Some Mathematical Tools for Machine Learning

Chris Burges
Microsoft Research

Max Planck Institute, Tübingen
August, 2003

Contents – Part 1

Lagrange Multipliers: An Overview, and Some Examples

Lagrange the Mathematician

- Born 1736 in Turin, one of two of 11 to survive infancy
- "Responsible for much fine mathematics published under the names of other mathematicians"
- *Believed that a mathematician has not thoroughly understood his own work till he has made it so clear that he can go out and explain it to the first person he meets on the street*
- Worked on mechanics, calculus, the calculus of variations astronomy, probability, group theory, and number theory
- At least partly responsible for the choice of base 10 for the metric system, rather than 12
- Supported by Euler and d'Alembert, financed by Frederick and Louis XIV, close to Lavoisier, Marie Antoinette

An indirect approach can be easier

Example: Minimize $f(x)$ subject to $c(x) \doteq x'Ax = 1,\ x \in \mathbb{R}^n$

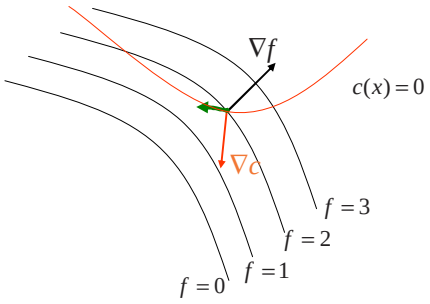
If $A \succ 0,$ could rotate to coordinate system and rescale so that constraints take the form $y'y = 1,$ substitute with a parameterization that encodes the constraints that x lives on $\mathbb{S}^{n-1}$:

$y_1 = \sin\theta_1 \sin\theta_2 \cdots \sin\theta_{n-1}$
$y_2 = \sin\theta_1 \sin\theta_2 \cdots \cos\theta_{n-1}$
$\cdots$
solve, and then apply the inverse mapping. But this can get very complicated!

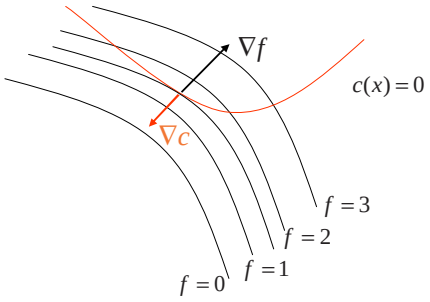
It is often not even possible to parameterize constraints (for example, polynomial constraints in several variables)

One equality constraint

Minimize $f(x)$ subject to $c(x) = 0,\ x \in \mathbb{R}^2$.



One equality constraint, cont.



Hence at the optimum, we must have $\nabla f \propto \nabla c,$ or:

$$\nabla f = \lambda \nabla c$$

$$\nabla L \doteq \nabla(f - \lambda c) = 0$$

Multiple equality constraints

n constraints: $c_i(x) = 0,\ i = 1, \cdots, n.$

Define gradients: $g_i(x) = \nabla c_i(x).$

Let S be the subspace spanned by the $g_i,$ and let $S_\perp$ be its orthogonal complement.

Suppose that at some point, all constraints hold, and $(\nabla f)_\perp \neq 0$

Then can increase (or decrease) f by moving along $(\nabla f)_\perp$

Hence $(\nabla f)_\perp = 0,$ or: $\nabla f = \sum_i \lambda_i \nabla c_i(x)$: $\nabla L \doteq \nabla(f - \sum_i \lambda_i c_i) = 0$

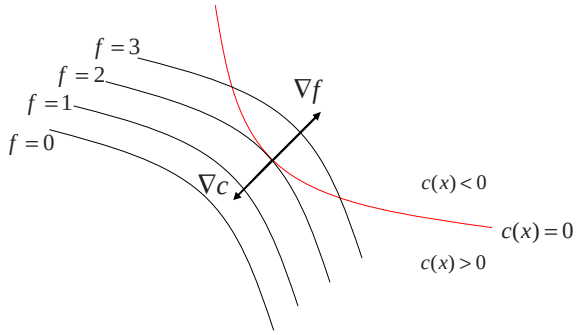
Puzzle: why not multiple Lagrangians?

One inequality constraint

Find x_* that minimizes $f(x)$ subject to $c(x) \le 0$.

What's new? At the solution, it's possible that $c(x) < 0$.

(Simple but not guaranteed: solve 'minimize $f(x)$', check that $c(x_*) \le 0$.)



$$\nabla f \propto -\nabla c: \ \nabla(f + \lambda c) = 0, \quad \lambda \ge 0$$

Multiple inequality constraints

$$\nabla(f + \sum_i \lambda_i c_i) = 0, \ \lambda_i \ge 0$$

Suppose that at the solution x_*, for the j^{th} constraint, $c_j < 0$.

Then removing c_j makes no difference, and we can drop ∇c_j from the sum in $\nabla f = \sum_i \lambda_i \nabla c_i$ (whether or not $\nabla c_j(x_*) = 0$).

Equivalently we can set $\lambda_j = 0$. In fact, in most cases, we *must* set $\lambda_j = 0$, and in all cases, we *can* set $\lambda_j = 0$.

Hence : always impose $\lambda_j c_j = 0 \ \forall \ j$

A simple example

Extremize the distance between two points on $\mathbb{S}^n$:

Embed in $\mathbb{R}^{n+1}$: extremize $f = \|x_1 - x_2\|^2$, $x_1, x_2 \in \mathbb{R}^{n+1}$

subject to $c_1(x_1, x_2) = 1 - \|x_1\|^2 = 0$, $c_2(x_1, x_2) = 1 - \|x_2\|^2$

$$L(x_1, x_2) = f - \sum_i \lambda_i c_i = \|x_1 - x_2\|^2 - \lambda_1(1 - \|x_1\|^2) - \lambda_2(1 - \|x_2\|^2)$$

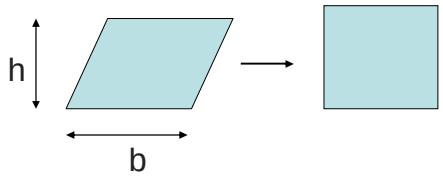
$$\nabla_1 L = 0 \Rightarrow (x_1 - x_2) + \lambda_1 x_1 = 0$$
$$\nabla_2 L = 0 \Rightarrow (x_2 - x_1) + \lambda_2 x_2 = 0$$

$$\Rightarrow x_2 = (1 + \lambda_1) x_1, \quad x_1 = (1 + \lambda_2) x_2$$

$\Rightarrow$ antipodal or equal: $\lambda_i = -2$ or 0.

Another simple example

Given a parallelogram whose sidelengths you can choose but whose perimeter c is fixed - what shape has the largest area?



Maximize bh subject to $2(b + h) = c$

$$L(b, h) = bh - \lambda(2(b + h) - c)$$
$$\nabla L = 0 \Rightarrow b = h$$

Again, λ not explicitly needed: hence "method of undetermined multipliers"

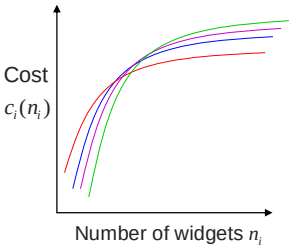
Simple exercises

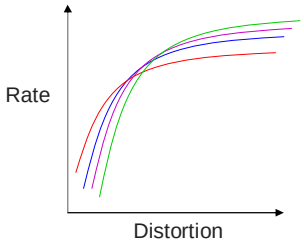
Puzzle: what coefficients maximize a convex sum of fixed numbers?

Puzzle: minimize $\sum_i x_i^2$ subject to $\sum_i x_i = 1$.

Puzzle: maximize $\sum_i x_i^2$ subject to $\sum_i x_i = 1$ and $x_i \geq 0$ (hint: use $\lambda_i x_i = 0$)

Resource allocation





Wish to manufacture N widgets, and minimize the total cost to do so. For factory i to produce n_i widgets costs $c_i(n_i)$.

Wish to minimize $\sum_i c_i(n_i)$ subject to $\sum_i n_i = N$.

$$L = \sum_i c_i(n_i) + \lambda(N - \sum_i n_i)$$

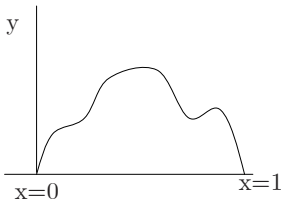
$$\nabla L = 0 \quad \Rightarrow \quad \boxed{\frac{\partial c_i}{\partial n_i} = \lambda \ \forall i}$$

A variational problem

An isoperimetric problem: find the curve of fixed length ρ and fixed endpoints $\{a,b\}$ that encloses maximum area above $[a,b]$.

Area $= \int_0^1 y\,dx$, length $\rho = \int_0^1 (1+y'^2)^{1/2}\,dx$

$$L = \int_0^1 y\,dx + \lambda\left(\int_0^1 (1+y'^2)^{1/2}\,dx - \rho\right)$$



$$\delta L = \int_0^1 \delta y\,dx + \lambda \int_0^1 (1+y'^2)^{-1/2}\, y'\,\delta y'\,dx$$

$$= \int_0^1 \left\{1 - \lambda \frac{d}{dx}\left(y'(1+y'^2)^{-1/2}\right)\right\} \delta y\,dx$$

$$= \int_0^1 \left(1 - \lambda y''(1+y'^2)^{-3/2}\right) \delta y\,dx$$

$$\Rightarrow \quad 1 - \lambda y''(1+y'^2)^{-3/2} = \boxed{1 - \lambda\kappa = 0}$$

…straight line, or arc of circle.

Which univariate distribution has max entropy?

Minimize $\int_{-\infty}^{\infty} f(x)\log f(x)\,dx$

subject to: $f(x) \geq 0 \quad \forall x,$

$$\int_{-\infty}^{\infty} f(x)\,dx = 1,$$

$$\int_{-\infty}^{\infty} x\,f(x)\,dx = c_1,$$

$$\int_{-\infty}^{\infty} x^2 f(x)\,dx = c_2$$

Need *functional derivative*:

$$\frac{\delta g(x)}{\delta g(y)} = \delta(x-y)$$

$$L = \int_{-\infty}^{\infty} f(x)\log f(x)\,dx + \lambda\left(1 - \int_{-\infty}^{\infty} f(x)\,dx\right) - \beta_1 \int_{-\infty}^{\infty} x\,f(x)\,dx - \beta_2 \int_{-\infty}^{\infty} x^2\,f(x)\,dx$$

Impose $\frac{\delta L}{\delta f(y)} = 0$, integrate w.r.t $x \Rightarrow \log f(y) + \log(e) - \lambda - \beta_1 y - \beta_2 y^2 = 0$

→ f must be Gaussian!

Which univariate distribution has max entropy?

Puzzle: I thought the uniform distribution has max entropy. What's going on?

Puzzle: What distribution do you get if you fix the mean, but not the variance?

Puzzle: What distribution do you get if you fix only the function's support?

Max Entropy for Discrete Distrib^n + Linear Constraints

Have discrete distribution P_i: $\sum_i P_i = 1, \quad P_i \geq 0$

Suppose you also have known linear constraints: $\sum_i a_{ij} P_j = C_j$

but you are maximally uncertain about everything else. So want max entropy distribution subject to these constraints.

$$L = \sum_i P_i \log P_i + \sum_j \lambda_j \left(\sum_i a_{ij} P_j - C_j \right) + \mu \left(\sum_i P_i - 1 \right) - \sum_i \delta_i P_i$$

$$\delta L = 0 \Rightarrow P_k = \exp(-1 - \mu + \delta_k - \sum_j \lambda_j a_{kj}) = (1/Z) \exp(-\sum_j \lambda_j a_{kj})$$

→ logistic regression!

Basic Concepts in Functional Analysis: A Brief Tour of Hilbert spaces, Norms, and All That

Bibliography

- Introduction to Hilbert Spaces, Debnath and Mikusinski, 2nd Edition, Academic Press , 1999
- Introduction to Functional Analysis, Taylor and Lay, 2nd Edition, Wiley , 1980
- Matrix Analysis, Horn and Johnson, Cambridge University Press , 1985
- Introductory Real Analysis, Kolmogorov and Fomin, Prentice-Hall, 1970
- Elements of the theory of functions and functional analysis, Kolmogorov and Fomin, Graylock Press, 1961
- Metric spaces, Copson, Cambridge University Press, 1968
- Intro. to metric and topological spaces, Sutherland, Clarendon Press, 1975

What is a Field?

$\{F,+,*\}$: F a set, $\{+, *\}$ operations

$\{F,+\}$ is an Abelian group with identity denoted by 0

$\{F-0,*\}$ is an Abelian group with identity denoted by 1

$x*(y+z) = x*y+x*z \quad \forall \{x,y,z \in F\}$

A field generalizes the notion of arithmetic on reals.

Field : Examples

With + meaning addition and $*$ meaning multiplication,

the reals: $F = \mathbb{R}$
the rationals: $F = \mathbb{Q}$
the complex numbers: $F = \mathbb{C}$

What's the smallest field?
$\mathbb{Z}_2 = \{0,1\}$ with + meaning "XOR" and $*$ meaning "AND"

Puzzle 1: For $\mathbb{Z}_2$, why doesn't using "OR" for + work?
Puzzle 2: Is $\mathbb{R}_+ = \{x : x \geq 0\}$ a field under $\{+,*\}$?

How Many Fields Are There?

Actually infinitely many - e.g. $\mathbb{Z}_p$, p prime.

However, can define an 'ordering;' for fields (like <, =, > for $\mathbb{R}$).

$\mathbb{R}$, $\mathbb{Q}$ can be ordered; $\mathbb{C}$ and $\mathbb{Z}_2$ cannot; in fact all finite fields cannot be ordered.

For ordered fields, 'supremum' and 'infimum' can be defined. An ordered field is 'complete' iff every nonempty subset of F that has an upper bound in F also has a supremum in F.

$\mathbb{Q}$ is not complete, but $\mathbb{R}$ is. In fact: every complete, ordered field is isomorphic to $\mathbb{R}$!

What is a Vector Space?

A vector space is a nonempty set E, a field F and operations 'addition' ($(x, y) \to x + y$ from $E \times E \to E$) and 'multiplication by a scalar' ($(\lambda, x) \to \lambda x$ from $F \times E$ into E) such that:

(a) $x + y = y + x$;
(b) $(x + y) + z = z + (y + z)$;
(c) For every $x, y \in E$, there exists $z \in E$ such that $x + y = z$;
(d) $\alpha(\beta x) = (\alpha\beta)x$;
(e) $(\alpha + \beta)x = \alpha x + \beta x$;
(f) $\alpha(x + y) = \alpha x + \alpha y$;
(g) $1x = x$.

A vector space generalizes the notion of vectors in $\mathbb{R}^n$

Vector Spaces: Field Matters!

$\{E, F\} = \{\mathbb{R}^N, \mathbb{R}\}$ (dimension N);
$\{E, F\} = \{\mathbb{C}^N, \mathbb{C}\}$ (dimension N);
$\{E, F\} = \{\mathbb{C}^N, \mathbb{R}\}$ (dimension $2N$);

'Linear dependence' depends on field F :

For the vector space $\{\mathbb{C}, \mathbb{R}\}$, vectors 1 and i are linearly independent.

For the vector space $\{\mathbb{C}, \mathbb{C}\}$ they are not.

Vector Spaces: More Examples

(1) Functions, whose range is a vector space, also themselves form a vector space:

$$(f+g)(x) = f(x) + g(x);$$
$$(\lambda f)(x) = \lambda f(x).$$

(2) M_{mn} (complex m by n matrices), over the field $\mathbb{C}$;

(3) l_p: for $p \geq 1$, the space of all infinite sequences of complex numbers such that $\sum_{n=1}^{\infty} |z_n|^p < \infty$

$$x + y = \sum_{n=1}^{\infty} (x_n + y_n), \quad \alpha x = \sum_{n=1}^{\infty} \alpha x_n$$

What is an Inner Product?

Let V be a vector space over $\mathbb{R}$ or $\mathbb{C}$. A function $\langle \cdot, \cdot \rangle : V \times V \to F$ is an inner product if for all $x, y, z \in V$,

(a) $\langle x, x \rangle \geq 0$
(b) $\langle x, x \rangle = 0 \Leftrightarrow x = 0$
(c) $\langle x + y, z \rangle = \langle x, z \rangle + \langle y, z \rangle$
(d) $\langle cx, y \rangle = c\langle x, y \rangle$ for all scalars $c \in F$
(e) $\langle x, y \rangle = \overline{\langle y, x \rangle}$

The inner product generalizes the notion of dot product

Inner Product: Examples

(1) Vector space $\{\mathbb{R}^N, \mathbb{R}\}$ with $\langle \cdot, \cdot \rangle$ defined by $\langle x, y \rangle \doteq \sum_i x_i y_i$

(2) Vector space l_2 over $\mathbb{R}$ with $\langle \cdot, \cdot \rangle$ defined by $\langle x, y \rangle \doteq \sum_{i=1}^{\infty} x_i y_i$

(3) Vector space of matrices over $\mathbb{R}$ with $\langle X, Y \rangle \doteq Trace(X^T Y)$
$(X, Y \in M_{pm})$

Inner Product: Trace

(a) $\langle X, X \rangle \geq 0$ $\quad Tr(X^T X) = \sum_i \|X_i\|^2 \geq 0$

(b) $\langle X, X \rangle = 0 \Leftrightarrow x = 0$ $\quad Tr(X^T X) = \sum_i \|X_i\|^2 \geq 0$

(c) $\langle x + y, z \rangle = \langle x, z \rangle + \langle y, z \rangle$: $\quad Tr((X+Y)^T, Z) = Tr(X^T Z) + Tr(Y^T Z)$

(d) $\langle cx, y \rangle = c\langle x, y \rangle$ for all scalars $c \in F$ $\quad Tr(\alpha X^T Y) = \alpha Tr(X^T Y)$

(e) $\langle x, y \rangle = \overline{\langle y, x \rangle}$ $\quad Tr(X^T Y) = Tr(Y^T X)$

Inner Product is General

Cauchy Schartz inequality holds for any inner product $\langle \cdot, \cdot \rangle$:

$|\langle x, y \rangle|^2 \leq \langle x, x \rangle \langle y, y \rangle$ for all $x, y \in V$

"Angle" between two vectors can be defined for any inner product:

$\theta \doteq \cos^{-1}\left(\frac{|\langle x, y \rangle|}{(\langle x, x \rangle \langle y, y \rangle)^{1/2}} \right)$: $\quad 0 \leq \theta \leq \frac{\pi}{2}$

E.g. the angle between $\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$ and $\begin{pmatrix} -1 & 0 \\ 1 & 2 \end{pmatrix}$ is 78.4 degrees !

What is a Norm?

Let V be a vector space over $\mathbb{R}$ or $\mathbb{C}$. A function $\|\cdot\|: V \to \mathbb{R}$ is a norm if for all $x, y \in V$,

(a) $\|x\| \geq 0$
(b) $\|x\| = 0 \Leftrightarrow x = 0$
(c) $\|cx\| = |c|\,\|x\|$ for all scalars $c \in F$
(d) $\|x + y\| \leq \|x\| + \|y\|$

If condition (b) is dropped, it's a 'seminorm'.

What is the simplest seminorm? $Ans:\ \|x\| = 0$

Seminorm splits the space

If $\|\cdot\|$ is a seminorm on V, then $V_0 \doteq \{v \in V : \|v\| = 0\}$ is a subspace (called the null space).

If V_1 is a subspace of V such that $V_0 \cap V_1 = \varnothing$, then $\|\cdot\|$ is a norm on V_1.

Define $x \sim y \Leftrightarrow \|x - y\| = 0$. The corresponding cosets form a vector space, and on that space, $\|\cdot\|$ is a norm.

Example seminorm on $\mathbb{R}^5$: $\langle \cdot, \cdot \rangle \doteq x'Dx,\ D \doteq diag(3,2,1,0,0)$, null space spanned by $(0,0,0,1,0)$ and $(0,0,0,0,1)$.

Norm Generalizes Length

$\|x\| = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2}$ for $x = (x_1, x_2, \cdots, x_n) \in \mathbb{R}^n$ is the Euclidean norm.

Let $z = \{z_n\} \in l_p$. The function defined by $\|z\| \doteq \left(\sum_{n=1}^{\infty} |z_n|^p \right)^{1/p}$ is a norm in l_p.

This works for finite sums too: define the l_p norm on $\mathbb{R}^n$ as:

$\|x\| \doteq \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}$

The l_1 norm is the 'Manhattan' norm $\|x\| = |x_1| + |x_2| + \cdots + |x_n|$

The l_∞ norm is the 'max' norm $\|x\| = \max(\{x_i\})$

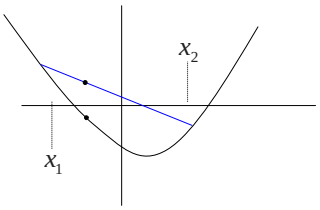
A normed vector space is a pair {vector space, norm}.

What is convexity?

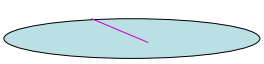
A function is convex if, between x_1 and $x_2 > x_1$, it lies below the chord joining $f(x_1)$ and $f(x_2)$. It is strictly convex if it lies strictly below.

$f((1-\lambda)x_1 + \lambda x_2) \le (1-\lambda) f(x_1) + \lambda f(x_2)$

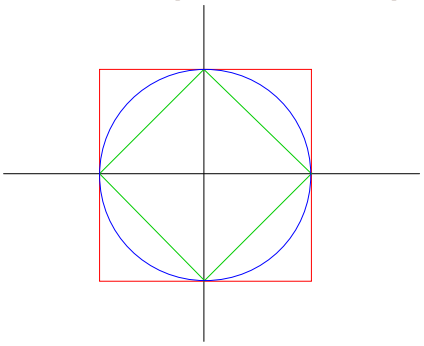
$0 \le \lambda \le 1$



A set S is convex if, for all $a \in S$ and $b \in S$, all points on the line joining a and b lie in S.



When is a sphere a square?

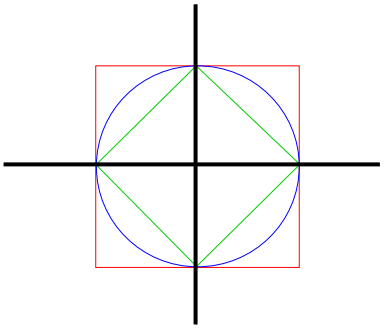


Red: the l_∞ 1-sphere
Blue: the l_2 1-sphere
Green: the l_1 1-sphere

Interesting... each ball looks convex

When is a sphere a cross?

? $\|x\|_0 \doteq \lim_{p\to 0}\left(\sum_i |x_i|^p\right)$?



In this context, never. The "l_0 norm" (count the number of nonzero elements) is not a norm (for $\mathbb{R}^n$ over $\mathbb{R}$ or $\mathbb{C}$).

$\|\alpha v\|_0 = 0$ if $\alpha = 0$,

$= \|v\|_0$ if $\alpha \neq 0$

$\neq |\alpha| \|v\|_0$ unless $\alpha = 1$.

All norms are convex

$$\begin{aligned}\|(1-\lambda)x_1+\lambda x_2\| &\le \|(1-\lambda)x_1\|+\|\lambda x_2\| \\ &= |(1-\lambda)|\|x_1\|+|\lambda|\|x_2\| \\ &= (1-\lambda)\|x_1\|+\lambda\|x_2\|\end{aligned}$$

If $f(x)$ is convex, then $f(x)\le 0$ defines a convex set:
let $R \doteq \{x : f(x)\le 0\}$. Then if $f(x_1)\le 0$ and $f(x_2)\le 0$,

$$f((1-\lambda)x_1+\lambda x_2) \le (1-\lambda)f(x_1)+\lambda f(x_2) \le 0$$

$\Rightarrow$ the unit ball, $\|x\|^2 \le 1$, for any norm is a convex region.

Open, Closed, Compact

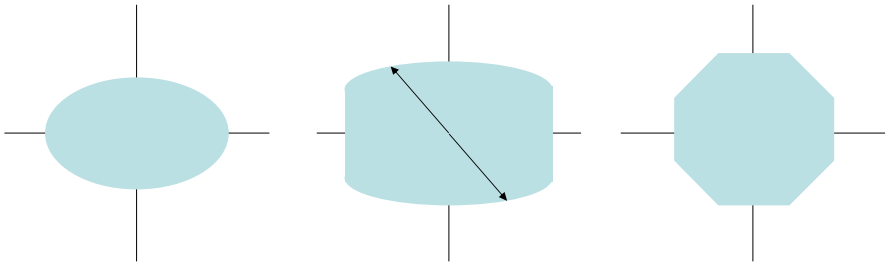
Given a norm $n \doteq \|\cdot\|$, and a set S in a vector space V:

Open: $\forall\ x\in S, \exists\ \varepsilon>0$ *s.t.* $B_n(\varepsilon, x)\subset S$
Closed : *complement of open*
Bounded : $\exists\ r>0$ *such that* $S\subset B_n(r,0)$
Compact : Every sequence $\{x_i\}$ *in S contains convergent subsequence with limit in S*
OR : Every cover has a finite sub - cover :
$\bigcup_\mu S_\mu \supset S,\ \exists\ S_{\alpha 1}, S_{\alpha 2}, \cdots, S_{\alpha N}, N$ *finite, such that* $\bigcup_{\alpha i}^{N} S_{\alpha i} \supset S$

Compact sets are closed and bounded.
For finite dimensional normed vector spaces, every closed bounded set is compact.
If for a normed vector space V, the unit ball is compact, then V has finite dimension.

Making your own norm

In fact, in real finite dimensional spaces, any symmetric, compact, convex region centered on the origin defines a norm (as the unit ball for that norm):



Puzzle: “compact” needed “norm” – is this circular?

Rescuing l_0: the Hamming norm

Consider n-tuples taking values in $\mathbb{Z}_2$. These form a vector space over the field $\mathbb{Z}_2$: for example,

$\alpha(x+y) = \alpha x + \alpha y$

$\alpha(\beta x) = (\alpha\beta)x$

$1 * x = x$

Now define the l_0 norm $\|x\|_0$ to be the number N of nonzero elements of x. Is this a norm?

(a) $\|x\|_0 \geq 0$

(b) $\|x\|_0 = 0 \Leftrightarrow x = 0$

(c) $\|cx\|_0 = |c| \|x\|_0$

(d) $\|x+y\|_0 \leq \|x\|_0 + \|y\|_0$

Hamming norm, cont.

Puzzle: $(N > 1) \notin \mathbb{Z}_2$ - how can this be correct?

Puzzle: What is the subtraction operation, in $\mathbb{Z}_2$?

Puzzle: What is the Hamming distance $\|x - y\|_0$?

When does a norm come from an inner product?

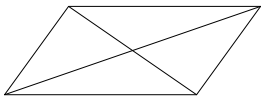
Every inner product defines a norm: $\|x\| = \sqrt{\langle x, x \rangle}$

Does every norm define an inner product? If so, for real vector spaces,

$$\langle x, y \rangle = \frac{1}{4}\left(\|x + y\|^2 - \|x - y\|^2\right)$$

No! A necessary and sufficient condition for a norm to correspond to an inner product is the paralellogram identity:

$$\frac{1}{2}\left(\|x + y\|^2 + \|x - y\|^2\right) = \|x\|^2 + \|y\|^2$$



(Jordan and von Neumann, 1935)

L_p norms, inner products

$\|f\|_{L_p} = \left(\int |f|^p\right)^{1/p}$, where $|f|^p$ is integrable

$\|f\|^2 = \left(\int |f|^p\right)^{2/p} = ? = \langle f, f\rangle$

E.g. try $\langle f, g\rangle \doteq \left(\int |fg|^{p/2}\right)^{2/p}$: then $\langle \lambda f, g\rangle = \lambda\langle f, g\rangle$ but

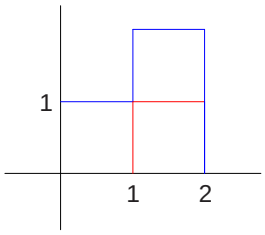
$\langle \alpha_1 f_1 + \alpha_2 f_2, g\rangle = \left(\int |(\alpha_1 f_1 + \alpha_2 f_2)g|^{p/2}\right)^{2/p} \neq \alpha_1\langle f_1, g\rangle + \alpha_2\langle f_2, g\rangle$

unless $p = 2$.

L_p norms, inner products cont.

Maybe we could find some other inner product that works for all $p \geq 1$?

No: if a norm is derivable from an inner product (over $\mathbb{R}$), then $\langle x, y\rangle = \frac{1}{4}\left(\|x+y\|^2 - \|x-y\|^2\right)$. Choose two functions:



$f_1(x) = 0 : x < 0$
$f_1(x) = 1 : 0 \leq x < 1$
$f_1(x) = 2 : 1 \leq x \leq 2$
$f_1(x) = 0 : x > 2$

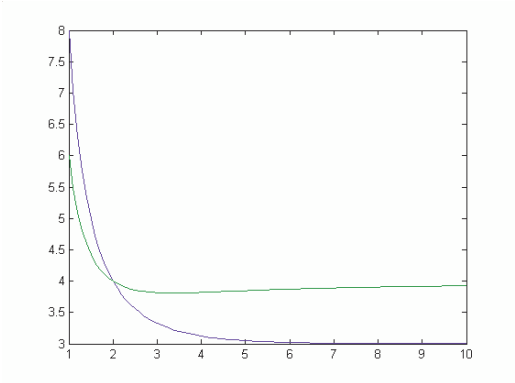
$f_2 = 0 : x < 1$
$f_2 = 1 : 1 \leq x < 2$
$f_2 = 0 : x > 2$

L_p norms, inner products cont.

Then:

$$4\langle \lambda f_1, f_2\rangle = \left(1+|2\lambda+1|^p\right)^{2/p} - \left(1+|2\lambda-1|^p\right)^{2/p}$$, and

$$4\lambda\langle f_1, f_2\rangle = \lambda\left(\left(1+3^p\right)^{2/p}\right) - 2^{2/p}$$



ℓ_∞ norm on $\mathbb{R}^n$ has no inner product

Example in $\mathbb{R}^2$: $x=[1,0],\ y=[0,2]$

Use parallelogram test:

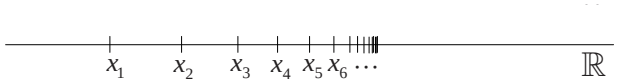
$$\frac{1}{2}\left(\|x+y\|^2+\|x-y\|^2\right) = ? = \|x\|^2+\|y\|^2$$

$$\frac{1}{2}\left(\|x+y\|^2+\|x-y\|^2\right) = \frac{1}{2}(2^2+2^2) = 4 \neq \|x\|^2+\|y\|^2 = 1+4 = 5$$

Extend to $\mathbb{R}^n$: $x=[1,0,0,\cdots,0],\ y=[0,2,0,\cdots,0]$

What is a Cauchy Sequence?

A sequence of vectors $\{x_n\}$ in a normed vector space is called a Cauchy sequence if for every ε>0 there exists a number M such that $\|x_m - x_n\| < \varepsilon$ for all $m, n > M$.



Key idea: the Cauchy sequence allows us to define notions of convergence *without ever leaving the space*

Cauchy sequences, cont.

Every convergent sequence is a Cauchy sequence.

Not every Cauchy sequence is a convergent sequence.

$\mathfrak{P}(0,1)$: the space of polynomials on $[0,1]$. Choose the l_∞ norm $\|P\| = \max_{[0,1]} |P(x)|$. Define

$$P_n(x) = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots + \frac{x^n}{n!}$$

for $n = 1, 2, \cdots$. Then $\{P_n\}$ is a Cauchy sequence but it does not converge in $\mathfrak{P}(0,1)$ because its limit is not a polynomial.

Note Cauchy sequence requires choice of norm!

The notion of completeness

A normed vector space E is called complete if every Cauchy sequence in E converges to an element of E.

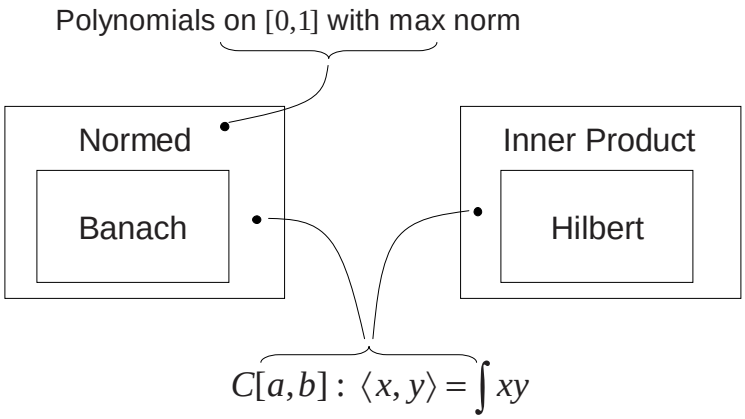
A complete normed space is called a Banach space.

$\mathbb{R}^n$ with l_p norm is complete, for all $1 \le p \le \infty$.
The sequence space l_p is complete, for all $1 \le p \le \infty$.
$C[a,b]$ with L_∞ norm is complete.
$C[a,b]$ with L_2 norm or L_1 norm is not complete.
"All square integrable functions with L_2 norm" is complete.

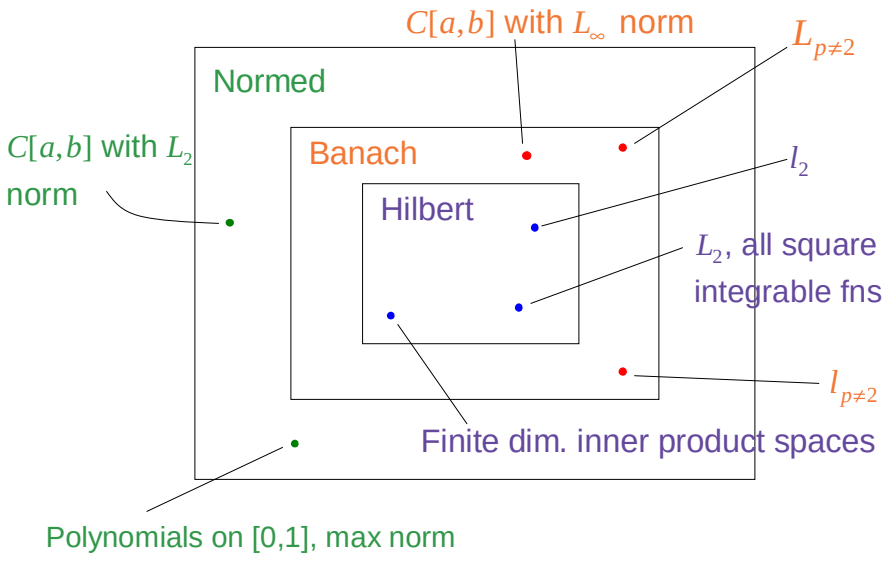
A complete space "contains all its limit points."
It is always possible to 'complete' a non-complete space.

Hilbert and Banach spaces

A Hilbert space is a complete inner product space.



Hilbert and Banach spaces, cont.



How many kinds of Hilbert spaces are there?

A mapping $T : E_1 \to E_2$, where $E_{1,2}$ are vector spaces, is called a linear mapping if $T(\alpha x + \beta y) = \alpha T(x) + \beta T(y) \ \forall \ x, y \in E_1$ and all scalars α, β.

A Hilbert space H_1 is said to be isomorphic to a Hilbert space H_2 if there exists a 1-1 linear mapping T from H_1 to H_2 such that

$$\langle T(x), T(y) \rangle = \langle x, y \rangle$$

for every $x, y \in H_1$. Such a T is called a Hilbert space isomorphism of H_1 onto H_2.

How many kinds of Hilbert spaces are there?

A Hilbert space H is called separable if it contains a complete orthonormal sequence. (Finite dimensional Hilbert spaces are considered separable).

E.g. l_2, $L_2[a,b]$ are separable Hilbert spaces.

Answer: 2...

Let H be separable:

If H is infinite dimensional, then it is isomorphic to l_2.

If H has dimension N, then it is isomorphic to $\mathbb{C}^N$.

The Riesz Representation Theorem

A linear functional on a normed vector space $\{V, F\}$ is a linear mapping $\phi : V \to F$.

The operator norm of a linear functional f is defined:

$$\|f\| \doteq \sup_{\|x\|=1} |f(x)|$$

A linear functional on a normed vector space is bounded ($\exists\ K\ s.t.\ |f(x)| \le K\|x\|\ \ \forall\ x \in V$) if and only if its operator norm is finite.

The Riesz Representation Theorem

Let f be a bounded linear functional on a Hilbert space H. Then there exists exactly one $x_0 \in H$ such that $f(x) = \langle x, x_0 \rangle$ for all $x \in H$, and in fact $\|f\| = \|x_0\|$.

Example: $H = L_2[a,b], \quad -\infty < a < b < \infty$. Define a linear functional by

$$f(x) \doteq \int_a^b x(t)dt$$

$$\textit{Linear?} \quad f(\lambda x + \mu y) = \int_a^b \lambda x(t) + \mu y(t)dt = \lambda \int_a^b x(t)dt + \mu \int_a^b y(t)dt = \lambda f(x) + \mu f(y)$$

$$\textit{Bounded?} \quad |f(x)| = \left| \int_a^b x(t)dt \right| \le \int_a^b |x(t)|\,dt = \int_a^b |x(t)1|\,dt$$

$$\le \left(\int_a^b x(t)^2 dt \right)^{\frac{1}{2}} \left(\int_a^b 1dt \right)^{\frac{1}{2}} = \sqrt{b-a}\,\|x\|$$

Riesz Representation Theorem, cont.

Can we find x_0? Try $x_0 \doteq 1$: $\langle x, 1 \rangle = \int_a^b x(t).1dt = \int_a^b x(t)dt = f(x)$

Check: $\|f\| = \|x_0\|$?

$$\|x_0\| = \left(\int_a^b 1^2 \right)^{1/2} = \sqrt{b-a}$$

$$\|f\| = \sup_{\|x\|=1} |f(x)| = \sup_{\|x\|=1} \left| \int_a^b x(t)dt \right| = \sup \left\{ \left| \int_a^b x(t)dt \right| : \int_a^b x(t)^2 dt = 1 \right\}$$

The sup is found by choosing $x = 1/\sqrt{b-a}, \; a \le t \le b, \; x = 0$ otherwise.

$\Rightarrow \|f\| = \sqrt{b-a}$

What is a metric space?

For any set E, let $\rho(x, y)$ be a function (with range in $\mathbb{R}$) defined on the set $E \times E$ of all ordered pairs (x, y) of members of E, satisfying:

(i) $\rho(x, y)$ is a finite real number for every pair (x, y) of $E \times E$;
(ii) $\rho(x, y) = 0 \Leftrightarrow x = y$;
(iii) $\rho(y, z) \leq \rho(x, y) + \rho(x, z), \ \{x, y, z\} \in E$.

Such a function $\rho : E \times E \to \mathbb{R}$ is called a metric on E; a set E with metric ρ is called a metric space. Different choices of metric on E give different metric spaces.

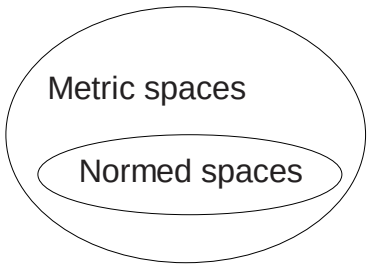
Puzzle: What about $\rho(x, y) \geq 0$?

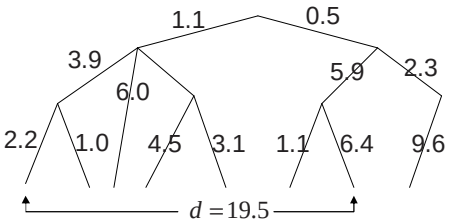
Puzzle: How about $\rho(x, y) = \rho(y, x)$?

Puzzle: Where's the triangle inequality: $\rho(x, y) \leq \rho(x, z) + \rho(z, y)$?

Metric versus norm

Every normed vector space is a metric space: define $\rho(x, y) \doteq \|x - y\|$
But metric spaces are much more general:





Metrics extend "continuity": $f(B_\delta(x)) \subset B_\varepsilon f(x)$

Some metrics on function spaces

Let A be the set of all bounded functions $f:[a,b]\to\mathbb{R}$. For two points $f,g\in A$, $\rho_\infty(f,g)\doteq \sup_{x\in[a,b]}|f(x)-g(x)|$ is a metric.

Let A be $C[a,b]$: $\rho_1(f,g)\doteq\int_a^b|f(x)-g(x)|dx$,

$\rho_2(f,g)\doteq\left(\int_a^b(f(x)-g(x))^2\,dx\right)^{\frac{1}{2}}$, $\rho_p(f,g)=?$

Suppose instead $A=C^r[a,b]$: then

$$\rho_{\infty,r}(f,g)=\sup_{x\in[a,b]}\left\{|f(x)-g(x)|,|f'(x)-g'(x)|,\cdots,\left|f^{(r)}(x)-g^{(r)}(x)\right|\right\}$$

Topological Spaces

A topological space $T=\{A,\mathcal{S}\}$: A is a non-empty set, $\mathcal{S}$ a fixed collection of subsets of A, satisfying

(1) $A,\varnothing\in\mathcal{S}$,

(2) Intersection of any two sets in $\mathcal{S}$ is in $\mathcal{S}$,

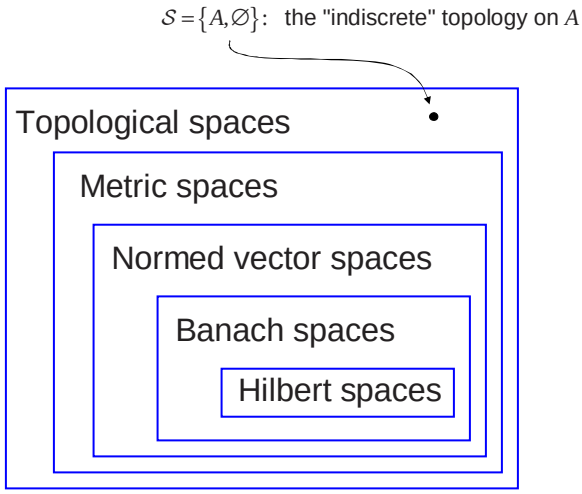
(3) Union of any collection of sets in $\mathcal{S}$ is in $\mathcal{S}$.

$\mathcal{S}$ is called a topology for A, and the members of $\mathcal{S}$ are called the open sets of T.

Topological spaces are more general than metric spaces!

Topological spaces extend "continuity": Given $T_1=\{A_1,\mathcal{S}_1\}$ and $T_2=\{A_2,\mathcal{S}_2\}$ and a map $\phi:A_1\to A_2$, ϕ is "continuous" if $U\in A_2\Rightarrow\phi^{-1}(U)\in A_1$

Putting Spaces in their Places…



Some Mathematical Tools for Machine Learning

Chris Burges
Microsoft Research

Max Planck Institute, Tübingen
August, 2003

Contents – Part 2

A dual basis

Orthonormal basis in $\mathbb{R}^n$: $e^{(a)}, a = 1, \cdots, n$: $e^{(a)} \cdot e^{(b)} = \delta_{ab}$

$e^{(a)}$ has components $e_i^{(a)}$, $i = 1, \cdots, n.$ Construct n new vectors: $\tilde{e}_i^{(a)} \doteq e_a^{(i)}$

Do the $\tilde{e}$ necessarily also form an orthonormal basis?

$$M_{ai} \doteq e_a^{(i)} = \begin{bmatrix} 0.8602 & 0.4155 & -0.2955 \\ -0.0619 & 0.6604 & 0.7483 \\ 0.5061 & -0.6254 & 0.5938 \end{bmatrix}_{ai}$$

$\tilde{e}_i^{(a)} \tilde{e}_i^{(b)} = e_a^{(i)} e_b^{(i)} = M_{ai} M_{bi} = (MM^T)_{ab}$. But $M^T M = \mathrm{I}$. M is full rank $\Rightarrow$ has inverse: $M^T M M^{-1} = M^{-1} = M^T$. So $MM^T = \mathrm{I}$.

E.g. $\cdots \sum_a e_i^{(a)} e_j^{(a)} \Lambda(i, j) \cdots = \Lambda(i, j)\, \delta_{ij}$

Matrix Mechanics: Products as Sums

Usual matrix multiplication: $X \in M_{mn},\ Y \in M_{np},\quad XY \in M_{mp}$

$(XY)_{ab} \doteq \sum_{i=1}^{n} X_{ai} Y_{ib}$ (Summand: numbers)

Let x_i be the i^{th} column of X and y_i^T be the i^{th} row of Y.

Then also $XY = \sum_{i=1}^{n} x_i y_i^T$ (Summand: matrices)

$$\begin{pmatrix} a & b & c \\ d & e & f \end{pmatrix}\begin{pmatrix} g & h \\ i & j \\ k & l \end{pmatrix} = \begin{pmatrix} a \\ d \end{pmatrix}\begin{pmatrix} g & h \end{pmatrix} + \begin{pmatrix} b \\ e \end{pmatrix}\begin{pmatrix} i & j \end{pmatrix} + \begin{pmatrix} c \\ f \end{pmatrix}\begin{pmatrix} k & l \end{pmatrix}$$

$$= \left\{\begin{pmatrix} a & 0 & 0 \\ d & 0 & 0 \end{pmatrix} + \begin{pmatrix} 0 & b & 0 \\ 0 & e & 0 \end{pmatrix} + \begin{pmatrix} 0 & 0 & c \\ 0 & 0 & f \end{pmatrix}\right\} \times \left\{\begin{pmatrix} g & h \\ 0 & 0 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ i & j \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 0 & 0 \\ k & l \end{pmatrix}\right\}$$

A Matrix Multiplication Invariant

$$AB = \begin{pmatrix} a & b & c \\ d & e & f \end{pmatrix}\begin{pmatrix} g & h \\ i & j \\ k & l \end{pmatrix} \in M_{22}$$

$$AB = \begin{pmatrix} a & b & c & x \\ d & e & f & y \end{pmatrix}\begin{pmatrix} g & h \\ i & j \\ k & l \\ 0 & 0 \end{pmatrix}$$

$$AB = \begin{pmatrix} a & b & c & 0 \\ d & e & f & 0 \end{pmatrix}\begin{pmatrix} g & h \\ i & j \\ k & l \\ x & y \end{pmatrix}$$

Totally Antisymmetric Shortcuts

(Just a quick reminder...)

$(a \wedge b) \cdot (c \wedge d) = ?$

$\varepsilon_{ijk}\varepsilon_{ilm} = \delta_{jl}\delta_{km} - \delta_{jm}\delta_{kl}$

$$(a \wedge b) \cdot (c \wedge d) = \varepsilon_{ijk} a_j b_k \varepsilon_{ilm} c_l d_m = (\delta_{jl}\delta_{km} - \delta_{jm}\delta_{kl}) a_j b_k c_l d_m$$
$$= (a \cdot c)(b \cdot d) - (a \cdot d)(b \cdot c)$$

Characterizing the Determinant

The determinant of a square matrix $A \in M_{nn}$ can be defined as:

$$|A| \doteq \frac{1}{n!} \varepsilon_{\alpha_1 \alpha_2 \cdots \alpha_n} \varepsilon_{\beta_1 \beta_2 \cdots \beta_n} A_{\alpha_1 \beta_1} A_{\alpha_2 \beta_2} \cdots A_{\alpha_n \beta_n}$$

Corollary: The determinant $|A|$ can also be represented as:

$$|A| \doteq \varepsilon_{\alpha_1 \alpha_2 \cdots \alpha_n} A_{1\alpha_1} A_{2\alpha_2} \cdots A_{n\alpha_n}$$ (Puzzle: Prove this!)

Theorem: The determinant and inverse are related as follows:

$$\frac{\partial \ln |A|}{\partial A_{ij}} = A^{-1}_{ji}$$

Proof: $\frac{\partial |A|}{\partial A_{ij}} = \varepsilon_{j\alpha_2 \cdots \alpha_n} \delta_{i1} A_{2a_2} \cdots A_{n\alpha_n} + \varepsilon_{\alpha_1 j \alpha_3 \cdots \alpha_n} A_{1\alpha_1} \delta_{i2} A_{3\alpha_3} \cdots A_{n\alpha_n} + \cdots$

$$A_{kj} \frac{\partial |A|}{\partial A_{ij}} = \varepsilon_{\alpha_1 \alpha_2 \cdots \alpha_n} (A_{k\alpha_1} \delta_{i1} A_{2\alpha_2} \cdots A_{n\alpha_n} + A_{1\alpha_i} A_{k\alpha_2} \delta_{i2} A_{3\alpha_3} \cdots + \cdots) = |A| \delta_{ki}$$

Characterizing the Inverse

The inverse of a square matrix $A \in M_{nn}$ can be defined by

$$A^{-1}_{ij} \doteq \frac{1}{|A|(n-1)!} \varepsilon_{j\alpha_1\alpha_2\cdots\alpha_{n-1}} \varepsilon_{i\beta_1\beta_2\cdots\beta_{n-1}} A_{\alpha_1\beta_1} A_{\alpha_2\beta_2} \cdots A_{\alpha_{n-1}\beta_{n-1}}$$

Puzzle: Check this by using $\frac{\partial \ln|A|}{\partial A_{ij}} = \cdots = A^{-1}_{ji}$

Theorem: For an arbitrary non-singular square matrix A,

$$\frac{\partial A^{-1}_{ij}}{\partial A_{\alpha\beta}} = -A^{-1}_{i\alpha} A^{-1}_{\beta j}$$

(in proof, only need to use $A^{-1}A = const$).

Taking Derivatives of Gaussians

$$p(x) \propto |\Sigma|^{-1/2} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right)$$

Assume IID: $p(x_1 x_2 \cdots x_n \mid \mu, \Sigma) = \prod_i p(x_i \mid \mu, \Sigma)$

$$\frac{\partial p}{\partial \mu} = \sum_i \Sigma^{-1}(x_i - \mu) p(x_1 x_2 \cdots x_n \mid \mu, \Sigma) = 0 \rightarrow \mu = \frac{1}{n}\sum_i x_i$$

How about the covariance?

$$\frac{\partial p(x_1 \cdots x_n \mid \mu, \Sigma)}{\partial \Sigma_{ab}} = 0 \rightarrow n\frac{\partial \log|\Sigma|}{\partial \Sigma_{ab}} = -\sum_{i,\alpha\beta} (x_i - \mu)_\alpha \frac{\partial \Sigma^{-1}_{\alpha\beta}}{\partial \Sigma_{ab}} (x_i - \mu)_\beta$$

$$n\Sigma^{-1}_{ab} = \sum_{i,\alpha\beta} \Sigma^{-1}_{a\alpha}\Sigma^{-1}_{\beta b}(x_i - \mu)_\alpha (x_i - \mu)_\beta \rightarrow \Sigma = \frac{1}{n}\sum_i (x_i - \mu)(x_i - \mu)^T$$

SVD in Seven Steps

Singular Value Decomposition: A generalization of eigendecomposition

$A \in M_{mn}$: Assume $m \le n$.

Key idea: focus on $AA^T \in M_{mm}$ and $A^T A \in M_{nn}$.
Let $AA^T x_i = \sigma_i^2 x_i$, $x_i \in \mathbb{R}^m$, $\sigma_i^2 > 0$, $i = 1, \cdots, k$

(1) AA^T *has the same nonzero eigenvalues as* $A^T A$
Let $y_i \doteq (1/\sigma_i)(A^T x_i)$ for $i : \sigma_i > 0$: $y \in \mathbb{R}^n$
$\Rightarrow A^T A y_i = (1/\sigma_i) A^T A A^T x_i = \sigma_i A^T x_i = \sigma_i^2 y_i$

(2) x_i *can be chosen orthonormal* $\Rightarrow$ y_i *also orthonormal*

(3) *Let* A *have rank* k. *Then* $rank(A) = rank(A^T) = rank(AA^T) = rank(A^T A) = k$

SVD in Seven Steps, cont.

(4) Let X (Y) be the matrix of cols of x_i (y_i): $X \in M_{mk}$, $Y \in M_{nk}$
$Y = A^T X \, diag(1/\sigma_i) \Rightarrow diag(\sigma_i) Y^T = X^T A$: If $m = k$, then $A = X \, diag(\sigma_i) Y^T$

(5) If $m < k$, add $m - k$ rows of orthonormal null vectors of A^T to bottom of X^T, and $m - k$ zero rows to bottom of $diag(\sigma_i)$. Then X is unitary and
$A = X \, diag(\sigma_i, 0) Y^T$, $X \in M_{mm}$, $diag(\sigma_i, 0) \in M_{mk}$, $Y \in M_{nk}$

(6) Want something closer to eigendecomposition? Add $n - k$ rows of orthonormal vectors (to the y_i) to bottom of Y^T and $n - k$ cols of zeros to right of $diag(\sigma_i, 0)$. Then Y also unitary and
$A = X \, diag(\sigma_i, 0, 0) Y^T$, $X \in M_{mm}$, $diag(\sigma_i, 0, 0) \in M_{mn}$, $Y \in M_{nn}$

(7) Want something closer to the sum of outer products? After (5), remove extra rows of X^T and diag(σ_i): then $A = \sum_{i=1}^{k} \sigma_i x_i y_i^T$

SVD: Moore-Penrose Generalized Inverse

Suppose $A \in M_{mm}$ is diagonalizable: $AE = E\Lambda$, or $A = E\Lambda E^T$.

Suppose further A is nonsingular. Then $A^{-1} = E\Lambda^{-1}E^T$.

Is there an SVD-like generalization of the inverse for arbitrary matrices?

$A = X\ diag(\sigma_i, 0, 0)Y^T$, $A \in M_{mn}$, $X \in U_m$, $Y \in U_n$, $diag(\sigma_i, 0, 0) \in M_{mn}$

Define $A^\dagger \doteq Y\ diag(1/\sigma_i, 0, 0)^T\ X^T \in M_{nm}$

Outer product form: $A^\dagger = \sum_{i=1}^{k} (1/\sigma_i) y_i x_i^T$

SVD: Moore-Penrose Generalized Inverse

Then:

(a) $AA^\dagger$ and $A^\dagger A$ are Hermitian;
(b) $AA^\dagger A = A$
(c) $A^\dagger AA^\dagger = A^\dagger$

In fact $A^\dagger$ is uniquely determined by (a), (b) and (c) as requirements.

If A is square and nonsingular, then $A^\dagger = A^{-1}$

If $(A^T A)^{-1}$ exists, then $A^\dagger = (A^T A)^{-1} A^T$

If $(AA^T)^{-1}$ exists, then $A^\dagger = A^T (AA^T)^{-1}$

Range and Null Space

$A \in M_{mn}$: A's 'range' $R \subset \mathbb{R}^m$ is spanned by $y = Ax$ (for all $x \in \mathbb{R}^n$)

A's 'null space' $N \subset \mathbb{R}^n$ is spanned by those x for which $Ax = 0$

Let $A_{\cdot i}$ denote columns of A. Then:

$$y = Ax = \begin{pmatrix} \vdots & & \vdots \\ A_{\cdot 1} & \cdots & A_{\cdot n} \\ \vdots & & \vdots \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = x_1 A_{\cdot 1} + \cdots + x_n A_{\cdot n}$$

$\rightarrow$ $\dim(range) = rank(A)$ and R is spanned by columns of A

$null\ space(A^T)$ spanned by those vectors orthog. to every row of A^T
(*i.e.* those vectors orthog to every column of A)

$\rightarrow$ $range(A) = orthogonal\ complement\ of\ null(A^T)$

Range and Null Space, cont.

$\dim(null(A)) =$ # lin. inδ^t vectors orthog. to every row of A
(recall, the row vectors in A are in $\mathbb{R}^n$):
$\dim(null(A)) = n - rank(A)$: $\dim(range(A)) + \dim(null(A)) = n$

In fact, using $A = \sum_{i=1}^{k} \sigma_i x_i y_i^T$:

The null space N of A is the subspace orthogonal to the k y_i, so $\dim(N) = n - k$.

The range R of A is spanned by the x_i, so $\dim(R) = k$

SVD: Characterizing Linear Maps

$Az = b,\ A \in M_{mn},\ z \in \mathbb{R}^n, b \in \mathbb{R}^m$ has $0, 1$ or ∞ many solutions.
(If z_1 and z_2 are solutions, so is $\alpha_1 z_1 + \alpha_2 z_2$).

When does a solution exist? Az is a linear combination of the columns of A, so b must lie in the span of the columns of A.

In fact $b \in R \ \rightarrow\ z_0 = A^\dagger b$ is a solution: $Az_0 = \sum_{i=1}^{k} \sigma_i x_i y_i^T \sum_{j=1}^{k} (1/\sigma_i) y_j x_j^T b$
$= \sum_{i=1}^{k} x_i x_i^T b = b,$ and the general solution is $z = A^\dagger b + N$

What if $b \notin R \ \rightarrow Az = b$ has no solution? (E.g. numerical problems!)
Then: minimize $\|z\|$ given that $\|Az - b\|$ is minimized. Solution: $z = A^\dagger b$.

SVD: The Unconstrained QP

Suppose you wish to minimize:

$$f(x) \doteq \frac{1}{2} xAx + bx, \quad x \in \mathbb{R}^n, \quad A \succeq 0.$$

Solution: $\nabla f = 0 \ \rightarrow\ Ax + b = 0$

If $b \notin range(A)$, no solution.
Puzzle: what happens if you try to minimize f anyway?

Else, the general solution is $x = A^\dagger b + N$

SVD and Matrix Norms

A function $\|\cdot\|: M_{mn} \to \mathbb{R}$ is a matrix norm if for all $A, B \in M_{mn}$,

(a) $\|A\| \geq 0$

(b) $\|A\| = 0 \Leftrightarrow A = 0$

(c) $\|cA\| = |c|\,\|A\|$ for all scalars $c \in F$

(d) $\|A+B\| \leq \|A\| + \|B\|$

SVD and Matrix Norms, cont.

A vector norm induces a matrix norm: $\|A\|_p \doteq \max_{\|x\|_p=1} \|Ax\|_p$

Frobenius (Hilbert - Schmidt) norm: $\|A\|_F \doteq \sqrt{\sum_{ij} |A_{ij}|^2} = \sqrt{Tr\left(A^T A\right)} = \sqrt{\sum_i \|Ae_i\|_2^2}$

Maximum Absolute Column Norm: $\|A\|_1 = \max_j \sum_i |A_{ij}|$

Maximum Absolute Row Norm: $\|A\|_\infty = \max_i \sum_j |A_{ij}|$

Spectral Norm: $\|A\|_2 = \sigma_1$

Also: $\|A\|_F = \sqrt{\sigma_1^2 + \cdots + \sigma_k^2}$: $\min_{x \neq 0} \frac{\|Ax\|_2}{\|x\|_2} = \sigma_m$

Positive (Semi)-Definite Matrices

Positive (semi)-definite $A: A \in S_n$ and $x^T Ax \geq 0$ (strict for $+ve\ def$)

Useful properties:

- All diag elements ≥ 0 for pos semidef, > 0 for pos def
- $A_{ii} = 1\ \forall i \Rightarrow |A_{ij}| \leq 1$
- A pos. semidef and rank one $\Leftrightarrow A = xx^T$ for some x
- If A, B pos semidef., then so is AB
- A symmetric, and all leading principal minors $> 0 \Leftrightarrow A$ pos def
- If symmetric, strictly diagonally dominant ($|A_{ii}| > \sum_{j \neq i} |A_{ij}|\ \forall i$), and $A_{ii} > 0$, then pos def.

$$\begin{pmatrix} \bullet & & \\ & \bullet & \\ & 0 & \\ & & \bullet \end{pmatrix} \Rightarrow \begin{pmatrix} \bullet & & 0 & \\ & \bullet & 0 & \\ 0 & 0 & 0 & 0 \\ & & 0 & \bullet \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} \bullet & & \\ y & & x \\ & \bullet & \\ & & z \end{pmatrix} \Rightarrow |x| \leq \sqrt{|yz|}$$

Gram Matrices

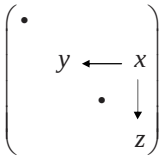
Let V be a vector space over some field, with inner product. The *Gram matrix* of a set of vectors $v_i \in V,\ i = 1, \cdots, n$ is defined: $G_{ij} \doteq \langle v_i, v_j \rangle$

Let $A \in S_n$. Then A is positive semidefinite with rank $r \leq n$ if and only if $\exists \{v_1, \cdots, v_n\},\ v_i \in V$ containing exactly r independent vectors such that $A_{ij} = v_i \cdot v_j$

Simple proofs!

$$\begin{pmatrix} \bullet & & \\ & \bullet & \\ & 0 & \\ & & \bullet \end{pmatrix} \Rightarrow \begin{pmatrix} \bullet & & 0 & \\ & \bullet & 0 & \\ 0 & 0 & 0 & 0 \\ & & 0 & \bullet \end{pmatrix} \quad : \quad \|x_i\|^2 = 0 \Rightarrow x_i \cdot x_j = 0$$

Gram Matrices, cont.



$\Rightarrow |x| \le \sqrt{|yz|}$: $\sqrt{|yz|} \ge \left|\sqrt{|yz|}\cos(\alpha)\right| \doteq |x|$

Suppose you're given a positive semidefinite matrix A. How to extract the Gram vectors v_i?

$A = E\Lambda E^T$; $E_{ij} = e_i^{(j)}$; $A_{ij} = E_{ik}\lambda_k\delta_{km}E_{jm} = e_i^{(k)}\lambda_k e_j^{(k)}$

Define $c_i^{(j)} \doteq \sqrt{\lambda_j}e_i^{(j)}$: then $A_{ij} = c_i^{(k)}c_j^{(k)} = \tilde{c}_k^{(i)}\tilde{c}_k^{(j)} = \tilde{c}^{(i)}\cdot\tilde{c}^{(j)}$

The Gram vectors are the dual basis to the scaled eigenvectors

Euclidean Distance Matrices

A 'Euclidean distance matrix' has the form: $D_{ij} \doteq \left\|x_i - x_j\right\|_2^2$

When is a symmetric matrix a Euclidean distance matrix?

$D \in S_n$ is a Euclidean distance matrix, for some points in $\mathbb{R}^k$, for some k, if and only if $D_{ii} = 0 \quad \forall i$, and for all y satisfying $1\cdot y = 0$, $y^T Dy \le 0$ *(Schoenberg 1935, Young and Householder 1938)*

E.g. $D_{ij} = \left\|x_i\right\|^2 + \left\|x_j\right\|^2 - 2x_i\cdot x_j$:

$$y_i D_{ij} y_j = \sum_i y_i \left\|x_i\right\|^2 \sum_j y_j + \sum_j y_j \left\|x_j\right\|^2 \sum_i y_i - 2\sum_{i,j} y_i (x_i \cdot x_j) y_j$$

Convex Optimization: A Brief Tour

Bibliography

- *Convex Optimization*, Boyd and Vandenberghe, Cambridge University Press, to appear 2003, also http://www.stanford.edu/~boyd/cvxbook.html (highly recommended)
- *Convex Analysis*, Rockafeller, Princeton University Press, 1970
- *Nonlinear programming*, Mangasarian, McGraw Hill, 1969
- *Practical Methods of Optimization*, Fletcher, Wiley 1987 (2nd Ed.)
- *Numerical Recipes in C++,* Press et. al, Cambridge University Press, 2002
- *Linear Programming*, Chvatal,W.H. Freeman and Co., 1980

Convexity, and some Definitions

Some Named Sets

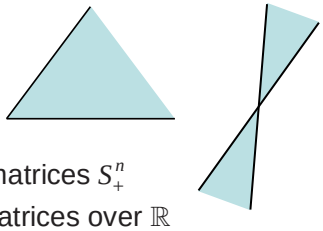
Affine set:

$\mathbb{S}_A \doteq \sum_{i=1}^{p} \theta_i x_i, \quad x_i \in \mathbb{R}^n, \quad \sum_{i=1}^{p} \theta_i = 1;$ choosing an origin gives subspace

Convex set:

$\mathbb{S}_C \doteq \sum_{i=1}^{p} \theta_i x_i, \quad x_i \in \mathbb{R}^n, \quad \sum_{i=1}^{p} \theta_i = 1, \quad \theta_i \geq 0$

Cone: If $x \in \mathbb{S}_O$ then so is $\theta x, \ \theta \geq 0.$



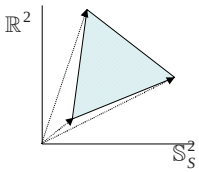
A cone not in $\mathbb{R}^n$: positive semidefinite matrices S_+^n
as a subset of the vector space of real matrices over $\mathbb{R}$

Some Named Sets, cont.

$x_0, x_1, \cdots, x_n$ are affinely independent if $x_1 - x_0, \; x_2 - x_0, \; \cdots, \; x_n - x_0$ are linearly independent. The x_i are then sometimes called 'in general position'.

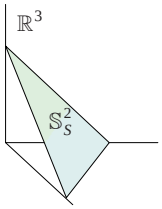
An n-dimensional simplex is just the set of convex combinations of a given set of affinely independent vectors:

$$\mathbb{S}_S^n \doteq \left\{ \sum_{i=0}^{n} \theta_i x_i : \; \theta_i \geq 0, \; \sum_{i=0}^{n} \theta_i = 1 \right\}$$



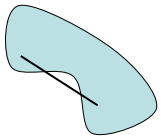
For the unit or regular n-simplex, convenient to embed $\mathbb{S}_S^n$ in $\mathbb{R}^{n+1}$: can take x_i to be orthonormal.

For regular n-simplex, the distance between any two vertices is $\sqrt{2}$.



Convexity: some definitions

A convex set S is a subset of a vector space such that if $x_1, \; x_2 \in S$ then $\theta_1 x_1 + (1 - \theta_2) x_2 \in S$



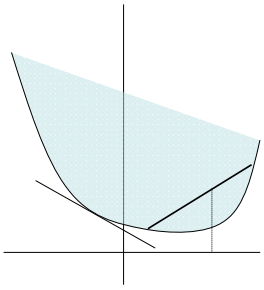
A convex set that is not in $\mathbb{R}^n$: positive semidefinite matrices over $\mathbb{R}$, S_+^n : if $S_1, \; S_2 \in S_+^n$ and $x \in \mathbb{R}^n$ then $x'(\theta S_1 + (1-\theta) S_2) x \geq 0$

Three equivalent definitions of a convex function (always require $domain(f)$ to be convex)

(1) $f : \mathbb{R}^n \to \mathbb{R} : \; f((1-\lambda)x_1 + \lambda x_2) \leq (1-\lambda) f(x_1) + \lambda f(x_2)$

(2) $f(y) \geq f(x) + \nabla f(x)'(y - x) \quad \forall x, y \in domain(f)$

(3) $\nabla^2 f(x) \succeq 0 \quad \forall x \in domain(f)$



Convexity, cont.

Recall f is strictly convex iff $\forall x_1 \neq x_2$,

$$f((1-\lambda)x_1 + \lambda x_2) < (1-\lambda) f(x_1) + \lambda f(x_2), \quad 0 \leq \lambda \leq 1$$

If $domain(f)$ is convex and $\nabla^2 f \succ 0$, then f is strictly convex.

However if f is strictly convex, $\nabla^2 f$ may not be positive definite (e.g. x^4)

But if f is quadratic, then $\nabla^2 f \succ 0 \Leftrightarrow f$ is strictly convex.

Examples:

Every norm is convex:

$$\|(1-\lambda)x_1 + \lambda x_2\| \leq |1-\lambda|\|x_1\| + |\lambda|\|x_2\| = (1-\lambda)\|x_1\| + \lambda\|x_2\|$$

Puzzle: Are there any norms which are strictly convex?

Puzzle: Show that for any distribution, $E[f(x)] \geq f(E[x])$ if f is convex.

Convexity: Examples, cont.

Some convex fns: $\exp(ax)$, $|x|^p$ for $p \geq 1$ on $\mathbb{R}$, $-\log(x)$, $x\log(x)$ on $\mathbb{R}_{++}$

Powers: x^a is convex for $a \geq 1$ or $a \leq 0$, and concave for $0 \leq a \leq 1$, on $\mathbb{R}_{++}$.

Some convex functions on $\mathbb{R}^n$:

Max: $f(x) = \max(x_1, \cdots, x_n)$

Log-sum-exp: $f(x) = \log(e^{x_1} + e^{x_2} + \cdots + e^{x_n})$ (smooth, convex approx to max)

Negative geometric mean (on $\mathbb{R}^n_{++}$): $f(x) = -\left(\prod_{i=1}^n x_i\right)^{1/n}$

$P(x_i|\alpha_i) = e^{\alpha_i x_i} / \sum_{j=1}^n e^{\alpha_j x_j}$ not, but $-\log(P(x_i \mid \alpha_i))$ is (hint: $E[v^2] \geq (E[v])^2$)

Some convex functions on matrices:

$\log(\det(X^{-1}))$, $X \succ 0$ (strict)

$Trace(X^{-1})$, $X \succ 0$ (strict)

Max. eigenvalue of a symmetric matrix

Proving Interesting Inequalities

The multi-λ version of $f((1-\lambda)x_1+\lambda x_2)\le(1-\lambda)f(x_1)+\lambda f(x_2)$:
For $\lambda_i\ge 0,\ \sum_i \lambda_i = 1,\ f(\lambda_1 x_1+\lambda_2 x_2+\cdots+\lambda_n x_n)\ \le\ \lambda_1 f(x_1)+\lambda_2 f(x_2)+\cdots+\lambda_n f(x_n)$

The geometric / arithmetic mean inequality: use the fact that $-\log(x)$ is convex:

$$-\log\left(\sum_{i=1}^n \frac{x_i}{n}\right)\le -\sum_{i=1}^n \frac{1}{n}\log(x_i) = -\log\left(\left(\prod_{i=1}^n x_i\right)^{\frac{1}{n}}\right) \Rightarrow \left(\prod_{i=1}^n x_i\right)^{\frac{1}{n}} \le \frac{1}{n}\sum_{i=1}^n x_i$$

Hölder's inequality: $\sum_{i=1}^n x_i y_i \le \left(\sum_{i=1}^n |x_i|^p\right)^{\frac{1}{p}}\left(\sum_{i=1}^n |y_i|^q\right)^{\frac{1}{q}},\quad \frac{1}{p}+\frac{1}{q}=1,\quad p>1$

Choose $0<\lambda \doteq \frac{1}{p}<1$, use $a^\lambda b^{1-\lambda}\le \lambda a+(1-\lambda)b,\quad a \doteq \frac{|x_i|^p}{\sum_{i=1}^n |x_i|^p},\ b \doteq \frac{|y_i|^q}{\sum_{i=1}^n |y_i|^q}$

Proving Interesting Inequalities, cont.

Minkowski's inequality: for $p\ge 1$, $\left(\sum_{i=1}^n |x_i+y_i|^p\right)^{\frac{1}{p}} \le \left(\sum_{i=1}^n |x_i|^p\right)^{\frac{1}{p}} + \left(\sum_{i=1}^n |y_i|^p\right)^{\frac{1}{p}}$

(triangle inequality for the l_p norm)

For $p=1$, use ordinary triangle inequality. For $p>1,\ \exists\ q>0 : \frac{1}{p}+\frac{1}{q}=1$

$$\sum_{i=1}^n |x_i+y_i|^p = \sum_{i=1}^n |x_i+y_i||x_i+y_i|^{p-1} \le \sum_{i=1}^n |x_i||x_i+y_i|^{p-1} + \sum_{i=1}^n |y_i||x_i+y_i|^{p-1}$$
$$\le \left(\sum_{i=1}^n |x_i|^p\right)^{\frac{1}{p}}\left(\sum_{i=1}^n |x_i+y_i|^{q(p-1)}\right)^{\frac{1}{q}} + \left(\sum_{i=1}^n |y_i|^p\right)^{\frac{1}{p}}\left(\sum_{i=1}^n |x_i+y_i|^{q(p-1)}\right)^{\frac{1}{q}}$$

Then since $q(p-1)=1$,

$$\sum_{i=1}^n |x_i+y_i|^p \le \left(\left(\sum_{i=1}^n |x_i|^p\right)^{\frac{1}{p}} + \left(\sum_{i=1}^n |y_i|^p\right)^{\frac{1}{p}}\right)\left(\sum_{i=1}^n |x_i+y_i|^p\right)^{\frac{1}{q}}$$

Convex Optimization

Convex Optimization

Minimize a convex function over a convex set.

Recall if $f(x)$ is convex, then $f(x) \le 0$ defines a convex set:

$$f(x) \le 0, \quad f(y) \le 0 \quad \rightarrow \quad f((1-\lambda)x + \lambda y) \le (1-\lambda) f(x) + \lambda f(y) \le 0$$

The intersection of two convex sets is convex:

Minimize $f(x)$ subject to $c_i(x) \le 0 \ \forall i$,
where f, c_i are convex functions.

Convex Optimization, cont.

For a convex optimization problem, all solutions are global, and the set of global solutions is convex.

Let S be the set of solutions. Suppose $x_* \in S$ is local but not global. Then $\exists\, x_1$ such that $f(x_1) < f(x_*)$. Let $z_\lambda \doteq (1-\lambda)x_* + \lambda x_1$. Then $f(z_\lambda) \leq (1-\lambda) f(x_*) + \lambda f(x_1) < f(x_*)$. Take limit as $\lambda \to 0 \Rightarrow$ local min is not a local min $\Rightarrow$ local solutions are global. Now let $x_0, x_1 \in S$, $z \doteq (1-\lambda)x_0 + \lambda x_1$. Then $f(z) \leq (1-\lambda) f(x_0) + \lambda f(x_1) = f(x_0)$. Hence the set of global solutions is convex.

A little uniqueness theorem

Let $x \in \mathbb{R}^a$ be a set S_x of variables and let $y \in \mathbb{R}^b$ be a second set S_y of variables. Let f be a real valued, strictly convex function defined on the y, and let g be a real valued convex function defined on the x. Let z be a third set S_z of variables defined by $S_z = S_x \bigcup S_y$. Let $c_i(z) \leq 0, \quad i = 1, \cdots, m$ define a convex set on $\mathbb{R}^n$, $n \leq a + b$. Define $F(z) \doteq f(y) + g(x)$. Then all solutions to the problem

Minimize $F(z)$ subject to $c_i(z) \leq 0$

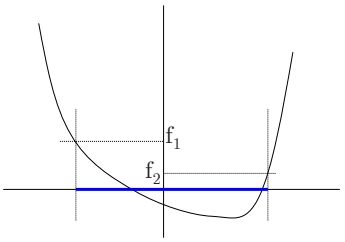
share the same value of y.

E.g. 'w' in SVM and ν-SVM classification / regression estimation is unique.

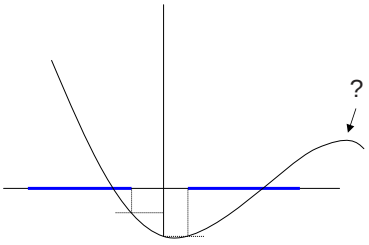
(*Burges & Crisp* 2003)

Stay on The Path

How about "maximize a convex function subject to convex constraints"? (-or, "Minimize a concave function subject to convex constraints"?)



Or maybe, "minimize a convex function over a non-convex set"?



Some simple examples†

For *any* unconstrained optimization problem (minimize f, f convex): Necessary and sufficient condition for solution is $\nabla f = 0$.

Unconstrained quadratic programs over $\mathbb{R}^n$:

$f(x) = \frac{1}{2} x'Hx + b'x + c$

$\nabla f = 0 \Leftrightarrow Hx + b = 0$: use SVD to find complete set of solutions

Convex optimization and norms: can appear in objective function as $\|your\ favorite\ machine\ learning\ norm\|$, and/or in the constraints as $\|another\ norm\| \le const.$

† *see Boyd and Vandenberghe, Convex Optimization, 2003*

Simple Linear Programming Examples

Minimize $c^T x$ subject to $Ax = b, \quad x \in \mathbb{R}^n$

Suppose A has rank k. Recall the SVD of A is $A = \sum_{i=1}^{k} \sigma_i z_i y_i^T$, and that $A^{\dagger} \doteq \sum_{i=1}^{k} (1/\sigma_i) y_i z_i^T$

If $b \notin range(A)$, infeasible. Else, $x = A^{\dagger} b + \sum_{i=1}^{n-k} \alpha_i e_i$, where e_i is any orthonormal set that is orthogonal to the k right singular vectors of A.

If $\langle c, null(A) \rangle \neq 0$, answer is unbounded below. Else, $x = A^{\dagger} b$ (solution doesn't depend explicitly on c!).

Simple LP Examples, cont

Minimize $c^T x$ subject to box constraints: $l \leq x \leq u$

$c_i > 0$: Take $x_i = l$
$c_i = 0$: Take $x_i = anything\ in\ [l, u]$
$c_i < 0$: Take $x_i = u$

Small changes can make it much harder:

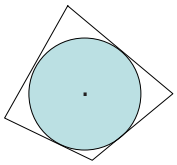
Minimize $x'Hx + b'x$ subject to $l \leq x \leq u, \quad x \cdot c = 0$

Chebyshev center

Largest Euclidean ball that fits inside a polyhedron. Center is farthest internal point from the boundary.

$$\mathcal{P} = \{x \in \mathbb{R}^n \mid a_i \cdot x \le b_i,\ i = 1, \cdots, m\}$$

$$\mathcal{B} = \{x_c + u,\ x_c \in \mathbb{R}^n, u \in \mathbb{R}^n \mid \|u\| \le R\}$$



Maximize R subject to $\mathcal{B} \subseteq \mathcal{P}$. $\mathcal{B}$ lies in halfspace $a_i \cdot x \le b$ if $a_i \cdot (x_c + u) \le b_i$ for all $u : \|u\| \le R$.

Since $\sup\{a_i \cdot u \mid \|u\| < R\} = R\|a_i\|$, constraint is: $a_i \cdot x_c + R\|a_i\| \le b_i$

Hence: Maximize R

subject to: $a_i \cdot x_c + R\|a_i\| \le b_i,\quad i = 1, \cdots, m$

Chebyshev center, cont.

Puzzle: Why is it OK to maximize R, not minimize it?

Puzzle: Can the minimum sphere enclosing $\mathcal{P}$ also be cast as an LP?

Puzzle: How about the minimum enclosing sphere for a set of points?

How about maximum *ellipsoid* $\mathcal{E}$ inside polytope $\mathcal{P}$?

$$\mathcal{E}(x_c, E) = \{v \in \mathbb{R}^n : (v - x_c)'(EE')^{-1}(v - x_c) \le 1\}$$

Minimize $-\log(\det(E))$ subject to: $b - Ax_c - h(E) \ge 0,\quad E \succ 0$

where $h(E) \doteq (\|Ea_1\|, \|Ea_2\|, \cdots, \|Ea_m\|)$ *(Zang and Gao, 2001)*

Duality

Duality

General optimization problem: Minimize: $f(x)$

subject to: $c_i(x) \le 0, \quad i = 1, \cdots, m$

$h_i(x) = 0, \quad i = 1, \cdots, p$

$f, c_i, h_i : \mathbb{R}^n \to \mathbb{R}$

Associated Lagrangian: $L(x, \lambda, \nu) \doteq f(x) + \sum_{i=1}^{m} \lambda_i c_i(x) + \sum_{i=1}^{p} \nu_i h_i(x)$

$\lambda_i \ge 0$

Define the Lagrange dual function $g(\lambda, \nu) : \mathbb{R}^{m+p} \to \mathbb{R}$ by

$$g(\lambda, \nu) \doteq \inf_{x \in D} \left\{ f(x) + \sum_{i=1}^{m} \lambda_i c_i(x) + \sum_{i=1}^{p} \nu_i h_i(x) \right\}$$

Lagrange Dual, Weak and Strong Duality

$$g(\lambda,\nu) \doteq \inf_{x\in D}\left\{ f(x) + \sum_{i=1}^{m} \lambda_i c_i(x) + \sum_{i=1}^{p} \nu_i h_i(x) \right\}$$

For any feasible x, $g(\lambda,\nu) \le f(x)$, in particular $\boxed{g(\lambda,\nu) \le f_*}$

Let $g_* \doteq \max_{\lambda,\nu} g(\lambda,\nu)$. Then $f_* - g_* \ge 0$ and $f_* - g_*$ is called the 'optimal duality gap'

Lagrange dual problem: Maximize $g(\lambda,\nu)$
subject to $\lambda \ge 0$

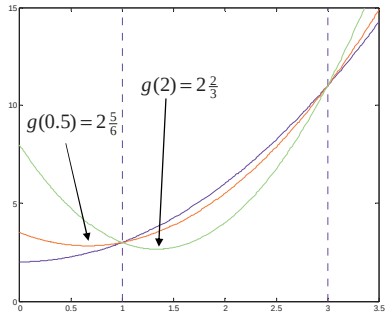
If $g_* = f_*$, "strong duality" holds. Does not always hold, even for convex optimization problems. But there are tests ("constraint qualifications").

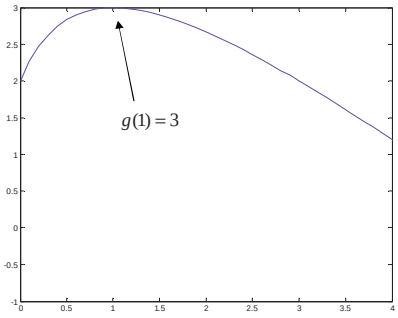
Example

Minimize $f(x) = x^2 + 2$ subject to $c_1(x) = (x-1)(x-3) \le 0, \quad x \in \mathbb{R}$

Feasible set: $1 \le x \le 3$
Optimal solution: $x_* = 1$
Optimal value: $f_* \doteq f(x_*) = x_*^2 + 2 = 3$
Lagrangian: $L(x,\lambda) = x^2 + 2 + \lambda(x-1)(x-3)$
Lagrange dual: $g(\lambda) = \inf_x L(x,\lambda) = \dfrac{-4\lambda^2}{1+\lambda} + 3\lambda + 2$
Dual problem: maximize $g(\lambda)$ subject to $\lambda \ge 0$ (*convex*)
$\dfrac{\partial g}{\partial \lambda} = 0$ at $\lambda = 1, \ \lambda = -3$ (not feasible)
At optimum, $g(\lambda) = g(1) = 3 \ \Rightarrow$ strong duality holds.

Example, cont.





Blue: objective function, $L(x, \lambda = 0)$

Red: $L(x, \lambda = 0.5)$

Green: $L(x, \lambda = 2.0)$

The Lagrange dual $g(\lambda)$

Puzzle: Why do all curves go through (1,3) and (3,11)?

Examples

Minimize $x^T x$ subject to $Ax = b,\ x \in \mathbb{R}^n,\ A \in M_{mn}$

$L(x, v) = x^T x + v^T (Ax - b), \quad v \in \mathbb{R}^m;\ \nabla_x L(x, v) = 2x + A^T v = 0$

$\Rightarrow\ x_{\min} = -(1/2) A^T v$

$g(v) = L(x_{\min}, v) = -(1/4) v^T A A^T v - b^T v$

The dual of an LP is an LP. For LPs, strong duality fails only if both primal and dual are infeasible. For LP, dual of dual is equivalent to primal.

Examples

A bound for a combinatoric problem:

Minimize $x^T A x$ subject to $x_i^2 = 1,\ i = 1, \cdots, n,\ x \in \mathbb{R}^n,\ A \in S^n$

Dual is $g(v) = \begin{cases} -\mathbf{1}^T v & A + diag(v) \succeq 0 \\ -\infty & otherwise \end{cases}$

Choose $v_c = -\lambda_{\min}(A)\mathbf{1}$ (dual feasible).

This gives a lower bound on the solution: $f_* \geq g(v_c) = n\lambda_{\min}$

Dual Norms

The dual norm is defined: $\|u\|^* \doteq \max_x \{u \cdot x \mid \|x\| = 1\},\ \ u, x \in \mathbb{R}^n$

In fact $\|x\|$ need only be a 'prenorm', in which case $\|u\|^*$ is still a norm

Theorem: Let $\|\cdot\|$ be a vector norm on $\mathbb{R}^n$, let $\|\cdot\|^*$ be its dual norm, and let $c > 0$. Then $\|x\| = c\|x\|^*\ \forall x \in \mathbb{R}^n \iff \|\cdot\| = \sqrt{c}\|\cdot\|_2$

For any $p \geq 1$, the dual of the l_p norm is the l_q norm, with $\frac{1}{p} + \frac{1}{q} = 1$

Conjugate Functions

The conjugate function f^* of a function $f : \mathbb{R}^n \to \mathbb{R}$ is defined:

$$f^* \doteq \sup_{x \in dom(f)} \left(y^T x - f(x) \right)$$

Domain of f^* defined to be those $y \in \mathbb{R}^n$ for which the $\sup$ is finite.

f^* is convex whether or not f is.

Examples: functions with domain $\mathbb{R}$:

$f_1(x) = ax + b$	$domain(f_1^*) = \{a\}$	$f_1^*(a) = b$
$f_2(x) = -\log x$	$domain(f_2^*) = \{y \mid y < 0\}$	$f_2^*(y) = -\log(-y) - 1$
$f_3(x) = \exp x$	$domain(f_3^*) = \mathbb{R}_+$	$f_3^*(y) = y \log y - y$
$f_4(x) = x \log x$	$domain(f_4^*) = \mathbb{R}$	$f_4^*(y) = \exp(y - 1)$

Conjugate functions: examples, cont.

Two examples in $\mathbb{R}^n$:

$$f(x) = \frac{1}{2} x^T H x,\ H \succ 0, \qquad domain(f^*) = \mathbb{R}^n \qquad f_5^*(y) = \frac{1}{2} y^T H^{-1} y$$

$$f(x) = \|x\|, \qquad domain(f^*) = \mathbb{R}^n \qquad f^*(y) = \begin{cases} 0 & \|y\|^* \le 1 \\ \infty & otherwise \end{cases}$$

The conjugate of a norm is the indicator function of the dual norm unit ball.

More general examples

Minimize $f(Ax+b)$

Recast: minimize $f(y)$ subject to $y = Ax+b$

$L(x,y,v) = f(y) - v^T(y - Ax - b)$

Lagrange dual function is $g(v) = \inf_{x,y}\left\{ f(y) - v^T y + v^T Ax + b^T v \right\}$

$\Rightarrow$ for feasible v, $v^T A = 0$

Then: $g(v) = -\sup_y \left\{ v^T y - f(y) \right\} + b^T v = b^T v - f^*(v)$

E.g. minimize $\|Ax+b\|$ for any norm $\|\cdot\|$

$\Rightarrow$ maximize $g(v) = b^T v$ subject to $\|v\|^* \le 1,\ v^T A = 0.$

Check: e.g. if A^{-1} exists, then $v = 0,\ g_* = 0$ and $\min\left(\|Ax+b\|\right) = 0.$

More generally if $b \in range(A)$ then $\min\left(\|Ax+b\|\right) = 0,$

and $v^T A = 0 \Rightarrow v \in null(A) \Rightarrow b'v = g = 0.$

More general examples, cont.

Minimize: $f(x)$

subject to: $Ax \le b,\quad Cx = d$

$domain(g) = \left\{ (\lambda, v) \mid -A'\lambda - C'v \in domain(f^*) \right\}$

$g(\lambda, v) = \inf_x \left(f(x) + \lambda'(Ax-b) + v'(Cx-d) \right)$

$= -b'\lambda - d'v - f^*(-A'\lambda - C'v)$

Notes on the Lagrange Dual

Stopping criterion: suppose x is primal feasible and (λ, ν) are dual feasible. Then for a problem for which strong duality holds, we know that

$f(x) - f(x_*) \leq f(x) - g(\lambda, \nu)$

A non-convex optimization problem can still have strong duality:

e.g. minimize $x^T A x + 2b^T x$, subject to $x^T x = 1$

Weak duality can be expressed as

$$\sup_{\lambda \geq 0} \inf_x L(x, \lambda) \leq \inf_x \sup_{\lambda \geq 0} L(x, \lambda) \qquad (\dagger)$$

In fact for any function $f(x, y)$, and any sets Y, Z, we have

$$\sup_{y \in Y} \inf_{z \in Z} f(y, z) \leq \inf_{z \in Z} \sup_{y \in Y} f(y, z)$$

If strong duality holds, the "$\leq$" in $(\dagger)$ becomes "$=$".

Puzzle: Find a function $f(y, z)$, and sets Y, Z containing just two points each, such that the weak max-min inequality is strict.

Karush -Kuhn -Tucker and Complementary Slackness

Complementary Slackness

Assume strong duality holds. Then:

$$f_* \doteq f(x_*) = g_* \doteq g(\lambda_*, \nu_*) = \inf_x \left(f(x) + \sum_{i=1}^m \lambda_{*i} c_i(x) + \sum_{i=1}^p \nu_{*i} h_i(x) \right)$$
$$\leq f_* + \sum_{i=1}^m \lambda_{*i} c_i(x_*) + \sum_{i=1}^p \nu_{*i} h_i(x_*)$$

$$\Rightarrow \sum_{i=1}^m \lambda_{*i} c_i(x_*) = 0$$

$$\Rightarrow \lambda_{*i} c_i(x_*) = 0 \ \forall i \ : \text{ "Complementary Slackness"}$$

Hence also at the solution $\inf_x L(x, \lambda_*, \nu_*) = L(x_*, \lambda_*, \nu_*)$

Karush-Kuhn-Tucker Conditions

For problems with strong duality (zero duality gap at opt.):

$$f_* = g_* = \inf_x \left(f(x) + \sum_{i=1}^m \lambda_{*i} c_i(x) + \sum_{i=1}^p \nu_{*i} h_i(x) \right)$$
$$= f(x_*) + \sum_{i=1}^m \lambda_{*i} c_i(x_*) + \sum_{i=1}^p \nu_{*i} h_i(x_*)$$

If f, c, h are differentiable, then

$\nabla f_* + \sum_{i=1}^m \lambda_{*i} \nabla c_{*i} + \sum_{i=1}^p \nu_{*i} \nabla h_{*i} = 0$ so:

Constraints met :	$h_i(x_*) = 0, \ c_i(x_*) \leq 0, \ \lambda_i \geq 0$
Complementary slackness :	$\lambda_{*i} c_i(x_*) = 0$
Gradient vanishes :	$\nabla f_* + \sum_{i=1}^m \lambda_{*i} \nabla c_{*i} + \sum_{i=1}^p \nu_{*i} \nabla h_{*i} = 0$

Karush-Kuhn-Tucker Conditions

For any optimization problem (convex or not), primal and dual optimal points with zero duality gap must satisfy KKT

For any convex optimization problem, if $\{x, \lambda_i, v_i\}$ satisfy KKT, then $\{x, \lambda_i, v_i\}$ are primal and dual optimal with zero duality gap.

Example: minimize $\frac{1}{2} x^T H x + q^T x$ subject to $Ax = b, \quad H \in S_+^n$

$$L = \tfrac{1}{2} x^T H x + q^T x + v^T (Ax - b)$$

KKT: (i) $Ax_* = b$

(ii) $\nabla L(x_*, v_*) = 0 = Hx_* + q + A^T v_*$

$$\Rightarrow \begin{pmatrix} H & A^T \\ A & 0 \end{pmatrix} \begin{pmatrix} x_* \\ v_* \end{pmatrix} = \begin{pmatrix} -q \\ b \end{pmatrix}$$

Linear system $\rightarrow$ get complete set of solutions with SVD