nPre lecture discussion

- in many machine learning systems, optimality is not the gold standard - with stochastic solvers you can get close enough, and that is good enough for low test set errors
- *strictly convex* functions do yield unique solutions - this is not true of all convex functions

Why is convexity important?

- we can broadly understand and solve covex optimization problems
- nonconvext are mostly treated on a case by case basis
- any local minimizer is a global minimizer

1. Convex sets

Convex combination of $x_1, ..., x_k \in \Re^n$: any linear combination $\sum \theta_i x_i$ with $\theta_i \geq 0$ and $\sum \theta_i = 1.$

Convex hull of a set C, $conv(C)$, is all convex combinations of elements in C. By defintion, it is always convex, even if C is not convex. The convex hull is the smallest convex set that contains a set.

Examples of convex sets

- **norm ball**
- **hyperplane**
- **halfspace** (everything on one side of a hyperplane)
- **affine space** (satisfies a set of linear equations)
- **polyhedron**: satisfies a set of linear inequalities
  - $\{x: Ax \leq b\}$, inequality interepreted component wise
  - Note that the set $\{x: Ax \leq b, Cx = d\}$ is also a polyhedron. This is because you can split the condition $Cx = d$ into $Cx \leq d$ and $-Cx \leq -d$, and stack them with the original condition to get new inequality conditions.
    - In linear programs, you dont need to consider equality constraints - just ineqaulity constraints, for this reason.
  - a polyhedron is also an intersection of half spaces
    - $a_i^T x \leq b_i, i = 1, ..., m$
- **simplex**: a special case of polyhedra, given by $conv(x_0, ..., x\}$ where these points are affinely independent.
  - the convex combination of $k+1$ points st there is no lower dimensional affine space that contains them
  - canonical example is the probability simplex:

$$conv\{e_1, \ ..., \ e_n\} \ = \ \left\{w: \ w \ \geq 0, \ 1^t w \ = \ 1\right\}$$

- **Cones**
  - Cone: $C \ \subseteq \ \Re^n$ such that $x \ \in C \ \implies \ tx \ \in C \ for \ all \ t \geq 0$
  - Take a point. If it is in the set, then a ray passing through that point is also in the set
  - so every cone has to contain 0
  - conic programming
- **Convex cone**: a cone that is also convex, i.e. $x_1, \ x_2 \ \in C \ \implies \ t_1 x_1 \ + \ t_2 x_2 \ \in C \ for \ all \ t_1, \ t_2 \ \geq 0$
  - Q: is every cone a convex cone? No!
  - Example of a cone that is not convex?
    * a cone that contains 2 and 3 but not 7 or 1? (my guess)
    * a 2D 'ice cream cone' - for x_1 and x_2 on the edge of the cone, $t_1 x_1 \ + \ t_2 x_2$ will not be in the cone!

- **Conic combination** of $x_1, \ ..., \ x_k \ \in \Re^n$: any linear combination $\sum \theta_i x_i$ with all $\theta_i \ \geq 0$.
  - no sum constraint here
- The **conic hull** collects all conic combinations.

Examples of convex cones

- **Norm cone**: $\{(x, \ t): \ ||x|| \ \leq t\}$ for any norm $||.||$
  - with the L2 norm on x, meaning a 3D cone with t being the third axis, this is called the second-order cone - informally the ice-cream cone
- **Normal cone**: given any set C and any point $x \ \in C$ we can define $N_C(x) \ = \ \left\{g: \ g^T x \ \geq g^T y\right\}$ for all $y \ \in C$.
  - The set of vectors g such that the inner product of g with x is largest for all points in the set
  - Every point in the set gets a unique cone
  - Show up in subgradients
  - This is always a convex cone - Why? Prove?
- **Positive semidefinite cone**: $S^n_+ \ = \ \left\{X \ \in S^n: \ X \ \succcurlyeq \ 0\right\}$
  - X is positive semidefinite = the smallest eigenvalue of $X, \ \lambda_{min} \ \geq \ 0$. (nonnegative)
  - an equivalent definition: $a^T X a \ \geq 0$ for all a

Notation: $A \ \succcurlyeq \ B \ \implies \ A \ - \ B \ \succcurlyeq \ 0$. That is, $A \ - \ B$ has a min eigenvalue that is non negative.

Proof that the psd cone is convex.

$S_+^n = \left\{X \in S^n : X \succcurlyeq 0\right\}$

Say we have $X,\ Y \in S_+^n$

is $t_1X + t_2Y \succcurlyeq 0$? That is, is this value also sdp? If true, then the psd cone is convex.

Proof:

$a^T(t_1X + t_2Y)a$

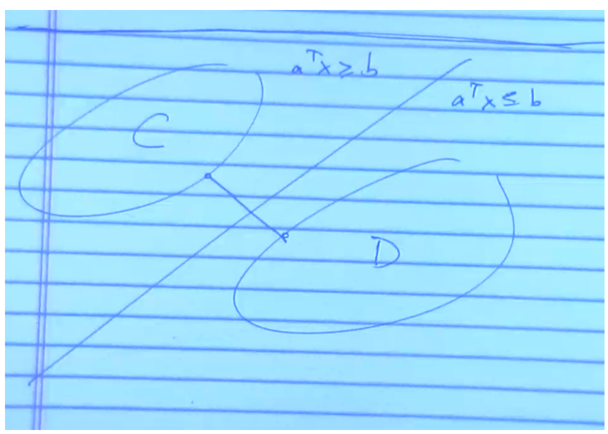
$= t_1\left(a^TXa\right) + t_2\left(a^TYa\right)$

≥ 0

because we know by definition that $\left(a^TXa\right) \succcurlyeq 0,\ X \in S_+^n$

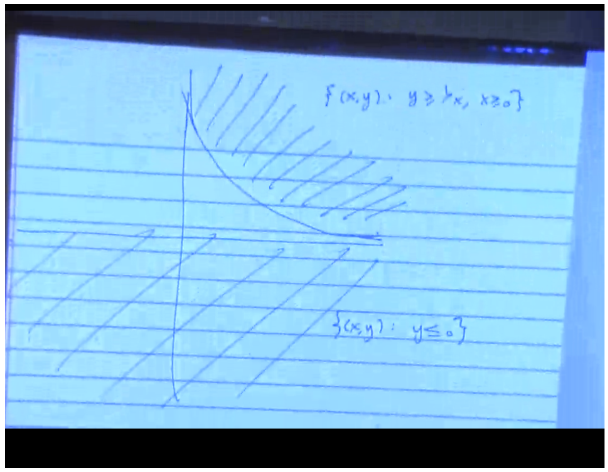
2. Key properties of convex sets

Separating hyperplane theorem: two disjoint convex sets have a separating hyperplane between them. (so the disjoint convex sets are halfspaces)



- Proof: find two points that are closest. Look at the line segment that joins them. And define the affine space defining the hyperplane to be orthogonal to this line.
- Q: Does this guarantee strict separation? ie can you be guaranteed that $a^Tx > b$ in one halfspace, and $a^Tx < b$ in the other?
  - No. Find an example to the contrary!
    * For 1 open 1 closed: take $\{(x,\ y) : y > 0\}$ and $\{(x,\ y) : y \leq 0\}$. The latter includes the x-axis so you get = b.
    * For two closed sets: Take $\{(x,\ y) : y \leq 0\}$ and $\{x,\ y) : y \geq 1/x,\ x \geq 0\}$. The latter is called the epigraph. Both sets are closed and convex, but there is no hyperplane that strictly separates them
    * The way to get strict separation is to have one set *bounded.*

.



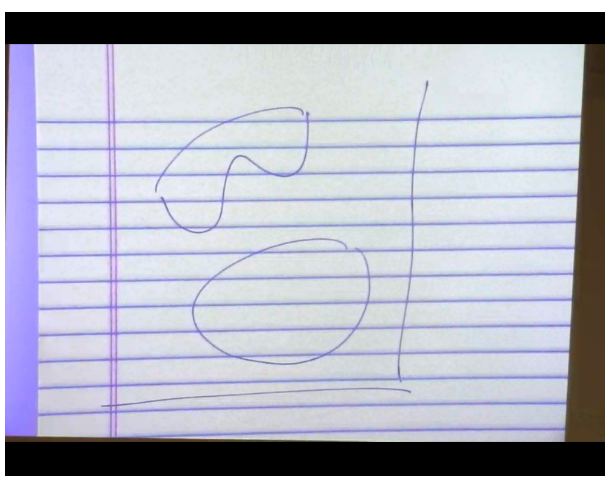
Supporting hyperplane theorem: a boundary point of a convex set has a supporting hyperplane passing through it

- Q: What is a boundary point?
- Formally: if C is a nonempty convex set and $x_0 \in bd(C)$ then thre exists a st $C \subseteq \left\{x: a^T x \leq a^T x_0\right\}$

Both the separating and supporting hyperplane theorems have partial converses. That is, if sets display these properties, then they are convex.

Edit: There is a gotcha with thinking through the converse of the separating hyperplane theorem!

These two set on top is definitely not convex, but it can be separated from the set at the bottom by a separating hyperplane.



Instead, if you have two convex sets, one open and one closed, and there exists a separating hyperplane, then they are disjoint.

Converse of the supporting hyperplane theorem is perhaps more interesting.
If a set is closed, has a non empty interior, and has a supporting hyperplane at every boundary point, then the set is convex.

3. Operations that preserve set convexity

- **Intersection**
- **Scaling and translation**
  - if C is convex, then $aC + b = \{ax + b : x \in C\}$
- **Affine images and preimages**
  - if $f(x) = Ax + b$ (affine function)
    - vs linear: $f(x) = Ax$. No constant
  - And C is convex, then $f(C) = \{f(x) : x \in C\}$ is convex
    - This is an affine image
  - And if D is convex then $f^{-1}(D) = \{x : f(x) \in D\}$ is convex
    - This is an affine preimage

A motivating example to show how these operaations are useful in determining when a function/solution is convex: linear matrix inequality.

The problem is: show that the set C of solutions $x \in \Re^k = \{x_1, ..., x_k\}$ to $x_1 A_1 + ... + x_k A_k \preccurlyeq B$ is a convex set, independent of the matrices.

- As a reminder the problem set up means that $B - \sum x_i A_i \succcurlyeq 0$, that is, the difference is psd.

There are two approaches here.
Approach 1: check the definition of convexity. Find two points in the set, and show that $tx + (1-t)y \in C$.

Approach 2: let $f : \Re^k \to S^n$, $f(x) = B - \sum x_i A_i$. We want to find the domain of $f(x)$ that guarantees $f(x) \succcurlyeq 0$.
Then $C = f^{-1}\left(S_+^n\right)$ i.e. it is an affine preimage of the psd cone, which we know is a convex set. So C is also a convex set!

$f^{-1}\left(S_+^n\right) = \left\{x : f(x) \in S_+^n\right\} = \left\{x : B - \sum x_i A_i \succcurlyeq 0\right\}$, which is what we are looking for!

3.1 More operations preserving set convexity

Perspective images and preimages: the perspective function is

$P: \Re^n * \Re_{++}\ (positive\ reals)\ \to\ \Re^n$.

$P(x,\ z)\ =\ x/z\ for\ z\ >\ 0$

If $C\ \subseteq dom(P)$ is convex then so is $P(C)$, and if D is convex then so is $P^{-1}(D)$.

Linear-fractional images and preimages: the perpsective map composed with an affine function

$$f(x)\ =\ \frac{Ax\ +\ b}{c^Tx\ +\ d}$$

If $C\ \subseteq dom(P)$ is convex then so is $P(C)$, and if D is convex then so is $P^{-1}(D)$.

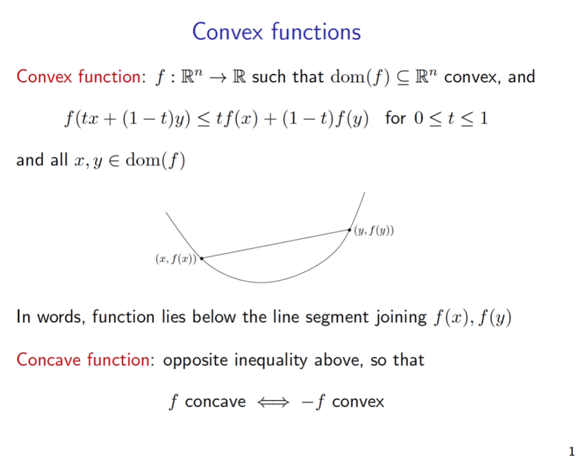
- fairly complicated transform that preserves convexity!
  - can save you a lot of work

4. Convex functions

Convex functions: $f:\ \Re^n\ \to\ \Re$ such that $dom(f)\ \subseteq \Re^n$ is conved AND

$f(tx\ +\ (t-t)y)\ \leq t(f(x)\ +\ (1-t)f(y))$

Graphically, the function lies below the line segment joining $f(x)$ and $f(y)$



Important modifiers:

1. Strictly convex: $f(tx\ +\ (t-t)y)\ <\ t(f(x)\ +\ (1-t)f(y))$. ie the function is strictly below the line segment. A linear function is convex but not strictly convex!

- Has strictly more curvature than a linear functions

2. Strongly convex with parameter $m > 0$: $f - \frac{m}{2}||x||_2^2$ is convex

- at least as curvey as a quadratic function
- it implies strict convexity

4.1 Examples of convex functions

1. **Exponential** e^{ax}
2. **Power functions** x^a is convex for $a \geq 1$ of $a \leq 0$ over $\Re_+$(nonnegative reals). But it is concave for $0 \leq a \leq 1$.

Example of **concave** functions:
1. **Logarithm**

Affine functions are both concave and convex. $a^Tx + b = c$ implies $a^Tx + b \leq c$ and $a^Tx + b \geq c$, so it satisfies both the concave and the convex inequality definitions.

Quadratic function $\frac{1}{2}x^TQx + b^Tx + c$ is convex if $Q \succcurlyeq 0$ ie Q is psd.

Least squares loss $||y - Ax||_2^2$ is always convex since $Q = A^TA$ is always psd!

- That is, $b^TA^TAb \succcurlyeq 0$
- Why? Because if $z = Ab$ then this is the same as $z^Tz = \sum z_i^2$ which must ≥ 0

Norm: $||x||$ is convex for any norm

$$l_p\ norms\ ||x||_p = \left(\sum_{i=1}^{n} x_i^p\right)^{1/p},\ p \geq 1,\ ||x||_\infty = max|x_i|$$

A norm is:

- zero iff x = 0
- positive homogeneity: norm(Ax) = A norm(x)
- triangle inequality

These three properties guarantee convexity.

The most common values for p are 1, 2, and inf.

- 1 is sparsity inducing - LASSO
- 2 norm is ubiquitous - use it in Euclidean space

What is the l_0 'norm'? Sometimes used to refer to the value $||x||_0 = \sum 1\{x_i \neq 0\}$

- this is not really a norm because it does not obey postiive homogeneity!
- and it is therefore not convex!

Norms also exist on matrices.
1. Operator norm: $||X||_{op} = \sigma_1(X)$. Biggest singular value.
2. Trace norm: $||X||_{tr} = \sum_{i=1}^{r} \sigma_r(X)$. Sum of the singular values
where $\sigma_1(X) \geq ... \geq \sigma_r(X)$ are the singular values of the matrix X.

Indicator function. If C is convex, then $I_C(x) = \{0 \; for \; x \in C, \; \infty \; for \; x \notin C\}$ is convex.
Two things to check for convexity of a function:
1. domain is convex
2. take any two points in the domain, and check that $f(tx + (1-t)y) \leq tf(x) + (1-t)f(y)$

For any two points in the domain, this is true since $f(x)$ is 0, $f(y)$ is 0, and $f(tx + (1-t)y)$ which is also in the domain if the domain is convex is also 0!

- Note that if the domain were not convex, then $f(tx + (1-t)y)$ would be infinite, so this inequality would not hold

Support function. For any set C (convex or not), its support function $I_c^*(x) = max \; x^T y$ is convex.
Max. $f(x) = \max\{x_1, ..., x_n\}$ is always convex.
Why?
Pick any two x, y in $\Re^n$.
Then $f(tx + (1-t)y) = max(tx + (1-t)y) = max_i(tx_i + (1-t)y_i)$

- The intuition is that picking the max of the vectors individually and adding them up is at least as large as picking the max of the combined vector from adding the two together!

But we know for any k, $tx_k + (1-t)y_k \leq t \; max_i x_i + (1-t) \; max_i y_i$, since all the coefficients are ≥ 0 and $x_k \leq max_i x_i$ for all k, and similarly for y
Therefore $\max_k tx_k + (1-t)y_k \leq t \; max_i x_i + (1-t) \; max_i y_i$, which makes it a convex function!

4.2 Important properties of convex functions

A function is convex iff its restriction to any line is convex.

- Look at convexity in a univariate space for every function
- Useful in showing the log determinant is convex

Epigraph characterization. A function f is convex iff its epigraph is a convex set.
$epi(f) = \{(x, t) \in dom(f) \; x \; \Re : \; f(x) \leq t\}$
Visually in a 2d plot, the epigraph is everything above the function curve.

- This is a key connect between convex sets and convex functions

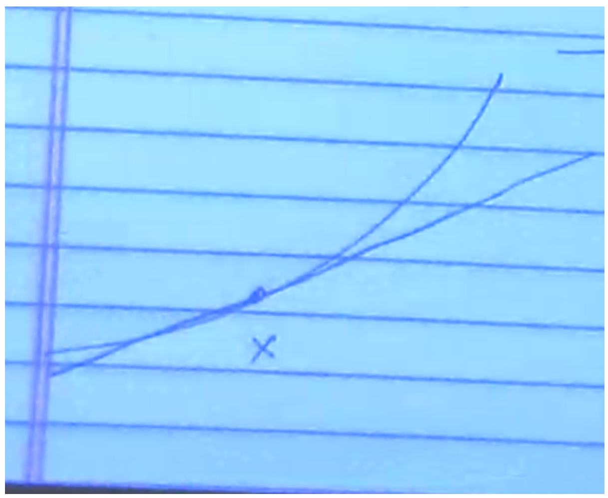
Convex sublevel sets:

5. Characterizations of convexity

First order characterization: if f is differentiable, then f is convex iff $dom(f)$ is convex, AND

$f(y) \geq f(x) + \nabla f(x)^T(y-x)$

- This defines a line in y. If $f(y)$ is above this line then f is convex
-



- Therefore for a differentiable convex function, $\nabla f(x) = 0 \leftrightarrow x\ minimizes\ f$

Second order characterization: If f is twice differentiable, then f is convex iff $dom(f$ is convex and $\nabla^2 f(x) \succcurlyeq 0$ for all $x \in dom(f)$. Ie a function is convex if the domain is convex and the second derivative is always greater than 0.

- If a function strictly convex, is the second derivative of a function always positive? No!
- $f(x) = x^4$ is strictly convex, but has $f''(0) = 0$!

But if the second derivative is always positive, it does imply strict convexity. Just not the other way round.

Jensen's inequality.

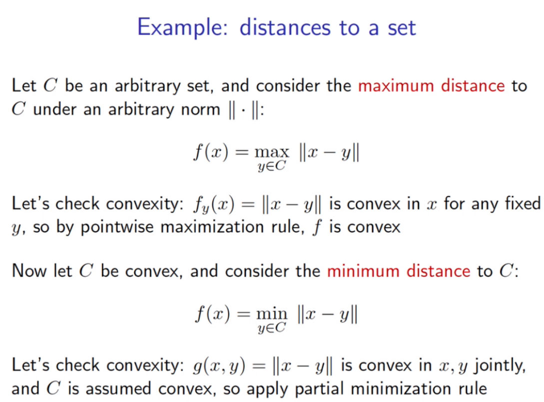
5. Operations that preserve the convexity of functions

1. **Nonnegative linear combination of convex functions.**
   - flipping a sign on a convex function makes it concave!

2. **Pointwise maximization**. If f_s is convex for any $s \in S$, the $f(x) = max_s f_s(x)$ is convex.

- maximization preserves convexity whenever the functions themseles are convex
- Example: distances to a set. The maximum distance to any set C under any arbitrary norm is convex

3. **Partial minimization**. If $g(x, y)$is convex jointly in x and y, and C is convex, then $f(x) = min_{y \in C} g(x, y)$ is convex

- minimization only preserves convexity when the minimizing set for one variable is convex
- Example: minimum distance to a set is convex only when C is convex to begin with
- 

Example: distances to a set

Let C be an arbitrary set, and consider the maximum distance to C under an arbitrary norm $\|\cdot\|$:

$$f(x) = \max_{y \in C} \|x - y\|$$

Let's check convexity: $f_y(x) = \|x - y\|$ is convex in x for any fixed y, so by pointwise maximization rule, f is convex

Now let C be convex, and consider the minimum distance to C:

$$f(x) = \min_{y \in C} \|x - y\|$$

Let's check convexity: $g(x, y) = \|x - y\|$ is convex in x, y jointly, and C is assumed convex, so apply partial minimization rule

-

6. Example: Show that softmax or log-sum-exp is convex.