Vector spaces and functions: Survey

vishvAs vAsuki

March 28, 2012

Contents

Part I

Introduction

1 Themes

1.1 References

Based on [5], [6], [2] [3], [4].

1.2 Themes

Vector spaces, their functions and functionals

1.2.1 Related surveys

For info on linear maps on vector spaces, see that survey.
For numerical analysis, conditioning, stability, differnce equations, differential equations: see Numerical analysis ref.

1.3 Characterization of research effort

See linear algebra survey.

2 Notation

See survey on linear maps on vector spaces.

Part II

Vectors and vector spaces

3 Vectors and their combinations

3.1 Definition, basic operations

A sequence of n elements of f, v, is a vector.

3.1.1 Addition, scalar multiplication

Addition of vectors is naturally defined: $(u+v)_i = u_i+v_i$. Scalar multiplication is similarly defined.

3.2 Geometric model: coefficient sequence

Can be interpreted as a point in the cartesian space F^n; but which point? That depends on what the basis vectors model, the units, the inner product between them.

3.2.1 As a combination of standard basis vectors

Define standard basis vectors: e_i as a vector with 1 in the ith position, 0 elesewhere.
So, v can be viewed as a combination v_ie_i of the standard basis vectors $\{e_i\}$.

3.2.2 Standard basis models what?

The standard basis are usually considered with the standard inner product: $\langle x, y\rangle = \sum_i x_i y_i$. But this need not be the case. Other bases are possible.

3.2.3 Equivalent representations of the same point

So, the vector representing a point in a coordinate space changes, when the basis chosen is different. Change of basis operation is a linear operation, for details see linear algebra ref.

3.3 Combinations of vectors

3.3.1 Linear combination

$\sum a_i v_i = p$.

3.3.2 Conic/ non-ve combination

Coefficients $a_i \geq 0$.

3.3.3 Affine combination

If coefficients $\sum a_i = 1$, this is an affine combination.

Colleniearity preserved Make affine combo $p = ax + (1-a)y$; take vector $x - p = (1-a)(x-y)$; this has same inner product with the basis vectors as x-y.

3.3.4 Convex combination

Affine combo where $a_i \geq 0$: both affine and conic.

4 Interpretations, applications

Not necessarily just an abstraction of geometric 3-D or 2-D spaces. Many other things modelled by fixed length sequences of numbers.

4.1 Modelling the real world

A number can represent a weight, height etc.. It could mean a certain combination of weight, height etc..

4.2 Functions as vectors

4.2.1 Using coefficients

Take the polynomial $P(x,t)$ of degree 6 with coefficients given by the vector t. Thus, one can

4.2.2 Using the domain

See function spaces section.

5 Vector sets

5.1 Properties

5.1.1 Linear independence

A set of vectors $\{t_i\}$ is linearly independent if, for all i, t_i can't be written as a linear combination of $\{t_j : j \neq i\}$.

5.1.2 Associated hyperplanes

Supporting hyperplane to C at boundary pt p: All C must lie on one side of the hyperplane.
Separating hyperplanes between sets.

5.2 Span of vectors in S

Contains all linear combos of vectors in S. Any linear subspace expressible as Ax = 0. Eg: $a^T x = 0$
$\langle a..b \rangle$ represents space spanned by vectors $a..b$.

5.3 Affine set X

X closed under affine combination. Eg: A line parallel to 1-d vector space, solution to Ax = b. Contains the line through any two points in X.
If it included the origin, it would be a linear subspace! Any affine set expressible as $\{x : Ax + b = 0\}$. Is convex.

5.3.1 Affine hull of S

aff(S): Smallest affine set which contains S.

Relative interior of S $relint(S) = \{x \in S : \exists \epsilon > 0 : N_\epsilon(x) \cap aff(S) \in S\}$. A straight line segment and a plane in 3-d space have no interior, but have a relative interior.

5.4 Convex cone C

If $x, y \in C$, $\forall t_1, t_2 \geq 0 : t_1x+t_2 y \in C$: encompasses all non-negative/ conic combinations of points. Is convex.
Thence, conic hull of S is defined.
Eg: Set of symmetric +ve semidefinite matrices.

5.4.1 Pointed cone C

If $p \in C$, $-p \notin C$. Smaller than halfspaces. Can delete 0 from them and still preserve convexity.

5.4.2 Proper cone C

C is closed, pointed, solid. Eg: non-ve orthant, S^n_+. Dual cones C' of proper cones are proper.

Generalized inequalities wrt C $x \leq_C y \equiv y - x \in C$;x$<_C y \equiv y - x \in int(C)$. For multiplication by scalar a, this behaves like inequalities on R.
$x \geq_C 0$ ***is a fancy way of saying that*** $x \in C$. Similarly, $x >_C 0$ means $x \in int(C)$.

Minima In general, not a complete ordering; so minimal and minimum elements defined as in ordered sets and partially ordered sets.

Minima and dual cone Minimum of C = $\arg\min_{x \in C} v^T x \forall v \in int(C')$.
Minimal element of C = $\arg\min_{x \in C} v^T x$ for some $v \in C'$: think of a dual as set of normals to supporting hyperplanes.

5.4.3 Norm cones

$\{(x, t) : \|x\| \leq t\}$: Epigraph of the norm. For euclidian norm, get 'ice cream cone': aka 2nd order cone.

5.4.4 Dual C^* of cone C

$C^* = \left\{y | y^T x \geq 0 \forall x\right\}$. This is the dual subspace of linear, nonnegative functionals. This is a cone too! C^* is set of normals to supporting hyperplanes of C.
Eg: R^n_+, S^n_+ are self dual.
Dual of a dual cone includes the primal cone.

Set of normals So, dual cone is actually the set of normal vectors defining all supporting hyperplanes of C, at its boundaries facing 0 in the first quadrant.

6 Convex set

6.1 Containment of convex combinations

X is convex if, for every $\{x_i\} \subseteq X$, its convex combination is in X. For convex combinations c, $[x, y \in C \implies c(x, y) \in C] \equiv [\{x_i\} \subseteq C \implies c((x_i)) \in C]$: from induction. extend this to possibly infinite number of points to get Jensen's inequality!

6.1.1 Containment of line-segments

Equivalently, convex combination of any pair of pts in X is in X: Can get the former condition by induction on number of points combined. So, join any 2 pts in X by a line, pick any pt p on that line; $p \in X$. So, easier to show that a set is non-convex than it is to show that it is convex.

6.2 Properties

6.2.1 Extreme points of convex set S

A corner of S; not in any line between two pts in S. If S also compact, S is the convex hull of the extreme points (Krein Milman).

6.2.2 Separating hyperplane

If C and D are 2 disjoint convex sets, then they are separable by a hyperplane. Strict separation need additional assumptions.

6.2.3 Supporting hyperplane

C has a supporting hyperplane at every boundary point.

6.2.4 Intersection of supporting half-spaces

If C is closed, it is intersection of halfspaces formed by supporting hyperplanes.

6.2.5 As domain of special barrier functionals

Any open convex set can be written as the domain of a self-concordant barrier functional.

6.3 Convex hull of a set of points X in a real vector space V

The minimal convex set containing X. $H_c(X)$. X is the boundary of the convex set.
If $|X|$ finite, convex hull is a polyhedron. Circle is the convex set of ∞ points.
Convex set is a set whose convex hull is itself.

6.4 Check convexity

Use defn. Start with convex sets, apply functions known to preserve convexity. Derive set using convexity prserving operations on other sets.

6.4.1 Functions which preserve convexity in image, inverse image

Affine fns: f(x) = Ax + b: see from defn. Perspective fns: see from defn. Linear fractional function: from composition of affine, perspective fn.

6.4.2 Convexity preserving operations

$\cap$.

6.4.3 Important convex sets

Sublevelsets of (quasi)convex fn.
Half-spaces, hyper-ellipsoids, polyhedra which are solutions of $Ax \leq b$. Norm ball.
The probability simplex: $p \geq 0; 1^T p = 1$.

7 Vector spaces and subtypes

7.1 Vector space V over field F

Vector space is closed under linear combination of a set of 'basis' vectors: A commutative group wrt +. Linear dependence of vectors: Any of the vectors expressible as a linear combo of the others.
Basis sets of n-degree polynomials (P_n) and matrices also define vector spaces.

7.2 Inner product space

A vector space V with an inner product $\langle . \rangle : V \times V \to F$.

7.3 Normed vector space

Space with a norm. Also a metric space. Thence inherit notion of completeness.

7.3.1 Lebesgue space

Aka L^p or l^p space. Infinite dimensional space with the p norm. (Minkowski) Triangle inequality still holds.

7.4 Banach space

Complete, normed, vector space.

7.5 Hilbert space

Hilbert noticed common theme: complete, normed, inner-product vector space.

8 Common vector spaces

8.1 Complex vector space

C is a field; so multiplication is defined for complex numbers. So, a complex vector space C^n is not equivalent to thinking about real vector space R^{2n}.

8.2 Functional space V

8.2.1 f() as dom(f) dim vector

Look upon f(x) as a vector whose dimensions = domain size d. A dimension for each value of x in a certain interval: $f(x_1)$ is the projection of f(x) along the x_1 direction.

Infinite dimensions Dimension of the function space could be ∞ or it could be finite depending on domain of f: see boolean functions ref.

8.2.2 Standard basis functionals

Let basis function/ direction along x_i be e_i: then by usual notion of inner product, $e_i \perp e_j$.
By geometric intuition, tringle inequality, cosine inequality hold. So, have a inner product vector space!

8.2.3 Restriction to finite length

Consider functionals f(x) which are of finite length, even if you are in an ∞ dimensional vector space : Eg: $f(x) = \sin x$ in $[0, 2\pi]$, not x^{-1} in $[0, 2\pi]$. Otherwise, hard to make sense of triangle inequality.

8.2.4 Other representations

The space of all polynomials can be represented both as a functional space described above, and as a vector space, where each polynomial is represented by the vector formed by its coefficients.

8.3 Euclidian space

R^n with the Euclidian structure (metric, inner product): $\langle a, b\rangle = \sum a_i b_i$.

8.3.1 Geometric properties

For geometric properties of various objects in 3-d euclidian space, see topology ref.
Orthant: a generalization of quadrant.

8.3.2 Box measure

Aka Lebesgue measure. This is the minimum cover measure described in the algebra survey.
There exist sets which are not box-measurable![**Find proof**]

8.3.2.1 Definition

For boxes, this is just the product measure: $m([a, b]) = \prod_{i=1}^{n}(b_i - a_i)$.
Let $B_i(S)$ be a set of disjoint boxes which cover S. In general, $m(S) = \inf \{B_i(S)\}$.

8.3.2.2 Properties

It has all properties of a measure. In addition, observe that $m([a, a]) = 0$. So, the measure of any countable set of points is 0. So, measure of rationals $m(Q) = 0$, whereas $m(R - Q) = 1$.

8.4 Dual vector space V^*

Vector space over F of all continuous linear (not affine) functionals: $V \to F$, with addition op: $(f + g)(x) = f(x) + g(x)$.
Linear functionals $f(x)$ can always be specified as $f^T x$.
If V has inner product, V^* has inner product.
Dual of a dual space includes the original space [**Check**].

This concept finds important applicaitons: Eg: Dual cone, dual norms.

8.4.1 Basis: $\left\{e^i\right\}$

$\left\{e^i : e^i(e_j) = 1\texttt{iff j = i}\right\}$. For the finite dimensional case, this is simply another finite dimensional vector space.

9 Topological properties of space V

For properties which arise when viewed as metric space, eg: compactness, boundedness, connectedness etc.., see topology ref.

9.1 Properties of R^n, C^n

See properties of k-cells in R in analysis of functions over fields ref.

9.1.1 Completeness of R^n, C^n

Take vector seq $(x_i) \to x$. If V is over R or C, $(x_i) \to x$ iff it is a Cauchy seq wrt norm: from equiv property over R and C.

9.2 Dimension of V

This is the maximal size of any linearly independent set of vectors in V.

9.3 Basis of vector space V

$T = \{t_i\} : \forall v \in V : v = \sum a_i t_i$, such that T is linearly independent. Often written as a matrix T, so that we can write $Ta = v$ for any v in V. Any maximal set of independent vectors in V is a basis. All bases (eg T and T') have the same size: otherwise, you would have a contradiction. So, $|T| = dim(V)$.

9.3.1 Orthonormal and standard basis

Orthonormal basis: $t_i^T t_j = 0, \langle t_i, t_i \rangle = 1$. Standard basis: see section on definiton of vectors.
Can get an orthogonal basis using QR making algorithms.

9.4 Subspaces

Aka Linear subspace. For subspaces associated with a linear operator, see linear algebra ref.

9.4.1 Membership conditions

A vector $v \neq 0$ is in a subspace S iff it is some linear combination of its basis Q; so $v = Qx$.
This happens only if $\exists i : v^T q_i \neq 0$: so v is not $\perp Q$ wrt standard inner product.

9.4.2 Invariant subspace

S is an invariant subspace of A if $AS \subseteq S$.

10 Inner products, norms

10.1 Inner products

10.1.1 Properties

Obeys Conjugate symmetry, bilinearity, homogeniety, non negativity, positive definiteness.
Bilinearity: linear in a and b separately: $(\alpha a)^*(\beta b) = \overline{\alpha}\beta a^* b$. ***Range of*** $\langle\rangle$ ***need not be*** $\Re$.
$\langle Ax + x, y\rangle = \langle x, A^* y + y\rangle$.

10.1.1.1 Orthogonality

If $\langle x, y\rangle = 0$, x orthogonal to y.

10.1.1.2 Associated norm

Defines norm $\|x\|^2 = \langle x, x\rangle$.
$\triangle$ ineq holds: Take $\|x - y\|^2 = \langle x - \mathrm{y}, x - y\rangle$, expand it, use cauchy schwartz.

10.1.1.3 2-norm Bound on size

(Cauchy, Schwarz). $|\langle a, b\rangle| \leq \|a\| \|b\|$.

Proof $0 \leq f(d) = \|u + dv\|^2 = \langle u, u\rangle - 2d\langle u, v\rangle + \langle v, v\rangle$.
Minimize f(d) wrt d to get: $d = \langle u, v\rangle\langle v, v\rangle^{-1}$.
So, $0 \leq \langle u, u\rangle - |\langle u, v\rangle|\langle v, v\rangle^{-1}$.

Tightness $|\langle a, b\rangle| = \|a\| \|b\|$ when $\langle a, b\rangle = 0$.

10.1.1.4 General norm-bound on size

(Aka Holder's inequality) For $p, q \geq 1$, $p^{-1} + q^{-1} = 1$: p, q are Holder conjugates; then $|\langle a, b\rangle| \leq \|a\|_p \|b\|_q$: a tight bound.

[**Proof**]: For $p, q > 1$, By Young's ineq,
$|a_i b_i| \leq \frac{|a_i|^p}{p} + \frac{|b_i|^q}{q}$;
$\frac{1}{\|a\|_p \|b\|_q} |\langle a, b \rangle| \leq p^{-1} + q^{-1} = 1$. $\square$
Taking the limiting case as $p \to 1$, we also have the $p = 1, q = \infty$ case.

10.1.2 Standard inner product

$\langle a, b \rangle = b^T a$. Can be generalized to $a, b \in C^m : b^* a$.

10.1.2.1 Geometric interpretation

$\langle a, b \rangle = b^T a = \|a\| \, \|b\| \cos \theta$.
So, orthogonality = perpendicularity.

Proof Prove for 2 dimensions by seeing: $\begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} \|a\| \cos \theta \\ \|a\| \sin \theta \end{bmatrix}$, using $cos(A - B) = \cos A \cos B + \sin A \sin B$.
Consider plane formed by a, b. Get new orthonormal basis Q $[QQ^* = I]$, so that q_1, q_2 span this plane; so $Qe_i = q_i$. The representations of the points a, b wrt this new basis is Qa, Qb; their norm remains same. By the 2 dimensional case, $\langle Qa, Qb \rangle = \|a\| \, \|b\| \cos \theta$. But, $\langle Qa, Qb \rangle = \langle a, b \rangle$ as $QQ^* = I$!

10.1.3 In function spaces

Consider functions with domain $X = [a, b]$, and let p be a probability measure on X: $\int_X \overline{f(x)} g(x) dx = \int_a^b \overline{f(x)} g(x) p(x) dx$ for complex valued $f(x)$; can include weight function W too. This defines norm too. Often scaled to make length of certain basis function vectors to be 1.

10.1.4 Orthogonality

Orthogonality of k vectors $\implies$ mutual independence - else contradiction. Orthogonal vector spaces. Orthogonality among bases $\implies$ orthogonality of vector spaces.

10.1.5 Weighted Inner product

Invertible matrix W, $A = W^* W$, A +ve semidefinite. Skew vectors before dot product: $\langle a, b \rangle_W = \langle Wa, Wb \rangle = b^T W^* W a = b^T A a$. Sometimes writ as $\langle a, b \rangle_A$. a and b are A conjugate if $\langle a, b \rangle_A = 0$.

10.1.6 Specify inner product using Gram matrix G

Aka Gramian matrix. Take symmetric +ve semidefinite G. $G_{i,j} = \langle x_i, x_j \rangle$ for some $\{x_i\}$; so $G = X^T X$ for $X = (x_i)$. Then, for any $u = \sum c_i x_i, v = \sum b_i x_i$, can rewrite in $\{x_i\}$ basis as c, b and find $\langle ., . \rangle$ using $\langle c, b \rangle = c^T G b$. As G +ve

semi-def, $c^T Gc \geq 0$; $c^T Gc = 0$ iff $c^T X^T Xc = 0$ or Xc=0: meaning of normness preserved.
G often normalized to make $G_{i,i} = 1$.
Extension to ∞ dimensions: Mercer's theorem.

10.2 Norms

10.2.1 Semi-norm properties

Aka pre-norm. Obeys triangle inequality, non negativity, homogeniety / scalability ($\|cv\| = |c| \|v\|$). But, not necessarily +ve definiteness ($\|x\| = 0$ iff $x = 0.$).

10.2.2 Norm: Defining properties

A seminorm which obeys **positive definiteness** ($\|x\| = 0$ iff $x = 0.$). Often triangle inequality is the only non trivial property to verify. Prenorm omits trinagle inequality.
All norms are metrics (see topology ref).

10.2.3 Variants

10.2.3.1 Absolute norms

$\|a\| = \||a|\|$.

Monotonicity connection $\|.\|$ monotone ($|x| \leq |y| \implies \|x\| \leq \|y\|$) iff it is absolute.
Pf: if monotone, take $y = |x|$; as $|y| = |x|$ get absoluteness. If absolute: take x; for $\alpha \in [0, 1]$, replace $x_i \to \alpha x_i$ to get x', replace $x_i \to -x_i$ to get x"; $\|x'\| = \left\|2^{-1}(1-\alpha)x'' + 2^{-1}(1-\alpha)x + \alpha x\right\| \leq 2^{-1}(1-\alpha)\|x''\| + 2^{-1}(1-\alpha)\|x\| + \alpha\|x\| = \|x\|$; by repetition get $\|(\alpha_i x_i)\| \leq \|x\|$; thus, for $|x| \leq |y|$: $\|x\| = \left\|(\alpha_i e^{i\theta_i} y_i)\right\| = \|(\alpha_i |y_i|)\| \leq \|y\|$.

10.2.4 Dual norm of pre-norm f

Take pre norm f, $f^D(y) = \sup_{x: f(x) \leq 1} \Re\langle x, y\rangle$: like the matrix norm. This is a norm defined on the dual space of $dom(f)$; so it measures the size of continuous linear functionals operating on $dom(f)$; so it is an operator norm - more precisely a functional norm. It is the maximum value attained by $y(x)$ in the unit ball defined by f.
For finite dimensional spaces, dual of dual norm is the original norm. [**Find proof**]

10.2.4.1 Normness proof

+ve definiteness comes from +ve definiteness of $\langle\rangle$. $\triangle$ inequality easily shown from linearity of $\langle\rangle$.

10.2.4.2 Geometric view

The greatest inner product with y of any x in the unit ball $\{x : f(x) = 1\}$.

10.2.4.3 Importance

Plays an important role in describing duality theory in optimization. Also important in describing the subdifferentials of many norms at 0.

10.2.4.4 Common dual norms

Dual of $\|.\|_1$ is $\|.\|_\infty$, and vice-versa: easily from definition.
Dual of $\|.\|_2$ is itself: from Cauchy Schwartz: $z^T x \leq \|z\| \|x\|$ is thight.

Applications Steepest descent method in optimization.

10.2.5 Properties

10.2.5.1 New norms out of old

Take norms $f_1..f_m$ on F; norm g on R with $g(x) \leq g(x+y)$ (monotonicity); then $\|x\| = g(f_i(x))$ also norm: Monotonicity ensures $\triangle$ ineq.
So, sum/ multiples/ max of norms is a norm.
Take $\|\|$: $\|x\|_T = \|Tx\|$ also norm for non singular T.
If vector space finite dimensional, $\|.\|, \|.\|'$ equivalent: see vector function properties. So, if vector seq $(x_k) \to x$ wrt 1 norm, it converges to the same vector wrt another.

10.2.5.2 Convexity of the norm function

Follows from homogeneity and $\triangle$ ineq. $\|\alpha a + (1-\alpha)b\| \leq \|\alpha a\| + \|(1-\alpha)b\| = \alpha \|a\| + (1-\alpha) \|b\|$.

10.2.5.3 The closed unit ball

$\{x : \|x\| = 1\}$. This is closed and bounded, so compact; has non empty interior; is convex; is symmetric about the origin. Conversely, any region with these properties is unit ball B of the norm $\|x\| = 1/(\sup_t (tx \in B))$.

10.2.5.4 Isometry for a norm

$B : \|Bx\| = \|x\|$. Set of isometries form a group: isometry group. [**Find proof**]

10.2.5.5 Equivalence of norms in finite dimension

$\forall \|.\|_1, \|.\|_2 : \exists a, b : \forall x : a\|x\|_1 \leq \|x\|_2 \leq b\|x\|_1$. [**Find proof**]So, one norm can be approximated by another within a factor of b/a.
Any norm can be approximated by a quadratic norm within a factor of $\sqrt{n}$.

Comaprison of p norms $\|x\|_\infty \leq \|x\|_p \leq \|x\|_1$. $\|x\|_1 \leq \sqrt{m}\|x\|_2$: By induction. $\|x\|_p \leq m^{\frac{1}{p}}\|x\|_\infty$.

10.2.6 p norm for p atleast 1

$\|x\|_p = (\sum |x_i|^p)^{\frac{1}{p}}$.

10.2.6.1 Normness proof

Only $\triangle$ inequality proof needs some steps.
(Minkowski): $\|a+b\|^p = \sum_i |a_i+b_i||a_i+b_i|^{p-1} \leq \sum_i |a_i||a_i+b_i|^{p-1} + \sum_i |b_i||a_i+b_i|^{p-1}$
$\leq (\|a\|_p + \|b\|_p)(\sum_i |a_i+b_i|^{(p-1)(\frac{p}{p-1})})^{\frac{p-1}{p}} = (\|a\|_p + \|b\|_p)\frac{\|a+b\|_p^p}{\|a+b\|_p}$.

10.2.6.2 The closed unit ball

Progression of shapes: $\|x\|_1$:rhombus, $\|x\|_2$:circle, $\|x\|_p$, $\|x\|_\infty$: square. 1 norm is $max_{i\in N}\|i\|_i$.
So, p norm not unitarily invariant. Take x; Ux, where U unitary; this is a combination of rotations and reflections; so projection Ux along various axes is different from that of x; so length differs: visualize with rhombus, square etc..

10.2.6.3 2 norm

Aka euclidian norm, $\|x\|_2 = \sqrt{x^*x}$. So, squared euclidian norm, $\|x\|_2^2$ corresponds to $\langle a, b\rangle := b^*a$.
For complex x: $\|x\|^2 = \|Re(x)\|^2 + \|Im(x)\|^2$.

Add orthogonal vectors (Pythagoras theorem) If $\langle x, y\rangle = 0, \|x+y\|_2^2 = \|x\|_2^2 + \|y\|_2^2$. Follows just from definition of the 2 norm.

10.2.6.4 1, infty norms

$\|x\|_\infty$: Max norm, aka Chebyshev norm. 1 norm: manhattan distance.

10.2.7 Lp norms in function spaces Wrt measure p

$(\int_X |f(x)|^p dv)^{1/p} = (\int_X |f(x)|^p p(x)dx)^{1/p}$. This is well defined if f is p-power integrable.

10.2.8 $p \in (0, 1)$ non-norms

Does not satisfy the triangle inequality. Not a norm, maybe a prenorm.

10.2.8.1 Zero (not a) norm

$\lim_{p \to 0} \|x\|_p^p$; if $0^0 := 0$: $\|x\|_0 = \sum_i x_i^0$: number of non zero elements.

10.2.9 Weighted p norm

Aka quadratic norm.
Corresponds to Weighted Inner product, W. $A = W^*W$ hermitian +ve definite.
If W a diagonal matrix: $\|x\|_{2,W}$, where W stretches x, shaped like ellipse.
Sometimes writ as $\|x\|_A$.

10.2.10 Non-norm from convex function f

(Bregman divergence). $d_f(x, y) = f(x) - f(y) - (x - y)^T \nabla f(y)$: Error at x in 1st order approximation of f rooted at y.
Not necessarily a norm: in general, not symmetric. $d_f(x, y) \geq 0$ from gradient inequality.
If $f(x) = \|x\|_2^2 : d_f = \|.\|_2$.
If $f(x) = \sum_i x_i \log x_i - x_i$: with -ve entropy term, get (Kullback, Leibler) $KL(x, y) = \sum x_i (\log \frac{x_i}{y_i} - x_i - y_i)$: has relative entropy term. For special case when applied to probability distrib vectors, see probability ref.

Part III

Vector functionals

11 Functionals and scalar functions over vector spaces

11.1 Introduction

11.1.1 Definition

Let $dim(V) = d$. Any $f : V \to F$ is called a functional.
If V is a Euclidian space, you have a scalar field: a scalar value attached with every value in the space.
Visualization, topological properties explored in another section.
Gradient of a scalar field is a vector field, obtained by taking derivative at all points.

11.1.2 Restriction to a line

Take $f : R^n \to R$, get $g : R \to R$ defined by $g_{x,u}(t) = f(x + tu)$. Thus the domain of g is dom(f) restricted to a single line in the vector space.

11.1.2.1 Importance

This is an important way of studying the properties of a functional. For example, the differential function corresponding to f can be defined by means of the differential function of g.

11.2 Properties

11.2.1 Superclasses

Properties of the more general class of real valued functions is described in the complex analysis survey.
Properties of the more general class of functionals over vector spaces are described in another part of the vector spaces survey.

11.2.2 Conjugate of f

$f'(y) = \sup_{x \in dom(f)}(y^T x - f(x))$. How far below the hyperplane $y^T x$ through 0 does f(x) go? Visualize for $f : R \to R :$ xy is a line, take intercept of farthest line which is tangent to f(x) and is parallel to xy.
Easy to make mistakes while finding conjugate: errors in specifying values of f at every point in domain, difficulty in maintaining rigor while maximizing some functions.
Observe: This is affine in terms of y.

11.2.3 Algebraic properties

11.2.3.1 Minimax vs maximin

Consider $\min_x \max_y f(x, y)$. This is not always equal to $\max_y \min_x f(x, y)$. Pf: Take binary x, y, Consider: $f(x, y) = [\begin{smallmatrix} 1 & 3 \\ 4 & 1.5 \end{smallmatrix}]$.

11.2.3.2 Homogeneity with degree a

$f(tu) = |t|^a f(u)$.

11.2.4 Domain: Important domains

11.2.4.1 Sub-level set

$\{x | f(x) \leq c\}$.

11.2.4.2 Stationary point

Local minima and maxima.

11.2.4.3 Critical points

Points of inflection, or maxima or minima: f'(x) = 0.

11.2.4.4 Equivalence

For finite dim V, take +ve definite, homogenous, continuous function f_1, f_2, $\exists a, b : af_1(x) \leq f_2(x) \leq bf_1(x)$. Pf: Take $\|\|_2$ unit sphere surface: compact set, $h = f_2/f_1$; use Weierstrass, take min, max of h as a, b.

11.2.4.5 Associated sets: Epigraph and subgraph

Given $g : R^D \to (0, B)$, can consider subgraph:
$\left\{(x, t) : \forall x \in R^D, t \in R :: g(x) \geq t\right\}$; visualize easily for $f : R \to R$: area below curve of f. Similar is epigraph.

11.3 Topological properties of functionals

Consider the properties of the more general class of real valued functions in the complex analysis survey.

11.3.1 Visualization

11.3.1.1 Plot in d+1 dim

Need d+1 dimensions. This is $\{(x, f(x))\}$. Even there, can consider restriction to various lines: eg: $f(x)\forall x : x_i = c$.

Tangent hyperplanes to epigraph Take a C-1 smooth point x_1. Then, the plot of the first order
approximation of f at x_1, $\left\{(x, y) | y = f(x_1) + \nabla f(x_1)^T (f(x) - f(x_1)\right\}$ is the tangent hyperplane to $\{(x, f(x))\}$; and $(\nabla f(x_1), -1)$ is the normal to this hyperplane.
For convex f, this becomes the supporting hyperplane, a universal lower bound to f. This concept can even be extended to non-differentiable points: see section on supporting hyperplanes to convex functionals.

11.3.1.2 Contour surfaces in d dim

Else, to visualize in d dimensions, can use contour lines corresponding to level sets $f(x) = a$: A 2 dim contour line for 2 dim functional. This is a d dimensional (possibly closed) surface: can view as a relation $x_d = h(x_1, ..x_{d-1})$. This is the boundary of the sublevel set.

For high-dimensional objects, see topology ref.
Also see geometric properties of the gradient to observe connection between the gradient (and subdifferentials) and sublevel-sets.

11.3.2 Bounding steepness

Take $f(x_1) - f(x_2) = \nabla f(x_3)(x_1 - x_2)$ using MVT, upper bound $\nabla f(x_3)$, and ye got a lipshcitz function (See toplogy ref).

11.3.3 Measure flexibility

Aka Rademacher average. Take Rademacher RV $a_i \to \{\pm 1\}$.
Given $\hat{R}_{\{X_i\}}(F) = E_a[\sup_{f \in F} n^{-1} \sum_i a_i f(X_i)]$. Sees how well functions in F can match sign of $\{a_i\}$ on these points; thence measures complexity/ flexibility of F.

11.4 Sequences, series of functionals on metric space X

Also see topology ref.
Series is just a sequence of partial sums.

11.4.1 (Weierstrass) M test

Let $|f_n| \leq M_n$. If (M_n) converges, $|\sum f_n|$ converges: use Cauchy criterion: $\forall \epsilon, \exists N : m, n > N \implies \sum_m^n |f_i| \leq \sum_m^n M_i \leq \epsilon$.

11.4.2 Space C(X) of continuous bounded complex valued fns on X

Take $\|f\| = \sum_{x \in X} |f(x)|$: positive definiteness, $\triangle$ ineq hold; so this defines a metric on C(X): $d(f, g) = \|f - g\|$. Like an $\leq |X|$ dimensional vector space.
In context of C(X), $lt_{n \to \infty} f_n \to$f$\equiv f_n \to f$ uniformly.
C(X) is a complete metric space. Pf: Take Cauchy sequence in C(X); so $\exists f : f_n \to f$; so $f_n \to f$ uniformly; thence see f is bounded, continuous; so $f \in C(X)$.

11.4.3 f as limit of uniformly convergent sequence of polynomials

(Stone, Weierstrass). f continuous. Polynomials p_n. Express f in terms of the basis functions p_n.

[**Incomplete**]

12 Differential function

12.1 Definition

12.1.1 Fixed direction differential fn

Aka directional derivative.
Fixing the direction v, $D_v(f)$ can be taken to map x to $D_v(f)(x)$. So, $D_v(f)(x) : V \to F$ is a constricted version of the differential function $D(f)(x)$.
$df(x;h) = D_h(f)(x) = \lim_{\Delta t \to 0} \frac{f(x+\Delta th)-f(x)}{\Delta t} = \frac{d}{dt}|_{t=0} f(x+th)$. Aka Gateaux differential.
Alternate notation: $\nabla_h(f(x))$: not the gradient vector, but its applicaiton in a certain direction.

12.1.1.1 Affine approximation view

This definition of the directional derivative is equivalent to the defining $D_h(f)$ as the function such that the following holds: $t \to 0$, $f(x+th) = f(x) + tD_h(f)(x)$.

12.1.1.2 R to R case

In this special case, there is just one direction: 1.

12.1.2 Directional differentiability

If, at x, the directional derivative exists in all directions, f is said to be Gateaux differentiable at x.
The differential of f at the point x in the direction v is a function of two variables: x, v. We regard $D(f)(x) : V \to F$, such that $D(f(x))[v] = \frac{df(x+tv)}{dt}$ is the directional derivative of f at x along v.
So, $D(f) : V \to L(V, F)$, where $L(V, F)$ is the space of continuous linear functionals $l : V \to F$. The fact that $D(f(x))$ is a linear functional follows from the affine approximation view of the directional derivative.
But, this is unsatisfactory as directional differentiability does not imply continuity. [**Find proof**]

12.1.3 Continuous differentiability

If at x, $\exists a$ such that $\forall c$, $\left\|f(x+c) - f(x) - a^T c\right\| = o(\|c\|)$, then f is differentiable at x; and the derivative is $Df(x)[c] := a^T c$, which maps $V \to F$. ***A measure of goodness of affine approximation!*** The view $D(f) : V \to L(V, F)$ still holds.
Aka Frechet derivative, total derivative.

12.1.3.1 Connection to directional differentiability

In non pathological cases, both notions of differentiability are equivalent: This comes from applying the polynomial approximation theorem for $g : R \to R$, $f(x+th) \to f(x) + tD_h(f)(x)$.
In the case of continuous differentiability, this follows from definition. In the case of directional differential functions, this can be seen using the polynomial approximation theorem for $f : R \to R$: $f(x+th) \to f(x)+tD_h(f)(x)$ as $t \to 0$.

12.1.4 Matrix functionals

Similar definition for differential functions for functionals over the vector space of matrices. Eg: See $\nabla tr(f(X))$ in linear algebra ref.

12.2 Linearity

The differential operator $D : f \to D(f)$ is linear: So $D(f+g) = D(f) + D(g)$: This follows from the affine approximation view of the differential function.
Note that this is separate from directional linearity.

12.3 Connection to partial derivatives

We suppose that linearity is established (simple in case of Frechet derivatives). From linearity, $D(f(x))[v] = \sum_i v_i D(f(x))[e_i]$. This can be written as a vector product: $D(f(x))e_i$, with D(f(x)) being a row vector. When written as a column vector, it is denoted by $\nabla(f(x))$, in which case, $\nabla_v f(x) = \nabla(f(x))^T v$.

12.3.1 Notation

$\nabla f(x) := \frac{df(x)}{dx} := (\frac{\partial f(x)}{\partial x_1}, \dots) = (\frac{\partial f(x)}{\partial x_1}, ..)$.

12.3.1.1 Note about representation

Note that, as explained there, 'gradients' are defined wrt to vectors - without differentiating between their representation as row or column vectors. Such representations are secondary to the correctness of their values, and can be altered as necessary for convenience of expression.

12.3.2 D(f) as a Vector field

Hence, the derivative operator $D(f)(x)$ can be viewed as a vector field, such that $D(f)(x) = \nabla f(x)$, a vector. However, often, following the convention used for vector to vector functions, $D(f)(x)$ is denoted by the row vector $\nabla f(x)^T$.

12.3.3 C1 smoothness

$f \in C^1$ if $\frac{\partial f}{\partial x_i}$ exists. Similarly, C^n, even C^∞ smoothness defined.

12.3.3.1 Differentiability vs smoothness

Gradient's existence does not guarantee differentiability; derivative must exist in all directions - in an open ball around c.

12.3.4 In contour graph

Perpendicular to contours

∇f is d dimensional vector. Always $\perp$ to every tangent to the contour of f in d dimensional space: else could move short distance along contour and increase value of f; or take x and $x+\epsilon$ on contour, take Taylor expansion: $f(x+\epsilon) = f(x) + x^T \nabla f(x)$; thence get $x^T \nabla f(x) = 0$.

Sublevel sets and gradient direction

Consider level-sets $f(x) = 0, f(x) = 0.1, f(x) = 0.2$. $\nabla f(x)$ will be oriented towards increasing $f(x)$, that is, away from the interior of the sublevel set $\{x : f_i(x) \leq 0\}$. So, points outwards if convex.

12.3.5 In the plot

Take the plot $(x, f(x))$. Then $\nabla f(x)$, if it exists, is sufficient to specify the tangent hyperplane to the plot at x: see subsection on tangent hyperplanes.

12.4 Subgradients at convex points

Extension of the gradient to non-differentiable functional f(x). See convex functional section.

12.5 Differential operator

Its general properties, including linearity, product rule and the chain rule, are considered under vector functions.

12.5.1 Derivatives of important functionals

For simplicity in remembering the rules it is easier to think in terms of the Differential operator, rather than the gradient (which is just $Df(x)^T$).

12.5.1.1 Linear functionals

$DAx = A : \nabla Ax = A^T, \nabla b^T x = b$ from Df(x) rules.

12.5.1.2 Quadratic functionals

$\nabla x^T Ax = (A^T + A)x$: [**Proof**]: expanding $(x+\delta x_i)^T A(x+\delta x_i)$. $\Box$ Alternate [**Proof**]: $D(x^T Ax) = x^T A + D(x^T A)x$ (product rule) $= x^T A + x^T D(A^T x) = x^T(A + A^T)$ $\Box$
If $A = A^T$: $D(x^T Ax) = x^T(2A)$.

12.6 Higher order differential functions

12.6.1 Definition

12.6.1.1 Linear map from V

Take the differential functional $D(f)(x) : V \to L(V, F)$. $L(V, F)$ is itself a vector space, and the space of continuous linear maps $L(V, L(V, F))$ is well defined. So, we can consider the differential function of $D(f)$. It is $D^2(f)(x) : V \to L(V, L(V, F))$.
Similarly, kth order differential function $D^k(f)(x)$ can be defined in general.
Differential operators, of which $D^k f(x)$ are special cases, for general functions between vector spaces are described elsewhere.

12.6.1.2 Directional higher order differential fn

With u fixed, $D_u(f)(x) = D(f)(x)[u]$ can be viewed as a functional: $D(f) : V \to F$. Once can consider the differential function of $D_u(f)$. Applying the definition, will be $D(D_u(f)) : V \to L(V, F)$ such that $D(D_u(f))(x)$ is specified by $D(D_u(f))(x)[v] = lt_{\Delta t_v \to 0} \frac{D_u(f)(x+\Delta t_v v) - D_u(f)(x)}{\Delta t_v} =$
$lt_{\Delta t_v, \Delta t_u \to 0} \frac{f(x+\Delta t_v v+\Delta t_u u) - f(x+\Delta t_u u) - f(x+\Delta t_v v) + f(x)}{\Delta t_u \Delta t_v}$
$= \frac{\partial^2}{\partial^2 t_u t_v}|_{t_u, t_v = 0} f(x + t_u u + t_v v)$.

12.6.1.3 Multi-Linear map from V^k

Note that, as defined here, $D^2(f)(x)[u]$ is a continuous linear functional, which when provided another argument $D^2(f)(x)[u][v]$ maps to a scalar.
So, using an isomorphism, it is convenient to view $D^2(f)(x) : V^2 \to F$.
Hence, $D^2(f) : V \to L^2(V, F)$, where $L^k(V, F)$ is the space spanned by k-linear maps $g : V^k \to F$. So, $D^2(f)$ maps each point x to a bilinear map.
Similarly, kth order differential functions can be defined in general.

12.6.2 Properties

12.6.2.1 Symmetry

$D^k f(x)$ is symmetric, except in pathological cases which can be eliminated by a good definition. This may follow by looking at the form of $D^2 f(x)[u, v]$ described earlier: $lt_{t_i \to 0} \frac{f(x+\sum t_i v_i)}{t_1 t_2}$.

12.6.2.2 Wrt basis vectors

The notation $D^2f(x)[e_i][e_j] = D_{ij}f(x)$ is used.

12.6.3 Tensor representation

$D^2f(x)[u][v] = \sum_{i,j} u_i v_j D^2_{i,j} f(x)$. [**Proof**]: By the distributive property of multilinear functions. This can also be proved by applying the chain rule, the directional linearity of the differential function and the linearity of the differential operator. □
Similarly $D^k f(x)$ can be completely specified using kth order derivatives along the basis vectors.

12.6.3.1 2nd order case

In the 2nd order case, this is aka Hessian matrix. $H_{i,j} = D_i D_j f(x)$: Always symmetric. Aka $\nabla^2 f(x) = \frac{\partial^2 f(x)}{\partial x \partial x^T} = D\nabla f(x)$, using the notation for derivatives of general vector to vector functions.
This matrix is important in tests for convexity at a critical point.

12.7 Polynomial approximation

See the 1-D case in complex analysis ref.
Restrict f to a line $g(t) = f(a + t(x - a))$. The polynomial approximation of this function leads us to: $f(a+v) = f(a) + \sum_{k \in ..n-1} \frac{1}{k!} D^k f(x)[v]^k + \frac{D^n f(c)[v]^n}{n!}$ for some $c \in hull(a, a+v)$ in the line segment.
$D^k f(a)[v]^k$ is often written using the product of k vectors with a k-th order tensor.

12.7.1 Polynomial approximation series

Aka Taylor series. Similarly, in the limit get: $f(x) = \sum_{|a|} D_a f(a)$. Here we have used the multi-index notation described below.

12.7.1.1 Multi-index notation

Take $b \in Z^n_+,\ x \in V$. Then, $b! := \prod b_i!,\ D_b := D_1^{b_1} .., x^b = \prod x_i^{b_i}$.

12.7.2 Connection with extreme values

See optimization ref.

13 Convexity and functionals

13.1 Convex functional f

Aka concave upwards, concave up or convex cup.

13.1.1 Domain, definitions

dom(f) always a convex set. Visualize as a cup. Epigraph of f (see analysis of functions over fields ref) is a convex set.
When they say 'convex function', they mean 'convex functional'.

13.1.1.1 Smoothness along line segment in the domain

Equivalent definition (easy pf): Convex function: $f(tx + (1-t)y) \leq tf(x) + (1-t)f(y)$: comparing f(some pt on segment [x, y]) with similar pt on secant between f(x) and f(y).

Many points extension By induction, can extend this to any convex combintation of n points. Jensen's inequality (see probability ref): simple extension of defn.

Convex sublevelsets The level set $\{x : f(x) \leq \alpha\}$ is a convex set: take $f(x), f(y) \leq \alpha$; take convex combo z, see $f(z) \leq \alpha$. This is important in convex optimization!

13.1.1.2 Extension of domain

For $x \notin dom(x) : \hat{f}(x) := +\infty$. Preserves $f(tx+(1-t)y) \leq tf(x)+(1-t)f(y)$, Simplifies notation. Similar extension for concave functions.

13.1.2 Strict and strong convexity

13.1.2.1 Strict convexity

A cup, not a hyperplane: imposing curvature. $f(ax+(1-a)x) < af(x)+(1-a)f(x) \forall a \in [0, 1]$.

13.1.2.2 Strong convexity with constant m

$\nabla^2 f(x) \succeq mI$. Implies strict convexity: consdier 2nd order approximation of convex function f.
Restricted strong convexity at t wrt C when d restricted to $d \in C$.

13.1.3 f restricted to a line

Take $f : R^n \to R$. f convex iff $\forall t : g(t) = f(x + tv)$ is convex. $x + tv$ is a line in the domain. Good for showing convexity!

13.1.4 Gradient tests

13.1.4.1 First order condition

If $f : R \to R$ differentiable, f convex iff f'(x) non decreasing: $f(y) \geq f(x) + f'(x)(y - x)$. Aka first order condition.

Pf: If f convex, taking the convex combination of y and x: $\forall t: f(x+t(y-x)) \leq (1-t)f(x)+tf(y)$. So consider $f(y) \geq f(x) + \frac{f(x+t(y-x))-f(x)}{t}$ as $t \to 0$.
Extend to $f: R^n \to R$ by considering f restricted to a line: $f(y) \geq f(x) + \langle \nabla f(x), (y-x) \rangle$. The RHS is the supporting hyperplane to the epigraph of f.

13.1.4.2 First order condition for non-smooth f

The subdifferential set $\partial f(x) \neq \emptyset$ for any x.

13.1.4.3 Second order condition

So, if $f(x)$ twice differentiable, f convex iff $f''(x) \geq 0$ ie: f' non decreasing. By techniques similar to first order condition, f convex if $\nabla^2 f(x) \preceq 0$. Aka second order condition.

13.1.5 Supporting hyperplane to the epigraph

13.1.5.1 Differntiable f

See description of tangent hyperplanes to epigraphs in the 'topological properties of functionals' subsection. For differentiable f, the tangent hyperplane at $(x_1, f(x_1))$ is specified by $(\nabla f(x_1), -1)$. **Pf**: From first order condition: $f(y) \geq f(x) + f'(x)(y-x)$.
So, $f(x) + f'(x)(y-x)$ is a global lower bound on f.

13.1.5.2 Non differentiable convex f: subdifferentials!

Can extend the idea of a gradient to non-differentiable convex f. The **subdifferential set** $\partial f(x) = \{z : \forall x, f(y) \geq f(x) + \langle z, (y-x) \rangle\}$.
Each $z \in \partial f(x)$ is a **subgradient**, a **generalization** of the concept of gradient. $(z, -1)$ is a supporting hyperplane to the epigraph of f at $(x, f(x))$. Also, $f(x) + \langle z, (y-x) \rangle$ is a global lower bound on f.
$\partial f(x)$ is a closed, convex set.
If $x \in int(dom(f))$, $\partial f(x)$ is non-empty and bounded.

Non subdifferentiable functionals Eg: $-\sqrt{x}$: supporting hyperplane at (0, 0) is vertical.

13.1.6 Supporting hyperplanes to sublevel sets

See the description of the connection between contour graphs and the gradient in the section on derivatives of functionals. At differentiable points, gradients define tangent hyperplanes to the contour; for convex functionals, these are also supporting hyperplanes to the sublevel set.
At non differentiable points, the gradient can be substituted with the subgradient.

13.1.7 Finding Subdifferentials

First identify the non-differntiable point, then just apply the definition.
If f is differentiable, $\partial f(x)$ contains just the gradient.

13.1.7.1 Unlike finding gradients

Where f is differentiable, to find $\nabla f(x)$, once can merely compute $\frac{\partial f(x)}{\partial x_i}\forall i$, and arrange the result in a vector.
For finding the subdifferential of f at a non-differentiable point, this does not work. Taking the OR of conditions describing subdifferentials of restrictions of a function to individual axes yields a superset of that function's subdifferential! For example, applying this flawed technique to finding $\partial \left\|0\right\|_2$ yields the unit ball of $\left\|x\right\|_\infty$. ***Can't do it one coordinate at a time.***

13.1.7.2 Of non-negative linear combo

Take $f(x) = ag(x) + a'r(x);\ a, a' \geq 0$. $\partial f(x) \supseteq \{z' + z : z' \in \partial g(x) z \in \partial r(x)\}$: applying dfn.

Of penalties plus regularizers Commonly encountered in statistics and modelling.
If $f(x) = g(x)+r(x)$ is convex, g is differentiable but not r: $\partial f(x) = \{\nabla f(x) + z : z \in \partial r(x)\}$.
eg: Take $f(x) = g(x) + l\left\|x\right\|_1$, where g is convex. Here,
$\partial f(x) = \{\nabla g(x) + lz | \forall i : x_i \neq 0, z_i = sgn(x_i);\ \texttt{else}\ : |z_i| <= 1\}$.

13.1.7.3 Of f(Ax + b)

$\partial f(Ax + b) = A^T \partial f(x)$. Pf: $\partial f(Ax + b) =$
$\{t' : \forall d : f(Ax + b + Ad) \geq f(Ax + b) + \langle t', d\rangle\}$
$= \left\{A^T t : \forall d : f(Ax + b + Ad) \geq f(Ax + b) + \langle A^T t, d\rangle\right\}$
$= A^T \partial f(x)$.
So, linear shift $x \to x + b$ does not change the subdifferential.

13.1.7.4 Of max of functionals

Take $f(x) = \max\{f_1(x)..f_m(x)\}$. Take $act(x) = \{i : f_i(x) = f(x)\}$; then $\partial f(x) = conv(\cup_{i \in act(x)} \partial f_i(x))$.
This property can be used to find subdifferential of the ∞ norm.

13.1.7.5 Subdifferentials of norms

For any norm, subdifferential at 0 is just the unit ball of the corresponding dual norm!
[**Proof**]: $\partial\left\|0\right\| = \{z : \forall d : \left\|0 + d\right\| \geq \left\|0\right\| + \langle d, z\rangle\} = \{z : \forall d : \left\|d\right\| \geq \langle d, z\rangle\}$
$= \{z : \forall \left\|d\right\| = 1 : 1 \geq \langle d, z\rangle\} = \left\{z : 1 \geq \left\|z\right\|^D\right\}$. $\square$

For $\|x\|_{\infty}$, many other points are non differentiable. Using the form of $\partial \max f_i(x)$, we get: $x \neq 0 : \partial \|x\|_{\infty} = conv(\{sgn(x_i)e_i : \forall |x_i| = \|x\|_{\infty}\})$.

13.1.8 Operations which preserve convexity

What preserves the curvature? Let f, g be convex.

13.1.8.1 Sum, max

$a_1 f(x) + a_2 g(x) : a_i \geq 0$ convex.
max(f, g)(x), $sup_{f \in F} f(x)$ convex. Eg: $\sup_{c \in C} \|c - x\|, \lambda_m ax(A)$. This is widely used: eg: concavity of dual function, conjugate function.

13.1.8.2 Minimization over some dimensions

Take f(x, y) convex in both x and y, C a convex set. From defn, $\inf_{y \in C} f(x, y)$ also convex.

13.1.8.3 Composition with affine transform

If f(x) is convex, so is g(x) = f(Ax+b): from defn; Even concavity preserved.

13.1.8.4 Other compositions

Take $h : R^k \to R$ convex. If $g : R^n \to R^k$ componentwise convex, $\hat{h}$ also componentwise non-decreasing, h(g(x)) convex: from defn. Also, if g concave, $\hat{h}$ also componentwise non-increasing h(g(x)) convex.

13.1.8.5 Transformations

Perspective of a functional (not same as perspective fn): See [1].
conjugate of f also convex: supremum of affine functions.

13.1.9 Important instances

13.1.9.1 In R

Affine functional: ax + b (also concave, so linear).
Exponential $e^{ax} \forall a$.
Powers $x^a \forall a \notin (0, 1)$.
Negative entropy: $x \log x$.

13.1.9.2 Matrix functionals

$f(Y) = \log \det Y; \nabla f(Y) = Y^{-1}$ [**Find proof**].

13.1.9.3 Log sum exponents

$f(x) = \log \sum e^{x_i}$ is convex: its Hessian $H \succeq 0$ (can't simply use composition rules!). Pf: Let $z_i = e^{x_i}$; then $\nabla f(x) = (\sum_i z_i)^{-1} z$, Hessian is $H = (\sum_i z_i)^{-2}((\sum_i z_i) diag(z) - zz^T)$. $(\sum_i z_i)^2 x^T H x = (\sum_i z_i)(\sum_i x_i^2 z_i) - (x^T z)^2$. Take $a_i = z_i^{1/2}, b_i = x_i z_i^{1/2}$, use Cauchy schwartz to see $x^T H x \geq 0$.
So, its composition with Affine transformation:
$\log \sum e^{a_i^T x + b_i}$.

Importance The convexity of these functionals is important because they are used to specify the -ve log likelihood functions of exponential families of distributions, and they need to be minimized during maximum likelihood estimation.

13.1.9.4 Other examples

Norms. Each component of an affine function: $(Ax + b)_i$.

13.1.9.5 Convex quadratic functionals

Like this: $f(x) = x^T A x + hx + c$, with $A \succeq 0$. Many level sets are ellipsoids.

13.2 Other Functional-classes defined using convexity

13.2.1 Concave functionals

Concave function: -f is convex, domain is still convex.
Any affine linear function y = ax+b is both concave and convex.
f(x, y) = xy not simultaneously convex in x and y: have a saddle point graph; but individually convex in x and y.

13.2.2 Linear functionals

These are both convex and concave.

13.2.3 Quasi-convex functionals

13.2.3.1 Convex sublevel sets

Any function with convex domain, convex sub-level sets. Eg: $|x|$. Similarly, quasi-concave, quasi-linear functionals are defined.
So, can replace each sublevel set with a sublevelset of a convex functionals. This is used in quasi-convex programming.

13.2.3.2 Smoothness along line segment in the domain

$\forall t \in [0,1] : f(tx+(1-t)y) \leq \max\{f(x), f(y)\}$: else $\max\{f(x), f(y)\}$ sublevelset of f would not be convex. So, there can't be local hills or craters, but there can be plateaus on the way down to any of the global minima.

13.2.3.3 First order condition

f quasiconvex iff $f(y) \leq f(x) \implies \nabla f(x)^T(y-x) \leq 0$: otherwise, could take a small step along the segment connecting x and y, and value of f would be greater than f(x). Visualize with contour maps in case of 2D functional.

13.2.3.4 Importance

Sublevel sets of quasi-convex functions, being convex, can be used to specify feasible region of convex optimization program.

13.2.4 Log concave functional f

$\log f$ is concave.

13.2.4.1 Importance

Many probability distributions log concave.

13.2.4.2 Properties

Directly from concavity of log f: $f(tx+(1-t)y) \geq f(x)^t f(y)^{1-t}$, if f differentiable: $f(x)\nabla^2 f(x) \preceq \nabla f(x) \nabla f(x)^T$.
$g(x) = \int f(x,y)dy$ is also log concave, but sum of log concave functionals not necessarily log concave.
Convolution of f also log concave.

14 Homogenous forms

14.1 As Polynomials

Homogenous Forms refer to homogenous polynomials of degree k. They can be viwed as $f : V^k \to F$.
They can be written as Tensor vector products, as in the case of quadratic forms.

14.1.1 Importance

Differential functions of order k are actually homogenous forms.

14.2 Quadratic form

14.2.1 Representation

$x^*Bx = \sum_{i,j} B_{i,j} x_i x_j$.
Reformulation: $tr(x^*Bx) = tr(Bxx^*)$.

14.2.2 Symmetrification

If $x^*Bx \in R$: As $B = H + H' = \frac{B+B^*}{2} + \frac{B-B^*}{2}$, skew hermitian part can be ignored: $x^*Bx = x^*Hx$.

14.2.3 Connection to triple matrix product

Similarly, in D = ABC has $D_{i,j} = a_{i,:}Bc_i$.

14.3 Generalizations

14.3.1 Monomial

$f(x) = c\prod x_i^{a_i}$.

14.3.2 Posynomial

Sum of monomials. Used to define geometric programming.

15 Other Important functional classes

15.1 Important functionals

15.1.1 Radial basis functionals

$f_c(x) = g(\|x - c\|)$. Gaussian radial basis function is used to define the Gaussian kernel.

15.1.2 Barrier functional

$f(x) \to \infty$ as $x \to bnd(dom(f))$. Eg: $\log(1 - x)$.
Used to charactarize feasible region in optimization problems. Any set in R^n is the domain as a barrier function.

15.2 Kernel function k

Implicitly (perhaps non-linearly) map x to $\phi(x)$ and give $\langle x, x' \rangle$ in that space.

15.2.1 Importance

See kernel trick in statistics ref.

15.2.2 Kernel fn

$k(x,x') = \phi(x')^T\phi(x)$: This can be -ve, but $k(x,x) \geq 0$ for norm notion in ftr space: k must be +ve semi-definite. So its Gram matrix K whose elements are $k(x_n, x_m)$ must be +ve semidefinite for all choices of $\{x_n\}$.

15.2.2.1 Association with kernels of integral transforms

See functional analysis ref. Integral transform: $T_K f(s) = \int_{x_1}^{x_2} K(x,s)f(x)dx$. Inner product $\int f(x)g(x)dx = \sum_{s,t} f(s)g(t)\int K(x,s)K(x,t)dx$: akin to inner product defined by gram matrix, which describes inner products between various basis vectors in the kernel space.

15.2.3 Kernel properties

Linear kernel: $k(x,x') = x^Tx'$. Stationary kernel: $k(x,x') = k(x-x')$; Homogenous kernel: $k(x,x') = k(\|x-x'\|)$.

15.2.4 Some kernels

Polynomial kernel (inhomogenous): $(\langle x,x'\rangle + 1)^d$; homogenous: $(\langle x,x'\rangle)^d$. Hyperbolic tangent: $tanh(\langle kx,x'\rangle + c)$ for some $k > 0, c < 0$.

15.2.4.1 Gaussian kernel

Using gaussian radial basis function:
$k(x,x') = e^{-\|x-x'\|^2/c}$. Everything mapped to the same quadrant in the associated feature space, as $k(x,x') \geq 0$.

15.3 Self concordance

15.3.1 Definition

15.3.1.1 R to R functions

$|D^3f(x)| \leq 2D^2f(x)^{3/2}$. This condition arises out of the need to bound the error term in the quadratic approximation to the functional f.

15.3.1.2 Functionals: restriction to a line

Functional f is self concordant if f restricted to every line is self concordant.

15.3.2 Examples in R

Linear, quadratic functions, -log x.

15.3.3 Invariance to operations

Let f be self concordant (sc).
If $a > 1, af(x)$ also sc. $f(Ax + b)$ also sc.

15.3.4 Importance

Any convex set is the domain of a self concordant barrier functional.

Part IV

Vector functions and relations

16 Vector-valued functions

16.1 Functions across vector spaces

$(y_1 \ldots) = f(x_1, \ldots) : C^n \to C^m$.
Called Operators on vector space by viewing vectors as functions.
If V is a Euclidian space, you have a vector field.
Also see functions over convex and affine spaces.

16.1.1 Functional sequence view

An important way to view a vector function is as a sequence of vector functionals. Thus, many properties of vector functions can be easily understood in terms of the properties of functionals. For example, the differential function is defined by combining differential functions corresponding to the constituent functionals.

16.2 Properties

Also see analysis of functions over fields.

16.2.1 Topological properties

For properties which arise when viewed as a metric space, see topology ref. Continuity properties of functions carry over from the continuity properties of functionals.

16.2.2 Linearity

See subsection on linear functions.

16.2.2.1 Bilinearity

Bilinear function: $f(a+b,c) = f(a,c) + f(b,c)$: if ye hold one ip fixed, ye get linearity wrt other var.
Eg: $f(x,y) = xy$.
Similarly, multilinearity is defined.

Distributive law Just like xy, the distributive law holds for all multilinear functions. [**Proof**]: Easy to see for bilinear function. The multilinear case then follows by induction. $\square$

16.2.3 Generalized convexity

Consider inequalities defined by a pointed cone. If $f(tx+(1-t)y) \preceq tf(x) + (1-t)f(y)$, f is convex. Many properties analogous to convexity of functionals, similarly proved. Epigraph of f is convex. Sublevel-sets of f are also convex.

17 Differential function

Aka derivative.

17.1 Directional differential function at x

This, with the various concepts of the differentiability, is simply defined using the sequence of directional derivatives of the corresponding functionals. See the section on functionals' derivative for details.

17.2 Derivative matrix

17.2.1 Motivation using directional derivatives

For every functional $f_i(x)$, we have $D(f_i)(x)[v] = \langle \nabla f_i(x), v \rangle$. So, this single functional $D(f_i)(x)$ is the row vector from the functional case.

17.2.2 Arrangement as rows

So, due to the definition of the differential function of vector valued functions, $D(f)(x)[v] = Jv$, where $J_{i,:} = D(f_i)(x)$. So, $D(f)(x)$ is completely specified by J, which may remind one of the fact that every linear operator can be represented by a matrix vector product.
This is aka Jacobian matrix. Notation: $J_f(x) = Df(x) = \frac{\partial(y_1 \dots)}{\partial(x_1 \dots)}$: $J_{i,j} = \frac{\partial y_i}{\partial x_j}$.

17.2.3 Note about dimensions

As explained in the case of derivatives of functionals, representations are secondary to the correctness of their values, and can be altered as necessary for convenience of expression. One must however pay attention to them to be consistent with other entities in the same algebraic expression.

17.3 Differential operator

Linearity follows from linearity of functional derivatives.

17.3.1 Row-valued functions

Sometimes, one encounters a function whose component functionals are arranged as a row vector $(f(x))^T$, rather than as a column vector $f(x)$. Though the actual derivative is the same, for the sake of consistency (eg: when it one wants to apply the product rule: $(x^T A)x$ and consider $D(x^T A)$), one can simply compute $[D_x(f(x))]^T$.

17.3.2 Product of functions

From scalar functional derivative product rule: $D_x f(x)^T g(x) = (D_x f(x))^T g(x) + f(x) D_x g(x)$. Note that this results in a column vector.

17.3.3 Composition of functions: chain rule

17.3.3.1 Directional differential functions

Take $h(x) = g(f(x))$. Then $Dh(x)[v] = D(g)[f(x)]D(f)[v]$.
[**Proof**]: We want $Dh(x)$ such that $lt_{t \to 0} g(f(x+tv)) = g(f(x)) + tDh(x)[v]$. We get the result using similar definitions for small t: $g(f(x+tv)) = g(f(x) + tD(f)(x)[v]) = g(f(x)) + tD(g)[f(x)]D(f)(x)[v]$ $\square$

17.3.3.2 In matrix representation

In terms of derivative matrices, this is a matrix product: $D(g)[D(f)(x)[v]] = J_g(f(x))J_f(x)v$! Note that order matters: first differentiate wrt outer function, then wrt inner function.
Observe how the dimensions match perfectly: for functional (function) compositions!

17.3.4 Linear and constant functions

$D(Ax)[v] = Av$, and $D(Ax) = A$: from the affine approximation definition of a derivative. $D(k) = 0$.

17.4 Non-triviality of inversion

COnsider $f(x) = Mx$.
If J is square and M is invertible: $J_{M^{-1}} = \frac{\partial(x_1 \ldots)}{\partial(y_1 \ldots)} = J_M^{-1}$: From inverse function thm [**Find proof**]. So, in general, $\frac{\partial y_j}{\partial x_i} = J_{j,i} \neq \frac{1}{\frac{\partial x_i}{\partial y_j}} = 1/J_{i,j}^{-1}$ unlike 1-D eqn $\frac{dx}{dy} = \frac{1}{\frac{dy}{dx}}$.

18 Notable types

18.1 Functions over N: Sequences of vectors

Many properties carry over from sequences on R or C. See analysis of functions over fields ref.

18.1.1 Convergence

The limit corresponds to the limit of projection sequence in each dimension. Sums, inner products, scalar multiples of convergent sequences converge.
(Bolzano, Weierstrass): Every bounded sequence has a subsequence which converges.

18.2 Curve

Continuous $f : R \to X$, where X is a topological space.

18.2.1 Plane curves

A curve in the Euclidian plane.
Affine plane over field F_q: A 2 D affine space.

18.2.1.1 Elliptic curve

A plane curve with the equation $y^2 = x^3 + ax + b$. The set of points on this curve, with the point at ∞ form a commutative additive group. ∞ point is 0.

18.2.1.2 $(E(F_p), +)$

If defined over an affine plane over field F_p, it is a finite group. 0 is not in affine plane. (Hasse) Number of points in the group is close to the size of F_p. [**Find proof**]
For use in cryptography, see cryptography ref.

18.3 Other V to V functions

18.3.1 Linear function

$f(ax+by) = af(x)+bf(y)$. Aka linear maps/ transformations. Preserves linear combinations of x, y. Note that linear functions do not include affine functions.
Equivalent to matrix multiplication Ax: see other section. See linear algebra ref.

18.3.2 Generalized projection operation

Take $A \subseteq V$, some norm: $\|.\|$. $P_A(u) = argmin_{v\in A}\|u-v\|$.

18.3.3 Perspective function

$P: R^{n+1} \to R^n : P(x,t) = x/t$, $dom(P) = \{(x,t) : t > 0\}$: note domain. Preserves convexity in images, inverse images.

18.3.4 Linear fractional fn

$f(x) = \frac{Ax+b}{c^Tx+d}$, $dom(x) = \left\{x : c^Tx + d > 0\right\}$: composition of affine and persepctive functions..

18.3.5 Affine functions between vector spaces

Aka affine
transformation/ map. ***Linear transformation followed by translations.*** Writeable as $y = Ax+b$; or as $y' = \left[\begin{smallmatrix} y \\ 1 \end{smallmatrix}\right] = \left[\begin{smallmatrix} A & b \\ 0 & 1 \end{smallmatrix}\right]\left[\begin{smallmatrix} x \\ 1 \end{smallmatrix}\right] = Mx'$, a linear transformation in a higher dimension space.
Somewhat preserves all affine combinations: $M(ax+(1-a)x') = aMx+(1-a)Mx' = y$ with last component in y being 1.
Invertible if A is invertible.

18.3.6 Soft-thresholding operator

Aka Winsorization. $f(v,l)_i = [v-l_i]_+$. This operator is often used in describing solutions to l_∞ regularized regression problems.

19 Relations over vector spaces

19.1 Majorization

Take $a, b \in C^m$, rearrange in descending
order to get $\left\{a_{[i]}\right\}, \left\{b_{[i]}\right\}$; and in ascending order to get $\left\{a_{(i)}\right\}, \left\{b_{(i)}\right\}$. $a \preceq b$ (b majorizes a) if $\sum_{i=1}^m a_i = \sum_{i=1}^m b_i, \sum_{i=1}^k a_{[i]} \leq \sum_{i=1}^k b_{[i]} \forall k$. $\equiv$ notion from using ascending order and saying a majorizes b.

19.1.1 Interleaving

If a majorizes $b \in R^n$, $\exists g \in R^{n-1}$ interleaved among a such that g majorizes $b' = b_{1:n-1}$.
Pf: True for 2; suppose $n \geq 2$; take $d \in R^{n-1}$ interleaved among a (ineq A) with $\forall k \in [1, n-2] : \sum_{i=1}^{k} d_{(i)} \leq \sum_{i=1}^{k} b_i$ (ineq B); take their set D; $a' = a_{1:n-1} \in D$, D bounded, closed: so compact; D convex; $\|a'\|_1 \leq \|b'\|_1$; take $d' = argmax_{d \in D} \|d\|_1$, set $g(t) = \|ta' + (1-t)d'\|_1$ is continuous over [0,1]; if $\|d'\|_1 \geq \|b'\|_1$, $\exists t : g(t) = \|b'\|_1$. To show $\|d'\| \geq \|b'\|$: if all ineq B are strict, all of ineq A must be equalities: else $\|d'\| \neq max_d \|d\|$: then, $\|a_{2:n}\| \geq \|b'\|$; if some ineq in ineq B is equality, take r = largest k for which this holds; then $\sum_i^r d'_i = \sum_i^r b_i$, $\forall k > r : d'_k = a_{k+1}$; thence again get $\|d'\| \geq \|b'\|$.

19.1.2 Connection with stochastic matrices

b majorizes a iff $\exists$ doubly stochastic S: $a = Sb$. Lem 1: If b maj a, can make real symmetric $B = Q\Lambda Q^*$ with diag a and ew b; B is normal matrix, so can say a = Sb for doubly stochastic S. Lem 2: Take a=Sb; as PSP' remains stochastic with permutation matrix ops P, P', wlog assume a, b in ascending order; take $w_j^{(k)} = \sum_{i=1}^{k} s_{i,j}$, with $\sum_{i=1}^{n} w_j^{(k)} = k$; see $\sum_{i=1}^{k}(a_i - b_i) = \sum_{j=1}^{n} w_j^{(k)} b_j - \sum_i^k b_i + b_k(k - \sum_{j=1}^{n} w_j^{(k)}) \geq 0$.
So, by Birkhoff, b maj a iff $a = Sb = \sum_i p_i(Pb)$.

19.1.3 Weak majorization

Weak majorization ($\succeq$) if $\sum_{i=1}^{m} a_i = \sum_{i=1}^{m} b_i$ condition omitted.

19.1.3.1 Connection with stochatic matrices

b weakly majorizes $a \geq 0$ iff $\exists$ doubly substochastic Q: $a = Qb$. Pf of if: $\exists$ doubly stochastic S: $Q \leq S$; so $\sum_{i=1}^{k}(Qb)_{[i]} \leq \sum_{i=1}^{k}(Sb)_{[i]} \leq \sum_{i=1}^{k} b_{[i]}$. Pf of $\rightarrow$: Let a have n nz elements; take $d = \sum b - \sum a$; extend b by adding m 0's to get b', extend a by adding m d/m valued entries; then $\exists$ dbl stoch S with $a' = Sb'$; then Q is $n \times n$ principle submatrix.
b weakly majorizes a iff $\exists$ doubly stochastic S: $a \leq Sb$. Pf of $\leftarrow$: If $a \leq Sb, a \preceq Sb \preceq b$. Pf of $\rightarrow$: Pick k to get $a' = a + k1, b' = b + k1$; If $a \preceq b$, for substochastic Q, $a + k1 = Q(b + k1)$; so $a = Qb \leq Sb$ where S is stochastic dilation of Q.

19.1.4 Weak Majorization and convex increasing fn

Take convex increasing scalar function f; b weakly majorizes a; then $f(b)$ weakly majorizes $f(a)$. Pf; For doubly stochastic Q, $a \leq Qb$; using monotonicity, $f(a) \leq f(Qb)$; so $f(a) \preceq f(Qb)$; but by Birkhoff $f(Qb) = f((\sum \alpha_i P_i)b) \leq \sum \alpha_i f((P_i)b) = \sum \alpha_i P_i f(b) = Qf(b)$, where $\sum \alpha_i = 1$; so $f(Qb) \preceq f(b)$.

If $0 \leq a$, $0 \leq b$, with entries in descending order, $\prod_{i=1}^{k} a_i \leq \prod_{i=1}^{k} b_i$, g is such that $g(e^x)$ is convex increasing, then g(b) weakly majorizes ($\succeq$) g(a): $\log a \preceq \log b$; use $f(x) = g(e^x)$; take care of cases where $a_{i>k} = 0$ using induction.

Bibliography

[1] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, March 2004.

[2] Horn and Johnson. *Matrix Analysis*. CUP, 1985.

[3] Horn and Johnson. *Topics in Matrix Analysis*. CUP, 1991.

[4] K. B. Petersen and M. S. Pedersen. The matrix cookbook, oct 2008. Version 20081110.

[5] Gilbert Strang. *Linear Algebra and its applications, 4e*. Thomson Brooks/ Cole, 2006.

[6] Lloyd N. Trefethen and David Bau III. *Numerical Linear Algebra*. Siam, 1997.