Applied Modeling and Optimization Part 2

(class 7) 03/25/2021

Conjugate Direction Method and Duality

Note: ① For multi-dimensional optimization, the search direction d_k at iteration k depends on the local properties of $f(x)$, gradient, Hessian

steepest descent

② Relationship between d_k and d_{k+1} is coincident, not by design

(e.g.) in steepest-descent, $d_k \perp d_{k+1}$

if α_k is s.t $\min_{\alpha_k} f(\underbrace{x_k + \alpha_k d_k}_{x_{k+1}})$

Can we impose relationships between d_k and d_{k+1} ?

Conjugate Directions

If $f(x) \in C^1$, $x \in \mathbb{R}^n$ $x = [x_1 \; x_2 \; \cdots x_n]^T$

and $\min f(x)$, $x \in \mathbb{R}^n$

Algorithm (co-ordinated - descent algorithm)

Step 1: input x_1 and initialize tolerance $\epsilon > 0$ set $k = 1$

Step 2: set $d_k = [0 \; 0 \cdots 0 \; d_k \; 0 \cdots 0]^T \in \mathbb{R}^n$

only the k^{th} element is non-zero

Step 3: Find α_k, the value of α that minimizes $f(x_k + \alpha d_k)$, using line search.

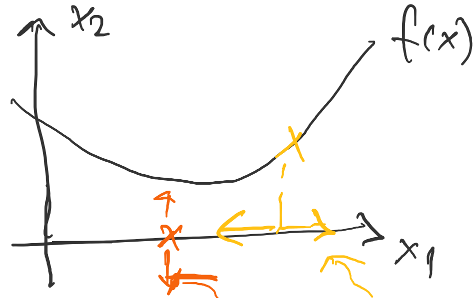
Set $x_{k+1} = x_k + \alpha_k d_k$

Calculate $f_{k+1} = f(x_{k+1})$

Step 4: If $\|\alpha_k d_k\| < \epsilon$ then output $x^* = x_{k+1}$ and $f(x^*) = f_{k+1}$ and stop.

Step 5: If $k = n$, set $x_1 = x_{k+1}$, $k = 1$ and go to step 2, otherwise put $k = k+1$ and go to step 2.

Example: $n=2$, $x=[x_1 \quad x_2]^T$



$$d_k = \begin{bmatrix} d_1 \\ d_2 \end{bmatrix}$$

$k=1$ $\quad d_1 = \begin{bmatrix} d_1 \\ 0 \end{bmatrix} \Rightarrow$ search only along the x_1-axis

$$x_2 = x_1 + \alpha_1 d_1 = \begin{bmatrix} x_1^{(1)} \\ x_2^{(1)} \end{bmatrix} + \alpha \begin{bmatrix} d_1 \\ 0 \end{bmatrix}$$

$$= \begin{bmatrix} x_1^{(1)} + \alpha_1 d \\ x_2^{(2)} \end{bmatrix}$$

$k=2$ $\quad d_2 = \begin{bmatrix} 0 \\ d_2 \end{bmatrix} \Rightarrow$ search only along x_2-axis.

Example: $d_k = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$, $d_{k+1} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$

Disadvantages:

1. not effective in practice
2. unreliable technique because of oscillations under some conditions.

Therefore conjugate directions → More effective algorithms.

Definitions: (a) Two distinct non-zero vectors d_1 and d_2 are said to be conjugate w.r.t a real symmetric matrix H if $d_1^T H d_2 = 0$ (generalized notion of ortogonality)

Recall:

$$d_1^T d_2 = 0 \quad \text{iff} \quad d_1 \perp^r d_2$$

(b) A finite set of distinct non-zero vectors $\{d_0, d_1, \ldots, d_k\}$ is said to ~~be~~ be conjugate w.r.t. a real symmetric matrix H if $d_i^T H d_j = 0$, $\forall i \neq j$

If we have several vectors, and if we choose two of them and they satisfies the condition $d_i^T H d_j = 0$ These are conjugate vectors.

Example: Let $H = I_n$ =

$$\begin{bmatrix} 1 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & & \cdots & 0 \\ 0 & 0 & 1 & & \cdots & \vdots \\ \vdots & & & & & \vdots \\ 0 & & \cdots & & & 1 \end{bmatrix}_{n \times n}$$

identity matrix

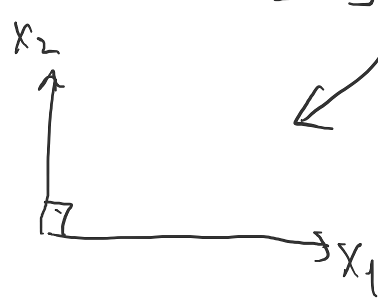
→ real and symmetric

$H^T = H$

Then $d_i^T H d_j = d_i^T I_n d_j = d_i^T d_j = 0$, $\forall i \neq j$

$$\Rightarrow d_i^T \perp^r d_j$$

$\begin{bmatrix} x_1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ x_2 \end{bmatrix}$



Example: Suppose d_j, $j=0,1,\ldots,k$ are eigenvectors of H.

eigenvalues of H

Then $H d_j = \lambda_j d_j$

eigenvectors of H

$j=0,1,\ldots,k$

$\therefore\ d_i^T H d_j = \lambda_j d_i^T d_j$ (from ①), $\forall i \neq j$

$= 0$ (eigenvectors are orthogonal to each other)

$\Rightarrow d_i$ and d_j are conjugate with respect to H.

Theorem: Conjugate vectors are linearly independent

Linear Independence

$a_1, \ldots, a_k$ are linearly independent

if $\alpha_1 a_1 + \alpha_2 a_2 + \cdots + \alpha_k a_k = 0$

$\alpha_1 = \alpha_2 = \cdots = \alpha_k = 0$

★ There is no linear relationship, it means linearly independent vector

Proof: $\sum_{j=0}^{k} \alpha_j d_j = 0$ ——— ①

$\alpha_j = 0$, $\forall j$

Pre-multiply (1) by $d_i^T H$, $0 \le i \le k$

$$\Rightarrow \sum_{j=0}^{k} \alpha_j d_i^T H d_j = \alpha_i \underbrace{d_i^T H d_i}_{\geq 0} = 0 \quad (\text{by } (1))$$

But, $d_i^T H d_i \geq 0$ because H is positive semi-defini

$$\therefore \alpha_i \underbrace{d_i^T H d_i}_{\geq 0} = 0 \Rightarrow \alpha_i = 0, \forall i$$

$\therefore \{d_j\}$ are linearly independent

Application Quadratic Programming

$$\min_x f(x) = a + x^T b + \frac{1}{2} x^T H x \quad (\text{Quadratic func})$$

$$(\text{e.g.}) = f(x_1, x_2) = x_1^2 + 2x_2^2 - 2x_1x_2 + \frac{1}{3}x_1 - 5x_2 + 6$$

Note: $a = f(0)$

$b = \nabla f(x)$ at $x = 0$

H is Hession of $f(x)$

$$\nabla f(x) = g(x) = b + Hx \quad (1)$$

→ gradient of the function

at the minimizer x^*, $\nabla f(x^*) = 0$ → first order nec. cond.

$$\Rightarrow b + H x^* = 0 \Rightarrow H x^* = -b \quad -(2)$$

If $d_0, d_1, \ldots, d_{n-1}$ are distinct conjugate directions in $\mathbb{R}^n$

$\Rightarrow$ they form a basis of $\mathbb{R}^n$

$\Rightarrow \{d_0, d_1, \ldots, d_{n-1}\}$ span $\mathbb{R}^n$

$$\therefore x^* \in \mathbb{R}^n \Rightarrow x^* = \sum_{i=0}^{n-1} \alpha_i d_i$$

If H is positive definite then

$$d_k^T H x^* = \sum_{i=0}^{n-1} \alpha_i d_k^T H d_i \quad - \text{from } (3)$$

$$= \alpha_k d_k^T H d_k$$

$$\Rightarrow \alpha_k = \frac{d_k^T H x^*}{d_k^T H d_k}$$

$$\boxed{\alpha_k = \frac{-d_k^T b}{d_k^T H d_k}} \quad - \text{from } (2)$$

basis of $\mathbb{R}^n \Rightarrow$ set of vectors that linearly independent and span $\mathbb{R}^n$

ie, generate every point in $\mathbb{R}^n$

linear combination.

(eg) Basic of $\mathbb{R}^2$

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = x_1 \begin{bmatrix} 1 \\ 0 \end{bmatrix} + x_2 \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

Basis $\left\{ \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix} \right\}$

Basis $\left\{ \begin{bmatrix} 2 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 2 \end{bmatrix} \right\}$

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \frac{x_1}{2} \begin{bmatrix} 2 \\ 0 \end{bmatrix} + \frac{x_2}{2} \begin{bmatrix} 0 \\ 2 \end{bmatrix}$$

$$\therefore \boxed{x^* = -\sum_{k=0}^{n-1} \frac{d_k^T b}{d_k^T H d_k} d_k}$$

close form solution

★★ If you have quadratic function, we don't have to implement numerical technic. we can simply take the Hessian of the function, computed eigen vectors and simply plug it in, you get optimal solution.

Duality (Lagrange Duality):

Primal Problem: $\min_x f_0(x)$ s.t $f_i(x) \leq 0$, $i=1,\dots,m$

* we are not going to do any assumption about concexity. If this is convex opt. problem obj. func. is convex constraints are convex. Nec. condition is also sufficient

1. No concexity assumption
2. Assume no equality constraints for simplicity (without loss of generality)

Re-write Primal Problem as: → Primal Problem

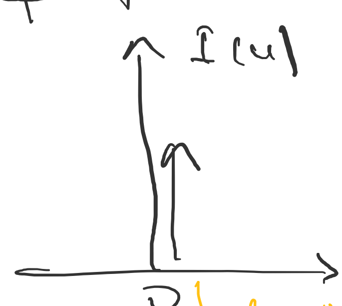
$$J(x) = \begin{cases} f(x), & \text{if } f_i(x) \leq 0, \ \forall i \\ \infty, & \text{otherwise} \end{cases}$$

If any of the constraints are not satisfied, we have infinite penalty (Huge penalty)

$$\Rightarrow J(x) = f_0(x) + \sum_{i=1}^{m} I[f_i(x)]$$

$I[u]$ is an infinite step function:

$$I[u] = \begin{cases} 0, & \text{if } u \leq 0 \\ \infty, & \text{otherwise} \end{cases}$$



↳ for anything else goes up

$I[u]$ is a penalty function

But $I[u]$ is discontinuous at $u=0$ and non-differentiable

$\therefore$ Replace $I[u]$ by a "nice" penalty function.

$\therefore$ Let's choose a linear penalty as $\lambda u, \lambda \geq 0$

$$L(x,\lambda) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) \quad (2)$$

$\rightarrow$ Lagrangian function

$\rightarrow \therefore$ Lagrange multipliers is the "cost"/penalty for violating a constraint

Standard form

min $f_0(x)$ s.t.

$f_i(x) \geq 0, \; i=1,\dots,m$

$$\Rightarrow L(x,\lambda) = f_0(x) - \sum_{i=1}^{m} \lambda_i f_i(x)$$

How do we relate this to the primal problem?

Claim: $\max\limits_{\lambda} L(x,\lambda) = J(x)$ (primal problem)

case (a) if all the constraints are satisfied,

$\lambda_i = 0, \forall i$ $\rightarrow$ penalty parameter

$\Rightarrow L(x,0) = f_0(x)$ if $f_i(x) \leq 0, \forall i$ (from (1))

$\rightarrow$ Primal Problem

case b : Suppose i^{th} constraint is not satisfied. then

$f_i(x) \geq 0$.

Then make the penalty λ_i very high, ie $\lambda_i \to \infty$

$\Rightarrow \mathcal{L}(x, \lambda) \to \infty$ since $\lambda_i \to \infty$

$\therefore$ From a and b $\max_{\lambda} \mathcal{L}(x, \lambda) = J(x)$

But goal is $\min_{x} J(x)$

$$\Rightarrow \min_{x} \underbrace{\left[\max_{\lambda} \mathcal{L}(x, \lambda) \right]}_{= J(x)}$$

$\to$ Hard to solve

$\therefore$ Try reverse direction:

$$\max_{\lambda} \underbrace{\left[\min_{x} \mathcal{L}(x, \lambda) \right]}_{= g(x)}$$

"Dual Function"

$$g(\lambda) = \min_{x} \mathcal{L}(x, \lambda)$$

$$\boxed{\max_{\lambda} \; g(\lambda)}$$ → "Dual Problem"

* if this is a concave func. It is easy to solve also solution is guarante. Max is guarante, uniqueness is guarantee

Note: $g(\lambda)$ is a point wise minimum of affine functions ($\because \mathcal{L}(x,\lambda)$ is affine or linear in λ)

$\Rightarrow g(\lambda)$ is a concave function

maximizing a concave function (Dual Problem) is easy. $\therefore$ Solving the dual problem is easy.

Note: $\lambda u \leq I(u)$

$\Rightarrow \mathcal{L}(x,\lambda) \leq J(x)$

Solution to Primal Problem

$\Rightarrow \min_{x} \mathcal{L}(x,\lambda) = g(\lambda) \leq \min_{x} J(x) = p^*$

$\Rightarrow g(x) \leq p^*$

$\Rightarrow d^* = \max_{\lambda} g(\lambda) \leq p^*$

Solution to dual problem

$$\boxed{\therefore \; d^* \leq p^*}$$ → "weak duality"

Solution of the dual problem always lower bound the primal problem.

($p^* - d^* \geq 0$ is called the duality gap)

↳ if the gap is very large is not a good approximation of primal problem.

$p^* = d^*$ then duality gap is zero for convex opt. problems. (strong duality)

Convex Opt. Problems (Strong Duality):

$$\boxed{\max_{\lambda} \min_{x} L(x,\lambda) = \min_{x} \max_{\lambda} L(x,\lambda)}$$

"$\leq$" weak duality

Duality, Inf, Sup, Subgradient

Recap: Lagrange Dual

Primal Problem: $\min_{x} f(x)$

s.t. $g(x) \leq 0$ and $h(x) = 0$

Lagrange Dual Problem: $\max\limits_{\lambda,\mu} \theta(\lambda,\mu)$ s.t $\mu \geq 0$

Step 2

Step 1

where $\theta(\lambda,\mu) = \min\left[f(x) + \mu^T g(x) + \lambda^T h(x)\right]$

Note: If $g(x) \geq 0$ then replace μ by $-\mu$

Example: $\min\limits_{x_1,x_2} \; x_1^2 + x_2^2$ s.t. $x_1 + x_2 \geq 4$

$x_1, x_2 \geq 0$

Primal Problem

Lagrangian: $L(x,\mu) = x_1^2 + x_2^2 + \mu\left[4 - x_1 - x_2\right]$

Feasible Set: $S = \left\{x \in \mathbb{R}^2 \mid x_1, x_2 \geq 0 \,;\, x_1 + x_2 \geq 4\right\}$

Lagrange Dual Function: $\theta(\mu) = \min\limits_{x_1,x_2 \in S} L(x,\mu)$

$$= \min_{x_1,x_2}\left[x_1^2 + x_2^2 + \mu(4 - x_1 - x_2)\right]$$

$$= 4\mu + \min_{x_1,x_2 \in S}\left[x_1^2 + x_2^2 - \mu x_1 - \mu x_2\right]$$

$$Q(\mu) = 4\mu + \min_{x_1 \geq 0}\left[x_1^2 - \mu x_1\right] + \min_{x_2 \geq 0}\left[x_2^2 - \mu x_2\right]$$

(seperable objective function)

For fixed value $\mu \geq 0$:

$$\min_{x_1 \geq 0}\left[x_1^2 - \mu x_1\right] \Rightarrow 2x_1 - \mu = 0$$ (derivative $= 0$) First ord. nec. condi

$$\Rightarrow x_1^* = \frac{\mu}{2} \quad \text{— } 1$$

$$x_1^2 - \mu x_1^* = \left(\frac{\mu}{2}\right)^2 - \mu\left(\frac{\mu}{2}\right) = \mu^2/4$$

2

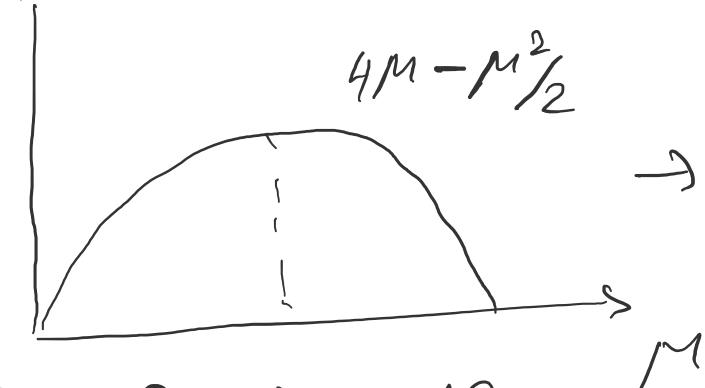
$$\min_{x_2 \geq 0}\left[x_2^2 - \mu x_2\right] \Rightarrow 2x_2 - \mu = 0 \Rightarrow x_2^* = \frac{\mu}{2}$$

$$x_2^* - \mu x_2^* = \frac{\mu^2}{4}$$

plunged in the function

$$\therefore\ Q(\mu) = L(x(\mu), \mu) = 4\mu - \frac{\mu^2}{2}, \quad \forall\, \mu \geq 0$$

$Q(\mu)$ → Dual Function



→ Concave Function

Step2:

$$\max_{\mu \geq 0} Q(\mu) \Rightarrow \frac{dQ}{d\mu} = 0$$

$$\Rightarrow 4 - \mu^* = 0$$

$$\Rightarrow \mu^* = 4 \quad , \quad \theta(\mu^*) = 8$$

from 1 and 2 $\quad x_1(\mu^*) = \frac{\mu^*}{2} = \frac{4}{2} = 2$

$$x_2(\mu^*) = \frac{4}{2} = 2$$

$$x^* = \begin{bmatrix} 2 \\ 2 \end{bmatrix}$$

Example: $\min\limits_{x_1, x_2, x_3} \; 3x_1 + 7x_2 + 10x_3$

$$\text{s.t.} \quad x_1 + 3x_2 + 5x_3 \geq 7$$

$$x_1, x_2, x_3 \in \{0, 1\}$$

(Binary Variables)

Feasible Set, $S = \{ x \in \mathbb{R}^3 \mid x_j \in \{0,1\}, \, j = 1,2,3 \}$

Lagrangian:

$$\mathcal{L}(x, \mu) = 3x_1 + 7x_2 + 10x_3 + \mu \left[7 - x_1 + 3x_2 + 5x_3 \right]$$

Lagrangian Dual:

$$\theta(\mu) = \min_{x \in S} \mathcal{L}(x,\mu)$$

$$= \min_{x_1,x_2,x_3 \in S} \left[3x_1 + 7x_2 + 10x_3 + \mu(7 - x_1 - 3x_2 - 5x_3)\right]$$

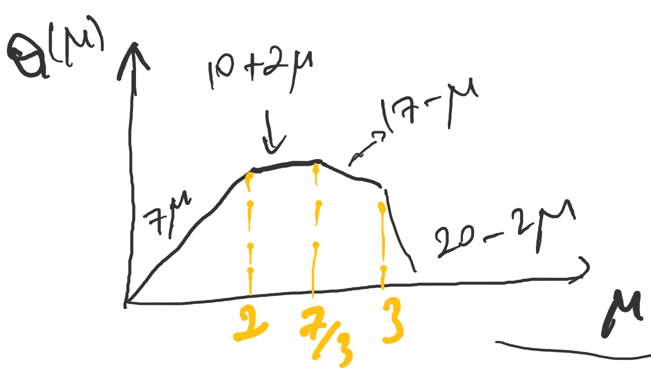
$$= 7\mu + \min_{x_1 \in \{0,1\}} \left[\overbrace{(3-\mu)}^{\text{co-eff}} x_1\right] + \min_{x_2 \in \{0,1\}} \left[\overbrace{(7-3\mu)}^{\text{co-eff}} x_2\right] +$$

$$\min_{x_3 \in \{0,1\}} \left[\overbrace{(10-5\mu)}^{\text{co-eff}} x_3\right]$$

Put $x_f = \begin{cases} 1, \text{ when co-eff} \le 0 \\ 0, \text{ when co-eff} \ge 0 \end{cases}$

$3-\mu \ge 0 \quad \mu \ge 3$
$7-3\mu \ge 0 \quad \mu \ge \frac{7}{3}$
$10-5\mu \ge 0 \quad \mu \le 2$

μ	$x_1(\mu)$	$x_2(\mu)$	$x_3(\mu)$	$\mathcal{L}(x(\mu))$	$\theta(\mu)$
$[-\infty, 2]$	0	0	0	7	7μ
$[2, 7/3]$	0	0	1	2	$10+2\mu$
$[\frac{7}{3}, 3]$	0	1	1	-1	$17-\mu$
$[3, \infty]$	1	1	1	-2	$20-2\mu$



→ $\theta(\mu)$ is concave but
→ non-differentiable at break points

→ Break points $\mu \in \{2, 7/3, 3\}$

→ Slope equals the value of the constraint function.

$\therefore\ \mu^* = \frac{7}{3}\ ,\ \theta(\mu^*) = \frac{44}{3}$

$\Rightarrow x^* = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}\ ,\ f(x^*) = 17$

→ Positive duality gap

→ Feasible solutions: $x(\mu) = \left\{ \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix} \right\}$

— Lagrandian Extra Examples in Canvas —

Infimum (inf) and Supremium (sup):

→ min, → max

→ Generalizations of min (inf) and max (sup)

→ Problems may not have min or max for different reasons.

Case (1): $\min_{x \in S} f(x)$: Suppose a min exists, $f^* = f(x^*)$

$\Rightarrow$ Every number $f_1 \le f^*$ is a lower bound for the problem.

$\therefore$ The greatest lower bound $= f^*$

$\Rightarrow f^*$ is the greatest lower bound for $\min_{x \in S} f(x)$

Example: $f(x) = x \quad 0 \le x \le 1$

$\min_x f(x) = 0$ Lower bound $\{\dots, -2, -1.5, -1, -0.5 \dots 0\} \Rightarrow glb = 0$

by definition min/max must be a finite number.

Case 2: $\min_{x \in S} f(x)$ does not exist but $f(x)$ is bounded below on S. The set of all lower bounds always has a well-defined largest element, the greatest lower bound for $\min_{x \in S} f(x)$

Example: $\min_{x} f(x) = x \quad , \quad 0 < x \leq 1$

$\Rightarrow \min_{x} f(x)$ does not exist since $x \neq 0$

but $0 < f(x)$ lower bound exists

$\therefore$ Lower bounds for $\min_{x} f(x) = \{\ldots -1, 0\}$

$\Rightarrow$ greatest lower bound $= 0$

$\therefore \inf_{x} f(x) \neq \min_{x} f(x)$

greatest lower bound

Case 3) : $\min_{x \in S} f(x)$ is unbounded below. Then there is no real number that is a lower bound to the problem

Example: $\min_x f(x) = x, \quad x \leq 1$

No min exists

Case 4) : Empty feasible set (ie, infeasible opt. problem)

$\Rightarrow S \neq \emptyset$

We adopt the convention that every real number satisfies the definition of a lower bound.

Consider extended real-line, $\bar{\mathbb{R}} = \mathbb{R} \cup \{-\infty, \infty\}$

Define: Let $S \subset \mathbb{R}^n$, $f: S \to \mathbb{R}$. Then $\inf_{x \in S} f(x)$ then infimum of $\min_{x \in S} f(x)$ is defined as:

$$\inf_{x \in S} f(x) = \begin{cases} \text{g.l.b for } \min_{x \in S} f(x) \text{ if } \min_{x \in S} f(x) \text{ is bounded below} \\ -\infty \text{ if the } \min_{x \in S} f(x) \text{ is unbounded below} \\ \infty, \text{ if } \min_{x \in S} f(x) \text{ is infeasible} \end{cases}$$

$\left| \; \min_{x \in S} f(x) \right.$ if a minimum exists

Note: $\inf_{x \in S} f(x)$ always exists even if $\min_{x \in S} f(x)$ exists or not.

Similarly, $\sup_{x \in S} f(x)$ <u>always exists</u> even if $\max_{x \in S} f(x)$ exists or not.

least upper bound

$\therefore$ write $\inf_{x \in S} f(x)$ instead of $\min_{x \in S} f(x)$

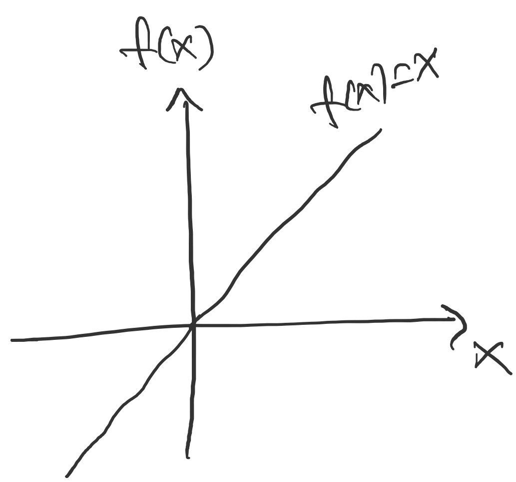
$\sup_{x \in S} f(x)$ instead of $\max_{x \in S} f(x)$

<u>**Example:**</u> Unconstrained problem with unbounded objective

$f: \mathbb{R} \to \mathbb{R}$ (ie, $S = \mathbb{R}$)

$f(x) = x, \quad \forall x \in \mathbb{R}$

$\min_{x \in \mathbb{R}} f(x) \Rightarrow$ no min exists.



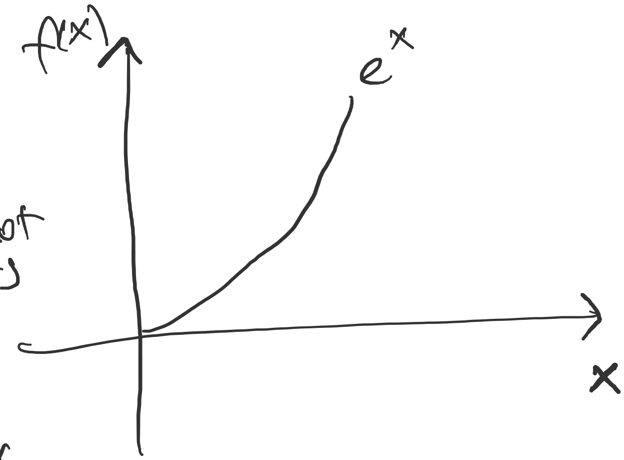
ie, $\nexists f^* \in \mathbb{R}$ s.t. $f^* \leq f(x), \; \forall x \in \mathbb{R}$

But $\inf_{x \in \mathbb{R}} f(x) = -\infty$

Example: $f : \mathbb{R} \to \mathbb{R}$

$f(x) = e^x \,, \; \forall x$

$\min_{x \in \mathbb{R}} f(x) = \min_{x \in \mathbb{R}} e^x$ does not exists



Lower bound: $f_1 = 0 \le \min_{x \in \mathbb{R}} e^x$

but $\nexists\, x \in \mathbb{R}$ that achieve the lower bound.

$\therefore$ no min exists.

but $\inf_{x \in \mathbb{R}} f(x) = 0$

Example: $f : \mathbb{R} \to \mathbb{R}$

$f(x) = x \,, \; \forall x \in \mathbb{R}$

s.t $x > 0$

$\min_{x > 0} f(x) \longrightarrow$ no minimum but bounded below

($\because f(x)$ cannot be negative)

$\therefore \inf_{x > 0} \{ f(x) \mid x > 0 \} = 0$ lower bound $= \{0, -0.5, -1, -1.5, \ldots$

$\Rightarrow \text{glb} = 0$

Example: $f: \mathbb{R} \to \mathbb{R}$

$f(x) = x$, s.t. $x+1=0$

$x-1=0$

$$S = \{x : x+1=0 \text{ and } x-1=0\}$$

$\Rightarrow S = \emptyset$ (infeasible problem)

$\therefore \inf_{x \in S} f(x) = \infty$ by definition.

Example: $f: \mathbb{R} \to \mathbb{R}$

$f(x) = x-1$, $x \in [0,1]$

$$\min_{x \in [0,1]} f(x) = -1 \Rightarrow \inf_{x \in [0,1]} f(x) = \min_{x \in [0,1]} f(x)$$

Example: Discontinuous objective function.

$S = \{x \in \mathbb{R} \mid x \geq 0\}$

$f : S \to \mathbb{R}$

$$f(x) = \begin{cases} 1, & x = 0 \\ x, & x \neq 0 \end{cases}$$

(jump discontinuity at $x=0$)

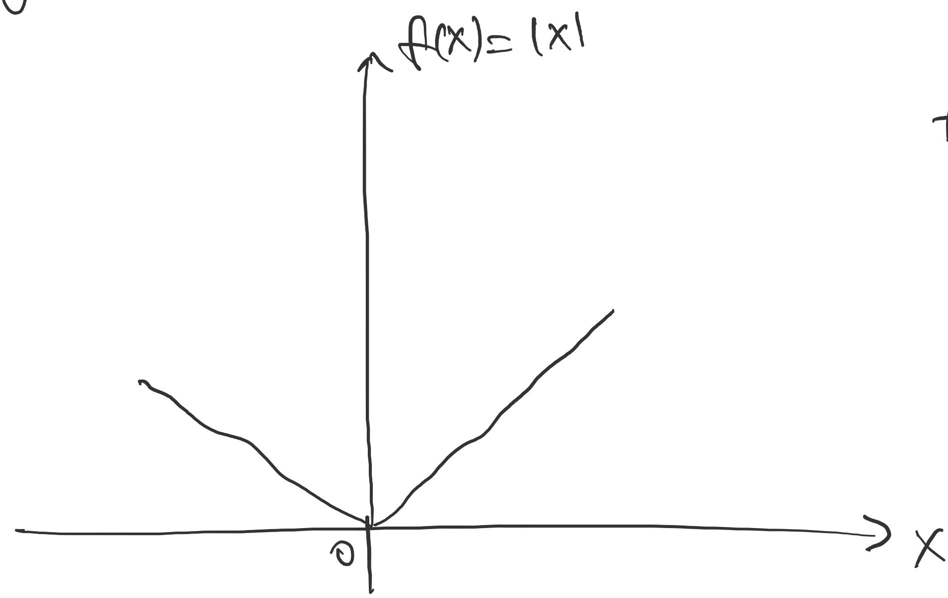
$\min_{x} f(x)$ does not exist

$\inf_{x \in \mathbb{R}} \{ f(x) \mid x \geq 0 \} = 0$

Subgradient (Subgradient Calculus)

what if the objective function is not differentiable?

- gradient ∇f does not exists.



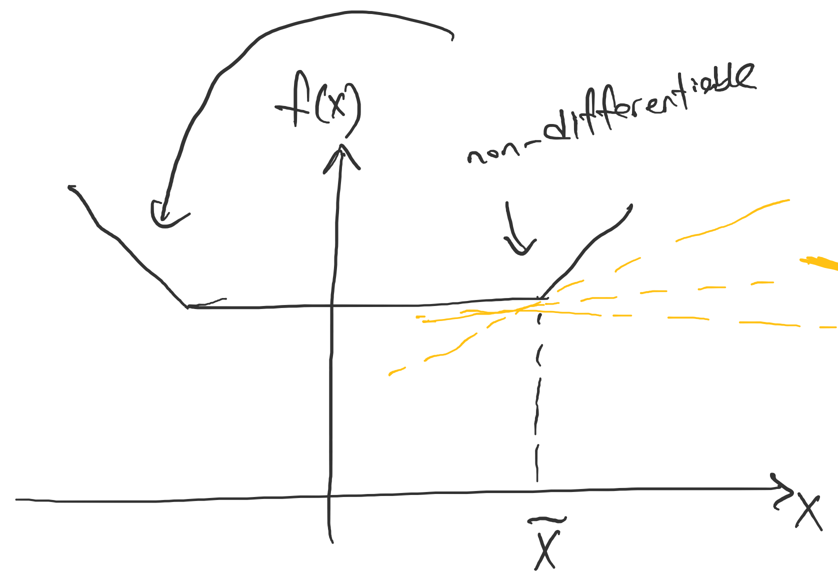
$$f(x) = |x| = \begin{cases} x, & \text{if } x > 0 \\ -x, & \text{if } x < 0 \\ 0, & \text{if } x = 0 \end{cases}$$

$\Rightarrow f$ not differentiable at $x = 0$

But $\min_{x} f(x) = 0$

↑

global minimum



① Subgradient lines with different slopes

② These lines are always below f(x)

③ $y = f(\tilde{x}) + v(x - \tilde{x})$ equation of a line passing through the point $(\tilde{x}, f(\tilde{x}))$ with slope v.

$\therefore\ y - f(x) = v(x - \tilde{x})$

The subgradient lines at $\tilde{x}$ always line below $f(x)$

$\Rightarrow y < f(x) \Rightarrow f(\tilde{x}) + v(x - \tilde{x}) \leq f(x)$

$$\Rightarrow \boxed{f(x) \geq f(\tilde{x}) + v(x - \tilde{x})}$$

Note: The collection of all slopes, v, of subgradient lines at $\tilde{x}$ is called the subdifferential of $f(x)$ at $\tilde{x}$.

Note: for $x \in \mathbb{R}^n$, $\boxed{f(x) \geq f(\tilde{x}) + v^T(x - \tilde{x})}$

Definition: Let $f: \mathbb{R}^n \to \mathbb{R}$ be a convex function and let $\tilde{x} \in \operatorname{dom} f$ (ie, $\operatorname{dom} f = \{x : f(x) < \infty\}$)

An element $v \in \mathbb{R}^n$ is a subgradient of $f(x)$ at $\tilde{x}$ if $v^T(x - \tilde{x}) \leq f(x) - f(\tilde{x})$, $\forall x \in \mathbb{R}^n$

The collection of subgradients of $f(x)$ at $\tilde{x}$, $\partial f(\tilde{x})$, is subdifferential of $f(x)$ at $\tilde{x}$.

Remark: The subgradient (similar to the gradient) is a global underestimator of $f(x)$.

Remark: If $f(x)$ is differentiable then the subgradient = gradient

Example: $f(x) = |x|$, $x \in \mathbb{R}$
subdifferential $\partial f(0)$ at $\tilde{x} = 0$?

Fix a slope $v \in \partial f(0)$

by definition: $v(x - \tilde{x}) \leq f(x) - f(\tilde{x})$

$$\Rightarrow v(x - 0) \leq |x| - 0$$

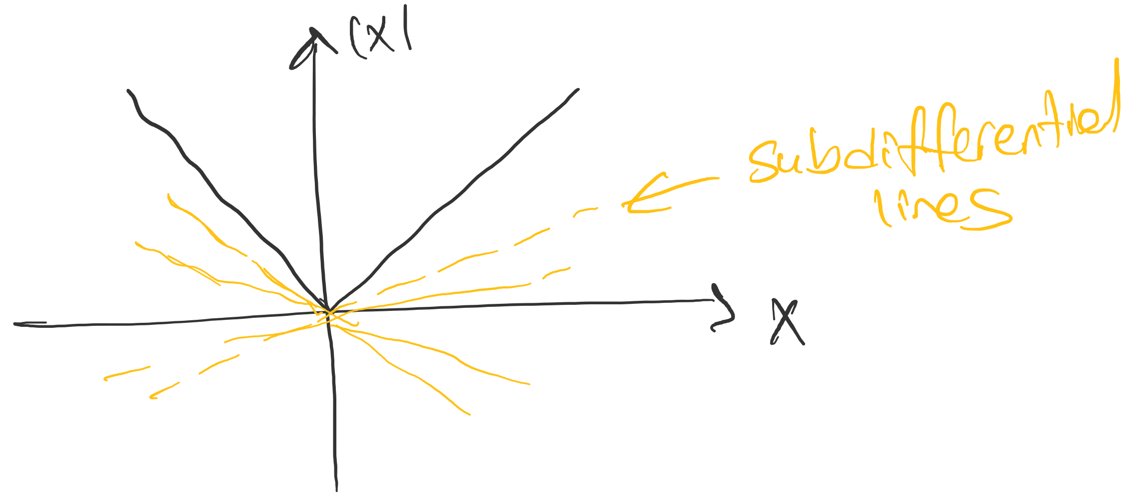
$$\Rightarrow vx \leq |x| \, , \quad \forall x \in \mathbb{R} \quad \text{——— } (1)$$

From (1), at $x = 1$, $v \leq 1$

at $x = -1$, $v(-1) \leq |-1| = 1$

$$\Rightarrow -v \leq 1 \Rightarrow v \geq -1$$

$$\therefore v \in [-1, 1] \Rightarrow \partial f(0) = [-1, 1]$$



Conversely, Let $v \in [-1,1] \Rightarrow |v| \le 1$

$$\therefore \forall x \in \mathbb{R}, \quad v(x-0) = vx \le |vx| = \underbrace{|v|}_{\le 1 \text{ from } 2}|x|$$

$$\Rightarrow vx \le |x|$$

$$\Rightarrow v(x-0) \le f(x) - f(0)$$

$$\Rightarrow v \in \partial f(0)$$

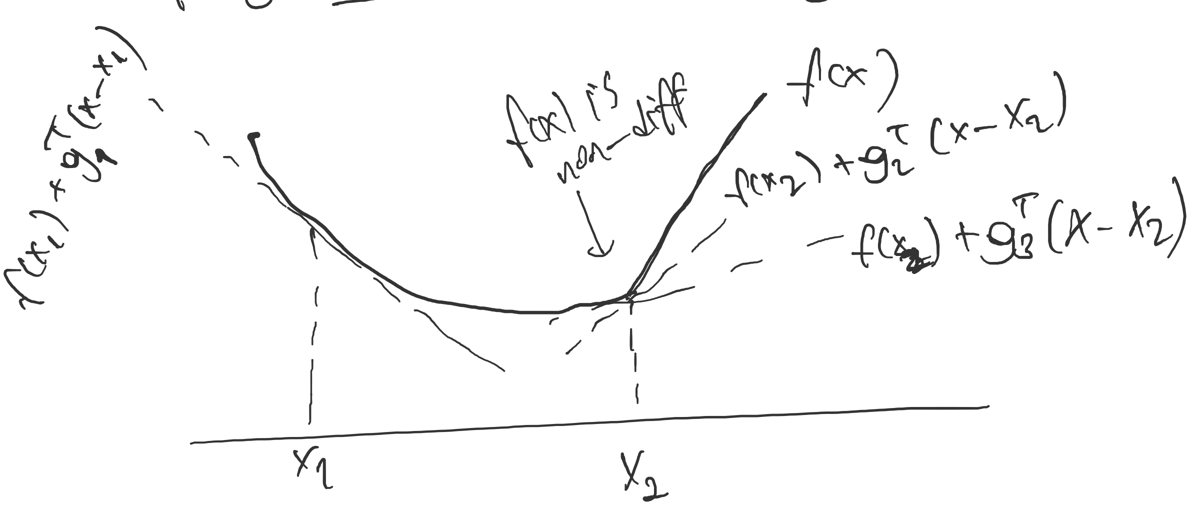
$$\Rightarrow [-1,1] \subseteq \partial f(0)$$

Lecture 8 — 04/08/2021

Subgrad Algorithm, Linear Programming

g is a subgradient of a function f at x, iff $\forall y$,

$$f(y) \ge f(x) + g^T(y-x)$$



Subgradient: g_1 at x_1 (infact $g_1(x_1) = f'(x_1)$)

g_2 and g_3 at x_3

Subgradient Method:

- No guarantee that the objective function is minimized at every iteration since subgradient is not a descent function.

$f : \mathbb{R}^n \longrightarrow \mathbb{R}$ is convex

$\min_{x \in \mathbb{R}} f(x) \rightarrow$ goal

Iteration: $x^{(k+1)} = x^{(k)} - \alpha_k g^{(k)}$

$g(x_k)$: Subgradient at x_k

$\alpha_k > 0$ (step size)

Keep track of the best solution found so far since subgradient is not a descent function.

$$f_{best}^{(k)} = \min\{f_{best}^{(k-1)}, f(x^{(k)})\}$$

and $i_{best}^{(k)} = k$ if $x^{(k)}$ is the best (smallest solution found so far.

$\Rightarrow f^{(k)}_{best} = \min\{f(x^{(1)}), f(x^{(2)}), ----, f(x^{(k)})\}$

$\Rightarrow f^{(k)}_{best}$ is a decreasing sequence $\Rightarrow$ it has a limit

Choose smallest one

$\left[f^{(k)}_{best}\right]_{k=1}^{\infty}$ is a decreasing sequence (could be $-\infty$ also)

Stopping Criterim: Stop the iterations after a certain number (eg, $k=1000$)

Step Size Choices:

1. Constant Stepsize, $\alpha_k = \underbrace{h}_{\text{constant}}$, $\forall k$. $h>0$

2. Constant steplength, $\alpha_k = \dfrac{h}{\|g^{(k)}\|_2}$

   (ie, $\|x^{(k+1)} - x^{(k)}\|_2 = h$)

3. Square summable but not summable

$$\underbrace{\sum_{k=1}^{\infty} \alpha_k^2 < \infty}_{\text{"finite"}}, \quad \sum_{k=1}^{\infty} \alpha_k = \infty$$

Example: $\alpha_k = \dfrac{a}{(b+k)}$, $a>0$, $b>0$

Example: $\alpha_k = \dfrac{1}{k}$

(4) Non Summable diminishing:

$$\lim_{k\to\infty} \alpha_k = 0, \quad \sum_{k=1}^{\infty} \alpha_k = \infty$$

Example: $\alpha_k = \dfrac{1}{k}$, $k=1,2,3,\ldots$

$$\alpha_k = \frac{a}{\sqrt{k}}, \quad a>0 \text{ (constant)}$$

Remark (1): For constant step size and constant step length rules, subgradient method is guaranteed to be "ϵ-optimal."

Example: $\lim_{k\to\infty} f_{best}^{(k)} - f^* < \epsilon$, $\epsilon>0$

$\epsilon(h)$ is a function of the step size.

(Example states that you can never achive the optimal solution but you can get arbitrarily close solution)

(this guarante that you can get very very close)

Remark

2) For diminishing step size rule and square summable but not summable rule, $\lim_{k \to \infty} f(x^k) = f^*$ (ie, convergence to optimal solution)

3) If f is differentiable (ie, ∇f exists) the subgradient method with constant step size converges to optimal value if step size h is small enough.

Linear Programming

Matrix theory recap:

A matrix $A \in \mathbb{R}^{m \times n}$; $A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & & & \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}_{m \times n}$

$m \rightarrow$ rows
$n \rightarrow$ columns

1) $C_{m \times p} = A_{m \times n} B_{n \times p}$ (Two matrices can be multiplied if their dimensions agree)

Inner dimensions

2) Transpose of $A = A^T$, swap rows and columns

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

$\Rightarrow (A^T)^T = A$

If $A \in \mathbb{R}^{m\times n} \Rightarrow A^T \in \mathbb{R}^{n\times m}$

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

(3) Column vectors $x \in \mathbb{R}^n$ is $x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}_{n\times 1}$

$\Rightarrow x^T = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix}_{1\times n}$

(4) Suppose $x, y \in \mathbb{R}^n$

<u>Inner Product:</u> $x^T y = [x_1\ x_2 \cdots x_n] \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} =$

(is a scalar)

("dot product")

$= x_1 y_1 + x_2 y_2 + \cdots + x_n y_n \in \mathbb{R}$

the same

$x \cdot y$ or $\langle x, y \rangle$

<u>Outer Product:</u> outputs an $n\times n$ matrix

$$x y^T = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} [y_1\ y_2 \cdots y_n] = \begin{bmatrix} x_1 y_1 & \cdots & x_1 y_n \\ \vdots & & \vdots \\ x_n y_1 & x_n y_2 \cdots & x_n y_n \end{bmatrix}_{n\times n}$$

5 Stacking matrices and vectors

If $x_1, x_2, ---, x_n \in \mathbb{R}^n \Rightarrow x_i = \begin{bmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{in} \end{bmatrix}$

$$X = \begin{bmatrix} x_1 & x_2 & --- & x_m \end{bmatrix} = \begin{bmatrix} x_{11} & x_{21} & -- & x_{m1} \\ x_{12} & x_{22} & & x_{m2} \\ \vdots & & & \vdots \\ x_{1n} & -- & & x_{mn} \end{bmatrix}_{n \times m} \in \mathbb{R}^{n \times m}$$

Concatenation of Blocks:

$$Y = \begin{bmatrix} A & B \\ C & D \end{bmatrix}$$

if A, C have same number of columns

A, B have same number rows . etc

Multiplication with block matrices

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} P \\ Q \end{bmatrix} = \begin{bmatrix} AP + BQ \\ CP + DQ \end{bmatrix}$$

Linear and Affine functions:

- A function $f(x_1, x_2, ---, x_n)$ is linear in the variables $x_1, x_2, ---, x_n$ if $\exists$ constant $a_1, a_2, ---, a_n$ such that

$$f(x_1, x_2, \dots, x_n) = a_1 x_1 + a_2 x_2 + \dots + a_m x_m$$

$$= a^T x \quad ; \quad a = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{bmatrix} \quad ; \quad x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix}$$

- A function $f(x_1, x_2, \dots, x_m)$ is affine in $x_1, x_2, \dots, x_m$ if $\exists$ constents $b, a_1, a_2, \dots a_m$ such that

$$f(x_1, x_2, \dots, x_m) = a_0 + a_1 x_1 + \dots + a_m x_m = a^T x + b$$

(linear function there is no extra value such b in the function but, in the Affine function there is b, This is the only difference betwen linear and Affine function)

Note: Linear and affine are used interchangeably

Example:

- $3x - y$ is linear in (x, y)
- $2xy + 4$ is affine in x (y is a constant)
  is affine in y (x is a constant)
  but not affine in x and y (x, y)
- $x^2 + y^2 + 2xy + 4$ not linear or affine

Combining several linear and affine functions

$$\left.\begin{aligned} &a_{11}x_1 + \cdots + a_{1n}x_n + b_1 \\ &a_{21}x_1 + \cdots + a_{2n}x_n + b_2 \\ &\vdots \\ &a_{m1}x_1 + \cdots + a_{mn}x_n + b_m \end{aligned}\right\} \text{ m affine functions}$$

$$\Rightarrow \underbrace{\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ \vdots & & & \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}}_{A_{m\times n}} \underbrace{\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}}_{X_{n\times 1}} + \underbrace{\begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}}_{b_{m\times 1}}$$

Therefore, $Ax + b$ is a compact notation for the system of m equations

- Vector-valued function, $F(x)$ is linear in x if $\exists$ a constant matrix A such that

$F(x) = AX$

- Vector-valued function $F(x)$ is affine in x if $\exists$ a constant matrix A and vector b s.t. $F(x) = Ax + b$

Geometry of Affine Equations

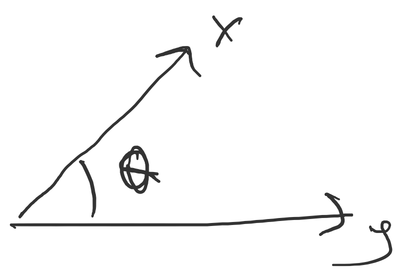
- Set of points $x \in \mathbb{R}^n$ that satisfy a linear equation

$$a_1 x_1 + \cdots + a_n x_n = 0$$

$(a^T x = 0)$ is called a hyperplane

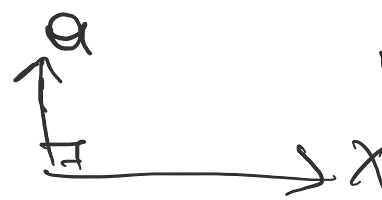
Therefore,

$$\text{Hyperplane} = \{ x \in \mathbb{R}^n \mid a^T x = 0 \}$$



$$x^T y = \|x\| \|y\| \cos\theta$$

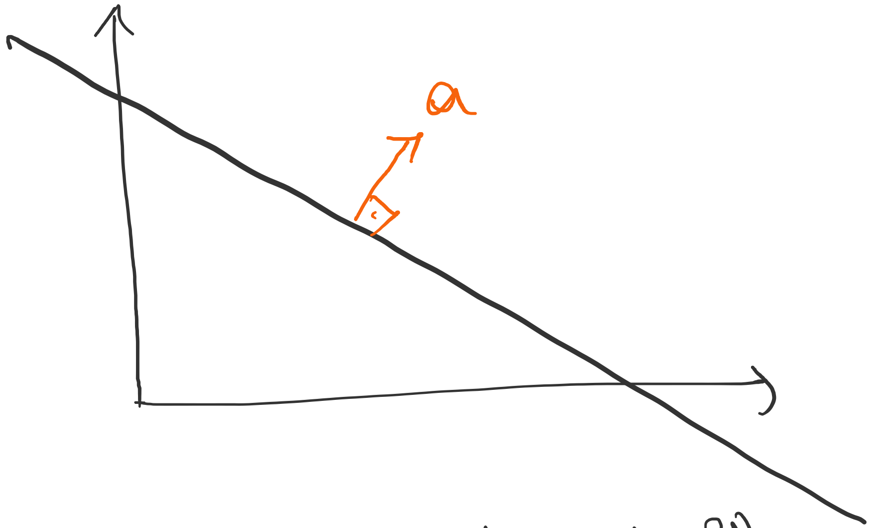
$$a^T x = 0 \Rightarrow a \perp x$$



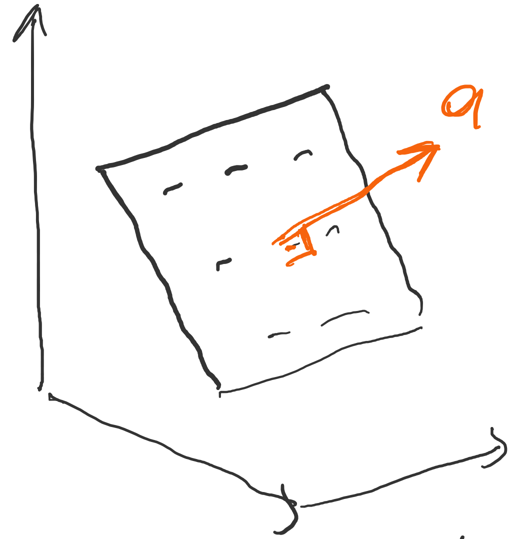
a is normal vector.

- If $a^T x = b$ then solution set is called an <u>affine hyperplane</u> (ie, it is a shifted hyperplane)

$$\text{Affine hyperplane} = \{ x \in \mathbb{R}^n \mid a^T x = b \}$$



Affine hyperplane in 2D



Affine hyperplane in 3D

- **<u>Subspace</u>** : set of points $x \in \mathbb{R}^n$ satisfying many linear equations

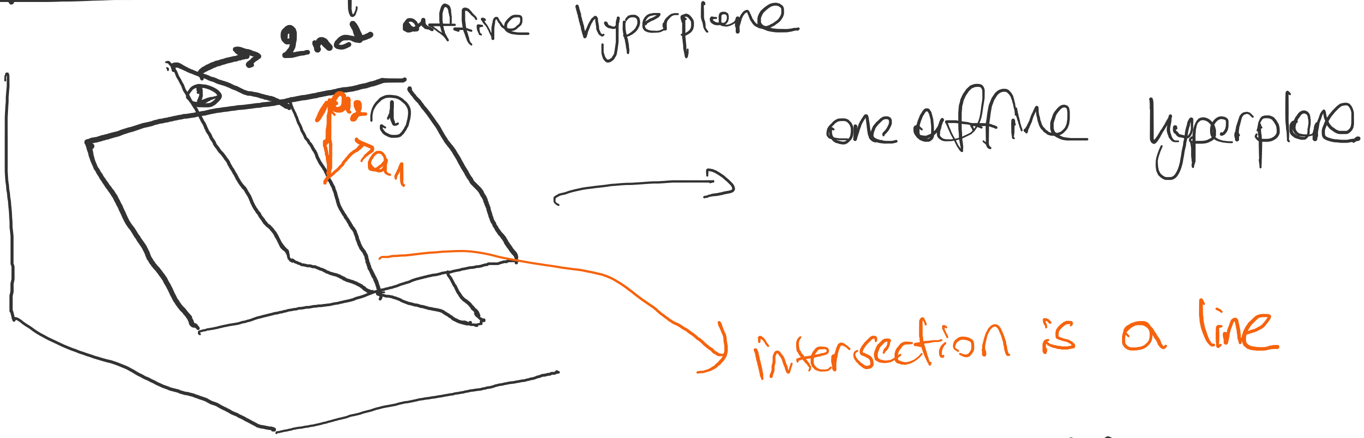
$$a_{i1}x_1 + \dots + a_{im}x_n = 0 \ , \ i = 1, 2, \dots m$$

ie $\{x \in \mathbb{R}^n \mid Ax = 0\}$ is a subspace

(Origin $x = 0$ is a point in this subspace)

$\longrightarrow$ intersection of many hyperplanes.

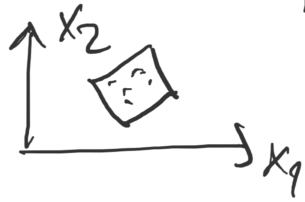
- For $Ax = b$ $(b \neq 0)$ the solution space is called an <u>affine subspace</u>.



Intersection of affine hyperplanes are affine subspace

- The dimension of a subspace is the number of independent directions it contains.

Example: Dimension of a line 1 

Dimension of a plane 2 

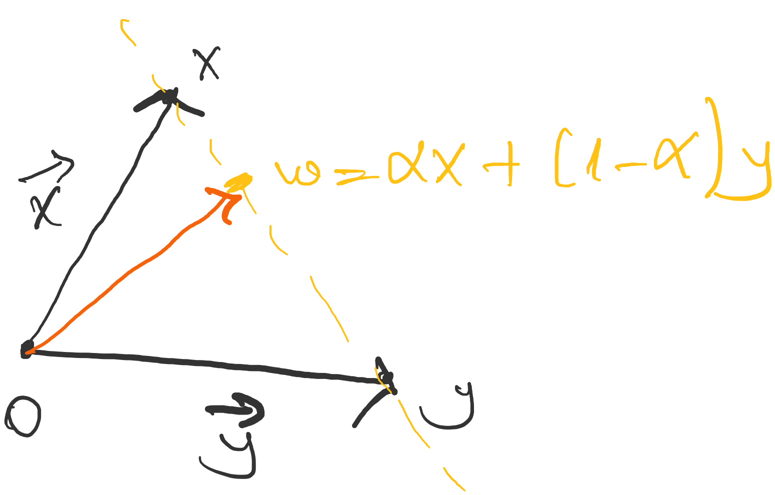
etc.

Hyperplanes are subspaces.

- A hyperplane in $\mathbb{R}^n$ is a subspace of dimension $n-1$
- The intersection of k hyperplanes has dimension at least $(n-k)$ ("at least" because of potential redundancy)

Affine Combinations:

If $x, y \in \mathbb{R}^n$ then $w = \alpha x + (1-\alpha) y$, for some $\alpha \in \mathbb{R}$.
is an <u>affine combination.</u>



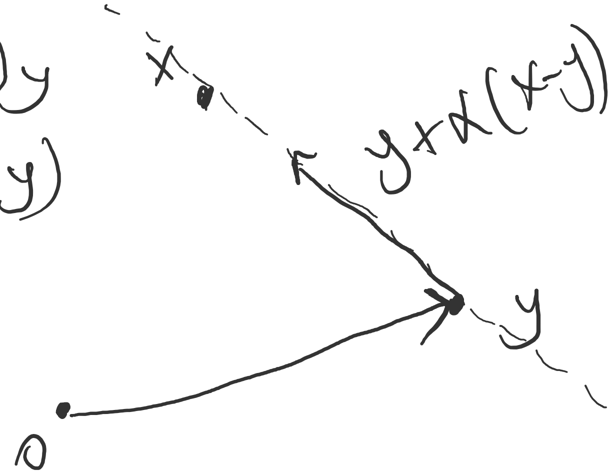
If $Ax=b$ and $Ay=b$

$$\Rightarrow Aw = A[\alpha x + (1-\alpha)y]$$
$$= \alpha \underbrace{Ax}_{b} + (1-\alpha)\underbrace{Ay}_{b}$$

$\Rightarrow Aw = \alpha b + (1-\alpha)b = b$ $\Rightarrow$ affine combinations of a points on affine subspace also belong to the subspace.

- If $x, y \in \mathbb{R}^n$ then $w = \alpha x + (1-\alpha)y$, $\alpha \in \mathbb{R}$

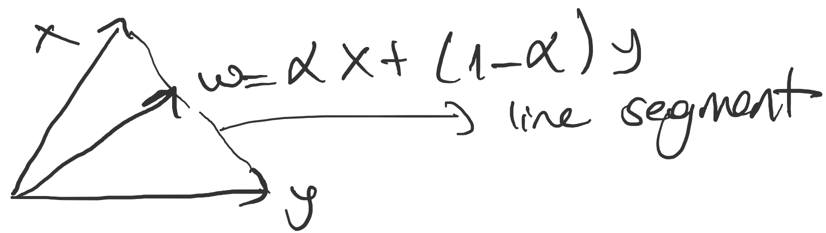
$$w = \alpha x + (1-\alpha)y$$
$$= y + \alpha(x-y)$$



Convex Combination

If $x, y \in \mathbb{R}^n$ then $w = \alpha x + (1-\alpha)y$ for some $0 \leq \alpha \leq 1$

is a convex combination.

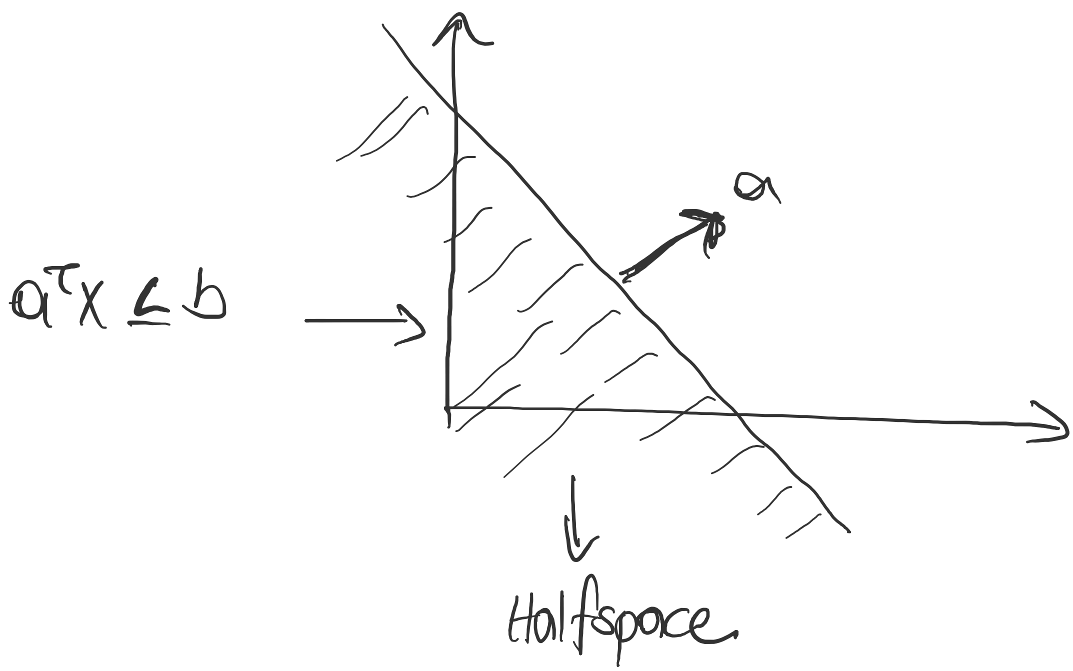


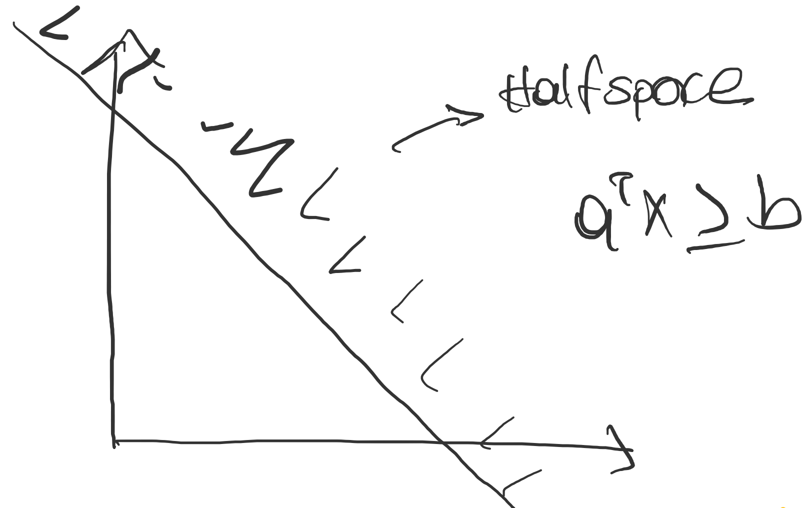
Halfspace: Set of points $x \in \mathbb{R}^n$ that satisfy a linear inequality

$$a_1x_1 + a_2x_2 + \dots + a_nx_n \le b \quad (\text{ie, } a^Tx \le b)$$

is called a halfspace.

- The vector a is normal to the halfspace and b shifts it.





For example, you are trying to classified files and above of the line means file is a dog, below of the line means that it is a cat. (linear classifier)

- Define $w = \alpha x + (1-\alpha) y$, $0 \le \alpha \le 1$

If $a^T x \le b$ and $a^T y \le b \Rightarrow a^T w \le b$

Geometry of affine inequalities

$x \in \mathbb{R}^n$ satisfies $a_{i1} x_1 + a_{i2} x_2 + \dots + a_{in} x_n \le b_i$

$i = 1, \dots, m$

(ie, $Ax \le b$)

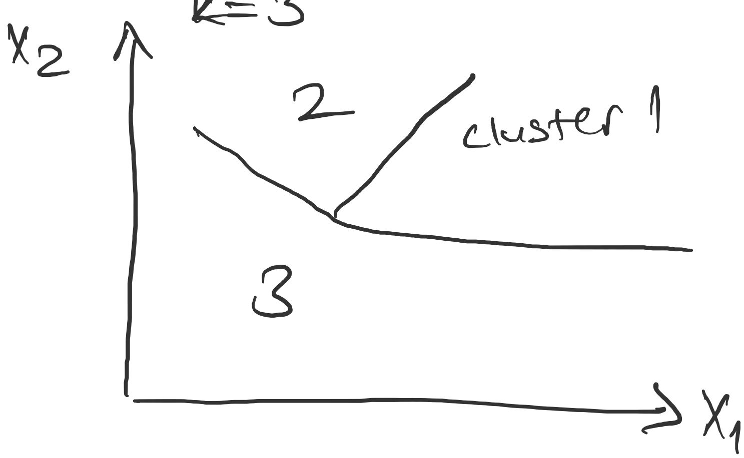
Polyhedron $= \{ x \in \mathbb{R}^n \mid Ax \le b \}$, intersection of many halfspaces

(or)

Polytope

K-means clustering

$K = 3$



- $w = \alpha x + (1-\alpha) y$, $0 \le \alpha \le 1$

If $Ax \le b$ and $Ay \le b \Rightarrow Aw \le b$

Linear Programming (LP)

Optimization model with:

- $x \in \mathbb{R}^n$ (independent variables)
- Objective function: $c^T x + d$ (affine func)
- min or max
- Constraints may be:
    - affine equations, ie, $Ax = b$
    - affine inequalities, ie, $Ax \leq b$ or $Ax \geq b$
    - Combinations of the above.
- Individual variables may have:
    - box constraints, ie, $p \leq x_i$; or $x_i \leq q$ or $p \leq x_i \leq q$
    - no constraints, ie, x_i is unconstrained

Note: Many different ways to represents LP

LP in Standard Form (one version):

Every LP can be put in the form:

$$\boxed{\begin{array}{ll} \max\limits_{x \in \mathbb{R}} & c^T x \\ \text{s.t} & Ax \leq b \\ & x \geq 0 \end{array}}$$

Example:

$$\max_{f,s} \quad 12f + 9s$$

$$\begin{aligned} \text{s.t} \quad 4f + 2s &\leq 4800 \\ f + s &\leq 1750 \\ 0 \leq f &\leq 1000 \\ 0 \leq s &\leq 1500 \end{aligned}$$

$$\Rightarrow \max_{f,s \in \mathbb{R}} \begin{bmatrix} 12 \\ 9 \end{bmatrix}^T \begin{bmatrix} f \\ s \end{bmatrix}$$

$$\begin{bmatrix} 12 & 9 \end{bmatrix} \begin{bmatrix} f \\ s \end{bmatrix} = 12f + 9s$$

$$\text{s.t.} \quad \begin{bmatrix} 4 & 2 \\ 1 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} f \\ s \end{bmatrix} \leq \begin{bmatrix} 4800 \\ 1750 \\ 1000 \\ 1500 \end{bmatrix}$$

$$\begin{bmatrix} f \\ s \end{bmatrix} \geq \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Therefore, in standard form:

$$A = \begin{bmatrix} 4 & 2 \\ 1 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad b = \begin{bmatrix} 4800 \\ 1750 \\ 1000 \\ 1500 \end{bmatrix}, \quad c = \begin{bmatrix} 12 \\ 9 \end{bmatrix},$$

$$x = \begin{bmatrix} f \\ s \end{bmatrix}$$

Tricks to transform non-standard LP to standard form LP

① $\min_x f(x) = -\max_x (-f(x))$

② $Ax \geq b \iff (-A)x \leq (-b)$

(equality to inequality)

③ $x = 0 \iff x \geq 0$ and $x \leq 0$

$\iff x \geq 0$ and $-x \geq 0$

(inequality to equality)

④ $f(x) \leq 0 \iff f(x) + s = 0$ and $s \geq 0$

$\uparrow$

slack variable

(5) unbounded to bounded:

$x \in \mathbb{R} \iff u \geq 0, \; v \geq 0$ and $x = u - v$

(6) bounded to unbounded (convert to inequality)

$$p \leq x \leq q \iff \begin{bmatrix} 1 \\ -1 \end{bmatrix} x \leq \begin{bmatrix} q \\ -p \end{bmatrix}$$

(7) Bounded to non-negative (shift the variable)

$p \leq x \leq q \iff 0 \leq (x - p)$ and $(x - p) \leq (q - p)$