Linear Algebra Primer

And a video discussion of linear algebra from EE263 is here (lectures 3 and 4):

https://see.stanford.edu/Course/EE263

Vector spaces

a *vector space* or *linear space* (over the reals) consists of

- a set $\mathcal{V}$
- a vector sum $+ : \mathcal{V} \times \mathcal{V} \to \mathcal{V}$
- a scalar multiplication : $\mathbb{R} \times \mathcal{V} \to \mathcal{V}$
- a distinguished element $0 \in \mathcal{V}$

which satisfy a list of properties

Vector space axioms

- $x + y = y + x$, $\forall x, y \in \mathcal{V}$ — *$+$ is commutative*
- $(x + y) + z = x + (y + z)$, $\quad \forall x, y, z \in \mathcal{V}$ — *$+$ is associative*
- $0 + x = x$, $\forall x \in \mathcal{V}$ — *0 is additive identity*
- $\forall x \in \mathcal{V} \;\; \exists (-x) \in \mathcal{V}$ s.t. $x + (-x) = 0$ — *existence of additive inverse*
- $(\alpha\beta)x = \alpha(\beta x)$, $\quad \forall \alpha, \beta \in \mathbb{R} \quad \forall x \in \mathcal{V}$ — *scalar mult. is associative*
- $\alpha(x + y) = \alpha x + \alpha y$, $\quad \forall \alpha \in \mathbb{R} \;\; \forall x, y \in \mathcal{V}$ — *right distributive rule*
- $(\alpha + \beta)x = \alpha x + \beta x$, $\quad \forall \alpha, \beta \in \mathbb{R} \;\; \forall x \in \mathcal{V}$ — *left distributive rule*
- $1x = x$, $\quad \forall x \in \mathcal{V}$ — *1 is multiplicative identity*

Examples

- $\mathcal{V}_1 = \mathbb{R}^n$, with standard (componentwise) vector addition and scalar multiplication
- $\mathcal{V}_2 = \{0\}$ (where $0 \in \mathbb{R}^n$)
- $\mathcal{V}_3 = \textbf{span}(v_1, v_2, \ldots, v_k)$ where

$$\textbf{span}(v_1, v_2, \ldots, v_k) = \{\alpha_1 v_1 + \cdots + \alpha_k v_k \mid \alpha_i \in \mathbb{R}\}$$

  and $v_1, \ldots, v_k \in \mathbb{R}^n$

Subspaces

- a *subspace* of a vector space is a *subset* of a vector space which is itself a vector space
- roughly speaking, a subspace is closed under vector addition and scalar multiplication
- examples $\mathcal{V}_1$, $\mathcal{V}_2$, $\mathcal{V}_3$ above are subspaces of $\mathbb{R}^n$

Vector spaces of functions

- $\mathcal{V}_4 = \{x : \mathbb{R}_+ \to \mathbb{R}^n \mid x \text{ is differentiable}\}$, where vector sum is sum of functions:

  $$(x + z)(t) = x(t) + z(t)$$

  and scalar multiplication is defined by

  $$(\alpha x)(t) = \alpha x(t)$$

  (a *point* in $\mathcal{V}_4$ is a *trajectory* in $\mathbb{R}^n$)

- $\mathcal{V}_5 = \{x \in \mathcal{V}_4 \mid \dot{x} = Ax\}$
  (*points* in $\mathcal{V}_5$ are *trajectories* of the linear system $\dot{x} = Ax$)

- $\mathcal{V}_5$ is a subspace of $\mathcal{V}_4$

Basis and dimension

set of vectors $\{v_1, v_2, \ldots, v_k\}$ is called a *basis* for a vector space $\mathcal{V}$ if

$$\mathcal{V} = \textbf{span}(v_1, v_2, \ldots, v_k)$$
and
$\{v_1, v_2, \ldots, v_k\}$ is independent

- equivalently, every $v \in \mathcal{V}$ *can be uniquely* expressed as

$$v = \alpha_1 v_1 + \cdots + \alpha_k v_k$$

- for a given vector space $\mathcal{V}$, the number of vectors in any basis is the same
- number of vectors in any basis is called the *dimension* of $\mathcal{V}$, denoted $\textbf{dim}\,\mathcal{V}$

Nullspace of a matrix

the *nullspace* of $A \in \mathbb{R}^{m \times n}$ is defined as

$$\textbf{null}(A) = \{\, x \in \mathbb{R}^n \mid Ax = 0 \,\}$$

- $\textbf{null}(A)$ is set of vectors mapped to zero by $y = Ax$
- $\textbf{null}(A)$ is set of vectors orthogonal to all rows of A

$\textbf{null}(A)$ gives *ambiguity* in x given $y = Ax$:

- if $y = Ax$ and $z \in \textbf{null}(A)$, then $y = A(x + z)$
- conversely, if $y = Ax$ and $y = A\tilde{x}$, then $\tilde{x} = x + z$ for some $z \in \textbf{null}(A)$

$\textbf{null}(A)$ is also written $\mathcal{N}(A)$

Zero nullspace

A is called *one-to-one* if 0 is the only element of its nullspace

$$\mathbf{null}(A) = \{0\}$$

Equivalently,

- x can always be uniquely determined from $y = Ax$
  ($i.e.$, the linear transformation $y = Ax$ doesn't 'lose' information)
- mapping from x to Ax is one-to-one: different x's map to different y's
- columns of A are independent (hence, a basis for their span)
- A has a *left inverse*, $i.e.$, there is a matrix $B \in \mathbb{R}^{n \times m}$ s.t. $BA = I$
- $A^{\mathsf{T}}A$ is invertible

Range of a matrix

the *range* of $A \in \mathbb{R}^{m \times n}$ is defined as

$$\textbf{range}(A) = \{Ax \mid x \in \mathbb{R}^n\} \subseteq \mathbb{R}^m$$

range(A) can be interpreted as

- the set of vectors that can be 'hit' by linear mapping $y = Ax$
- the span of columns of A
- the set of vectors y for which $Ax = y$ has a solution

range(A) is also written $\mathcal{R}(A)$

Outline

- Vectors and matrices
  - Basic Matrix Operations
  - Determinants, norms, trace
  - Special Matrices
- Transformation Matrices
  - Homogeneous coordinates
  - Translation
- Matrix inverse
- Matrix rank
- Eigenvalues and Eigenvectors
- Matrix Calculus

Outline

- Vectors and matrices
  - Basic Matrix Operations
  - Determinants, norms, trace
  - Special Matrices
- Transformation Matrices
  - Homogeneous coordinates
  - Translation
- Matrix inverse
- Matrix rank
- Eigenvalues and Eigenvectors(SVD)
- Matrix Calculus

Eigenvector and Eigenvalue

- An eigenvector **x** of a linear transformation *A* is a non-zero vector that, when *A* is applied to it, does not change direction.

$$Ax = \lambda x, \quad x \neq 0.$$

Eigenvector and Eigenvalue

- An eigenvector **x** of a linear transformation *A* is a non-zero vector that, when *A* is applied to it, does not change direction.
- Applying *A* to the eigenvector only scales the eigenvector by the scalar value λ, called an eigenvalue.

$$Ax = \lambda x, \quad x \neq 0.$$

Eigenvector and Eigenvalue

- We want to find all the eigenvalues of A:

$$Ax = \lambda x, \quad x \neq 0.$$

- Which can we written as:

$$Ax = (\lambda I)x \quad x \neq 0.$$

- Therefore:

$$(\lambda I - A)x = 0, \quad x \neq 0.$$

Eigenvector and Eigenvalue

- We can solve for eigenvalues by solving:

$$(\lambda I - A)x = 0, \quad x \neq 0.$$

- Since we are looking for non-zero **x**, we can instead solve the above equation as:

$$|(\lambda I - A)| = 0.$$

Spectral theory

- We call an eigenvalue λ and an associated eigenvector an **eigenpair**.
- The space of vectors where (A − λI) = 0 is often called the **eigenspace** of A associated with the eigenvalue λ.
- The set of all eigenvalues of A is called its **spectrum**:

$$\sigma(A) = \{\lambda \in \mathbb{C} : \lambda I - A \text{ is singular}\}.$$

Diagonalization

- An n × n matrix A is diagonalizable if it has n linearly independent eigenvectors.
- Most square matrices (in a sense that can be made mathematically rigorous) are diagonalizable:
  - Normal matrices are diagonalizable
  - Matrices with n distinct eigenvalues are diagonalizable

Lemma: Eigenvectors associated with distinct eigenvalues are linearly independent.

Diagonalization

- An n × n matrix A is diagonalizable if it has n linearly independent eigenvectors.
- Most square matrices are diagonalizable:
  - Normal matrices are diagonalizable
  - Matrices with n distinct eigenvalues are diagonalizable

Lemma: Eigenvectors associated with distinct eigenvalues are linearly independent.

Diagonalization

- Eigenvalue equation:

$$AV = VD$$

$$A = VDV^{-1}$$

  - Where D is a diagonal matrix of the eigenvalues

$$\begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix}$$

Diagonalization

- Eigenvalue equation:

$$AV = VD$$

$$A = VDV^{-1}$$

- Assuming all λ_i's are unique:

$$A = VDV^{T}$$

  - Remember that the inverse of an orthogonal matrix is just its transpose and the eigenvectors are orthogonal

Symmetric matrices

- Properties:
  - For a symmetric matrix A, all the eigenvalues are real.
  - The eigenvectors of A are orthonormal.

$$A = V D V^T$$

Symmetric matrices

- Therefore:

$$x^T A x = x^T V D V^T x = y^T D y = \sum_{i=1}^{n} \lambda_i y_i^2$$

  – where $y = V^T x$

- So, if we wanted to find the vector x that:

$$\max_{x \in \mathbb{R}^n} \; x^T A x \quad \text{subject to } \|x\|_2^2 = 1$$

Symmetric matrices

- Therefore:

$$x^T A x = x^T V D V^T x = y^T D y = \sum_{i=1}^{n} \lambda_i y_i^2$$

  - where $y = V^T x$

- So, if we wanted to find the vector x that:

$$\max_{x \in \mathbb{R}^n} \; x^T A x \quad \text{subject to } \|x\|_2^2 = 1$$

  - Is the same as finding the eigenvector that corresponds to the largest eigenvalue.

Properties

- The trace of a A is equal to the sum of its eigenvalues:

$$\mathrm{tr}A = \sum_{i=1}^{n} \lambda_i.$$

- The determinant of A is equal to the product of its eigenvalues

$$|A| = \prod_{i=1}^{n} \lambda_i.$$

Properties

- The trace of a A is equal to the sum of its eigenvalues:

$$\mathrm{tr}A = \sum_{i=1}^{n} \lambda_i.$$

- The determinant of A is equal to the product of its eigenvalues

$$|A| = \prod_{i=1}^{n} \lambda_i.$$

- The rank of A is equal to the number of non-zero eigenvalues of A.

Properties

- The trace of a A is equal to the sum of its eigenvalues:

$$\mathrm{tr}A = \sum_{i=1}^{n} \lambda_i.$$

- The determinant of A is equal to the product of its eigenvalues

$$|A| = \prod_{i=1}^{n} \lambda_i.$$

- The rank of A is equal to the number of non-zero eigenvalues of A.
- The eigenvalues of a diagonal matrix D = diag(d1, . . . dn) are just the diagonal entries d1, . . . dn

Spectral theory

- We call an eigenvalue λ and an associated eigenvector an **eigenpair**.
- The space of vectors where (A − λI) = 0 is often called the **eigenspace** of A associated with the eigenvalue λ.
- The set of all eigenvalues of A is called its **spectrum**:

$$\sigma(A) = \{\lambda \in \mathbb{C} : \lambda I - A \text{ is singular}\}.$$

Spectral theory

- The magnitude of the largest eigenvalue (in magnitude) is called the spectral radius

$$\rho(A) = \max\{|\lambda_1|, \ldots, |\lambda_n|\}$$

  - Where C is the space of all eigenvalues of A

Spectral theory

- The spectral radius is bounded by infinity norm of a matrix: $\rho(A) = \lim_{k \to \infty} \|A^k\|^{1/k}$
- Proof: Turn to a partner and prove this!

Spectral theory

- The spectral radius is bounded by infinity norm of a matrix: $\rho(A) = \lim_{k\to\infty} \|A^k\|^{1/k}$
- Proof: Let λ and v be an eigenpair of A:

$$|\lambda|^k \|\mathbf{v}\| = \|\lambda^k \mathbf{v}\| = \|A^k \mathbf{v}\| \leq \|A^k\| \cdot \|\mathbf{v}\|$$

and since $\mathbf{v} \neq 0$ we have

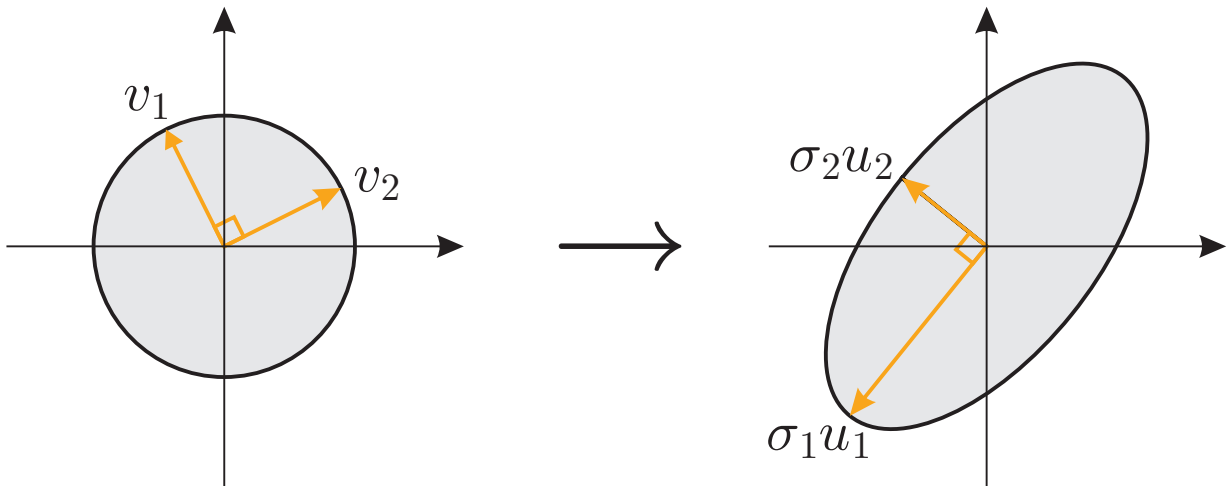
$$|\lambda|^k \leq \|A^k\|$$

and therefore

$$\rho(A) \leq \|A^k\|^{\frac{1}{k}}.$$

Singular Value Decomposition

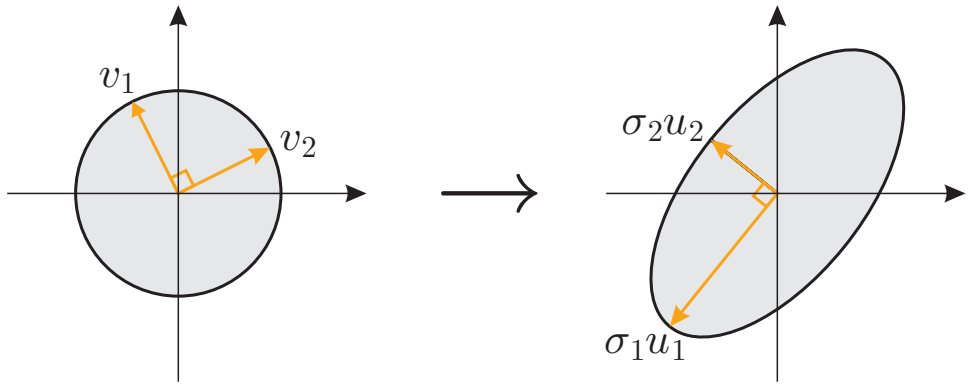
Geometry of linear maps



every matrix $A \in \mathbb{R}^{m \times n}$ maps the unit ball in $\mathbb{R}^n$ to an ellipsoid in $\mathbb{R}^m$

$$S = \left\{ x \in \mathbb{R}^n \;\middle|\; \|x\| \leq 1 \right\} \qquad AS = \left\{ Ax \;\middle|\; x \in S \right\}$$

Singular values and singular vectors



- first, assume $A \in \mathbb{R}^{m \times n}$ is skinny and full rank
- the numbers $\sigma_1, \ldots, \sigma_n > 0$ are called the *singular values* of A
- the vectors $u_1, \ldots, u_n$ are called the *left* or *output singular vectors* of A. These are *unit vectors* along the principal semiaxes of AS
- the vectors $v_1, \ldots, v_n$ are called the *right* or *input singular vectors* of A. These map to the principal semiaxes, so that

$$Av_i = \sigma_i u_i$$

Thin singular value decomposition

$$Av_i = \sigma_i u_i \qquad \text{for } 1 \leq i \leq n$$

For $A \in \mathbb{R}^{m \times n}$ with **Rank**$(A) = n$, let

$$U = \begin{bmatrix} u_1 & u_2 & \cdots & u_n \end{bmatrix} \qquad \Sigma = \begin{bmatrix} \sigma_1 & & & \\ & \sigma_2 & & \\ & & \ddots & \\ & & & \sigma_n \end{bmatrix} \qquad V = \begin{bmatrix} v_1 & v_2 & \cdots & v_n \end{bmatrix}$$

the above equation is $AV = U\Sigma$ and since V is orthogonal

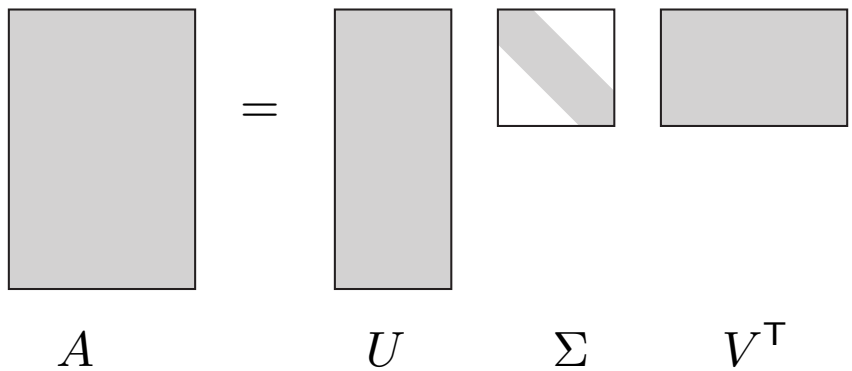
$$A = U\Sigma V^{\mathsf{T}}$$

called the *thin SVD* of A

Thin SVD

For $A \in \mathbb{R}^{m \times n}$ with $\mathbf{Rank}(A) = r$, the *thin SVD* is

$$A = U \Sigma V^{\mathsf{T}} = \sum_{i=1}^{r} \sigma_i u_i v_i^{\mathsf{T}}$$



here

- $U \in \mathbb{R}^{m \times r}$ has orthonormal columns,
- $\Sigma = \mathbf{diag}(\sigma_1, \ldots, \sigma_r)$, where $\sigma_1 \geq \cdots \geq \sigma_r > 0$
- $V \in \mathbb{R}^{n \times r}$ has orthonormal columns

SVD and eigenvectors

$$A^\mathsf{T} A = (U\Sigma V^\mathsf{T})^\mathsf{T}(U\Sigma V^\mathsf{T}) = V\Sigma^2 V^\mathsf{T}$$

hence:

- v_i are eigenvectors of $A^\mathsf{T} A$ (corresponding to nonzero eigenvalues)
- $\sigma_i = \sqrt{\lambda_i(A^\mathsf{T} A)}$ (and $\lambda_i(A^\mathsf{T} A) = 0$ for $i > r$)
- $\|A\| = \sigma_1$

SVD and eigenvectors

similarly,

$$AA^\mathsf{T} = (U\Sigma V^\mathsf{T})(U\Sigma V^\mathsf{T})^\mathsf{T} = U\Sigma^2 U^\mathsf{T}$$

hence:

- u_i are eigenvectors of AA^T (corresponding to nonzero eigenvalues)
- $\sigma_i = \sqrt{\lambda_i(AA^\mathsf{T})}$ (and $\lambda_i(AA^\mathsf{T}) = 0$ for $i > r$)

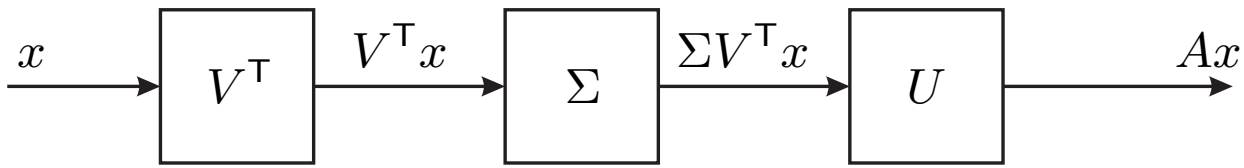
SVD and range

$$A = U\Sigma V^{\mathsf{T}}$$

- $u_1, \ldots u_r$ are orthonormal basis for **range**(A)
- $v_1, \ldots v_r$ are orthonormal basis for **null**$(A)^{\perp}$

Interpretations

$$A = U\Sigma V^\mathsf{T} = \sum_{i=1}^{r} \sigma_i u_i v_i^\mathsf{T}$$



linear mapping $y = Ax$ can be decomposed as

- compute coefficients of x along input directions $v_1, \ldots, v_r$
- scale coefficients by σ_i
- reconstitute along output directions $u_1, \ldots, u_r$

difference with eigenvalue decomposition for symmetric A: input and output directions are *different*

General pseudo-inverse

if $A \neq 0$ has SVD $A = U\Sigma V^{\mathsf{T}}$, the *pseudo-inverse* or *Moore-Penrose inverse* of A is

$$A^{\dagger} = V\Sigma^{-1}U^{\mathsf{T}}$$

- if A is skinny and full rank,

  $$A^{\dagger} = (A^{\mathsf{T}}A)^{-1}A^{\mathsf{T}}$$

  gives the least-squares approximate solution $x_{\mathrm{ls}} = A^{\dagger}y$
- if A is fat and full rank,

  $$A^{\dagger} = A^{\mathsf{T}}(AA^{\mathsf{T}})^{-1}$$

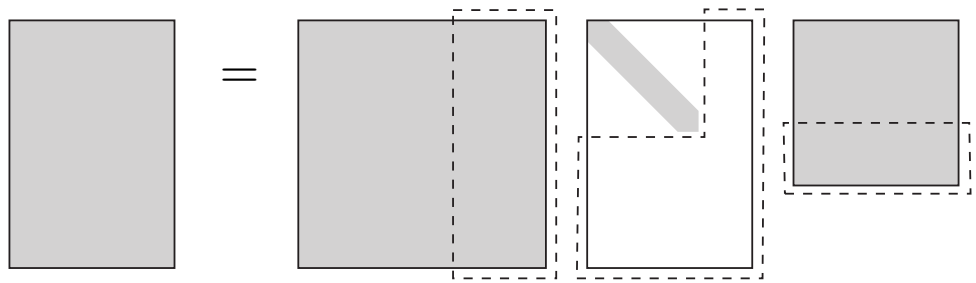
  gives the least-norm solution $x_{\mathrm{ln}} = A^{\dagger}y$

Full SVD

SVD of $A \in \mathbb{R}^{m \times n}$ with **Rank**$(A) = r$

$$A = U_1 \Sigma_1 V_1^{\mathsf{T}} = \begin{bmatrix} u_1 & \cdots & u_r \end{bmatrix} \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_r \end{bmatrix} \begin{bmatrix} v_1^{\mathsf{T}} \\ \vdots \\ v_r^{\mathsf{T}} \end{bmatrix}$$

Add extra columns to U and V, and add zero rows/cols to Σ_1



Full SVD

- find $U_2 \in \mathbb{R}^{m \times (m-r)}$ such that $U = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \in \mathbb{R}^{m \times m}$ is orthogonal
- find $V_2 \in \mathbb{R}^{n \times (n-r)}$ such that $V = \begin{bmatrix} V_1 & V_2 \end{bmatrix} \in \mathbb{R}^{n \times n}$ is orthogonal
- add zero rows/cols to Σ_1 to form $\Sigma \in \mathbb{R}^{m \times n}$

$$\Sigma = \left[\begin{array}{c|c} \Sigma_1 & 0_{r \times (n-r)} \\ \hline 0_{(m-r) \times r} & 0_{(m-r) \times (n-r)} \end{array}\right]$$

then the full SVD is

$$A = U_1 \Sigma_1 V_1^{\mathsf{T}} = \left[\begin{array}{c|c} U_1 & U_2 \end{array}\right] \left[\begin{array}{c|c} \Sigma_1 & 0_{r \times (n-r)} \\ \hline 0_{(m-r) \times r} & 0_{(m-r) \times (n-r)} \end{array}\right] \left[\begin{array}{c} V_1^{\mathsf{T}} \\ \hline V_2^{\mathsf{T}} \end{array}\right]$$

which is $A = U \Sigma V^{\mathsf{T}}$

example: SVD

$$A = \begin{bmatrix} 1 & 2 \\ 3 & 1 \\ 4 & 2 \end{bmatrix}$$

SVD is

$$A = \begin{bmatrix} -0.319 & 0.915 & -0.248 \\ -0.542 & -0.391 & -0.744 \\ -0.778 & -0.103 & 0.620 \end{bmatrix} \begin{bmatrix} 5.747 & 0 \\ 0 & 1.403 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} -0.880 & -0.476 \\ -0.476 & 0.880 \end{bmatrix}$$

Image of unit ball under linear transformation

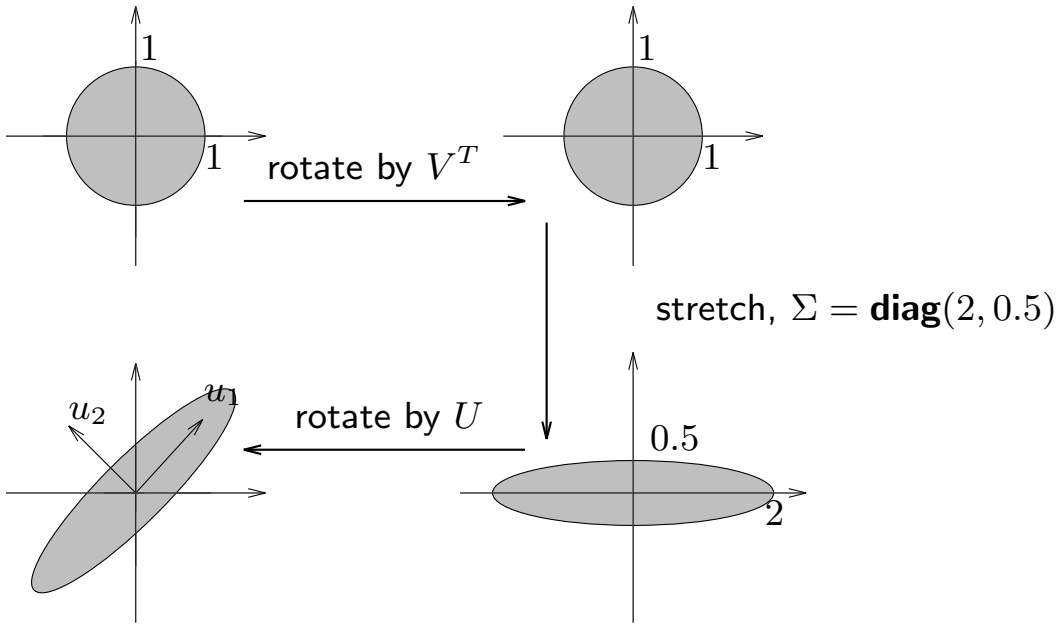
full SVD:

$$A = U\Sigma V^{\mathsf{T}}$$

gives intepretation of $y = Ax$:

- rotate (by V^{T})
- stretch along axes by σ_i ($\sigma_i = 0$ for $i > r$)
- zero-pad (if $m > n$) or truncate (if $m < n$) to get m-vector
- rotate (by U)

Image of unit ball under A



$\{Ax \mid \|x\| \leq 1\}$ is *ellipsoid* with principal axes $\sigma_i u_i$.

Some applications of Eigenvalues

- PageRank
- Schrodinger's equation
- PCA

Outline

- Vectors and matrices
  - Basic Matrix Operations
  - Determinants, norms, trace
  - Special Matrices
- Transformation Matrices
  - Homogeneous coordinates
  - Translation
- Matrix inverse
- Matrix rank
- Eigenvalues and Eigenvectors(SVD)
- Matrix Calculus

Matrix Calculus – The Gradient

- Let a function $f : \mathbb{R}^{m \times n} \rightarrow \mathbb{R}$ take as input a matrix A of size m × n and returns a real value.
- Then the **gradient** of **f:**

$$\nabla_A f(A) \in \mathbb{R}^{m \times n} = \begin{bmatrix} \frac{\partial f(A)}{\partial A_{11}} & \frac{\partial f(A)}{\partial A_{12}} & \cdots & \frac{\partial f(A)}{\partial A_{1n}} \\ \frac{\partial f(A)}{\partial A_{21}} & \frac{\partial f(A)}{\partial A_{22}} & \cdots & \frac{\partial f(A)}{\partial A_{2n}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f(A)}{\partial A_{m1}} & \frac{\partial f(A)}{\partial A_{m2}} & \cdots & \frac{\partial f(A)}{\partial A_{mn}} \end{bmatrix}$$

Matrix Calculus – The Gradient

- Every entry in the matrix is: $\nabla_A f(A))_{ij} = \frac{\partial f(A)}{\partial A_{ij}}.$
- the size of $\nabla_A f(A)$ is always the same as the size of A. So if A is just a vector x:

$$\nabla_x f(x) = \begin{bmatrix} \frac{\partial f(x)}{\partial x_1} \\ \frac{\partial f(x)}{\partial x_2} \\ \vdots \\ \frac{\partial f(x)}{\partial x_n} \end{bmatrix}$$

Exercise

- Example:

For $x \in \mathbb{R}^n$, let $f(x) = b^T x$ for some known vector $b \in \mathbb{R}^n$

$$f(x) = \begin{bmatrix} b_1 & b_2 & \dots & b_n \end{bmatrix}^T \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

- Find: $\dfrac{\partial f(x)}{\partial x_k} = ?$

$$\nabla_x f(x) = ?$$

Exercise

- Example:

For $x \in \mathbb{R}^n$, let $f(x) = b^T x$ for some known vector $b \in \mathbb{R}^n$

$$f(x) = \sum_{i=1}^{n} b_i x_i$$

$$\frac{\partial f(x)}{\partial x_k} = \frac{\partial}{\partial x_k} \sum_{i=1}^{n} b_i x_i = b_k.$$

- From this we can conclude that: $\nabla_x b^T x = b$

Matrix Calculus – The Gradient

- Properties
  - $\nabla_x(f(x) + g(x)) = \nabla_x f(x) + \nabla_x g(x).$
  - For $t \in \mathbb{R}$, $\nabla_x(t\ f(x)) = t\nabla_x f(x).$

Matrix Calculus – The Hessian

- The Hessian matrix with respect to x, written $\nabla_x^2 f(x)$ or simply as H is the n × n matrix of partial derivatives

$$\nabla_x^2 f(x) \in \mathbb{R}^{n \times n} = \begin{bmatrix} \frac{\partial^2 f(x)}{\partial x_1^2} & \frac{\partial^2 f(x)}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f(x)}{\partial x_2 \partial x_1} & \frac{\partial^2 f(x)}{\partial x_2^2} & \cdots & \frac{\partial^2 f(x)}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f(x)}{\partial x_n \partial x_1} & \frac{\partial^2 f(x)}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_n^2} \end{bmatrix}$$

Matrix Calculus – The Hessian

- Each entry can be written as: $\nabla_x^2 f(x))_{ij} = \frac{\partial^2 f(x)}{\partial x_i \partial x_j}$

- Exercise: Why is the Hessian always symmetric?

Matrix Calculus – The Hessian

- Each entry can be written as: $\nabla_x^2 f(x))_{ij} = \frac{\partial^2 f(x)}{\partial x_i \partial x_j}$.

- The Hessian is always symmetric, because

$$\frac{\partial^2 f(x)}{\partial x_i \partial x_j} = \frac{\partial^2 f(x)}{\partial x_j \partial x_i}.$$

- This is known as Schwarz's theorem: The order of partial derivatives don’t matter as long as the second derivative exists and is continuous.

Matrix Calculus – The Hessian

- Note that the hessian is not the gradient of whole gradient of a vector (this is not defined). It is actually the gradient of **every entry** of the gradient of the vector.

$$\nabla_x^2 f(x) \in \mathbb{R}^{n \times n} = \begin{bmatrix} \frac{\partial^2 f(x)}{\partial x_1^2} & \frac{\partial^2 f(x)}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f(x)}{\partial x_2 \partial x_1} & \frac{\partial^2 f(x)}{\partial x_2^2} & \cdots & \frac{\partial^2 f(x)}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f(x)}{\partial x_n \partial x_1} & \frac{\partial^2 f(x)}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_n^2} \end{bmatrix}$$

Matrix Calculus – The Hessian

- Eg, the first column is the gradient of $\dfrac{\partial f(x)}{\partial x_1}$

$$\nabla_x^2 f(x) \in \mathbb{R}^{n \times n} = \begin{bmatrix} \dfrac{\partial^2 f(x)}{\partial x_1^2} & \dfrac{\partial^2 f(x)}{\partial x_1 \partial x_2} & \cdots & \dfrac{\partial^2 f(x)}{\partial x_1 \partial x_n} \\ \dfrac{\partial^2 f(x)}{\partial x_2 \partial x_1} & \dfrac{\partial^2 f(x)}{\partial x_2^2} & \cdots & \dfrac{\partial^2 f(x)}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{\partial^2 f(x)}{\partial x_n \partial x_1} & \dfrac{\partial^2 f(x)}{\partial x_n \partial x_2} & \cdots & \dfrac{\partial^2 f(x)}{\partial x_n^2} \end{bmatrix}$$

Exercise

- Example:

consider the quadratic function $f(x) = x^T Ax$

$$f(x) = \sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij} x_i x_j$$

$$\frac{\partial f(x)}{\partial x_k} = \frac{\partial}{\partial x_k} \sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij} x_i x_j$$

Exercise

$$\frac{\partial f(x)}{\partial x_k} = \frac{\partial}{\partial x_k} \sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij} x_i x_j$$

Exercise

$$\frac{\partial f(x)}{\partial x_k} = \frac{\partial}{\partial x_k} \sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij} x_i x_j$$

$$= \frac{\partial}{\partial x_k} \left[\sum_{i \neq k} \sum_{j \neq k} A_{ij} x_i x_j + \sum_{i \neq k} A_{ik} x_i x_k + \sum_{j \neq k} A_{kj} x_k x_j + A_{kk} x_k^2 \right]$$

Divide the summation into 3 parts depending on whether:

- i == k or
- j == k

Exercise

$$\begin{aligned}
\frac{\partial f(x)}{\partial x_k} &= \frac{\partial}{\partial x_k} \sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij} x_i x_j \\
&= \frac{\partial}{\partial x_k} \left[\sum_{i \neq k} \sum_{j \neq k} A_{ij} x_i x_j + \sum_{i \neq k} A_{ik} x_i x_k + \sum_{j \neq k} A_{kj} x_k x_j + A_{kk} x_k^2 \right] \\
&= \sum_{i \neq k} A_{ik} x_i + \sum_{j \neq k} A_{kj} x_j + 2 A_{kk} x_k
\end{aligned}$$

Exercise

$$\frac{\partial f(x)}{\partial x_k} = \frac{\partial}{\partial x_k} \sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij} x_i x_j$$

$$= \frac{\partial}{\partial x_k} \left[\sum_{i \neq k} \sum_{j \neq k} A_{ij} x_i x_j + \sum_{i \neq k} A_{ik} x_i x_k + \sum_{j \neq k} A_{kj} x_k x_j + A_{kk} x_k^2 \right]$$

$$= \sum_{i \neq k} A_{ik} x_i + \sum_{j \neq k} A_{kj} x_j + 2 A_{kk} x_k$$

Exercise

$$\frac{\partial f(x)}{\partial x_k} = \frac{\partial}{\partial x_k} \sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij} x_i x_j$$

$$= \frac{\partial}{\partial x_k} \left[\sum_{i \neq k} \sum_{j \neq k} A_{ij} x_i x_j + \sum_{i \neq k} A_{ik} x_i x_k + \sum_{j \neq k} A_{kj} x_k x_j + A_{kk} x_k^2 \right]$$

$$= \sum_{i \neq k} A_{ik} x_i + \sum_{j \neq k} A_{kj} x_j + 2 A_{kk} x_k$$

Exercise

$$\frac{\partial f(x)}{\partial x_k} = \frac{\partial}{\partial x_k} \sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij} x_i x_j$$

$$= \frac{\partial}{\partial x_k} \left[\sum_{i \neq k} \sum_{j \neq k} A_{ij} x_i x_j + \sum_{i \neq k} A_{ik} x_i x_k + \sum_{j \neq k} A_{kj} x_k x_j + A_{kk} x_k^2 \right]$$

$$= \sum_{i \neq k} A_{ik} x_i + \sum_{j \neq k} A_{kj} x_j + 2A_{kk} x_k$$

Exercise

$$
\begin{aligned}
\frac{\partial f(x)}{\partial x_k} &= \frac{\partial}{\partial x_k} \sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij} x_i x_j \\
&= \frac{\partial}{\partial x_k} \left[\sum_{i \neq k} \sum_{j \neq k} A_{ij} x_i x_j + \sum_{i \neq k} A_{ik} x_i x_k + \sum_{j \neq k} A_{kj} x_k x_j + A_{kk} x_k^2 \right] \\
&= \sum_{i \neq k} A_{ik} x_i + \sum_{j \neq k} A_{kj} x_j + 2 A_{kk} x_k
\end{aligned}
$$

Exercise

$$
\begin{aligned}
\frac{\partial f(x)}{\partial x_k} &= \frac{\partial}{\partial x_k} \sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij} x_i x_j \\
&= \frac{\partial}{\partial x_k} \left[\sum_{i \neq k} \sum_{j \neq k} A_{ij} x_i x_j + \sum_{i \neq k} A_{ik} x_i x_k + \sum_{j \neq k} A_{kj} x_k x_j + A_{kk} x_k^2 \right] \\
&= \sum_{i \neq k} A_{ik} x_i + \sum_{j \neq k} A_{kj} x_j + 2 A_{kk} x_k \\
&= \sum_{i=1}^{n} A_{ik} x_i + \sum_{j=1}^{n} A_{kj} x_j = 2 \sum_{i=1}^{n} A_{ki} x_i,
\end{aligned}
$$

Exercise

$$f(x) = x^T A x$$

$$f(x) = \sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij} x_i x_j$$

$$\frac{\partial^2 f(x)}{\partial x_k \partial x_\ell} = \frac{\partial}{\partial x_k} \left[\frac{\partial f(x)}{\partial x_\ell} \right] = \frac{\partial}{\partial x_k} \left[\sum_{i=1}^{n} A_{\ell i} x_i \right]$$

Exercise

$$f(x) = x^T A x$$

$$f(x) = \sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij} x_i x_j$$

$$\frac{\partial^2 f(x)}{\partial x_k \partial x_\ell} = \frac{\partial}{\partial x_k} \left[\frac{\partial f(x)}{\partial x_\ell} \right] = \frac{\partial}{\partial x_k} \left[\sum_{i=1}^{n} A_{\ell i} x_i \right]$$

$$= 2A_{\ell k} = 2A_{k\ell}.$$

Exercise

$$f(x) = x^T A x$$

$$f(x) = \sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij} x_i x_j$$

$$\frac{\partial^2 f(x)}{\partial x_k \partial x_\ell} = \frac{\partial}{\partial x_k} \left[\frac{\partial f(x)}{\partial x_\ell} \right] = \frac{\partial}{\partial x_k} \left[\sum_{i=1}^{n} 2 A_{\ell i} x_i \right]$$

$$= 2A_{\ell k} = 2A_{k\ell}.$$

$$\nabla_x^2 f(x) = 2A$$

What we have learned

- Vectors and matrices
  - Basic Matrix Operations
  - Special Matrices
- Transformation Matrices
  - Homogeneous coordinates
  - Translation
- Matrix inverse
- Matrix rank
- Eigenvalues and Eigenvectors
- Matrix Calculate