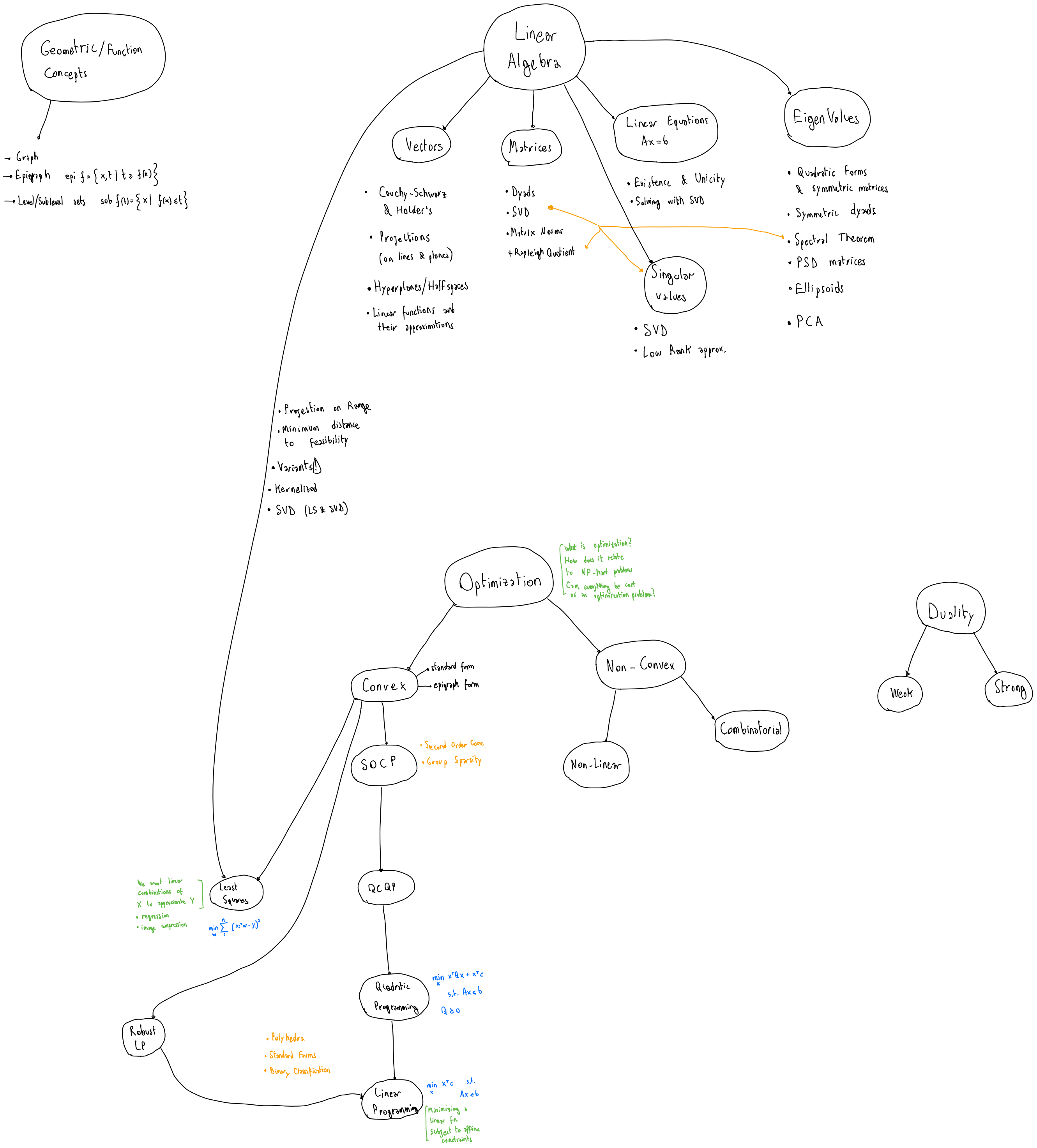

Geometric/Function Concepts
→ Graph
→ Epigraph epi $f = \{x,t \mid t \geq f(x)\}$
→ Level/Sublevel sets sub $f(t) = \{x \mid f(x) \leq t\}$
Linear Algebra
Vectors
• Cauchy-Schwarz & Holder's
• Projections (on lines & planes)
• Hyperplanes/Halfspaces
• Linear functions and their approximations
Matrices
• Dyads
• SVD
• Matrix Norms
+ Rayleigh Quotient
Linear Equations $Ax=b$
• Existence & Unicity
• Solving with SVD
Singular values
• SVD
• Low Rank approx.
Eigen Values
• Quadratic Forms & symmetric matrices
• Symmetric dyads
• Spectral Theorem
• PSD matrices
• Ellipsoids
• PCA
• Projection on Range
• Minimum distance to feasibility
• Variants ⚠
• Kernelized
• SVD (LS & SVD)
Optimization
What is optimization?
How does it relate to NP-hard problems
Can everything be cast as an optimization problem?
Convex
standard form
epigraph form
Non - Convex
Non-Linear
Combinatorial
Duality
Weak
Strong
SOCP
• Second Order Cone
• Group Sparsity
QCQP
Least Squares
We want linear combinations of X to approximate Y
• regression
• image compression
$\min_w \sum_i^n (x_i^T w - y_i)^2$
Quadratic Programming
$\min_x x^T Q x + x^T c$
s.t. $Ax \leq b$
$Q \geq 0$
Robust LP
• Polyhedra
• Standard Forms
• Binary Classification
Linear Programming
$\min_x x^T c$ s.t. $Ax \leq b$
Minimizing a linear fn. subject to affine constraints

Linear Algebra

Linear Independence

A set of vector are lin. ind. if no vector in the set can be expressed as a linear combination of the others:

$$\sum_i^n \alpha_i \vec{x}_i = 0 \Leftrightarrow \alpha_i = 0 \ \forall i$$

Subspace of $\mathbb{R}^n$

→ Subset of $\mathbb{R}^n$ closed under addition and scalar multiplication

Any subspace can be represented as the span of vectors $x_i \in \mathbb{R}^n$

Affine sets

any point in the shifted span of some vectors

$$A = \{x_0 + \sum_i^n \alpha_i \vec{x}_i\}$$

(translation of a subspace)

Basis

set of n indep. vectors → if $u_1, \dots, u_n$ form a basis for a subspace S, we can express any vector of S as a linear combination of $u_1, \dots, u_n$

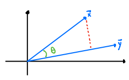
Standard Basis

A basis where all vectors are of the form $\begin{bmatrix} 1 \\ 0 \\ \vdots \\ \vdots \end{bmatrix}$

Inner Product

$x, y \in \mathbb{R}^n$

$$\langle x, y \rangle = x^T y = \sum_i^n x_i y_i$$



Angles

angle between vectors

$$\frac{x^T y}{\|x\|_2 \|y\|_2} = \cos\theta$$

$$x^T y = 0 \Leftrightarrow x \perp y$$

Cauchy-Schwarz

for two vectors $x, y \in \mathbb{R}^n$

we know that $|x^T y| \leq \|x\|_2 \|y\|_2$ with equality obtained when x collinear with y → by $\|x\|\|y\| \cos\theta = x^T y$ so $\theta = 0$ or $\theta = \pi$ for $\|x\|\|y\| = |x^T y|$

Holder's Inequality

$$\max_{\|x\|_\infty \leq 1} |x^T y| = \|x\|_0 \|y\|_\infty$$

$$\max_{\|x\|_0 \leq 1} |x^T y| = \|x\|_\infty \|y\|_0$$

Hyperplanes & Half-spaces

Hyperplanes are subspaces of dimension n-1 (defined by one single affine constraint)

∴ a hyperplane in $\mathbb{R}^n$ is defined as

$$H = \{x \mid a^T x = b \quad \text{for } a \in \mathbb{R}^n\}$$

if b=0 a is a vector NORMAL to hyperplane: $a \perp x \quad \forall x \in H$!

if $b \neq 0$ we have a translation by b in the direction of a of those points

Note that if x_0 is any other element in H

$$a^T x = a^T x_0 \Rightarrow a^T(x - x_0) = 0 \Rightarrow H = \{x \mid a^T(x - x_0) = 0\}$$

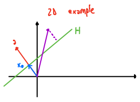
We know that at least one point in the plane will be along $\vec{a}$ as $a^T\left(\frac{a}{\|a\|_2^2} b\right) = b$

We usually say that $x_0 = \frac{b}{\|a\|_2^2} a$

Half spaces are just subsets (not subspaces) of $\mathbb{R}^n$ defined by an inequality constraint

$$H = \{x \mid a^T x \geq b\} = \{x \mid a^T(x - x_0) \geq 0\}$$

∴ for any point $x - x_0$ that forms an acute angle with $\vec{a}$, it will be on ⊕ side, and obtuse it will be on negative side



Subtracting x_0 off makes geometric sense as above.

Projections

Inner Products are closely tied to projection:

$$\text{proj}(x \text{ onto } y) = \text{proj}_y x = \frac{x^T y}{\|y\|_2} y$$

This idea generalize also to projections on lines.

Firstly, lines in $\mathbb{R}^n$ can be defined as $L = \{x_0 + ut \ \forall t \in \mathbb{R}\}$ where $x_0 \in \mathbb{R}^n$ (passes through x_0) an $v \in \mathbb{R}^n$ is direction

Projecting on a line

Projection of a point $x \in \mathbb{R}^n$ on a line can be cast as the following optimization problem:

finding t that defines the point with min. distance to x

$$\arg\min_t \|x_0 + tu - x\|_2 = \underset{t}{\text{argmin}} \|(x_0 - x) + tu\|_2^2 \Rightarrow \frac{\partial}{\partial t} = 2t\|u\|_2^2 + 2u^T(x_0 - x) = 0$$

$$\Rightarrow t^* = \frac{u^T(x - x_0)}{\|u\|_2^2} \Rightarrow x^* = x_0 + t^* u = x_0 + \frac{u^T(x - x_0)}{\|u\|_2^2} u$$

Min. distance to line: $\|\text{proj} - x\|_2 = \|x_0 + \frac{u^T(x - x_0)}{\|u\|_2^2} - x\|_2$ $\boxed{?}$

Projection on line: $a = x_0 + \frac{u^T(x - x_0)}{\|u\|_2^2} u$

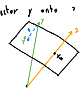
Projection on a hyperplane

Similarly to above we would like to cast the projection of a vector y onto a hyperplane $H = \{x \mid x^T a = b\}$ as an opt. problem. However we need to change form:

$$\underset{x}{\text{argmin}} \|y - ?$$

$$\Rightarrow \min_x \|x\| \quad \text{s.t.}$$

projection residual



as x_0 on hyperplane

$$\text{Length/Distance} \quad \frac{|(y - x_0)^T a|}{\|a\|_2} = \frac{|y^T a - x_0^T a|}{\|a\|_2} = \frac{|y^T a - b|}{\|a\|_2}$$

$$\text{Projection} \quad y^* = y - \frac{y^T a - b}{\|a\|_2} \frac{a}{\|a\|_2} = y - \frac{y^T a - b}{\|a\|_2^2} a$$

Linear & Affine Functions

A linear function preserves scaling and addition of the input argument

$$\begin{cases} f(x + y) = f(x) + f(y) \\ f(\alpha x) = a f(x) \end{cases}$$

Affine functions $f(x) = a^T x + b$ are technically not linear, but by augmenting the space to $\mathbb{R}^{n+1}$ one could view them as such $f(x) = [a \ b]\begin{bmatrix} x \\ 1 \end{bmatrix}$

Approximations to linear functions

Linear functions have very nice properties and so a lot of times we approximate non-linear functions with linear ones:

Close to x_0: $f(x) \approx f(x_0) + \nabla_x f(x_0)^T (x - x_0)$

Matrices

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

Can be viewed as

- → a Transformation
- → a list of vectors
- → n points in m dim space

$$A = [x_1 \ \cdots \ x_n]$$

or m points in n-dim space

$$A = \begin{bmatrix} x_1^T \\ \vdots \\ x_n^T \end{bmatrix}$$

Matrix-Vector Product

Ax

Linear combination of cols. of A

$x^T A$ → lin. comb. of rows of A $\overset{\shortparallel}{(A^T x)^T}$

Matrix-Matrix Product

$$AB = [Ab_1 \ \cdots \ \underset{\text{cols. B}}{Ab_n}]$$

$$= \begin{bmatrix} a_1^T B \\ \vdots \\ a_n^T B \end{bmatrix}$$

rows of A

Dyads

matrices of the form

$$A = uv^T$$

→ output Ax in direction of u

→ u is evector

u is customer choices

v is a "profile"

Linear Maps

$f: \mathbb{R}^n \to \mathbb{R}^m$, $A \in \mathbb{R}^{m \times n}$

$f(x) = Ax$ Affine → $f(x) = Ax + b$

A_{ij} is "influence" of x_j onto y_i ⚠

First order approximations

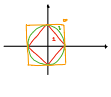
Norms

A measure of distance/size

$$\|x\|_2 = \sqrt{\sum x_i^2}$$

$$\|x\|_1 = \sum |x_i|$$

$$\|x\|_\infty = \max_i |x_i|$$



Matrix Norms

RMS Gain

$$\|A\|_F = \sqrt{\sum_i^n \|a_i\|_2^2} = \sqrt{\sum_i^n \sum_j^m A_{ij}^2} = \sqrt{Tr(A^T A)}$$

cols/rows

RMS root mean square (average size of how much a unit norm error would affect output)

$$= \sqrt{Tr(U\Sigma^2 U^T)} = \sqrt{Tr(\Sigma^2)} = \sqrt{\sum_i^n \sigma_i(A)^2} = \sqrt{\sum_i^n \lambda_i(A^T A)}$$

if A symmetric

$$= \sqrt{\sum_i^n |\lambda_i(A)|}$$

if $\lambda > 0$ (PSD)

$$= \sqrt{\sum_i \lambda_i}$$

Peak Gain (worst case approach (w bounded noise)

$$\|A\|_{LSV} = \max_v \|Av\|_2 \text{ s.t. } \|v\|_2 \leq 1$$

error bounded in terms of norm

$$= \sigma_1(A)$$

largest singular value!

Condition Number

: ratio btw largest and smallest singular value!

(needs invertability) $\overset{\shortparallel}{\text{sq. + full rank!}}$

$$K(A) = \frac{\sigma_1}{\sigma_n} = \sigma_1(A) \cdot \sigma_n(A) = \sigma_1(A) \cdot \sigma_1(A^{-1}) = \|A\|\|A^{-1}\|$$

measure of sensitivity of solution to changes in A

(. . .) 2.6.3 livebook

Linear Equations

$Ax = b$

We want to be able to determine

- if there is a solution
- uniqueness
- compute it
- (find orthonormal basis for set of solutions)

if there is a solution, the set of solutions will form a subspace

why?

Range / Rank / Null Space

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

$\text{Range}(A) = \{Ax \ \forall x \in \mathbb{R}^n\}$ = span of colums of A

if $y \neq R(A)$ then infeasible, no solution

$\text{Rank}(A) = \dim(R(A))$

"how many dimensions does the range span?"

$\text{Nullspace}(A) = \mathcal{N}(A) = \{x \mid Ax = 0\}$

if $\mathcal{N}(A) = \{\vec{0}\} \Rightarrow \text{Rank}(A) = m$

∴ $y \in R(A) \Rightarrow \exists$ solution [unique]

one can think of $\mathcal{N}$ as "ambiguity" in x:

$y = Ax = A(x + z) \quad \forall z \in \mathcal{N}(A)$

so if $\mathcal{N} = \{\vec{0}\}$ there are ∞ solutions as $\mathcal{N}$ is subspace

with SVD: $Ax = b \Rightarrow U\Sigma V^T x = b$

$$\Rightarrow \Sigma V^T x = U^T b \Rightarrow \Sigma \tilde{x} = \tilde{y}$$

$\tilde{x} = V^T x$

$V\tilde{x} = x$

$$\Rightarrow \begin{cases} \sigma_i \tilde{x}_i = \tilde{y}_i & \forall i = 1, \dots, r \\ 0 = \tilde{y}_i & \forall i = r+1, \dots, n \end{cases}$$

∴ ∃ solution ⇔ all $\tilde{y}_i = 0 \quad \forall i = r+1, \dots, n$

if not b not in range of A

$$\tilde{x}_i = \begin{cases} \frac{\tilde{y}_i}{\sigma_i} = \frac{u_i^T b}{\sigma_i} & i = 1 \dots r \\ \text{free} & i = r+1 \dots n \end{cases}$$

go to nullspace!

$$\Rightarrow x = V\tilde{x}$$

Least Squares

It turns out that a lot of problems can be expressed as LS problems: trying to express a good approximation of a vector y with a linear combination of colums of A.

There are various interpretations:

Projection on the range

We know that we want to express y as a lin. comb. of the columns of A, and minimize the residuals $y - Ax$ (some kind of reconstruction error).

We also know that we can take into account any lin. comb. by considering the Range of A.

In this context minimizing the residuals is equivalent to finding the point in R(A) with min distance to y: this is equivalent to finding the projection of y on A (and y lives outside the m-dim subspace: i.e. there is no solution with 0 error)

$A^T(Ax - y) = 0$ (from opt. conditions)

residuals orthogonal to row space

We usually have $y \in \mathbb{R}^n$ and $A \in \mathbb{R}^{n \times m}$ where $m < n$

If $n = m$, there is a solution with 0 error ∀y ⇔ A is full rank

If $m > n$, there is a 0-error solution ∀y if A full row rank?

what does it mean for subspace

m-subspace can span a smaller space that doesn't include y → can still do it? → yes! but notice that that would mean that your features (cols) are lin. dependent → therefore there are infinitely many ways to express the projection!

min norm solution ?

Minimum Distance to Feasibility

Usually $Ax = y$ is not feasible so in a way we can interpret the problem as finding "the min distance to feasibility"

⇒ finding the distance from y to R(A) → area that would give feasibility!

Variants

Linearly Constrained Least Squares

$$\min_x \|Ax - y\| \quad \text{s.t.} \quad Cx = d$$

Constraining our weights to be in a specific subspace of $\mathbb{R}^m$ ($A \in \mathbb{R}^{n \times m}$, $x \in \mathbb{R}^m$)

interpretation: $Cx = d$

- could have unique sol. → then subspace in which x is contrained to be is just a point
- infinitely many sol. → set of possible solutions would be $x = z + x_0$ where $z \in \mathcal{N}(C)$ and $Cx_0 = d$ → we have constrained our solution for LS to a nullity(C)-dimensional subspace!
- infeasible (no sol.)

⇕

$X = \{x_0 + Nz \ \forall z\}$ where N spans $\mathcal{N}(C)$

feasible X

can rewrite in terms of free variable z without constraint

$$\Rightarrow \min_z \|A(x_0 + Nz) - y\|_2 = \min_z \|ANz + Ax_0 - y\|_2$$

$$= \min_z \|\tilde{A}z + \tilde{y}\|_2$$

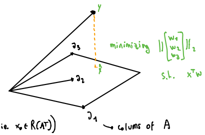
Minimum Norm Solution

When collinear features we can still find the minimum norm solution!

$$\min_x \|x\|_2 \quad \text{s.t.} \quad Ax = y$$

(implicitly assuming y is in colspace of A)

$$\Rightarrow \min_x \max_\lambda \|x\|_2 + \lambda(Ax - y)$$



By FTLA, any solution of $Ax = y$ will be of the form $x = x_0 + z$ where $z \in \mathcal{N}(A)$ and $Ax_0 = y$ (i.e. $x_0 \in R(A^T)$)

$= x_0 + Nz \ \forall z$ where $R(N) = \mathcal{N}(A)$

if a vector in $\mathbb{R}^n$ is not in the null space, it's in the range of transpose by FTLA

To be min. norm, we know that $z = 0$, i.e. $x^* \perp \mathcal{N}(A) \Rightarrow x^* \in R(A^T)$

∴ $x^* = A^T \xi$

$$\Rightarrow \min_\xi \|A^T \xi\|_2 \quad \text{s.t.} \quad AA^T \xi = y$$

$$\Downarrow$$

$$\xi^* = (AA^T)^{-1} y$$

$A = [\quad\quad]$

$AA^T = [\quad]$ full rank if A full row rank

$$\Rightarrow x^* = A^T \xi^* = A^T(AA^{-1})y$$

Regularized Least Squares

If A not full column rank, closed form cannot be applied

$$\min_x \|Ax - y\|_2^2 + \lambda\|x\|_2^2$$

$$= \min_x x^T(A^T A + \lambda I)x - y^T y - 2y^T x$$

tricky transition

$$= \min_x \left\|\begin{bmatrix} A \\ \sqrt{\lambda} I \end{bmatrix} x - \begin{bmatrix} I \\ 0 \end{bmatrix} y\right\|_2^2$$

the presence of this identity ensures full column rank!

$$= \min_x \|\tilde{A}x - \tilde{Y}\|_2^2$$

$A = [\quad]$ lin dep or $= [\quad]$

Can't we still project on the line? (throw away lin. dep columns?)

A	LS	Geom	Idea
Full row rank ⊕ $y \in R(A)$	min norm solution	cols of A are dependent, infinite sols. want to find smallest	FTLA
not full col rank	Ridge	Biasing result with 0 observations	Put back into LS form after adding regularization (ensures full rank)

Weighted Least Squares

$$\min_x \|Ax - y\|_2^2 + x^T W x$$

$$x^* = (A^T A + W)^{-1} A^T y$$

$$\tilde{\approx} \min \|W(Ax - y)\|_2^2$$

residuals not given same importance

Linear Algebra - Evals / S.Vals

Quadratic Functions

We call a function "quadratic" if it can be expressed as:

$$q(x) = x^T Ax + 2b^T x + c = \sum_i^n \sum_j^n A_{ij} x_i x_j + 2\sum_i^n b_i x_i + c$$

note that to be quadratic A has to be square (and also symmetric by definition of Q.F.)

Link with Symmetric Matrices

Any quadratic fn can be written as $\begin{pmatrix} x \\ 1 \end{pmatrix}^T \begin{pmatrix} A & b \\ b & c \end{pmatrix} \begin{pmatrix} x \\ 1 \end{pmatrix}$ where $A \in S^n$

in quadratic form (no constants or lin. terms)

Spectral Theorem

We can decompose any symmetric matrix A as:

$$A = U \Lambda U^T$$

U is orthonormal matrix of eigenvectors, Λ is diagonal of eigenvalues!

Second order approx. of non-quadr.

We saw that we could approximate non-lin. fns with first order lin. approx.

Now we can extend to second order lin. approx. for non quadratic fns:

$$f(x) \approx f(x_0) + \nabla f(x_0)^T (x - x_0) + \frac{1}{2}(x-x_0)^T \underset{\nabla^2 f(x_0)}{\underset{\parallel}{H(x_0)}} (x - x_0)$$

Symmetric Dyads

Matrices of the form: $A = uu^T$ are called symmetric dyads.

They correspond to quadratic terms that are squared linear terms!

$$x^T Ax = x^T uu^T x = (u^T x)^2$$

Rayleigh Quotients

Given a symm. mtx we can express smallest and largest eigen values as:

$$\lambda_{min}(A) = \min_{\|x\| \leq 1} x^T Ax = \min_{\|\tilde{x}\| \leq 1} \tilde{x}^T \Lambda \tilde{x}$$

→ Interpretation: min & max evals are a measure of the range of the quadratic form $x^T Ax$ over the euclidean norm ball !

connecting back to LSV norm $\|A\| = \max_{\|x\| \leq 1} \|Ax\|_2 \Rightarrow \|A\|^2 = \max_{\|x\|_2 \leq 1} x^T A^T Ax = \lambda_{max}(A^T A)$

$$\Rightarrow \|A\| = \sqrt{\lambda_{max}(A^T A)}$$

for symm. matrices LSV is sqrt of L eval.

PSD matrices

A matrix is PSD if $x^T Ax \geq 0 \ \forall x \Leftrightarrow \sum_i^n x_i^2 \lambda_i \geq 0 \Leftrightarrow \lambda_i \geq 0 \ \forall i$

In the case of symmetric dyads this is even easier to see $x^T B^T Bx = \|Bx\|_2^2 \geq 0 \ \forall x.$

If A PSD → AA^T & $A^T A$ PSD

If full column rank $A^T A$ PSD

If full row rank AA^T PSD

Cholesky Decomposition & Square Root

We can generalize the notion of square root to PSD matrices:

$$A = U\Lambda U^T = U\Lambda^{\frac{1}{2}} \Lambda^{\frac{1}{2}} U^T = \| \Lambda^{\frac{1}{2}} U^T \|_2^2 = A^{\frac{1}{2}T} A^{\frac{1}{2}} \quad \text{where } A^{\frac{1}{2}} = \Lambda^{\frac{1}{2}} U^T$$

Any PSD matrix can be written as $A = LL^T$

Decomposition is not unique → if PD, L can be written as lower triangular

Cholesky Decomposition

Ellipsoids

There is a strong connection between PSD matrices and ellipsoids. In particular, the range of quadratic forms (PSD) over the unit norm Euclidean ball form ellipsoids

We can see this in various ways. One is to simply consider an ellipse

$$E = \{(x-x_0)^T A^{-1} (x-x_0) \leq 1\}$$

$$= \{(x-x_0)^T (LL^T)^{-1} (x-x_0) \leq 1\}$$

$$= \{\|L^{-1}(x-x_0)\|_2^2 \leq 1\}$$

$$= \{x \text{ s.t. } \|L^{-1}(x-x_0)\|_2 \leq 1\}$$

$$= \{x_0 + Lz \text{ s.t. } \|z\|_2 \leq 1\}$$

translation; transformed unit ball (stretched & rotated)

L transf matrix, A^{-1} quadratic matrix } $LL^T = A$

smart transition:

$$\boxed{\begin{aligned} z &= L^{-1}(x - x_0) \\ \Rightarrow Lz &= x - x_0 \\ \Rightarrow x &= x_0 + Lz \end{aligned}}$$

Now, considering the set E in it's first form:

$$E = \{(x-x_0)^T A^{-1} (x-x_0) \leq 1\}$$

$$= \{(x-x_0)^T (U\Lambda U^T)^{-1} (x-x_0) \leq 1\}$$

$$= \{(x-x_0)^T U\Lambda^{-1} U^T (x-x_0) \leq 1\}$$

$$= \{\|\Lambda^{-\frac{1}{2}} U^T (x-x_0)\|_2^2 \leq 1 \ \forall x\}$$

$$= \{\|\Lambda^{-\frac{1}{2}} \tilde{x}\|_2^2 \leq 1\} \quad \text{where } \tilde{x} = U^T(x-x_0)$$

$$= \left\{\sum_i^n \frac{\tilde{x}_i^2}{\lambda_i} \leq 1\right\}$$

$$= \left\{\sum_i^n \bar{x}_i^2 = \bar{x}^T \bar{x} \leq 1\right\} \quad \text{for } \bar{x}_i := \frac{\tilde{x}_i}{\sqrt{\lambda_i}}$$

ball in $\bar{x}$ bar space in $\tilde{x}_i = \sqrt{\lambda_i}\,\bar{x}_i$ → axis

$\therefore \|\tilde{x}_i\| = \sqrt{\lambda_i}\,\|\bar{x}_i\| \leq \sqrt{\lambda_i} \cdot 1 = \sqrt{\lambda_i}$

$\therefore \|U^T(x-x_0)\| = \|\tilde{x}_i\| \leq \sqrt{\lambda_i}$

⇒ axis lengths are $\sqrt{\lambda_i} \ \forall i$ and directions will be u_i

why? $\tilde{x}_i = \sqrt{\lambda_i}$ will saturate the inequ.

$\tilde{x} = U^T(x-x_0)$

$\Rightarrow U\tilde{x} + x_0 = x$

projecting $(x-x_0)$ on $u_1, \dots u_n$ so these will be directions

$$\begin{bmatrix} \sqrt{\lambda_1} \\ 0 \\ \vdots \\ 0 \end{bmatrix} = U^T(x-x_0) \qquad x - x_0 \perp u_{2,\dots n}$$

Relies on A PD, if not → degenerate ellipse:

Types of degenerate? { slab:

Exercise:

$$\left\|\begin{bmatrix} 2 \\ 1 \end{bmatrix}\right\| = \sqrt{2^2+1^2} = \sqrt{5}$$

$$A = \frac{1}{\sqrt{10}} \begin{pmatrix} 2 \\ 1 \end{pmatrix} (1 \ -1) + \frac{2}{\sqrt{10}} \begin{pmatrix} -1 \\ 2 \end{pmatrix} (1 \ \ 1) = \overbrace{\frac{1}{\sqrt{10}} \sqrt{5}\sqrt{2}}^{1}\, u_1 v_1^T + \overbrace{\frac{2}{\sqrt{10}} \sqrt{5}\sqrt{2}}^{2}\, u_2 v_2^T$$

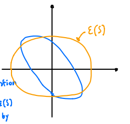
a) what is SVD of A? → $= \sigma_2 u_2 v_2^T + \sigma_1 u_1 v_1^T$

$$= [u_1 \ u_2] \begin{bmatrix} \sigma_1 & \\ & \sigma_2 \end{bmatrix} [v_1 \ v_2]^T$$

$$= \frac{1}{\sqrt{5}} \begin{bmatrix} -1 & 2 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} 2 & \\ & 1 \end{bmatrix} \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$$

order them!

b) Find the semi-axis lengths and principal axes of the ellipsoid



$$\mathcal{E}(A) = \{Ax : x \in \mathbb{R}^2, \ \|x\|_2 \leq 1\}$$

$$= \{U\Sigma V^T x : x \in \mathbb{R}^2 \ \|x\|_2 \leq 1\}$$

$$= \{U\Sigma\bar{x} : \bar{x} \in \mathbb{R}^2 \ \|\bar{x}\|_2 \leq 1\}$$

smaller set, one less rotation

$$= \{U\bar{y} : \bar{y} \in \mathcal{E}(S)\}$$

lets just look at ε(S) and analyse $U\bar{y}$ by using $U\bar{y}$ later!

$$\mathcal{E}(S) = \{\Sigma\bar{x} \ \text{ s.t. } \|\bar{x}\|_2 \leq 1\}$$

finally a nice form!

$$= \{\underset{\begin{bmatrix} \sigma_1\bar{x}_1 \\ \sigma_2\bar{x}_2 \end{bmatrix}}{\overset{}{\sigma_1 \bar{x}_1 e_1 + \sigma_2 \bar{x}_2 e_2}} \quad \sqrt{\bar{x}_1^2 + \bar{x}_2^2} \leq 1\}$$

V doesn't end up affecting our ellipse! this is because it's second!

$$\Rightarrow \mathcal{E}(A) = \{\sigma_1 \bar{x}_1 u_1 + \sigma_2 \bar{x}_2 u_2 \quad \bar{x}_1^2 + \bar{x}_2^2 \leq 1\}$$

As $\bar{x}_1^2 + \bar{x}_2^2 \leq 1$ we know: semi-axis lengths will be given by: σ_1, σ_2

directions u_1, u_2

Assume input in unit ball:

$$x \in B^n \qquad B^n = \{z \in \mathbb{R}^n : \|z\|_2 \leq 1\} \leftarrow \text{Ball}$$

$$\Rightarrow \text{output} \qquad \mathcal{E}_y = \{y = Ax \quad \forall x \in B^n\} \leftarrow \text{Ball output}$$

$$\Rightarrow \text{input} \qquad \mathcal{E}_x = \{x \mid Ax \in B^m\} \leftarrow \text{input to ball}$$

$$\Downarrow$$

$$\mathcal{E}_x = \{x \mid Ax \in \{Ax \mid \|Ax\|_2 \leq 1\}\}$$

$$= \{x \mid x^T A^T Ax \leq 1\}$$

ellipsoid unbounded in directions x in nullspace of A (degenerate)

PCA

We might want to find the directions that explain most of the variance in our data

For this we use PCA $X = \begin{bmatrix} | & & | \\ x_1 & \dots & x_n \\ | & & | \end{bmatrix}$ datapoints! ⇒ max Variance of $u^T X = [u^T x_1 \ \dots \ u^T x_n]$. We don't have variance but can estimate as $\text{Var}(x) = \frac{1}{n}\sum_i^n (x_i - \bar{x})^2$

$$\Rightarrow \max_u \hat{\text{Var}}(u^T X) = \max_u \frac{1}{n} \sum_i^n (u^T x_i - u^T \bar{x})^2 = \max_u \sum_i^n (u^T(x_i - \bar{x}))^2$$

$$= \max_u \left\| \begin{bmatrix} (x_1-\bar{x})^T u \\ \vdots \\ (x_n - \bar{x})^T u \end{bmatrix} \right\|_2^2 = \max_u u^T \sum_i^n (x_i - \bar{x})(x_i - \bar{x})^T u = \max_u u^T \hat{\Sigma} u$$

$$[(x_i-\bar{x})^T u \ \dots \ (x_i-\bar{x})^T u] \begin{bmatrix} (x_i-\bar{x})^T u \\ \\ (x_i - \bar{x})^T u \end{bmatrix}$$

$$u^T [(x_i - \bar{x}) \dots (x_i - \bar{x})] \begin{bmatrix} (x_i-\bar{x})^T \\ \vdots \\ (x_i-\bar{x})^T \end{bmatrix} u$$

$\therefore$ if $\bar{x} = 0$: $\hat{\Sigma} = \sum_i^n x_i x_i^T = XX^T = \begin{bmatrix} | & & | \\ x_1 & \dots & x_n \\ | & & | \end{bmatrix} \begin{bmatrix} - x_1^T - \\ - x_n^T - \end{bmatrix}$

nicer representation if assume $\bar{x} = 0$

$$\text{Cov Mtx} = \begin{bmatrix} x_1^T x_1 & \\ & x_n^T x_n \end{bmatrix}_{d \times d}$$

features!

$$= \max_u u^T U \Lambda U^T u \qquad \tilde{u} = U^T u, \ u = U\tilde{u}$$

$$= \max_u \tilde{u}^T \Lambda \tilde{u}$$

$$= \max_u \sum \tilde{u}_i^2 \lambda_i \qquad \text{if } \|u\|_2 = 1$$

we will pick direction corresponding to max eval λ_1

in $\tilde{u}$ space it's e_1 $(\tilde{u}^* = e_1)$

$\Rightarrow u = Ue_1 = u_1$ ✓

Explained Variance

We can estimate the amount of variance explained through PCA by considering the trace of the covariance matrix; note that $\text{Tr}(\Sigma)$ = Total variance = sum of var for each feature

Also,

$$\Rightarrow \text{Tr}(\Sigma) = \text{Tr}(U\Lambda U^T) = \text{Tr}(U^T U\Lambda) = \text{Tr}(\Lambda) = \sum_i^n \lambda_i$$

$$\therefore \frac{\text{var explained}}{\text{total variance}} = \frac{\text{Tr}(P\Sigma P^T)}{\text{Tr}(\Sigma)} = \frac{\lambda_1 + \dots + \lambda_p}{\lambda_1 + \dots + \lambda_n} \quad \text{where } P = [u_1, \dots, u_p]$$

top p evects

K-PCA

SVD

By the SVD, we know that any matrix can be expressed as:

$$A = U\Sigma V^T \qquad A \in \mathbb{R}^{m \times n}$$

$$= \sum_i^n \sigma_i u_i v_i^T \qquad \text{where } u_i^T A = \sigma_i v_i^T \Leftrightarrow A^T u_i = \sigma_i v_i$$

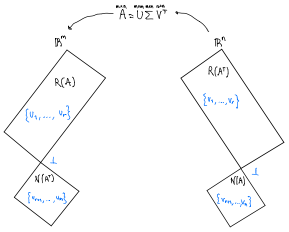
$$= \sum_i^r \sigma_i u_i v_i^T \qquad Av_i = \sigma_i u_i$$

$$= U_r \Sigma_r V_r^T$$

$$Ax = \sum \sigma_i u_i v_i^T x$$

$$= \sum (\sigma_i v_i^T x) u_i$$

$$Ax \in \text{span}\{u_1, \dots, u_r\}$$



A full rank if r = n

SVD vs EVD

$$A = U\Sigma V^T \qquad \Rightarrow A^T A = V^T \Sigma^2 V \text{ svd} \Rightarrow \sigma_i(A)^2 = \lambda_i(A^T A) = \lambda_i(AA^T) \Rightarrow \sigma_i(A) = \sqrt{\lambda_i(A^T A)}$$

(s.vals of A)²

$$= U\Lambda U^T \qquad A^T A = U^T \Lambda U \text{ evd}$$

evals of $A^T A$

$$= U^T \Lambda^2 U \ (1)$$

evals of A! (if A symmetric!)

if symmetric (1) $= \sqrt{\lambda_i^2(A)} = |\lambda_i(A)|$

∴ if symm:	otherwise:
$\sigma_i(A) = \lvert\lambda_i(A)\rvert$	$\sigma_i^2(A) = \lambda_i(A^T A)$

→ no relationship with $\lambda(A)$!

$\Rightarrow Au_i = \sigma_i u_i$ if $\lambda_i(A) > 0 \ \forall i$

FTLA & SVD

$\mathbb{R}^n = N(A) + R(A^T)$ so any vector in $\mathbb{R}^n$ $x = y + z$ where $y \in N(A)$ and $z \in R(A^T)$

$$\Rightarrow y^T z = 0$$

$\therefore$ a vector $x \perp N(A) \Leftrightarrow x \in R(A^T) \Rightarrow$ $\boxed{x^T y = 0 \text{ wherever } Ay = 0 \Leftrightarrow \exists \lambda \text{ s.t. } x = A^T \lambda}$

2.6.4

Low-Rank Approximations

We might want to approximate a matrix using less rank

$$\min_{\hat{X}} \|X - \hat{X}\|_F \qquad \text{s.t. } \text{rank}(\hat{X}) = K$$

Best K-rank approx: $\hat{X} = \sum_i^K \sigma_i u_i v_i^T = \underset{n\times K}{U_K} \underset{K\times K}{\Sigma_K} \underset{K\times m}{V_K^T}$ ← ordered!

Note $\|X - \hat{X}\|_F = \sqrt{\|X - \hat{X}\|_F^2} = \sqrt{\|U(\Sigma - \Sigma_K)V^T\|_F^2}$

$$= \sqrt{\|\Sigma - \Sigma_K\|_F^2} = \sqrt{\sum_{K+1}^n \sigma_i}$$

error!

PCA is basically low rank approx. → PCA also provides evects (as directions) so more useful!

(2.6.5 rows?)

$$\text{explained variance: } \frac{\|\hat{A}_K\|_F^2}{\|A\|_F^2} = \frac{\sigma_1^2 + \dots + \sigma_K^2}{\sigma_1^2 + \dots + \sigma_r^2}$$

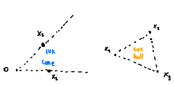
Convex Sets

A set is said to be convex if all convex combinations of its items are also in the set

C cvx set $\Rightarrow \left[x_1, \dots, x_n \in C \Leftrightarrow \sum_i^n \lambda_i x_i \in C \quad \text{s.t.} \sum_i^n \lambda_i = 1 \right]$

lines, planes etc. are convex
↳ contain the entire line i.e. any linear combination! ($\sum \lambda_i = 1$ not necessary)

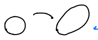
C convex cone $\Leftrightarrow \begin{cases} C \text{ convex} \\ x \in C \Leftrightarrow \alpha x \in C \quad \forall \alpha \geq 0 \end{cases}$



Operations that preserve convexity

- Intersection (infinite)
- Affine transformation: $f(C) = \{ f(x) \ \forall x \in C \}$ will be convex

$\left(f \text{ affine so } f(x) = Ax + b \overset{?}{=} U \Sigma V^T x + b \right)$ ← rotation / rescaling / translation



→ projection of a cvx set on a subspace / affine set is convex
↳ UNCLEAR WHY

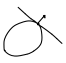
Projection of x on C $\quad \pi_C(x) = \underset{y}{\operatorname{argmin}} \|x - y\| \quad \text{s.t. } y \in C$

Supporting/Separating Hyperplane

Convex sets that do not intersect can be separated by a hyperplane

Supporting hyperplane

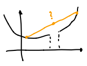
If C cvx, for any x_0 on boundary, $\exists$ a supporting hyperplane to C at x_0, i.e. $\exists a \in \mathbb{R}^n$ s.t. $(x - x_0)^T a \leq 0$



Separating Hyperplane

For any two convex sets with no intersection, $\exists a$ s.t.

$a^T x \leq b \quad \forall x \in C_1 \quad$ and $\quad a^T x \geq b \quad \forall x \in C_2$



dom ≠ convex

Convex Functions

We say that a function is convex if $\begin{cases} \text{the dom } f \text{ is convex} \\ f(\sum \lambda_i x_i) \leq \sum \lambda_i f(x_i) \end{cases}$

equivalent definitions:
- basic f property
- epigraph
- SOConstraints
- tangent

$\Updownarrow$

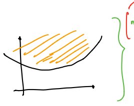
$\text{epi} f$ is convex set (i.e. $\text{epi} f = \{ (x,t) : f(x) \leq t \ \forall x, t \}$)

$\Updownarrow$

Second Order Condition: $\frac{\partial}{\partial^2 x} f \geq 0 \Leftrightarrow \nabla_x^2 f(x) \succeq 0$ PSD Hessian

$\Updownarrow$

Any point on any tangent to f will be below f at that point $\Rightarrow f(y) \geq f(x) + \nabla f(x)^T (y - x) \quad \forall x, y$



WRONG: (comes from sublevel sets!) → sublevel sets of cvx fns are convex sets!

[note that this enables us to relate constraints for feasibility /w convex sets

$f_i(x) \leq 0 \quad f$ convex

$\Rightarrow S = \{ x \mid f_i(x) \leq 0 \}$ is convex set!

Operations that preserve convexity of functions

- Restriction to a line of domain why?
- pointwise maximum: $f(x) = \max_t g(x,t)$ where g is convex in t
- perspective $g(x,t) = \begin{cases} t f(\frac{x}{t}) & \text{for } t > 0 \quad \frac{x}{t} \in \text{dom} f \\ +\infty & \text{otherwise} \end{cases}$
- non-negative weighted sum
- partial minimum: $f(x) = \min_t g(x,t)$ where g is Jointly convex in (x,t) why?
  $\Rightarrow f$ is convex in x
- composition with monotone cvx fns: $f(x) = (\overset{\text{cvx}}{g} \circ \overset{\text{increasing}}{h})(x)$

Convex Problems

Convex optimization problems can be written in the form

$\min_x f_0(x) \quad \text{s.t.} \quad \left. \begin{array}{l} f_i(x) \leq 0 \\ g_i(x) = 0 \end{array} \ \forall i \right\}$ equivalent to $\min_{x \in X} f_0(x)$ ← feasible set

where $f_0(x)$ convex
$f_i(x)$ convex
$g_i(x)$ affine

Tractability

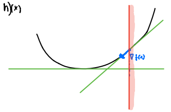
Not a property of problems but of formulations (reminiscent of CS 170 → finding efficient formulations)

$\max x_1 x_2^3 x_3 \quad x_i \geq 0 \quad x_1 x_2 \leq 2$

$\Updownarrow$

$\max z_1 + 3z_2 + z_3 \quad z_1 + z_2 \leq \log 2$

Optimality Conditions



We know that if f diff. & convex

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

$\therefore x \in X$ optimal if $\nabla f(x)^T (y - x) \geq 0 \quad \forall y \in X$ (any other y you pick will lead you to an increase)

$\Leftrightarrow \nabla f(x)^T y \geq \nabla f(x)^T x$

↳ idea underlying this is that gradient can only get smaller as you approach the optimal point in a convex fn given that second derivative is zero — not really

Optimization Algorithms

OLS for problems of form $q(x) = x^T A x + g^T x$ when $A \succeq 0$

Unconstrained Minimization: Newton's Method
Constrained Minimization: Interior Point Methods
+ Gradient Methods

Gradient Methods for constrained case: $x \in C$

$x_{t+1} = P(x_t - \alpha_t \nabla f(x)_t)$

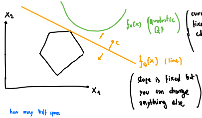
↑ projection! OMP?
↓ closest point in C!

Convex Models

LP & QP

All function involved are affine. Feasible set is of form $Ax \leq b$ $\Rightarrow$ feasible set will be sublevel set of cvx fn $\Rightarrow$ convex!

not just convex! that's for all cvx problems! → LP we have that feasible set is a polyhedron, intersection of half spaces $a_i^T x \leq b_i \ \forall i$



$f_0(x)$ (quadratic QP) (curvature is fixed but can change anything else)

$f_0(x)$ (line) (slope is fixed but you can change anything else)

Standard form

$\min_x f_0(x) \quad \left. \begin{array}{l} f_i(x) \leq 0 \\ g_i(x) = 0 \end{array} \right\}$

LP → all affine

QP → $f_0(x)$ quadratic ($x^T Q x + c^T x$) $Q \succeq 0$
$f_i, \dots, g_i \dots$ affine

aside — how many half spaces do you need to make a closed object in dimension d?
$d=2 \quad 3$ △
$d=3 \quad 4$ △
$d=4$?

Example Projection of origin on a polyhedron

$\underset{x \in P}{\operatorname{argmin}} \|x\|_2 \quad$ where $P = \{ x \mid Ax \leq b \}$

$\Updownarrow$

$\underset{x \in P}{\operatorname{argmin}} \|x\|_2^2 = \underset{x \in P}{\operatorname{argmin}} x^T x$ — QP⚠

Minimizing Polyhedral Functions

We say a function is polyhedral if epigraph f is a polyhedron

i.e. $\exists C$ s.t. $\text{epi} f = \{ (x,t) \mid t \geq f(x) \} = \{ (x,t) \mid C \begin{bmatrix} x \\ t \end{bmatrix} \leq d \}$

Maxima of affine fns:

$f(x) = \max_i (a_i^T x + b_i)$ is polyhedral: $\text{epi} f = \{ (x,t) \mid t \geq \max (a_i^T x + b_i) \}$
$= \{ (x,t) \mid t\mathbb{1} \geq Ax + b \} = \{ (x,t) \mid Ax - \mathbb{1}^T t \leq -b \}$
$= \{ (x,t) \mid [A \ -\mathbb{1}] \begin{bmatrix} x \\ t \end{bmatrix} \leq -b \}$

Sums of maxima of affine fns:

$f(x) = \sum_j^p \max_i a_{ij}^T x + b_{ij} \Rightarrow \text{epi} f = \{ (x,t) \mid t \geq \sum_j^p \max_i a_{ij}^T x + b_{ij} \}$
$= \{ (x,t) \mid t \geq \sum_j^p g_j(x) \}$ where $g_j(x)$ is some polyhedral fn s.t. $g_j(x) \geq a_{ij}^T x + b_{ij} \ \forall i$

Cardinality Minimization

A lot of times problems of this form arise

$\min_{x \in P} f_0(x) + \lambda \text{card}(x)$ where P is a polyhedron

Idea: tradeoff between sparsity and value of objective function

card(·) is not convex $\Rightarrow$ we replace with $\|x\|_1$

$\|x\|_1 = y^T |x| \leq \|y\|_2 \|x\|_2 = \sqrt{k} \|x\|_2$

$\Rightarrow$ lower bound on Euclid norm of any vector x with cardinality k
$\|x\|_2^2 \geq \frac{1}{k} \|x\|_1^2$

Also $\|x\|_1 \leq k \|x\|_\infty$

$\Rightarrow \min_x c^T x + \frac{1}{k} \|x\|_1^2 \quad \text{s.t. } Ax \leq b$

$\Downarrow$

$\min_{x,u,t} c^T x + \frac{1}{k} t^2 \quad \text{s.t.} \begin{cases} Ax \leq b \\ t \geq \sum_i^n |x_i| \end{cases}$

$\Updownarrow$

$\begin{cases} t \geq \sum_i^n u_i \\ u_i \geq x_i \\ u_i \geq -x_i \\ Ax \leq b \end{cases}$

Piecewise constant fitting

$\min_{\hat{x}} \|y - \hat{x}\|_2^2 \quad \text{card}(D\hat{x} \leq k) \Rightarrow$ relax to $\|D\hat{x}\|_1 \leq q$
↖ estimate of original signal st. few changes in consecutive steps

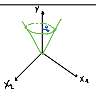
D diff. matrix $\begin{bmatrix} -1 & 1 & & 0 \\ & -1 & 1 & \\ 0 & & & \ddots \end{bmatrix}$

LASSO

Also falls under this category

SOCP

Second Order Cone



$x \in \mathbb{R}^p$

$K = \{ (x,y) \in \mathbb{R}^{p+1} \mid \|x\|_2 \leq y \}$

$= \bigcap_{\|u\|_2 \leq 1} \{ (x,y) \in \mathbb{R}^{p+1} \mid x^T u \leq y \}$ (can be expressed as intersection of ∞ halfspaces)

Rotated Second Order Cone

$K_r = \{ (x,y,z) \in \mathbb{R}^{p+2} \quad x^T x \leq yz, \ y \geq 0, \ z \geq 0 \}$

$\begin{cases} x^T x \leq yz \\ y \geq 0 \\ z \geq 0 \end{cases} \Leftrightarrow \left\| \begin{bmatrix} 2x \\ y - z \end{bmatrix} \right\|_2 \leq y + z$

(if we square we get
$4x^2 + y^2 + z^2 - 2zy \leq y^2 + z^2 + 2zy$
$x^2 \leq zy$)

$\therefore \begin{cases} zy \geq 0 \\ y + z \geq 0 \end{cases} \Leftrightarrow \begin{cases} y \geq 0 \\ z \geq 0 \end{cases}$

Rotated SOC useful to describe quadratic convex inequalities:

$c^T x + x^T Q x \leq t \Leftrightarrow x^T Q x \leq t - c^T x \Rightarrow w^T w \leq yz$

where
$w = Q^{\frac{1}{2}} x$ if PSD
$y = t - c^T x$
$z = 1 \geq 0$

Intersection of Second Order Cone with Affine Sets

Standard Form

Standard forms for SOCP are:

Inequality form: $\min_x c^T x \quad \|Ax + b\|_2 \leq Cx + d$ → can write this as many inequalities. Each one is a SOC constraint } Intersection of convex sets

Conic Form: $\min_x c^T x \quad \text{s.t.} \quad (A_i x + b_i, c_i^T x + d_i) \in K_{p_i}$

- many different cones
- each cone is convex
- mappings are affine

} feasible set is intersection of cvx sets

Magnitude Constraints on affine complex vectors

Mag. of cplx num: $z = z_R + z_I i$ is:

$|z| = \sqrt{z_R^2 + z_I^2} = \left\| \begin{bmatrix} z_R \\ z_I \end{bmatrix} \right\|_2$

$\therefore$ If $f : \mathbb{R}^n \to \mathbb{C}$

$f(x) = (a_R^T x + b_R) + i (a_I^T x + b_I)$

$= \underbrace{\left\| \begin{bmatrix} a_R^T x + b_R \\ a_I^T x + b_I \end{bmatrix} \right\|_2 \leq t}_{\text{SOCP form}}$

Linear Programming as SOCP

$\min_x c^T x \quad \text{s.t.} \quad Ax \leq b$

$\Updownarrow$

$\min_x c^T x \quad \text{s.t.} \quad \|Cx - d\|_2 \leq -Ax + b$

$\Updownarrow$

$\|c_i^T x - d_i\|_2 \leq -a_i^T x + b_i \quad \forall i$ ← negative sign ok (affine)

$c_i = 0 \quad d_i = 0 \quad \forall i$

Quadratic Programming as SOCP

$\min_x c^T x + x^T Q x \quad \text{s.t.} \quad Ax \leq b$

$\Updownarrow$

naive approach

$\min_{x,t} c^T x + t \quad \begin{cases} t \geq x^T Q x \Leftrightarrow \begin{cases} w^T w \leq t \Leftrightarrow \|w\|_2^2 \leq t \\ w = Q^{\frac{1}{2}} x \end{cases} \\ \|0x\|_2 \leq -Ax - b \end{cases}$

Doesn't work as: both variables. Also ! squared!
There might be some possible way of expressing it but easier with rotated cone

$\min_{x,w} c^T x + w^T w \quad \begin{cases} Ax \leq b \\ w = Q^{\frac{1}{2}} x \end{cases}$

$\Updownarrow$

$\min_{x,w,y,z} c^T x + y \quad \begin{cases} y \geq w^T w \geq 0 \\ w = Q^{\frac{1}{2}} x \\ Ax \leq b \end{cases} \Leftrightarrow \begin{cases} yz \geq w^T w \\ z = 1 \geq 0 \\ w = Q^{\frac{1}{2}} x \\ Ax \leq b \\ y \geq 0 \end{cases}$ ← Rotated cone constraint!

Remember
$y \geq 0$
$z \geq 0$
for rotated constraints

ROBUST & LS

QCQPs

$\min_x c^T x + x^T Q_0 x \quad \text{s.t.} \quad x^T Q_i x + a_i^T x \leq b_i \quad \forall i$

$\Updownarrow$

$\min_{x,w,y_i,z_i} c^T x + w^T w \quad \text{s.t.} \begin{cases} w = Q^{\frac{1}{2}} x \\ w_i^T w_i \leq -a_i^T x + b_i \quad \forall i \\ w_i = Q_i^{\frac{1}{2}} x \end{cases} \Leftrightarrow \begin{cases} w = Q^{\frac{1}{2}} x \\ w_i^T w_i \leq y_i z_i \\ y_i \geq 0 \\ y_i = -a_i^T x + b_i \\ w_i = Q_i^{\frac{1}{2}} x \\ z_i = 1 \geq 0 \end{cases}$ (rotated; affine)

$\Updownarrow$

$\min_{x, w_0, w_i, y_i, z, y_0} c^T x + y_0 \quad \text{s.t.} \begin{cases} y_0 z \geq w_0^T w_0 \\ y_i z \geq w_i^T w_i \\ w_0 = Q_0^{\frac{1}{2}} x \\ w_i = Q_i^{\frac{1}{2}} x \\ z = 1 \\ y_i \geq 0 \\ y_0 \geq 0 \end{cases}$ SOCP form

Group Sparsity

Sometimes want entire blocks of zeros in decision vector

$X = (x_1, \dots, x_K)$ where $x_i \in \mathbb{R}^{n_i}$

$\min_x \|Ax - b\|_2 + \sum_{i=1}^K \|x_i\|_2 \sim$ equivalent to L1 norming the norm vector $\| [\|x_1\|_2 \dots \|x_K\|_2] \|_1$ } drives entire columns to zero!

to SOCP → $\min_{x, t_1, t_{2i} \ \forall i} t_1 + \sum_i^K t_{2i} \quad \begin{cases} t_1 \geq \|Ax - b\|_2 \\ t_{2i} \geq \|x_i\|_2 \end{cases}$

Robust Models

General Uncertainty

There are many ways we can model uncertainty.

A basic way is to consider a finite set of possibilities. $U = \{a^1, \dots, a^K\}$

The robust counterpart for a half space constraint $x^T a \leq b$ is $x^T a \leq b \quad \forall a \in U$

$$\Updownarrow$$

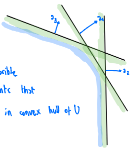
$$x^T a^k \leq b \quad \forall k = 1, \dots |U|$$

$$\overset{\shortparallel}{K}$$

$\therefore \min_x x^T c \qquad a_i^T x \leq b_i \ \forall i, \quad \forall a_i \in U_i$

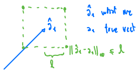
$$\Downarrow$$

$$(a_i^k)^T x \leq b \quad k = 1, \dots K_i, \ \forall i$$



only feasible for points that satisfy any line in convex hull of U

Box Uncertainty Model



$\hat{a}_1$ what we observe

a_1 true vector $\quad a_1 = \hat{a}_1 + \rho u$

$\|u\|_\infty \leq 1$

Assumes that the coefficient vectors lie in a box (hyper-rectangle in $\mathbb{R}^n$)

$\therefore$ we have $\forall a \in U \quad a^T x \leq b$ and we know $a^k = \hat{a} + \rho u^k$ ← unit vectors in each direction

$$\Downarrow$$

$$(a^k)^T x \leq b$$

LP Formulation

A better way to encode the above is as an LP: We assume $a = \hat{a} +$

the constraint $a^T x \leq b \ \forall a \in U \iff b \geq \max_{a \in U} a^T x = \max_{\|u\|_\infty \leq 1} \hat{a}^T x + \rho u^T x \leq \hat{a}^T x + \rho \|x\|_1$

↳ equality when u aligned with x

$$= \hat{a}^T x + \rho \|x\|_1$$

$\therefore$ box uncertainty becomes:

$$U_i = \{a \mid \|a - \hat{a}_i\|_\infty \leq \rho\} \ \forall i$$

→ all vectors in the box close to $\hat{a}_i$

$\min_x c^T x \qquad \forall a_i \in U_i \quad a_i^T x \leq b_i \ \forall i$

$$\Downarrow$$

$\min_x c^T x \qquad b_i \geq \hat{a}_i^T x + \rho \|x\|_1 \quad \forall i$

Note: $\{a = \hat{a}_i + \rho u \ \ \|u\|_K \leq 1\}$

$= \{a \mid a - \hat{a}_i = \rho u \mid \|u\|_K \leq 1\}$

$= \{a \mid \|a - \hat{a}_i\|_K \leq \rho\}$

Spherical Uncertainty

$U_i = \{a \mid \|a - \hat{a}_i\|_2 \leq \rho\} \quad \forall i$

$\min_x c^T x \qquad \forall a_i \in U_i \quad a_i^T x \leq b$

$$\Downarrow$$

$\min_x c^T x \qquad b \geq \max_{\|u\|_2 \leq 1} \hat{a}_i^T x + \rho u^T x = \hat{a}_i^T x + \rho \|x\|_2$

$$\Downarrow$$

$\min_x c^T x \qquad$ s.t. $\quad b \geq \hat{a}_i^T x + \rho \|x\|_2$

Ellipsoidal Uncertainty

$U_i = \{a = \hat{a}_i + R_i u \quad \|u\|_2 \leq 1\}$

$\min_x c^T x \qquad a_i^T x \leq b \quad \forall a_i \in U_i$

$$\Updownarrow$$

$\min_x c^T x \qquad b \geq \hat{a}_i^T x + (R_i u)^T x \quad \forall \|u\|_2 \leq 1$

$$\geq \max_{\|u\|_2 \leq 1} a_i^T x + (R_i u)^T x$$

$$\geq a_i^T x + \max_{\|u\|_2 \leq 1} u^T (R_i^T x)$$

$$= a_i^T x + \|R_i^T u\|_2$$

$\min_x x^T S x \qquad \mu \leq r^T x \qquad Ax \leq b \qquad S \succ 0$

$r \in R \quad \Rightarrow \quad \mu \leq r^T x \qquad \forall r \in R \Rightarrow \boxed{\mu \leq \min_{r \in R} r^T x}$

a) $R = \{r : \|r - \hat{r}\|_\infty \leq \rho\} \quad \hat{r} \in \mathbb{R}^n \quad \rho > 0$

$\Rightarrow \quad r = \hat{r} + \rho u \quad$ s.t. $\quad \|u\|_\infty \leq 1$

$$\mu \leq \min_{\|u\|_\infty \leq 1} \hat{r}^T x + \rho u^T x$$

$$= \hat{r}^T x - \rho \|x\|_1$$

b) $R = \{r : \|r - \hat{r}\|_2 \leq \rho\}$

$r = \hat{r} + \rho u \quad \|u\|_2 \leq 1$

$$\mu \leq \min_{\|u\| \leq 1} \hat{r}^T x + \rho u^T x = \hat{r}^T x - \rho \|x\|_2$$

c) $R = \{r = \hat{r} + Ru \quad \|u\|_2 \leq \rho\}$

$$\mu \leq \min_{\|u\|_2 \leq \rho} \hat{r}^T x + u^T R^T x$$

$$= \hat{r}^T x + \min_{\|u\|_2 \leq 1} u^T (\rho R^T x) = \hat{r}^T x - \rho \|R^T x\|_2$$

$$\min_w \|X^T w - y\|_2 + \lambda \|w\|_1$$

Example: Square-Root LASSO

(true lasso would have squared!)

$$\min_w \|X^T w - y\|_2 + \underbrace{\lambda \|w\|_1}_{\text{enforces sparcity}}$$

← no square

$\sum \max(w_i, -w_i)$

turn this into SOCP

↓

not LS as doesn't have squared norm + has L1 norm

$= \min_{w,u,t} t + \lambda \sum u_i \qquad \underbrace{t \geq \|X^T w - y\|_2}_{\text{SOCP covers this!}} \quad u_i \geq w_i, \ u_i \geq -w_i \quad \forall i$

} Converted to SOCP

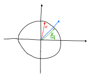
Dual ?

From Cauchy Schwartz

$$\boxed{\|z\|_2 = \max_v v^T z \quad \text{s.t.} \quad \|v\|_2 \leq 1}$$

↳ WANT TO WRITE WITH MIN-MAX

↓ trying to get min-max expression



SO:

$$p^* = \min_w \|X^T w - y\|_2 + \lambda \|w\|_1$$

$$= \min_w \max_{\|v\|_2 \leq 1} v^T (X^T w - y) + \lambda \max_{\|p\|_\infty \leq 1} p^T w$$

$$= \min_w \max_{\substack{\|v\|_2 \leq 1 \\ \|p\|_\infty \leq 1}} v^T (X^T w - y) + \lambda p^T w$$

Also $|w_i| = \max_{p_i} p_i w_i \qquad |p_i| \leq 1$

and $\sum_i |w_i| = \sum_i \max_{p_i} p_i w_i$

$= \max_p \underbrace{\sum_i p_i w_i}_{p^T w}$ as all p_is independent

$= \max_p p^T w$

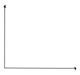
s.t. $\|p\|_\infty \leq 1$

↳ just as $|p_i| \leq 1$ from before

$$\boxed{\therefore \|w\|_1 = \sum_i |w_i| = \max_{\|p\|_\infty \leq 1} p^T w}$$

generalized Cauchy Schwartz

Assume strong duality



$$d^* = \max_{\substack{\|v\|_2 \leq 1 \\ \|p\|_\infty \leq 1}} \min_w v^T (X^T w - y) + \lambda p^T w$$

$$= \max_{\substack{\|v\|_2 \leq 1 \\ \|p\|_\infty \leq 1}} \underbrace{\min_w \overbrace{w^T (X v + \lambda p) - v^T y}^{\text{affine fn of } w \to \text{straight line!}}}_{\text{Dual function}}$$

trying to minimize a line

Dual function → minimizing fn of w, all other variables are fixed

Note

$$\begin{cases} -v^T y & \text{if} \quad Xv + \lambda p = 0 \\ -\infty & \text{otherwise} \end{cases}$$

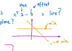
Conv way

$$\frac{\partial}{\partial w} = Xv + \lambda p = 0$$

when minimizing a line this is what you get

why is $x^T a - b$ a line? Hyperplane?

slope offset



$$\Downarrow$$

$= \max_{\substack{\|v\|_2 \leq 1 \\ \|p\|_\infty \leq 1}} -v^T y \qquad$ s.t. $\quad Xv + \lambda p = 0 \Rightarrow$ eliminate p! as only in constraint

$$\Downarrow$$

$$\boxed{\|Xv\|_\infty \leq \lambda}$$

$$\Downarrow$$

as $\|p\|_\infty \leq 1$ then taking $\|\|_\infty$ of both sides $\|Xv\| = \|-\lambda p\|_\infty$

$= \max_v -v^T y \qquad$ s.t. $\quad \|Xv\|_\infty \leq \lambda$ ← scalar, $\|v\|_2 \leq 1$

try to think about it geometrically

↓

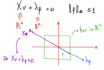
algebraically it becomes a mess

↓

El Ghaoui just says → .... in the box

$Xv + \lambda p = 0 \qquad \|p\|_\infty \leq 1$

$\mathbb{R}^n \quad \mathbb{R}^n$



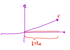
idea is $\underbrace{\frac{Xv}{\lambda}}_{\mathbb{R}^n} + \underbrace{p}_{\mathbb{R}^n} = 0$

$\|p\|_\infty \leq 1$

$\therefore \|-\frac{Xv}{\lambda}\|_\infty \leq 1$

$\Rightarrow \|Xv\|_\infty \leq \lambda$

What does $\|\ \|_\infty$ do? → takes largest component



Note $\|\lambda p\|_\infty = \lambda \|p\|_\infty \leq 1$

And as $Xv + \lambda p = 0 \Rightarrow \|Xv\|_\infty = \|\lambda p\|_\infty \leq \lambda$