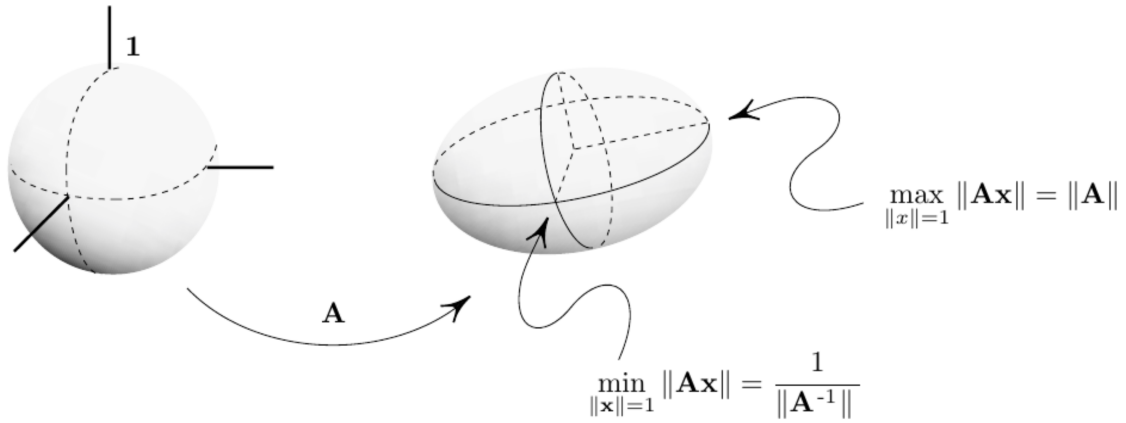


(b) equivalent system

Linear Systems: Interpretation under SVD

- SVD reveals the geometry about linear transformation $\mathbf{y} = \mathbf{A}\mathbf{x}$
- **Example:** consider the transformation of a unit sphere in $\mathbb{R}^3$ under a nonsingular $\mathbf{A} \in \mathbb{R}^{3\times 3}$ and the singular values tell how much distortion can occur under $\mathbf{A}$

$$1 \geq \|\mathbf{x}\|_2^2 = \|\mathbf{A}^{-1}\mathbf{A}\mathbf{x}\|_2^2 = \|\mathbf{A}^{-1}\mathbf{y}\|_2^2 = \|\mathbf{V}\mathbf{\Sigma}^{-1}\mathbf{U}^T\mathbf{y}\|_2^2 = \|\mathbf{\Sigma}^{-1}\mathbf{U}^T\mathbf{y}\|_2^2$$



(recall the result $\sigma_{\min}(\mathbf{A})\|\mathbf{x}\|_2^2 \leq \|\mathbf{y}\|_2^2 = \|\mathbf{A}\mathbf{x}\|_2^2 \leq \sigma_{\max}(\mathbf{A})\|\mathbf{x}\|_2^2$ for $m \geq n$)

- similar results apply to rectangular and singular $\mathbf{A}$
- **Fact:** the image of the unit sphere under *any* linear map $\mathbf{A}$ is a hyperellipse
- **Fact:** the amount of distortion of unit sphere under transformation $\mathbf{A}$ determines the degree to which uncertainties in a linear system $\mathbf{y} = \mathbf{A}\mathbf{x}$ can be magnified

Linear Systems: Sensitivity Analysis

- Scenario:
  - let $\mathbf{A} \in \mathbb{R}^{n \times n}$ be nonsingular, and $\mathbf{y} \in \mathbb{R}^n$. Let $\mathbf{x}$ be the solution to

$$\mathbf{y} = \mathbf{A}\mathbf{x}.$$

  - it is a well-determined linear system
  - consider a perturbed version of the above system: $\hat{\mathbf{A}} = \mathbf{A} + \Delta\mathbf{A}, \hat{\mathbf{y}} = \mathbf{y} + \Delta\mathbf{y}$, where $\Delta\mathbf{A}$ and $\Delta\mathbf{y}$ are errors. Let $\hat{\mathbf{x}}$ be a solution to the perturbed system

$$\hat{\mathbf{y}} = \hat{\mathbf{A}}\hat{\mathbf{x}}.$$

- Problem: analyze how the solution error $\|\hat{\mathbf{x}} - \mathbf{x}\|_2$ scales with $\Delta\mathbf{A}$ and $\Delta\mathbf{y}$
- remark: $\Delta\mathbf{A}$ and $\Delta\mathbf{y}$ may be floating point errors, measurement errors, etc

Linear Systems: Sensitivity Analysis

- the condition number of a given nonsingular matrix $\mathbf{A}$ for a given matrix operator norm is defined as

$$\kappa(\mathbf{A}) = \|\mathbf{A}\|\|\mathbf{A}^{-1}\|$$

  - $\kappa(\mathbf{A}) \geq 1$
  - $\mathbf{A}$ is said to be well-conditioned if $\kappa(\mathbf{A})$ is small
  - $\mathbf{A}$ is said to be ill-conditioned if $\kappa(\mathbf{A})$ is very large; that refers to cases where $\mathbf{A}$ is close to singular (high linear dependence between columns or rows of $\mathbf{A}$)
  - it is customary to denote $\kappa(\mathbf{A}) = \infty$ if $\mathbf{A}$ is a singular matrix

- the 2-norm condition number of a given nonsingular matrix $\mathbf{A}$ is given by

$$\kappa_2(\mathbf{A}) = \|\mathbf{A}\|_2\|\mathbf{A}^{-1}\|_2 = \frac{\sigma_{\max}(\mathbf{A})}{\sigma_{\min}(\mathbf{A})}$$

  - $\kappa_2(\mathbf{A}) = 1$ if $\mathbf{A}$ is a multiple of an orthogonal matrix (perfectly conditioned)

- if not specially specified, the condition number is commonly referred to as $\kappa_2(\mathbf{A})$

Linear Systems: Sensitivity Analysis

Theorem 2. If $\mathbf{A}$ is known exactly and there is an uncertainty $\Delta\mathbf{y}$, then

$$\kappa_2^{-1}(\mathbf{A})\frac{\|\Delta\mathbf{y}\|_2}{\|\mathbf{y}\|_2} \leq \frac{\|\hat{\mathbf{x}}-\mathbf{x}\|_2}{\|\mathbf{x}\|_2} \leq \kappa_2(\mathbf{A})\frac{\|\Delta\mathbf{y}\|_2}{\|\mathbf{y}\|_2}.$$

(requires a proof)

- if $\mathbf{A}$ is well-conditioned, a small uncertainty in $\mathbf{y}$ cannot produce a very large solution error
- if $\mathbf{A}$ is ill-conditioned, a small uncertainty in $\mathbf{y}$ can produce a very large solution error, or a large uncertainty in $\mathbf{y}$ can produce a very small solution error, which depends on the "direction" of $\Delta\mathbf{y}$

Theorem 3. If $\mathbf{y}$ is known exactly and there is an uncertainty $\Delta\mathbf{A}$, then

$$\frac{\|\hat{\mathbf{x}}-\mathbf{x}\|_2}{\|\hat{\mathbf{x}}\|_2} \leq \kappa_2(\mathbf{A})\frac{\|\Delta\mathbf{A}\|_2}{\|\mathbf{A}\|_2} \quad \text{and} \quad \frac{\|\hat{\mathbf{x}}-\mathbf{x}\|_2}{\|\mathbf{x}\|_2} \leq \frac{1}{1-\kappa_2(\mathbf{A})\frac{\|\Delta\mathbf{A}\|_2}{\|\mathbf{A}\|_2}}\kappa_2(\mathbf{A})\frac{\|\Delta\mathbf{A}\|_2}{\|\mathbf{A}\|_2}.$$

(proof by yourself)

Linear Systems: Sensitivity Analysis

Theorem 4. If there are uncertainties $\Delta\mathbf{A}$ and $\Delta\mathbf{y}$, then

$$\frac{\|\hat{\mathbf{x}}-\mathbf{x}\|_2}{\|\hat{\mathbf{x}}\|_2} \leq \kappa_2(\mathbf{A})\left(\frac{\|\Delta\mathbf{A}\|_2}{\|\mathbf{A}\|_2}+\frac{\|\Delta\mathbf{y}\|_2}{\|\mathbf{A}\|_2\|\hat{\mathbf{x}}\|_2}\right)$$

and

$$\text{or} \quad \frac{\|\hat{\mathbf{x}}-\mathbf{x}\|_2}{\|\mathbf{x}\|_2} \leq \frac{1}{1-\kappa_2(\mathbf{A})\frac{\|\Delta\mathbf{A}\|_2}{\|\mathbf{A}\|_2}}\kappa_2(\mathbf{A})\left(\frac{\|\Delta\mathbf{A}\|_2}{\|\mathbf{A}\|_2}+\frac{\|\Delta\mathbf{y}\|_2}{\|\mathbf{y}\|_2}\right).$$

(proof by yourself)

Linear Systems: Sensitivity Analysis

Theorem 5. Let $\varepsilon > 0$ be a constant such that

$$\frac{\|\Delta \mathbf{A}\|_2}{\|\mathbf{A}\|_2} \leq \varepsilon, \qquad \frac{\|\Delta \mathbf{y}\|_2}{\|\mathbf{y}\|_2} \leq \varepsilon.$$

If ε is sufficiently small such that $\varepsilon \kappa_2(\mathbf{A}) < 1$, then

$$\frac{\|\hat{\mathbf{x}} - \mathbf{x}\|_2}{\|\mathbf{x}\|_2} \leq \frac{2\varepsilon \kappa_2(\mathbf{A})}{1 - \varepsilon \kappa_2(\mathbf{A})}.$$

(requires a proof)

- Implications:
  - for small errors and in the worst-case sense, the relative error $\|\hat{\mathbf{x}} - \mathbf{x}\|_2 / \|\mathbf{x}\|_2$ tends to increase with the condition number
  - in particular, for $\varepsilon \kappa_2(\mathbf{A}) \leq \frac{1}{2}$, the error bound can be simplified to

$$\frac{\|\hat{\mathbf{x}} - \mathbf{x}\|_2}{\|\mathbf{x}\|_2} \leq 4\varepsilon \kappa_2(\mathbf{A})$$

    where the error bound scales linearly with the condition number

Linear Systems: Sensitivity Analysis

- Scenario:
  - let $\mathbf{A} \in \mathbb{R}^{m \times n}$ be nonsingular, and $\mathbf{y} \in \mathbb{R}^m$. A vector $\mathbf{x}$ is an optimal solution to the least squares problem

$$\min_{\mathbf{x} \in \mathbb{R}^n} \ \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2$$

    if and only if it satisfies the normal equation

$$\mathbf{A}^T \mathbf{A}\mathbf{x} = \mathbf{A}^T \mathbf{y}.$$

  - consider a perturbed version of the above system: $\hat{\mathbf{A}} = \mathbf{A} + \Delta\mathbf{A}, \hat{\mathbf{y}} = \mathbf{y} + \Delta\mathbf{y}$, where $\Delta\mathbf{A}$ and $\Delta\mathbf{y}$ are errors. Let $\hat{\mathbf{x}}$ be a solution to the perturbed system

$$\hat{\mathbf{A}}^T \hat{\mathbf{A}}\hat{\mathbf{x}} = \hat{\mathbf{A}}^T \hat{\mathbf{y}}.$$

- Problem: analyze how the solution error $\|\hat{\mathbf{x}} - \mathbf{x}\|_2$ scales with $\Delta\mathbf{A}$ and $\Delta\mathbf{y}$

Linear Systems: Sensitivity Analysis

- note that the condition number

$$\kappa_2(\mathbf{A}^T\mathbf{A}) = (\kappa_2(\mathbf{A}))^2$$

- implication: we should avoid directly solving the normal equation

- when the QR decompostion $\mathbf{A} = \mathbf{QR}$ is applied for least squares solving, we have

$$\kappa_2(\mathbf{Q}) = 1 \quad \text{and} \quad \kappa_2(\mathbf{A}) = \kappa_2(\mathbf{Q}^T\mathbf{A}) = \kappa_2(\mathbf{R})$$

  in which case the influence of $\Delta\mathbf{A}$ and $\Delta\mathbf{y}$ to the solution error in least squares is propotional to $\kappa_2(\mathbf{A})$ in the same way as in the linear system

- implication: least squares via QR is more numerically stable

- Question: how to tackle the ill-conditioned $\mathbf{A}$? one solution is the total least squares method (in Least Squares Revisited Topic) which relies on the SVD