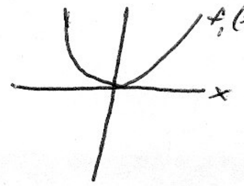
Recitation 2-28-08: Subgradient & Convexity (1)

Gradient (of $f(x)$ at x_0): $\frac{\partial f(x_0)}{\partial x}$

ex $f_1(x) = x^2$

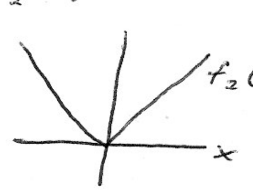


gradient $\frac{\partial f}{\partial x} = 2x$

$\uparrow$

defined everywhere

$f_2(x) = |x|$

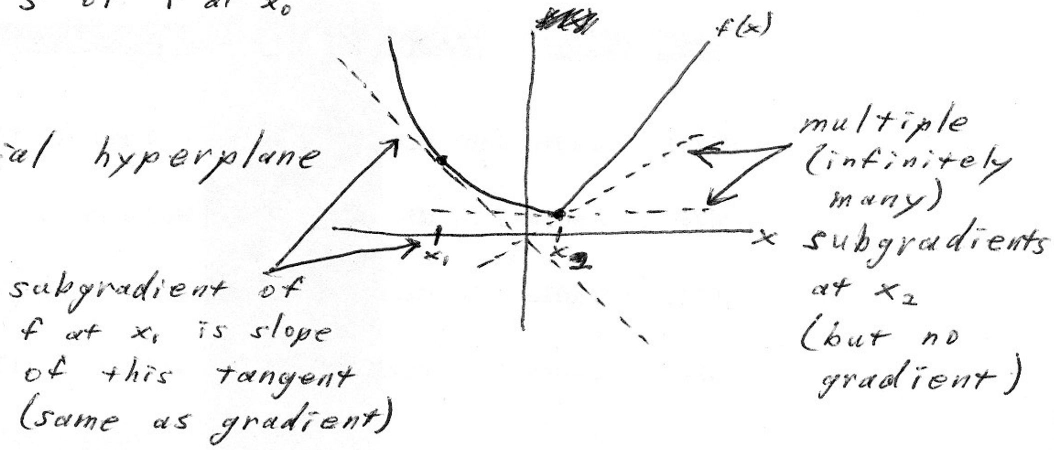


- gradient not defined at $x=0$

Subgradient s of f at x_0

Definition:

① tangential hyperplane



② subgradient s of f at x_0 is s such that

$$f(x) \geq f(x_0) + (x - x_0)^T s, \quad \forall x$$

Note: $\partial f(x)$ is set of all subgradients of f at x

ex $f(x) = |x|$

$$\partial f(x) = \begin{cases} -1 & \text{if } x < 0 \\ [-1, 1] & \text{if } x = 0 \\ 1 & \text{if } x > 0 \end{cases}$$

Note: 1) If f differentiable at x_0, then $\partial f(x_0) = \left\{ \frac{\partial f(x_0)}{\partial x} \right\}$

2) Minimum of convex f is at x_0 iff 0 is a subgradient of f at x_0.

ex $\rightarrow$ LASSO

estimate $y = w \cdot x$ from examples $(x_1, y_1), \ldots, (x_n, y_n)$

$$\min_w \sum_{i=1}^n (y_i - w \cdot x_i)^2 + \sum_j |w_j|$$

Subgradient of this w.r.t. w is

$$\sum_{i=1}^n 2(y - w \cdot x_i)(-x_i) + v$$

where v: $v_j = \begin{cases} -1 & \text{if } w_j < 0 \\ [-1, 1] & \text{if } w_j = 0 \\ 1 & \text{if } w_j > 0 \end{cases}$ (from before)

Subgradient descent (to minimize f)

Start at x_0; use learning rate η_t

Iterate:

$g_t \leftarrow$ estimate $\partial f(x_t)$

$x'_{t+1} \leftarrow x_t - \eta_t g_t$

$x_{t+1} \leftarrow$ Project x'_{t+1} onto feasible region F

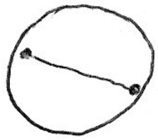
- Choosing η_t:
  - need $\eta_t \to 0$ as $t \to \infty$
  - e.g. $\eta_t = \eta_0 / t$

Convex optimization

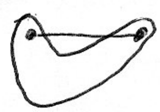
- More general than LPs, QPs; solve more problems.
- Still efficient with guarantees.
- Convex relaxations of non-convex (hard) problems.

Convex sets

ex



vs.



Definition: Set C is convex if

$$\forall x_1, x_2 \in C, \ \forall \theta \in [0,1], \ \theta x_1 + (1-\theta) x_2 \in C$$

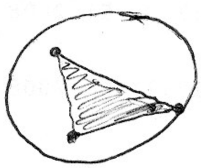
(See line segments above.)

Generalize this to convex combination:

$\forall \theta_1, \dots, \theta_k \geq 0$ such that $\sum_{i=1}^{k} \theta_i = 1$,

if $x_1, \dots, x_k \in C$, then $\sum_{i=1}^{k} \theta_i x_i \in C$

ex



Probabilistic interpretation:

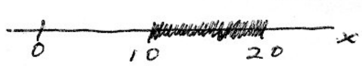
Define distribution $P(x)$ over $x \in C$.

(So $P(x) \geq 0$, and $\int_C P(x)\,dx = 1$.

This generalizes the convex combination idea from above.)

Then expectation $E[X]$ is in C.

ex $C =$ line segment $\{x : 10 \leq x \leq 20\}$



Weighted average of ~~Averaging~~ $x \in [10, 20]$ will give $x' \in [10, 20]$.

To prove a set C is convex,
you can use definition ($\theta x_1 + (1-\theta)x_2 \in C$).

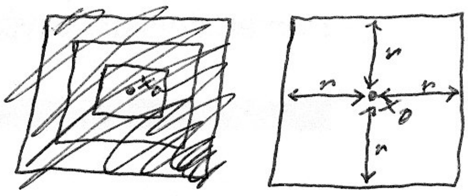
ex → $\mathbb{R}^n$, point, line segment, ray, line, halfspace, polyhedron, linear subspace

- same as $\mathbb{R}^n$, really
- ex → $y = 3x+1$ defines line, which is a linear subspace in $\mathbb{R}^2$

ex → C is $L\infty$-norm ball centered at x_o with radius r:

$$C = \{x : \|x - x_c\|_\infty \le r\} = \{x : \max_j |x_j - x_{cj}| \le r\}$$

Show C is convex.



i.e. show $\forall x_1, x_2 \in C$,
$\forall \theta \in [0,1]$,
$\theta x_1 + (1-\theta)x_2 \in C$

We know $\max_j |x_{1j} - x_{cj}| \le r$ and $\max_j |x_{2j} - x_{cj}| \le r$.

So

$$\max_j \left| [\theta x_1 + (1-\theta)x_2]_j - x_{cj} \right|$$
$$= \max_j \left| [\theta x_{1j} - \theta x_{cj}] + [(1-\theta)x_{2j} - (1-\theta)x_{cj}] \right|$$
$$\le \max_j |\theta x_{1j} - \theta x_{cj}| + |(1-\theta)x_{2j} - (1-\theta)x_{cj}|$$
$$\le \max_j \theta |x_{1j} - x_{cj}| + \max_j (1-\theta)|x_{2j} - x_{cj}|$$
$$\le \theta r + (1-\theta) r = r$$

<u>Definition</u>: <u>Convex hull</u> conv(C) of C is

$$\text{conv}(C) = \{x : x = \sum_i \theta_i x_i,\ x_i \in C,\ \theta_i \ge 0,\ \sum_i \theta_i = 1\}$$

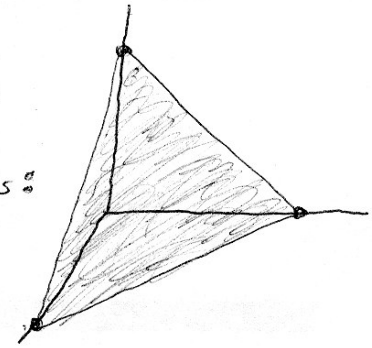
<u>Note</u>: conv(C) is smallest convex set containing C.

<u>Idempotency</u>: If C is convex, $C = \text{conv}(C)$

ex $\rightarrow$ Define n points in $\mathbb{R}^n$:

$$\left.\begin{aligned} x_1 &= (1,0,0,0,\dots,0) \\ x_2 &= (0,1,0,0,\dots,0) \\ &\vdots \\ x_n &= (0,0,\dots,0,0,1) \end{aligned}\right\}$$

Convex hull of the n points is:



We can think of this convex hull as a probability simplex:

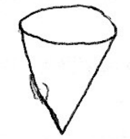
- Each x_i gives probability 1 to event i.
- All points in convex hull represent probability distributions ~~a~~ over n events.

Convex Cones

Definition: Set C is a cone if it is invariant to non-negative scaling:

If $x \in C$ and $\theta \geq 0$, then $\theta x \in C$

ex $\rightarrow$ empty ice cream cone ~~[illegible]~~

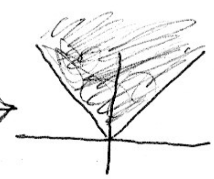


Definition: Set C is a convex cone if it is a cone and is convex.

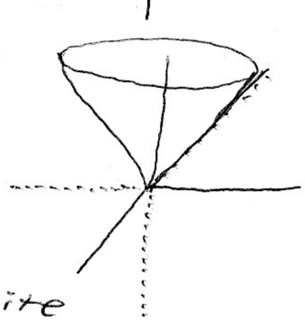
ex $\rightarrow$ filled-in ice cream cone (aka 2^{nd}-order cone)

$C = \{(x,t) \in \mathbb{R}^{n+1} : \|x\|_2 \leq t\}$

In $\mathbb{R}^2$, $C = \{(x,y) : |x| \leq y\}$ $\longrightarrow$



In $\mathbb{R}^3$, $C = \{(x,y,z) : \sqrt{x^2+y^2} \leq z\}$



More generally,
replace $\|x\|_2 \leq t$
with $x^T \Sigma^{-1} x \leq t^2$
where Σ positive semidefinite

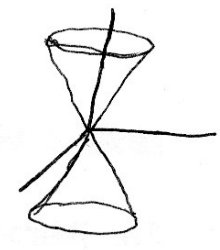
ex → Weirder examples of convex cones:

① ray from origin:

② line

ex → Question: Is this a convex cone?

No: cone, but not convex



Positive Semidefinite Cone

Definition: Matrix Σ is positive semidefinite if ① it is symmetric: $\Sigma^T = \Sigma$ and ② $\forall x,\ x^T \Sigma x \geq 0$

We write $\Sigma \in S_+^n$ or $\Sigma \succeq 0$.

We say definite if strict inequality and write $\Sigma \in S_{++}^n$.

Theorem: $\Sigma \succeq 0$ iff eigenvalues of Σ are ≥ 0

Note: $\Sigma \succeq 0$ iff $\Sigma^{-1} \succeq 0$

ex → ellipsoids & positive definite matrices

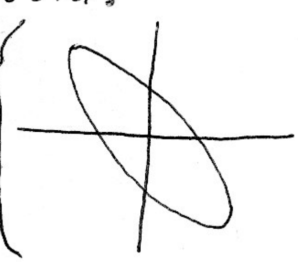
If Σ is in S_{++}^n, then $x^T \Sigma x = 1$ defines a (rotated) ellipsoid.

Say $\Sigma = \begin{pmatrix} 5 & 4 \\ 4 & 5 \end{pmatrix}$, then

$$x^T \Sigma x = 1$$

$$\Updownarrow$$

$$5x^2 + 8xy + 5y^2 = 1$$



If $\Sigma \not\succeq 0$, then $x^T \Sigma x = 1$ is not an ellipsoid.

e.g. $\Sigma = \begin{pmatrix} 1 & 0 \\ 0 & -9 \end{pmatrix}$ gives $x^T \Sigma x = x^2 - 9y^2 = 1$ which is a hyperbola.

<u>Definition</u>: The <u>positive semidefinite cone</u> S_+^n is $\{\Sigma : x^T \Sigma x \geq 0, \forall x\}$.

S_+^n is convex.

Examples of $\Sigma \in S_+^n$: covariance matrix, kernel matrix

<u>Theorem</u>: $\Sigma \succeq 0$ iff $\Sigma = R^T R$ for some matrix R

<u>ex</u> Suppose R has columns $r_j \in \mathbb{R}^m$ which are points in $\mathbb{R}^m$. Then $\Sigma_{ij} = r_i \cdot r_j$, so Σ is a matrix of <u>dot products</u> of points.

The above theorem shows the reverse:
If $\Sigma \succeq 0$, then there exists a matrix R such that $\Sigma_{ij} = r_i \cdot r_j$.
So all $\Sigma \in S_+^n$ represent dot product matrices of n points in some space $\mathbb{R}^m$.

How can we show $\Sigma = R^T R \Rightarrow \Sigma \succeq 0$?

$$x^T (R^T R) x = (Rx)^T (Rx) \geq 0$$

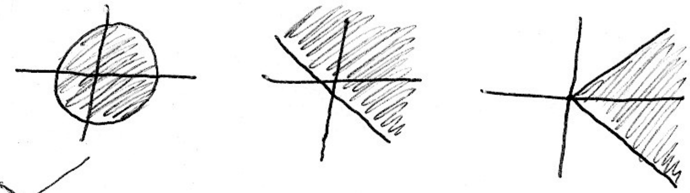
<u>Note</u>: You can sometimes prove $\Sigma \succeq 0$ by showing there is an R such that $\Sigma = R^T R$. Useful since many optimization problems are phrased in terms of $\Sigma \succeq 0$.

<u>Note</u>: The Cholesky decomposition of $\Sigma \succ 0$ gives R which is upper-triangular.

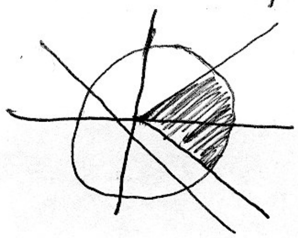
Convexity-preserving operations

Intersection

ex: Convex sets in $\mathbb{R}^2$:



Intersection is convex:



Affine function: $f(x) = Ax + b$ (for any A, b)

If S is convex, then image under f is convex:

$$\{y : y = f(x),\ x \in S\} = \{Ax + b : x \in S\}$$

Corresponds to translation (b), scaling (A), projection.

Proof: Let $y_1, y_2 \in \{Ax + b : x \in S\}$.

Then $\exists\, x_1, x_2$ such that $f(x_1) = y_1$, $f(x_2) = y_2$

So for $\theta \in [0,1]$, we have

$$\theta y_1 + (1-\theta) y_2 = A(\theta x_1) + \theta b + A((1-\theta)x_2) + (1-\theta) b$$
$$= A(\theta x_1 + (1-\theta) x_2) + b$$

Separating hyperplane theorem

If C, D convex sets which do not intersect, they have a separating hyperplane.

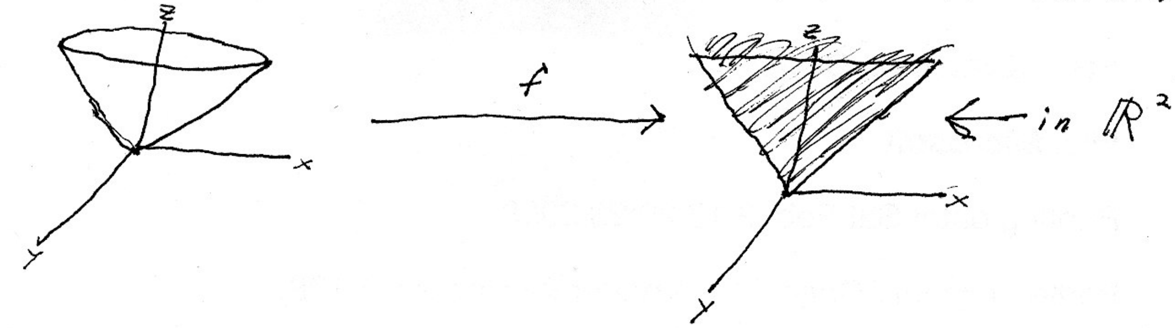
Supporting hyperplane of C at x_0 is tangent at x_0.

ex Prove projection of cone
$C = \{(x, y, z) : \sqrt{4x^2+y^2} \le z\}$ onto
subspace $y = 0$ is convex.

Write projection as an affine function

$$f(x,y,z) = A\begin{pmatrix} x \\ y \\ z \end{pmatrix} + b = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

So we know image of C under f is convex.



We could also say:

1) $f(x,y,z) = f(x,0,z)$
2) So image of C under f is same as image of $\{(x,0,z) : \sqrt{4x^2} \le z\}$ under f
3) This new set is exactly our image in $\mathbb{R}^2$.