Some Mathematical Tools for Machine Learning

Chris Burges
Microsoft Research
February 13th, 2009

Contents – Part 1

A Brief Diversion

Leslie Lamport: *“How to Write a Proof”*: DEC tech report, Feb 14, 1993

Abstract: A method of writing proofs is proposed that makes it much harder to prove things that are not true. The method, based on hierarchical structuring, is simple and practical.

“Anecdotal evidence suggests that as many as a third of all papers published in mathematical journals contain mistakes – not just minor errors, but incorrect theorems and proofs.”

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 $c(x) = x'Ax = 1,\ x \in R^n$

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 y lives on S^{n-1}:

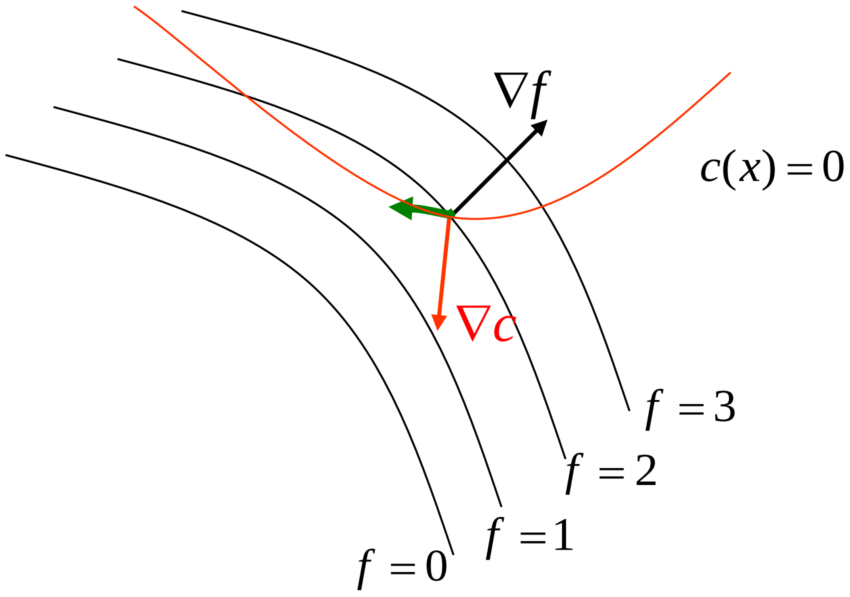
$$y_1 = \sin\theta_1 \sin\theta_2 \cdots \sin\theta_{n-1}$$
$$y_2 = \sin\theta_1 \sin\theta_2 \cdots \cos\theta_{n-1}$$
$$\ldots$$

solve, and then apply the inverse mapping. But this can get very complicated!

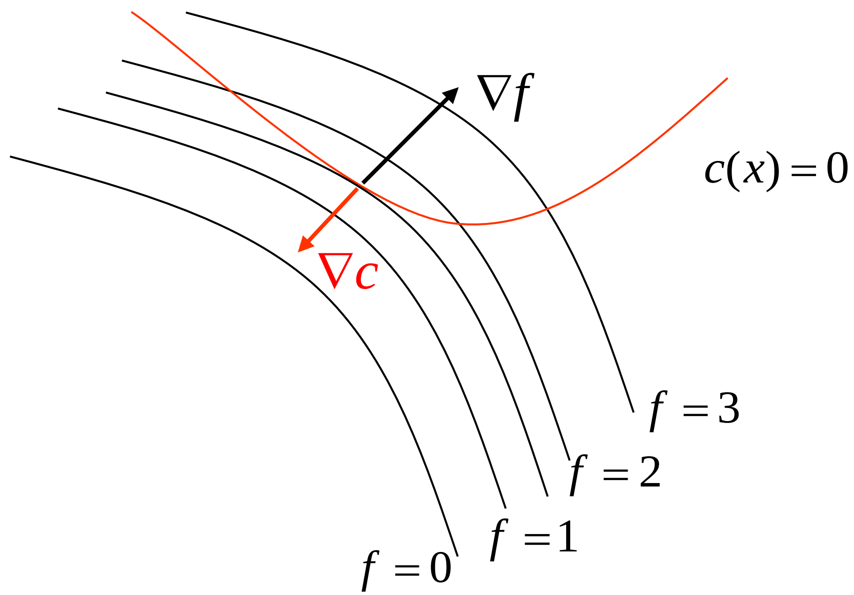
It is often not even possible to parameterize constraints (for example, polynomial constraints in several variables)

One equality constraint

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



One equality constraint, cont.



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

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

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

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 $\left(\nabla f\right)_\perp \neq 0$

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

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

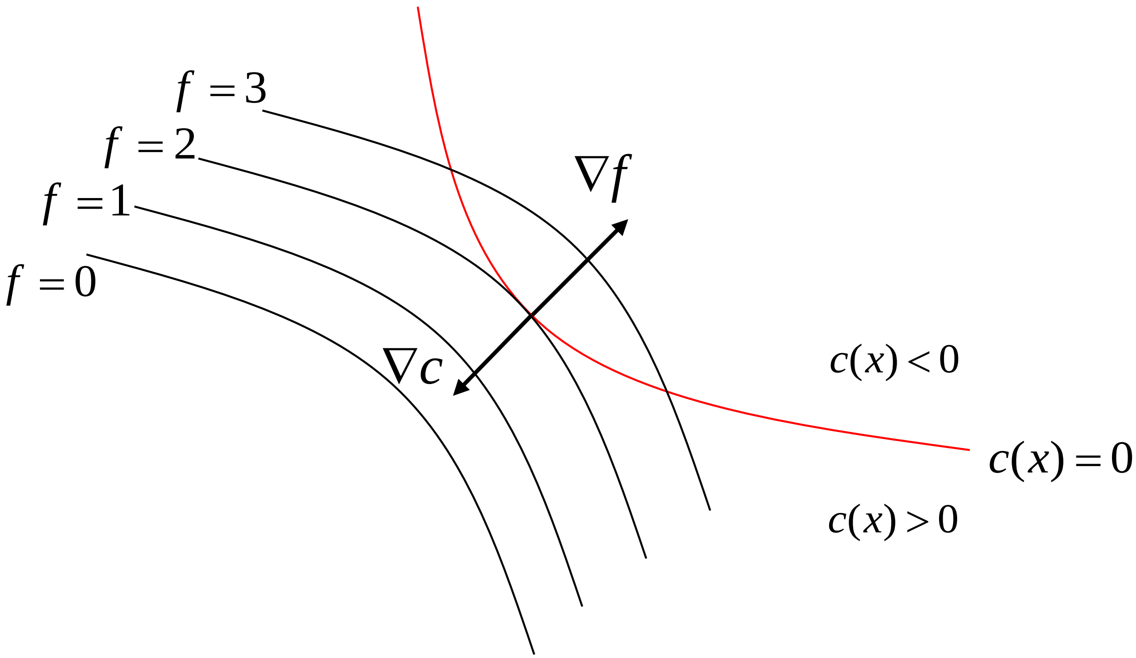
Puzzle: why not multiple Lagrangians?

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

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

Suppose that at the solution x_*, $c_j(x_*) < 0$.

Then removing c_j makes no difference, and we must drop ∇c_j from the sum in

$$\nabla f = -\sum_i \lambda_i \nabla c_i$$

Equivalently we can set $\lambda_j = 0$

Hence, always impose $\lambda_i c_i(x_*) = 0$

A simple example

Extremize the distance between two points on S^n :

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

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

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

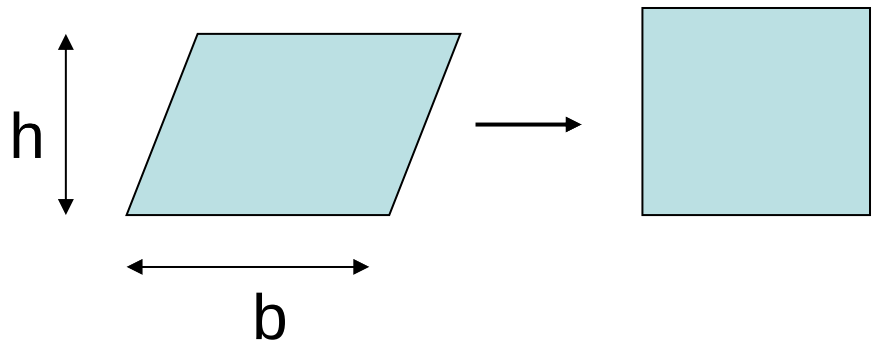
$$\nabla_1 L = 0 \;\Rightarrow\; (x_1 - x_2) + \lambda_1 x_1 = 0$$
$$\nabla_2 L = 0 \;\Rightarrow\; (x_2 - x_1) + \lambda_2 x_2 = 0$$

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

$\Rightarrow$ 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"

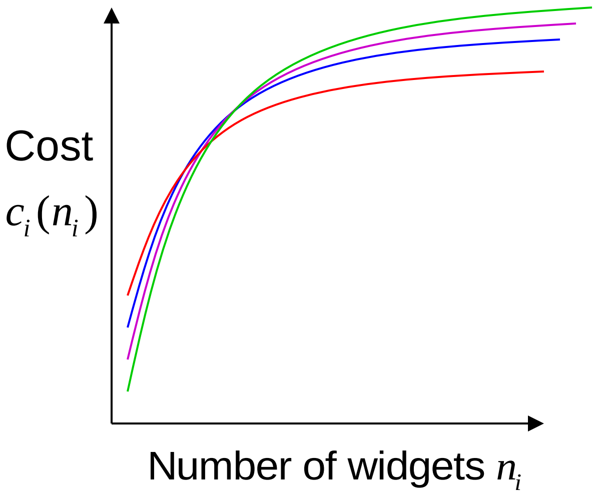
Simple exercises

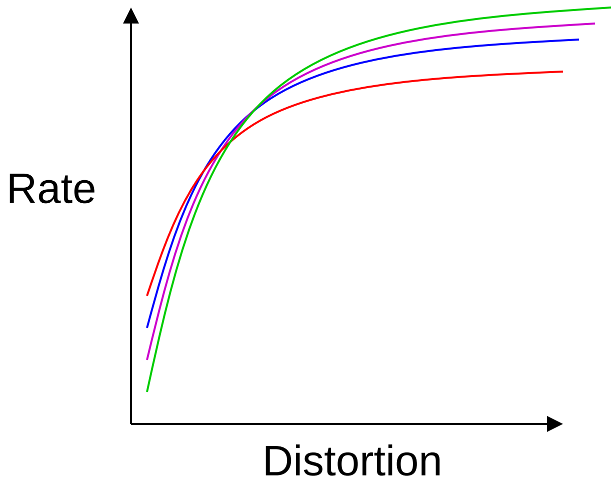
Puzzle: what coefficients maximize a convex sum of fixed numbers?

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





Puzzle: Does this work for economics, too?

Fiber i has n_i bit errors per second, and sends $c_i(n_i)$ bits total, per second. We wish to maximize the bit rate at a fixed distortion rate:

$$L = \sum_{i=1}^{4} c_i(n_i) - \lambda \left(N - \sum_{i=1}^{4} n_i \right)$$

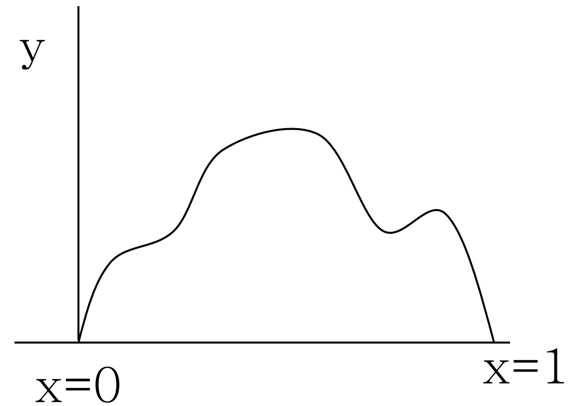
$$\nabla L = 0 \;\rightarrow\; \frac{\partial c_i}{\partial n_i} = \lambda \;\; \forall i$$

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

$\rightarrow$ 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: $\sum_i a_{ij} P_i = C_j$

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

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

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

$\rightarrow$ logistic regression!

Are Lagrange Multipliers Really That Common?

Yes.

For example:

- Most flavors of Support Vector Machines
- Principal Component Analysis
- Canonical Correlation Analysis
- Locally Linear Embedding
- Laplacian Eigenmaps
- …

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

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: $F = \mathrm{R}$

the rationals: $F = \mathrm{Q}$

the complex numbers: $F = \mathrm{C}$

What's the smallest field?

$\mathbf{Z}_2 = \{0,1\}$ with + meaning "XOR" and $*$ meaning "AND"

Puzzle 1: For Z_2, why doesn't using "OR" for + work?

Puzzle 2: Is $R_+ = \{x : x \geq 0\}$ a field under $\{+,*\}$?

How Many Fields Are There?

Actually infinitely many - e.g. Z_p, p prime.

However, can define an ordering for some fields.

R, Q can be ordered; C and 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.

Q is not complete, but R is. In fact: every complete, ordered field is isomorphic to 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 R^n

Vector Spaces: Field Matters!

$\{E, F\} = \{R^n, R\}$ (dimension N);

$\{E, F\} = \{C^n, C\}$ (dimension N);

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

'Linear dependence' depends on field F :

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

For the vector space $\{C, C\}$ 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 C;

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

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

What is an Inner Product?

Let V be a vector space over R or 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 $\{R^n, R\}$ with $\langle \cdot,\cdot \rangle$ defined by $\langle x, y \rangle = \sum_i x_i y_i$

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

(3) Vector space of matrices over R with $\langle X, Y \rangle = Trace(X^T Y)$

$(X,\ Y \in M_{pm})$

Inner Product: Trace

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

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

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

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

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

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:

$$\theta = \cos^{-1}\left(\frac{\langle x, y\rangle}{\sqrt{\langle x, x\rangle\langle y, y\rangle}}\right):\ 0 \leq \theta \leq \pi$$

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 R or C. A function $\|\cdot\|: V \to R$ 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 $V_0 = \{v \in V : \|v\| = 0\}$ 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 R^5: $\langle \cdot, \cdot \rangle = x' Dx,\ D = diag(3,2,1,0,0),$ 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 $x = (x_1, x_2, \cdots, x_n) \in R^n$ is the Euclidean norm.

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

This works for finite sums too: define the l_p norm on R^n as:

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

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

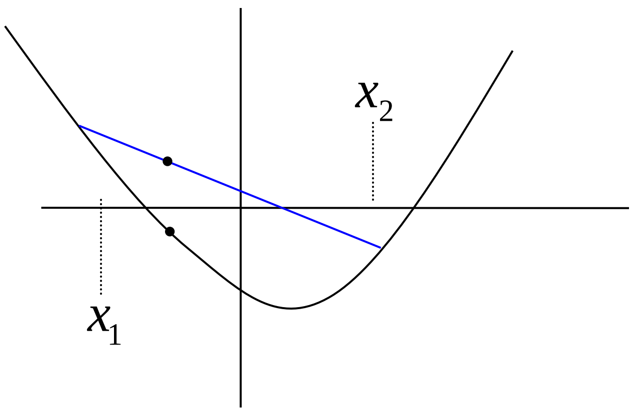
The l_∞ norm is the 'max' norm $\|x\| = \max(\{x_i\})$

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

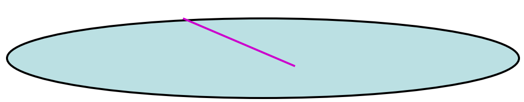
What is convexity?

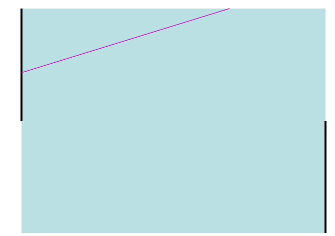
A function is convex if, between x_1 and $x_2 > x_1$, it lies below the chord joining $f(x_1)$ and $f(x_2)$. It is strictly convex if it lies strictly below.

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

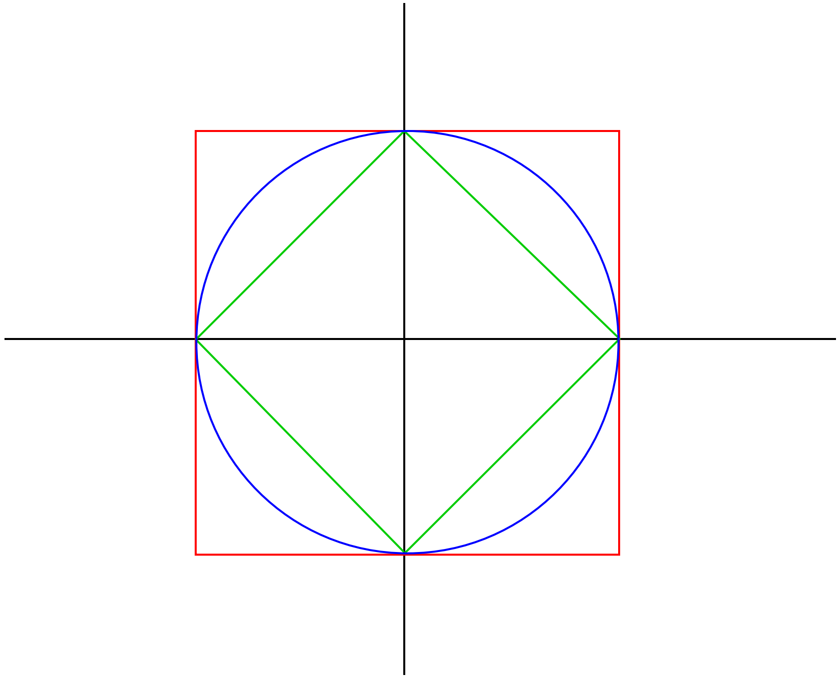


A set S is convex if, for all $a \in S$ and $b \in S$, all points on the line joining a and b lie in S.





When is a sphere a square?



Red: the l_∞ 1-sphere

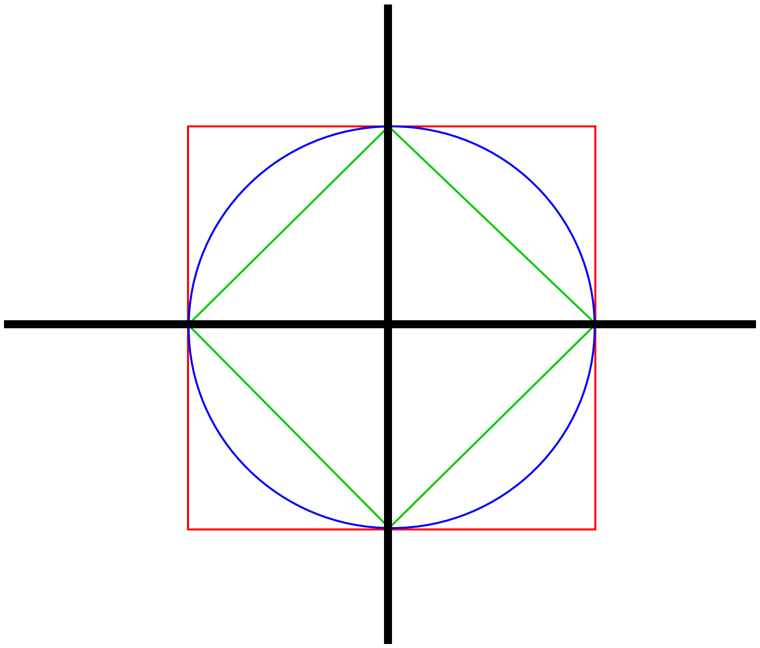
Blue: the l_2 1-sphere

Green: the l_1 1-sphere

Interesting... each ball looks convex

When is a sphere a cross?

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



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

$$\begin{aligned}\|\alpha v\|_0 &= 0 \text{ if } \alpha = 0,\\ &= \|v\|_0 \text{ if } \alpha \neq 0\\ &\neq |\alpha|\,\|v\|_0 \text{ unless } \alpha = 1.\end{aligned}$$

All norms are convex functions

$$\|(1-\lambda)x_1+\lambda x_2\| \le \left\|\left(1-\lambda\right)x_1\right\| + \|\lambda x_2\|$$
$$= \left|\left(1-\lambda\right)\right| \|x_1\| + |\lambda| \|x_2\|$$
$$= \left(1-\lambda\right)\|x_1\| + \lambda \|x_2\|$$

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

let $S = \{x : f(x) \le 0\}$. 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 $n = \| \cdot \|$, and a set S in a vector space V :

$Open: \ \forall \ x \in S, \exists \ \varepsilon > 0 \ s.t. \ B_n(\varepsilon, x) \subset S$

$Closed : complement \, of \ open$

$Bounded : \ \exists \ r > 0 \ such \, that \, S \subset B_n(r, 0)$

$Compact : \ Every \, sequence \left\{ x_i \right\} in \, S \, contains \, convergent \, subsequence$

$with \, limit \ in \ S$

$OR : Every \, cover \, has \, a \ finite \, sub\text{-}cover :$

$$\bigcup_\mu S_\mu \supset S, \ \exists \ S_{\alpha 1}, S_{\alpha 2}, \ \cdots, S_{\alpha N}, \ N \ finite, \ such \, that \ \bigcup_{\alpha i}^N S_{\alpha i} \supset S$$

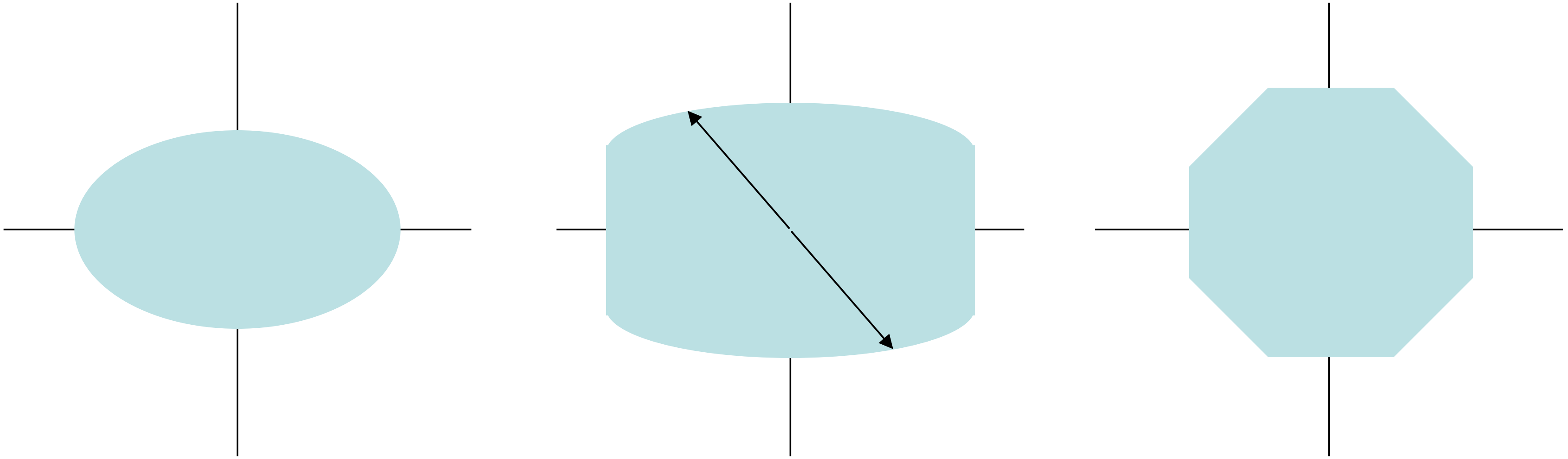
Compact sets are closed and bounded.

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):



Rescuing l_0: the Hamming norm

Consider n-tuples taking values in Z_2. These form a vector space over the field 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: $(N > 1) \notin Z_2$ - how can this be correct?

Puzzle: What is the subtraction operation, in Z_2?

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

When does a norm come from an inner product?

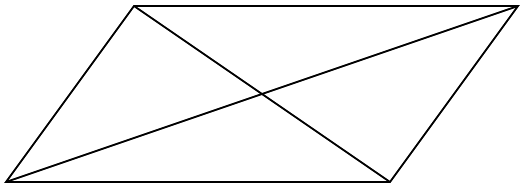
Every inner product defines a norm: $\|x\| = \sqrt{\langle x, x \rangle}$

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

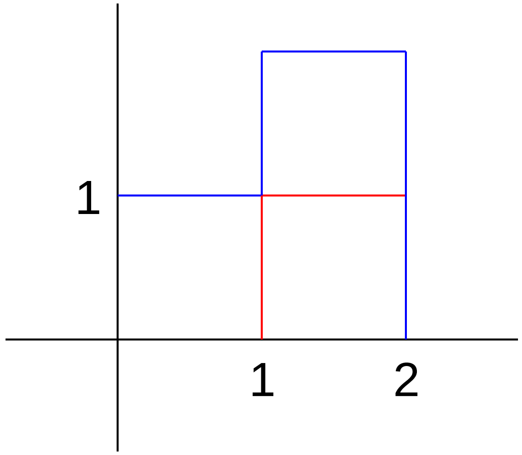
$\langle \alpha_1 f_1 + \alpha_2 f_2, g \rangle = \left(\int ((\alpha_1 f_1 + \alpha_2 f_2) g)^{p/2} \right)^{2/p} \neq \alpha_1 \langle f_1, g \rangle + \alpha_2 \langle f_2, g \rangle$

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 R), then $\langle x, y\rangle = \frac{1}{4}\left(\|x+y\|^2 - \|x-y\|^2\right)$. Choose two functions:



$f_1(x) = 0 : x < 0$

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

$f_1(x) = 2 : 1 \leq x \leq 2$

$f_1(x) = 0 : x > 2$

$f_2 = 0 : x < 1$

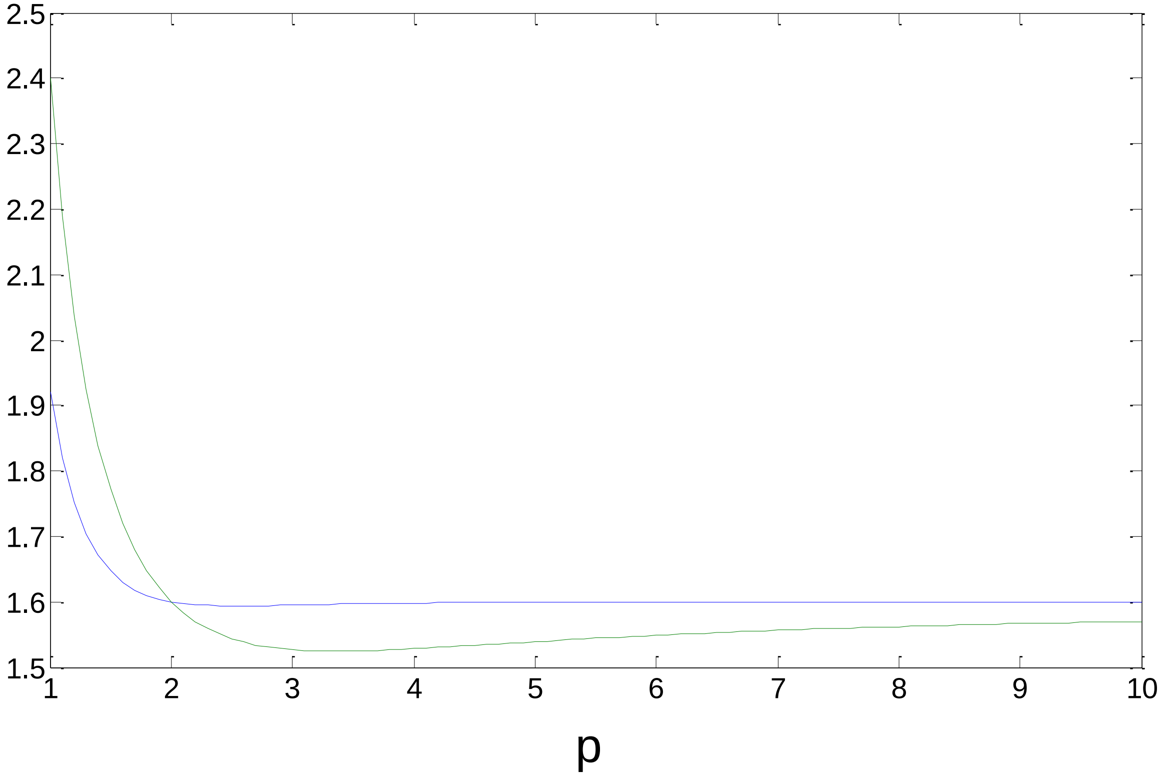
$f_2 = 1 : 1 \leq x < 2$

$f_2 = 0 : x > 2$

L_p norms, inner products cont.

We'd like:

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



l_∞ norm on R^n has no inner product

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

Use parallelogram test:

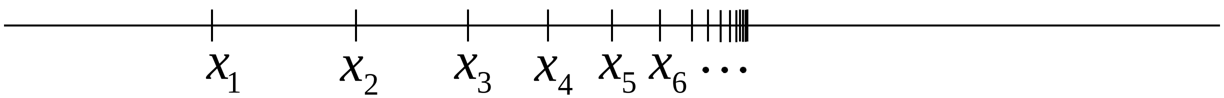
$$\frac{1}{2}\left(\|x+y\|^2 + \|x-y\|^2\right) = ? = \|x\|^2 + \|y\|^2$$

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

Extend to 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.

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 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.

R^n with l_p norm is complete, for all $1 \leq p \leq \infty$.

The sequence space l_p is complete, for all $1 \leq p \leq \infty$.

$C[a,b]$ with L_∞ norm is complete.

$C[a,b]$ with L_2 norm or L_1 norm is not complete.

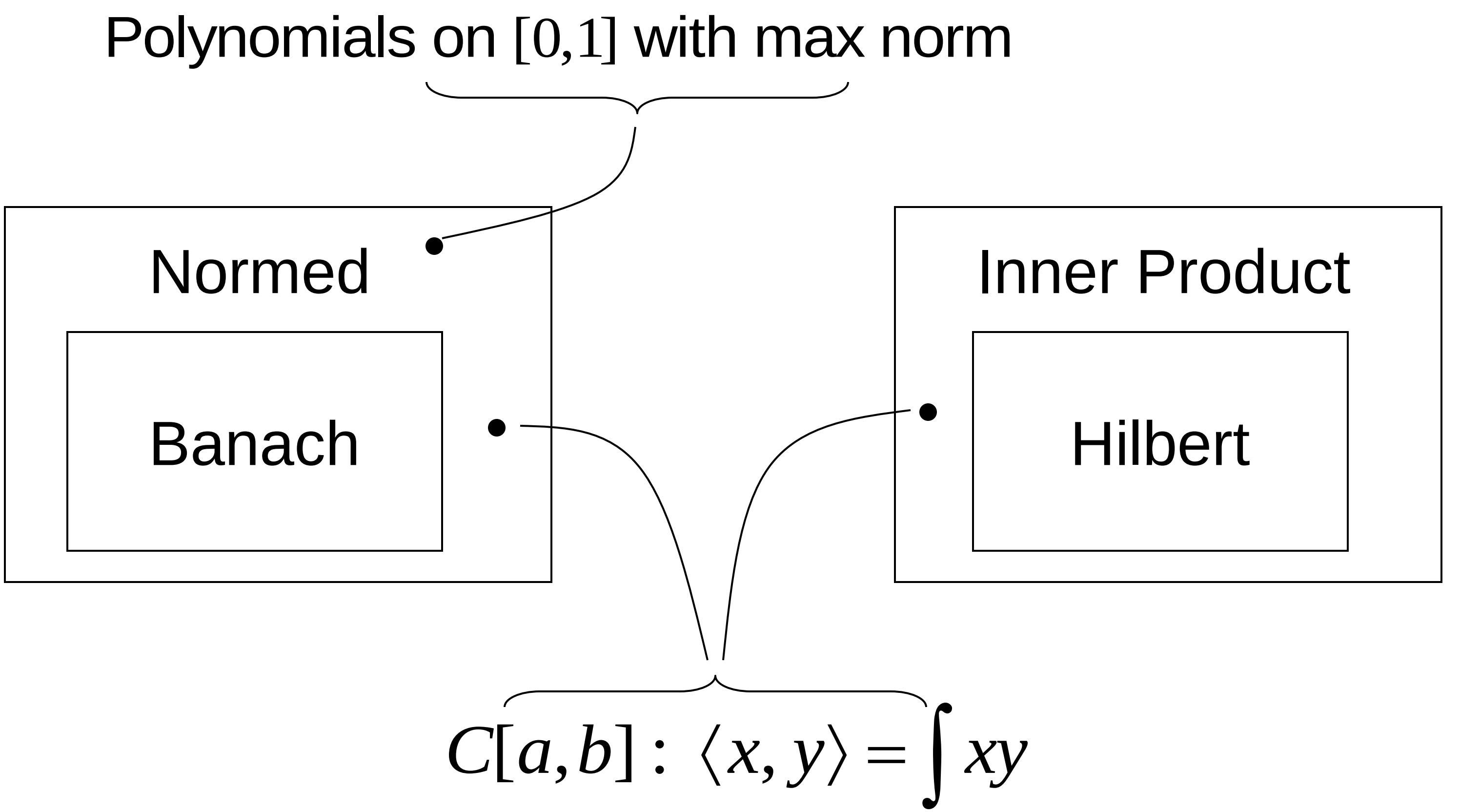
"All square integrable functions with L_2 norm" is complete.

A complete space "contains all its limit points."

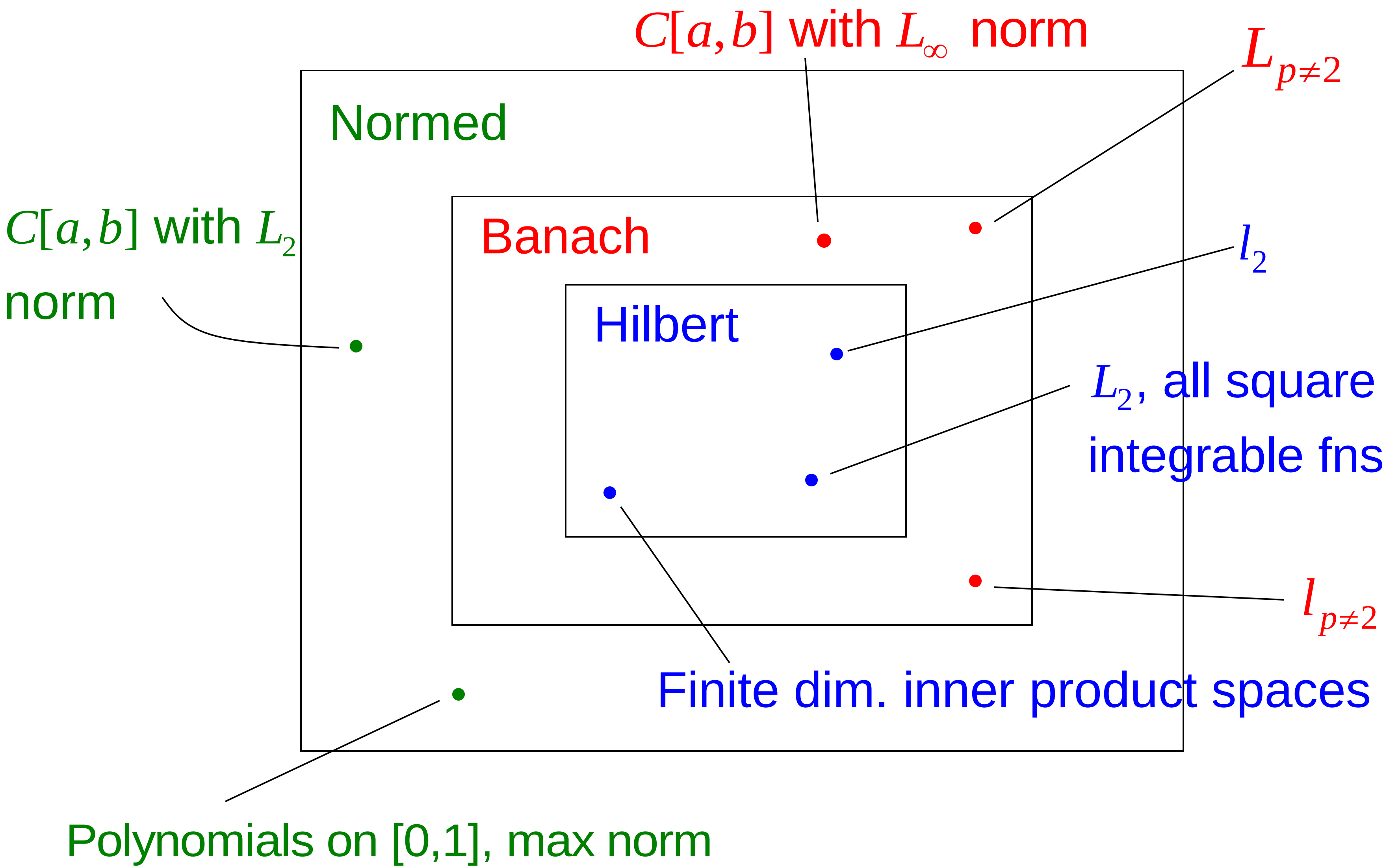
It is always possible to 'complete' a non-complete space.

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 and only if it admits a countable orthonormal basis. (So, all finite dimensional Hilbert spaces are 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 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:

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

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

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

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

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

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

Riesz Representation Theorem, cont.

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

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.

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

A Brief Look Ahead

The Riesz representation theorem can be used to show that any Hilbert space for which the evaluation functional is continuous, is a Reproducing Kernel Hilbert Space: there exists *K* such that

$$\langle f, K(x,.)\rangle = f(x)$$

In particular:

$$\langle K(x_1,.), K(x_2,.)\rangle = K(x_1, x_2)$$

Representer Theorem (Kimeldorf and Wahba, 1971; Schölkopf and Smola, 2002): Let $\Omega\colon R_+ \rightarrow R$ be a strictly monotonic increasing function and let c be an arbitrary loss function. Then each minimizer $f \in \mathcal{H}$ of the “regularized risk”

$$c\big(x_1, y_1, f(x_1), \ldots, x_m, y_m, f(x_m)\big) + \Omega(\|f\|_{\mathcal{H}})$$

admits a representation of the form $f(x) = \sum_{i=1}^{m} \alpha_i K(x, x_i)$

What is a metric space?

For any set E, let $\rho(x, y)$ be a function (with range in 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 $\rho : E \times E \to R$ 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 $\rho(x, y) \geq 0$?

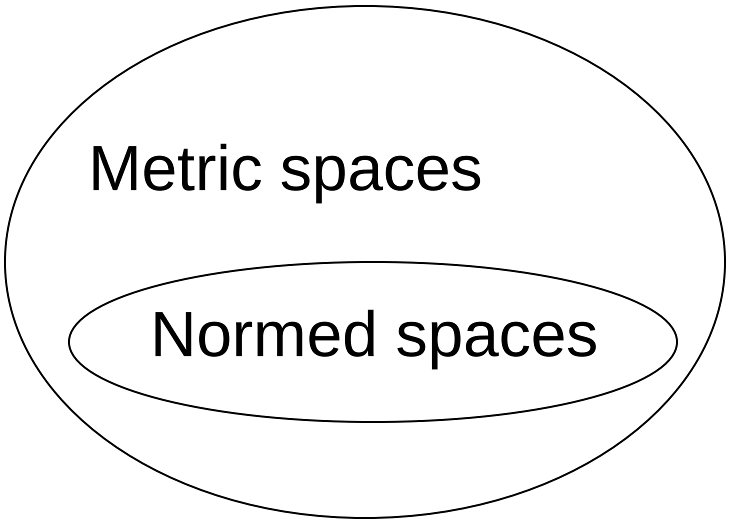
Puzzle: How about $\rho(x, y) = \rho(y, x)$?

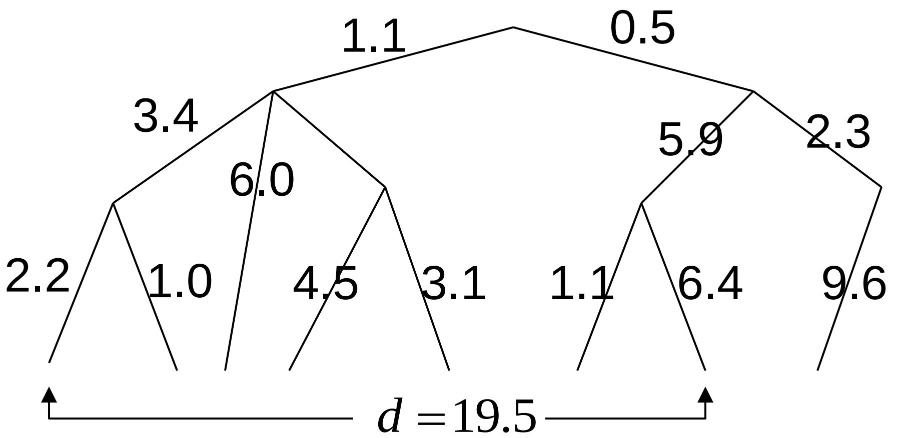
Puzzle: Where's the triangle inequality: $\rho(x, y) \leq \rho(x, z) + \rho(z, y)$?

Metric versus norm

Every normed vector space is a metric space: define $\rho(x,y) = \|x-y\|$

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 $f:[a,b]\to \mathrm{R}$. For two points $f,g\in A$, $\rho_\infty(f,g)=\sup_{x\in[a,b]}\left|f(x)-g(x)\right|$ is a metric.

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

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

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

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

Topological Spaces

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

(1) $A, \varnothing \in S$,

(2) Intersection of any two sets in S is in S,

(3) Union of any collection of sets in S is in S.

S is called a topology for A, and the members of S are called the open sets of T.

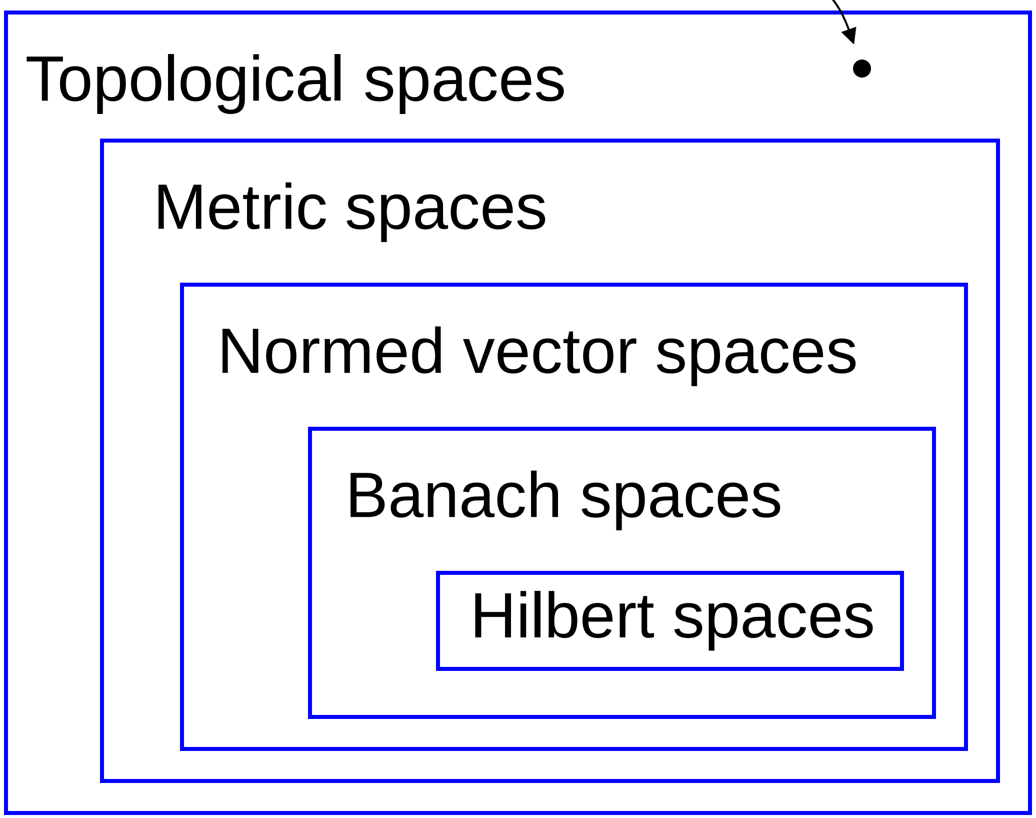
Topological spaces are more general than metric spaces.

Topological spaces extend "continuity": Given $T_1 = \{A_1, S_1\}$ and $T_2 = \{A_2, S_2\}$ and a map $\phi : A_1 \to A_2$, ϕ is "continuous" if $U \in A_2 \Rightarrow \phi^{-1}(U) \in A_1$

Continuity, convergence, connectedness… without distance!

Putting Spaces in their Places…

$S = \{A, \varnothing\}$: the "indiscrete" topology on A



~ The Middle ~

Thanks!