

## Notes on Vector and Matrix Norms

These notes survey most important properties of norms for vectors and for linear maps from one vector space to another, and of maps norms induce between a vector space and its dual space.

### Dual Spaces and Transposes of Vectors

Along with any space of real vectors  $\mathbf{x}$  comes its dual space of linear functionals  $\mathbf{w}^T$ . The representation of  $\mathbf{x}$  by a column vector  $x$ , determined by a coordinate system or *Basis*, is accompanied by a corresponding way to represent functionals  $\mathbf{w}^T$  by row vectors  $w^T$  so that always  $w^T x = \mathbf{w}^T \mathbf{x}$ . A change of coordinate system will change the representations of  $\mathbf{x}$  and  $\mathbf{w}^T$  from  $x$  and  $w^T$  to  $\bar{x} = C^{-1}x$  and  $\bar{w}^T = w^T C$  for some suitable nonsingular matrix  $C$ , keeping  $\bar{w}^T \bar{x} = \mathbf{w}^T \mathbf{x}$ . But between vectors  $\mathbf{x}$  and functionals  $\mathbf{w}^T$  no relationship analogous to the relationship between a column  $x$  and the row  $x^T$  that is its transpose necessarily exists. Relationships can be invented; so can any arbitrary maps between one vector space and another.

For example, given a coordinate system, we can define a functional  $\mathbf{x}^T$  for every vector  $\mathbf{x}$  by choosing arbitrarily a nonsingular matrix  $T$  and letting  $\mathbf{x}^T$  be the functional represented by the row  $(Tx)^T$  in the given coordinate system. This defines a linear map  $\mathbf{x}^T = \mathbf{T}(\mathbf{x})$  from the space of vectors  $\mathbf{x}$  to its dual space; but whatever change of coordinates replaces column vector  $x$  by  $\bar{x} = C^{-1}x$  must replace  $(Tx)^T$  by  $(T\bar{x})^T = (Tx)^T C = (TC\bar{x})^T C$  to get the same functional  $\mathbf{x}^T$ . The last equations can hold for all  $\bar{x}$  only if  $T = C^T T C$ . In other words, the linear map  $\mathbf{T}(\mathbf{x})$  defined by the matrix  $T$  in one coordinate system must be defined by  $\bar{T} = C^T T C$  in the other. This relationship between  $T$  and  $\bar{T}$  is called *Congruence* (Sylvester's word for it).

Evidently matrices congruent to the same matrix are congruent to each other; can all matrices congruent to a given matrix  $T$  be recognized? Only if  $T = T^T$  is real and symmetric does this question have a simple answer; it is *Sylvester's Law of Inertia* treated elsewhere in this course.

The usual notation for complex vector spaces differs slightly from the notation for real spaces. Linear functionals are written  $\mathbf{w}^H$  or  $\mathbf{w}^*$  instead of  $\mathbf{w}^T$ , and row vectors are written  $w^H$  or  $w^*$  to denote the complex conjugate transpose of column  $w$  instead of merely its transpose  $w^T$ . (Matlab uses "w.'" for  $w^T$  and "w'" for  $w^*$ .) We'll use the  $w^*$  notation because it is older and more widespread than  $w^H$ . Matrix  $T$  is congruent to  $C^* T C$  whenever  $C$  is any invertible matrix and  $C^*$  is its complex conjugate transpose. Most theorems are the same for complex as for real spaces; for instance *Sylvester's Law of Inertia* holds for congruences among complex *Hermitian* matrices  $T = T^*$  as well as real symmetric. Because many proofs are simpler for real spaces we shall stay mostly with them.

Not all maps from a vector space to its dual have been found useful; some useful maps are not linear. Among the most useful maps, linear and nonlinear, are the ones derived from the *metrics* or *norms* associated with different notions of *length* in vector spaces. Applications of these norms and of their derived maps are the subject matter of the following notes.

## Norms

A *norm* is a scalar function  $\|\mathbf{x}\|$  defined for every vector  $\mathbf{x}$  in some vector space, real or complex, and possessing the following three characteristic properties of *length* :

Positivity:  $0 < \|\mathbf{x}\| < \infty$  except that  $\|\mathbf{o}\| = 0$  .

Homogeneity:  $\|\lambda\mathbf{x}\| = |\lambda| \|\mathbf{x}\|$  for every scalar  $\lambda$  .

Triangle inequality:  $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$  . ( Equality need not imply parallelism!)

**Exercise 1:** Prove  $|\|\mathbf{w}-\mathbf{x}\| - \|\mathbf{y}-\mathbf{z}\|| \leq \|\mathbf{w}-\mathbf{y}\| + \|\mathbf{x}-\mathbf{z}\|$  for any  $\mathbf{w}, \mathbf{x}, \mathbf{y}, \mathbf{z}$  in a normed space.

Three examples of norms defined on the space of column vectors  $\mathbf{x}$  with elements  $\xi_1, \xi_2, \dots, \xi_n$  are  $\|\mathbf{x}\|_p := (\sum_k |\xi_k|^p)^{1/p}$  for  $p = 1$  or  $2$  , and  $\|\mathbf{x}\|_\infty := \max_k |\xi_k|$  . ( Can you verify that these three  $\|\dots\|_p$  are norms? The triangle inequality is the hard part; see below.) In this course we shall discuss mostly these three norms, but there are lots of others. Every *nonsingular* linear operator  $\mathbf{L}$  converts one norm  $\|\mathbf{x}\|$  into another norm  $\|\mathbf{x}\| := \|\mathbf{L}\mathbf{x}\|$  . ( Why nonsingular?) Also the maximum of two norms is a third and the sum of two norms is another. ( Can you see why?)

## The Norm's Unit-ball $\Omega$

Every norm has its own *Unit-ball*  $\Omega$  defined as the set of all vectors  $\mathbf{x}$  with  $\|\mathbf{x}\| \leq 1$  . Some writers use the words "Unit-sphere" to mean what we call its *boundary*  $\partial\Omega$  , consisting of all the norm's *unit vectors*  $\mathbf{u}$  with  $\|\mathbf{u}\| = 1$  . Our unit ball  $\Omega$  turns out to be a bounded closed centrally symmetric convex body with an interior:

"Bounded" means for every  $\mathbf{x} \neq \mathbf{o}$  that  $\lambda\mathbf{x}$  lies outside  $\Omega$  for all  $\lambda > 1/\|\mathbf{x}\|$  .

"Closed" means that  $\Omega$  includes its boundary  $\partial\Omega$  .

"Centrally Symmetric" means that if  $\mathbf{x}$  lies in  $\Omega$  so does  $\lambda\mathbf{x}$  whenever  $|\lambda| = 1$  ;  
for real vector spaces  $\lambda = \pm 1$  .

"Convex" means that if  $\mathbf{x}$  and  $\mathbf{y}$  both lie in  $\Omega$  then so must the line segment traced by  $\lambda\mathbf{x} + (1-\lambda)\mathbf{y}$  for  $0 \leq \lambda \leq 1$  ; it's because  $\|\lambda\mathbf{x} + (1-\lambda)\mathbf{y}\| \leq \lambda\|\mathbf{x}\| + (1-\lambda)\|\mathbf{y}\| \leq 1$  .

"Interior" to  $\Omega$  is where  $\mathbf{o}$  lies; this means for every  $\mathbf{x}$  and all nonzero  $\lambda$  chosen with  $|\lambda|$  tiny enough ( smaller than  $1/\|\mathbf{x}\|$  ) that  $\lambda\mathbf{x}$  lies in  $\Omega$  too.

Conversely, given a bounded closed centrally symmetric convex body  $\Omega$  with an interior, a norm  $\|\dots\|$  can be so defined that  $\Omega$  is its unit-ball. In fact, define  $\|\mathbf{o}\| := 0$  and for nonzero vectors  $\mathbf{x}$  define  $\|\mathbf{x}\|$  to be that positive value of  $\xi$  that puts  $\mathbf{x}/\xi$  on the boundary  $\partial\Omega$  . Such a  $\xi$  must exist because  $\mathbf{x}/\xi$  lies interior to  $\Omega$  for all  $\xi$  big enough, and lies outside  $\Omega$  for all  $\xi$  tiny enough. Central symmetry implies homogeneity of  $\|\dots\|$  . Convexity of  $\Omega$  implies the triangle inequality thus: For any nonzero  $\mathbf{x}$  and  $\mathbf{y}$  we know that  $\mathbf{x}/\|\mathbf{x}\|$  and  $\mathbf{y}/\|\mathbf{y}\|$  both lie on  $\partial\Omega$  . Therefore  $\lambda\mathbf{x}/\|\mathbf{x}\| + (1-\lambda)\mathbf{y}/\|\mathbf{y}\|$  lies in  $\Omega$  whenever  $0 < \lambda < 1$  , and surely lies there if  $\lambda = \|\mathbf{x}\|/(\|\mathbf{y}\| + \|\mathbf{x}\|)$  , whereupon  $\|\lambda\mathbf{x}/\|\mathbf{x}\| + (1-\lambda)\mathbf{y}/\|\mathbf{y}\|\| \leq 1$  and the rest follows easily.

Unit-balls can be very diverse. For instance, the unit-balls  $\Omega_p$  belonging to the norms  $\|\dots\|_p$  defined above for  $p = 1, 2$  and  $\infty$  have very different shapes when the dimension  $n$  is large.  $\Omega_\infty$  has  $2n$  facets and  $2^n$  vertices, whereas  $\Omega_1$  has  $2^n$  facets and  $2n$  vertices, and  $\partial\Omega_2$  is a very smooth sphere in between. To appreciate these shapes draw pictures of  $\Omega_p$  for  $n = 2$  or  $3$  ; then  $\Omega_\infty$  is a square or cube,  $\Omega_1$  is a diamond or octahedron, and  $\Omega_2$  is a circular disk or solid sphere respectively. Shapes like these will predominate in the following notes.

## Continuity and Topological Equivalence of Norms

Despite the diverse shapes of unit-balls, all vector norms  $\|\dots\|$  have many common properties. One is *Continuity*; this is proved by substituting  $\mathbf{w} = \mathbf{y} = \mathbf{o}$  in Exercise 1 above to deduce that  $|\|\mathbf{x}\| - \|\mathbf{z}\|| \leq \|\mathbf{x} - \mathbf{z}\|$ . This shows that  $\|\mathbf{x}\|$  can be kept as close to  $\|\mathbf{z}\|$  as one likes by keeping  $\|\mathbf{x} - \mathbf{z}\|$  small enough, by keeping  $\mathbf{x}$  in a sufficiently tiny ball shaped like  $\Omega$  centered at  $\mathbf{z}$ . But if  $\Omega$  can be arbitrarily needle-like or arbitrarily flattened, why can't  $\|\mathbf{x}\|$  change arbitrarily violently when  $\mathbf{x}$  changes arbitrarily little measured by some other metric? That can happen in infinite-dimensional spaces but not in a space of finite dimension  $n$ , and here is why:

First choose any basis  $\mathbf{B} = [\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_n]$  and then substitute  $\mathbf{b}_j/\|\mathbf{b}_j\|$  for every  $\mathbf{b}_j$  to force every  $\|\mathbf{b}_j\| = 1$ . In this basis the components  $\xi_j$  of any vector  $\mathbf{x} = \sum_j \mathbf{b}_j \xi_j = \mathbf{B}\mathbf{x}$  form a column vector  $\mathbf{x}$ . Define  $\|\mathbf{x}\|_\infty := \|\mathbf{B}^{-1}\mathbf{x}\|_\infty = \|\mathbf{x}\|_\infty = \max_j |\xi_j|$  and  $\|\mathbf{x}\|_1 := \|\mathbf{B}^{-1}\mathbf{x}\|_1 = \|\mathbf{x}\|_1 = \sum_j |\xi_j|$ , two new norms to compare with  $\|\mathbf{x}\| = \|\sum_j \mathbf{b}_j \xi_j\| \leq \sum_j \|\mathbf{b}_j\| |\xi_j| = \sum_j |\xi_j| = \|\mathbf{x}\|_1 \leq n \|\mathbf{x}\|_\infty$ . Then  $|\|\mathbf{x}\| - \|\mathbf{z}\|| \leq \|\mathbf{x} - \mathbf{z}\| \leq \|\mathbf{x} - \mathbf{z}\|_1 \leq n \|\mathbf{x} - \mathbf{z}\|_\infty$ , which confirms that  $\|\mathbf{x}\|$  is a continuous function of the components  $\xi_j$  of  $\mathbf{x}$  and of  $\mathbf{x}$  in *every* basis. (If  $n$  were infinite,  $\|\mathbf{x}\|$  might change arbitrarily violently even though every change in every component  $\xi_j$  of  $\mathbf{x}$  is arbitrarily tiny.)

Because every  $\mathbf{x}$  satisfies  $\|\mathbf{x}\| \leq \|\mathbf{x}\|_1 \leq n \|\mathbf{x}\|_\infty$  the unit-ball  $\Omega$  of  $\|\mathbf{x}\|$  contains the unit-ball  $\Omega_1 = \mathbf{B}\Omega_1$  of  $\|\mathbf{x}\|_1$ , and  $\Omega_1$  contains  $(1/n)\Omega_\infty$ , a fractional copy of the unit-ball  $\Omega_\infty = \mathbf{B}\Omega_\infty$  of  $\|\mathbf{x}\|_\infty$ . (Can you see why?) This phenomenon is typical; given any two norms for a finite-dimensional vector space, some small positive multiple of either's unit-ball always fits inside the other. Here is why:

For any two norms  $\|\dots\|$  and  $|\dots|$ , let's consider the quotient  $\|\mathbf{x}\|/|\mathbf{x}|$ . As  $\mathbf{x}$  runs through all nonzero vectors this quotient sweeps through a range of positive values which is the same range as  $\|\mathbf{u}\|/|\mathbf{u}|$  sweeps out while  $\mathbf{u} := \mathbf{x}/\|\mathbf{x}\|_\infty$  runs through all unit vectors on  $\partial\Omega_\infty$ . Every such  $\mathbf{u} = \mathbf{B}\mathbf{u}$  for a unit column  $\mathbf{u}$  on  $\partial\Omega_\infty$ , and vice-versa, so the range in question is swept out by the quotient  $\|\mathbf{B}\mathbf{u}\|/|\mathbf{B}\mathbf{u}|$  while  $\mathbf{u}$  runs over all of  $\Omega_\infty$ . Two paragraphs ago we saw why  $\|\mathbf{B}\mathbf{u}\|$  must be a continuous function of  $\mathbf{u}$ , and the same goes for  $|\mathbf{B}\mathbf{u}|$ ; and since both norms are positive their quotient is continuous too. Boundary  $\partial\Omega_\infty$  is a closed bounded set in a finite-dimensional space, so every continuous function thereon achieves its maximum and minimum values somewhere on  $\partial\Omega_\infty$ ; in particular the quotient's maximum  $M = \|\mathbf{B}\hat{\mathbf{u}}\|/|\mathbf{B}\hat{\mathbf{u}}| > 0$  and minimum  $\mu = \|\mathbf{B}\hat{\mathbf{u}}\|/|\mathbf{B}\hat{\mathbf{u}}| > 0$  are achieved respectively at some unit columns  $\hat{\mathbf{u}}$  and  $\hat{\mathbf{u}}$  (not determined uniquely). Therefore  $0 < \mu \leq \|\mathbf{x}\|/|\mathbf{x}| \leq M$  for all  $\mathbf{x} \neq \mathbf{o}$ , and each " $\leq$ " sign turns into " $=$ " for some  $\mathbf{x}$ . These inequalities tell us something geometrical about the norms' unit-balls,  $\Omega$  for  $\|\dots\|$  and  $\Omega$  for  $|\dots|$ ; you should confirm now that  $M\Omega$  barely contains  $\Omega$  which barely contains  $\mu\Omega$ . Here "barely" means boundaries touch.

The foregoing paragraph is important for two reasons. First, its style of reasoning will recur. Second, it shows that all finite-dimensional vector norms are *Topologically Equivalent*: if an infinite sequence of vectors converges when distance from its limit is measured in one norm, then convergence occurs no matter what norm is used to measure distance. Do you see why?

(Different norms defined for an infinite-dimensional vector space do not have to be Topologically Equivalent.)

## Lagrange's Identities, and Cauchy's and Hölder's Inequalities

These are stated and proved here for columns  $w = \{w_j\}$  and  $x = \{\xi_j\}$  of complex numbers of which  $\bar{w}_j$  and  $\bar{\xi}_j$  are their complex conjugates and  $\bar{w}_j w_j = |w_j|^2$  and  $\bar{\xi}_j \xi_j = |\xi_j|^2$  their squared magnitudes. First, a generalization of  $\|x\|^2 \cdot \|w\|^2 - |x \cdot w|^2 = \|x \times w\|^2$  from Euclidean 3-space is

$$\text{Lagrange's Identity: } w^* w x^* x - |w^* x|^2 = \sum_j \sum_k |w_j \xi_k - w_k \xi_j|^2 / 2 .$$

It is proved by expanding the double-sum's  $|\dots|^2$  and collecting terms. In matrix terms it says  $w^* w x^* x - |w^* x|^2 = \text{Trace}((wx^T - xw^T) \cdot (wx^T - xw^T)) / 2$ . Another version, more natural, is  $w^* w x^* x - \text{Re}((w^* x)^2) = \text{Trace}((wx^* - xw^*) \cdot (wx^* - xw^*)) / 2$ . Since  $\text{Trace}(M^* M)$  is the sum of squared magnitudes of any matrix  $M$ 's elements it must be nonnegative, whence follows

$$\text{Cauchy's Inequality: } |w^* x| \leq \sqrt{(w^* w x^* x)} = \|w^*\|_2 \|x\|_2 .$$

This becomes equality only if  $w$  or  $x$  is a scalar multiple of the other. Cauchy's Inequality implies (and can be proved equivalent to) the triangle inequality for  $\|x\|_2 = \sqrt{(x^* x)}$  because

$$(\|w\|_2 + \|x\|_2)^2 - \|w + x\|_2^2 = 2(\|w\|_2 \|x\|_2 - \text{Re}(w^* x)) \geq 2(\|w^*\|_2 \|x\|_2 - |w^* x|) \geq 0 .$$

Note the implicit definition of  $\|w^*\|_2 := \|w\|_2 = \sqrt{(w^* w)}$  here. It is a special case. We shall see that other norms  $\|w^*\|$  of rows are not computed from the same formulas  $\|w\|$  as for columns.

Angle  $\arccos(\text{Re}(w^* x) / (\|w\| \cdot \|x\|))$  between vectors  $w$  and  $x$  in a Euclidean or Unitary space is real because of Cauchy's Inequality, which was proved by H.A. Schwarz for integrals as well as sums, and was discovered also by Bunyakovsky; all three names get attached to it. Analogous inequalities apply to  $\|\dots\|_p$  for every  $p \geq 1$ ; its triangle inequality is also called **Minkowski's Inequality**, and its analog of Cauchy's Inequality is called

$$\text{Hölder's Inequality: } |w^* x| \leq \|w^*\|_p \|x\|_q := \|w\|_q \|x\|_p \text{ for } q := 1 + 1/(p-1) \geq 1 .$$

Note that the formula to compute  $\|w^*\|_p$  from row  $w^*$  is *not* the same as the formula to compute  $\|w\|_p$  from column  $w$  unless  $q = p = 2$ ; see below. Class notes on *Jensen's Inequality*, or texts about Normed Spaces, or Functional Analysis, or Inequalities, supply proofs of Minkowski's and Hölder's inequalities, either of which can be deduced from the other. Neither will figure much in these notes for  $p$  and  $q$  different from 1, 2 and  $\infty$ .

## The Dual Norm

Given a norm  $\|x\|$  for a real space of vectors  $x$ , its *Dual Norm* is another norm induced over the dual space of linear functionals  $w^T$  thus:

$$\|w^T\| := \max |w^T x| = \max w^T x \text{ maximized over all } x \text{ in } \Omega .$$

(For complex spaces  $\|w^*\| := \max |w^* x| = \max \text{Re}(w^* x)$  over all  $x$  in  $\Omega$ .) Please do verify that these definitions have all three of the characteristic properties norms must have, and that  $\max |w^* x|$  really equals  $\max \text{Re}(w^* x)$ ; IT'S IMPORTANT! Provided the vector space has finite dimension, the asserted maxima exist because they are maxima of continuous functions of  $x$  over a closed bounded region  $\Omega$ ; but no simple formula for  $\|w^T\|$  need exist. Fortunately, a simple formula does exist for the norms dual to  $\|\dots\|_p$  defined above:

Let row  $w^T = [w_1, w_2, \dots, w_n]$ ; then  $\|w^T\|_p$  turns out to be just  $\|w\|_q$  with  $q := 1 + 1/(p-1)$  for every  $p \geq 1$ , though we care about only  $p = 1, 2$  and  $\infty$ . In these cases observe that

$$\begin{aligned} \|w^T\|_p &= (\max w^T x \text{ over all } x \text{ with } \|x\|_p \leq 1) = (\max w^T u \text{ over all } u \text{ with } \|u\|_p = 1) \\ &= \max_k |w_k| \quad \text{when } p = 1, && \text{(You can verify this easily, so do so.)} \\ &= \sum_k |w_k| \quad \text{when } p = \infty, && \text{(You can verify this easily, so do so.)} \\ &= \sqrt{(\sum_k |w_k|^2)} \quad \text{when } p = 2. && \text{(You can verify this easily, so do so.)} \end{aligned}$$

The case  $p = 2$  follows from Cauchy's Inequality, which is the special case for  $\|\dots\|_2$  of what is called, for norms in general,

**Hölder's Inequality:**  $\mathbf{w}^T \mathbf{x} \leq \|\mathbf{w}^T\| \|\mathbf{x}\|$  for all  $\mathbf{w}^T$  and  $\mathbf{x}$ .

This follows immediately from the definition of the dual norm  $\|\mathbf{w}^T\|$ . Moreover we may verify easily (and you should do so) for all three norms  $\|\dots\|_p$  that

$$\|\mathbf{x}\|_p = \max \mathbf{w}^T \mathbf{x} \text{ over all } \mathbf{w}^T \text{ with } \|\mathbf{w}^T\|_p \leq 1,$$

which suggests that the relationship between dual norms  $\|\dots\|$  and  $\|\dots\|^T$  may be symmetrical.

The last relation is true not just for  $\|\dots\|_p$  and  $\|\dots\|_p^T$  but for all pairs of dual norms, though the general proof must be postponed until the Hahn-Banach theorem has been presented.

### Support-Planes

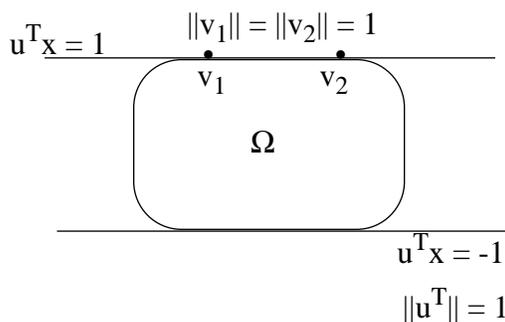
Now some geometrical properties of dual norms can be described. They will be described for real 2- and 3-dimensional spaces though analogous descriptions apply to all finite dimensions.

Let  $\mathbf{u}^T$  be an arbitrarily chosen unit-functional ( $\|\mathbf{u}^T\| = 1$ ). Let  $\mathbf{v}$  be a unit-vector ( $\|\mathbf{v}\| = 1$ ) on the boundary  $\partial\Omega$  that maximizes

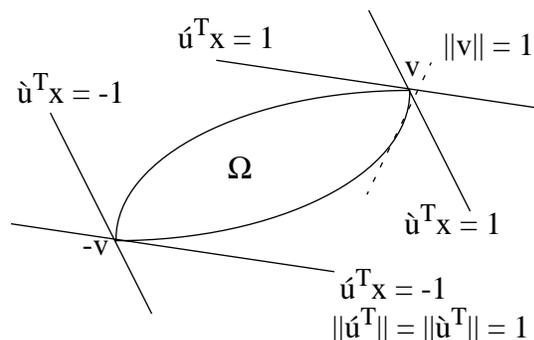
$$\mathbf{u}^T \mathbf{v} = \|\mathbf{u}^T\| \|\mathbf{v}\| = 1 = \max \mathbf{u}^T \mathbf{x} \text{ over all } \mathbf{x} \text{ with } \|\mathbf{x}\| = 1.$$

For each constant  $\lambda$  the equation  $\mathbf{u}^T \mathbf{x} = \lambda$  describes a line or plane in the space of vectors  $\mathbf{x}$ . Corresponding to different values  $\lambda$  are different members of a family of parallel lines or planes. Two of those lines or planes touch the unit-ball  $\Omega$  and sandwich it between them; their equations are  $\mathbf{u}^T \mathbf{x} = \pm 1$ . To confirm that  $\Omega$  lies between them, observe for every  $\pm \mathbf{x}$  in  $\Omega$  that  $\mathbf{u}^T(\pm \mathbf{x}) \leq \|\mathbf{u}^T\| \|\pm \mathbf{x}\| \leq 1$ , so  $-1 \leq \mathbf{u}^T \mathbf{x} \leq 1$ . And to verify that those two lines or planes touch  $\partial\Omega$  note that  $\mathbf{u}^T(\pm \mathbf{v}) = \pm 1$ ; each of them touches  $\partial\Omega$  but not the interior of  $\Omega$ .

The line or plane whose equation is  $\mathbf{u}^T \mathbf{x} = \pm 1$  is said to *support*  $\Omega$  at  $\pm \mathbf{v}$  respectively; it is *tangent* to  $\partial\Omega$  there only if  $\mathbf{v}$  is not a vertex (corner) of  $\Omega$ . Thus the association of  $\Omega$ 's support-lines or support-planes with their points of contact can be viewed as an association of unit-functionals  $\mathbf{u}^T$  with unit-vectors  $\mathbf{v}$  on  $\partial\Omega$ . This association is one-to-one only if  $\Omega$  is *rotund*, which means *smooth* (no vertices nor edges) and *strictly convex* (no facets nor edges); otherwise  $\mathbf{v}$  cannot determine  $\mathbf{u}^T$  uniquely at edges or vertices, and  $\mathbf{u}^T$  cannot determine  $\mathbf{v}$  uniquely at edges or facets of  $\partial\Omega$ , as these diagrams show.



First choose  $\mathbf{u}^T$ ; then find  $\mathbf{v}$ .



First choose  $\mathbf{v}$ ; then find  $\mathbf{u}^T$ .

One of these diagrams takes for granted something unobvious that requires proof. In the first diagram, an arbitrary unit-functional  $\mathbf{u}^T$  is chosen first, so  $\|\mathbf{u}^T\| = 1$ , and then at least one unit-vector  $\mathbf{v}$  on  $\partial\Omega$  is found to maximize  $\mathbf{u}^T\mathbf{v} = \|\mathbf{v}\| = 1$ . All other vectors  $\mathbf{x}$  must satisfy  $\mathbf{u}^T\mathbf{x} \leq \|\mathbf{x}\|$ ; in other words, the interior of  $\Omega$  lies entirely on one side of the support-line or support-plane whose equation is  $\mathbf{u}^T\mathbf{x} = 1$  and which touches  $\partial\Omega$  at the point(s)  $\mathbf{v}$  thus found.

For the second diagram an arbitrary unit-vector  $\mathbf{v}$  with  $\|\mathbf{v}\| = 1$  is chosen first on  $\partial\Omega$ , and then at least one unit-functional  $\mathbf{u}^T$  (so  $\|\mathbf{u}^T\| = 1$ ) is found to maximize

$$\mathbf{u}^T\mathbf{v} = \max \mathbf{w}^T\mathbf{v} \text{ over all } \mathbf{w}^T \text{ with } \|\mathbf{w}^T\| = 1;$$

this maximum  $\mathbf{u}^T\mathbf{v} \leq \|\mathbf{u}^T\| \cdot \|\mathbf{v}\| = 1$ . But the diagram assumes  $\mathbf{u}^T\mathbf{v} = 1$ . Why isn't  $\mathbf{u}^T\mathbf{v} < 1$ ?

The dotted line shows what would happen were the maximized  $\mathbf{u}^T\mathbf{v} < 1$ : The support-line or support-plane whose equation is  $\mathbf{u}^T\mathbf{x} = 1$  would touch  $\partial\Omega$  elsewhere than at  $\mathbf{v}$ . What seems so obvious in the diagram, namely that every  $\mathbf{v}$  on the boundary  $\partial\Omega$  is a point of contact with at least one of  $\Omega$ 's support planes, needs a proof, and it is difficult enough to deserve being named after the people who first got it right in the late 1920s.

**The Hahn-Banach Theorem:**  $\|\mathbf{v}\| = \max \mathbf{w}^T\mathbf{v}$  over all  $\mathbf{w}^T$  with  $\|\mathbf{w}^T\| = 1$ ; in other words, every point on  $\partial\Omega$  is touched by at least one support-line or support-plane of  $\Omega$ .

Proof: Since this  $\max \mathbf{w}^T\mathbf{v} \leq \max \|\mathbf{w}^T\| \cdot \|\mathbf{v}\| = \|\mathbf{v}\|$ , the proof merely requires the construction of a maximizing unit-functional  $\mathbf{u}^T$  with  $\|\mathbf{u}^T\| = 1$  and  $\mathbf{u}^T\mathbf{v} = \|\mathbf{v}\|$ . No generality is lost by assuming  $\|\mathbf{v}\| = 1$ . The construction proceeds through a sequence of subspaces of ever greater dimensions. The first subspace is 1-dimensional consisting of scalar multiples  $\mu\mathbf{v}$  of  $\mathbf{v}$ . On this subspace  $\mathbf{u}^T(\mu\mathbf{v}) = \mu$  follows from an initial assignment  $\mathbf{u}^T\mathbf{v} := \|\mathbf{v}\| = 1$  consistent with the requirement that  $\|\mathbf{u}^T\| = 1$ . Subsequent subspaces will be spanned by more leading elements of an arbitrary basis  $[\mathbf{v}, \mathbf{b}_2, \mathbf{b}_3, \dots]$  while the components  $\mathbf{u}^T\mathbf{v} = 1, \mathbf{u}^T\mathbf{b}_2, \mathbf{u}^T\mathbf{b}_3, \dots$  of  $\mathbf{u}^T$  for that basis are determined in turn until the definition of  $\mathbf{u}^T$  extends over the whole space.

Suppose  $\mathbf{u}^T$  has been defined upon a subspace  $S$  that includes  $\mathbf{v}$ , and  $|\mathbf{u}^T\mathbf{x}| \leq \|\mathbf{x}\|$  for every  $\mathbf{x}$  in  $S$  as well as  $\mathbf{u}^T\mathbf{v} = \|\mathbf{v}\| = 1$ . If  $S$  is not yet the whole vector space there must be some nonzero vector  $\mathbf{b}$  not in  $S$ . Our first task is to choose  $\mathbf{u}^T\mathbf{b} := \beta$ , without changing  $\mathbf{u}^T S$ , in such a way that  $|\mathbf{u}^T(\mathbf{s}+\mathbf{b})| \leq \|\mathbf{s}+\mathbf{b}\|$  for every  $\mathbf{s}$  in  $S$ . We already know  $\mathbf{u}^T(\mathbf{s}-\mathbf{t}) \leq \|\mathbf{s}-\mathbf{t}\|$  for all  $\mathbf{s}$  and  $\mathbf{t}$  in  $S$ , and this implies  $\mathbf{u}^T(\mathbf{s}-\mathbf{t}) \leq \|(\mathbf{s}+\mathbf{b})-(\mathbf{t}+\mathbf{b})\| \leq \|\mathbf{s}+\mathbf{b}\| + \|\mathbf{t}+\mathbf{b}\|$ , which implies in turn that  $-\|\mathbf{t}+\mathbf{b}\| - \mathbf{u}^T\mathbf{t} \leq \|\mathbf{s}+\mathbf{b}\| - \mathbf{u}^T\mathbf{s}$ . Therefore the least upper bound of the last inequality's left-hand side cannot exceed the greatest lower bound of its right-hand side as  $\mathbf{s}$  and  $\mathbf{t}$  run independently through  $S$ . Any number  $\beta$  between those bounds must satisfy

$$-\|\mathbf{t}+\mathbf{b}\| - \mathbf{u}^T\mathbf{t} \leq \beta \leq \|\mathbf{s}+\mathbf{b}\| - \mathbf{u}^T\mathbf{s} \text{ for every } \mathbf{s} \text{ and } \mathbf{t} \text{ in } S.$$

This choice for  $\mathbf{u}^T\mathbf{b} := \beta$  ensures that

$-\|\mathbf{t}+\mathbf{b}\| \leq \mathbf{u}^T\mathbf{t} + \beta = \mathbf{u}^T(\mathbf{t}+\mathbf{b})$  and  $\mathbf{u}^T(\mathbf{s}+\mathbf{b}) = \mathbf{u}^T\mathbf{s} + \beta \leq \|\mathbf{s}+\mathbf{b}\|$  for every  $\mathbf{s}$  and  $\mathbf{t}$  in  $S$ , which boils down to  $|\mathbf{u}^T(\mathbf{s}+\mathbf{b})| \leq \|\mathbf{s}+\mathbf{b}\|$  for every  $\mathbf{s}$  in  $S$ , as desired. For every  $\mathbf{x} = \mathbf{s} + \mu\mathbf{b}$  in the bigger subspace  $S + \{\mu\mathbf{b}\}$  we find  $|\mathbf{u}^T\mathbf{x}| = |\mu| \cdot |\mathbf{u}^T(\mathbf{s}/\mu + \mathbf{b})| \leq |\mu| \cdot \|\mathbf{s}/\mu + \mathbf{b}\| = \|\mathbf{x}\|$  again.

Thus can the components  $\mathbf{u}^T \mathbf{v}$ ,  $\mathbf{u}^T \mathbf{b}_2$ ,  $\mathbf{u}^T \mathbf{b}_3$ , ... of  $\mathbf{u}^T$  be chosen successively until all of them have been so defined that  $|\mathbf{u}^T \mathbf{x}| \leq \|\mathbf{x}\|$  for every  $\mathbf{x}$  in the whole vector space and  $\mathbf{u}^T \mathbf{v} = \|\mathbf{v}\|$ , which is what the theorem asserts. End of proof.

The assertion just proved is a special case, attributed to Minkowski, that conveys the essence of the Hahn-Banach theorem, which is usually stated in a more general way valid for infinite-dimensional spaces and for other convex bodies besides unit-balls of norms. The theorem was first proved only for real vector spaces; Bohnenblust and Sobczyk proved its validity for complex spaces too in 1938. The following simplification of their approach began to appear in texts like W. Rudin's *Real and Complex Analysis* 2d. ed. (1974, McGraw-Hill) in the 1970s.

The norm on the dual  $Z^*$  of a complex normed vector space  $Z$  was defined above to be

$$\|\mathbf{w}^*\| := \max |\mathbf{w}^* \mathbf{z}| = \max \operatorname{Re}(\mathbf{w}^* \mathbf{z}) \text{ over all } \mathbf{z} \text{ in } Z \text{ with } \|\mathbf{z}\| = 1.$$

(Did you verify this?) Now, given any nonzero complex vector  $\mathbf{t}$  in  $Z$ , we shall prove that

$$\|\mathbf{t}\| = \max |\mathbf{w}^* \mathbf{t}| = \max \operatorname{Re}(\mathbf{w}^* \mathbf{t}) \text{ over all } \mathbf{w}^* \text{ in } Z^* \text{ with } \|\mathbf{w}^*\| = 1.$$

Proof: The complex version shall be deduced from the real by associating the complex spaces  $Z$  and  $Z^*$  with real counterparts  $\mathbb{Z}$  and  $\mathbb{Z}^T$  that have respectively the same vectors and functionals with the same norms. Begin by choosing any basis for  $Z$  that has  $\mathbf{t}$  among the basis vectors. The set of all real linear combinations of these basis vectors (multiplying them by only real scalar coefficients) constitutes a real vector space  $X$  with  $\mathbf{t}$  among its vectors. Each  $\mathbf{z}$  in  $Z$  is a linear combination of basis vectors with complex coefficients whose real and imaginary parts, taken separately, decompose  $\mathbf{z}$  into  $\mathbf{z} = \mathbf{x} + i\mathbf{y}$  where  $i = \sqrt{-1}$  and  $\mathbf{x}$  and  $\mathbf{y}$  come from  $X$  and are determined uniquely by  $\mathbf{z}$ . This decomposition associates each  $\mathbf{z}$  in  $Z$  with  $\mathbb{z} := [\mathbf{x}; \mathbf{y}]$  in a real space  $\mathbb{Z}$  of pairs of vectors from  $X$ ; real  $\mathbb{Z}$  has twice the dimension of complex  $Z$  and inherits its norm thus:  $\|\mathbb{z}\| := \|\mathbf{z}\|$ . And  $\mathbb{Z}$  inherits  $\mathbb{t} := [\mathbf{t}; \mathbf{o}]$ . (Although  $\mathbb{Z}$  and  $Z$  seem to have the same vectors under different names, the spaces are different because multiplying a vector in  $Z$  by a complex scalar multiplies the vector's associate in  $\mathbb{Z}$  by a linear operator not a real scalar:  $(\beta + i\mu)(\mathbf{x} + i\mathbf{y})$  in  $Z$  associates with  $[\beta\mathbf{x} - \mu\mathbf{y}; \mu\mathbf{x} + \beta\mathbf{y}]$  in  $\mathbb{Z}$ .)

What about dual spaces? Space  $X^T$  dual to  $X$  consists of real-valued linear functionals obtained by decomposing complex linear functionals from  $Z^*$  thus: Applying each  $\mathbf{c}^*$  in  $Z^*$  to any  $\mathbf{x}$  in  $X$  defines  $\mathbf{a}^T \mathbf{x} := \operatorname{Re}(\mathbf{c}^* \mathbf{x})$  and  $\mathbf{b}^T \mathbf{x} := -\operatorname{Im}(\mathbf{c}^* \mathbf{x})$ ; that  $\mathbf{a}^T$  and  $\mathbf{b}^T$  really are linear functionals in  $X^T$  is easy to verify. Conversely every two members  $\mathbf{a}^T$  and  $\mathbf{b}^T$  of  $X^T$  determine a linear functional  $\mathbf{c}^* := \mathbf{a}^T - i\mathbf{b}^T$  in  $Z^*$  whose value at any  $\mathbf{z} = \mathbf{x} + i\mathbf{y}$  in  $Z$  is  $\mathbf{c}^* \mathbf{z} = \mathbf{a}^T \mathbf{x} + \mathbf{b}^T \mathbf{y} + i(\mathbf{a}^T \mathbf{y} - \mathbf{b}^T \mathbf{x})$ . These formulas also associate  $\mathbb{c}^T := [\mathbf{a}^T, \mathbf{b}^T]$  in  $\mathbb{Z}^T$  with each  $\mathbf{c}^*$  in  $Z^*$  thus:  $\mathbb{c}^T \mathbb{z} := \operatorname{Re}(\mathbf{c}^* \mathbf{z}) = \mathbf{a}^T \mathbf{x} + \mathbf{b}^T \mathbf{y}$ . Note that  $\operatorname{Im}(\mathbf{c}^* \mathbf{z}) = -\operatorname{Re}(i\mathbf{c}^* \mathbf{z}) = -\operatorname{Re}(\mathbf{c}^*(i\mathbf{z})) = -\mathbb{c}^T \mathbb{s}$  where  $\mathbb{s}$  in  $\mathbb{Z}$  is the associate of  $i\mathbf{z}$  in  $Z$ . Conversely each  $\mathbb{c}^T = [\mathbf{a}^T, \mathbf{b}^T]$  in  $\mathbb{Z}^T$  determines  $\mathbf{c}^*$  in  $Z^*$  from the preceding formula for  $\mathbf{c}^* \mathbf{z}$ . The same result can be obtained without decomposing  $\mathbb{c}^T$  from a definition  $\operatorname{Re}(\mathbf{c}^* \mathbf{z}) := \mathbb{c}^T \mathbb{z}$  and an identity  $\mathbf{c}^* \mathbf{z} = \operatorname{Re}(\mathbf{c}^* \mathbf{z}) - i\operatorname{Re}(\mathbf{c}^*(i\mathbf{z}))$ ; this identity requires  $\mathbb{c}^T$  to be applied twice, first to the real associate  $\mathbb{z}$  of  $\mathbf{z}$ , and second to the real associate of  $i\mathbf{z}$ . Finally  $\|\mathbb{c}^T\| = \max_{\|\mathbb{z}\|=1} \mathbb{c}^T \mathbb{z} = \max_{\|\mathbf{z}\|=1} \operatorname{Re}(\mathbf{c}^* \mathbf{z}) = \|\mathbf{c}^*\|$ .

Strictly speaking, spaces  $X$  and  $X^T$  are extraneous, introduced here only in the hope that they help make the relationship between  $Z$  and  $\mathbb{Z}$  easier to understand by making it more concrete. This relationship amounts to two one-to-one, invertible and norm-preserving maps, one map between all of complex space  $Z$  and all of real space  $\mathbb{Z}$ , the other map between their dual spaces, such that  $\mathbb{c}^T \mathbb{z} = \operatorname{Re}(\mathbf{c}^* \mathbf{z})$ .

Back to the proof of the complex Hahn-Banach theorem. Its real version provides at least one real  $\mathbb{c}^T$  in  $\mathbb{Z}^T$  to satisfy  $\mathbb{c}^T \mathbb{t} = \|\mathbb{t}\|$  and  $|\mathbb{c}^T \mathbb{z}| \leq \|\mathbb{z}\|$  for every  $\mathbb{z}$  in  $\mathbb{Z}$ , so  $\|\mathbb{c}^T\| = 1$ . The associated  $\mathbf{c}^*$  in  $Z^*$  has  $\|\mathbf{c}^*\| = \|\mathbb{c}^T\| = 1$  and  $\operatorname{Re}(\mathbf{c}^* \mathbf{t}) = \mathbb{c}^T \mathbb{t} = \|\mathbb{t}\| = \|\mathbf{t}\|$ ; moreover  $\operatorname{Im}(\mathbf{c}^* \mathbf{t}) = 0$  because otherwise  $|\mathbf{c}^* \mathbf{t}| = \|\mathbf{t}\| + \operatorname{Im}(\mathbf{c}^* \mathbf{t})$  would exceed  $\|\mathbf{t}\|$  contradicting  $\|\mathbf{c}^*\| = 1$ . Therefore  $\mathbf{c}^* \mathbf{t} = \|\mathbf{t}\| = \max_{\|\mathbf{w}^*\|=1} \operatorname{Re}(\mathbf{w}^* \mathbf{t})$  as claimed. End of proof.

## Duality or Polarity with respect to the Norm

Analogous to the involutory ( self-inverting ) map between columns and rows effected by complex conjugate transposition is a map between any normed space  $Z$  of vectors  $z$  and its dual space  $Z^*$  of functionals  $w^*$  inspired by the symmetry we have just established in the formulas  $\|w^*\| = \max_{\|z\|=1} \operatorname{Re}(w^*z)$  and  $\|z\| = \max_{\|w^*\|=1} \operatorname{Re}(w^*z)$ . The constraints

$\|\dots\| = 1$  are inessential in these formulas, which can be rewritten in the equivalent forms

$$\|w^*\| = \max_{z \neq 0} \operatorname{Re}(w^*z)/\|z\| \quad \text{and} \quad \|z\| = \max_{w^* \neq 0} \operatorname{Re}(w^*z)/\|w^*\|$$

to show how each nonzero  $w^*$  determines at least one maximizing direction  $z$ , and each nonzero  $z$  determines at least one maximizing direction  $w^*$ . When this maximization occurs, nonzero lengths can be assigned to satisfy the

$$\text{Duality Equations:} \quad w^*z = \|w^*\| \cdot \|z\| \quad \text{and} \quad \|w^*\| = \|z\| \neq 0,$$

and then  $w^*$  and  $z$  are said to be *Dual with respect to the norm*. This kind of duality is also called *Polarity* sometimes. These duality equations determine either  $w^*$  or  $z$  as a generally nonlinear function of the other, and not always uniquely; for instance, given a nonzero  $w^*$ , choose any unit-vector  $u$  that maximizes  $\operatorname{Re}(w^*u)$  to determine  $z := \|w^*\|u$ .

**Examples:** First for  $p = 2$ , then for  $p = 1$ , and then for  $p = \infty$ , we shall see how, given either of  $w^*$  and  $z$ , to determine the other so that they will be dual with respect to the norm  $\|\dots\|_p$ . In all cases the column vector  $z$  has components  $\zeta_1, \zeta_2, \zeta_3, \dots$ , and the row  $w^*$  has components  $\bar{\omega}_1, \bar{\omega}_2, \bar{\omega}_3, \dots$  where  $\bar{\omega}_j$  is the complex conjugate of  $\omega_j$ .

For  $p = 2$ ,  $\|w^*\|_2 = \sqrt{\sum_j |\omega_j|^2}$  and  $\|z\|_2 = \sqrt{\sum_j |\zeta_j|^2}$ ; duals have  $\omega_j = \zeta_j$ .

For  $p = 1$ ,  $\|w^*\|_1 = \max_j |\omega_j|$  and  $\|z\|_1 = \sum_j |\zeta_j|$ ; duals have either

$$\omega_j = \|z\|_1 \cdot \zeta_j / |\zeta_j| \quad \text{whenever } \zeta_j \neq 0, \quad \text{and otherwise any } |\omega_j| \leq \|z\|_1 \text{ will do,} \quad \text{or} \\ \zeta_j = 0 \quad \text{unless } |\omega_j| = \|w^*\|_1, \quad \text{and then any } \zeta_j / \omega_j \geq 0 \text{ with } \sum_j \zeta_j / \omega_j = 1 \text{ will do.}$$

For  $p = \infty$  swap  $w$  and  $z$  in the case  $p = 1$ .

The cases  $p = 1$  and  $p = \infty$ , like the two diagrams earlier, illustrate how neither  $z$  nor  $w^*$  need determine its dual uniquely if the unit-ball has vertices or edges or flat facets.

**Exercise 2:** Verify that the Duality Equations are satisfied by the alleged dual pairs  $w^*$  and  $z$  defined above.

**Exercise 3:** Tabulate, for  $p$  and  $q$  taking values 1, 2 and  $\infty$  independently,  $\mu_{pq} := \max \|z\|_p / \|z\|_q$  as  $z$  runs over all nonzero complex  $n$ -dimensional column vectors.

**Exercise 4:** Two given norms  $\|\dots\|$  and  $\|\dots\|$  on a finite-dimensional vector space induce norms  $\|\dots\|^*$  and  $\|\dots\|^*$  respectively on the dual space; explain why  $\max_{z \neq 0} \|z\| / \|z\| = \max_{w^* \neq 0} \|w^*\| / \|w^*\|$ .

**Exercise 5:** Given one norm  $\|\dots\|$  and an invertible linear operator  $R$ , define a new norm  $\|z\| := \|Rz\|$  for all vectors  $z$  in some space. How is  $\|\dots\|^*$  related to  $\|\dots\|^*$  on the dual space? Given also nonzero  $z$  and  $w^*$  dual with respect to  $\|\dots\|$ , find a nonzero pair dual with respect to  $\|\dots\|$ .

**Exercise 6:** Explain why the set of vectors  $z$  dual to a given nonzero functional  $w^*$  must constitute a convex set.

**Exercise 7:** Show that  $\|z + \mu v\| \geq \|z\|$  for all scalars  $\mu$  if and only if  $w^*v = 0$  for a  $w^*$  dual to  $z$ . Then  $v$  is called *orthogonal* to  $z$ . In that case, must  $z$  be orthogonal to  $v$ ? Justify your answer. *This exercise is hard!*

## Duality or Polarity, and the Derivative of a Norm

The non-uniqueness in the determination of either of a pair of dual vectors by the other, and of either of a support line/plane and its point of contact by the other, is a complication that afflicts notation, terminology and proofs to the detriment of pedagogy. Let's leave these complications to courses more advanced than this one; see texts about convexity and convex bodies.

For the sake of simplicity in our discussion of differentiability, we shall assume a real vector space of finite dimension with a rotund unit-ball  $\Omega$ . If it is not rotund,  $\Omega$  can be made that way by slightly rounding its vertices and edges and slightly puffing out its facets, just as perfect cubes must be rounded off a little to make a usable pair of dice. Then the lines/planes that support  $\Omega$  can be regarded all as tangents, each one touching  $\partial\Omega$  at just one point. A tangent's orientation and its point of contact determine each other uniquely and continuously each as a function of the other; since an analytical proof of this claim is too tedious an argument about compact sets to reproduce here its confirmation is left to courses and texts about real analysis and convex bodies, or to the reader's geometrical intuition about smooth rotund bodies.

The continuous one-to-one association between tangents to rotund  $\partial\Omega$  and their points of tangency, between dual pairs of unit-functionals  $\mathbf{u}^T$  and unit-vectors  $\mathbf{v}$ , extends to all pairs of dual functionals  $\mathbf{w}^T$  and vectors  $\mathbf{x}$  satisfying the

$$\text{Duality Equations } \mathbf{w}^T \mathbf{x} = \|\mathbf{w}^T\| \cdot \|\mathbf{x}\| \quad \text{and} \quad \|\mathbf{w}^T\| = \|\mathbf{x}\|$$

because these now determine either  $\mathbf{w}^T$  or  $\mathbf{x}$  uniquely from the other. Can you see how? The importance of this duality to applications of norms can be gauged from the fact that  $\|\mathbf{x}\|$  is now differentiable if nonzero, and its derivative involves the functional  $\mathbf{w}^T$  dual to  $\mathbf{x}$  thus:

$$d\|\mathbf{x}\| = \mathbf{w}^T d\mathbf{x} / \|\mathbf{x}\| .$$

To see why this formula is valid, fix nonzero vectors  $\mathbf{x}$  and  $\mathbf{h}$  arbitrarily and, for any real scalar  $\lambda$ , let  $\mathbf{w}^T_\lambda$  be dual to  $\mathbf{x} + \lambda\mathbf{h}$  so  $\mathbf{w}^T_\lambda (\mathbf{x} + \lambda\mathbf{h}) = \|\mathbf{w}^T_\lambda\| \cdot \|\mathbf{x} + \lambda\mathbf{h}\|$  and  $\|\mathbf{w}^T_\lambda\| = \|\mathbf{x} + \lambda\mathbf{h}\|$ .

The rotundity of  $\Omega$  implies  $\mathbf{w}^T_\lambda \rightarrow \mathbf{w}^T$  as  $\lambda \rightarrow 0$ . For all  $\lambda \neq 0$ , Hölder's inequality says

$$\|\mathbf{x} + \lambda\mathbf{h}\| - \|\mathbf{x} - \lambda\mathbf{h}\| \leq \mathbf{w}^T_\lambda (\mathbf{x} + \lambda\mathbf{h}) / \|\mathbf{w}^T_\lambda\| - \mathbf{w}^T_\lambda (\mathbf{x} - \lambda\mathbf{h}) / \|\mathbf{w}^T_\lambda\| = 2\lambda \mathbf{w}^T_\lambda \mathbf{h} / \|\mathbf{w}^T_\lambda\|$$

and similarly with  $-\lambda$  in place of  $\lambda$ . Hence if  $\lambda > 0$

$$\mathbf{w}^T_{-\lambda} \mathbf{h} / \|\mathbf{w}^T_{-\lambda}\| \leq (\|\mathbf{x} + \lambda\mathbf{h}\| - \|\mathbf{x} - \lambda\mathbf{h}\|) / (2\lambda) \leq \mathbf{w}^T_\lambda \mathbf{h} / \|\mathbf{w}^T_\lambda\| .$$

Letting  $\lambda \rightarrow 0+$  implies  $\mathbf{w}^T \mathbf{h} / \|\mathbf{x}\| = \mathbf{w}^T \mathbf{h} / \|\mathbf{w}^T\| = d\|\mathbf{x} + \lambda\mathbf{h}\| / d\lambda$  at  $\lambda = 0$ . Since  $\mathbf{h}$  is arbitrary, this confirms that  $d\|\mathbf{x}\| = \mathbf{w}^T d\mathbf{x} / \|\mathbf{x}\|$ .

Derivatives and *Gradients* are often mixed up. Let  $f(\mathbf{x})$  be any differentiable real scalar function defined over a space of real vectors  $\mathbf{x}$ . The derivative  $f'(\mathbf{x})$  belongs to the dual space because scalar  $df(\mathbf{x}) = f'(\mathbf{x})d\mathbf{x}$ . The gradient  $\text{Grad } f(\mathbf{x})$  is a vector in the same space as  $\mathbf{x}$ , *not* in its dual space, defined to have the direction in which  $f(\mathbf{x})$  increases most rapidly, and a norm equal to that rate of change. More precisely,  $\text{Grad } f(\mathbf{x})$  is parallel to the unit-vector  $\mathbf{u}$  that maximizes  $df(\mathbf{x} + \lambda\mathbf{u}) / d\lambda$  at  $\lambda = 0$ , and this maximum equals  $\|\text{Grad } f(\mathbf{x})\|$ .

**Exercise 8:** Show why  $\text{Grad } f(\mathbf{x})$  is the vector dual to  $f'(\mathbf{x})$  with respect to the norm.

Therefore  $\text{Grad } \|\mathbf{x}\| = \mathbf{x} / \|\mathbf{x}\|$ . And when  $\mathbf{x}$  runs over a Euclidean space,  $\text{Grad } f(\mathbf{x}) = f'(\mathbf{x})^T$ .

**Auerbach's Theorem for a Real Rotund Unit-ball  $\Omega$  :** It can be circumscribed by at least one parallelepiped whose faces support  $\Omega$  at their midpoints.

Proof: Of all the parallelepipeds circumscribing (barely containing)  $\Omega$ , the one(s) with minimum content (area or volume) has/have the desired supporting properties, as we shall see after finding one. (Others have the desired properties too but one is all we need.)

Choose an arbitrary basis in which vectors  $\mathbf{c}$  can be represented by column vectors  $\mathbf{c}$  and functionals  $\mathbf{r}^T$  by row vectors  $\mathbf{r}^T$ ; of course  $\|\mathbf{c}\| = \|\mathbf{c}\|$  and  $\|\mathbf{r}^T\| = \|\mathbf{r}^T\|$ . For any square matrix  $\mathbf{C} := [\mathbf{c}_1, \mathbf{c}_2, \dots, \mathbf{c}_n]$  whose columns all have length  $\|\mathbf{c}_j\| = 1$  there is a square matrix  $\mathbf{R}$  whose rows  $\mathbf{r}_j^T$  are dual to the corresponding columns  $\mathbf{c}_j$  of  $\mathbf{C}$ ; this means every  $\mathbf{r}_j^T \mathbf{c}_j = 1$  and  $\|\mathbf{r}_j^T\| = \|\mathbf{c}_j\| = 1$ . Also, unless  $\det(\mathbf{C}) = 0$ , there is an inverse matrix  $\mathbf{Q} = \mathbf{C}^{-1}$  whose rows  $\mathbf{q}_i^T$  satisfy  $\mathbf{q}_i^T \mathbf{c}_j = (1 \text{ if } i=j, \text{ else } 0)$ . Now choose  $\mathbf{C}$  to maximize  $\det(\mathbf{C})$ . The maximum is achieved because it is the maximum of a continuous function  $\det(\dots)$  over a closed bounded set of columns all of length 1. Varying columns  $\mathbf{c}_j$  infinitesimally by  $d\mathbf{c}_j$  causes  $\log \det(\mathbf{C})$  to vary, according to Jacobi's formula for a determinant's derivative, by  $d \log \det(\mathbf{C}) = \text{Trace}(\mathbf{C}^{-1} d\mathbf{C}) = \sum_j \mathbf{q}_j^T d\mathbf{c}_j$ . Keeping  $\|\mathbf{c}_j + d\mathbf{c}_j\| = 1$  constrains  $d\mathbf{c}_j$  to satisfy  $d\|\mathbf{c}_j\| = \mathbf{r}_j^T d\mathbf{c}_j / \|\mathbf{c}_j\| = 0$ . Whenever every  $d\mathbf{c}_j$  satisfies this constraint,  $\sum_j \mathbf{q}_j^T d\mathbf{c}_j = 0$  because of the maximality of  $\log \det(\mathbf{C})$ . Therefore every pair  $\{\mathbf{r}_j^T, \mathbf{q}_j^T\}$  must be a pair of parallel rows; and since  $\mathbf{r}_j^T \mathbf{c}_j = \mathbf{q}_j^T \mathbf{c}_j = 1$  we conclude that  $\mathbf{R} = \mathbf{Q} = \mathbf{C}^{-1}$  row by row.

Now we revert from rows and columns to functionals and vectors, and interpret geometrically the configuration that has been found. A basis  $\mathbf{C} = [\mathbf{c}_1, \mathbf{c}_2, \dots, \mathbf{c}_n]$  of unit vectors  $\mathbf{c}_j$  and a dual basis  $\mathbf{R}$  of unit functionals  $\mathbf{r}_i^T$  have been found to satisfy  $\mathbf{r}_i^T \mathbf{c}_j = (1 \text{ if } i=j, \text{ else } 0)$  and  $\|\mathbf{r}_j^T\| = \|\mathbf{c}_j\| = 1$ . The  $2n$  equations  $\mathbf{r}_i^T \mathbf{x} = \pm 1$  are the equations of  $2n$  (hyper)planes that support  $\Omega$ , touching its boundary  $\partial\Omega$  at  $2n$  points  $\pm \mathbf{c}_i$  respectively, each exactly midway between all the other planes because  $\mathbf{r}_j^T \mathbf{c}_i = 0$  for all  $j \neq i$ . Thus, each contact point  $\pm \mathbf{c}_i$  is the midpoint of a face of a parallelepiped  $\partial\Omega_\infty$  that circumscribes  $\Omega$ ; that face lies in the plane  $\mathbf{r}_i^T \mathbf{x} = \pm 1$  between  $n-1$  pairs of planes  $\mathbf{r}_j^T \mathbf{x} = \pm 1$  for  $j \neq i$ . End of proof.

Incidentally, the content (volume) inside the circumscribing parallelepiped  $\partial\Omega_\infty$  can be shown to be  $2^n / \det(\mathbf{R})$ , which is minimized when  $\det(\mathbf{R})$  is maximized. An argument very similar to the foregoing leads to the same conclusion  $\mathbf{R} = \mathbf{C}^{-1}$  where  $\mathbf{C}$  is the matrix whose columns are dual to the rows of  $\mathbf{R}$  with respect to the norm.

An *Auerbach Basis* is a basis like  $\mathbf{C} = [\mathbf{c}_1, \mathbf{c}_2, \dots, \mathbf{c}_n]$  above of unit vectors each *orthogonal* (in the sense of Exercise 7) to all the others. Auerbach's theorem, which has been proved also for non-rotund unit-balls, complex spaces, and infinite dimensions, ensures that at least one such  $\mathbf{C}$  exists. It associates every column  $n$ -vector  $\mathbf{x}$  with a vector  $\mathbf{x} = \mathbf{C}\mathbf{x}$  for which  $\|\mathbf{x}\|_\infty := \|\mathbf{x}\|_\infty$  is a new norm whose unit-ball's boundary  $\partial\Omega_\infty$  is the theorem's circumscribing parallelepiped. A third norm is  $\|\mathbf{x}\|_1 := \|\mathbf{x}\|_1$ ; its unit-ball is the convex hull of all vectors  $\pm \mathbf{c}_j$ .

**Exercise 9:** Show that  $\|\mathbf{x}\|_1/n \leq \|\mathbf{x}\|_\infty \leq \|\mathbf{x}\| \leq \|\mathbf{x}\|_1 \leq n\|\mathbf{x}\|_\infty$  for every  $\mathbf{x}$ . (See p. 3.)

## Matrix Norms

The trouble with matrix norms is their abundance. They are so numerous that they overwhelm typography's capacity to differentiate among them without a plethora of subscripts and/or nonce-symbols. However, for two reasons, only a few of those many matrix norms get much use:

- Only a few matrix norms are easy to compute, and another is not difficult; and ...
- These can approximate the rest well enough for most practical purposes.

Justification for the last reason explains why a substantial fraction of the theory of matrix norms concerns the closeness with which one can approximate another. It's a vast theory overflowing with terms and concepts. Only a few pages' worth can be touched in these notes.

Auerbach's Theorem and Ex. 9 have shown why, with the choice of an apt Auerbach Basis, any vector norm  $\|\mathbf{x}\|$  for abstract  $n$ -dimensional vectors  $\mathbf{x}$  can be approximated within a factor  $n^{\pm 1/2}$  by  $\|\mathbf{x}\|_{\infty}$  or  $\|\mathbf{x}\|_1$  applied to the column  $\mathbf{x}$  that represents  $\mathbf{x}$  in the apt coordinate system. Finding an apt basis is the costly part of this approximation. Fortunately, simply rescaling whatever basis is already in use (this is tantamount to a change to more convenient units for variables) often gets adequate approximations from simple norms like  $\|\mathbf{x}\|_{\infty}$  and  $\|\mathbf{x}\|_1$ .

Since  $m$ -by- $n$  matrices  $\mathbf{L}$ , representing linear operators  $\mathbf{L}$  in chosen bases for their domain and target-space, also constitute a vector space of dimension  $m \cdot n$ , either the largest magnitude or the sum of magnitudes of  $m \cdot n$  linear combinations of the elements of  $\mathbf{L}$  can approximate any norm  $\|\mathbf{L}\|$  within a factor  $(m \cdot n)^{\pm 1/2}$ , according to Auerbach's Theorem. Actually, far less work suffices to approximate any  $\|\mathbf{L}\|$  far better, the more so as dimensions  $m$  and  $n$  get bigger. This will be just one payoff from a relatively brief foray into the theory of matrix norms.

In this theory the norm function  $\|\dots\|$  is heavily *Overloaded*; this means that the function's definition and computation depend upon the *Linguistic Type* of its argument. We have seen this already;  $\|\mathbf{r}^T\|$  may be computed from the elements of the row  $\mathbf{r}^T$  representing functional  $\mathbf{r}^T$  very differently than  $\|\mathbf{x}\|$  is computed from the elements of the column  $\mathbf{x}$  representing vector  $\mathbf{x}$ . Now another  $\|\mathbf{L}\|$  will be computed from the matrix  $\mathbf{L}$  representing linear operator  $\mathbf{L}$ . And if vector  $\mathbf{y} = \mathbf{L} \cdot \mathbf{x}$  in the range of  $\mathbf{L}$  is represented by a column  $\mathbf{y}$  its elements may figure in the computation of  $\|\mathbf{y}\|$  differently than column  $\mathbf{x}$  figures in the computation of  $\|\mathbf{x}\|$  for  $\mathbf{x}$  in the domain of  $\mathbf{L}$ ; even if its range and domain have the same dimension the two spaces can have different norms. An assertion so simple as " $\|\mathbf{L} \cdot \mathbf{x}\| \leq \|\mathbf{L}\| \cdot \|\mathbf{x}\|$ " is likely to involve three different norms, and their computation from representing arrays  $\mathbf{L} \cdot \mathbf{x}$ ,  $\mathbf{L}$  and  $\mathbf{x}$  will generally change when bases for the domain and target space of  $\mathbf{L}$  change. What a mess!

READERS BEWARE! Always keep the context of  $\|\dots\|$ 's argument in mind to identify its linguistic type and hence determine which norm is at issue.

The mental burden imposed because the meaning of " $\|\dots\|$ " depends upon the linguistic type of its argument has turned out to be less annoying than the notational clutter that used to be created when each linear space's norm had to have its own distinctive name. Some unavoidable clutter will afflict these notes whenever different norms for the same linear space are compared.

The oldest matrix norm is the *Frobenius Norm*  $\|\mathbf{B}\|_F := \sqrt{(\text{trace}(\mathbf{B}^T \cdot \mathbf{B}))} = \sqrt{(\sum_i \sum_j b_{ij}^2)}$ . (For complex matrices  $\mathbf{B}$  replace  $b_{ij}^2$  by  $|b_{ij}|^2$ .) Since  $\|\mathbf{B}\|_F$  is the Euclidean norm applied to a

column containing all the elements of  $B$ , it satisfies the three laws that every norm must satisfy:

$$\begin{array}{ll} \text{Positivity:} & \infty > \|B\| > 0 \text{ except } \|O\| = 0. \\ \text{Homogeneity:} & \|\mu \cdot B\| = |\mu| \cdot \|B\| \text{ for every scalar } \mu. \\ \text{Triangle Inequality:} & \|B+C\| \leq \|B\| + \|C\|. \end{array}$$

Another property, desirable for every matrix norm, is possessed by the Frobenius norm:

$$\text{Multiplicative Dominance: } \|B \cdot C\|_F \leq \|B\|_F \cdot \|C\|_F \text{ for all multipliable } B \text{ and } C.$$

Proof: Let  $b_i^T$  denote row #i of  $B$ , and let  $c_j$  denote column #j of  $C$ . Then

$$\begin{aligned} \|B \cdot C\|_F^2 &= \sum_i \sum_j (b_i^T \cdot c_j)^2 \leq \sum_i \sum_j (b_i^T \cdot b_i) \cdot (c_j^T \cdot c_j) \text{ by Cauchy's inequality} \\ &= (\sum_i b_i^T \cdot b_i) \cdot (\sum_j c_j^T \cdot c_j) = \|B\|_F^2 \cdot \|C\|_F^2, \text{ as claimed.} \end{aligned}$$

**Exercise 10:** For a rank 1 matrix  $b \cdot c^T$  show that  $\|b \cdot c^T\|_F = \|b\|_2 \cdot \|c^T\|_2 = \|b\|_F \cdot \|c^T\|_F$ .

By analogy with the Frobenius norm's descent from the Euclidean vector norm, two other matrix norms are the biggest-magnitude norm  $\|B\|_M := \max_i \max_j |b_{ij}|$  and the total-magnitudes norm  $\|B\|_\Sigma := \sum_i \sum_j |b_{ij}|$ . However, for dimensions 2 or greater, ...

**Exercise 11:** Show that  $\|\dots\|_\Sigma$  possesses Multiplicative Dominance but  $\|\dots\|_M$  does not. Evaluate  $\max_{B,C \neq O} \|B \cdot C\|_\Sigma / (\|B\|_\Sigma \cdot \|C\|_\Sigma)$  and  $\max_{B,C \neq O} \|B \cdot C\|_M / (\|B\|_M \cdot \|C\|_M)$ , and describe the maximizing matrices  $B$  and  $C$ . Then show that  $\|B\|_\mu := \|B\|_M \cdot (\text{number of } B \text{'s columns})$  is a matrix norm that does possess Multiplicative Dominance. The number of rows works too.

The four norms  $\|\dots\|_F$ ,  $\|\dots\|_\Sigma$ ,  $\|\dots\|_M$  and  $\|\dots\|_\mu$  act upon matrices of arbitrary dimensions (finite in the case of  $\|\dots\|_\mu$ ). This will be true of practically all the matrix norms discussed in these notes, though a matrix norm can be defined more narrowly to act only upon matrices of a specified shape and/or dimensions. In particular, a norm  $\|\dots\|$  that acts only upon square matrices of a specific dimension can always be scaled, as was done in Ex. 11, to possess Multiplicative Dominance thusly: Find  $\bar{\mu} := \max_{B,C \neq O} \|B \cdot C\| / (\|B\| \cdot \|C\|)$  for square matrices  $B$  and  $C$  of that specific dimension, and then replace  $\|\dots\|$  by  $\|\dots\|_{\bar{\mu}} := \|\dots\| \cdot \bar{\mu}$ .

Consequently there is little incentive to study matrix norms lacking multiplicative dominance, and we shall avoid most of them.

A property resembling multiplicative dominance is possessed by any matrix norm  $\|\dots\|$  that is *Subordinate* to two vector norms in the sense that  $\|L \cdot x\| \leq \|L\| \cdot \|x\|$  for all matrices  $L$  and column vectors  $x$  of dimensions acceptable to their respective norms.  $\|\dots\|_F$  is subordinate to  $\|\dots\|_2$ . Many writers prefer the phrase "*Compatible with*" instead of "*Subordinate to*". What happens next depends upon which came first, the vector norms or the matrix norm.

## Operator Norms

Given two normed vector spaces with one norm  $\|\mathbf{x}\|$  for vectors  $\mathbf{x}$  in one space and another possibly different norm  $\|\mathbf{y}\|$  for vectors  $\mathbf{y}$  in the other space, a norm  $\|\mathbf{L}\| := \max_{\mathbf{x} \neq \mathbf{0}} \|\mathbf{L} \cdot \mathbf{x}\| / \|\mathbf{x}\|$  is induced upon linear maps  $\mathbf{L}$  from the first space to the second, and another perhaps fourth norm  $\|\mathbf{K}\| := \max_{\mathbf{y} \neq \mathbf{0}} \|\mathbf{K} \cdot \mathbf{y}\| / \|\mathbf{y}\|$  is induced upon linear maps  $\mathbf{K}$  from the second to the first. These induced norms are called “Operator Norms” and “Lub Norms” (after “Least Upper Bound”, which replaces “max” when dimensions are infinite) and “Sup Norms” (likewise).

**Exercise 12:** Confirm that each operator norm obeys all norms’ three laws, namely Positivity, Homogeneity and the Triangle Inequality. Confirm that, as a linear map from a space to itself, the identity  $\mathbf{I}$  has operator norm  $\|\mathbf{I}\| = 1$ , and that therefore  $\|\dots\|_F$ ,  $\|\dots\|_\Sigma$  and  $\|\dots\|_\mu$  from Ex. 11 cannot be operator norms for dimensions 2-by-2 or greater.

What about Multiplicative Dominance? Operator norms have a similar but subtly different multiplicative property whose description requires a third space of vectors  $\mathbf{z}$  and its norm  $\|\mathbf{z}\|$ . Three operator norms are induced for maps from the first to the second, from the second to the third, and from the first to the third; and three more are induced in the opposite directions. All six are given the same heavily overloaded name  $\|\dots\|$ . If  $\mathbf{z} = \mathbf{B} \cdot \mathbf{y}$  and  $\mathbf{y} = \mathbf{C} \cdot \mathbf{x}$  then we find

$$\text{Operator Norms are } \textit{Multiplicative}: \|\mathbf{B} \cdot \mathbf{C}\| \leq \|\mathbf{B}\| \cdot \|\mathbf{C}\|.$$

Proof: If  $\mathbf{B} \cdot \mathbf{C} \neq \mathbf{0}$  then

$$\begin{aligned} \|\mathbf{B} \cdot \mathbf{C}\| &= \max_{\mathbf{x} \neq \mathbf{0}} \|\mathbf{B} \cdot \mathbf{C} \cdot \mathbf{x}\| / \|\mathbf{x}\| = \max_{\mathbf{x} \neq \mathbf{0} \& \mathbf{y} = \mathbf{C} \cdot \mathbf{x}} (\|\mathbf{B} \cdot \mathbf{y}\| / \|\mathbf{y}\|) \cdot \|\mathbf{C} \cdot \mathbf{x}\| / \|\mathbf{x}\| \\ &\leq \max_{\mathbf{x} \neq \mathbf{0} \& \mathbf{y} \neq \mathbf{0}} (\|\mathbf{B} \cdot \mathbf{y}\| / \|\mathbf{y}\|) \cdot \|\mathbf{C} \cdot \mathbf{x}\| / \|\mathbf{x}\| = \|\mathbf{B}\| \cdot \|\mathbf{C}\| \text{ as claimed.} \end{aligned}$$

Note that this proof involves five generally different norm functions, namely  $\|\mathbf{B} \cdot \mathbf{C}\|$ ,  $\|\mathbf{x}\|$ ,  $\|\mathbf{B} \cdot \mathbf{y}\| = \|\mathbf{z}\|$ ,  $\|\mathbf{C} \cdot \mathbf{x}\| = \|\mathbf{y}\|$ ,  $\|\mathbf{B}\|$  and  $\|\mathbf{C}\|$ . Somehow the grammar of abstract linear algebra has ensured that the right ones were invoked. Still, when you use the inequality  $\|\mathbf{B} \cdot \mathbf{C}\| \leq \|\mathbf{B}\| \cdot \|\mathbf{C}\|$  you should check that the norm applied to the target space of  $\mathbf{C}$  is the same as is applied to the domain of  $\mathbf{B}$  just as you would check that their spaces’ dimensions match before multiplying.

**Exercise 13:** Show that  $\|\mathbf{B}\|_M := \max_i \max_j |b_{ij}|$  is the operator norm for matrices  $\mathbf{B}$  construed as linear maps  $\mathbf{y} = \mathbf{B} \cdot \mathbf{x}$  from a space of columns  $\mathbf{x}$  normed by  $\|\mathbf{x}\|_1$  to a space of columns  $\mathbf{y}$  normed by  $\|\mathbf{y}\|_\infty$ , so this operator norm is as multiplicative as are all operator norms, though it lacks multiplicative dominance if the matched dimension is 2 or bigger.

**Exercise 14:** Explain why, if a linear operator  $\mathbf{B}$  maps a normed space to itself and has a compatible  $\|\mathbf{B}\| < 1$ , then  $\mathbf{I} - \mathbf{B}$  is invertible. However, if a square matrix  $\mathbf{B}$  maps one normed space to another with a different norm, then  $\mathbf{I} - \mathbf{B}$  can be non-invertible despite that, for a compatible matrix norm,  $\|\mathbf{B}\| < 1 = \|\mathbf{I}\|$ ; give an example. Hint: 2-by-2, and Ex. 13.

**Exercise 15:** Confirm these equivalent expressions for an operator norm:

$\|\mathbf{B}\| = \max_{\mathbf{x} \neq \mathbf{0}} \|\mathbf{B} \cdot \mathbf{x}\| / \|\mathbf{x}\| = \max_{\|\mathbf{x}\|=1} \|\mathbf{B} \cdot \mathbf{x}\| = \max_{\|\mathbf{x}\|=1, \|\mathbf{w}^T\|=1} \mathbf{w}^T \cdot \mathbf{B} \cdot \mathbf{x} = \max_{\mathbf{w}^T \neq \mathbf{0}^T} \|\mathbf{w}^T \cdot \mathbf{B}\| / \|\mathbf{w}^T\|$ . Five different norms appear here; identify each one’s domain. Confirm also that every operator norm satisfies  $\|\mathbf{x} \cdot \mathbf{w}^T\| = \|\mathbf{x}\| \cdot \|\mathbf{w}^T\|$  and identify the three norms’ domains.

### When a Matrix Norm is Given First

Operator norms are induced by vector norms in the operators' domain and target spaces. This process is partly reversible. Suppose  $\|B\|$  is a given matrix norm specified for square matrices  $B$  and possessing multiplicative dominance:  $\|B \cdot C\| \leq \|B\| \cdot \|C\|$  for all square matrices  $B$  and  $C$  of the same dimension. Choose a nonzero row  $r^T$  of the same dimension and let  $\|x\| := [x \cdot r^T]$  for column vectors  $x$  of the same dimension. This vector norm induces a new operator norm:

$$\|B\| := \max_{x \neq 0} \|B \cdot x\| / \|x\| = \max_{x \neq 0} [B \cdot x \cdot r^T] / [x \cdot r^T] \leq \|B\|.$$

Thus does a given multiplicatively dominant matrix norm  $\| \dots \|$  induce a vector norm  $\| \dots \|$  that induces an operator norm  $\| \dots \|$  no greater than the given matrix norm  $\| \dots \|$  no matter which arbitrarily fixed row  $r^T$  was chosen. Consequently there is a sense in which operator norms are minimal among matrix norms possessing multiplicative dominance.

**Exercise 16:** Confirm that if  $\| \dots \|$  above is already an operator norm then its induced operator norm  $\| \dots \| = \| \dots \|$ . Why is this equation violated when the given operator norm  $\| \dots \| = \| \dots \|_M$ ?

**Exercise 17:** Describe the operator norm  $\| \dots \|$  induced by  $\| \dots \| = \| \dots \|_\Sigma$ . Then describe the operator norm  $\| \dots \|$  induced by  $\| \dots \| = \| \dots \|_F$ . In each case the induced  $\| \dots \|$  costs a lot more than its given  $\| \dots \|$  to compute when the dimension is big.

### Formulas for Familiar Operator Norms

Each subscript  $p$  and  $q$  below will be chosen from the set  $\{1, 2, \infty\}$  to specify familiar vector norms like  $\|x\|_p$  and  $\|w^T\|_q$  defined in these notes' early sections titled "Norms" and "The Dual Norm". These norms' induced operator norms  $\|B\|_{pq} := \max_{x \neq 0} \|B \cdot x\|_p / \|x\|_q$  will be identified by pairs of subscripts except when the two subscripts are equal, in which case we may sometimes abbreviate " $\|B\|_{pp}$ " to " $\|B\|_p$ " at the risk of conflicts with a few other writers who sometimes write " $\|B\|_1$ " for our " $\|B\|_\Sigma$ ", " $\|B\|_\infty$ " for our " $\|B\|_M$ ", and more often " $\|B\|_2$ " for our " $\|B\|_F$ " introduced two pages ago.

**Exercise 18:** Confirm the following five formulas in which matrix  $B$  has elements  $b_{ij}$ :

- $\|B\|_\infty = \max_i \sum_j |b_{ij}| = \|B^T\|_1$ , the biggest row-sum of magnitudes.
- $\|B\|_1 = \max_j \sum_i |b_{ij}| = \|B^T\|_\infty$ , the biggest column-sum of magnitudes.
- $\|B\|_2 = (\text{the biggest singular value of } B) = \|B^T\|_2$ , the biggest eigenvalue of  $\begin{bmatrix} 0 & B^T \\ B & 0 \end{bmatrix}$ .
- $\|B\|_{\infty 1} = \max_i \max_j |b_{ij}|$  is the biggest magnitude. (Ex. 13.)
- $\|B\|_{\infty 2} = \max_i \sqrt{(\sum_j |b_{ij}|^2)}$  is the biggest row-length.

When dimensions are big the other four norms  $\|B\|_{pq}$  get little use because they cost much too much to compute. Do you see why  $\|B\|_2 = \sqrt{(\text{the biggest eigenvalue of } B^T \cdot B)}$ ? It costs a little too much too but has too many uses, both in theory and practice, to be disregarded.

## Maxima of Ratios of Matrix Norms

When dimensions are not too big, and sometimes even if they are, different norms may approximate each other well enough that only the one easier to compute gets used. Suppose two norms  $\|\mathbf{x}\|_a$  and  $\|\mathbf{x}\|_b$  are given for the same space of vectors  $\mathbf{x}$ . Let  $\mu_{ab} := \max_{\mathbf{x} \neq \mathbf{0}} \|\mathbf{x}\|_a / \|\mathbf{x}\|_b$ . For  $\|\mathbf{x}\|_1$ ,  $\|\mathbf{x}\|_2$  and  $\|\mathbf{x}\|_\infty$  these ratios should already have been tabulated in Ex. 3 in the section of these notes titled “Duality or Polarity with respect to the Norm”. And Ex. 4 there established that  $\max_{\mathbf{w}^T \neq \mathbf{0}^T} \|\mathbf{w}^T\|_a / \|\mathbf{w}^T\|_b = \mu_{ba}$  for dual norms. Of course  $\mu_{ab} \cdot \mu_{ba} \geq 1$ .

There are analogous maxima for ratios of matrix norms.

**Exercise 19:** What are all six nontrivial maxima for ratios of matrix norms  $\|\dots\|_F$ ,  $\|\dots\|_\Sigma$  and  $\|\dots\|_M$ ? (Yes, the matrices’ dimensions will figure in some of those maxima.)

Bounds for ratios of vector norms turn out to provide bounds also for ratios of induced operator norms. Subscripts clutter the notation unavoidably: We write  $\|\mathbf{L}\|_{cd} := \max_{\mathbf{x} \neq \mathbf{0}} \|\mathbf{L} \cdot \mathbf{x}\|_c / \|\mathbf{x}\|_d$  and then  $\mu_{abcd} := \max_{\mathbf{L} \neq \mathbf{0}} \|\mathbf{L}\|_{ab} / \|\mathbf{L}\|_{cd}$ . This can be found with the aid of Exs. 4 and 15:

$$\mu_{abcd} := \max_{\mathbf{L} \neq \mathbf{0}} \|\mathbf{L}\|_{ab} / \|\mathbf{L}\|_{cd} = \mu_{ac} \cdot \mu_{db}.$$

Proof: Let  $\mathbf{L} \neq \mathbf{0}$  be a linear operator that maximizes  $\|\mathbf{L}\|_{ab} / \|\mathbf{L}\|_{cd}$ , and then choose  $\mathbf{w}^T \neq \mathbf{0}^T$  and  $\mathbf{x} \neq \mathbf{0}$  to maximize  $\mathbf{w}^T \cdot \mathbf{L} \cdot \mathbf{x} / (\|\mathbf{w}^T\|_a \cdot \|\mathbf{x}\|_b) = \|\mathbf{L}\|_{ab}$  as Ex. 15 permits. It also provides  $\|\mathbf{L}\|_{cd} = \max_{\mathbf{y}^T \neq \mathbf{0}^T \& \mathbf{z} \neq \mathbf{0}} \mathbf{y}^T \cdot \mathbf{L} \cdot \mathbf{z} / (\|\mathbf{y}^T\|_c \cdot \|\mathbf{z}\|_d) \geq \mathbf{w}^T \cdot \mathbf{L} \cdot \mathbf{x} / (\|\mathbf{w}^T\|_c \cdot \|\mathbf{x}\|_d)$ , which implies that

$$\begin{aligned} \mu_{abcd} = \|\mathbf{L}\|_{ab} / \|\mathbf{L}\|_{cd} &\leq (\mathbf{w}^T \cdot \mathbf{L} \cdot \mathbf{x} / (\|\mathbf{w}^T\|_a \cdot \|\mathbf{x}\|_b)) / (\mathbf{w}^T \cdot \mathbf{L} \cdot \mathbf{x} / (\|\mathbf{w}^T\|_c \cdot \|\mathbf{x}\|_d)) \\ &= (\|\mathbf{w}^T\|_c / \|\mathbf{w}^T\|_a) \cdot (\|\mathbf{x}\|_d / \|\mathbf{x}\|_b) \leq \mu_{ac} \cdot \mu_{db}. \end{aligned}$$

To promote this inequality  $\mu_{abcd} \leq \mu_{ac} \cdot \mu_{db}$  up to equality, we construct a maximizing  $\mathbf{L}$  out of a functional  $\mathbf{y}^T \neq \mathbf{0}^T$  chosen to maximize  $\|\mathbf{y}^T\|_b / \|\mathbf{y}^T\|_d = \mu_{db}$  (cf. Ex. 4) and a vector  $\mathbf{z} \neq \mathbf{0}$  chosen to maximize  $\|\mathbf{z}\|_a / \|\mathbf{z}\|_c = \mu_{ac}$ . Let  $\mathbf{L} := \mathbf{z} \cdot \mathbf{y}^T$  to find from Ex. 15 again that

$$\mu_{abcd} \geq \|\mathbf{L}\|_{ab} / \|\mathbf{L}\|_{cd} = (\|\mathbf{z}\|_a \cdot \|\mathbf{y}^T\|_b) / (\|\mathbf{z}\|_c \cdot \|\mathbf{y}^T\|_d) = \mu_{ac} \cdot \mu_{db}. \quad \text{End of proof.}$$

**Exercise 20:** Use the six maxima  $\mu_{pq}$  tabulated in Ex. 3 to tabulate all 72 nontrivial maxima  $\mu_{abcd}$  of ratios of pairs of operator norms obtained when  $a, b, c$  and  $d$  range over the set  $\{1, 2, \infty\}$ . This task is for a computer program that manipulates symbols like “ $\sqrt{m}$ ”.

## Isometries

An *Isometry*  $\mathbf{Q}$  is a linear map from a normed vector space to itself that preserves the norm;  $\|\mathbf{Q} \cdot \mathbf{x}\| = \|\mathbf{x}\|$  for all vectors  $\mathbf{x}$ . The space’s isometries form a *Multiplicative Group* because a product of isometries is an isometry. Operator norm  $\|\mathbf{L}\|$  is unchanged if  $\mathbf{L}$  is postmultiplied by an isometry for the domain of  $\mathbf{L}$  and/or premultiplied by an isometry for its target-space.

For  $\|\dots\|_\infty$  and  $\|\dots\|_1$  the group is generated by all *Permutations* and all *Sign Changers* — diagonal matrices whose diagonal entries all have magnitude 1 — so the group consists of all square matrices of the right dimension in whose every row and every column only one element is nonzero and its magnitude is 1. These spaces have  $2^n \cdot n!$  real isometries of dimension  $n$ .

Those and infinitely many more belong to the group of isometries  $Q$  for  $\|\dots\|_2$ ; these are the *Orthogonal* matrices  $Q^T = Q^{-1}$  for real spaces, *Unitary* matrices  $Q^* = Q^{-1}$  for complex spaces. Orthogonal matrices represent linear operators that rotate and/or reflect real Euclidean space. *Proper Rotations* are generated by either  $Q = \exp(S)$  or  $Q = (I+S)^{-1} \cdot (I-S)$  as  $S = -S^T$  runs through all real *Skew-Symmetric* matrices. Neither of these formulas for proper rotations  $Q$ , each of which must have  $\det(Q) = +1$ , is fully satisfactory. The formula  $Q = \exp(S)$  is many-to-one;  $\log(\exp(S)) \neq S$  if  $\|S\|_2 > \pi$ . The *Cayley Transform* formula  $Q = (I+S)^{-1} \cdot (I-S)$  is one-to-one because  $S = (I+Q)^{-1} \cdot (I-Q)$ , but cannot generate any proper rotation  $Q$  that has  $-1$  as an eigenvalue (necessarily of even multiplicity) except by taking a limiting value as the elements of  $S$  approach infinities in suitably correlated ways.

A simple orthogonal reflection  $W = W^T = W^{-1} = I - w \cdot w^T$  is determined by its mirror-plane whose equation is  $w^T \cdot x = 0$  and whose normal  $w$  has been scaled to have length  $\|w\|_2 = \sqrt{2}$ . You should confirm easily that  $W \cdot w = -w$  but  $W \cdot x = x$  if  $w^T \cdot x = 0$ . Numerical analysts call them “Householder Reflections” because Alston S. Householder demonstrated their virtues for solving Least-Squares problems on computers in the mid 1950s, and then they became staples for eigenvalue and singular value computations too. Every  $n$ -by- $n$  orthogonal matrix  $Q$  can be expressed as a product of at most  $n$  such reflections, and an even number of them if  $Q$  is a proper rotation, but the reflections in the product need not be determined uniquely by  $Q$ .

Any linear map  $L$  from one Euclidean space to another can be reduced to its unique canonical form by isometries in its domain and target spaces. This canonical form of the matrix  $L$  is a similarly dimensioned (perhaps not square) diagonal matrix  $V$  of sorted nonnegative *Singular Values* satisfying  $L = Q \cdot V \cdot P^T$  (the *Singular-Value Decomposition*) in which  $Q$  and  $P$  are (square) orthogonal matrices not necessarily determined uniquely by  $L$  though its singular values on the diagonal of  $V$  are determined uniquely if sorted in descending order. In a more compact *SVD*, diagonal  $V$  is square of dimension  $r := \text{rank}(L)$ , so only the nonzero singular values of  $L = Q \cdot V \cdot P^T$  appear, and  $Q^T \cdot Q = P^T \cdot P = I$  ( $r$ -by- $r$ ). This compact *SVD* asserts algebraically a geometrical relationship called “Autonne’s Theorem”:

Every linear map  $L$  of rank  $r$  from one Euclidean space to another is a *Dilatation* described as follows:  $L$  selects  $r$  vectors (columns of  $P$ ) from an orthonormal basis for  $\text{Domain}(L)$  and associates them one-to-one with  $r$  vectors (columns of  $Q$ ) constituting an orthonormal basis for  $\text{Range}(L)$ ; then  $L$  projects its domain orthogonally onto its  $r$ -dimensional subspace spanned by the selected  $r$  vectors, stretches or squashes each of these coordinate directions by its corresponding singular value in  $V$ , and copies the result onto  $\text{Range}(L)$  after aligning the result’s stretched-or-squashed coordinate directions along their associated coordinate directions in  $\text{Range}(L)$ . If  $L$  maps a Euclidean space to itself the last realignment amounts to a rotation.

## Fritz John's Ellipsoid Theorem

His contribution to the *1948 Courant Anniversary Volume* (InterScience/Wiley, New York) was a proof of a slightly more general statement than the following ...

**Theorem:** Any given centrally symmetric convex body  $\mathbb{B}$  in  $n$ -space can be circumscribed by an ellipsoid  $\mathbb{E}$  closely enough that  $\sqrt{n} \cdot \mathbb{B} \supseteq \mathbb{E} \supseteq \mathbb{B}$ .

The constant  $\sqrt{n}$  cannot be reduced without falsifying the theorem when  $\mathbb{B}$  is a hypercube or more general parallelepiped. Compare this constant with the bigger constant  $n$  in Auerbach's theorem where  $\mathbb{E}$  is drawn from parallelepipeds and  $\mathbb{B}$  can be the *hyperoctahedron* which is the unit ball for the norm  $\|\dots\|_1$  that figures in our Ex. 9. Fritz John's theorem can be restated in norm terms by interpreting  $\mathbb{B}$  as the unit ball of a given norm  $\|\dots\|$ . The restatement is ...

Any norm  $\|\dots\|$  in  $n$ -space can be approximated by  $\|\text{Ex}\|_2 := \sqrt{(\text{Ex})^T(\text{Ex})}$  closely enough, if matrix  $E$  is chosen appropriately, that  $1/\sqrt{n} \leq \|\text{Ex}\|_2/\|x\| \leq 1$  for every vector  $x \neq 0$ .

Fritz John's ellipsoid  $\mathbb{E} = \{x: (\text{Ex})^T(\text{Ex}) \leq 1\} = E^{-1} \cdot \{y: y^T y \leq 1\} = E^{-1} \cdot \Omega_2$  is the unit ball for the vector norm  $\|\text{Ex}\|_2$  just as  $\mathbb{B} = \{x: \|x\| \leq 1\}$  is the unit ball for the given norm  $\|x\|$ .

His more general statement covered arbitrary convex bodies  $\mathbb{B}$  for which  $\sqrt{n}$  was increased to  $n$ . Restricting his theorem to centrally symmetric bodies simplifies its proof to fit with what has already been presented in class. As he did, we shall characterize  $\mathbb{E}$  as the ellipsoid of least *Content* (area, volume, ...) circumscribing  $\mathbb{B}$ . Because  $\text{Content}(\mathbb{E}) = \text{Content}(\Omega_2)/|\det(E)|$  we seek, as he did, a matrix  $E$  that maximizes  $\det(E)$  subject to the constraint  $\|\text{Ex}\|_2/\|x\| \leq 1$  for all  $x \neq 0$ . But our argument, first presented in lecture notes for Math. 273 in 1974, will go more directly than his did.

First observe that two matrix norms  $\|Z\|_{2\bullet} := \max_{x \neq 0} \|Zx\|_2/\|x\|$  and  $\|Z\|_{\bullet 2} := \max_{x \neq 0} \|Zx\|/\|x\|_2$  are induced by the two vector norms in question. Now we seek a characterization of those matrices  $E$  that maximize  $\det(E)$  over the ball  $\|E\|_{2\bullet} \leq 1$  in  $n$ -by- $n$  matrix space, and hope to infer from that characterization that  $\|E^{-1}\|_{\bullet 2} \leq \sqrt{n}$ , which will imply  $\sqrt{n} \cdot \mathbb{B} \supseteq \mathbb{E} \supseteq \mathbb{B}$ .

At least one maximizing  $E$  must exist because  $\det(E)$  is a continuous function on a compact set, the unit ball  $\|E\|_{2\bullet} \leq 1$  in  $n$ -by- $n$  matrix space. For such a maximizing  $E$  we find that

$$\|E^{-1}\|_{\bullet 2} = \max \|E^{-1}v\| = \max w^T E^{-1}v \text{ over } \|w^T\| = \|v\|_2 = 1,$$

and this maximum is achieved at some  $w^T$  and  $v$  determined here as in Ex. 15 to satisfy

$$\|w^T\| = \|v\|_2 = 1 \text{ and } \|E^{-1}\|_{\bullet 2} = w^T E^{-1}v. \text{ and } w^T E^{-1} = \|E^{-1}\|_{\bullet 2} v^T.$$

The last equation is satisfied because, in order to achieve this maximum,  $w^T E^{-1}$  and  $v$  must be dual to each other with respect to the norm  $\|\dots\|_2$ . Meanwhile, because  $\|w^T\| = 1$ , every vector  $y$  has  $\|y\| \geq w^T y = w^T E^{-1} E y = \|E^{-1}\|_{\bullet 2} v^T E y$ . This will be used twice in  $(\dagger)$  below.

Now let  $Z := vw^T/\|E^{-1}\|_{\bullet,2} - E/(n+\beta)$  for any tiny  $\beta > 0$ , and let  $f(\mu) := \log(\det(E + \mu Z))$  be examined at tiny values  $\mu > 0$ . Jacobi's formula for the derivative of a determinant says that  $f'(\mu) := df(\mu)/d\mu = \text{Trace}((E + \mu Z)^{-1}Z)$  provided  $\mu$  is tiny enough that  $(E + \mu Z)^{-1}$  still exists. Therefore  $f'(0) = \text{Trace}(E^{-1}Z) = w^T E^{-1} v / \|E^{-1}\|_{\bullet,2} - n/(n+\beta) = \beta/(n+\beta) > 0$ . Since  $E$  maximizes  $f(0) = \log(\det(E))$  subject to the constraint  $\|E\|_{2,\bullet} \leq 1$ , it is violated by  $E + \mu Z$  for every sufficiently tiny  $\mu > 0$ ; in other words,  $\|E + \mu Z\|_{2,\bullet} > 1$  for every sufficiently tiny  $\mu > 0$ . For every such  $\mu$  some maximizing vector  $y = y(\mu)$  exists with  $\|y\| = 1$  and

$$\|(E + \mu Z)y\|_{2,\bullet}^2 = \|E + \mu Z\|_{2,\bullet}^2 \cdot \|y\|^2 > \|y\|^2 = 1 \geq \|E\|_{2,\bullet}^2 \geq \|Ey\|_{2,\bullet}^2.$$

Rearranging this algebraically produces a strict inequality

$$\begin{aligned} 0 < (\|(E + \mu Z)y\|_{2,\bullet}^2 - \|Ey\|_{2,\bullet}^2) / \mu &= 2(Ey)^T Zy + \mu \|Zy\|_{2,\bullet}^2 \\ &= 2v^T Ey \cdot w^T y / \|E^{-1}\|_{\bullet,2} - 2\|Ey\|_{2,\bullet}^2 / (n+\beta) + \mu \|Zy\|_{2,\bullet}^2 \quad (\dagger) \\ &= 2(w^T y / \|E^{-1}\|_{\bullet,2})^2 - 2\|Ey\|_{2,\bullet}^2 / (n+\beta) + \mu \|Zy\|_{2,\bullet}^2 \quad (\ddagger) \\ &\leq 2/\|E^{-1}\|_{\bullet,2}^2 - 2\|Ey\|_{2,\bullet}^2 / (n+\beta) + \mu \|Z\|_{2,\bullet}^2. \end{aligned}$$

Combine this with another inequality

$$\|Ey\|_{2,\bullet} \geq \|(E + \mu Z)y\|_{2,\bullet} - \mu \|Zy\|_{2,\bullet} > \|y\| - \mu \|Z\|_{2,\bullet} = 1 - \mu \|Z\|_{2,\bullet}.$$

to infer that

$$0 < 2/\|E^{-1}\|_{\bullet,2}^2 - 2(1 - \mu \|Z\|_{2,\bullet})^2 / (n+\beta) + \mu \|Z\|_{2,\bullet}^2 \rightarrow 2/\|E^{-1}\|_{\bullet,2}^2 - 2/(n+\beta) \text{ as } \mu \rightarrow 0+.$$

Consequently  $\|E^{-1}\|_{\bullet,2}^2 \leq n+\beta$  for every  $\beta > 0$ , which proves  $\|E^{-1}\|_{\bullet,2} \leq \sqrt{n}$  as we had hoped.

• • • •

Fritz John's Ellipsoid Theorem has far-reaching implications; here briefly are three of them:

• **Ellipsoidal Bounds for Errors in Computed Approximations**

A norm chosen to gauge computational errors should ideally have this property: All errors of about the same norm are about equally (in)consequential. Such a norm may be difficult if not impossible to find, and may be found only after at least one attempt at the computation has been tried. But the norms implicit in many approximate computational algorithms typically resemble the vector norms  $\|\dots\|_p$  discussed extensively in these notes; their common characteristic is that the norm of perturbations of a vector are little changed by permutations, which means that errors in one component of a vector will be deemed roughly as (in)consequential as errors of the same size in any other component. This can be a serious mistake.

For instance, a mathematical model of the processes that control the growth of an elephant from a fertilized ovum will involve amounts of materials ranging from micrograms of hormones to tons of flesh. Were all these amounts reckoned in grams and then arranged in a column vector representing the state of the organism's development, the vector's elements could range from 0.0000001 to 10000000.0. An error of the order of 1.0 committed during an equation-solving process could alter some elements imperceptibly and alter others overwhelmingly. To govern numerical errors better, and to bring variables closer to values humans can assess easily, we choose different units — micrograms, milligrams, grams, kilograms, tonnes — for different variables. And we may change units as the growth process passes through different phases like

implantation in the endometrium, birth, and maturation. The choice of units is tantamount to premultiplying the state-vector by a diagonal matrix of scale factors before applying a familiar  $\|\dots\|_p$ -norm to its perturbations. Sometimes diagonal premultiplication is not general enough.

Sometimes later states of an evolving process respond to early perturbations far more severely in some directions than others, and those directions need not be parallel to coordinate axes. In such cases the severity of an early error  $x$  should be gauged by a norm  $\|x\|$  whose unit ball is squashed in some of those directions, elongated in others. According to Fritz John's theorem, a premultiplying matrix  $E$  can be so chosen that  $\|x\|$  is approximated by  $\|E \cdot x\|_2$  to within a factor no worse than  $(\text{Dimension})^{\pm 1/4}$ . This much uncertainty about an error estimate is often tolerable provided the dimension of  $x$  is not too big. The hard part is finding a satisfactory  $E$ .

Over the several decades since Fritz John's Theorem was published, ellipsoidal error-bounds like  $\|E \cdot x\|_2$  have come to be appreciated for qualities not apparent from his theorem. See ...

- <http://www.cs.berkeley.edu/~wkahan/Math128/Ellipsoi.pdf> and [.../ODEintvl.pdf](http://www.cs.berkeley.edu/~wkahan/ODEintvl.pdf).
- *Uncertain Dynamic Systems* by Fred. C. Schewpe (1973, Prentice-Hall, NJ).
- "The wrapping effect, ellipsoid arithmetic, stability and confidence regions" by Arnold Neumaier, pp. 175-190 in *Computing Supplementum 9* (1993).

### • The Banach Space Projections Constant

A *Projection* is a linear map  $\mathbf{P}$  of a vector space into itself satisfying  $\mathbf{P}^2 = \mathbf{P}$ . To avoid trivialities we assume also that  $\mathbf{I} \neq \mathbf{P} \neq \mathbf{O}$ . Then  $\mathbf{P}$  cannot have an inverse (otherwise it would imply  $\mathbf{I} = \mathbf{P}^{-1} \cdot \mathbf{P} = \mathbf{P}^{-1} \cdot \mathbf{P}^2 = \mathbf{I} \cdot \mathbf{P} = \mathbf{P}$ ), so the range of  $\mathbf{P}$  is the proper subspace onto which  $\mathbf{P}$  projects the whole space.  $\mathbf{I} - \mathbf{P}$  is a projection onto a complementary subspace. An example is  $\mathbf{P} = \begin{bmatrix} 1 & 2 \\ 0 & 0 \end{bmatrix}$ . Note that  $\mathbf{P}$  need not be an *Orthogonal Projection*.

Orthogonal projections are peculiar to Euclidean spaces and are special there too. Projection  $\mathbf{P}$  is orthogonal just when  $\mathbf{P}^2 = \mathbf{P} = \mathbf{P}^T$ ; and then  $\|\mathbf{P}\|_2 = 1$ . This follows from observing that every eigenvalue of  $\mathbf{P}^T \cdot \mathbf{P} = \mathbf{P}^2 = \mathbf{P}$  is either 0 or 1. Any other projection  $\mathbf{Q}$  onto the same subspace as  $\text{Range}(\mathbf{P})$  must have  $\|\mathbf{Q}\|_2 > 1$ . This follows from equations  $\mathbf{Q} \cdot \mathbf{P} = \mathbf{P} = \mathbf{P}^T$  and  $\mathbf{P} \cdot \mathbf{Q} = \mathbf{Q}$  that say  $\mathbf{P}$  and  $\mathbf{Q}$  are projections each onto the other's range, and from a change to new orthonormal coordinates that transform  $\mathbf{P}$  into  $\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$  but  $\mathbf{Q}$  into  $\begin{bmatrix} 1 & R \\ 0 & 0 \end{bmatrix}$  with  $R \neq 0$ .

A non-Euclidean normed space is called a *Banach* space after Stefan Banach, who studied them intensively in the 1920s and 1930s until the Nazis overran Poland and killed as many of its intellectuals as they could before 1945; he outlasted the Nazis only to die later that year from lung cancer. He had greatly advanced the study of infinite-dimensional spaces. In these notes all spaces' dimensions are finite.

A Banach space's norm violates the ...

*Parallelogram Law:*  $\|\mathbf{x} + \mathbf{y}\|^2 + \|\mathbf{x} - \mathbf{y}\|^2 = 2\|\mathbf{x}\|^2 + 2\|\mathbf{y}\|^2$  for all  $\mathbf{x}$  and  $\mathbf{y}$  satisfied by Euclidean norms even if the coordinate system is not orthonormal. Consequently

orthogonality is almost entirely absent from Banach spaces. See Ex. 7 for a feeble exception. See the class notes on “How to Recognize a Quadratic Form”, <.../MathH110/QF.pdf>, for a proof that explains why only Euclidean norms honor the Parallelogram Law.

Each Banach space has a multiplicative operator norm induced by the vector norm, and when computed for a projection  $\mathbf{P} = \mathbf{P}^2$  the norm must satisfy  $\|\mathbf{P}\| = \|\mathbf{P}^2\| \leq \|\mathbf{P}\|^2$ , so  $\|\mathbf{P}\| \geq 1$ . How much bigger than 1 must  $\|\mathbf{P}\|$  be? This question, posed by Banach, was first answered in 1972 by Yehoram Gordon who established, with the aid of Fritz John’s Ellipsoid Theorem, that any  $r$ -dimensional subspace in any Banach space is the range of at least one projection  $\mathbf{P}$  of rank  $r$  and norm  $\|\mathbf{P}\| \leq \sqrt{r}$ ; and no constant smaller than  $\sqrt{r}$  can be valid for every  $r$ -dimensional subspace of every Banach space. Gordon’s proof is too long to reproduce here.

### • The Smallest Generalized Inverse

Every (possibly rectangular) matrix  $F$  has at least one *Generalized Inverse*  $G$  satisfying the one equation  $F \cdot G \cdot F = F$  that every generalized inverse must satisfy.  $x = G \cdot y$  is a solution of the possibly over- or under-determined linear equation  $F \cdot x = y$  if a solution  $x$  exists; and if not,  $G \cdot y$  is an approximate solution in some sense. If  $F$  is rectangular or rank deficient it has infinitely many generalized inverses  $G$  almost all of which have enormous magnitudes  $\|G\|$  gauged by any norm. This follows from the observation that  $G+Z$  is another generalized inverse whenever  $Z$  satisfies either  $F \cdot Z = O$  or  $Z \cdot F = O$ . An oversized  $\|G\|$  induces severe numerical misbehavior because it amplifies small errors in  $y$  when  $G \cdot y$  is computed; none of its computed digits will contain useful information if  $\|G\|$  is too big. There are extreme cases when every  $\|G\|$  is too big. Every generalized inverse  $G$  of  $F$  must satisfy

$$\|G\| \geq 1 / (\text{minimum } \|\Delta F\| \text{ for which } \text{rank}(F - \Delta F) < \text{rank}(F))$$

in which the two matrix norms need only be compatible with the vector norms in the domain and target spaces of  $F$ . The foregoing assertions are Lemma 1 and Theorem 5 in the class notes on “Huge Generalized Inverses of Rank-Deficient Matrices”, <.../MathH110/Gllite.pdf>. Theorem 8 stated but not proved in those notes asserts, for the operator norms induced by the vector norms, that at least one generalized inverse  $G$  also satisfies

$$\|G\| \leq \sqrt{\text{rank}(F)} / (\text{minimum } \|\Delta F\| \text{ for which } \text{rank}(F - \Delta F) < \text{rank}(F)).$$

My proof uses Fritz John’s Ellipsoid Theorem but is still too long to reproduce here.

The foregoing two bounds upon  $\|G\|$  have valuable practical implications when the data in  $F$  are uncertain enough that some nearby  $F - \Delta F$  of lower rank differs from  $F$  by less than its uncertainty. Changing the data  $F$  to a nearly indistinguishable matrix  $F - \Delta F$  of lowest rank may reduce the norm of its nearly minimal generalized inverse enough to forestall numerical obscurity. If this can be accomplished, we can accomplish it by means of a Singular Value Decomposition after applying whatever coordinate changes in the domain and target spaces of  $F$  are necessary to make the spaces’ norms approximately Euclidean. Provided dimensions are not too big, Fritz John’s Ellipsoid Theorem says that these necessary coordinate changes exist without saying how to find them. Let’s hope they amount only to diagonal scaling.

... ..

More applications of Fritz John's Ellipsoid Theorem and another longer proof for its centrally symmetric case can be found in Keith Ball's lecture notes "An Elementary Introduction to Modern Convex Geometry", pp. 1-58 of *Flavors of Geometry*, MSRI Publications - Volume 31, Edited by Silvio Levy for Cambridge University Press, Cambridge, 1997. Ball's notes are also posted at <http://www.msri.org/publications/books/Book31/files/ball.pdf>. Don't read too much into the title's word "Elementary".