Matrix Algebra Review

Kevin Kircher — Cornell MAE — Spring '14

These notes collect some useful facts from finite dimensional linear algebra.

Contents

1 Working with matrices and vectors

A vector $\mathbf{v} \in \mathbf{R}^n$ is a column of n real scalars:

$$\mathbf{v} = \begin{bmatrix} v_1 \\ \vdots \\ v_n \end{bmatrix} \qquad \Longleftrightarrow \qquad \mathbf{v}^T = \begin{bmatrix} v_1 & \dots & v_n \end{bmatrix} \in \mathbf{R}^{1 \times n}$$

where $\mathbf{v}^T$ is the **transpose** of $\mathbf{v}$. In these notes vectors are denoted by bold, lower case letters.

A matrix $A \in \mathbf{R}^{m \times n}$ contains n columns of m real scalars:

$$A = \begin{bmatrix} a_{11} & \dots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \dots & a_{mn} \end{bmatrix} \qquad \Longleftrightarrow \qquad A^T = \begin{bmatrix} a_{11} & \dots & a_{m1} \\ \vdots & \ddots & \vdots \\ a_{1n} & \dots & a_{mn} \end{bmatrix} \in \mathbf{R}^{n \times m}$$

We sometimes write $A = \begin{bmatrix} a_{ij} \end{bmatrix}$ or $(A)_{ij} = a_{ij}$ to mean that A is built from the elements a_{ij}. In these notes matrices are denoted by non-bold capital letters.

Some definitions and properties of operations between scalars, vectors and matrices follow.

- Vector addition for $\mathbf{u}, \mathbf{v} \in \mathbf{R}^n$:

$$\mathbf{u} + \mathbf{v} = \begin{bmatrix} u_1 + v_1 \\ \vdots \\ u_n + v_n \end{bmatrix}$$

- Addition of matrices $A, B \in \mathbf{R}^{m \times n}$:

$$A + B = \begin{bmatrix} a_{11} + b_{11} & \dots & a_{1n} + b_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} + b_{m1} & \dots & a_{mn} + b_{mn} \end{bmatrix}$$

- Vector-scalar multiplication for $\alpha \in \mathbf{R}$, $\mathbf{v} \in \mathbf{R}^n$:

$$\alpha \mathbf{v} = \begin{bmatrix} \alpha v_1 \\ \vdots \\ \alpha v_n \end{bmatrix}$$

- Matrix-scalar multiplication for $\alpha \in \mathbf{R}$, $A \in \mathbf{R}^{n \times m}$:

$$\alpha A = \begin{bmatrix} \alpha a_{11} & \dots & \alpha a_{1n} \\ \vdots & \ddots & \vdots \\ \alpha a_{m1} & \dots & \alpha a_{mn} \end{bmatrix}$$

- The **inner product** of vectors $\mathbf{u}, \mathbf{v} \in \mathbf{R}^n$ is

$$\langle \mathbf{u}, \mathbf{v} \rangle = \mathbf{u} \cdot \mathbf{v} = \mathbf{u}^T \mathbf{v} = \sum_{i=1}^{n} u_i v_i \in \mathbf{R}$$

  Two vectors are **orthogonal** iff their inner product is zero.

- The **angle** between $\mathbf{u}$ and $\mathbf{v} \in \mathbf{R}^n$ is

$$\angle(\mathbf{u}, \mathbf{v}) = \cos^{-1}\left(\frac{\mathbf{u}^T \mathbf{v}}{\sqrt{(\mathbf{u}^T \mathbf{u})(\mathbf{v}^T \mathbf{v})}} \right) \in [0, \pi]$$

- The **orthogonal projection** of $\mathbf{v} \in \mathbf{R}^n$ onto $\mathbf{u} \in \mathbf{R}^n$ is

$$\mathrm{proj}_{\mathbf{u}} \mathbf{v} = \frac{\mathbf{v}^T \mathbf{u}}{\mathbf{u}^T \mathbf{u}} \mathbf{u}$$

  This allows decomposition of $\mathbf{v}$ into components parallel and orthogonal to $\mathbf{u}$:

$$\mathbf{v} = \mathbf{v}_\perp + \mathbf{v}_\parallel$$

  where

$$\mathbf{v}_\perp = \mathrm{proj}_{\mathbf{u}} \mathbf{v}, \quad \mathbf{v}_\parallel = \mathbf{v} - \mathbf{v}_\perp$$

- The **outer product** of vectors $\mathbf{u} \in \mathbf{R}^n$ and $\mathbf{v} \in \mathbf{R}^m$ is

$$|\mathbf{u}\rangle\langle\mathbf{v}| = \mathbf{u}\mathbf{v}^T = \begin{bmatrix} u_i v_j \end{bmatrix} \in \mathbf{R}^{n \times m}$$

- Matrix multiplication of $A \in \mathbf{R}^{m \times n}$, $B \in \mathbf{R}^{n \times p} \Rightarrow AB \in \mathbf{R}^{m \times p}$

$$(AB)_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj}, \quad i \in \{1, \ldots, m\}, \quad j \in \{1, \ldots, p\}$$

  It can be useful to view matrix multiplication in terms of rows and columns: if

$$A = \begin{bmatrix} \boldsymbol{\alpha}_1^T \\ \vdots \\ \boldsymbol{\alpha}_m^T \end{bmatrix}, \quad B = \begin{bmatrix} \mathbf{b}_1 & \ldots & \mathbf{b}_p \end{bmatrix}$$

  then

$$(AB)_{ij} = \boldsymbol{\alpha}_i^T \mathbf{b}_j \quad \Longleftrightarrow \quad AB = \begin{bmatrix} \boldsymbol{\alpha}_1^T B \\ \vdots \\ \boldsymbol{\alpha}_m^T B \end{bmatrix} = \begin{bmatrix} A\mathbf{b}_1 & \ldots & A\mathbf{b}_p \end{bmatrix}$$

Matrix multiplication can also be viewed in terms of outer products. If

$$A = \begin{bmatrix} \mathbf{a}_1 & \dots & \mathbf{a}_n \end{bmatrix}, \quad B = \begin{bmatrix} \boldsymbol{\beta}_1^T \\ \vdots \\ \boldsymbol{\beta}_n^T \end{bmatrix}$$

then

$$AB = \mathbf{a}_1 \boldsymbol{\beta}_1^T + \dots + \mathbf{a}_n \boldsymbol{\beta}_n^T$$

The transpose of a product is

$$(AB)^T = B^T A^T$$

- Multiplication by the **identity** matrix $I_n \in \mathbf{R}^{n \times n}$, for $A \in \mathbf{R}^{n \times m}$:

$$I_n = \begin{bmatrix} 1 & & \\ & \ddots & \\ & & 1 \end{bmatrix} \quad \Rightarrow \quad I_n A = A I_m = A$$

The identity matrix can be expressed in terms of the **Cartesian basis vectors** $\mathbf{e}_1, \dots, \mathbf{e}_n$:

$$\mathbf{e}_1 = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad \dots, \quad \mathbf{e}_n = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} \quad \Rightarrow \quad I_n = \begin{bmatrix} \mathbf{e}_1 & \dots & \mathbf{e}_n \end{bmatrix}$$

- Matrix-vector multiplication of $A \in \mathbf{R}^{m \times n}$, $\mathbf{x} \in \mathbf{R}^n \Rightarrow A\mathbf{x} \in \mathbf{R}^m$:

$$(A\mathbf{x})_i = \sum_{j=1}^{n} a_{ij} x_j$$

It can be useful to view matrix-vector multiplication as mixture of the columns of A:

$$A = \begin{bmatrix} \mathbf{a}_1 & \dots & \mathbf{a}_n \end{bmatrix} \quad \Rightarrow \quad A\mathbf{x} = x_1 \mathbf{a}_1 + \dots + x_n \mathbf{a}_n$$

- A set of vectors $S = \{\mathbf{a}_1, \dots, \mathbf{a}_n\}$ is **linearly dependent** if there exists an $\mathbf{x} \in \mathbf{R}^n$ such that $\mathbf{x} \neq \mathbf{0}$ and

$$x_1 \mathbf{a}_1 + \dots + x_n \mathbf{a}_n = \mathbf{0}.$$

If no such $\mathbf{x}$ exists, *i.e.* if

$$A\mathbf{x} = \mathbf{0} \quad \Rightarrow \quad \mathbf{x} = \mathbf{0}$$

where $A = \begin{bmatrix} \mathbf{a}_1 & \dots & \mathbf{a}_n \end{bmatrix}$, then S is **linearly independent**.

A linearly independent set $\{\mathbf{a}_1, \dots, \mathbf{a}_k\}$ with $\mathbf{a}_i \in \mathbf{R}^n$ can have at most n elements, *i.e.* $k \leq n$. Equivalently, any set of k n-vectors, with $k > n$, is linearly dependent.

- A set of n linearly independent n-vectors is a **basis** for $\mathbf{R}^n$. If $\{\mathbf{a}_1, \ldots, \mathbf{a}_n\}$ is a basis for $\mathbf{R}^n$, then any vector in $\mathbf{R}^n$ can be written as a linear combination of $\mathbf{a}_1, \ldots, \mathbf{a}_n$.

- The **column space** or **range** of $A \in \mathbf{R}^{n\times m}$ is the set

$$\mathrm{col}(A) = \{\mathbf{b} \in \mathbf{R}^n \mid A\mathbf{x} = \mathbf{b}, \mathbf{x} \in \mathbf{R}^n\}$$

  The column space of A is the span (the set of all linear combinations) of the columns $\mathbf{a}_1, \ldots, \mathbf{a}_n$ of A.

- The **null space** or **kernel** of $A \in \mathbf{R}^{n\times m}$ is the set

$$\mathrm{null}(A) = \{\mathbf{x} \in \mathbf{R}^m \mid A\mathbf{x} = \mathbf{0}\}$$

  Fact: $\dim(\mathrm{null}(A)) + \dim(\mathrm{col}(A)) = n$

- The **rank** of $A \in \mathbf{R}^{m\times n}$ is the number of linearly independent columns of A, *i.e.* $\mathrm{rank}(A) = \dim(\mathrm{col}(A)) \leq \min\{m, n\}$. If $\mathrm{rank}(A) = \min\{m, n\}$, then A is called **full rank**.

  Facts:

  - ◇ the row rank and column rank of a matrix are equal
  - ◇ if A is full rank, then $A\mathbf{x} = \mathbf{0} \iff \mathbf{x} = \mathbf{0}$
  - ◇ if $A \in \mathbf{R}^{n\times n}$ and $B \in \mathbf{R}^{n\times p}$, then $\mathrm{rank}(AB) \leq \min\{\mathrm{rank}(A), \mathrm{rank}(B)\}$
  - ◇ if $A \in \mathbf{R}^{m\times n}$ is full rank and $B \in \mathbf{R}^{n\times p}$, then $\mathrm{col}(AB) = \mathrm{col}(B)$ and $\mathrm{rank}(AB) = \mathrm{rank}(B)$
  - ◇ the rank of a diagonal matrix is the number of its nonzero elements
  - ◇ the rank of a symmetric matrix is the number of its nonzero eigenvalues

2 Square matrices

The definitions and results in this section apply only to **square** matrices, *i.e.* to matrices $A \in \mathbf{R}^{n\times n}$.

- The **trace** of $A \in \mathbf{R}^{n\times n}$ is

$$\operatorname{tr}(A) = \sum_{i=1}^{n} a_{ii}$$

  Cyclic permutations: for $A \in \mathbf{R}^{m\times n}$, $B \in \mathbf{R}^{n\times p}$, $C \in \mathbf{R}^{p\times m}$

$$\operatorname{tr}(ABC) = \operatorname{tr}(CAB) = \operatorname{tr}(BCA)$$

- The **determinant** of $A \in \mathbf{R}^{n\times n}$ is

$$\det(A) = \sum_{i=1}^{n} a_{ij}(-1)^{i+j}C^{ij}$$

  where the **cofactor** C^{ij} is the determinant of the $(n-1)\times(n-1)$ matrix formed by deleting the i^{th} row and j^{th} column of A. For $n = 2$,

$$\det(A) = a_{11}a_{22} - a_{12}a_{21}$$

  Facts:

  - ⋄ if $A, B \in \mathbf{R}^{n\times n}$, then $\det(AB) = \det(A)\det(B)$
  - ⋄ $\det(A^T) = \det(A)$
  - ⋄ if A^{-1} exists, then $\det(A^{-1}) = \frac{1}{\det(A)}$

- A matrix $A \in \mathbf{R}^{n\times n}$ is **singular** if $\det(A) = 0$ or, equivalently, if $\operatorname{rank}(A) < n$.

- If $A \in \mathbf{R}^{n\times n}$ is nonsingular, then there exists a matrix A^{-1}, called the **inverse** of A, such that

$$A^{-1}A = AA^{-1} = I$$

  In practice, computing A^{-1} is expensive (order n^3 flops) and may be inaccurate if A is ill-conditioned. For small matrices, A^{-1} can be found by hand using Cramer's rule,

$$A^{-1} = \frac{1}{\det(A)}\operatorname{adj}(A)^T$$

  where $\operatorname{adj}(A)$, called the **adjutant** of A, is the matrix of the cofactors of A:

$$\operatorname{adj}(A) = \left[(-1)^{i+j}C^{ij}\right] = \begin{bmatrix} C^{11} & -C^{12} & \dots & (-1)^{n+1}C^{1n} \\ -C^{12} & C^{22} & & \\ \vdots & & \ddots & \vdots \\ (-1)^{n+1}C^{n1} & \dots & & C^{nn} \end{bmatrix}$$

so for $n = 2$,

$$A^{-1} = \frac{1}{a_{11}a_{22} - a_{12}a_{21}} \begin{bmatrix} a_{22} & -a_{21} \\ -a_{12} & a_{11} \end{bmatrix}$$

If $A, B \in \mathbf{R}^{n \times n}$ are both non-singular, then

$$(AB)^{-1} = B^{-1}A^{-1}$$

- **Matrix inversion lemma**: if all the relevant inverses exist, then

$$(A + BCD)^{-1} = A^{-1} - A^{-1}B(DA^{-1}B + C^{-1})^{-1}DA^{-1}$$

- Let A be a square block matrix, and let $A^{-1} = B$, *i.e.*

$$A^{-1} = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}^{-1} = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix} = B$$

where A_{11} and A_{22} are square, $\det(A_{11}) \neq 0$, and $\dim(A_{ij}) = \dim(B_{ij})$ for all i, j. Assume as well that $\det(\Delta) \neq 0$, where $\Delta = A_{22} - A_{21}A_{11}^{-1}A_{12}$. Then the inverse of A is given by

$$\begin{aligned} B_{11} &= A_{11}^{-1} + A_{11}^{-1}A_{12}\Delta^{-1}A_{21}A_{11}^{-1} \\ B_{12} &= -A_{11}^{-1}A_{12}\Delta^{-1} \\ B_{21} &= -\Delta^{-1}A_{21}A_{11}^{-1} \\ B_{22} &= \Delta^{-1} \end{aligned}$$

- A matrix $A \in \mathbf{R}^{n \times n}$ is **idempotent** if $A^2 \equiv AA = A$.

  Facts:

  ◇ if A is idempotent, then $\mathrm{tr}(A) = \mathrm{rank}(A)$

  ◇ if A is idempotent, then any eigenvalue of A is either 0 or 1

  Examples:

  ◇ $\bar{J} = \frac{1}{n}\mathbf{1}\mathbf{1}^T \in \mathbf{R}^{n \times n}$ is idempotent. $\bar{J}$ is called an **averaging matrix** because

$$\bar{J}\mathbf{x} = \frac{1}{n}\mathbf{1}\mathbf{1}^T\mathbf{x} = \left(\frac{1}{n}\mathbf{1}^T\mathbf{x}\right)\mathbf{1} = \bar{x}\mathbf{1}$$

  where $\bar{x}$ is the mean of the elements of $\mathbf{x}$.

  ◇ $C = I - \bar{J}$ is idempotent. C is called a **centering matrix** because

$$C\mathbf{x} = (I - \bar{J})\mathbf{x} = \mathbf{x} - \bar{x}\mathbf{1}$$

- (Cayley-Hamilton theorem: A solves its own characteristic polynomial)

  Let $A \in \mathbf{R}^{n\times n}$ and $p(\lambda) = \det(\lambda I - A)$, *i.e.* let $p : \mathbf{C} \to \mathbf{R}$ be the characteristic polynomial of A. Overload $p : \mathbf{R}^{n\times n} \to \mathbf{R}$ such that

$$\begin{aligned} & p(\lambda) = \alpha_0 + \alpha_1\lambda + \cdots + \alpha_n\lambda^n \\ \iff\quad & p(A) = \alpha_0 I + \alpha_1 A + \cdots + \alpha_n A^n \end{aligned}$$

  then $p(A) = 0$. This result allows A^p to be expressed as a linear combination of $I, A, \ldots, A^{n-1}$ for any $p \in \{1, 2, \ldots\}$ (and for negative integers, if A^{-1} exists).

2.1 Eigenstuff and diagonalization

If A is square, then there exist at least one $\lambda_i \in \mathbf{R}$ and $\mathbf{v}_i \in \mathbf{R}^n$ such that $\mathbf{v}_i \neq \mathbf{0}$ and $A\mathbf{v}_i = \lambda_i\mathbf{v}_i$, *i.e.* multiplication by A stretches $\mathbf{v}_i$ but doesn't rotate it. These special vectors $\mathbf{v}_i$ and scalars λ_i are called the **eigenvectors** and **eigenvalues** of A, respectively. Interesting fact: the image of the unit sphere, transformed by A, is an ellipsoid with principle axes $\lambda_i\mathbf{v}_i$.

We can find the eigenstuff of A by solving $\det(\lambda I - A) = 0$, an n^{th}-order polynomial in λ, for $\lambda_1, \ldots, \lambda_n$. The eigenvalues of A are generally complex and not necessarily distinct.

Let λ_i be an eigenvalue of $A \in \mathbf{R}^{n\times n}$ with corresponding eigenvector $\mathbf{v}_i$. Then

- $\operatorname{tr}(A) = \sum_i \lambda_i$

  (so the eigenvalues of A^T and A are the same)

- $\det(A) = \prod_i \lambda_i$

  (so A is singular iff $\lambda_i = 0$ for some $i \in \{1, \ldots, n\}$)

- if $\det(A) \neq 0$, then $1/\lambda_i$ is an eigenvalue of A^{-1}

- the eigenvalues of A^k are $\lambda_1^k, \ldots, \lambda_n^k$

  (so the linear system $\mathbf{x}(k+1) = A\mathbf{x}(k)$ is stable iff $|\lambda_i| < 1$ for all i)

- $\alpha\lambda_i + \beta$ is an eigenvalue of $\alpha A + \beta I$ with eigenvector $\mathbf{v}_i$

Other facts:

- if $A \in \mathbf{S}^n$, then the eigenvalues $\lambda_1, \ldots, \lambda_n$ of A are real (but not necessarily distinct), and A has a full set of n linearly independent eigenvectors.

- if $A \in \mathbf{S}^n_+$, then the eigenvalues of A are real and nonnegative

- if $A \in \mathbf{S}^n_{++}$, then the eigenvalues of A are real and positive

- the eigenvalues of the block triangular matrix

$$\begin{bmatrix} A & B \\ 0 & C \end{bmatrix}$$

are the eigenvalues of A and the eigenvalues of C

- if every row of A sums to c, then c is an eigenvalue of A with eigenvector $\begin{bmatrix}1, \ldots, 1\end{bmatrix}^T$
- if every column of A sums to c, then c is an eigenvalue of A (no info about the corresponding eigenvector)

A major application of eigenstuff is **diagonalization**, also called **spectral decomposition**. For most, *but not all*, $A \in \mathbf{R}^{n\times n}$, there exist nonsingular T and diagonal Λ such that $A = T^{-1}\Lambda T$. Diagonalization greatly facilitates computations and proofs. Examples:

- **matrix exponential**

$$e^{At} = e^{T^{-1}\Lambda T t} = T^{-1} \begin{bmatrix} e^{\lambda_1 t} & & \\ & \ddots & \\ & & e^{\lambda_n t} \end{bmatrix} T$$

- **powers**

$$A^k = (T^{-1}\Lambda T)^k = T^{-1} \begin{bmatrix} {\lambda_1}^k & & \\ & \ddots & \\ & & {\lambda_n}^k \end{bmatrix} T$$

(always valid for $k \geq 1$, valid for $k < 1$ if A is invertible)

Theorem 1 (Spectral decomposition) *Let $A \in \mathbf{R}^{n\times n}$ have the n linearly independent eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$ with associated eigenvalues $\lambda_1, \ldots, \lambda_n$, and let $T = \begin{bmatrix}\mathbf{v}_1 \ldots \mathbf{v}_n\end{bmatrix}$ and $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$. Then $A = T\Lambda T^{-1}$.*

Corollary 2 *If $A \in \mathbf{R}^{n\times n}$ has n distinct eigenvalues, then A is diagonalizable.*

Note that the converse is not true: there are diagonalizable matrices that do not have full sets of distinct eigenvalues.

It's hard to tell whether a general matrix A satisfies these conditions without computing its eigenstuff. An exception is if A is real and symmetric.

Lemma 3 (Principle axis theorem) *If $A = A^T \in \mathbf{R}^{n\times n}$, then A is orthogonally diagonalizable, i.e.*

$$A = T\Lambda T^T,$$

where T is orthogonal ($TT^T = T^T T = I$) and Λ is diagonal.

3 Norms

- A **norm** $\|\cdot\| : \mathbf{R}^n \to \mathbf{R}$ of a vector satisfies

  1. nonnegativity
  $$\|\mathbf{v}\| \geq 0 \text{ for all } \mathbf{v} \in \mathbf{R}^n$$
  2. definiteness
  $$\|\mathbf{v}\| = 0 \iff \mathbf{v} = \mathbf{0}$$
  3. homogeneity
  $$\|\alpha \mathbf{v}\| = |\alpha| \, \|\mathbf{v}\| \text{ for all } \mathbf{v} \in \mathbf{R}^n, \alpha \in \mathbf{R}$$
  4. triangle inequality
  $$\|\mathbf{u} + \mathbf{v}\| \leq \|\mathbf{u}\| + \|\mathbf{v}\|$$

  There are many vector norms. The most common is the **Euclidean norm** (also called the l_2 norm):
  $$\|\mathbf{v}\|_2 = \sqrt{\mathbf{v}^T \mathbf{v}}$$
  Other common examples are the l_1 norm:
  $$\|\mathbf{v}\|_1 = \sum_{i=1}^{n} |v_i|$$
  and the l_∞ norm:
  $$\|\mathbf{v}\|_\infty = \max_{i \in \{1,\ldots,n\}} |v_i|$$
  These can be viewed as members of the family of l_p norms for $p \geq 1$,
  $$\|\mathbf{v}\|_p = \left(\sum_{i=1}^{n} |v_i|^p \right)^{1/p}$$
  with l_∞ the limiting case as $p \to \infty$.

- (Schwartz inequality) For vectors $\mathbf{u}, \mathbf{v} \in \mathbf{R}^n$,
  $$\left|\mathbf{u}^T \mathbf{v}\right| \leq \|\mathbf{u}\| \, \|\mathbf{v}\|$$

- The **induced norm** of a matrix $A \in \mathbf{R}^{m \times n}$ is
  $$\|A\| = \max_{\|\mathbf{v}\|=1} \|A\mathbf{v}\| = \max_{\mathbf{v} \neq \mathbf{0}} \frac{\|A\mathbf{v}\|}{\|\mathbf{v}\|}$$
  Loosely, this measures "the biggest stretch" that multiplication by A induces on a unit vector, with respect to the norm $\|\cdot\|$. In the special case of the l_2 norm,
  $$\|A\|_2 = \sqrt{\lambda_{\max(} A^T A)} = \sigma_{\max}(A)$$
  where $\lambda_{\max}$ and $\sigma_{\max}$ denote the maximum eigenvalue and singular value, respectively.

- The **Frobenius norm** of $A \in \mathbf{R}^{m \times n}$ is

$$\|A\|_F = \left(\sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij}^2 \right)^{1/2} = \sqrt{\operatorname{tr}(A^T A)}$$

4 Useful sets and decompositions

- Call the set of $n \times n$ real, **symmetric** matrices $\mathbf{S}^n$, where

$$\mathbf{S}^n = \{A \in \mathbf{R}^{n \times n} \mid A = A^T\}$$

- Call the set of **positive semi-definite** (psd) matrices $\mathbf{S}^n_+$, where

$$\mathbf{S}^n_+ = \{A \in \mathbf{S}^n \mid \mathbf{v}^T A \mathbf{v} \geq 0 \text{ for all } \mathbf{v} \neq \mathbf{0}\}$$

  If $A \in \mathbf{S}^n_+$, then

  - we write $S \succeq 0$
  - all eigenvalues of A are nonnegative

- Call the set of **positive definite** (pd) matrices $\mathbf{S}^n_{++}$, where

$$\mathbf{S}^n_{++} = \{A \in \mathbf{S}^n \mid \mathbf{v}^T A \mathbf{v} > 0 \text{ for all } \mathbf{v} \neq \mathbf{0}\}$$

  If $A \in \mathbf{S}^n_{++}$, then

  - we write $S \succ 0$. Note that $\mathbf{S}^n_{++} \subset \mathbf{S}^n_+ \subset \mathbf{S}^n$
  - A is invertible and $A^{-1} \in \mathbf{S}^n_{++}$
  - all eigenvalues of A are positive

- If $P \in \mathbf{S}^n_{++}$, the **weighted inner product** of vectors $\mathbf{u}, \mathbf{v} \in \mathbf{R}^n$ is

$$\mathbf{u}^T P \mathbf{v} = \sum_{i=1}^{n} \sum_{j=1}^{n} u_i p_{ij} v_j \in \mathbf{R}$$

- If $P \in \mathbf{S}^n$, then the product

$$\mathbf{v}^T P \mathbf{v} = \sum_{i=1}^{n} \sum_{j=1}^{n} v_i p_{ij} v_j \in \mathbf{R}$$

  is called a **quadratic form**. Without loss of generality, can assume $P = P^T$. If $P \neq P^T$, then replace P by $\frac{1}{2}(P + P^T)$:

$$\frac{1}{2}\mathbf{v}^T(P + P^T)\mathbf{v} = \frac{1}{2}(\mathbf{v}^T P \mathbf{v} + \mathbf{v}^T P^T \mathbf{v}) = \frac{1}{2}\left[\mathbf{v}^T P \mathbf{v} + (\mathbf{v}^T P \mathbf{v})^T\right] = \mathbf{v}^T P \mathbf{v}$$

- If $Q \in \mathbf{R}^{n \times n}$ and $Q^T Q = Q Q^T = I$, then Q is **orthogonal**.

  Properties:

- $\diamond$ for all column vectors $\mathbf{q}_i$ of Q, $\mathbf{q}_i^T\mathbf{q}_j = \delta_{ij}$
- $\diamond$ multiplication by Q preserves length: $\|Q\mathbf{x}\|_2^2 = \mathbf{x}^T Q^T Q\mathbf{x} = \mathbf{x}^T\mathbf{x} = \|x\|_2^2$
- $\diamond$ multiplication by Q also preserves angles
- $\diamond$ rotation and reflection are orthogonal transformations

- A matrix $A \in \mathbf{R}^{m\times n}$ with $m \geq n$ can be written as the product of an orthogonal matrix $Q \in \mathbf{R}^{m\times m}$ and a block upper triangular matrix $R \in \mathbf{R}^{m\times n}$:

$$A = \begin{bmatrix} a_{11} & \dots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \dots & a_{nn} \\ \vdots & & \vdots \\ a_{m1} & \dots & a_{mn} \end{bmatrix} = \begin{bmatrix} q_{11} & \dots & q_{1m} \\ & & \\ \vdots & \ddots & \vdots \\ & & \\ q_{m1} & \dots & q_{mm} \end{bmatrix} \begin{bmatrix} r_{11} & \dots & r_{1n} \\ & \ddots & \vdots \\ & & r_{nn} \\ & & \\ & & \end{bmatrix} = QR$$

This is called the 'full' **QR decomposition** of A. It's useful for fast, numerically stable algorithms, particularly for matrix inversion and least squares estimation.

The 'compact' QR decomposition of $A \in \mathbf{R}^{m\times n}$, $m \geq n$, can also be computed, and yields $Q \in \mathbf{R}^{m\times n}$ and $R \in \mathbf{R}^{n\times n}$, where $Q^TQ = I$ and R is upper triangular:

$$A = \begin{bmatrix} a_{11} & \dots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \dots & a_{nn} \\ \vdots & & \vdots \\ a_{m1} & \dots & a_{mn} \end{bmatrix} = \begin{bmatrix} q_{11} & \dots & q_{1n} \\ & & \\ \vdots & \ddots & \vdots \\ & & \\ q_{m1} & \dots & q_{mn} \end{bmatrix} \begin{bmatrix} r_{11} & \dots & r_{1n} \\ & \ddots & \vdots \\ & & r_{nn} \end{bmatrix} = QR$$

How to produce Q and R? By a series of Householder transformations,

$$Q = \prod_{i=1}^{p} H_i$$

where $p = \min\{m, n\}$.

- For all $\mathbf{u}, \mathbf{v} \in \mathbf{R}^n$ with $\|\mathbf{u}\| = \|\mathbf{v}\|$, there exists a **Householder transformation** $H \in \mathbf{R}^{n\times n}$, where $HH^T = H^TH = I$ (*i.e.* H is orthogonal), such that

$$H\mathbf{u} = \mathbf{v}$$

Householder transformations are a key component of many fast, numerically stable algorithms (*e.g.* for solving systems of linear equations). We can produce the matrix H as follows:

1. define $\mathbf{w} = \mathbf{u} - \mathbf{v}$

2. if $\|\mathbf{w}\|_2 = 0$, then $H = I$
3. if $\|\mathbf{w}\|_2 \neq 0$, then

$$H = I - \left(\frac{2}{\|\mathbf{w}\|_2^2}\right) \mathbf{w}\mathbf{w}^T$$

- The **Cholesky decomposition** of any $P \in \mathbf{S}^n_{++}$ is

$$P = R^T R$$

where $R \in \mathbf{R}^{n \times n}$ is upper triangular and nonsingular.

- The **condition number** of $A \in \mathbf{R}^{n \times n}$ is

$$\text{cond}(A) = \frac{\lambda_{\max}}{\lambda_{\min}}$$

where $\lambda_{\max}$ and $\lambda_{\min}$ are the largest and smallest eigenvalues of A, respectively. The condition number gives an idea of how close A is to being singular. For instance, in double precision arithmetic (where calculations are accurate to 16 digits), $16 - \log \text{cond}(A)$ *useful* digits are preserved when solving $A\mathbf{x} = \mathbf{b}$.

- Let $X \in \mathbf{S}^n$ be the block matrix

$$X = \begin{bmatrix} A & B \\ B^T & C \end{bmatrix}$$

(so A and C are symmetric). The **Schur complement** of A in X is

$$S = C - B^T A^{-1} B$$

Facts:

$$X \succ 0 \iff (A \succ 0 \text{ and } S \succ 0)$$
$$X \succ 0 \iff (C \succ 0 \text{ and } A - BC^{-1}B^T \succ 0)$$
$$A \succ 0 \Rightarrow (X \succeq 0 \iff S \succeq 0)$$
$$C \succ 0 \Rightarrow (X \succeq 0 \iff A - BC^{-1}B^T \succeq 0)$$

The Schur complement is a useful tool for breaking apart quadratic matrix inequalities.

- The 'full' **singular value decomposition** (SVD) of $A \in \mathbf{R}^{m \times n}$ with $\text{rank}(A) = r$ is $A = U\Sigma V^T$, where $U \in \mathbf{R}^{m \times m}$ and $V \in \mathbf{R}^{n \times n}$ are orthogonal and Σ is an $m \times n$ matrix with the singular values $\sigma_1, \ldots, \sigma_r$ of A on the main diagonal:

$$
\begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{1n} & \cdots & a_{nn} \\ \vdots & & \vdots \\ a_{mn} & \cdots & a_{mn} \end{bmatrix} = \begin{bmatrix} u_{11} & \cdots & u_{1m} \\ \vdots & \ddots & \vdots \\ u_{m1} & \cdots & u_{mm} \end{bmatrix} \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_r \\ & & \\ & & \end{bmatrix} \begin{bmatrix} v_{11} & \cdots & v_{1n} \\ \vdots & \ddots & \vdots \\ v_{n1} & \cdots & v_{nn} \end{bmatrix}^T
$$

where $\sigma_1 \geq \cdots \geq \sigma_r \geq 0$.

The full SVD gives a nice interpretation of multiplication by A: to compute $A\mathbf{x}$, first rotate $\mathbf{x}$ to $\mathbf{y} = V^T\mathbf{x}$; then scale components $y_1, \ldots, y_r$ by $\sigma_1, \ldots, \sigma_r$ (and zero out the other components) by computing $\mathbf{z} = \Sigma\mathbf{y}$; then rotate $\mathbf{z}$ to $\mathbf{x} = U\mathbf{z}$. The image of the unit sphere, transformed by A, is an ellipsoid with principle axes $\sigma_i\mathbf{u}_i$.

The SVD is costly to compute, but gives a lot of useful information:

- $r = \operatorname{rank}(A)$ is the number of nonzero elements of Σ (this is how Matlab computes rank)
- $\operatorname{col}(A) = \operatorname{span}(\mathbf{u}_1, \ldots, \mathbf{u}_r)$
- $\operatorname{null}(A) = \operatorname{span}(\mathbf{v}_{r+1}, \ldots, \mathbf{v}_n)$
- $\sigma_1^2, \ldots, \sigma_r^2$ are the eigenvalues of A^TA:

$$
\begin{aligned}
A^TA = (U\Sigma V^T)^T(U\Sigma V^T) &= V\Sigma^T U^T U\Sigma V = V\Sigma^T\Sigma V^T \\
= V \begin{bmatrix} \sigma_1^2 & & \\ & \ddots & \\ & & \sigma_r^2 \end{bmatrix} V^T &= V\Lambda V^T
\end{aligned}
$$

  (which is the spectral decomposition of A^TA)
- the squared Frobenius norm of A is the sum of its squared singular values:

$$
\|A\|_F^2 = \operatorname{tr}(A^TA) = \operatorname{tr}(V\Sigma^T\Sigma V^T) = \sum_{i=1}^{r} \sigma_i^2
$$

As with the QR decomposition, there's a 'compact' form of the SVD: $A = U\Sigma V^T$, where $U \in \mathbf{R}^{m \times r}$, $U^T U = I$, $V \in \mathbf{R}^{n \times r}$, $V^T V = I$, and $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_r)$

$$\begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{1n} & \cdots & a_{nn} \\ \vdots & & \vdots \\ a_{mn} & \cdots & a_{mn} \end{bmatrix} = \begin{bmatrix} u_{11} & \cdots & u_{1r} \\ \vdots & & \vdots \\ u_{m1} & \cdots & u_{mr} \end{bmatrix} \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_r \end{bmatrix} \begin{bmatrix} v_{11} & \cdots & v_{1r} \\ \vdots & \ddots & \vdots \\ v_{n1} & \cdots & v_{nr} \end{bmatrix}^T$$

5 Derivatives

- The **gradient** of a function $f : \mathbf{R}^n \to \mathbf{R}$ with respect to a vector $\mathbf{v} \in \mathbf{R}^n$ is

$$\nabla f(\mathbf{v}) = \begin{bmatrix} \frac{\partial f}{\partial v_1} \\ \vdots \\ \frac{\partial f}{\partial v_n} \end{bmatrix} \in \mathbf{R}^n$$

The gradient of a vector-valued function $\mathbf{f} : \mathbf{R}^n \to \mathbf{R}^m$ is

$$\mathbf{f}(\mathbf{v}) = \begin{bmatrix} f_1(\mathbf{v}) \\ \vdots \\ f_m(\mathbf{v}) \end{bmatrix} \quad \Rightarrow \quad \nabla \mathbf{f}(\mathbf{v}) = \begin{bmatrix} \nabla f_1(\mathbf{v}) & \dots & \nabla f_m(\mathbf{v}) \end{bmatrix} = \begin{bmatrix} \frac{\partial f_1}{\partial v_1} & \dots & \frac{\partial f_m}{\partial v_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_1}{\partial v_n} & \dots & \frac{\partial f_m}{\partial v_n} \end{bmatrix} \in \mathbf{R}^{n \times m}$$

- **Alternate notation:** we sometimes want to work with rows of derivatives instead of columns Why? Because it can make vector-matrix derivatives look more like their scalar analogs. This motivates the notation

$$Df(\mathbf{v}) = \frac{\partial f}{\partial \mathbf{v}} = \nabla f(\mathbf{v})^T = \begin{bmatrix} \frac{\partial f}{\partial v_1} & \dots & \frac{\partial f}{\partial v_n} \end{bmatrix} \in \mathbf{R}^{1 \times n}$$

$$D\mathbf{f}(\mathbf{v}) = \frac{\partial \mathbf{f}}{\partial \mathbf{v}} = \nabla \mathbf{f}(\mathbf{v})^T = \begin{bmatrix} \nabla f_1(\mathbf{v})^T \\ \vdots \\ \nabla f_m(\mathbf{v})^T \end{bmatrix} = \begin{bmatrix} \frac{\partial f_1}{\partial v_1} & \dots & \frac{\partial f_1}{\partial v_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial v_1} & \dots & \frac{\partial f_m}{\partial v_n} \end{bmatrix} \in \mathbf{R}^{m \times n}$$

In either case, $D\mathbf{f}(\mathbf{v})$ is called the **Jacobian matrix** of $\mathbf{f}$ with respect to $\mathbf{v}$.

- **Chain rule**. Let $\mathbf{f} : \mathbf{R}^p \to \mathbf{R}^m$, $\mathbf{g} : \mathbf{R}^n \to \mathbf{R}^p$ and $\mathbf{h} = \mathbf{f}(\mathbf{g}(\mathbf{x}))$ (so $\mathbf{h} : \mathbf{R}^n \to \mathbf{R}^m$). Then

$$\nabla \mathbf{h}(\mathbf{x}) = \nabla \mathbf{g}(\mathbf{x}) \nabla \mathbf{f}(\mathbf{g}(\mathbf{x})) \in \mathbf{R}^{n \times m}$$
$$D\mathbf{h}(\mathbf{x}) = D\mathbf{f}(\mathbf{g}(\mathbf{x})) D\mathbf{g}(\mathbf{x}) \in \mathbf{R}^{m \times n}$$

- The first-order Taylor expansion of $\mathbf{f} : \mathbf{R}^n \to \mathbf{R}^m$, for $\mathbf{x}$ near $\mathbf{x_0}$, is

$$\mathbf{f}(\mathbf{x}) \approx \mathbf{f}(\mathbf{x_0}) + \nabla \mathbf{f}(\mathbf{x_0})^T (\mathbf{x} - \mathbf{x_0})$$

- The **Hessian matrix** is a generalized second derivative. For $f : \mathbf{R}^n \to \mathbf{R}$,

$$\nabla^2 f(\mathbf{v}) = \frac{\partial^2 f}{\partial \mathbf{v}^2} = \begin{bmatrix} \frac{\partial^2 f}{\partial v_1^2} & \dots & \frac{\partial^2 f}{\partial v_1 \partial v_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial v_n \partial v_1} & \dots & \frac{\partial^2 f}{\partial v_n^2} \end{bmatrix} \in \mathbf{R}^{n \times n}$$

Note that the equivalence of cross-partials implies that the Hessian is symmetric.

- The second-order Taylor expansion of $\mathbf{f} : \mathbf{R}^n \to \mathbf{R}^m$, for $\mathbf{x}$ near $\mathbf{x_0}$, is

$$\mathbf{f}(\mathbf{x}) \approx \mathbf{f}(\mathbf{x_0}) + \nabla \mathbf{f}(\mathbf{x_0})^T(\mathbf{x} - \mathbf{x_0}) + \frac{1}{2}(\mathbf{x} - \mathbf{x_0})^T \nabla^2 \mathbf{f}(\mathbf{x_0})(\mathbf{x} - \mathbf{x_0})$$

- **Mean value theorem.** If $f : \mathbf{R}^n \to \mathbf{R}$ is continuously differentiable, then for any $\mathbf{x}, \mathbf{y} \in \mathbf{R}^n$ there exists a $\lambda \in [0, 1]$ such that

$$f(\mathbf{y}) - f(\mathbf{x}) = \nabla f(\lambda \mathbf{x} + (1 - \lambda)\mathbf{y})^T(\mathbf{y} - \mathbf{x})$$

- **Generalized mean value theorem.** If $f : \mathbf{R}^n \to \mathbf{R}$ is twice continuously differentiable, then for any $\mathbf{x}, \mathbf{y} \in \mathbf{R}^n$ there exists a $\lambda \in [0, 1]$ such that

$$f(\mathbf{y}) - f(\mathbf{x}) = \nabla f(\mathbf{x})^T(\mathbf{y} - \mathbf{x}) + \frac{1}{2}(\mathbf{y} - \mathbf{x})^T H(\lambda \mathbf{x} + (1 - \lambda)\mathbf{y})(\mathbf{y} - \mathbf{x})$$

- **Derivatives of quadratic forms.** Let $f(\mathbf{v}) = \frac{1}{2}\mathbf{v}^T P \mathbf{v} + \mathbf{g}^T \mathbf{v}$, where $P \in \mathbf{S}^n_{++}$ and $\mathbf{g}, \mathbf{v} \in \mathbf{R}^n$. Then

$$\nabla f(\mathbf{v}) = P\mathbf{v} + \mathbf{g} \quad \text{and} \quad \nabla^2 f(\mathbf{v}) = P$$

- **Derivative of matrix inverse.** Let $A \in \mathbf{R}^{n \times n}$ be nonsingular and depend on a parameter t. Then

$$\frac{\mathrm{d}}{\mathrm{d}t} A^{-1} = -A^{-1} \Big(\frac{\mathrm{d}}{\mathrm{d}t} A \Big) A^{-1}$$