MAT 180: Mathematics of Deep Learning

Greg DePaul

September 23, 2022

1 Linear Algebra Review

1.1 Vector Spaces

Definition 1.1. *A* <u>*vector space*</u> *over* $\mathbb{R}$ *is a set* $\mathcal{V}$ *closed under addition and scalar multiplication satisfying the usual properties.*

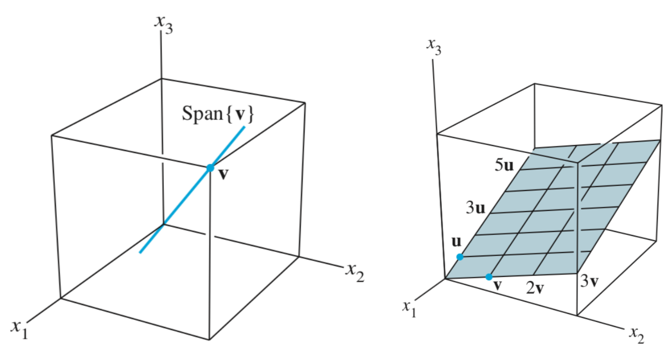
Example 1.2. *Given* $\mathbb{R}^n$ *where* $n = 2$ *we see*

$$(1, 2) + (-3, 4) = (-2, 6)$$

$$10 \cdot (-1, 1) = (-10, 10)$$

Definition 1.3. *Suppose we are given a set of vectors* $S = \{v_1, \ldots, v_n\} \subset \mathcal{V}$. *Then the* <u>*span*</u> *of* S *is the set*

$$span(S) := \{a_1 v_1 + \ldots + a_k v_k : v_i \in S, a_i \in \mathbb{R} \ \forall i \ \}$$



(a) Span of one vector (b) Span of two vectors

Figure 1: Linear Spans

Example 1.4. *The typical well at a cocktail bar contains at least four ingredients at the bartender's disposal; vodka, tequila, orange juice, and grenadine. Assuming we have this well, we can represent drinks as points in* $\mathbb{R}^4$, *with one element for each ingredient. For instance, a tequila sunrise can be represented using the point*

$$(0, 1.5, 6, 0.75)$$

representing amounts of vodka, tequila, orange juice, and grenadine (in ounces). The set of drinks can be represented as the span

$$span(\{(1, 0, 0, 0), (0, 1, 0, 0), (0, 0, 1, 0), (0, 0, 0, 1)\})$$

Further, the bartender might be able to save time by making the observation that many drinks have the same orange juice - to - grenadine ratio, and therefore mix the two. So they might simplify their well by mixing the two:

$$span(\{(1, 0, 0, 0), (0, 1, 0, 0), (0, 0, 6, 0.75)\})$$

Notice, it is now easier to pour drinks but this bartender can no longer make as many drinks, such as a screwdriver which contains orange juice but no grenadine.

Definition 1.5. *A set $S \subset \mathcal{V}$ of vectors is* linearly dependent *if there exists a non-empty linear combination of elements $v_k \in S$ yielding*

$$\sum_{k=1}^{m} c_k v_k = 0$$

where $c_k \neq 0$ for all k. A set that is not linearly dependent is called linearly independent.

Definition 1.6. *The* dimension *of $\mathcal{V}$ is the maximal size $|S|$ of a linearly independent set $S \subset \mathcal{V}$ such that $span(S) = \mathcal{V}$.*

Definition 1.7. *Any linearly independent set S of maximal size $|S|$ with $span(S) = \mathcal{V}$ is a* basis *of $\mathcal{V}$*

Example 1.8. *The standard basis for $\mathbb{R}^n$ is the set of vectors of the form*

$$e_k = (\underbrace{0, \ldots, 0,}_{k-1 \text{ elements}} \ 1, \ \underbrace{0, \ldots, 0}_{n-k \text{ elements}})$$

1.2 Vector Norms

Definition 1.9. *A* vector norm *is a function $\|\cdot\| : \mathbb{R}^n \to [0, \infty)$ satisfying the following conditions:*

1. *$\|x\| = 0$ if and only if $x = 0$ (Nondegeneracy)*
2. *$\|cx\| = |c| \, \|x\|$ for all scalars $c \in \mathbb{R}, x \in \mathbb{R}^n$ (Absolutely scalability)*
3. *$\|x + y\| \leq \|x\| + \|y\|$ for all $x, y \in \mathbb{R}^n$ (Triangle Inequality)*

Definition 1.10. $\|x\|_p := (|x_1|^p + |x_2|^p + \ldots + |x_n|^p)^{\frac{1}{p}}$

Definition 1.11. $\|x\|_1 := \sum_{k=1}^{n} |x_k|$

Definition 1.12. $\|x\|_\infty := \max(|x_1|, |x_2|, \ldots, |x_n|)$

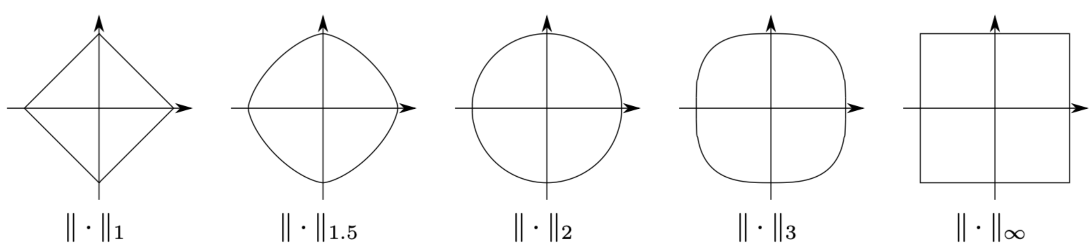


Figure 4.7 The set $\{\vec{x} \in \mathbb{R}^2 : \|\vec{x}\| = 1\}$ for different vector norms $\|\cdot\|$.

Definition 1.13. *Two norms $\|\cdot\|$ and $\|\cdot\|'$ are* equivalent *if there exists constants c_{low} and c_{high} such that*

$$c_{low} \|x\| \leq \|x\|' \leq c_{high} \|x\|$$

for all $x \in \mathbb{R}^n$.

Theorem 1.14 (Equivalence of norms in finite dimension)**.** *All norms on $\mathbb{R}^n$ are equivalent.*

1.3 The Inner Product Space $\mathbb{R}^n$

Definition 1.15. *The* dot product *of two vectors a and b in $\mathbb{R}^n$ is given by*

$$a \cdot b = \sum_{k=1}^{n} a_k b_k$$

Example 1.16. *In* $\mathbb{R}^2$, *we see*

$$(1,2) \cdot (-2,6) = 10$$

Topologically, the dot product $\cdot$ induces a norm:

$$\|a\|_2 = \sqrt{a_1^2 + \ldots + a_n^2} = \sqrt{a \cdot a}$$

Geometrically, we may remember the following relationship:

$$a \cdot b = \|a\|_2 \|b\|_2 \cos\theta$$

where θ is the angle between a and b. Now, we $\cos\theta = 0$, we see that the dot product is also zero. This motivates the following definition:

Definition 1.17. *Two vectors* $a, b \in \mathbb{R}^n$ *are* <u>*orthogonal*</u> *when* $a \cdot b = 0$.

1.4 Linear Functions

Definition 1.18. *Suppose* $\mathcal{V}$ *and* $\mathcal{V}'$ *are vector spaces. Then an operator* $\mathcal{L} : \mathcal{V} \to \mathcal{V}'$ *is* <u>*linear*</u> *if it satisfies the following for all* $v_1, v_2 \in \mathcal{V}$ *and* $c \in \mathbb{C}$*:*

- $\mathcal{L}[v_1 + v_2] = \mathcal{L}[v_1] + \mathcal{L}[v_2]$
- $\mathcal{L}[cv] = c\mathcal{L}[v]$

Example 1.19.

$$\begin{aligned} f &: \mathbb{R}^2 \to \mathbb{R}^3 \\ (x,y) &\to (3x, 2x+y, -y) \end{aligned}$$

Proposition 1.20. *A linear operator* $\mathcal{L}$ *on* $\mathbb{R}^n$ *is completely determined by its action on the standard basis vectors* e_k*.*

Proof.

$$\mathcal{L}[a] = \mathcal{L}[\sum_k a_k e_k] = \sum_k \mathcal{L}[a_k e_k] = \sum_k a_k \mathcal{L}[e_k]$$

■

Example 1.21. *Returning the previous example:*

$$f(x,y) = xf(e_1) + yf(e_2) = x \begin{pmatrix} 3 \\ 2 \\ 0 \end{pmatrix} + y \begin{pmatrix} 0 \\ 1 \\ -1 \end{pmatrix}$$

1.5 Matrices

Definition 1.22. *The* <u>*space of matrices*</u> $\mathbb{R}^{m \times n}$ *is the set composed of matrices of the form:*

$$\begin{pmatrix} \vdots & \vdots & \ldots & \vdots \\ v_1 & v_2 & \ldots & v_n \\ \vdots & \vdots & \ldots & \vdots \end{pmatrix} = \begin{pmatrix} v_{11} & v_{12} & \ldots & v_{1n} \\ v_{21} & v_{22} & \ldots & v_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ v_{m1} & v_{m2} & \ldots & v_{mn} \end{pmatrix}$$

with $v_{i,j} \in \mathbb{R}$.

Definition 1.23. *Matrix to vector multiplication is defined by*

$$\begin{pmatrix} \vdots & \vdots & \dots & \vdots \\ v_1 & v_2 & \dots & v_n \\ \vdots & \vdots & \dots & \vdots \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{pmatrix} = c_1 v_1 + c_2 v_2 + \ldots + c_n v_n$$

Example 1.24. *The previous function f can be written*

$$f(x, y) = \begin{pmatrix} 3 & 0 \\ 2 & 1 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}$$

Theorem 1.25 (Riesz Representation)**.** *Every linear operator from a vector space $\mathcal{V}$ to $\mathcal{V}'$, there exists a linear matrix A on a chosen basis such that*

$$\mathcal{L}(x_1, \ldots, x_n) = A \cdot \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$$

Definition 1.26. *Matrix to matrix multiplication is defined by natural extension of matrix to vector multiplication*

$$M \begin{pmatrix} \vdots & \vdots & \dots & \vdots \\ v_1 & v_2 & \dots & v_n \\ \vdots & \vdots & \dots & \vdots \end{pmatrix} = \begin{pmatrix} \vdots & \vdots & \dots & \vdots \\ Mv_1 & Mv_2 & \dots & Mv_n \\ \vdots & \vdots & \dots & \vdots \end{pmatrix}$$

Example 1.27. *Returning to the cocktail example, suppose we make two drinks from our 3 defined wells of liquid (vodka, tequila, and the mix of grenadine and orange juise). Then to find the basic ingredients we simply use matrix multiplication:*

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 6 \\ 0 & 0 & 0.75 \end{pmatrix} \begin{pmatrix} 0 & 0.75 \\ 1.5 & 0.75 \\ 1 & 2 \end{pmatrix} = \begin{pmatrix} 0 & 0.75 \\ 1.5 & 0.75 \\ 6 & 12 \\ 0.75 & 1.5 \end{pmatrix}$$

1.6 Matrix Transpositions

Definition 1.28. *The transpose of a matrix $A \in \mathbb{R}^{m \times n}$ is the matrix $A^T \in \mathbb{R}^{n \times m}$ with elements defined by*

$$[A^T]_{i,j} = [A]_{j,i}$$

Definition 1.29. *We can redefine the dot product in terms of matrix multiplication $a \cdot b = a^T b$.*

Proposition 1.30. $(A^T)^T = A$

Proposition 1.31. $(A + B)^T = A^T + B^T$

Proposition 1.32. $(AB)^T = B^T A^T$

1.7 Trace of a Matrix

Definition 1.33. *The trace of a given matrix $Tr(A) = \sum_i A_{ii}$.*

Proposition 1.34. $Tr(A) = Tr(A^T)$

Proposition 1.35. $Tr(AB) = Tr(BA)$

1.8 Special Matrices

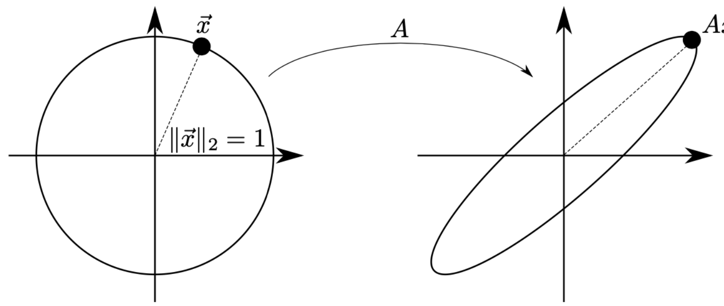
Definition 1.36. *A matrix is* symmetric *if* $A = A^T$.

Definition 1.37. *A matrix* $A \in \mathbb{R}^{n\times n}$*is* positive definite *if for every* $x \in \mathbb{R}^n, x \neq 0 \implies x^T Ax > 0$.

1.9 Matrix Norms

Definition 1.38. *The matrix norm on* $\mathbb{R}^{m\times n}$ induced *by a vector norm* $\|\cdot\|$ *is given by*

$$\|A\| = \max_{\|x\|=1} \|Ax\|$$



Definition 1.39. $\|A\|_1 = \max_{1\leq j\leq n} \sum_{i=1}^m |a_{i,j}|$

Definition 1.40. $\|A\|_\infty = \max_{1\leq i\leq m} \sum_{j=1}^n |a_{i,j}|$

Definition 1.41. $\|A\|_2 = \sqrt{\lambda : \exists x \in \mathbb{R}^n \text{ with } A^T Ax = \lambda x}$

Definition 1.42. $\|A\|_{Fro} := \sqrt{\sum_{i,j} a_{i,j}^2}$ *is the Frobenius norm.*

Note 1.43. *The Frobenius norm cannot be induced from a vector norm.*

Proposition 1.44. $\|A\|_{Fro} = \sqrt{Tr(AA^T)}$

Definition 1.45. *Given a positive definite matrix* A*, we can define a vector norm* $\|\cdot\|_A$ *by*

$$\|x\| = \sqrt{x^T Ax}$$

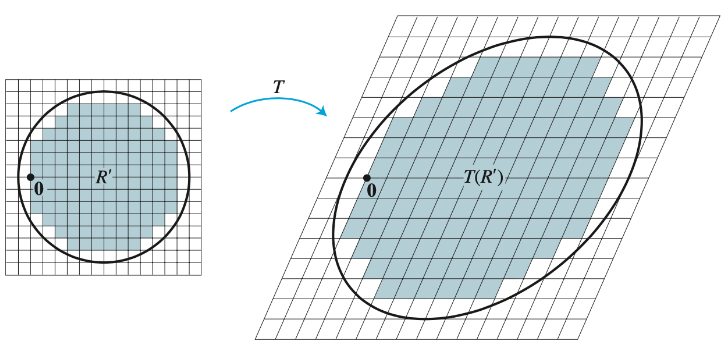
1.10 Determinants and Invertibility

Definition 1.46. *Given a matrix* $A \in \mathbb{R}^{n\times n}$*, we define*

$$det(A) = \sum_{\pi\in S_n} (-1)^{sign(\pi)} \prod_i A_{i,\pi(i)}$$

Example 1.47.

$$det\left(\begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix}\right) = 1$$



If A is a 2×2 matrix, the area of the parallelogram determined by the columns of A is $|\det A|$. If A is a 3×3 matrix, the volume of the parallelepiped determined by the columns of A is $|\det A|$.

Proposition 1.48. $det(AB) = det(A)det(B)$

Proposition 1.49. *A matrix has nonzero determinant if and only if that matrix is invertible.*

1.11 Orthogonality

Definition 1.50. *A set of vectors $v_1, \ldots, v_k$ is orthonormal if $\|v_i\|_2 = 1$ for all i and $v_i \cdot v_k = 0$.*

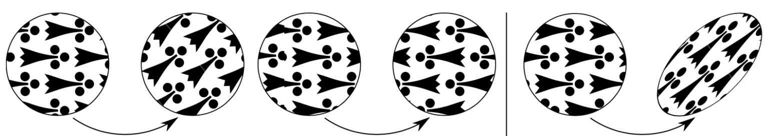
Definition 1.51. *A square matrix whose columns are orthonormal is called an orthogonal matrix.*

Proposition 1.52. *A matrix Q is orthogonal if and only if Q is invertible and $Q^{-1} = Q^T$.*

Proposition 1.53. *If matrix Q is orthogonal, then $Q^T Q = QQ^T = I$*

Definition 1.54. *An isometry on $\mathbb{R}^n$ is a distancen preserving bijection $f : \mathbb{R}^n \to \mathbb{R}^n$. That is,*

$$\|f(\vec{x}) - f(\vec{y})\| = \|x - y\|$$



(a) Isometric (b) Not isometric

Proposition 1.55. *If Q is orthogonal, then the function $x \to Qx$ is an isometry on $\mathbb{R}^n$.*

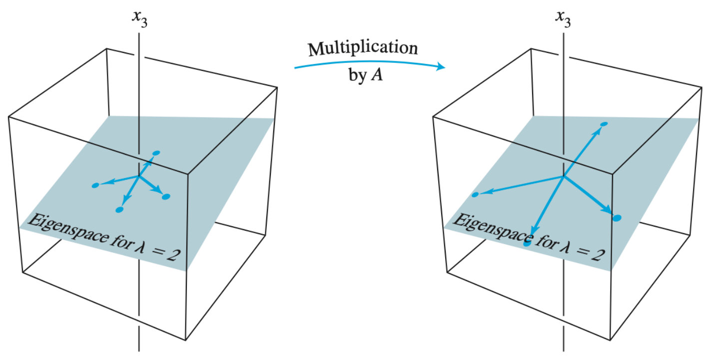
Proposition 1.56. *Every isometry on $\mathbb{R}^n$ can be written as the the function $x \to A + Qx$ for some $A, Q \in \mathbb{R}^{n \times n}$ with Q orthogonal.*

2 Eigenvalue and Singular Value Decompositions

2.1 Eigenvalues and Eigenvectors

Definition 2.1. *Suppose T is a linear operator from a vector space V to V. A subspace $U \subset V$ is called invariant under T is $u \in U$ implies $Tu \in U$.*

Definition 2.2. *An eigenvector of a matrix $A \in \mathbb{R}^{n \times n}$ is a nonzero vector v such that $Av = \lambda v$ for some scalar λ. A scalar λ is called an eigenvalue of A if there is a nontrivial solution v such that $Av = \lambda v$.*



Theorem 2.3. *Every matrix $A \in \mathbb{R}^{n \times n}$ has at least one (potentially complex) eigenvector.*

Proof. Take any vector $x \in \mathbb{R}^n \setminus \{0\}$ and assume $A \neq 0$ since this matrix trivially has eigenvalue 0. The set

$$\{x, Ax, A^2x, \ldots, A^n x\}$$

must be linearly dependent because it contains $n+1$ vectors in n dimensions. So there exists constants $c_0, \ldots, c_n \in \mathbb{R}$ not all zero such that

$$0 = c_0 x + c_1 Ax + c_2 A^2 x + \ldots + c_n A^n x$$

We define the polynomial

$$f(z) = c_0 + c_1 z + \ldots + c_n z^n$$

By the fundamental theorem of Algebra, there exist $m \geq 1$ roots $z_i \in \mathbb{C}$ and $c \neq 0$ such that

$$f(z) = c(z - z_1)(z - z_2) \ldots (z - z_m)$$

Applying this factorization, we see that

$$\begin{aligned} 0 &= c_0 x + c_1 Ax + c_2 A^2 x + \ldots + c_n A^n x \\ &= (c_0 I + c_1 A + \ldots + c_n A^n)x \\ &= f(A)x \\ &= c(A - z_1 I) \ldots (A - z_m I)x \end{aligned}$$

In this form, at least one $A - z_i I$ has a null space, since otherwise each term would be invertible, forcing $x = 0$, which we already assumed against. So if we take v to be a nonzero vector in the null space of $A - z_i I$, then by construction

$$Av = z_i v$$

■

Theorem 2.4. *Suppose V is a complex vector space and $T \in \mathcal{L}(V)$. Let $\lambda_1, \ldots, \lambda_m$ denote the distinct eigenvalues of T, with multiplicities $d_1, \ldots, d_m$. Then the polynomial*

$$(z - \lambda_1)^{d_1} \ldots, (z - \lambda_m)^{d_m}$$

is called the characteristic polynomial of T.

Proof. 3 key steps to proof, each of which should be proven carefully as well:

1. Suppose $T \in \mathcal{L}(V)$. Then there is a unique monic polynomial p of smallest degree such that $p(T) = 0$. This is called the minimal polynomial.
2. Suppose $T \in \mathcal{L}(V)$. Then the zeros of the minimal polynomial are exactly the eigenvalues of T.
3. Suppose $\mathbb{F} = \mathbb{C}$ and $T \in (V)$. Then the characteristic polynomial of T is a polynomial multiple of the minimal polynomial.

∎

Theorem 2.5. *A scalar λ is an eigenvalue of and $n \times n$ matrix A if and only if λ satisfies the characteristic equation*

$$det(A - \lambda I) = 0$$

Theorem 2.6. *The eigenvectors corresponding to distinct eigenvalues of a given matrix are linearly independent.*

Proof. Suppose otherwise. Then there exists eigenvectors $v_1, \ldots, v_k$ with distinct eigenvalues $\lambda_1, \ldots, \lambda_k$ that are linearly dependent. This implies that there are coefficients $c_1, \ldots, c_k$ not all zero such that

$$0 = c_1 v_1 + \ldots + c_k v_k$$

Notice, for any $i \neq j$, we see that

$$(A - \lambda_i I)x_j = Ax_j - \lambda_i x_j = (\lambda_j - \lambda_i)x_j$$

We can then isolate one of the coefficients

$$0 = (A - \lambda_2 I) \ldots (A - \lambda_k I)c_1 v_1 + \ldots + c_k v_k = c_1(\lambda_1 - \lambda_2) \ldots (\lambda_1 - \lambda_k)$$

Since λ_1 does not equal any of the other distinct eigenvalues, then

$$c_1 = 0$$

We can repeat this for all the either eigenvalues. ∎

2.2 Hermitian and Positive Semi-definite Matrices

Definition 2.7. *The conjugate transpose of $A \in \mathbb{C}^{m \times n}$ is $A^H = \overline{A}^T$.*

Definition 2.8. *A matrix $A \in \mathbb{R}^{n \times n}$is hermitian if $A = A^H$.*

Theorem 2.9. *All eigenvalues of Hermitian matrices are real.*

Proof. Suppose $A \in \mathbb{C}^{n \times n}$ is Hermitian with $Av = \lambda v$. By scaling, we assume $\|v\|_2^2 = v^T \overline{v} = 1$. Then

$$\begin{aligned}
\lambda &= \lambda v^T \overline{v} \\
&= (\lambda v)^T \overline{v} \\
&= (Av)^T \overline{v} \\
&= v^T \overline{A^T v} \\
&= v^T A \overline{v} \\
&= \overline{\lambda} v^T \overline{v} \\
&= \overline{\lambda}
\end{aligned}$$

∎

Theorem 2.10. *Eigenvectors corresponding to distinct eigenvalues of Hermitian matrices must be orthogonal.*

Proof. Suppose $A \in \mathbb{C}^{n\times n}$ is Hermitian and suppose $\lambda \neq \mu$ with $Ax = \lambda x$ and $Ay = \mu y$. By the previous theorem, we know that $\lambda, \mu \in \mathbb{R}$. Then $x^T A^T y = \lambda x^T y$. But since A is Hermitian, we can also write

$$x^T A^T y = x^T A^H y = x^T A y = \mu x^T y$$

Therefore, $\lambda x^T y = \mu x^T y$. Since $\lambda \neq \mu$, then $x^T y = 0$. ■

Definition 2.11. *A matrix $A \in \mathbb{R}^{n\times n}$ is positive semi-definite if for every $x \in \mathbb{R}^n \implies x^T A x \geq 0$.*

Theorem 2.12. *If A is positive-semidefinite, then A has nonnegative real eigenvalues.*

Proof. Take $A \in \mathbb{R}^{n\times n}$ to be positive definite, and suppose $Ax = \lambda x$ with $\|x\|_2 = 1$. By positive definiteness, we know that $x^T A x \geq 0$. But notice,

$$0 \leq x^T A x = x^T(\lambda x) = \lambda \|x\|_2^2 = \lambda$$

■

Corollary 2.12.1. *If A is positive-definite, then A has positive eigenvalues.*

Proposition 2.13. *For any $A \in \mathbb{R}^{m\times n}$, the matrix $A^T A$ is positive semidefinite.*

Proof. Take any $x \in \mathbb{R}^n$. Then

$$x^T(A^T A)x = (Ax)^T(Ax) = \|Ax\|_2^2 \geq 0$$

■

Corollary 2.13.1. *For any $A \in \mathbb{R}^{m\times n}$, the matrix $A^T A$ is positive definite provided the columns of A are linear independent.*

Proof. Suppose the columns of A are linearly independent. If A were only semi-definite, then there $\exists x \neq 0$ such that

$$\begin{aligned} x^T A^T A x &= 0 \\ \implies \|Ax\|_2 &= 0 \\ \implies Ax &= 0 \\ \implies \text{columns of } A &\text{ are not linearly independent.} \end{aligned}$$

Contradiction! ■

2.3 Eigenvalue Diagonalization and Spectral Theorem

Definition 2.14. *A matrix $A \in \mathbb{R}^{n\times n}$ is diagonalizable if $A = PDP^{-1}$ for some invertible matrix $P \in \mathbb{R}^{n\times n}$ and some diagonal matrix $D \in \mathbb{R}^{n\times n}$.*

Theorem 2.15. *An $n \times n$ matrix $A \in \mathbb{R}^{n\times n}$ is diagonalizable if and only if A has n linearly independent. Further,*

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

where the columns of P are exactly the eigenvectors of A and the diagonal entries of D are the eigenvalues of A.

Theorem 2.16 (Spectral Theorem)**.** *Suppose $A \in \mathbb{C}^{n\times n}$ is Hermitian. Then A has exactly n orthonormal eigenvectors $v_1, \ldots, v_n$ with (possibly repeated) eigenvalues $\lambda_1, \ldots, \lambda_n$. In other words, there exists an orthogonal matrix Q of eigenvectors and a diagonal matrix D such that*

$$A = QDQ^T$$

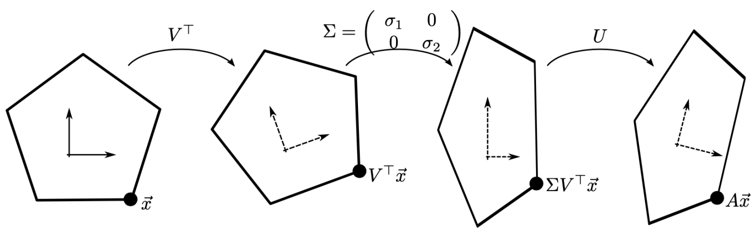
2.4 Singular Value Decomposition

Theorem 2.17. *$A \in \mathbb{R}^{m\times n}$, then there exist orthogonal $U \in \mathbb{R}^{m\times m}$ and $V \in \mathbb{R}^{n\times n}$ such that*

$$UAV^T = \Sigma := diag(\sigma_1, \sigma_2, \ldots, \sigma_p) \in \mathbb{R}^{m\times n}$$

where $\sigma_1 \geq \sigma_2 \geq \sigma_3 \geq \ldots \geq \sigma_p \geq 0$.

Visually,



Proof. Since A^TA is a symmetric matrix, we can apply spectral theorem to get

$$A^TA = VDV^T$$

Further, since A^TA is positive semidefinite, we know all of the entries of D are nonnegative. We now define the set $\{v_i\}_{i=1}^n$ to be the set of eigenvectors of that make up the columns of Q. Define

$$u_i = \frac{Av_i}{\sqrt{D_{ii}}}$$

Observe,

$$\|u_i\| = \left|\frac{1}{\sqrt{D_{ii}}}\right| \|Av_i\| = \left|\frac{1}{\sqrt{D_{ii}}}\right| \sqrt{v_i^TA^TAv_i} = \left|\frac{1}{\sqrt{D_{ii}}}\right| \sqrt{D_{ii}v_i^Tv_i} = 1$$

$$u_i^Tu_j = \frac{v_i^TA^TAv_j}{\sqrt{D_{ii}D_{jj}}} = \frac{v_i^TD_{jj}v_j}{\sqrt{D_{ii}D_{jj}}} = \frac{D_{jj}}{\sqrt{D_{ii}D_{jj}}}v_i^Tv_jj = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$$

So we see that $\{u_i\}_{i=1^m}$ is a set of orthonormal vectors. Lastly, define

$$\Sigma_{i,j} = \sqrt{D_{ii}}$$

Then we see that

$$U\Sigma = \begin{pmatrix} \vdots & \vdots & \cdots & \vdots \\ \frac{Av_1}{\sqrt{D_{11}}} & \frac{Av_2}{\sqrt{D_{22}}} & \cdots & \frac{Av_n}{\sqrt{D_{nn}}} \\ \vdots & \vdots & \cdots & \vdots \end{pmatrix} \begin{pmatrix} \sqrt{D_{11}} & 0 & \cdots & 0 \\ 0 & \sqrt{D_{22}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sqrt{D_{nn}} \end{pmatrix} = \begin{pmatrix} \vdots & \vdots & \cdots & \vdots \\ Av_1 & Av_2 & \cdots & Av_n \\ \vdots & \vdots & \cdots & \vdots \end{pmatrix} = AV$$

∎

Corollary 2.17.1. *If $U^TAV = \Sigma$ for $A \in \mathbb{R}^{m\times n}$, then for $1 \leq i \leq n, Av_i = \sigma_iu_i$ and $A^Tu_i = \sigma_iv_i$.*

Corollary 2.17.2. *If $A \in \mathbb{R}^{m\times n}$, then $\|A\|_2 = \sigma_1$ and $\|A\|_{Fro} = \sqrt{\sigma_1^2 + \ldots + \sigma_p^2}$*

Proof.

$$\|A\|_{Fro} = \sqrt{\sum_{i,j} a_{ij}^2} = \sqrt{tr(A^TA)} = \sqrt{tr(V\Sigma U^TT\Sigma V^T)} = \sqrt{tr(V\Sigma^2V^T)} = \sqrt{tr(VV^T\Sigma^2)} = \sqrt{tr(\Sigma^2)} = \sqrt{\sigma_1^2 + \ldots + \sigma_p^2}$$

∎

Note 2.18. $\|\Sigma\| = \left\|UAV^T\right\| = \|A\|$.

Corollary 2.18.1. *If A has r positive singular values, then $rank(A) = r$ and*

$$null(A) = span\{v_{r+1}, \dots, v_n\}$$

$$col(A) = span\{u_1, \dots, u_r\}$$

Proof. $rank(A) = rank(\Sigma) = r$. ■

Corollary 2.18.2. *If $A \in \mathbb{R}^{m \times n}$, with $rank(A) = r$, then $A = \sum_{k=1}^{r} \sigma_i u_i v_i^T$.*

Proof.

$$(U\Sigma)V^T = \begin{pmatrix} \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \sigma_1 u_i & \dots & \sigma_r u_r & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{pmatrix} \begin{pmatrix} \dots & v_1^T & \dots \\ & \vdots & \\ \dots & v_n^T & \dots \end{pmatrix} = \sum_{k=1}^{r} \sigma_i u_i v_i^T$$

■

3 Principal Component Analysis