1 Vectors

Inner Product: $\langle a, b\rangle = a^T b = \sum_{i=1}^{n} a_i b_i$ Magnitude: $\|v\|^2 = \langle v, v\rangle$ Angle: $\cos\theta = \frac{\langle v, w\rangle}{\|v\|\|w\|}$

1.1 Linear Combination

A linear combination of vectors $\{v_1, \ldots, v_n\}$ is a sum $c_1 v_1 + \cdots + c_n v_n$, where each c_i is a real number.

1.2 Linear Independence

Vectors are LI iff no vector in the set can be expressed as a linear combination of the other vectors. Equivalently:

$$\text{LI iff} \quad \lambda_1 v_1 + \cdots + \lambda_k v_k = 0 \implies \lambda_1 = 0, \ldots, \lambda_k = 0$$

2 Matrix Properties

A matrix $M \in \mathbb{R}^{m\times n}$ has m rows and n columns.

2.1 Matrix-Matrix Multiplication

- $A(BC) = (AB)C$
- $A(B + C) = AB + AC$, $(A + B)C = AC + AB$
- $AB \neq BA$ generally

2.2 Inverse

R is a right inverse if $AR = I$, L is left inverse if $LA = I$. If L *and* R exist, then $L = R = A^{-1}$ (and A is square).

- For square invertable A and B, $(AB)^{-1} = B^{-1}A^{-1}$.
- If A is orthogonal then $A^{-1} = A^T$ (since $A^T A = I$).

To compute the inverse, use Gaussian Elimination to transform $[A\, I]$ to $[I\, B]$. Then $B = A^{-1}$.

For a 2×2 matrix, $\left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right]^{-1} = \frac{1}{ad-bc}\left[\begin{smallmatrix} d & -b \\ -c & a \end{smallmatrix}\right]$

2.3 Transpose

The *transpose* of an $m \times n$ matrix A is an $n \times m$ matrix A^T where $(A^T)_{ij} = Aji$.

- $(AB)^T = B^T A^T$
- $(A^T)^T = A$
- $(A + B)^T = A^T + B^T$
- For square and invertible A, $(A^{-1})^T = (A^T)^{-1}$

2.4 Rank

The rank of a matrix is given by:

- The max number of L.I. columns ($= \dim C(A)$)
- The max number of L.I. rows ($= \dim C(A^T)$)
- The number of pivots in the RREF of A.

2.4.1 Rank-Nullity Theorem

For any $m \times n$ matrix A, $rank(A) + dim(N(A)) = n$.

3 Gaussian Elimination

To solve a system $Ax = b$ (find x), we can perform row operations on the augmented matrix to simplify without changing the solution set.

Swap Rows 1 & 2 $\begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$ $r_1' = r_1 - r_2$ $\begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$ $r_1' = 2 * r_1$ $\begin{bmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$

A matrix is in row-echelon form when all pivots (leftmost nonzero in a row) go from left to right. A matrix is in reduced REF if each pivot has only zeros in its column.

Example. Let's say we have the following system, which we write in matrix-vector form:

$$\begin{cases} x_1 - x_2 & +2x_3 = 1 \\ -2x_1 + 2x_2 & -3x_3 = -1 \\ -3x_1 - x_2 & +2x_3 = -3 \end{cases}$$

To keep track of operations, we put the matrix in augmented form by appending b, and then perform operations to convert it to REF: $r_2' = r2 + 2r_1 \to r_3' = r_3 + 3r_1 \to$ swap r_2, r_3.

$$\left[\begin{array}{ccc|c} 1 & -1 & 2 & 1 \\ -2 & 2 & -3 & -1 \\ -3 & -1 & 2 & -3 \end{array}\right] \to \left[\begin{array}{ccc|c} 1 & -1 & 2 & 1 \\ 0 & -4 & 8 & 0 \\ 0 & 0 & 1 & 1 \end{array}\right]$$

We can then create simpler equations:

$$\begin{bmatrix} 1 & -1 & 2 \\ 0 & -4 & 8 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} \to \begin{cases} x_1 - x_2 & +x_3 = 1 \\ -4x_2 & +8x_3 = 0 \\ & x_3 = 1 \end{cases}$$

A matrix is singular if it has a row of all zeros—then it has many or no solutions.

4 Operations on Vectors

4.1 Permutations

A permutation is a function that changes the order of elements in a vector. In matrix form, it has a single 1 in each row and column and zeros elsewhere. Applying a permutation matrix preserves length and relative angle between vectors, and if A is a permutation matrix then $A^{-1} = A^T$.

4.2 Rotation

To rotate a two-dimensional vector x by an angle θ, we can use the matrix below. Conveniently, $R_\theta^{-1} = R_{-\theta} = R_\theta^T$ (and this holds in higher dimensions):

$$R_\theta = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \quad R_\theta^{-1} = R_{-\theta} = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}$$

4.3 Projection

We can project a point onto a line by minimizing the distance to the line, i.e. $\text{proj}_l(a) = \arg\min_{b\in l} \|a - b\|$. To do this more easily, we can rotate the line to be on the x axis and zero the y coordinate. With $D = \left[\begin{smallmatrix} 1 & 0 \\ 0 & 0 \end{smallmatrix}\right]$, a matrix that projects onto a line with angle θ is then $P = R_\theta D R_{-\theta}$, where Pv gives the projection. Note that a projection matrix cannot have an inverse and $P^2 = P$.

$$\text{proj}_w v = \frac{v \cdot w}{\|w\|^2} w$$

4.4 Reflection

A reflection across a line l can be expressed as $2(b-a)+a$, where b is the closest point on l to a. Since $2(b-a)+a = 2b-a$, we can write this as a matrix product with $2P-I$. A reflection across a line with angle θ can also be written as:

$$\begin{bmatrix} \cos 2\theta & \sin 2\theta \\ \sin 2\theta & -\cos 2\theta \end{bmatrix}$$

5 Vector Space

A vector space is a set V on which addition and scaling are closed: $\forall v, w \in V, v+w \in V$ and $\forall v \in V, \alpha \in \mathbb{R}, \alpha v \in V$. If S and V are vector spaces, then $S \subseteq V$ is also a subspace.

- If S_1 and S_2 are subspaces then $S = S_1 \cap S_2$ is too.
- $S = S_1 + S_2$ is too (NOT union).

5.1 Columnspace

The columnspace of a matrix is the span of its n columns, where the span of a set of vectors is the set of all linear combinations of vectors in the set. If $A \in \mathbb{R}^{m \times n}$, then $C(A) \in \mathbb{R}^m$.

Given a matrix A, a basis for the columnspace of A is given by the columns corresponding to the pivots in $rref(A)$.

5.2 Nullspace

The nullspace of A is $N(A) = \{x | Ax = 0\}$.

- If B is square and invertible, then $N(A) = N(BA)$.
- For any B, if $A = BA'$ then $N(A') \subseteq N(A)$.

To find the nullspace of a matrix, we can take advantage of the fact that $N(A) = N(BA)$ if B is invertible. We can reduce A to rref, then form equations from R and write the vector in terms of the free variables.

5.3 Generators and Bases

A set of vectors $\mathcal{V} \subseteq \mathcal{S}$ *generates* a subspace $\mathcal{S}$ iff every vector in $\mathcal{S}$ can be written as a linear combination of vectors in $\mathcal{V}$. If the vectors in $\mathcal{V}$ are linearly independent, then $\mathcal{V}$ is a *basis* for $\mathcal{S}$. All bases have the same cardinality—the dimension of the subspace.

5.4 Projections

For a subspace V defined by orthonormal basis vectors $\{v_1, \dots v_i\}$, we can compute $proj_V x = \sum_{i=1}^{k} (x \cdot v_i) v_i$. Expanding this:

$$proj_V x = \sum_{i=1}^{k} (x \cdot v_i) v_i = \sum_{i=1}^{k} v_i (v_i^T x) = (\sum_{i=1}^{k} v_i v_i^T) x$$

We can then equivalently define $P = \sum_{i=1}^{k} v_i v_i^T = VV^T$, where $V = [v_1 \dots v_k]$. Then, $proj_V x = VV^T x$.

Note that the rank of P is equal to the dimension of the subspace.

5.5 Orthogonality

Two subspaces can be orthogonal, which means that $\forall v \in V, \forall w \in W, v \cdot w = 0$ (i.e. all vectors are orthogonal). Given $V \in \mathbb{R}^n$, its orthogonal complement $V^\perp$ is the set of vectors that are orthogonal to all $v \in V$. Importantly, $V^\perp$ is a subspace.

- $dim(V) + dim(V^\perp) = n$ (where $V \in \mathbb{R}^n$).
- $(V^\perp)^\perp = V$

If V and W are orthogonal complements of each other, they form an orthogonal decomposition and every vector $x \in R^n$ can be written uniquely as $x = v + w$ where $v \in V, w \in W, v \cdot w = 0$.

5.6 Matrix Decomposition

$N(A)$ and $C(A^T) \in \mathbb{R}^n$ are orthogonal complements of each other, as are $C(A) \in \mathbb{R}^m$ and $N(A^T) \in \mathbb{R}^m$.

For a matrix $A \in \mathbb{R}^{m \times n}$ with $rank(A) = m$, we can decompose a point x into these two subspaces:

- $P = A^T(AA^T)^{-1}A$ is the orthogonal projection onto $C(A^T)$
- $Q = I - A^T(AA^T)^{-1}A$ is the projection onto $N(A)$.

Notice that $P^2 = P$ and $P + Q = I$.

6 Equivalences

6.1 Tall Matrices ($A \in \mathbb{R}^{m \times n}, m \geq n$)

1. Columns of A are LI
2. If solvable, $Ax = b$ has a unique solution
3. A has a left inverse
4. $N(A) = \{0\}$
5. rank$(A) = n$ (this is called "full column rank")

6.2 Wide Matrices ($A \in \mathbb{R}^{m \times n}, m \leq n$)

1. Rows of A are LI
2. $Ax = b$ is solvable for every b
3. A has a right inverse
4. $C(A) = \mathbb{R}^m$
5. rank$(A) = m$ (this is called "full row rank")

6.3 Square Matrices ($A \in \mathbb{R}^{n \times n}$)

1. Rows of A are LI
2. Columns of A are LI
3. $Ax = b$ always has a unique solution
4. A is invertible (has a left and a right inverse)
5. $N(A) = \{0\}$
6. $C(A) = \mathbb{R}^n$
7. rank$(A) = n$ (this is called "full rank")
8. $\det A \neq 0$

7 Determinant

The determinant of a square matrix, geometrically, gives the ratio between the *volume* of a set of points before and after application of the matrix.

- Rotations preserve the determinant
- A matrix is invertible/nonsingular iff $\det \neq 0$.
- For square A, B, $\det(AB) = \det(A)\det(B)$.
- $\det(A+B) \neq \det(A) + \det(B)$.
- $\det(A) = \det(A^T)$
- For a triangular (or diagonal) matrix, $\det A$ is the product of the diagonal.
- For a block triangular matrix, such as $\begin{smallmatrix} B & C \\ 0 & D \end{smallmatrix}$, $\det A = \det(B)\det(D)$.

- Swapping rows negates, adding rows preserves. Multiplying row by c multiplies by a factor of c.

We can also compute the determinant using cofactor expansion. Where $\text{sub}(A, i, j)$ gives A except for row i and column j, and for any row i (or column, via A^T):

$$\det A = \sum_{j=1}^{n} (-1)^{i+j} A_{ij} \det(\text{sub(A, i, j)})$$

8 Least Squares

For $Ax = b$, the following minimizes the error if overdetermined, or the norm if underdetermined.

$$x^* = (A^T A)^{-1} A^T b$$

Sometimes, we care about the size of x, and want to instead minimize $\|Ax - b\|^2 + \lambda \|x\|^2$, where λ is a regularization parameter. $x^* = (A^T A + \lambda I)^{-1} A^T b$ always exists and gives this.

9 Singular Value Decomposition

For any matrix $A \in \mathbb{R}^{n \times m}$, we can write:

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

Where $U \in \mathbb{R}^{n \times n} = [v_1 \dots v_n]$ and $V \in \mathbb{R}^{m \times n} = [v_1 \dots v_m]$ are orthogonal matrices (orthonormal columns) and $\Sigma \in \mathbb{R}^{n \times m}$ is a rectangular matrix with $\sigma_1, \sigma_2, \dots, \sigma_r, \sigma_{r+1}, \dots$ along the main diagonal. These σ_i are in decreasing order up to some σ_r, and then are all zero.

The σ_i are singular values and u_i and v_i are singular vectors. We can think of $U \Sigma V^T x$ as first rotating x by V^T, then scaling it by Σ, and finally rotating it by U.

- $\text{rank}(A)$ is the number of nonzero singular values.
- $u_1, \dots, u_r$ are an orthonormal basis for $C(A)$.
- $v_{r+1}, \dots, v_m$ are an orthonormal basis for $N(A)$.
- If $A = U \Sigma V^T$, then $A^T = (U \Sigma V^T)^T = V \Sigma U^T$

9.1 Pseudoinverse

If A is square and invertible (full column rank), then $A^{-1} = V \Sigma^{-1} U^T$. We can derive this: $A^{-1} = (U \Sigma V^T)^{-1} = (V^T)^{-1} \Sigma^{-1} U^{-1}$ and since U and V^T are orthogonal, $A^{-1} = V \Sigma^{-1} U^T$. We can extend this for non-invertible matrices, defining the *pseudoinverse*:

$$A^+ = \sum_{i=1}^{r} \sigma_i^{-1} v_i u_i^T$$

- $AA^+ = \sum_{i=1}^{r} u_i u_i^T$ gives the projector onto $C(A)$.
- $A^+A = \sum_{i=1}^{r} v_i v_i^T$ gives the projector onto $N(A)^\perp$.

The following two properties uniquely define it:

- $AA^+A = A$ and $A^+AA^+ = A^+$.
- AA^+ and A^+A are symmetric.

9.2 Operator Norm

The operator norm of $A \in \mathbb{R}^{m \times n}$ is the maximum norm result of multiplying by a unit vector:

$$\|A\| := \max_{x \in \mathbb{R}^n, \|x\| = 1} \|Ax\|$$

- $\|A\|$ is equal to the largest singular value of A.
- If P and Q are orthogonal, then $\|PA\| = \|AQ\| = \|A\|$.
- $|Ax| \le \|A\||x|$, $\|AB\| \le \|A\|\|B\|$

9.3 Low Rank Approximation

If we want to create a k-rank matrix C that approximates $B = \sum_{i=1}^{r} \sigma_i u_i v_i^T$ such that $\|B - C\|$ is minimized, that minimum is σ_{k+1} and it is achieved by the truncated SVD:

$$C = \sum_{i=1}^{k} \sigma_i u_i v_i^T$$

9.4 Principal Component Analysis

For analyzing datasets, it can be useful to find a projection onto a lower-dimensional space that maximizes the spread of the data. That is, finding $c_1, \dots c_k$ such that:

$$\max_{c_1, \dots, c_k} \frac{1}{p} \sum_{i=1}^{p} \|C^T y_i\|^2$$

Where $C = [c_1 \ \dots \ c_k]$ and $c_1, \dots c_k$ are orthonormal. The best set of vectors is $u_1, \dots, u_k$ from the SVD of the data. This basis also minimizes the reconstruction error $\frac{1}{p} \sum_{i=1}^{p} \|y_i - \hat{y}_i\|^2$ where $\hat{y}_i$ is the projection onto the subspace defined by the c_is.

10 Eigenvalues and Eigenvectors

For a square $n \times n$ matrix A, we can find eigenvalues λ and corresponding eigenvectors $x \in \mathbb{R}^n$ where $Ax = \lambda x$. To find the eigenvalues, we find the roots of the characteristic polynomial, given by $\det(A - \lambda I)$, which is degree n. Given an eigenvalue, we can find the corresponding eigenvector by finding a vector in the nullspace of $A - \lambda I$, i.e. a vector for which the above holds.

Note that for 2×2 A, $p(\lambda) = \lambda^2 - \text{Tr}(A)\lambda + \det(A)$ and $\lambda = \frac{1}{2}\left(\text{tr(A)} \pm \sqrt{\text{tr}(A)^2 - 4\det(A)}\right)$.

- λ is an eigenvalue of $A \iff Av = \lambda v \iff (\lambda I - A)v = 0 \iff N(\lambda I - A) \neq \{0\} \iff \lambda I - A$ not invertible $\iff \det(\lambda I - A) = 0 \iff p(\lambda) = 0$.
- $\det A = \prod \lambda_i$, $\text{tr}(A) = \sum \lambda_i$

10.1 Diagonalization

For a matrix $A \in \mathbb{R}^{n \times n}$, it can sometimes be written as $A = TDT^{-1}$, where $D = \text{Diagonal}(\lambda_1, \dots, \lambda_n)$, and $T = [v_1 \cdots v_n]$ where v_i is the eigenvector associated with λ_i.

A is diagonalizable iff every eigenvalue has the same algebraic multiplicity and geometric multiplicity. λ has alg. mult. k if $p(t)$ has a factor $(t - \lambda)^k$, and λ has geom. mult. K if $\dim N(A - \lambda I) = K$.

- If A is diagonalizable s.t. $A = TDT^{-1}$ as above, then $A^k = TD^kT^{-1}$.
- For a polynomial or convergent power series, $q(A) = TD_qT^{-1}$ where $D_q = \text{Diagonal}(q(\lambda_1), \ldots)$.
- Cayley-Ham.: $p(\lambda) = \det(A - \lambda I) \implies p(A) = 0$.

11 Symmetric Matrices

A square matrix is symmetric if $A = A^T$. Then:

- The eigenvalues of A are real.
- The eigenvectors of A can be chosen to be orthogonal.
- A is diagonalizable, and $A = TDT^T$.
- The nonzero singular values of A are the square root of the eigenvalues of MM^T or M^TM.

11.1 Positive (Semi-)Definite

A symmetric matrix $A \in \mathbb{R}^{n\times n}$ is PSD if $x^TAx \geq 0 \forall x \in \mathbb{R}^n$. A is positive definite if $x^TAx > 0$ for nonzero x. A matrix is PSD if:

- All eigenvalues are ≥ 0 (> 0 for PD). For 2×2, it is equivalent that $\text{trace}(A) \geq 0$ and $\det A \geq 0$.
- It is the sum of two PSD matrices.
- It can be expressed as AA^T for $A \in \mathbb{R}^{n\times m}$.

12 Power Method

By choosing a random vector and repeatedly applying A to it and normalizing the result, we will converge on the largest eigenvector x_k and eigenvalue x_kAx_k. The convergence rate depends on the spectral gap $\frac{|\lambda_1|}{|\lambda_2|}$. We apply $x_{k+1} = \frac{Ax_k}{\|Ax_k\|}$.

13 Linear Dynamical Systems

If we have a system $x_k = Ax_{k-1}$, equivalently $x_k = A^kx_0$, then for large k the system evolves at λ_1^k where λ_1 is the largest eigenvalue.

14 Linear ODEs

For a system of linear ordinary differential equations:

$$\frac{d}{dt}x(t) = Ax(t), \quad x(0) = x_0$$

The solution is $x(t) = e^{At}x_0$. We define the matrix exponential equivalently to the scalar exponential:

$$e^{At} = I + At + \frac{A^2t^2}{2} + \frac{A^3t^3}{3!} + \cdots$$

$$= I + VDV^{-1}t + \frac{VD^2V^{-1}t^2}{2!} + \cdots = V(I + Dt + \frac{D^2t^2}{2!} + \cdots)V^{-1}$$

15 Convexity

A *set* C is convex if it contains the line segment between any two points in the set:

$$ta + (1-t)b \in C \forall a, b \in C, \forall t \in [0, 1]$$

- If C_1, C_2 are convex, then $C_1 \cap C_2$ is convex.
- If C is convex and ϕ is a linear (Ax) or affine ($Ax+b$) map, then $\phi(C)$ is convex.

15.1 Convex Functions

A function f is convex if the function is below the chord between any two points:

$$f(ta + (1-t)b) \leq tf(a) + (1-t)f(b) \forall t \in [0, 1]$$

$f(x)$ is convex if:

- $f(x) = c_1f_1(x) + c_2f_2(x)$ with f_1, f_2 convex.
- $f(x) = \max\{f_1(x), \ldots, f_n(x)\}$ with f_i convex.

If $f(x)$ is convex, then the following are convex sets:

- $S_\gamma = \{x \in \mathbb{R}^n : f(x) \leq \gamma\}$
- $\text{epi}f = \{(x, y) : x \in \mathbb{R}^n, y \in \mathbb{R} : f(x) \leq y\}$

f is convex if $f'' \geq e0$ or $\mathbf{H}_f$ PSD. For convex f, local minima are global so $\nabla f = 0$ is a minimum.

16 Gradient & Hessian

The gradient of a function $f(x_1, x_2, \ldots, x_d)$ is the d-dimensional vector with $\frac{\partial f}{\partial x_i}$ at each entry. The Hessian $\mathbf{H}_f$ is the $d \times d$ matrix with $\frac{\partial f}{\partial x_i x_j}$ at each i, j.

17 Quadratic Functions

Quadratic functions are of the form:

$$f(x) = x^TAx + b^Tx$$

Then, $\nabla f = 2Ax + b$ and $H(x) = 2A$. So, if A is PSD then f is convex. It is always possible to write A as a symmetric matrix.

17.1 Quadratic Programming

If we have a quadratic program like $\min \frac{1}{2}x^TPx + q^Tx$, we can solve it by setting the gradient equal to zero. If we add an equality constraint $Ax = b$, then we can solve:

$$\begin{bmatrix} P & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} x \\ c \end{bmatrix} = \begin{bmatrix} -q \\ b \end{bmatrix}$$

And the solution x will be optimal.

18 Gradient Descent

Gradient descent repeatedly applies the following:

$$x_{k+1} = x_k - \gamma \nabla f(x_k)$$

If f is a convex function with $\nabla^2 f = H$, and $\lambda_1 \geq \cdots \geq \lambda_n \geq 0$:

- If f is smooth and convex, with $\lambda_i \leq L$, then should choose $\gamma = \frac{1}{L}$. If f is quadratic, then it converges for any $0 < \gamma < \frac{2}{\lambda_1}$. This converges according to $f(x_k) - f(x^*) \leq \frac{L}{2k}\|x_0 - x^*\|^2$.
- If f is smooth and strongly convex, the best step size is $\gamma = \frac{2}{m+L}$ and we define $Q = \frac{L}{m}$. If f is quadratic, $m = \lambda_n$, $L = \lambda_1$. This converges according to $f(x_k) - f(x^*) \leq \frac{L}{2}\left(\frac{Q-1}{Q+1}\right)^{2k}\|x_0 - x^*\|^2$: much faster.