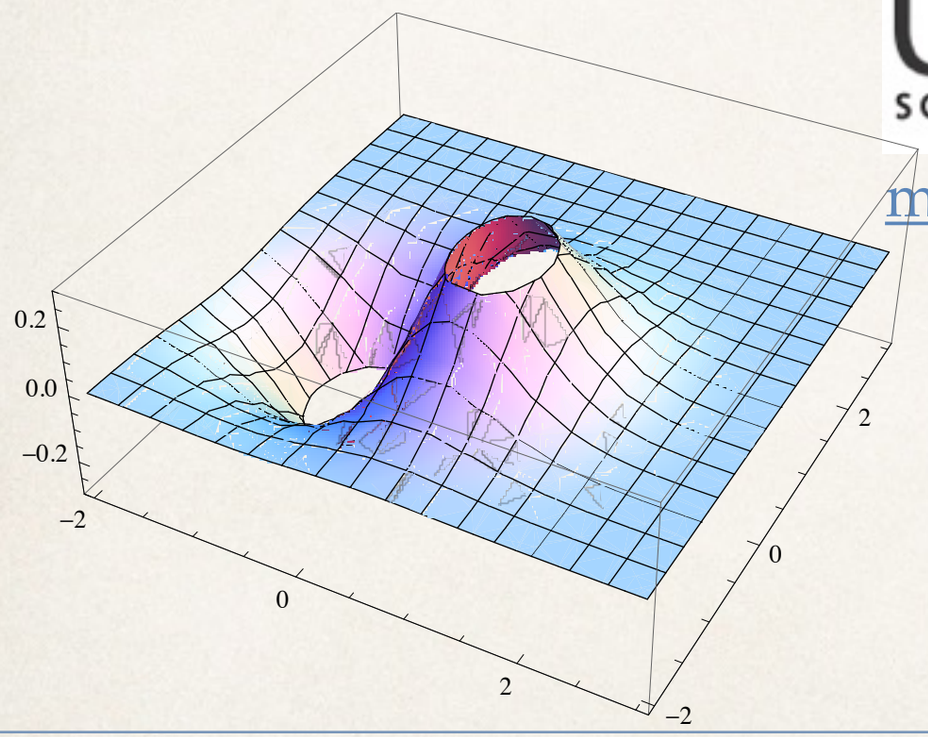
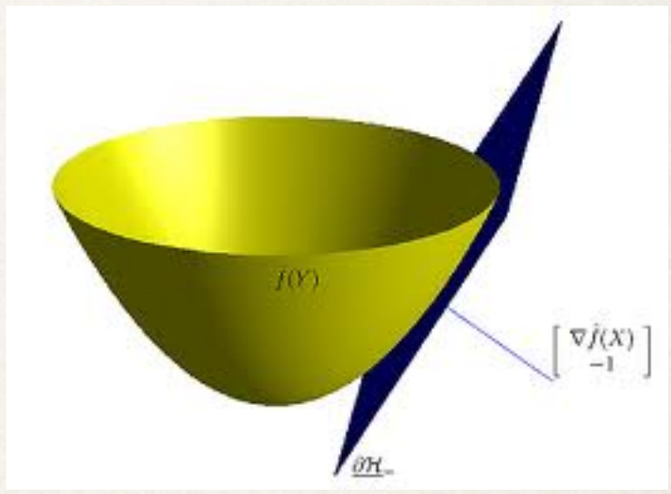
Optimization for CS

Sridhar Mahadevan

UMASSCS
SCHOOL OF COMPUTER SCIENCE

mahadeva@cs.umass.edu





Lecture 2

First, a motivating example….

- As we begin the study of optimization, it may be useful to begin with a real-world motivating example
- I am going to use an exciting problem from natural language processing
  - How to construct the “meanings” of words?
  - How to reason with word meanings about sentences?
  - I will illustrate how these problems are solved using ideas from optimization

The Jabberwocky song

Twas brillig, and the slithy toves
Did gyre and gimble in the wabe;
All mimsy were the borogoves,
And the mome raths outgrabe.

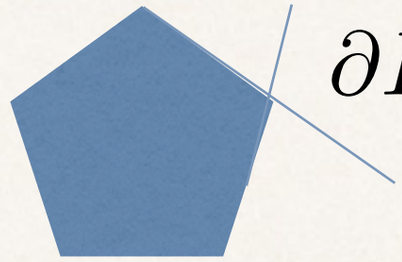
Explanation

- Brillig: Following the poem, the character of Humpty Dumpty comments: " 'Brillig' means four o'clock in the afternoon, the time when you begin broiling things for dinner."[16] According to *Mischmasch*, it is derived from the verb to *bryl* or *broil*.
- Slithy: Humpty Dumpty says: " 'Slithy' means 'lithe and slimy'. 'Lithe' is the same as 'active'. You see it's like a portmanteau, there are two meanings packed up into one word."[16] The original in *MischMasch* notes that 'slithy' means "smooth and active"[17] The *i* is long, as in *writhe*
- Tove: Humpty Dumpty says " 'Toves' are something like badgers, they're something like lizards, and they're something like corkscrews. [...] Also they make their nests under sun-dials, also they live on cheese."[16] Pronounced so as to rhyme with *groves*.[19] They "gyre and gimble," i.e. rotate and bore. Toves are described slightly differently in *Mischmasch*: "a species of Badger [which] had smooth white hair, long hind legs, and short horns like a stag [and] lived chiefly on cheese".[17

Wabe: The characters in the poem suggest it means "The grass plot around a sundial", called a 'wa-be' because it "goes a long way before it, and a long way behind it".[16] In the original

In convex optimization, we see Jabberwocky statements often

Subdifferential of the indicator function of a convex set C is the normal cone of C



$$\partial I_C = \mathcal{N}_C$$

Words and their meanings

WORD OF THE DAY

JANUARY 26, 2016

zeugma

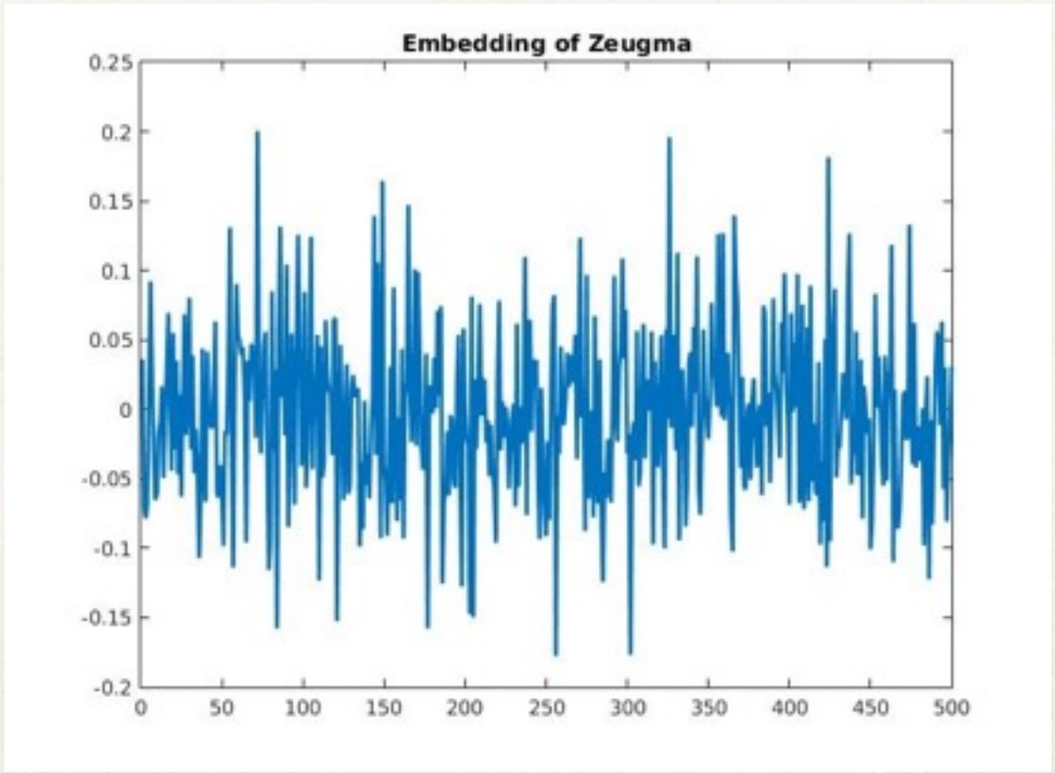
the use of a word in more than one sense

Representation discovery of word meanings

You shall know a word by the company it keeps

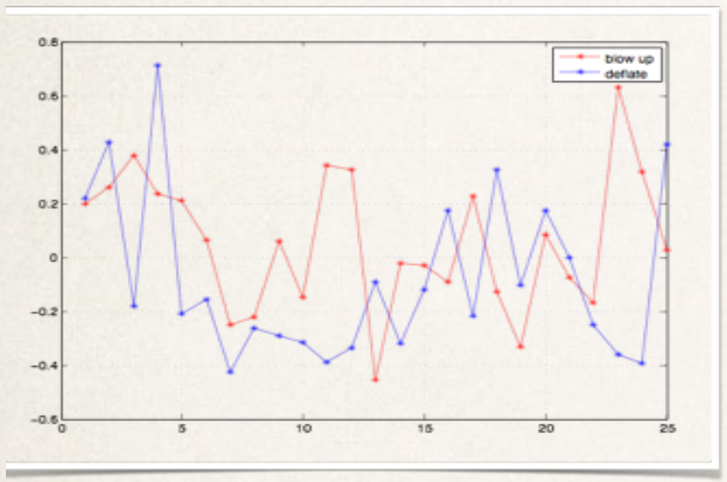
zeugma

vector in n-dimensions



500-dim embedding using word2vec (Mikolov, 2013)

NLP as optimization



Wikipedia Dataset

Create Word Embedding

Google, MSR

Classification of synonyms etc.

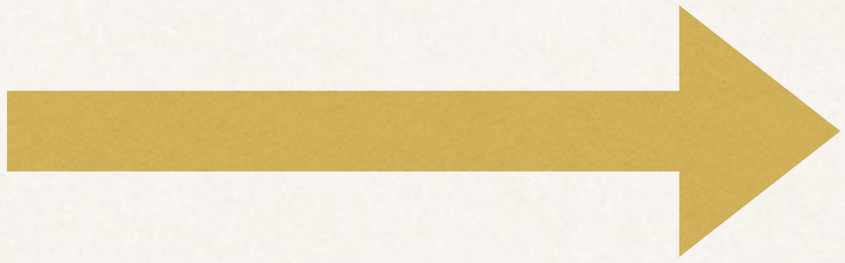
Wordnet Synonym

Relationship prediction

Baghdad is to Iraq as London is to ?

How to construct word vectors?

I love Amherst
I love shopping
I hate math



Cooccurrence table

Word Cooccurrence matrix

Counts	I	love	Amherst	shopping	hate	math
I	0	2	0	0	1	0
love	2	0	1	1	0	0
Amherst	0	1	0	0	0	0
shopping	0	1	0	0	0	0
hate	1	0	0	0	0	1
math	0	0	0	0	0	0

Best Rank-K Approximation

Given a matrix M, find another matrix M* of rank k such that

$$\|M - M^*\|_F^2$$

Non-convex!

Solution: SVD of M = U S V'

Word embedding: first k entries of the singular vector in U corresponding to word w

GLOVE Algorithm

(Pennington, Socher, Manning)

$$J(\theta) = \frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c} \log p(w_{t+j}|w_t)$$

$$p(w_o|w_c) = \frac{\exp^{v_o^T \tilde{v}_j}}{\sum_{k=1}^{V} \exp v_o^T \tilde{v}_k}$$

Reasoning about analogies

King is to Man as Queen is to ?

COSADD

$$\text{argmax}_{y \in V} \delta(\omega_y, \omega_x - \omega_a + \omega_b), \text{ where } \delta(i,j) = \frac{\omega_i^T \omega_j}{\|\omega_i\|_2 \|\omega_j\|_2}$$

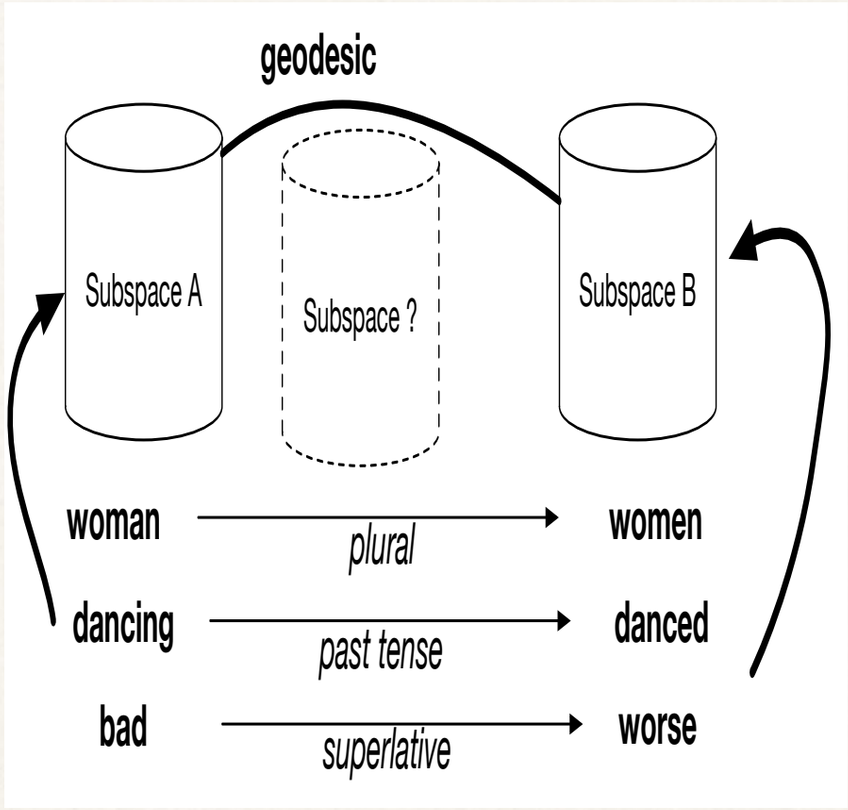
COSMUL

$$\text{argmax}_{y \in V} \frac{\delta(y,b)\delta(y,x)}{\delta(y,a) + \epsilon}$$

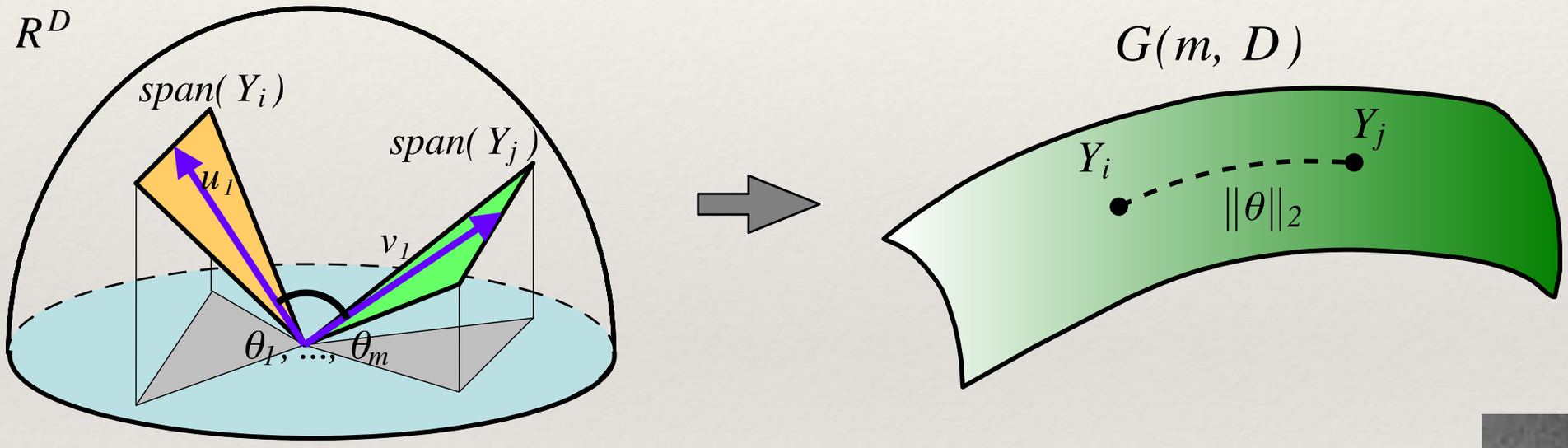
Subspace Representations in NLP

(Mahadevan & Chandar, Arxkiv, 2015)

	Relation
GOOGLE	capital-common-countries
	capital-world
	city-in-state
	currency
	family (gender inflections)
	gram1-adjective-to-adverb
	gram2-opposite
	gram3-comparative
	gram4-superlative
	gram5-present-participle
	gram6-nationality-adjective
	gram7-past-tense
	gram8-plural (nouns)
	gram9-plural-verbs
MSR	adjectives
	nouns
	verbs

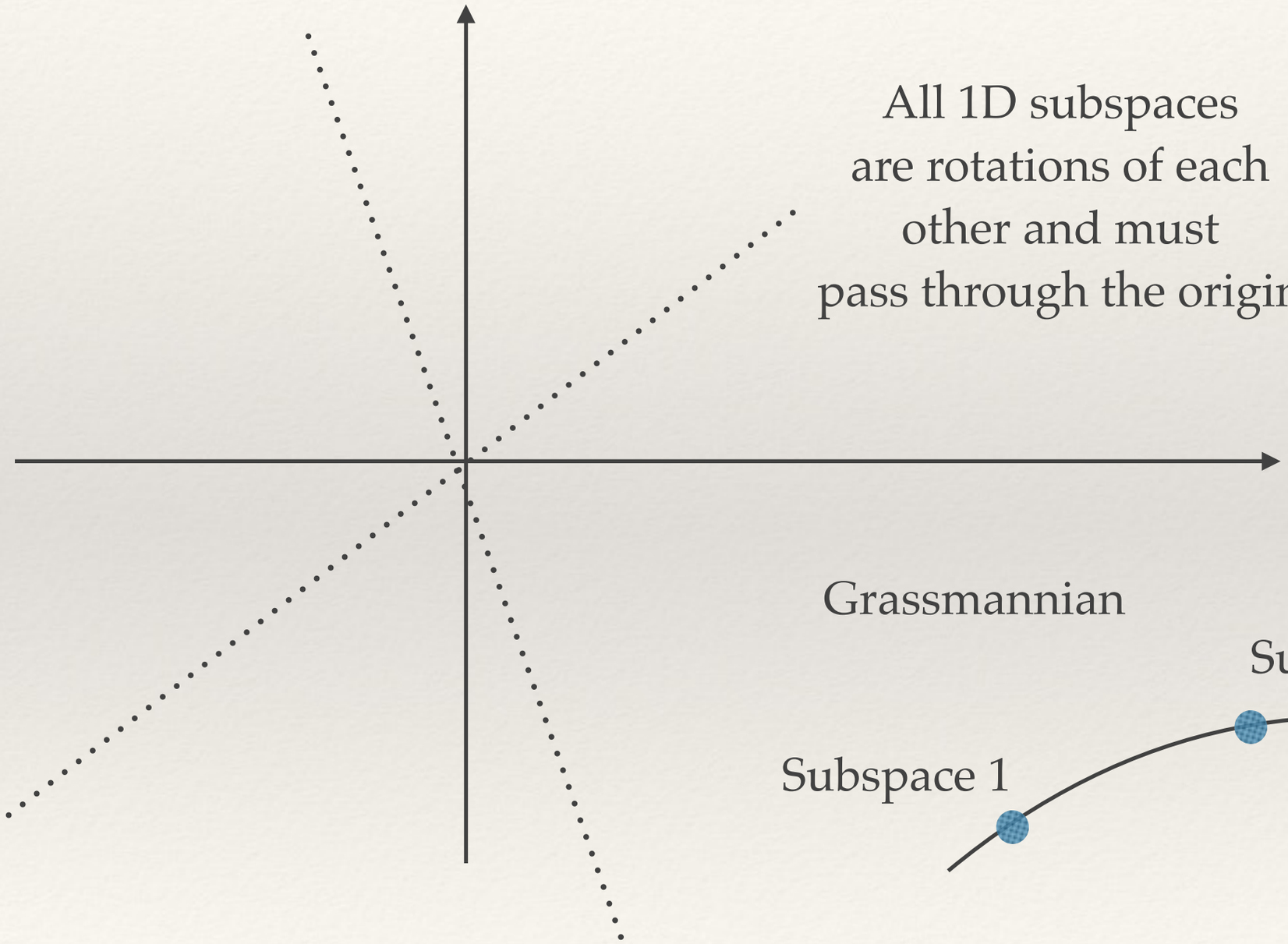


Grassmannian Manifolds



1809-1877

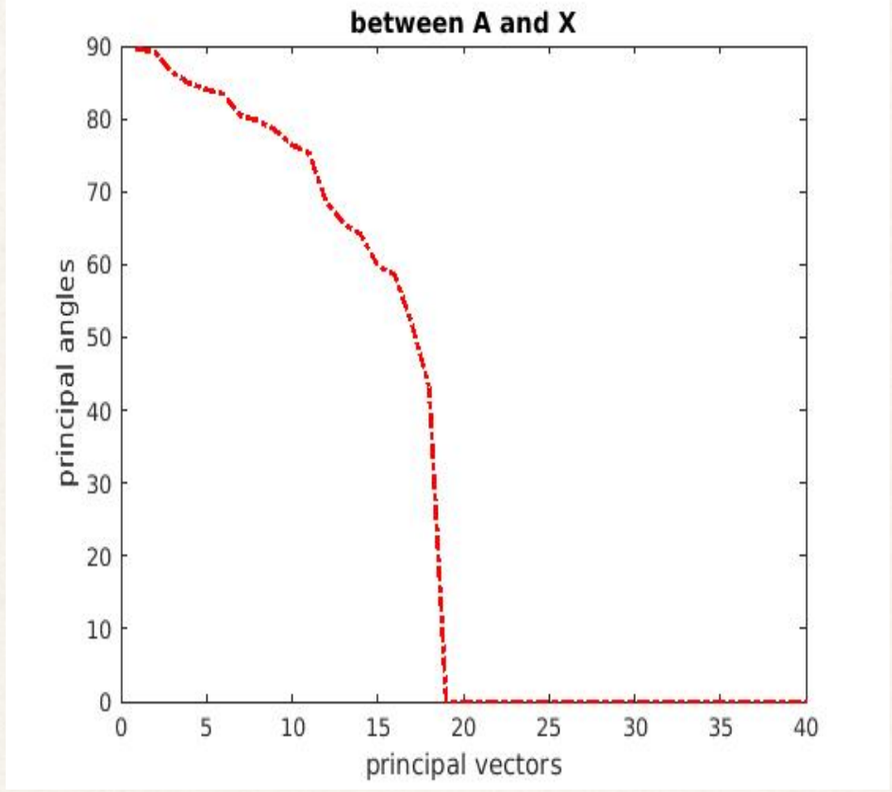
2D Example

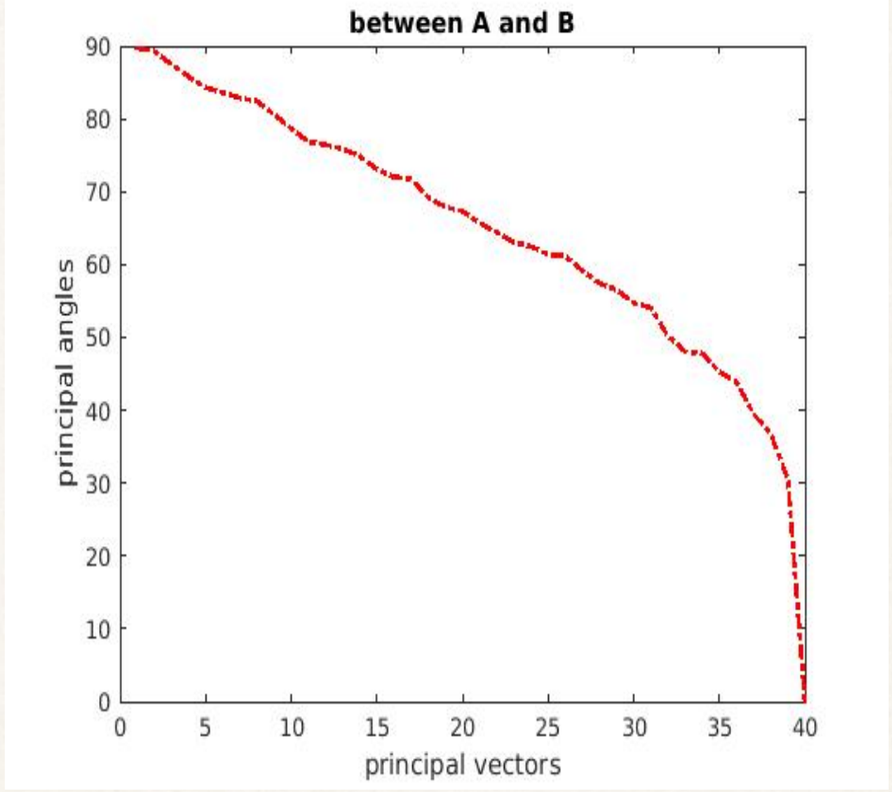


All 1D subspaces
are rotations of each
other and must
pass through the origin

Grassmannian

Principal Angles between Subspaces





$$P_H^T P_T = U_1 \Gamma V^T, \quad R_H^T P_T = -U_2 \Sigma V^T$$

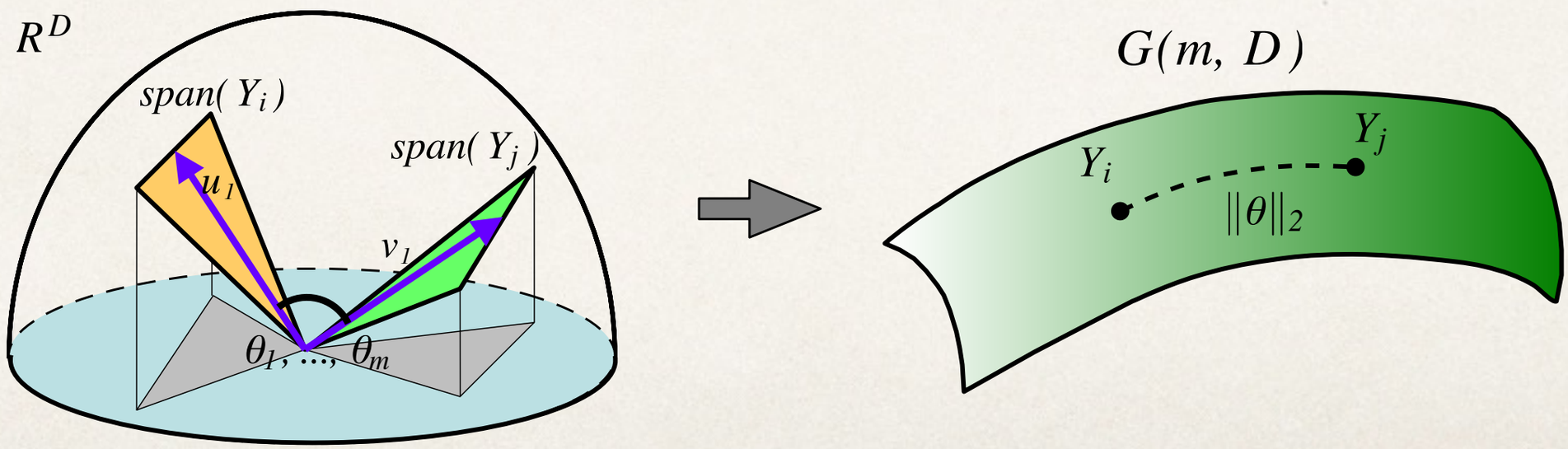
The $d \times d$ diagonal matrices Γ and Σ are particularly important since they represent $\cos(\theta_i)$ and $\sin(\theta_i), i = 1, \ldots, d$, where θ_i are the so-called *principal angles* between the subspaces P_H and P_T.

Geodesic Flow Kernels

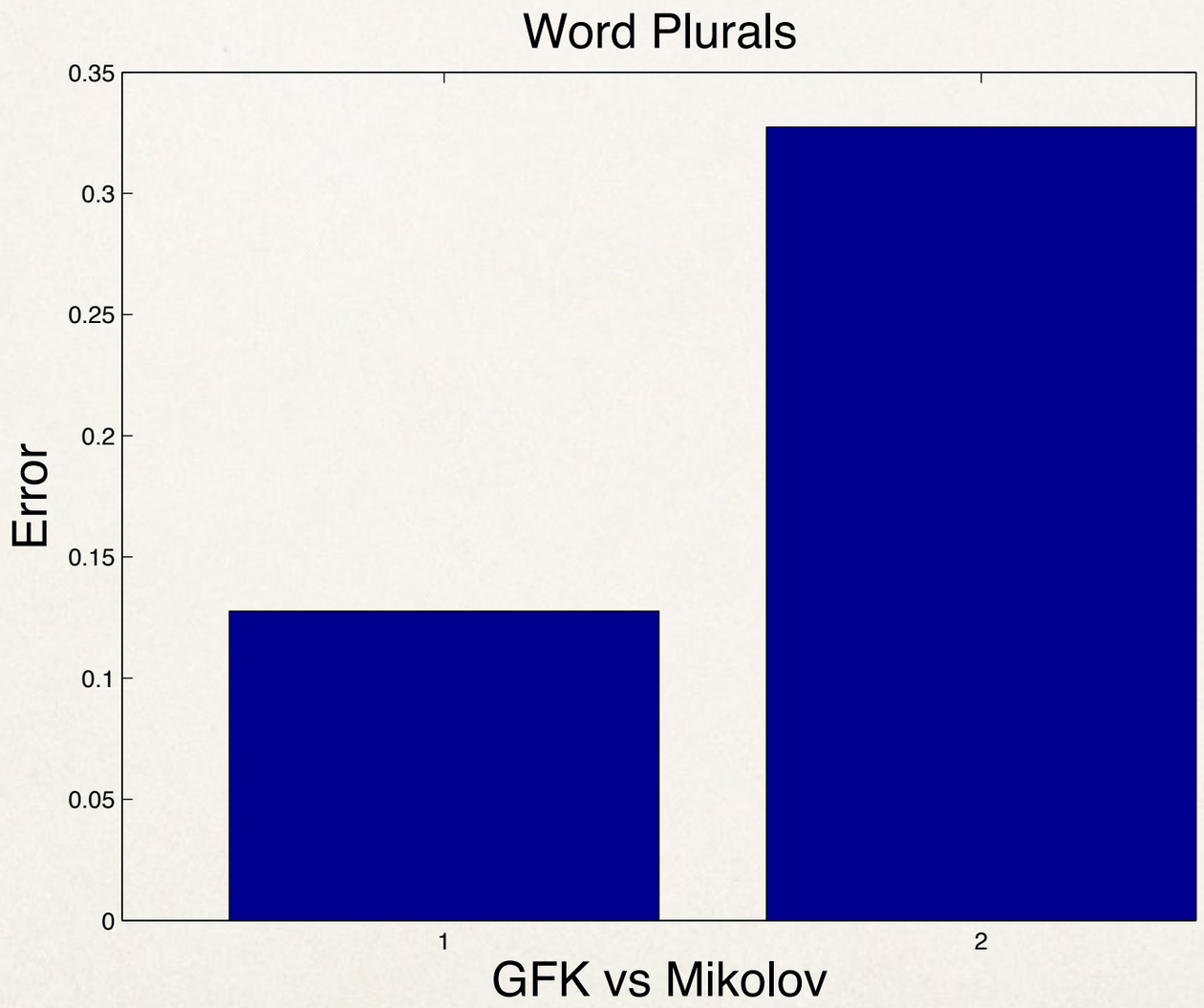
(Gong et al., 2012)

$$G_R = \begin{pmatrix} P_S U_1 & R_S U_2 \end{pmatrix} \begin{pmatrix} \Lambda_1 & \Lambda_2 \\ \Lambda_2 & \Lambda_3 \end{pmatrix} \begin{pmatrix} U_1^T P_S^T \\ U_2^T R_S^T \end{pmatrix}$$

$$\langle z_i, z_j \rangle_R = \int_0^1 (\Phi(t)_R^T x_i)^T (\Phi(t)_R^T x_j) dt = x_i^T G_R x_j$$



Word Analogies using Subspace Flow Kernels

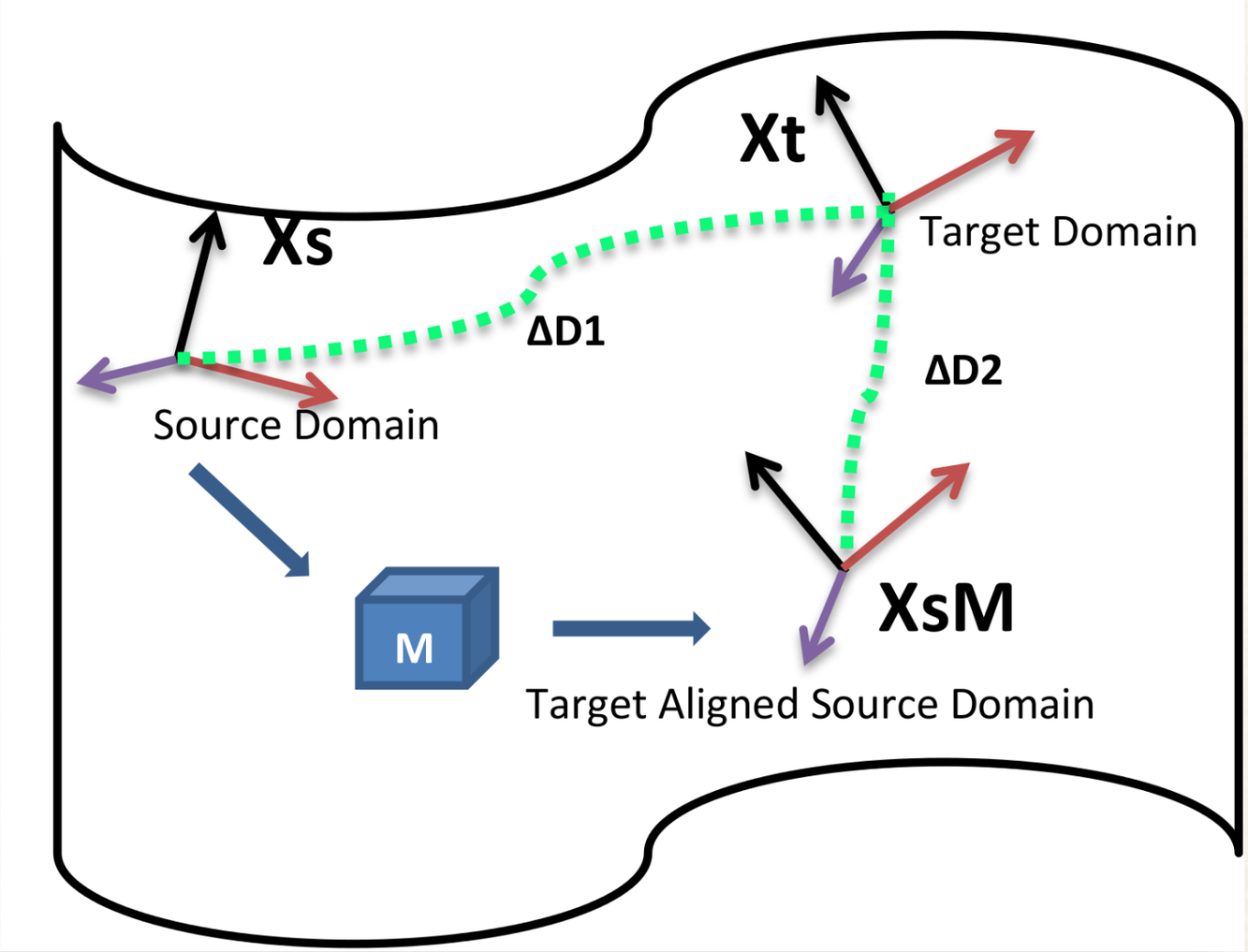


Word Analogy Experiments (Mahadevan and Chandar, Arxiv)

Config	Model	CosADD	CosMUL	GFKCosADD	GFKCosMUL
win=2, pos=True	SGNS	45.15%	54.27%	57.62%	**62.35%**
	SVD	43.66%	60.05%	58.66%	**65.91%**
win=5, pos=True	SGNS	53.17%	62.19%	67.68%	**71.70%**
	SVD	52.14%	71.34%	62.46%	**74.18%**
win=2, pos=False	SGNS	49.41%	63.21%	71.17%	**76.01%**
	SVD	50.87%	65.82%	67.11%	**72.45%**
win=5, pos=False	SGNS	56.14%	74.43%	81.06%	**84.64%**
	SVD	60.82%	75.14%	72.29%	**79.15%**

Google Dataset

Subspace Alignment
(Fernando et al, 2014)



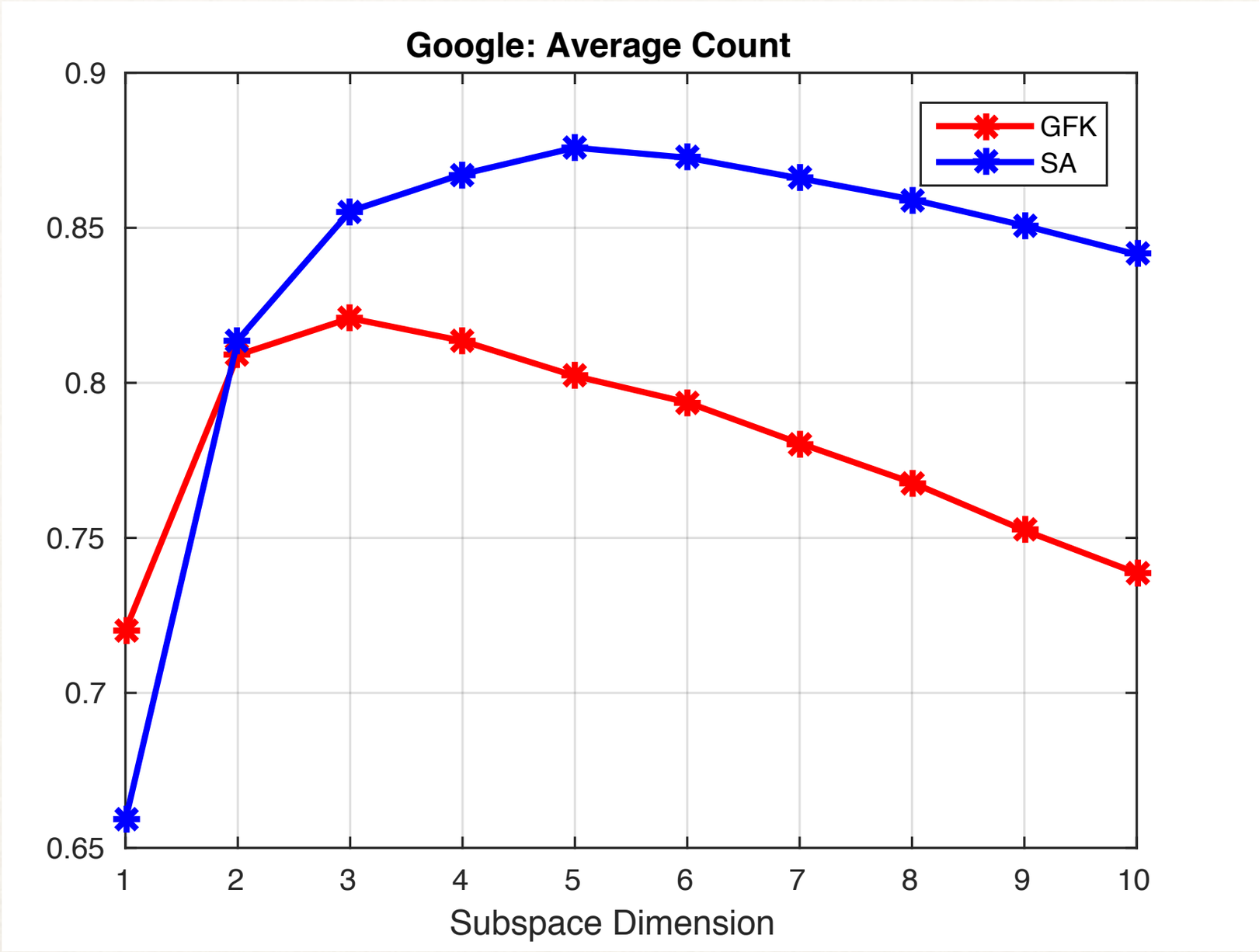
Subspace Alignment

$$F(M) = ||X_S M - X_T||_F^2 \tag{1}$$

$$M^* = argmin_M(F(M)) \tag{2}$$

$$\begin{aligned} M^* &= argmin_M ||X_S' X_S M - X_S' X_T||_F^2 \qquad (3) \\ &= argmin_M ||M - X_S' X_T||_F^2. \end{aligned}$$

Word Analogies using Subspace Alignment (Mahadevan, 2016)



Summary

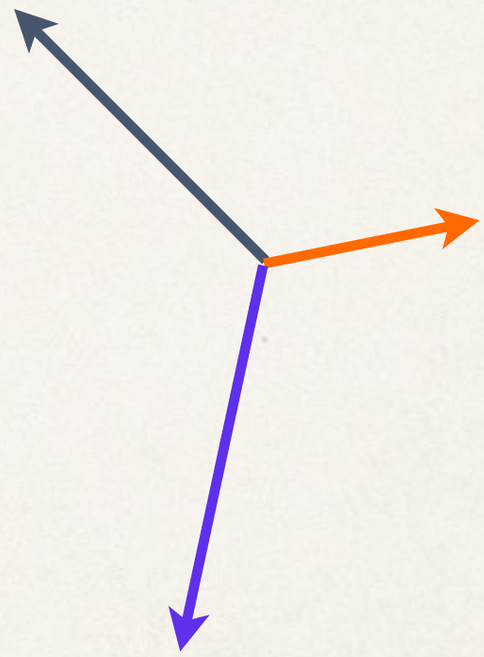
- The goal was to illustrate how optimization plays a central role in an exciting problem in NLP: reasoning about words and text
- Many open problems
  - Does this approach extend to metaphors ("The stock market crashed")
  - How to represent documents using more complex representations?

Outline

- Normed vector spaces: minimum norm problems
- Projections
- Convex sets and functions
- Subgradients and subdifferential calculus
- Conjugate functions
- Duality theory

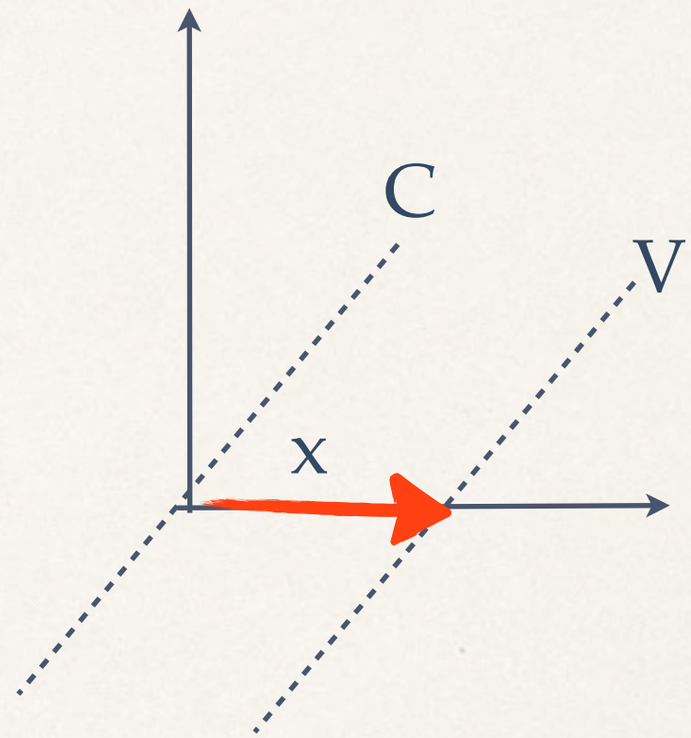
Vector spaces

- We will be primarily investigating optimization in **normed** vector spaces (or linear spaces)
  - Euclidean n-dimensions
  - Continuous functions
  - Matrices
- **Problem**: find a vector x in X whose norm is minimum



Linear Varieties

- A subspace C of a vector space X is a subset of X that forms a vector space in its own right



- Linear varieties are ``displaced'' subspaces of a vector space

$$V = C + x = \{y + x | y \in C, x \in X\}$$

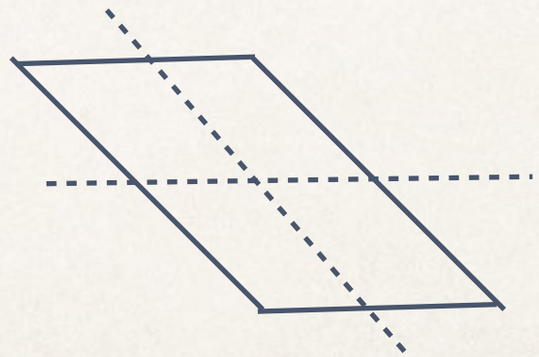
Linear functionals

- A linear functional T:X => R on a vector space X is defined as:

$$T(\alpha x + \beta y) = \alpha T(x) + \beta T(y), \alpha, \beta \in \mathbb{R}, x, y \in X$$

- Linear functionals will play a critical role in optimization

- The level sets of linear functionals are called hyperplanes



$$H = \{x | T(x) = c, x \in X, c \in \mathbb{R}\}$$

Dual spaces

- Given a vector space X, the dual space X* is defined as the space of all real-valued linear functionals on X
- Dual spaces play a crucial role in the theory of optimization
- Example:
  - In Euclidean n-dimensions, $X^* = X$
  - Lp and l_p spaces

Normed linear space

- A normed linear space is a vector space X on which is defined a **sublinear functional** called a norm:

$$\|x\| \geq 0 \text{ for all } x \in X, \|x\| = 0 \Leftrightarrow x = 0$$

$$\|x + y\| \leq \|x\| + \|y\|$$

$$\|\alpha x\| = |\alpha| \|x\|, \alpha \in \mathbb{R}$$

Examples of norms

- L1 norm: $\|x\|_1 = \sum_i |x_i|$
- L2 norm: $\|x\|_2 = \left(\sum_i x_i^2 \right)^{\frac{1}{2}}$
- L-infinity or max-norm: $\|x\|_\infty = \max_i |x_i|$
- Frobenius norm of a matrix $\|A\|_F = \left(\sum_{i,j} a_{i,j}^2 \right)^{\frac{1}{2}}$
- Spectral norm of a matrix:

$$\|A\|_2 = \max_{\|x\|>0} \frac{\|Ax\|_2}{\|x\|_2}$$

l_p spaces

- Consider the set of all sequences of real numbers
  - x = (x_0, x_1, ...) such that $\|x\|_p = \left(\sum_i |x_i|^p \right)^{\frac{1}{p}}$ is bounded
  - Here, $p \in (1, \infty)$ where $\|x\|_\infty = \sup_i |x_i|$

L_p spaces

- L_p spaces are defined correspondingly for function spaces
- The $L_p(a,b)$ space is defined as the vector space of all functions for which

$$\|f\|_p = \left(\int_a^b |f(x)|^p dx \right)^{\frac{1}{p}} < \infty$$

Holder inequality

- One of the most important inequalities in normed linear spaces is the Holder inequality
- Given two positive numbers p and q such that $1/p + 1/q = 1$
- and $x = (x_0, x_1, ...)$ is a vector in l_p space
- and $y = (y_0, y_1, ...)$ is a vector in l_q space
- then

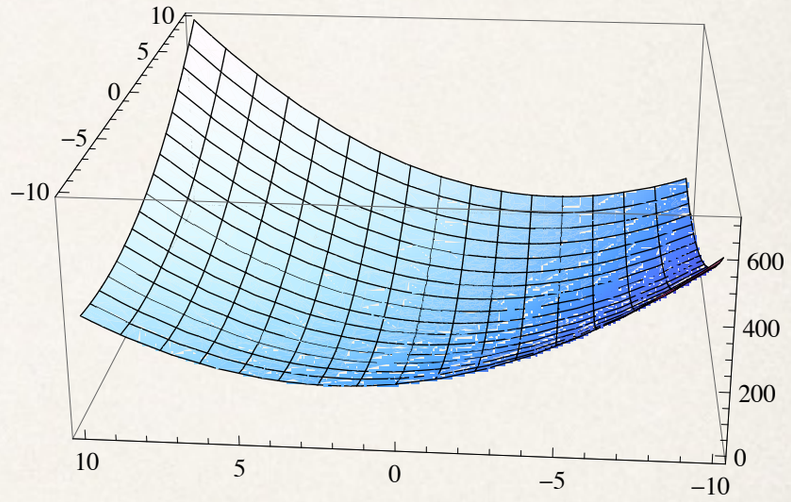
$$\sum_{i=1}^{\infty} |x_i y_i| \leq \|x\|_p \|y\|_q$$

Example: Linear Equation Solving

- Consider the simple system of linear equations

$$f(x) = \frac{1}{2}x^T Ax - b^T x + c$$

$$\begin{pmatrix} 3 & 2 \\ 2 & 6 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 2 \\ -8 \end{pmatrix}$$



- It is easy to see that the solution is

$$x = \begin{pmatrix} 2 \\ -2 \end{pmatrix}$$

Example: Solving linear equations

- One of the most widely used applications of optimization is solving Ax = b

- This system of linear equations is the mainstay of scientific computing

- How do we convert this into an optimization problem?

$$f(x) = \frac{1}{2}x^T Ax - b^T x + c$$

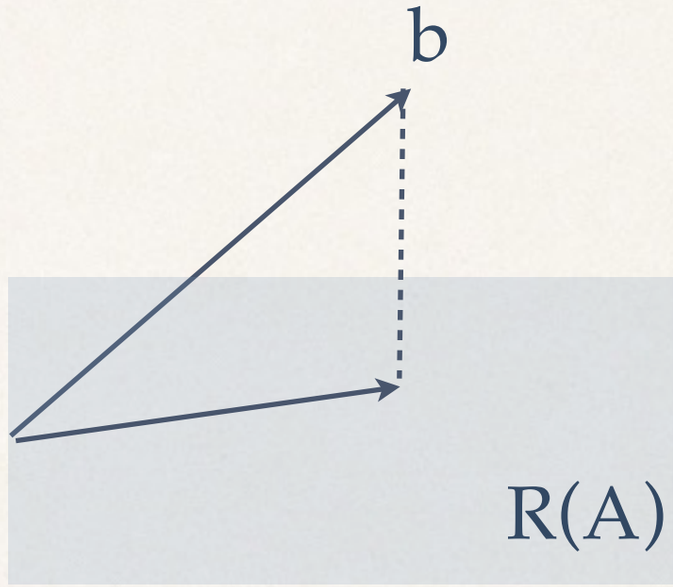
If A is symmetric

$$\frac{\partial f}{\partial x} = Ax - b$$

General Solutions to Ax=b

- We consider the general case where A is non-symmetric

- Also, A is not square and the vector b lies outside the column space of A



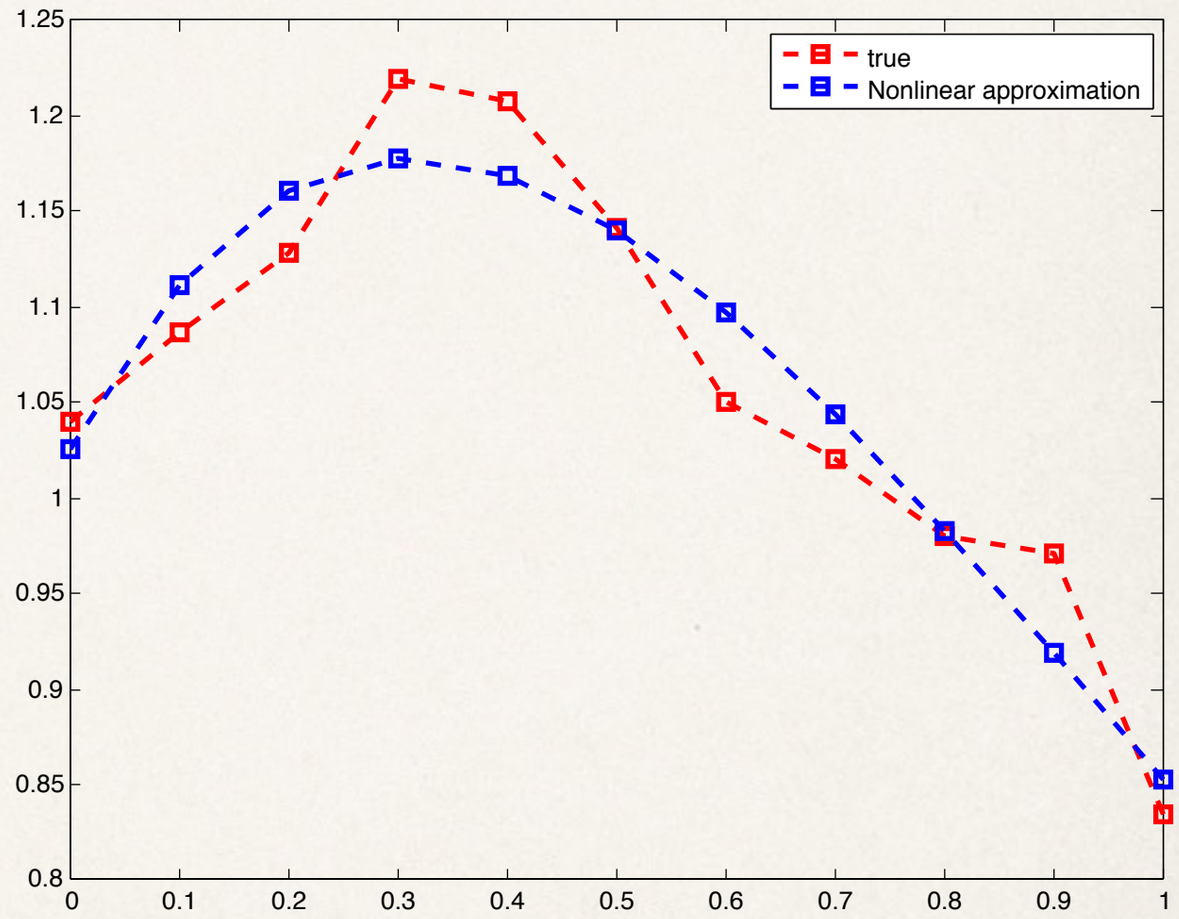
$$\min_{x \in \mathbb{R}^n} \|b - Ax\|_2$$

Show that $x^* = A^\dagger b = (A'A)^{-1} A'b$

Example: Nonlinear regression

- Regression is the problem of fitting a parametric function to a set of data points (x,y)

- In nonlinear regression, the parameters influence the function in a complex manner



$$y = w_1 \exp^{-\alpha_2 t} \cos(\alpha_3 t) + w_2 \exp^{-\alpha_1 t} \cos(\alpha_2 t)$$

Nonlinear Regression

- Let us express this problem using our generic problem specification

$$\min_{x \in \Omega} f(x)$$

- The feasible region is $\Omega = \mathbb{R}^5$

- We define a smooth differentiable (convex) loss function

$$J(\alpha, w) = \|y - \Phi(\alpha)w\|^2 = \| \left(I - \Phi(\alpha)\Phi^\dagger(\alpha)\right) y\|^2$$

Matrix Completion Problem

- The Netflix™ competition
  - Given a set of user ratings of movies
  - Learn a recommendation function over all movies
- The input is a very large sparse matrix
- The goal is to fill in missing values

- Sensor network triangulation
  - Suppose we have a network of low power sensors
  - Each sensor computes distances to nearby sensors
  - Problem is to fill in distance matrix over all sensors

Matrix Completion Problem

- Consider a square matrix of size N x N
- This contains N^2 numbers
- However, if the matrix is of low rank (say R << N)
- Then, the number of parameters is much less
- (2N - R)*R

- Formulation:

Minimize rank(X)

such that $X_{i,j} = M_{i,j}$

- NP-hard problem!
- Convex relaxation:

Minimize $|X|_*$

such that $X_{i,j} = M_{i,j}$

Well-defined optimization problem

- Typically, we are interested in minimizing (or maximizing) some function f over some desired subspace Y of vector space X

- The structure of these subspaces needs to be specified carefully

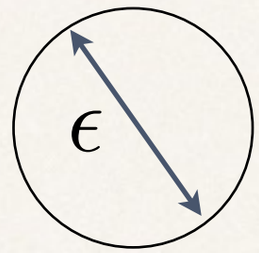
  - Example: minimize $3x + 4y$ such that $x + y > 0$

  - This is an ill-defined optimization problem!

Topology of Normed Spaces

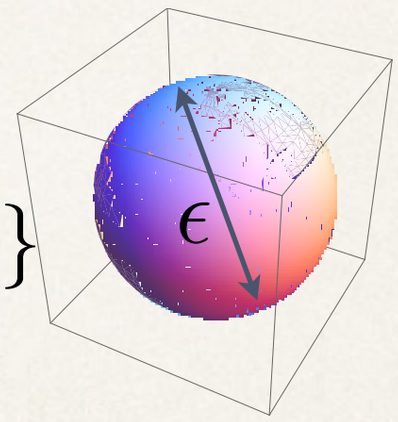
- Convergence: a sequence of vectors $\{x_n\} \to x^*$ if for any $\epsilon > 0$,

$$\|x_n - x^*\| \leq \epsilon, \forall n \geq M$$



- x^* is the limit point

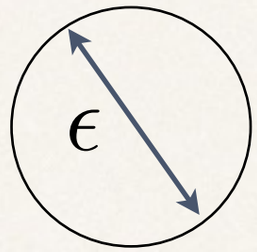
- A ball or sphere around x is the set of all points $N(x, \epsilon) = \{y \in X | \|x - y\| < \epsilon\}$

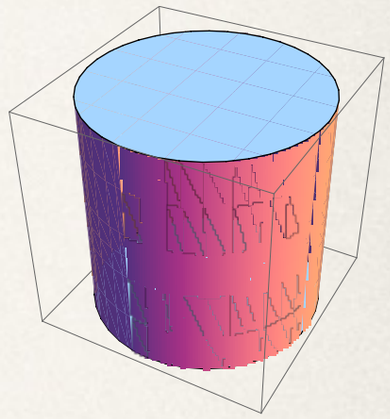


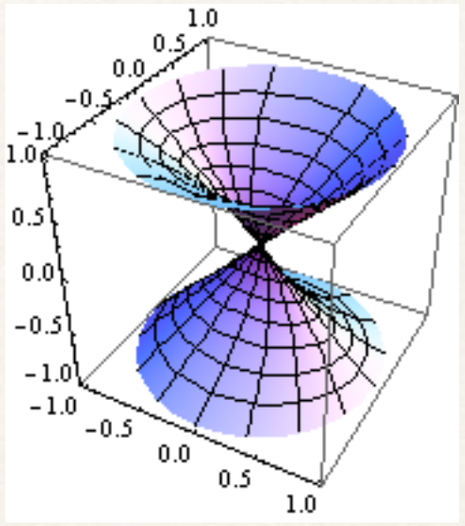
- A set S is open if every point is contained in a sphere in S. A set is closed if its complement is open

Vector Spaces: Topology

- Open set: {x: |x| < 1}
- Closed set: complement of open set
- Interior: all points around which a sphere exists inside
- Boundary: all points inside S excluding the interior
- Compact: closed and bounded set







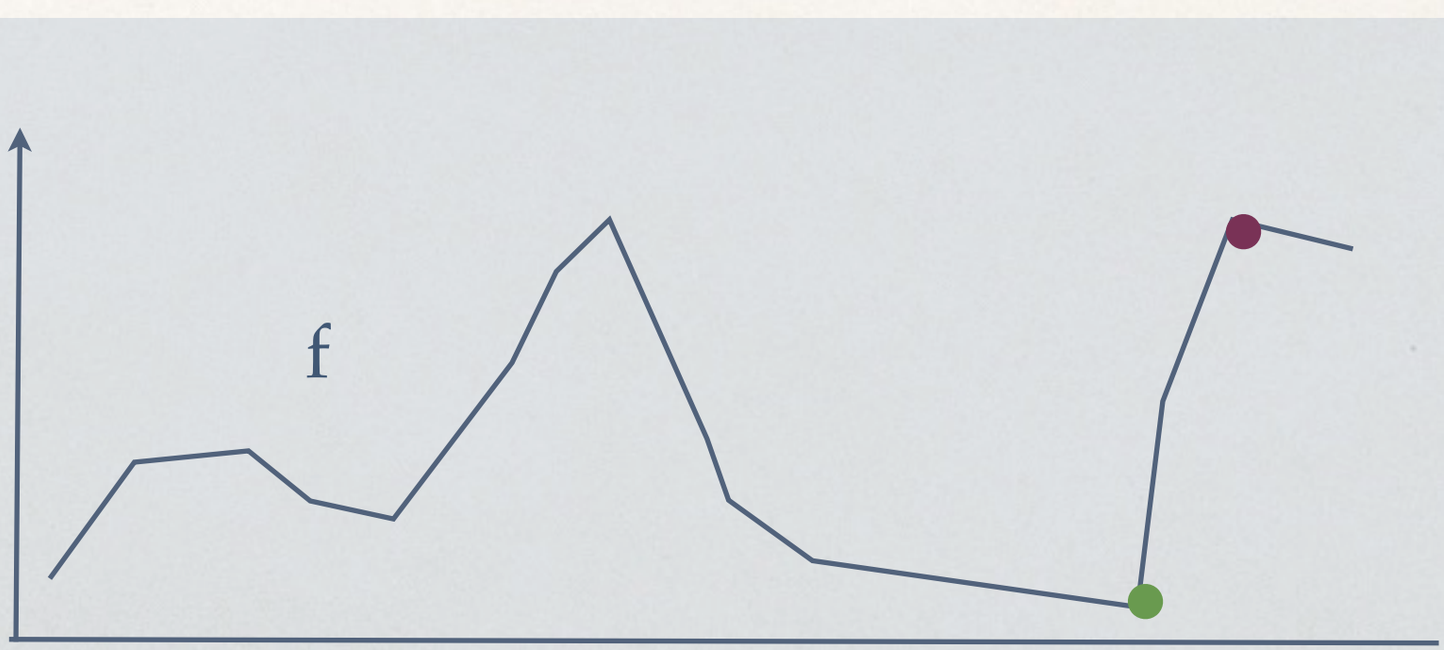
Functions: Continuity

- A function on a normed space is **continuous** if $\|x_n - x^*\| < \delta \Rightarrow \|f(x_n) - f(x^*)\| < \epsilon$

- Weierstrass Theorem: A continuous function on a compact set S has a minimum in S

Weierstrass Theorem

If $f : \mathbb{R}^n \to \mathbb{R}$ is a continuous function and if S is a non-empty, closed, and bounded subset of $\mathbb{R}^n$, there exists some $x^* \in S$ such that $f(x^*) \leq f(x), \forall x \in S$. Similarly, there exists $y^* \in S$ such that $f(y^*) \geq f(x), \forall x \in S$.



Inner product spaces

- An inner product space is a vector space equipped with an inner product

$$\langle x, y \rangle : X \times X \to \mathbb{R}$$

$$\langle \alpha x + \beta y, z \rangle = \alpha \langle x, z \rangle + \beta \langle y, z \rangle$$

$$\langle x, x \rangle = 0 \Leftrightarrow x = 0$$

$$\langle x, x \rangle \geq 0, \|x\| = \sqrt{\langle x, x \rangle}$$

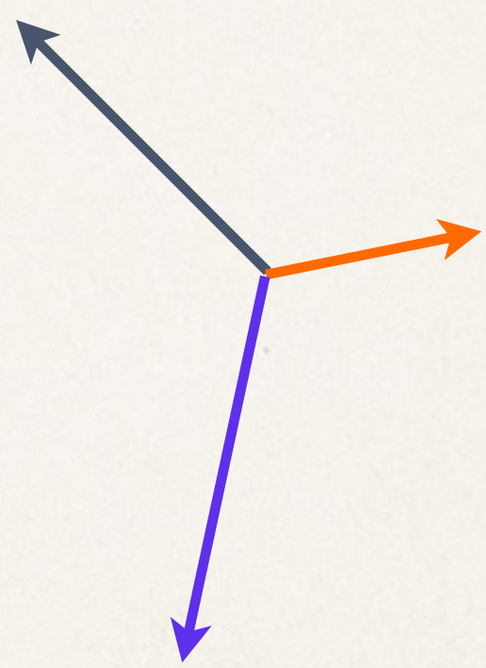
$$\langle x, y \rangle = \langle y, x \rangle$$

- Examples:
  - Euclidean space
  - All continuous functions
  - Symmetric matrices

Examples of Inner product spaces

- Euclidean n-dimensional space

$$\langle x, y \rangle = x^T y = \sum_{i=1}^{n} x_i y_i$$

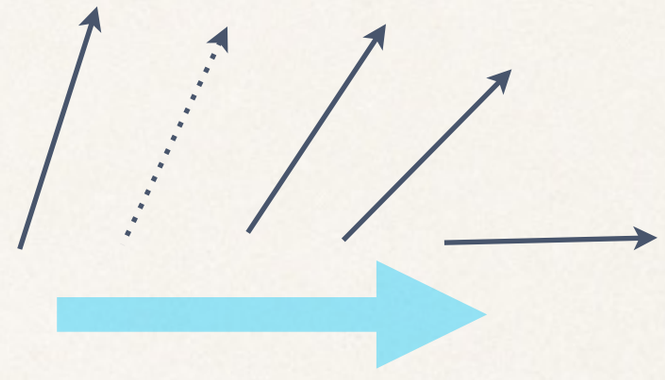


- All continuous functions C[a,b]

$$\langle f, g \rangle = \int_a^b f(x) g(x) dx$$

Complete Spaces

- Cauchy sequence: $\|x_n - x_m\| \to 0$



- A space S is complete if all Cauchy sequences converge to an element of S

- Examples
  - Real numbers
  - N-dimensional Euclidean space

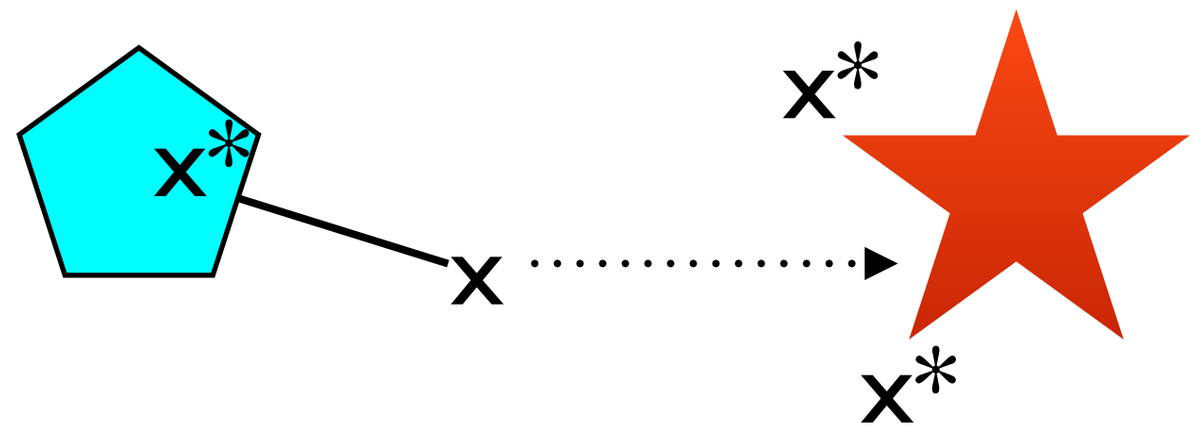
The space $\{x \in \mathbb{R}^2 : \|x\| < 0\}$ is not complete

Banach and Hibert Spaces

- A normed vector space that is complete is called a Banach space
  - Euclidean, n-dimensions
  - All real-valued functions on a Markov chain M
- A complete vector space with an inner product <x, y> is a Hilbert space
- Banach spaces
  - Most general space for continuous optimization
  - Hahn-Banach theorem: foundational theorem
- Hilbert spaces
  - Used in ML, statistics, etc.
  - Projections are defined

Projections

- The concept of a projection is fundamental to many optimization methods



$$x^* = \Pi_C(x) = \text{argmin}_{u \in C} \|u - x\|_2^2$$

Projections onto a subspace

- Given a complete subspace M of a Hilbert space X, and an element x, the minimum norm problem is to find the element in M closest to x

- **Key idea:** the vector connecting x and the closest element in M must be orthogonal to every vector in M

$$d = \min_{y \in M} \|x - y\|$$

$$\langle x - \hat{x}, y \rangle = 0, \text{ for all } y \in M$$

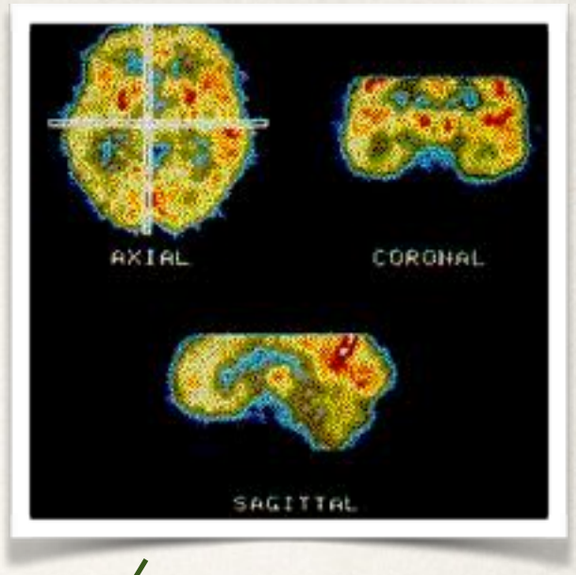
$$\hat{x} = \sum_i \alpha_i y_i$$

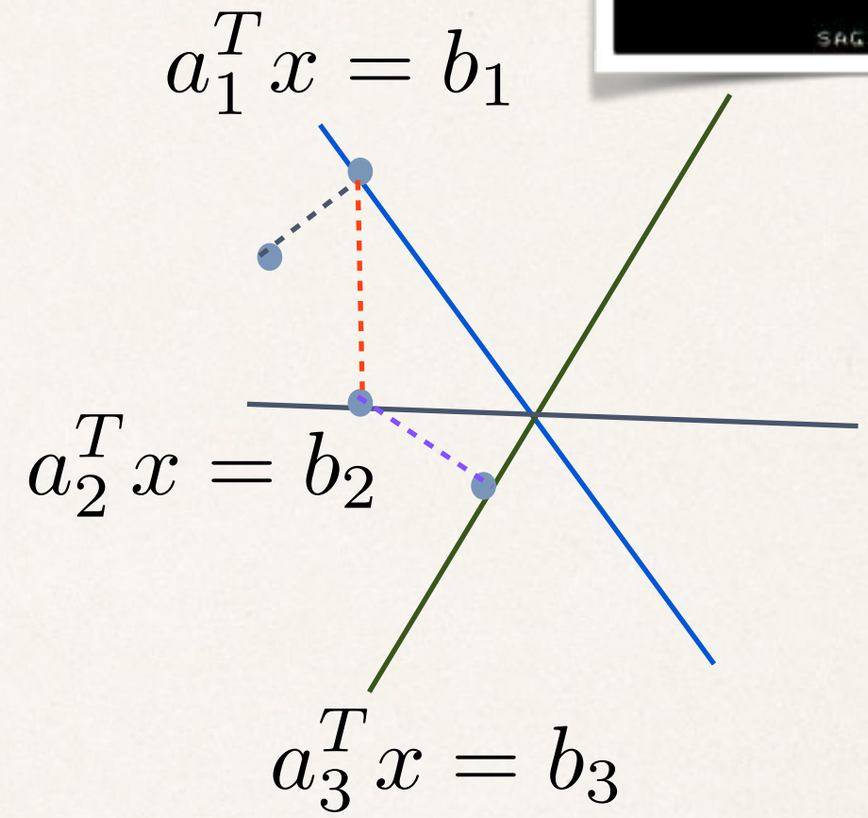
$$\langle x, y_i \rangle = \sum_j \alpha_j \langle y_j, y_i \rangle$$

Kaczmarz algorithm

- Consider a system of ``overdetermined'' linear equations A x = b
  - Number of columns n <= number of rows m
- One of the most beautiful algorithms for solving Ax=b

Medical imaging (CAT, PET)

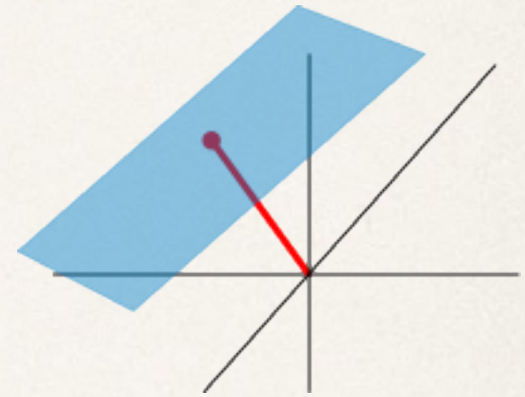




$$x^{k+1} = x^k + \lambda_k \frac{b_i - \langle a_i, x^k \rangle}{\|a_i\|^2} a_i$$

Affine Spaces

- A subset C of a vector space X is called affine if



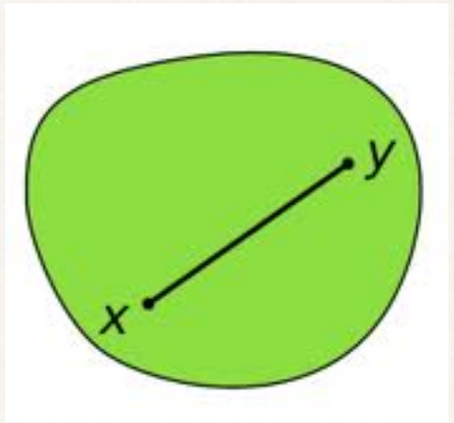
$$x_1, x_2 \in C \Rightarrow \alpha x_1 + (1-\alpha)x_2 \in C,\ \alpha \in \mathbb{R}$$

- The solution set of a system of linear equations is an affine set

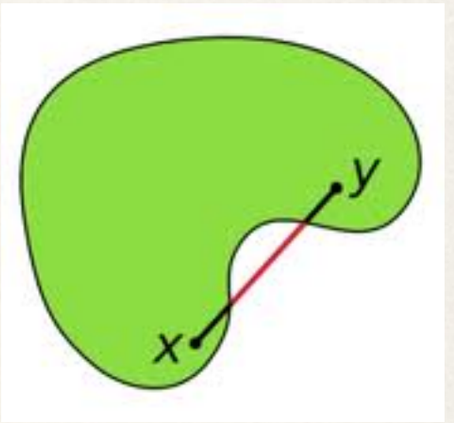
$$C = \{x | Ax = b, A \in \mathbb{R}^{m \times n}, b \in \mathbb{R}^m\}$$

Convex Set

convex



non-convex



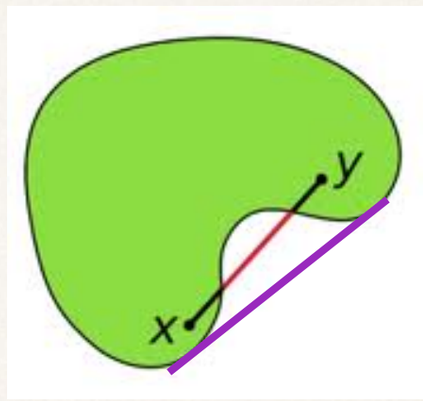
- A subset C of a vector space X is called convex if

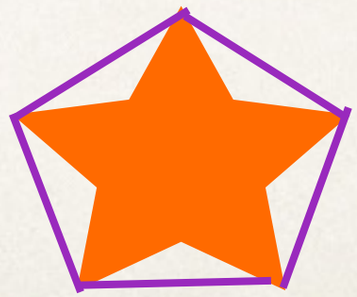
$$x_1, x_2 \in C \Rightarrow \alpha x_1 + (1-\alpha)x_2 \in C, \ \alpha \in (0,1)$$

- Examples:
  - System of linear inequalities $C = \{x | Ax \leq b\}$
  - Norm balls: $C = \{x | \|x\| \leq K\}$
  - Set of positive semi-definite matrices

Convex Hull

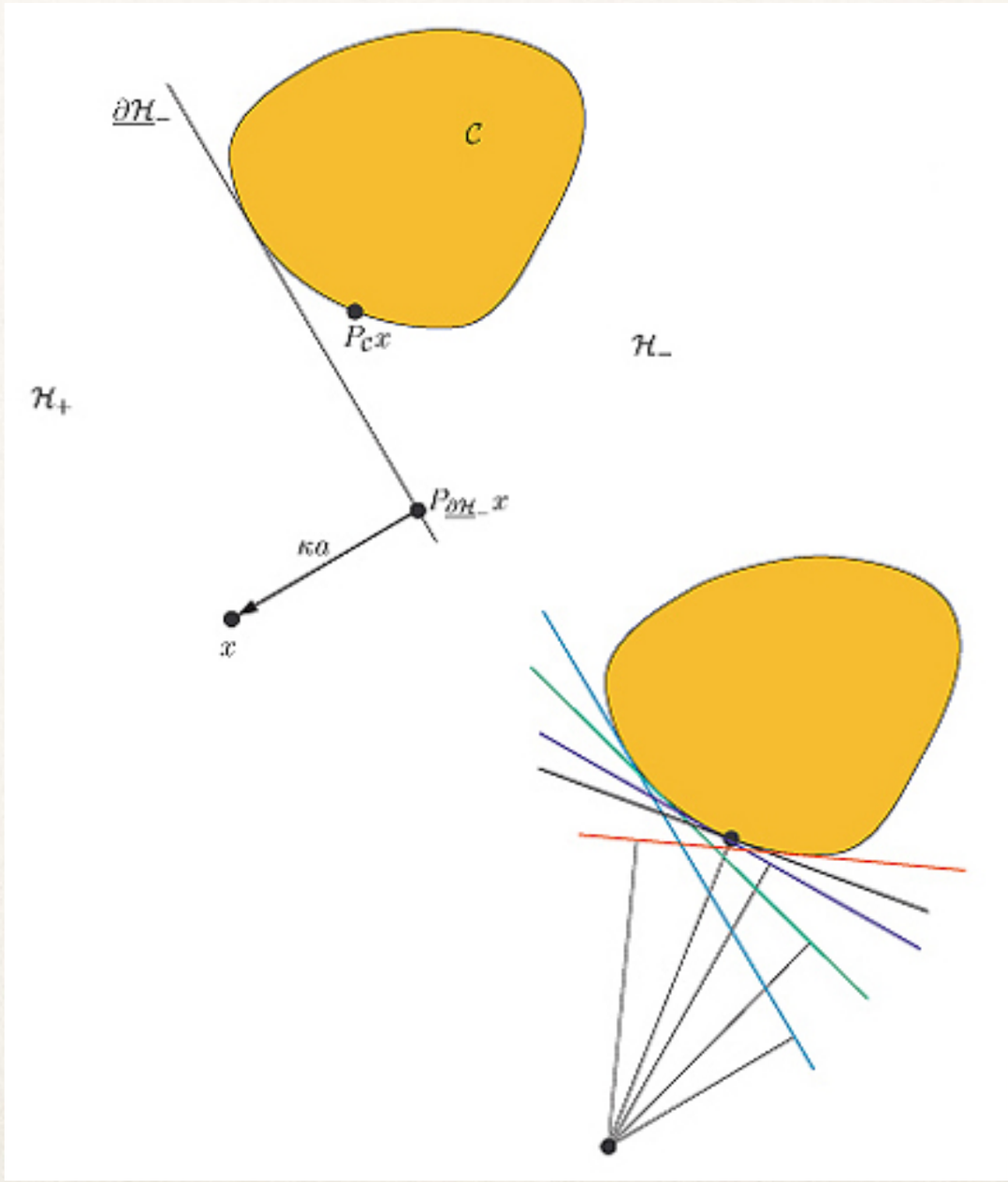
- The convex hull of a set of points C is the set of all points that can be expressed as a convex combination of a finite subset of points within C



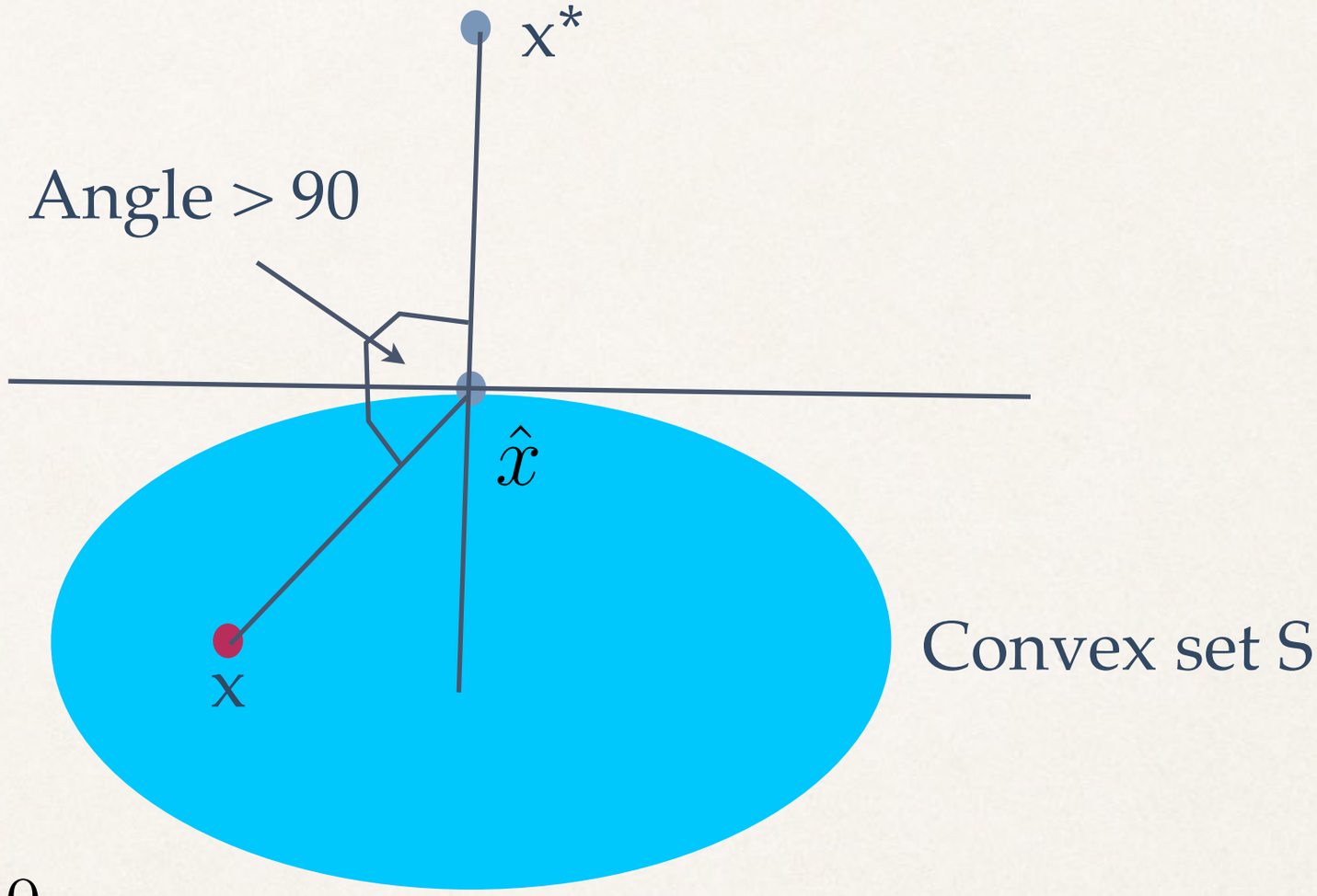


Geometric Hahn-Banach Theorem

Hyperplanes separate points from convex sets



Separating Hyperplane Theorem



$$(x^* - \hat{x})'(x - \hat{x}) \leq 0$$

Proof of Geometric HB Theorem

Consider the optimization problem

$$\hat{x} = \text{argmin}_{x \in S} \|x - x^*\|$$

Since S is a closed bounded set, according to Weierstrass' theorem, this optimization problem is well-defined. Let x be any point in S. Since S is convex,

$$(1-\lambda)\hat{x} + \lambda x \in S$$

It follows that

$$\begin{aligned} \|\hat{x} - x^*\|^2 &\leq \|\hat{x} + \lambda(x - \hat{x}) - x^*\|^2 \\ &= \|\hat{x} - x^*\|^2 + 2\lambda(\hat{x} - x^*)'(x - \hat{x}) + \lambda^2\|x - \hat{x}\|^2 \end{aligned}$$

which implies

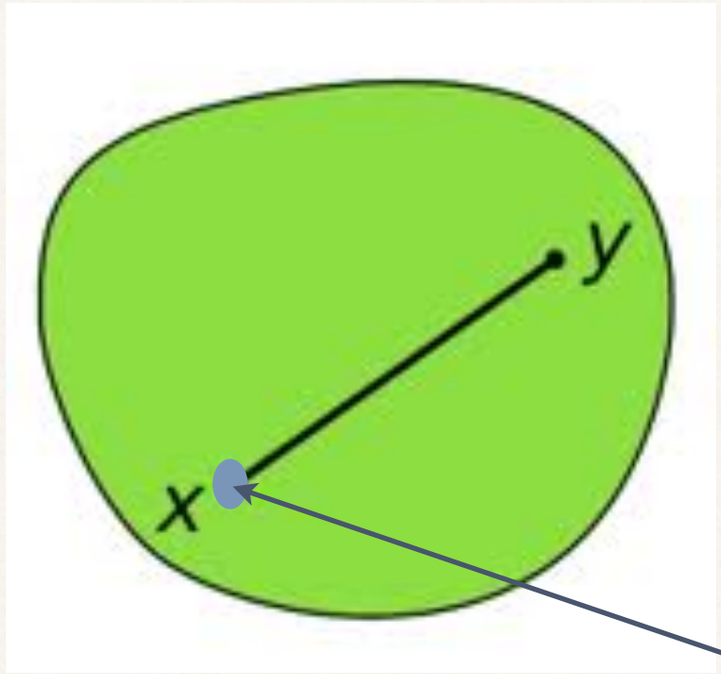
$$2\lambda(\hat{x} - x^*)'(x - \hat{x}) + \lambda^2\|x - \hat{x}\|^2 \geq 0$$

Consequently

$$(\hat{x} - x^*)'(x - \hat{x}) \geq 0 \Rightarrow (x^* - \hat{x})'(x - \hat{x}) \leq 0$$

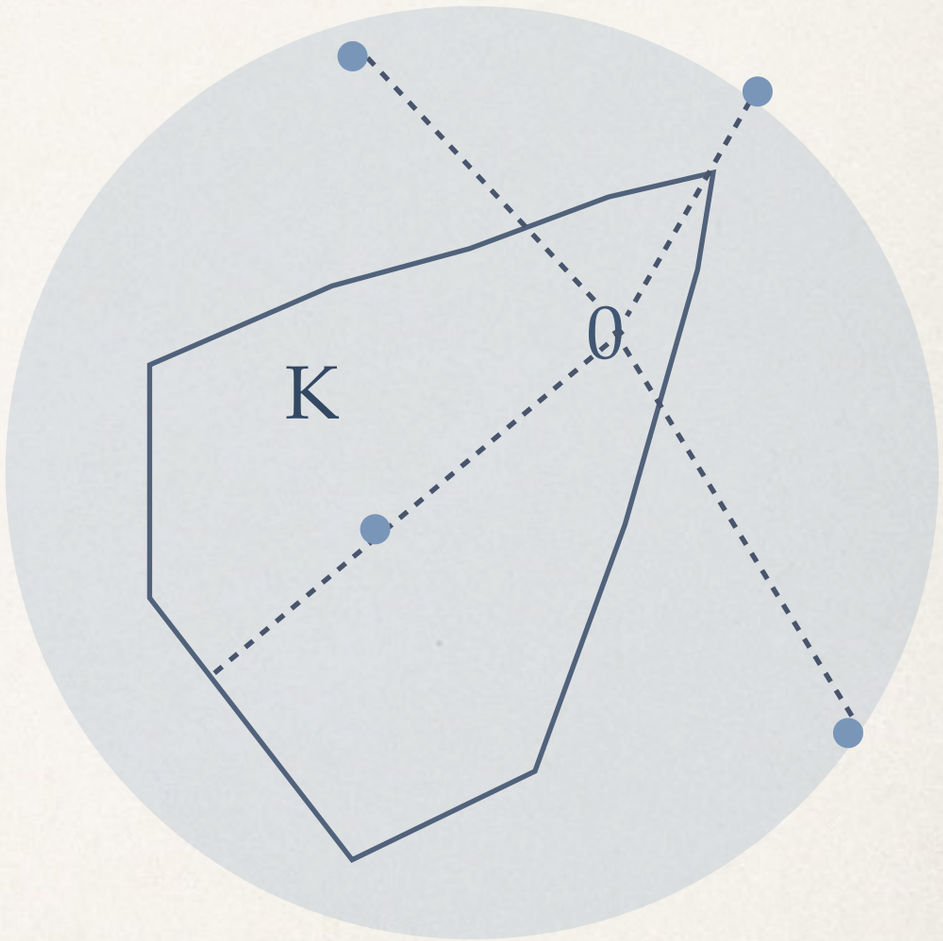
Relative Interior

A point x in a convex set C is in its **relative interior** if for each y in the convex hull of C, the line segment joining x and y is in C



Minkowski functional of a convex set

- Let K be a convex set in a linear normed space X, such that 0 is an interior point of K

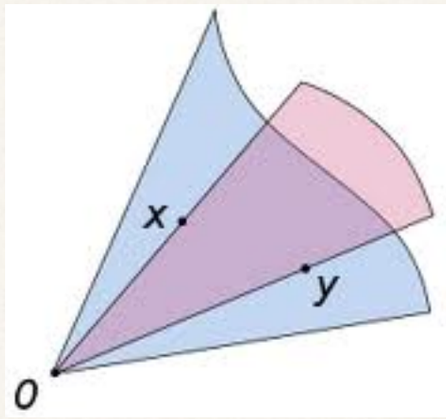


- The **Minkowski functional** p of K is defined on X as

$$p(x) = \inf\{r : \frac{x}{r} \in K, r > 0\}$$

Cones

- A subset C of a vector space X is a cone if $x \in C, \theta \geq 0 \Rightarrow \theta x \in C$

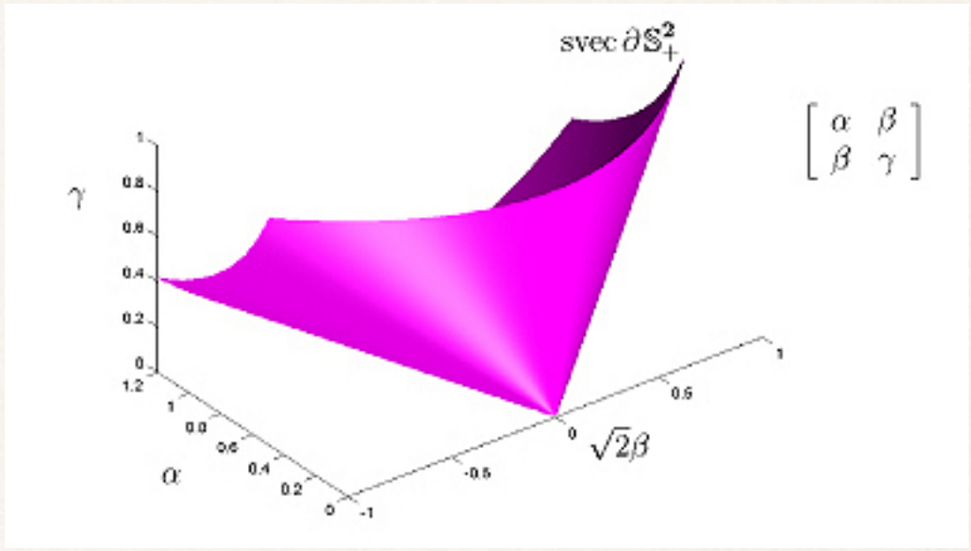


- A convex cone C is both a convex set and a cone

- Examples:

  - The set of all non-negative matrices

Positive semidefinite cones

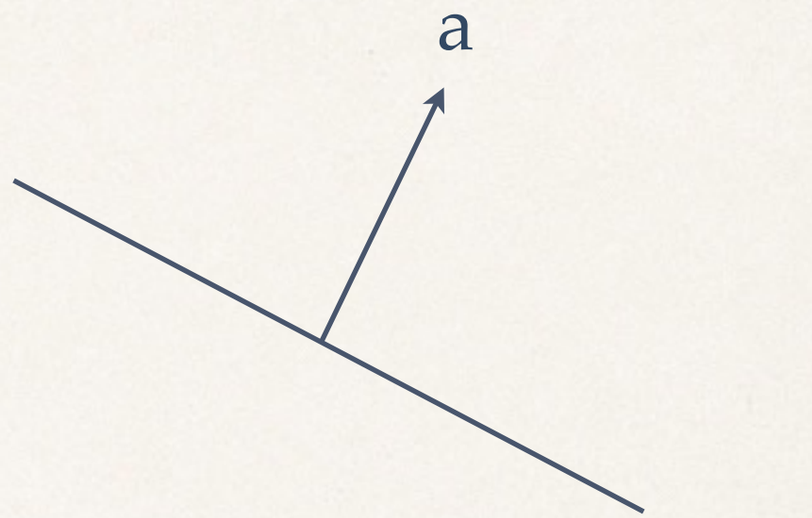


- The set of symmetric positive semi-definite matrices forms a cone

$$S_+^n = \{X \in S^n | X \succeq 0\}$$

$$x^T(\theta_1 A + \theta_2 B)x = \theta_1 x^T A x + \theta_2 x^T B x \geq 0$$

Hyperplanes



- A hyperplane H is also a maximal proper subset of a vector space X

- H divides X into two ``half spaces''

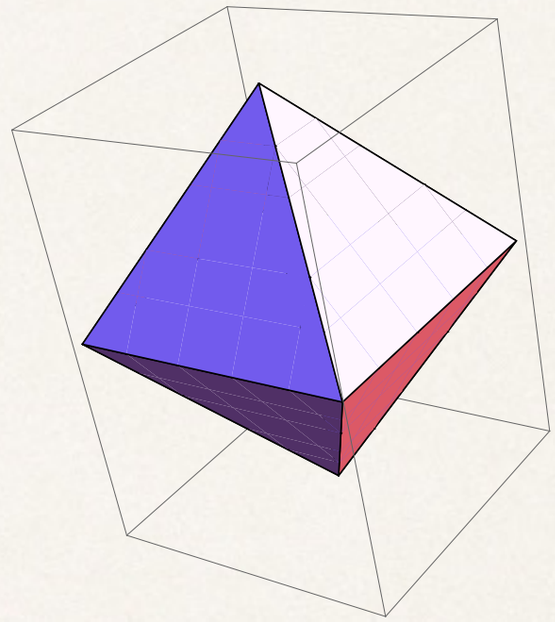
$$V_1 = \{x | a^T x < b\}$$

$$V_2 = \{x | a^T x > b\}$$

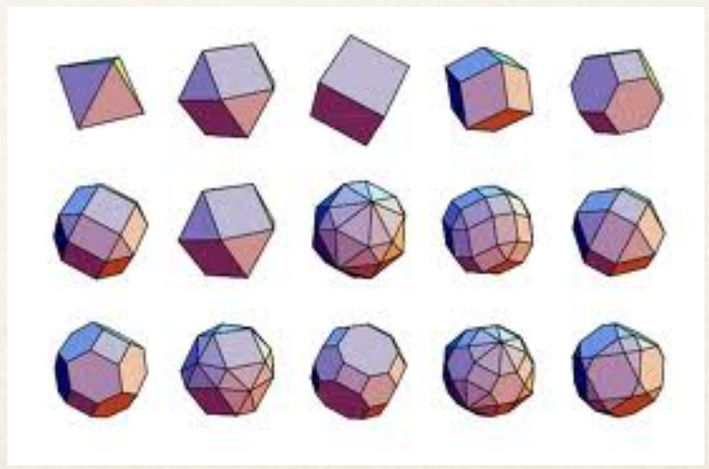
Polyhedra

- A polyhedron is a subset P that satisfies a set of linear inequalities

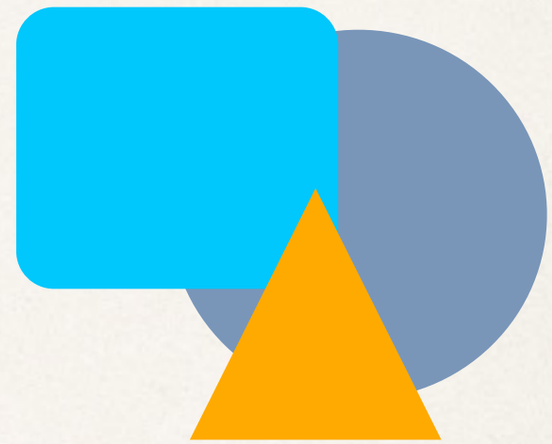
$$C = \{x | Ax \leq b\}$$



- Example:
  - Unit simplex



Operations preserving convexity



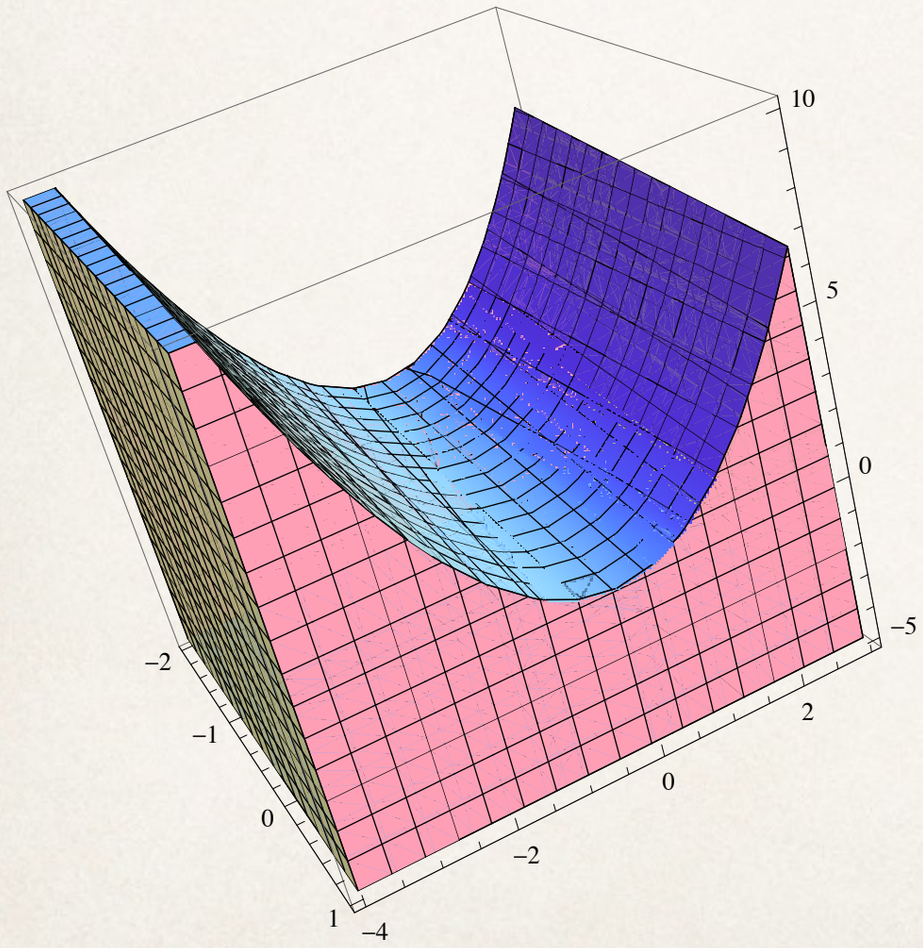
Convex feasibility problem

- The intersection of convex sets is also convex

$$S_+^n = \bigcap_{z \neq 0} \{X \in S^n | z^T X z \geq 0\}$$

Convex Functions

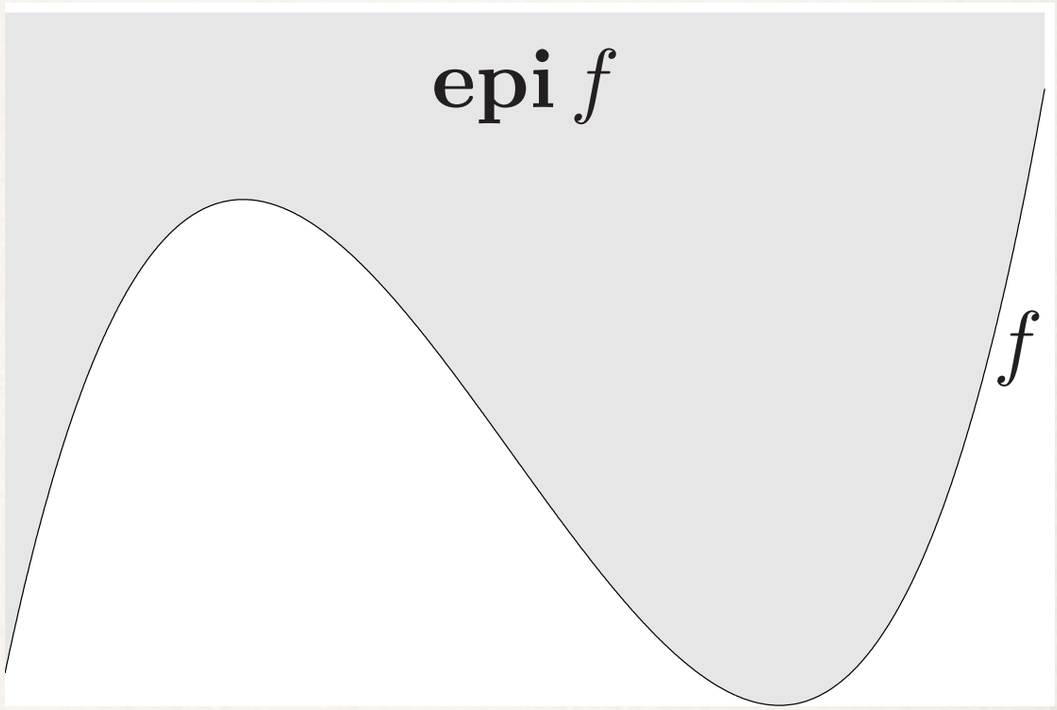
$$f(\lambda x + (1-\lambda)y) \leq (1-\lambda)f(x) + \lambda f(y)$$



Convex functions: Examples

- Linear and affine functions are convex
- Exponentials, negative logarithm, and positive entropy
- Norms are convex

Epigraph of a convex function



- A function is convex if and only if its epigraph is a convex set

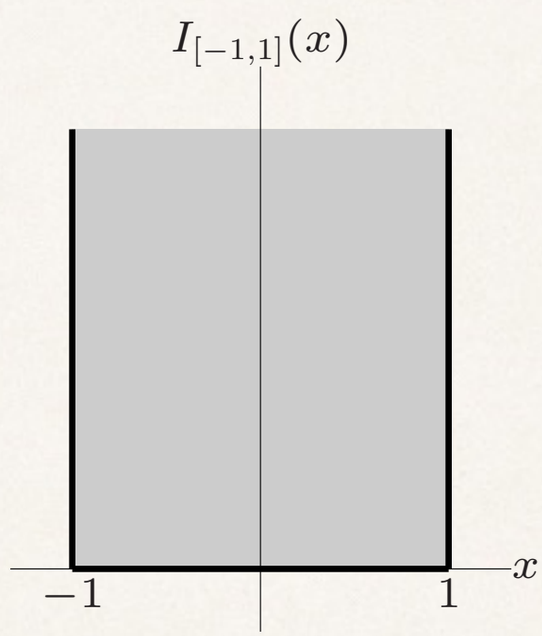
$$\text{epi} f = \{(x,t) | x \in \text{dom}(f), f(x) \leq t\}$$

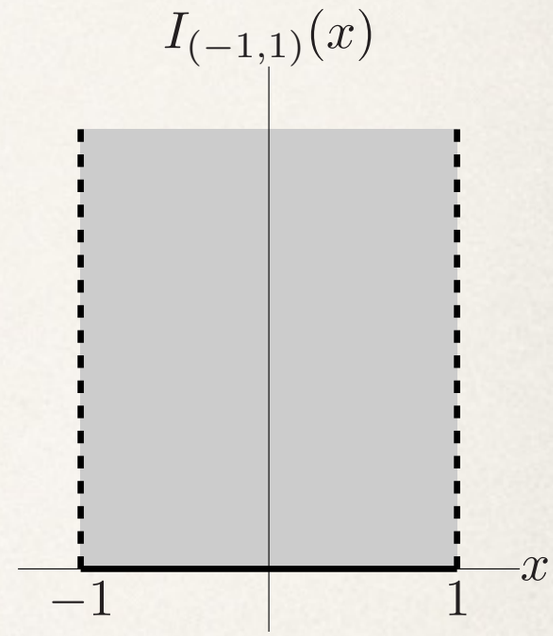
Indicator Function

- The indicator function of a set C is defined as:

$$I_C(x) = 0 : x \in C$$

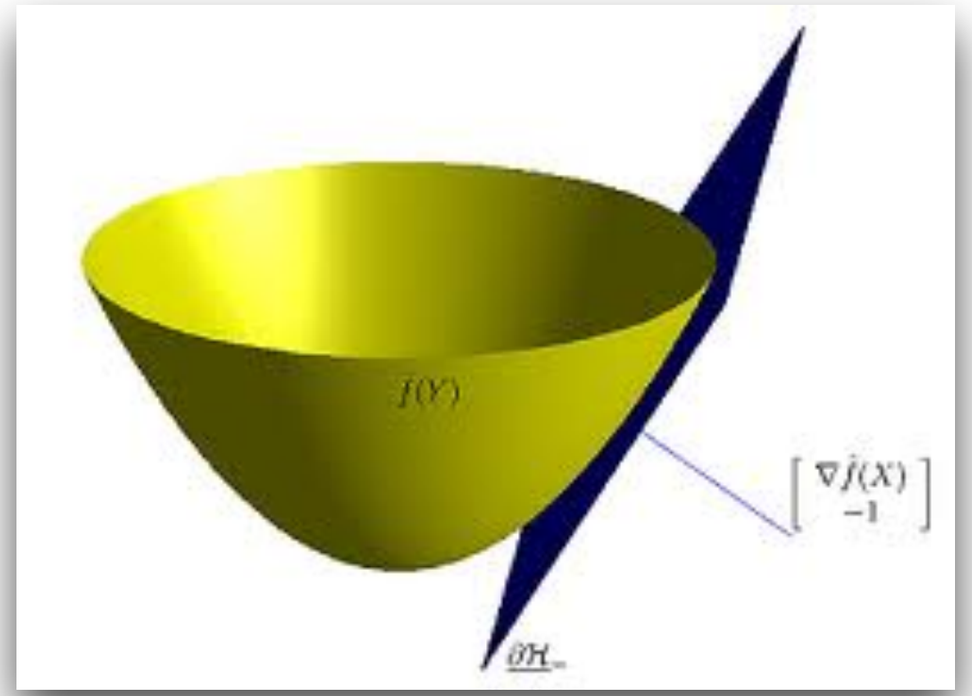
$$I_C(x) = \infty : x \notin C$$



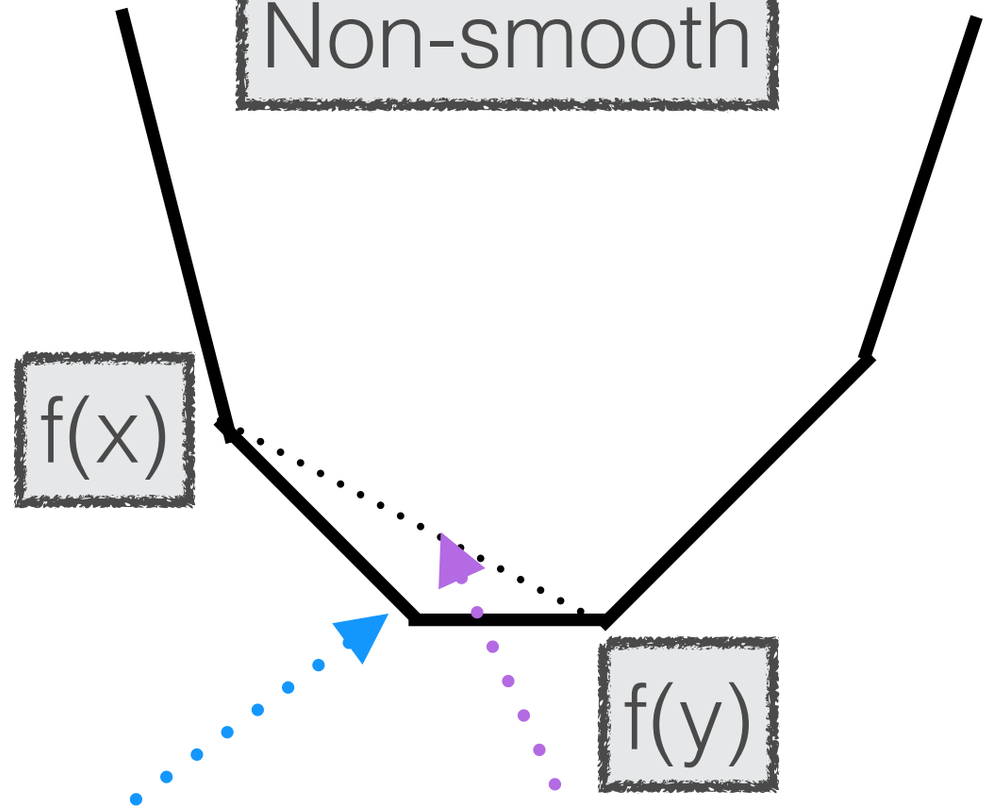


Convex functions

Smooth

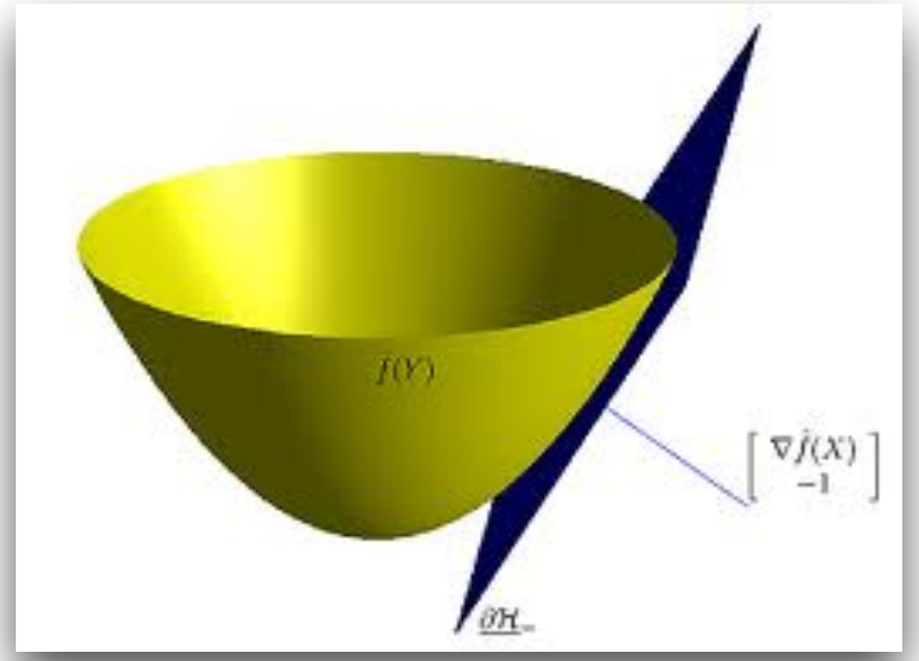


Non-smooth



$$f((1-\lambda)x + \lambda y) \leq (1-\lambda)f(x) + \lambda f(y)$$

Differentiable convex functions



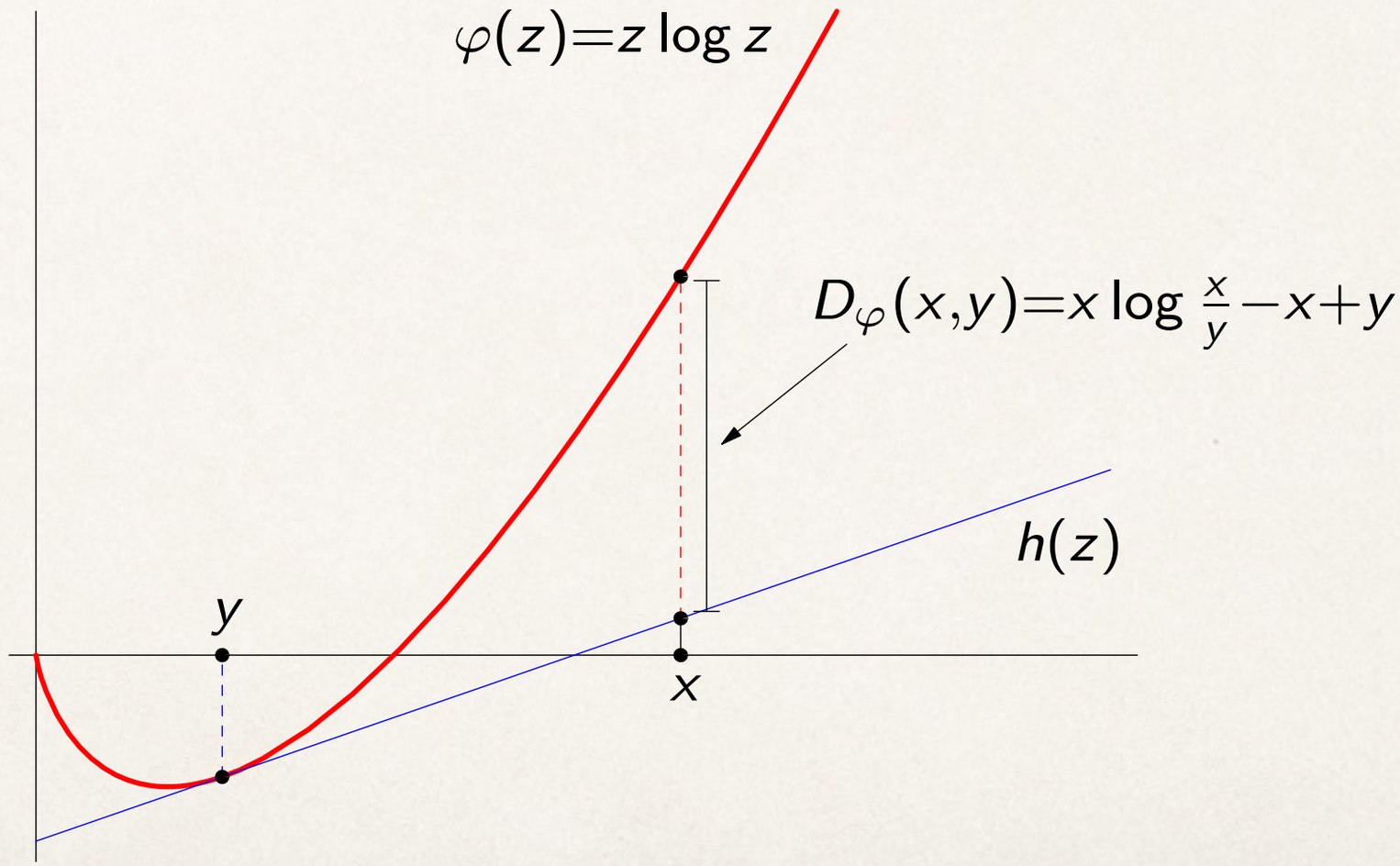
- A basic identity for convex functions

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

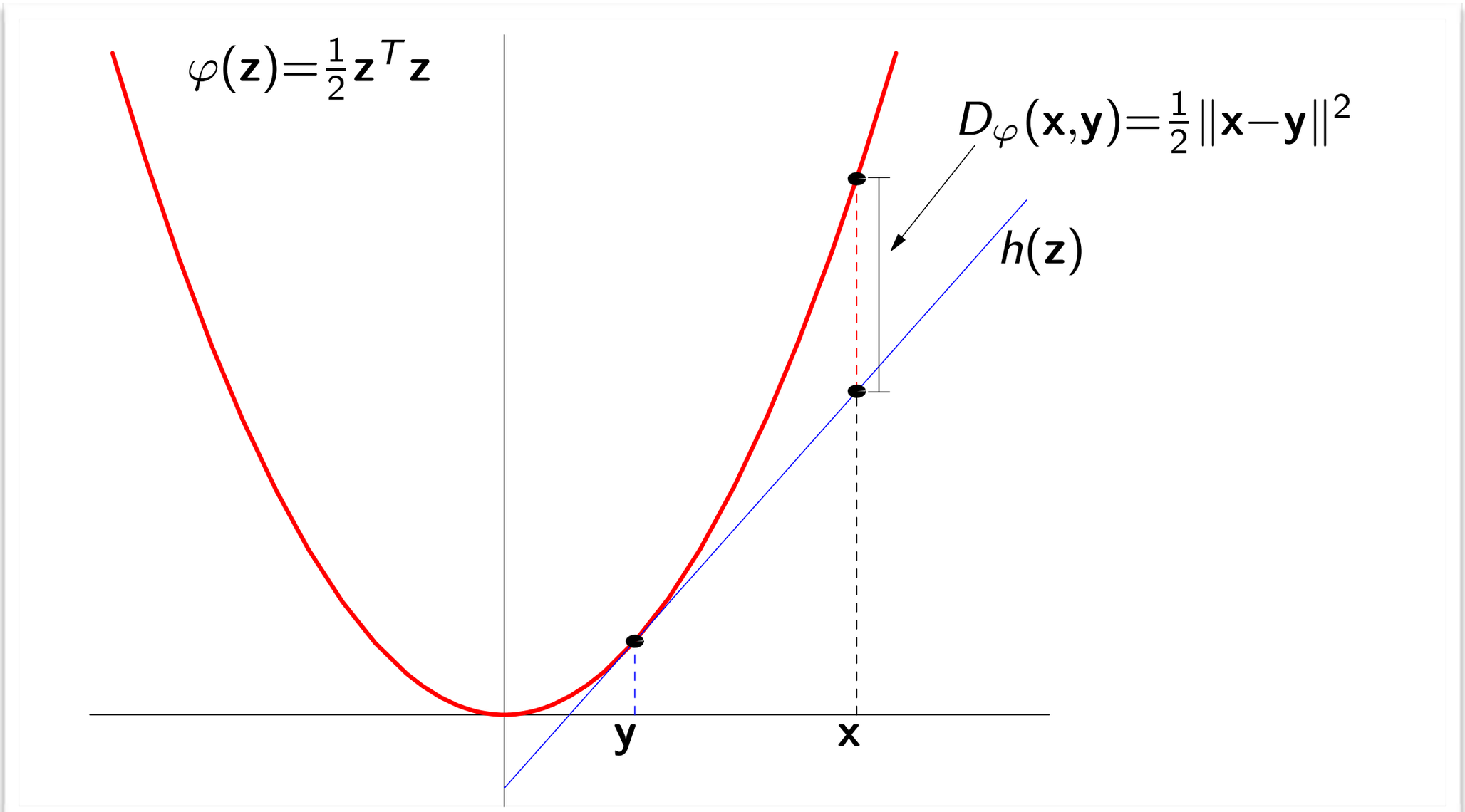
$$\nabla^2 f(x) \succeq 0 \quad \text{for all } x \in \mathrm{dom} f$$

Bregman Divergence

$$D_\varphi(\mathbf{x}, \mathbf{y}) = \varphi(\mathbf{x}) - \varphi(\mathbf{y}) - (\mathbf{x} - \mathbf{y})^T \nabla\varphi(\mathbf{y})$$



Euclidean Distance



Euclidean Distance

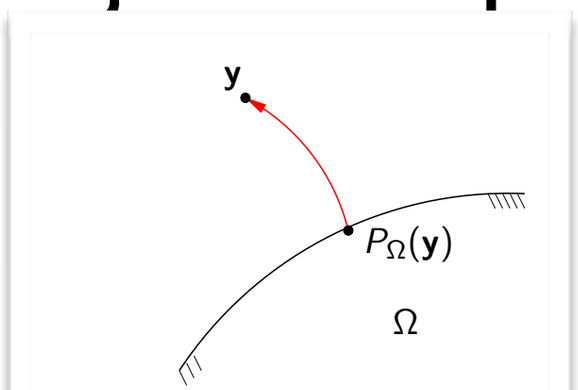
$$D_\phi(x, y) = \frac{1}{2}x^T x - \frac{1}{2}y^T y - y^T(x - y)$$

$$D_\phi(x, y) = \frac{1}{2}x^T x + \frac{1}{2}y^T y - y^T x$$

$$D_\phi(x, y) = \frac{1}{2}\|x - y\|_2^2$$

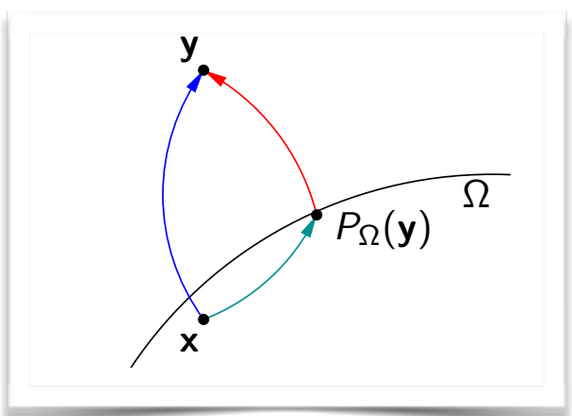
Generalized Projections

- Bregman Divergence leads to a generalized projection operation



$$P_\Omega(y) = \text{argmin}_{w \in \Omega} D_\phi(w, y)$$

- Generalized Pythagorean theorem



$$D_\phi(x, y) \geq D_\phi(x, P_\Omega(y)) + D_\phi(P_\Omega(y), y)$$

Operations preserving convexity of functions

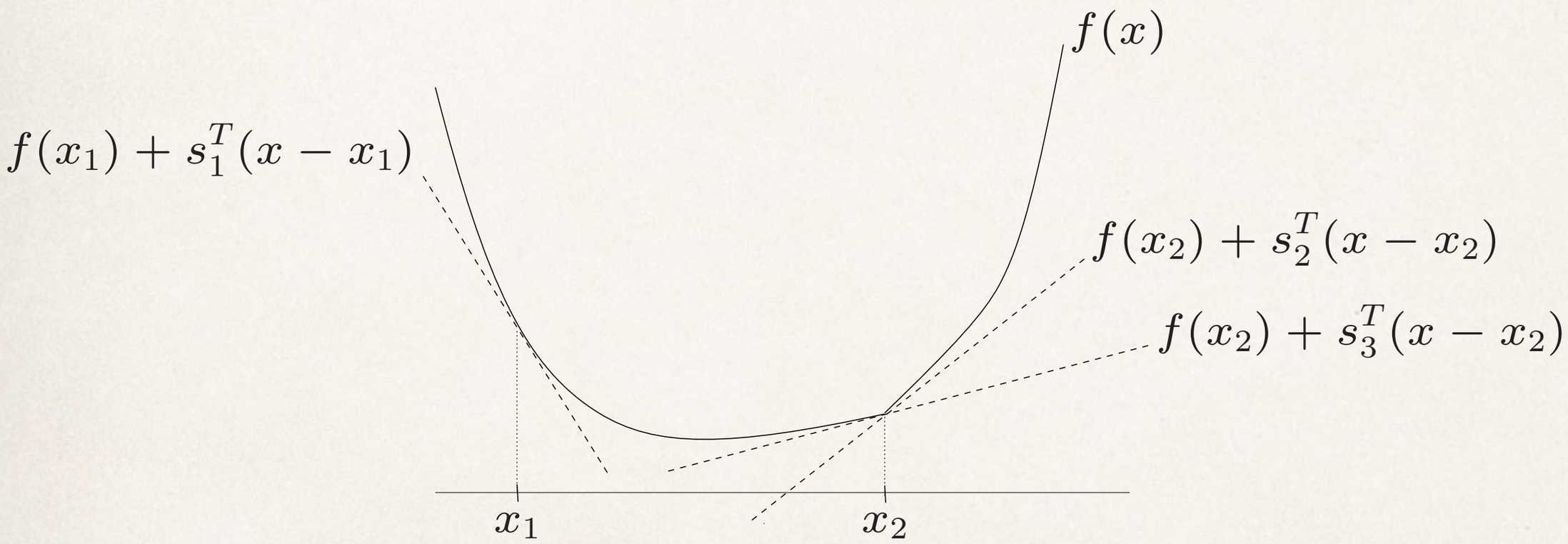
- Nonnegative weighted sums: 3f + 2g is convex if f and g are
- Composition with affine functions: f(Ax + b) is convex if f is
- Pointwise maximum and supremum: f(x) = max{g(x), h(x), q(x)}
  - Maximum eigenvalue of symmetric matrix $\lambda_{\max}(X) = \sup_{\|y\|_2=1} y^T X y$
- Minimization: distance to convex set $h(x) = \inf_{y \in C} \|x - y\|$
- Composition

Matrix completion revisited

- Recall that the matrix completion problem was reformulated as
  - Minimize $|X|_*$ such that $X_{i,j} = M_{i,j}$ (over observed samples)
  - Trace norm $|X|_*$ = sum of singular values of X
- Why is this a convex optimization problem?
  - Show maximum singular value can be written as

$$\sigma \max = \sup_{\|x\|=1, \|y\|=1} x' A y$$

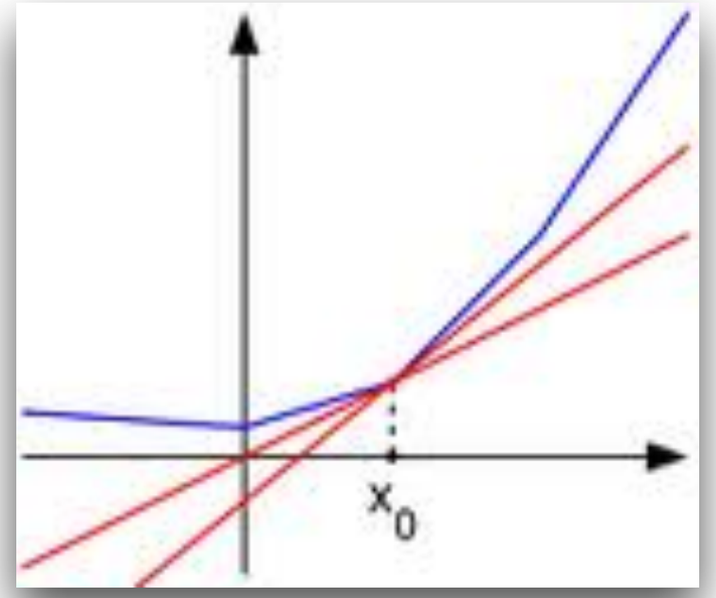
Subgradient of a function



Subgradients and Subdifferentials

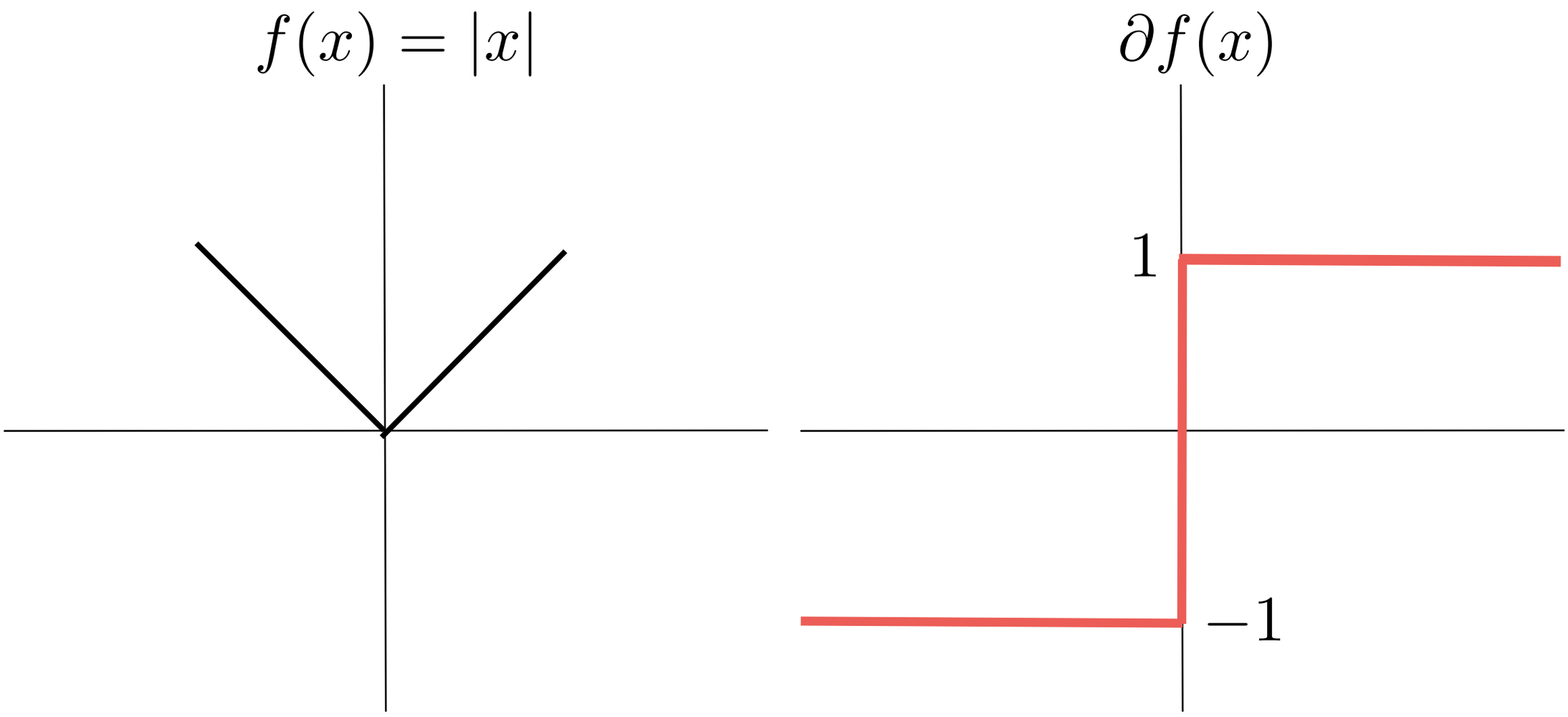
- Convex functions may not be differentiable
- Subgradient of a convex function:

$$f(y) \geq f(x) + \langle s, y - x \rangle$$



- Subdifferential: set of all subgradients

Subdifferential of |x|



$$f(x) = |x| \geq f(0) + v^t(x - 0) \Rightarrow |v| \leq 1$$

Euclidean Norm

$$f(x) = \|x\|_2$$

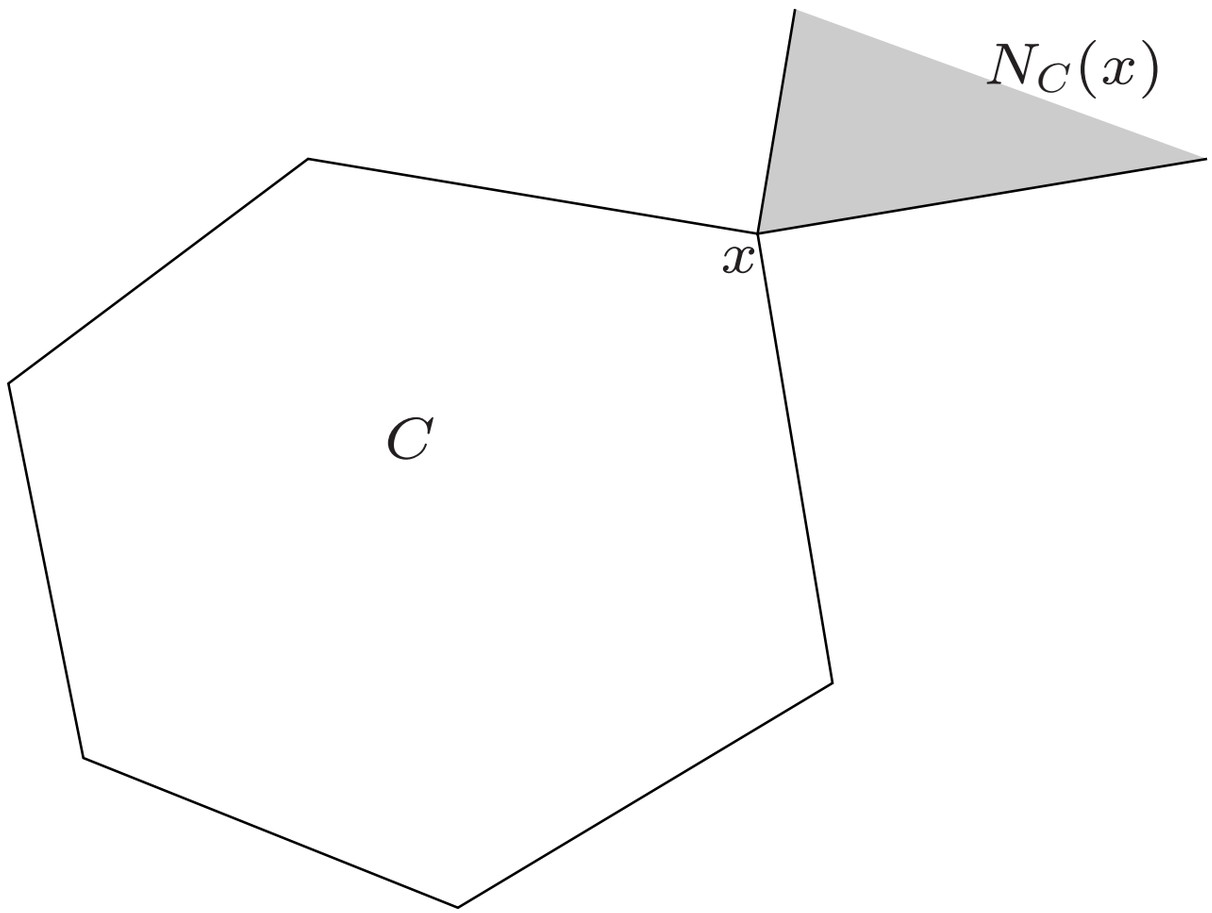
$$\partial f(x) = \frac{1}{\|x\|_2} x \text{ if } x \neq 0$$

$$\partial f(x) = \{g | \|g\|_2 \leq 1\} \text{ if } x = 0$$

Indicator Function

$$\partial I_C(x) = \{s \mid s^T(y-x) \leq 0, \text{ for all } y \in C\}$$

this is known as the *normal cone* to C at x (notation: $N_C(x)$)



Monotonicity of Subdifferentials

if $s \in \partial f(x)$ and $\hat{s} \in \partial f(\hat{x})$, then

$$(\hat{s} - s)^T (x - \hat{x}) \geq 0$$

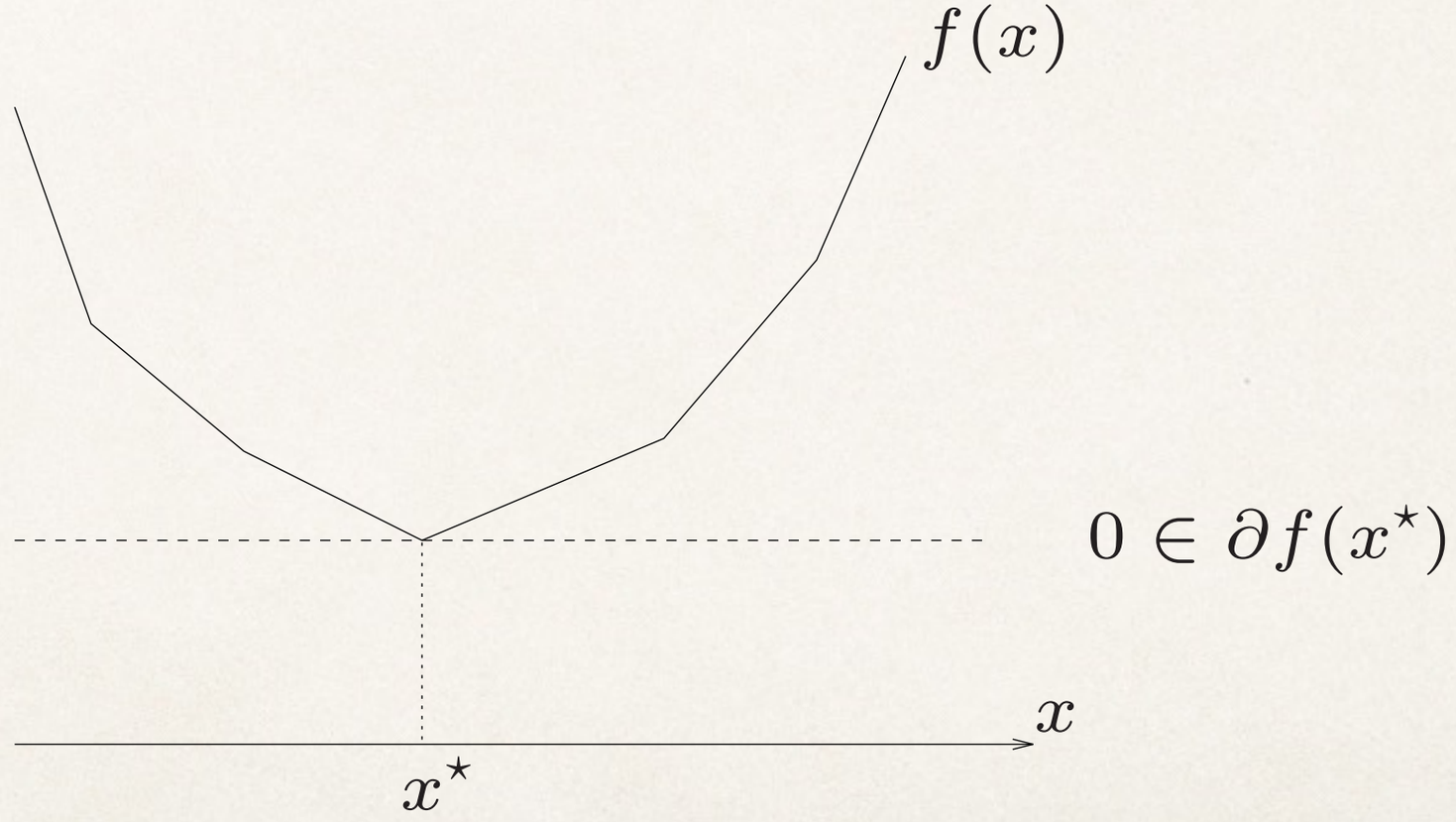
this property is called *monotonicity* of the (multivalued) mapping ∂f

proof: add left and righthand sides of the two inequalities

$$f(x) \geq f(\hat{x}) + \hat{s}^T (x - \hat{x})$$
$$f(\hat{x}) \geq f(x) + s^T (\hat{x} - \hat{x})$$

Optimality Conditions

$$0 \in \partial f(x^\star)$$



$$f(y) \geq f(x^*) + 0^T(y - x^*)$$

Proximal Mapping

- The proximal mapping of a convex function is defined as

$$\text{prox}_h(x) = \text{argmin}_u \left(h(u) + \frac{1}{2}\|u - x\|_2^2 \right)$$

- Examples:

$$h(x) = 0, \text{prox}_h(x) = x$$

$$h(x) = I_C(x), \text{prox}_h(x) = P_C(x) = \text{argmin}_{u \in C}\|u - x\|_2^2$$

Proximal Mappings

if h is convex and closed, then

$$\mathbf{prox}_h(x) = \underset{u}{\operatorname{argmin}} \left(h(u) + \frac{1}{2}\|u - x\|_2^2 \right)$$

exists and is unique for all x

subgradient characterization

from optimality conditions of minimization in the definition:

$$\begin{aligned} u = \mathbf{prox}_h(x) \quad &\iff \quad x - u \in \partial h(u) \\ &\iff \quad h(z) \geq h(u) + (x-u)^T(z-u) \quad \forall z \end{aligned}$$

Projections and Proximal Mapping

proximal mapping of indicator function I_C is Euclidean projection on C

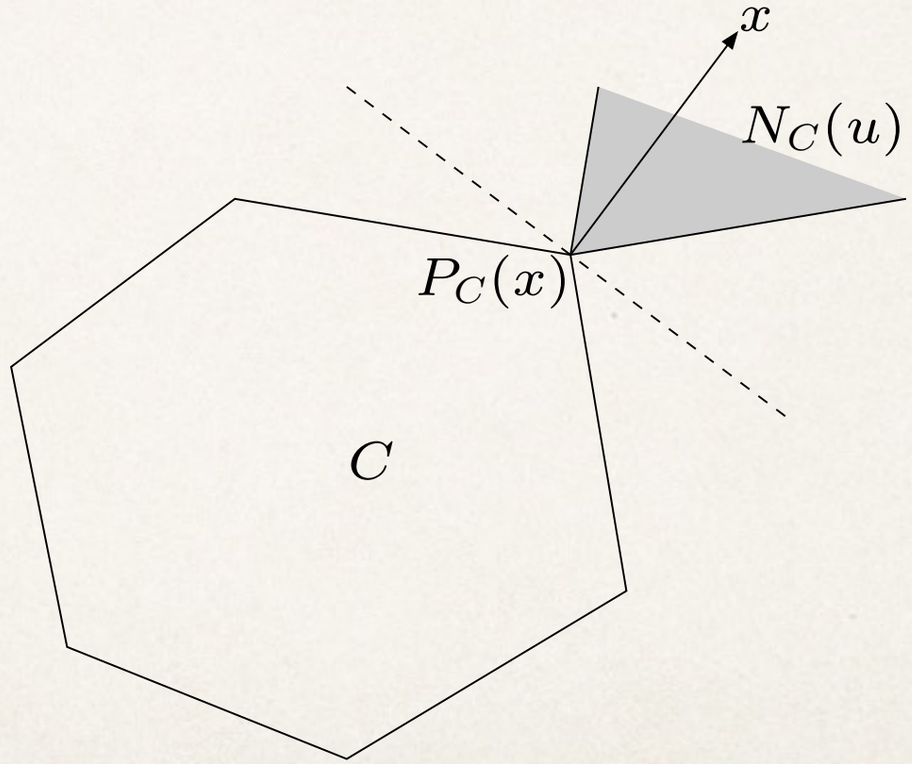
$$\mathbf{prox}_{I_C}(x) = \operatorname*{argmin}_{u \in C} \|u - x\|_2^2 = P_C(x)$$

subgradient characterization

$$u = P_C(x)$$

$$\Updownarrow$$

$$(x - u)^T (z - u) \leq 0 \quad \forall z \in C$$



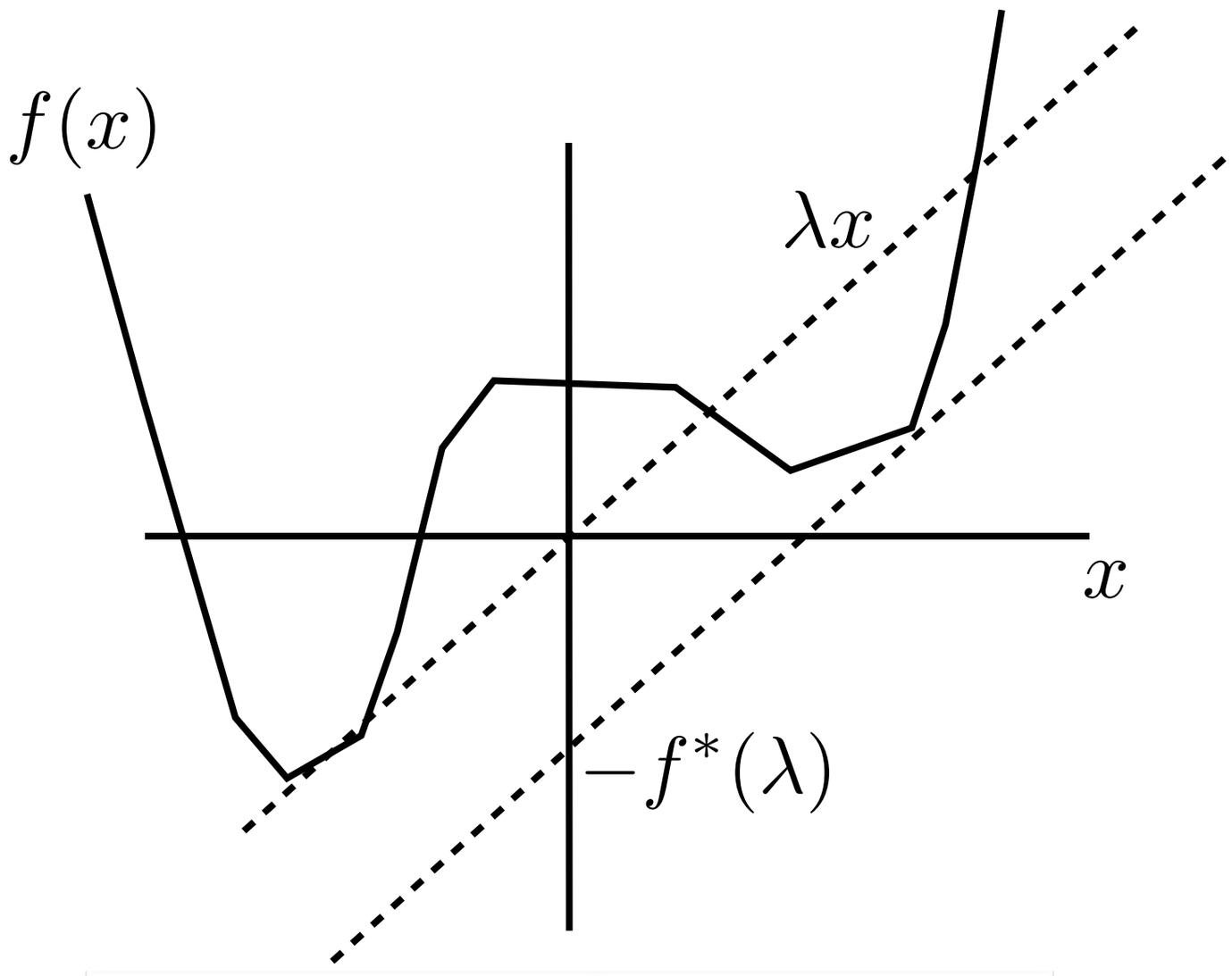
Conjugate Function

- A core concept in convex analysis is the notion of conjugate functions

$$f^*(\lambda) = \sup_x(\langle x, \lambda\rangle - f(x))$$

- The conjugate function is always convex, even if the original function is not
- Legendre Transform: conjugate of differentiable function $\lambda = \nabla f(x)$

Example

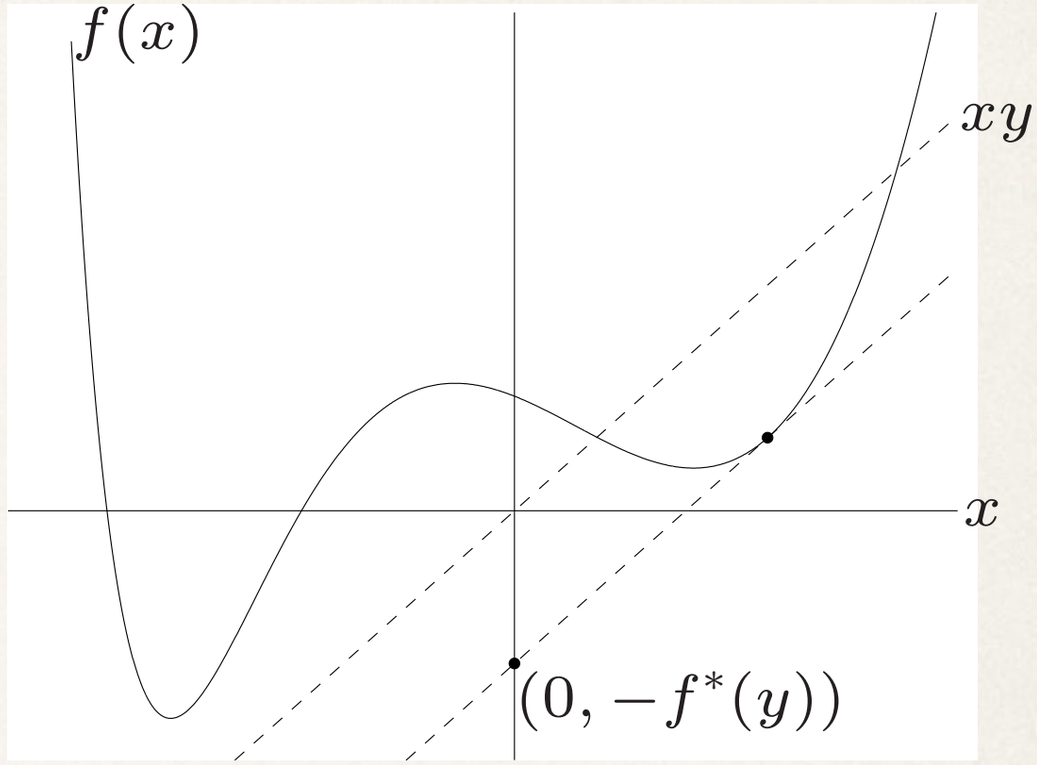


$$f^*(\lambda) = \sup_x(\langle x, \lambda \rangle - f(x))$$

The conjugate function

the **conjugate** of a function f is

$$f^*(y) = \sup_{x \in \mathbf{dom}\, f} (y^T x - f(x))$$



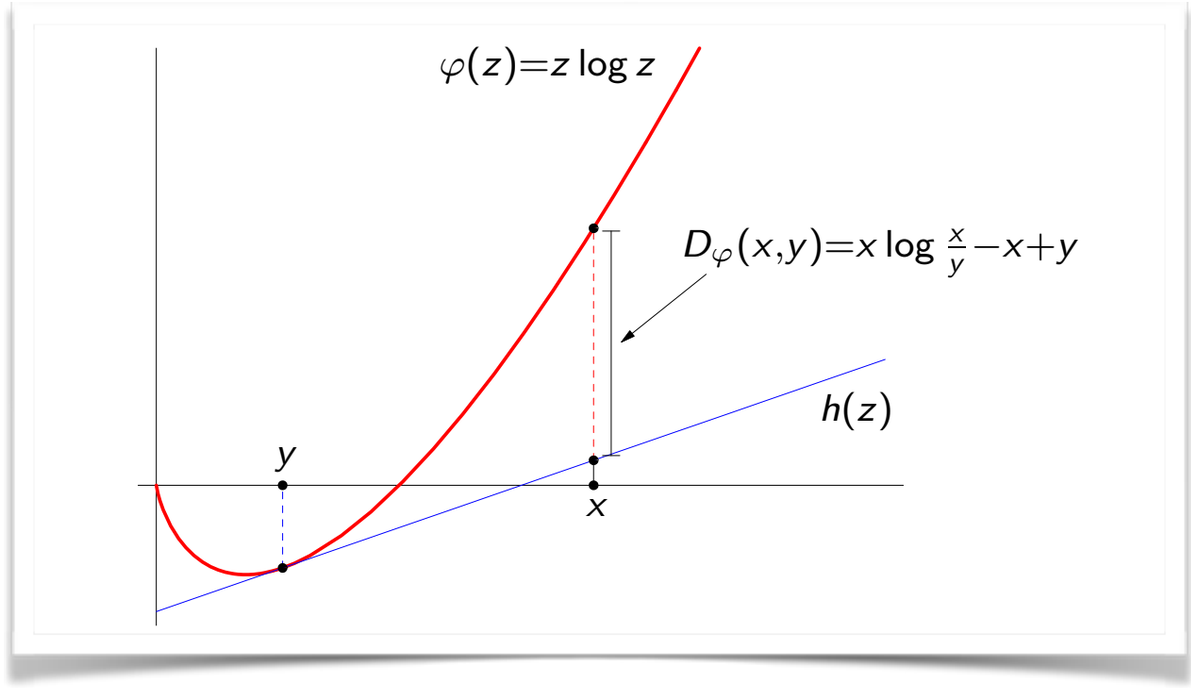
f^* is closed and convex (even if f is not)

Fenchel's inequality

$$f(x) + f^*(y) \geq x^T y \quad \forall x, y$$

Conjugate Functions

$$\phi^*(z) = \log \sum_i e^{z_i}$$



Examples

negative logarithm $f(x) = -\log x$

$$\begin{aligned} f^*(y) &= \sup_{x>0} (xy + \log x) \\ &= \begin{cases} -1 - \log(-y) & y < 0 \\ \infty & \text{otherwise} \end{cases} \end{aligned}$$

quadratic function $f(x) = (1/2)x^T Q x$ with $Q \in \mathbf{S}^n_{++}$

$$\begin{aligned} f^*(y) &= \sup_x (y^T x - (1/2)x^T Q x) \\ &= \frac{1}{2} y^T Q^{-1} y \end{aligned}$$

Examples

indicator function

$$I_C^*(y) = \sup\left(y^T x - I_C(x)\right) = \sup_{x \in C} y^T x$$

this is known as the *support function* of C

norm $f(x) = \|x\|$

$$f^*(y) = \sup_x \left(y^T x - \|x\|\right) = \begin{cases} 0 & \|y\|_* \leq 1 \\ +\infty & \text{otherwise} \end{cases}$$

$i.e.$, the indicator function of the norm ball of the *dual norm*

$$\|y\|_* = \sup_{\|x\| \leq 1} y^T x$$

Subgradient of Conjugate Function

$$f^*(y) = \sup_{x \in \mathbf{dom}\, f} (x^T y - f(x))$$

weak subgradient rule

if $\hat{x}$ maximizes $x^T\hat{y} - f(x)$ over $x \in \mathbf{dom}\, f$, then $\hat{x} \in \partial f^*(\hat{y})$

$$\begin{aligned} f^*(y) = \sup_{x \in \mathbf{dom}\, f} (x^T y - f(x)) &\geq \hat{x}^T y - f(\hat{x}) \\ &= \hat{x}^T\hat{y} - f(\hat{x}) + \hat{x}^T(y - \hat{y}) \\ &= f^*(\hat{y}) + \hat{x}^T(y - \hat{y}) \end{aligned}$$

Reading Assignment

- Read Chapters 2 and 3 of Boyd and Vandenberghe
- Work out all the examples given in this lecture
- Implement the Kacmarz algorithm