1 Linear Algebra

- Linear Independent: m vectors are **l.i.** if $c_1\mathbf{x_1}+...+c_m\mathbf{x_m} = \mathbf{0}$, only when $c_1 = ... = c_d = \mathbf{0}$
- Standard basis of $\mathbb{R}^d$ is composed of $e_1, ..., e_d$
- Euclidean length: $\sqrt{x_1^2 + \cdots + x_d^2}$

Norm function

A function $\|\cdot\| : \mathbb{R}^d \to \mathbb{R}$ is called a norm if it satisfies

1. $\|x\| \geq 0 \forall \mathbf{x} \in \mathbb{R}^d$ and for $\mathbf{x} \neq \mathbf{0}$ we have $\|x\| > 0$
2. $\forall c \in \mathbb{R}, \|cx\| = |c| \, \|x\|$
3. $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d, \|x+y\| \leq \|x\| + \|y\|$

l_p-norm function

The norm is defined as $\|x\|_p := (|x_1|^p + \cdots + |x_d|^p)^{\frac{1}{p}}$

Note: $\|x\|_p = \lim_{p\to\infty} \|x\|_p = \underset{1\leq i\leq d}{max}|x_i|$

Also, $l_p - norms$ are decreasing in $p \geq 1$, namely $1 \leq p \leq q \leq \infty \Rightarrow \|x\|_p \geq \|x\|_q$.

The inner product $\langle x, y\rangle = x^T y$ is positive-definite for all $\langle x, x\rangle$, symmetric for all $\mathbf{x}, \mathbf{y}$, and linear for all $\mathbf{x_1}, \mathbf{x_2}, \mathbf{y}, c \in \mathbb{R}$.

Cauchy-Schwartz inequality: $|\langle x, y\rangle| \leq \|x\| \, \|y\|$. As a result we have $\cos\theta = \frac{\langle x,y\rangle}{\|x\|\|y\|}$

Mutual Orthogonal

A set of vectors is called **mutually orthogonal** if $\forall i \neq j :< x^{(i)}, x^{(j)} >= 0$ and a set of mutually orthogonal vectors is linearly independent.

Note: we call a set orthonormal if $\langle x^{(i)}, x^{(j)}\rangle = \begin{cases} 1 & \text{if i =j} \\ 0 & \text{if i } \neq \text{j} \end{cases}$

Matrix notation: $A = \begin{pmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m,1} & a_{m,2} & \cdots & a_{m,n} \end{pmatrix}$

The product is defined as $[AB]_{i,j} = \sum_{k=1}^n a_{i,k} b_{k,j}$

- Range of matrix, the subspace of all vectors following from linear combinations of A's columns: $\mathcal{R}(A) := \{Ax : x \in \mathbb{R}^n\}$
- Rank of matrix: dimension of subspace $\mathcal{R}(A)$
- $Rank(A_{m\times n}) \leq min(m, n)$ and full-rank if $Rank(A) = min(m, n)$
- Null space of a matrix, the subspace of all vectors which A maps to $\mathbf{0}$: $\mathcal{N}(A) := \{x : Ax = \mathbf{0}\}$
- Determinant: $det(A) = \sum_{j=1}^n (-1)^{i+j} a_{i,j} det(A_{(i,j)})$
- Inverse of $\left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right] = \frac{1}{ad-bc}\left[\begin{smallmatrix} d & -b \\ -c & a \end{smallmatrix}\right]$

Equivalent sayings:

- A is invertible
- A is non-singular, $det(A) \neq 0$
- A is full rank, $Rank(A) = n$
- A has linearly independent rows or columns
- A has a zero null space $\mathcal{N}(A) = 0$
- A has full range $\mathcal{R}(A) = 0$

Some statements:

- Wrong: If $x \perp y$ and $x \perp z$ then $y \perp z$
- Wrong: If **x,y** are linearly independent, and **x,z** as well, then **y,z** are also linearly independent.
- Correct: If $x \perp y$ and $x \perp z$, then $x \perp (y+z)$
- Wrong: If **x,y,** are linearly independent and **x,z** as well, then $x, (y+z)$ are also linearly independent.

Eigenvectors and Eigenvalues

For a vector $\mathbf{v} \neq \mathbf{0}$, it's an eigenvector for its eigenvalue λ such that: $Av = \lambda v$

Can be calculated by solving $det(A - \lambda I_n)$

Spectral theorem

For a symmetric matrix A there exists a spectral decomposition. Such that

$$A = \sum_{i=1}^n \lambda_i v_i v_i^T = V\Lambda V^T.$$

, where $V = [v_1, ..., V_n]$ and $\Lambda = diag(\lambda_1, \ldots, \lambda_n)$ such that $\lambda_1 \geq \cdots \geq \lambda_n$ and v_i is normalized

A symmetric matrix is called PSD if for every vecotr $x \in \mathbb{R}^n$ we have $x^T Ax \geq 0$. Strictly definite if it holds strictly for $\mathbf{x} \neq \mathbf{0}$

Theorem: A is psd $\Leftrightarrow$ all its eigenvalues are non-negative $\Leftrightarrow$ we have a matrix H such that $A = HH^T$

Partial Order for Matrices

- $A \succeq B$ if $A - B$ is PSD
- $A \succ B$ if $A - B$ is positive definite (PD)
- $A \preceq B$ if $A - B$ is negative semidefinite
- $A \prec B$ if $A - B$ is negative definite

$A \not\succeq B$ does not imply $A \preceq B$

2 (Multivariable) Calculus Recap

Special sets

Epigraph: $epif := \{(\mathbf{x}, t) : \mathbf{x} \in \mathbb{R}^n, t \geq f(\mathbf{x})\}$
Contour set: $C_f(t) := \{\mathbf{x} \in \mathbb{R}^n : f(\mathbf{x}) = t\}$
Hyperplane: $H = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{a}^T x = b\}$
Halfspace (change $\leq$ to $\geq$ for $+$) $H_- = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{a}^T x \leq b\}$

A function is linear if $\forall x, y \in \mathbb{R}^n, c \in \mathbb{R} f(cx+y) = cf(x) + f(y)$

General Quadratic Form

We can write a quadratic function, in terms of $A \in \mathbb{R}^{n\times n}, b \in \mathbb{R}^n, c \in \mathbb{R}$:

$$f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T A\mathbf{x} + \mathbf{b}^T\mathbf{x} + c.$$

Gradient and Hessians

A function $f : \mathbb{R}^n \to \mathbb{R}$ has a gradient defined as $\nabla f(x) = \left[\frac{\delta f(x)}{\delta x_1}, \ldots, \frac{\delta f(x)}{\delta x_n}\right]^T$ and hessian is the second derivative (square matrix.)

Math Rules:

- Derivative for $f : \mathbb{R}^n \to \mathbb{R}^m$: $df(x) = \begin{bmatrix} \frac{df_1}{dx_1} & \cdots & \frac{df_1}{dx_n} \\ \frac{df_2}{dx_1} & \cdots & \frac{df_2}{dx_n} \\ \vdots & \ddots & \vdots \\ \frac{df_m}{dx_1} & \cdots & \frac{df_m}{dx_n} \end{bmatrix} \in \mathbb{R}^{m\times n}$
- Gradient: we can calculate gradient for $f : \mathbb{R}^n \to \mathbb{R}$. It's defined as $\nabla f(x) = df(x)^T$
- Chain rule: $d(g \circ f)(x) = dg(f(x)) \cdot df(x)$
- Product rule: $g(x)h(x) = dg(x)h(x) + g(x)dh(x)$
- $f(x) = \|Ax - b\|^2, \nabla f(x) = 2A^T(Ax - b), Hess_f(x) = 2A^T A$
- $f(x) = \|Ax - b\|, \nabla f(x) = \frac{A^T(Ax-b)}{\|Ax-b\|}$, $Hess_f(x) = \frac{A^T A}{\|Ax-b\|} - \frac{(A^T Ax-b)((x^T A^T - b^T)A)}{\|Ax-b\|^3}$
- $d\mathbf{0} = 0, d(\alpha\mathbf{X}) = \alpha d\mathbf{X}, d(\mathbf{X^{-1}}) = -\mathbf{X^{-1}}(\mathbf{dX})X^{-1}$, $d\mathbf{X}^T = (d\mathbf{X})^T$, $\frac{dx^T a}{dx} = \frac{da^T x}{dx} = a$, $\frac{dx^T Ax}{dx}(A + A^T)x$, $\frac{d}{ds}(x - As)^T W(x - As) = -2A^T W(x - As)$, $\frac{d}{dx}(x - As)^T W(x - As) = 2W(x - As)$

We can approximate functions using gradients and hessians: **First order Taylor:** $f(x) = f(\mathbf{x_0}) + \nabla f(\mathbf{x_0})^T(\mathbf{x} - \mathbf{x_0}) +$

$\epsilon(x)$ it holds that $lim_{x \to x_0} \frac{\epsilon(x)}{\|x-x_0\|} = 0$ **Second order Taylor:** $f(x) = f(\mathbf{x_0}) + \nabla f(\mathbf{x_0})^T(\mathbf{x}-\mathbf{x_0}) + \frac{1}{2}(\mathbf{x}-\mathbf{x_0})^T H_f(\mathbf{x_0})(\mathbf{x}-\mathbf{x_0}) + \epsilon_2(x)$ it holds that $lim_{x \to x_0} \frac{\epsilon(x)_2}{\|x-x_0\|^2} = 0$

Affine Functions

Let $f \in \mathbb{R}^d \to \mathbb{R}$ be a multivariable function. Then f is an affine function iff: $\forall x, y \in \mathbb{R}^d, \theta \in [0,1] : f(\theta x + (1-\theta)y) = \theta f(x) + (1-\theta) f(y)$

Convex function (one dimensional): if we can replace $=$ above with $\leq$. A function f is **concave** if $-f(x)$ is convex.
Convex (multivariable): $||\mathbf{x}||$, $||\mathbf{x}||^p, p \geq 1$, $\log(\sum_i^d e^{x_i})$.
Concave: $\sum_i^d x_i log(\frac{1}{x_i})$,
$dom(f) = R_+^d$, $(x_1 x_2 \dots x_d)^{\frac{1}{d}}, dom(f) = \mathbb{R}_+^d$
Proposition: **A set S is convex** iff $\forall x, y \in S, \theta \in [0,1] : \theta x + (1-\theta)y \in S$
Examples of convex functions: $x^p, p \geq 1$ or $p \leq 0$, $|x|^p, p \geq 1$, e^{ax+b}. **Concave:** $x^p, x \in \mathbb{R}_+, 0 \leq p \leq 1$

Sublevel set

Let $S_f(t) = \{\mathbf{x} \in dom(f) : f(\mathbf{x}) \leq t$. If f is convex, then $S_f(t)$ is a convex set for every t.

Examples of convex sets: hyperplanes, halfspaces, norm balls ($\{\mathbf{x} : \|x\| \leq \epsilon$)
Convexity preserving operations: intersection, affine and inverse-affine mappings, linear fractional functions
How to proof a function is convex:

1. Verify the inequality
2. Proof over epigraph and sub-level sets
3. Gradients and hessians
4. Convexity preserving operations

Convexity preserving operations: positive scalar multiplication, addition of two convex functions, composition with affine functions $f(\mathbf{Ax}+\mathbf{b})$, pointwise maximum $max\{f_1(x), \dots, f_k(\mathbf{x})\}$ if each f_i is convex, composition $f(g(x))$, where f,g both convex and f non-decreasing in every entry.

First-order convexity condition

A differentiable function f is convex iff its domain is convex and $\forall x, y \in dom(f) : f(y) \geq f(x) + \nabla f(x)^T(y-x)$

Second-order convexity

Same as above, but this $\forall \mathbf{x} \in dom(f) : H_f(\mathbf{x}) \succeq 0$ (so PSD)

3 Optimization Problems

Let the optimization problem be formulated as $\underset{x \in \mathbb{R}^d}{min} f(x)$ subject to $g_i(x) \leq 0$ for all i

- NP-Hard problems: *proven* to be intractable
- Linear Programming Problems: if a problem can be rewritten as $\underset{x \in \mathbb{R}^d}{min} \mathbf{c}^T\mathbf{x}$ subject to $A\mathbf{x} \leq \mathbf{b}$ and $\mathbf{x} \geq \mathbf{0}$, with f and $g_i \forall i$ affine. If f is non-Affine, then it's a non linear programming task
- Convex optimization problem: if f, g_i are all convex and we can rewrite it as $\min_{x \in \mathbb{R}^d} f(x)$ subject to $g_(x) \leq 0$ for all i or $Ax = b$, else non-convex. Also if we have a constraint with equality, it must be affine.

How to

- Define optimization variables, i.e. $x \in \mathbb{R}^d$
- Define objective function
- Define feasible set, or constraint functions, also must $x_i \geq 0$ for example?

Examples:

- LP: Transport task, manufacturing task, sorting task
- Convex problems: LP-problems, projection problem, distance computation problem, ridge regression

Definition: two problems are called *equivalent* if their optimal solutions are in one-to-one correspondence.

- Two problems are **equivalent** if their optimal solutions are in one-to-one correspondence
- For above, sometimes a non-convex task we can find an equivalent convex task.
- **Feasible set:** $S = \{x \in \mathbb{R}^d : g_i(x) \leq$ for all $1 \leq i \leq m\}$ (the set satisfying the constraint functions)
- Locally optimal solution $(x^*) : S \cap \{x : \|x - x^*\| \leq \epsilon\}$ for some $\epsilon > 0$ (usually easy to compute)
- Globally optimal solution $(x^*) : x^* \in S : \forall x \in S : f(x^*) \leq f(x)$
- In convex optimization problems, every local optimum is also a global optimum. **Proof by contradiction:** assume x_0 is locally optimal, and x^* globally optimal. We know $\theta x_0 + (1-\theta)x^* \in S$, thus $f(\theta x_0 + (1-\theta)x^*) \leq \theta f(x_0) + (1-\theta) f(x^*) < \theta f(x^*) + (1-\theta) f(x^*) = f(x^*)$, which means x^* isn't the global optimum.
- The feasible set, is also a convex set

4 Extra

Prove PSD (Sylvesters Criterion)

One way to show PSD, is by x^TAx and usually showing it's a norm squared. Let $A^{(k)}$ be the $k \times k$ submatrix from topleft, let $A^{(1)} = [a_{11}], A^{(n)} = A$, and $\Delta_k = det(A^{(k)})$. (NOT ALWAYS CONCLUSIVE METHOD)

- A spd $\Leftrightarrow \Delta_i > 0, ..., \Delta_n > 0$
- A snd $\Leftrightarrow (-1)^1\Delta_1 > 0, ..., (-1)^n\Delta_n > 0$

Gram-Schmidt: orthogonalize a basis

Given some vectors for a basis of a subspace $S \subseteq \mathbb{R}^n$, to get an orthogonal basis we can use the following on all vectors.

1. Let $proj_u(v) = \frac{\langle u,v \rangle}{\langle u,u \rangle} u$
2. $u_1 = v_1 \Rightarrow e_1 = \frac{u_1}{\|u_1\|}$
3. $u_2 = v_2 - proj_{u_1}(v_2) \Rightarrow e_2 = \frac{u_2}{\|u_2\|}$
4. ...
5. $u_k = v_k - \sum_{j=1}^{k-1} proj_{u_j}(v_k) \Rightarrow e_k = \frac{u_k}{\|u_k\|}$

Find eigenvectors

1. Solve for λ in $det(A - \lambda I) \stackrel{!}{=} 0$ (these are eigenvalues)
2. To get all eigenvectors, set $\lambda = \lambda_i$ in $(A - \lambda I)\mathbf{x} \stackrel{!}{=} \mathbf{0}$ for all eigenvalues we got.
3. Perform gauss elimination, and then get the systems of equations wrt one variable and set the variable to 1
4. If necessary, normalize (for ex. spectral dec.)

Example: $\begin{bmatrix} 2 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 2 \end{bmatrix}$ has eigenvectors $\lambda_1 = 2, \lambda_2 = 2 - \sqrt{2}, \lambda_3 = 2 + \sqrt{2}$. Plugging in λ_1 gives $\begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} \Rightarrow y = 0, x = -z, y = 0$ thus the eigenvector $[1, 0, -1]^T$

Suppose $A, B \in \mathbb{R}^{n \times n}$ are symmetric PSD matrices:

- also PSD: $A + B$, $A + I_n$, A^{-1}
- not PSD: AB (only when the product is symmetric)

Extra: For a unit vector $\mathbf{v}(\| \mathbf{v} \| = 1), \epsilon > 0$ it holds that $f(x_0 + \epsilon v) \approx f(x_0) + \epsilon \nabla f(x_0)^T \mathbf{v}$. We have a *maximal rate of local variation* along the gradient. In contrast, *zero rate of variation* along any direction orthogonal to the gradient. Gradient are orthogonal to contour sets