A.1.1 Matrices and Vectors

Definition of **Matrix**. An MxN *matrix* $\mathbf{A}$ is a two-dimensional array of numbers

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1N} \\ a_{21} & a_{22} & \dots & a_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ a_{M1} & a_{M2} & \dots & a_{MN} \end{bmatrix}$$

A matrix can also be written as $\mathbf{A} = (a_{nm})$ where $n = 1, 2, \dots, N$ and $m = 1, 2, \dots, M$. Matrices are usually written as boldface, upper-case letters.

Definition of **Matrix Transpose**. The transpose of $\mathbf{A}$, denoted by $\mathbf{A}^t$, is the NxM matrix (a_{mn}) or

$$\mathbf{A}^t = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1M} \\ a_{21} & a_{22} & \dots & a_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ a_{N1} & a_{N2} & \dots & a_{NM} \end{bmatrix}$$

Example. A example of a matrix A and it's transpose A^t is:

$$A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{bmatrix} \text{ and } A^t = \begin{bmatrix} 1 & 3 & 5 \\ 2 & 4 & 6 \end{bmatrix}$$

Definition of **Vectors**.. A *column vector* is a $Bx1$ matrix and a *row vector* is a $1xB$ matrix, where $B \geq 1$. Column and row vectors are usually simply referred to as *vectors* and they are assumed to be column vectors unless they are explicitly identified as row vector or if it is clear from the context. Vectors are denoted by boldface, lower-case letters.

$$\mathbf{x} = [x_1, x_2, \dots, x_B]^t = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_B \end{bmatrix}$$

Definition of **Dimension**. The integer B is called the *dimension* of the vector $\mathbf{x}$ in the definition above. The phrase $\mathbf{x}$ is $B - dimensional$ is equivalent to stating that the dimension of $\mathbf{x}$ is B.

Definition of **0 and 1 vectors**.. The vectors

$$\mathbf{0} = (0, 0, \dots, 0)^t \text{ and } \mathbf{1} = (1, 1, \dots, 1)^t$$

are called the *Origin* or *Zero Vector* and the *One Vector*, respectively.

Definition of **Vector Addition and Scalar Multiplication**.. The addition of two vectors $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^B$ is defined as: $\mathbf{x}_3 = [x_{11} + x_{21}, x_{12} + x_{22}, \dots, x_{1B} + x_{2B}]^t$. The word "scalars" is another word for numbers. Scalar multiplication is the product of a number α and a vector $\mathbf{x}$, defined by $\alpha\mathbf{x}$ = $[\alpha x_1, \alpha x_2, \ldots, \alpha x_B]^t$.

Example. If $\mathbf{x}_1 = [1, 2]$ and $\mathbf{x}_2 = [3, 4]$ then $\mathbf{x}_3 = \mathbf{x}_1 + \mathbf{x}_2 = [1 + 3, 2 + 4]$ = $[4, 6]$. If $\alpha = 3$, then $\alpha\mathbf{x}_1$ = $[3, 6]$.

Definition of **Matrix Multiplication**. If $\mathbf{A}$ and $\mathbf{B}$ are M×N and N×P matrices, then the *matrix product*, $\mathbf{C} = \mathbf{AB}$, is the M×P matrix defined by

$$c_{mp} = \sum_{n=1}^{N} a_{mn} b_{np}$$

If $A_{3\times 2}$ and $B_{2\times 2}$ are the following matrices:

$$A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{bmatrix} \quad \text{and } B = \begin{bmatrix} 11 & 12 \\ 13 & 14 \end{bmatrix}$$

then $C = AB$ is the 3×2 matrix:

$$C = \begin{bmatrix} 37 & 40 \\ 85 & 92 \\ 109 & 118 \end{bmatrix}$$

For example, $c_{2,1}$ = (3)(11) + (4)(13) = 33 + 52 = 85.

Definition of **Dot, or Inner, Product**. If $\mathbf{x}$ and $\mathbf{y}$ are B-dimensional vectors, then the matrix product $\mathbf{x}^t\mathbf{y}$ is referred to as the *Dot* or *Inner Product* of $\mathbf{x}$ and $\mathbf{y}$.

Example. If $\mathbf{x}$ = $[-2, 0, 2]^t$ and $\mathbf{y}$ = $[4, 1, 2]^t$, then $\mathbf{x}^t\mathbf{y}$ = (-2)(4) + (0)(1) + (2)(2) = -4.

Note that a linear system of equations can be represented as a matrix multiplication. For example, the system

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 &= b_1 \\ a_{21}x_1 + a_{22}x_2 &= b_2 \end{aligned}$$

and be written in matrix-vector form as

$$\mathbf{Ax} = \mathbf{b}.$$

Definition of **Diagonal Matrix**. A diagonal matrix $\mathbf{D} = (d_{mn})$ is an N×N matrix with the property that $d_{mn} = 0$ if $m \neq n$.

Definition of **Identity Matrix**. The N×N *Identity Matrix*, denoted by $\mathbf{I}_N$ or just $\mathbf{I}$ if N is known, is the N×N diagonal matrix with $d_{nn} = 1$ for $n = 1, 2, \ldots, N$. Notice that if $\mathbf{A}$ is any N×N matrix, then $\mathbf{AI}_N$ = $\mathbf{I}_N\mathbf{A}$ = $\mathbf{A}$.

Definition of **Inverse Matrix**. Let $\mathbf{A}$ be an N×N square matrix. If there is a matrix $\mathbf{B}$ with the property that $\mathbf{AB}$ = $\mathbf{BA}$ = $\mathbf{I}_N$, then $\mathbf{B}$ is called *the inverse of* $\mathbf{A}$ or $\mathbf{A}$ *inverse*.

It is denoted by $\mathbf{A}^{-1}$. The inverse of a square matrix $\mathbf{A}$ does not always exist. If it does exist, then $\mathbf{A}$ is *invertible*.

A.1.2 Vector Spaces

Definition of **Vector, or Linear, Space**. Let B denote a positive integer. A finite-dimensional *Vector, or Linear, Space* with *dimension* B is a collection of $B - dimensional$ vectors, V that satisfies commutative, associative, and distributive laws and with the properties that:

1. For every $\mathbf{x} \in V$ and $\mathbf{y} \in V$, the sum $(\mathbf{x} + \mathbf{y}) \in V$.
2. For every real number s and $\mathbf{x} \in V$, the product $s\mathbf{x} \in V$.
3. For every $\mathbf{x} \in V$, the additive inverse $-\mathbf{x} \in V$, which implies that $\mathbf{0} \in V$.

Notation. A B-dimensional vector space is often denoted by $\mathbb{R}^B$.

Definition of **Subspace**. A *Subspace* is a subset of a vector space that is also a vector space. Notice that subspaces of vector spaces always include the origin. Although it is not common, we us the notation $\mathcal{S} \sqsubset \mathcal{V}$ to state that $\mathcal{S}$ is a subspace of a vector space $\mathcal{V}$.

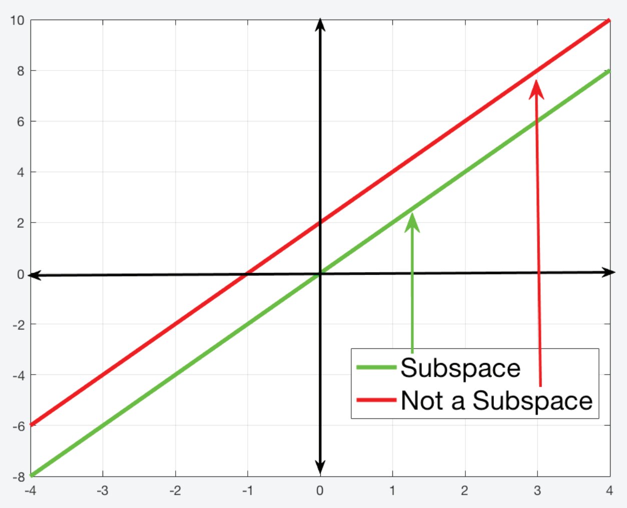

Figure A.1 Subspace and non-Subspace

Subspaces are often used in algorithms for analyzing image spectrometer data. The motivation is that spectra with certain characteristics might all be contained in a subspace.

Definition of **Linear Combination** .

If $\mathbf{y}$, $\{\mathbf{x}_d\}_{d=1}^{D} \in \mathbb{R}^B$ and $\mathbf{y} = a_1\mathbf{x}_1 + a_2\mathbf{x}_2 + \cdots + a_D\mathbf{x}_D$ then $\mathbf{y}$ is a *linear combination* of $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_D\}$. The numbers $a_1, a_2, \ldots, a_D$ are called the *coefficients* of the linear combination. Notice that the linear combination can be written in matrix-vector form as:

$$\mathbf{y} = \mathbf{X}\mathbf{a}$$

where $\mathbf{X}$ is the B×D matrix with $\mathbf{x}_d$ as the d^{th} column and $\mathbf{a}$ is the vector of coefficients.

Definition of **Linear Transformation**.. Let $\mathcal{V}$ and $\mathcal{W}$ denote two vector spaces. A function $f : \mathcal{V} \to \mathcal{W}$ is called a *Linear Transformation* if $\forall a, b \in \mathbb{R}$ and $\forall \mathbf{x}, \mathbf{y} \in \mathcal{V}$, the statement $f(a\mathbf{x} + b\mathbf{y})$ = $af(\mathbf{x}) + bf(\mathbf{y})$ is true.

If A is an $m \times n$ matrix and $\mathbf{x} \in \mathbb{R}^m$, then $f(\mathbf{x}) = A\mathbf{x}$ is a linear transformation.

Definition of **Linear Independence**. If $\mathbb{B}$=$\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_D\}$ is a set of vectors with the property that

$$a_1\mathbf{x}_1 + a_2\mathbf{x}_2 + \cdots + a_D\mathbf{x}_D = 0 \implies a_1 = a_2 = \ldots = a_D = 0$$

then the vectors in the set $\mathcal{B}$ are called *Linearly Independent*. Informally, no vector in a set of linearly independent vectors can be written as a linear combination of the other vectors in the set.

Fact.There can be no more than B linearly independent vectors in a $B - dimensional$ vector space. If $\mathcal{B}$ = $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_B\}$ is a set of B linearly independent vectors in a $B - dimensional$ space $\mathcal{V}$, then every $\mathbf{x} \in \mathcal{V}$ can be written uniquely as a linear combination of the vectors in $\mathcal{B}$.

Definition of **Basis**. A *Basis* of a subspace of B-dimensional space, $\mathcal{S}$, is a set of vectors $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_D\}$ with the property that every vector $\mathbf{x} \in \mathcal{S}$ can be written one and exactly one way as a linear combination of the elements of the basis. In other words, if $\mathbf{x} \in \mathcal{S}$ then there is a unique set of coefficients, $a_1, a_2, \ldots, a_D$ such that $\mathbf{y} = a_1\mathbf{v}_1 + a_2 + \mathbf{v}_2, \cdots + a_D\mathbf{v}_D$. It must be the case that $D \leq B$.

Fact.There are infinitely many bases for any subspace but they all have the same number of elements. The number of elements is called the dimension of the subspace.

Definition of **Standard Basis**. The standard basis of a D-dimensional subspace is the set of vectors $\{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_D\} \subset \mathbb{R}^D$ where the d^{th} element of the d^{th} vector $\mathbf{s}_d$, s_{dd}, is equal to one and the rest of the elements are equal to zero. That is

$$s_{dj} = 1 \text{ if } d = j \text{ and } s_{dj} = 0 \text{ if } d \neq j.$$

Example. . The standard basis for $\mathbb{R}^3$ is

$$\mathbf{s}_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \text{ and } \mathbf{s}_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \text{ and } \mathbf{s}_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

Any 3-Dimensional vector can be represented as a linear combination of the standard basis vectors:

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = x_1\mathbf{s}_1 + x_2\mathbf{s}_2 + x_3\mathbf{s}_3 = x_1 \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + x_2 \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} + x_3 \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

Fact.If the columns of an N×N matrix $\mathbf{A}$ form a basis for $\mathbb{R}^N$, then A is invertible.

Bases can be though of as coordinate systems. It can be useful to represent vectors in terms of different bases, or coordinate systems. Therefore, the method for changing the coordinates used to represent a vector from one basis to another is first described. Then, a method for determining a linear transformation for mapping one basis to another basis is described. The two methods are referred to as Change of Coordinates or Change of Basis.

Definition of **Change of Basis**. The process of changing the representation of a vector $\mathbf{x}$ from a representation in terms of one basis to a representation in terms of a different basis is called a *Change of Basis* transformation.

Calculating a change of basis transformation may require solving a linear system. To see this, let $\mathcal{U} = \{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_B\}$ and $\mathcal{V} = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_B\}$ be two bases for $\mathbb{R}^B$. Suppose the representation of a vector $\mathbf{x}$ in terms of the basis $\mathcal{U}$ is known to be

$$\mathbf{x} = a_1\mathbf{u}_1 + a_2\mathbf{u}_2 + \cdots + a_B\mathbf{u}_B = \mathbf{Ua}$$

and that the representation of $\mathbf{x}$ in terms of $\mathcal{V}$ is unknown

$$\mathbf{x} = c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \cdots + c_B\mathbf{v}_B = \mathbf{Vc}.$$

In other words, $\mathbf{a} = (a_1, a_2, \ldots, a_B)$ is known and $\mathbf{c} = (c_1, c_2, \ldots, c_B)$ is unknown. Then $\mathbf{c} = \mathbf{V}^{-1}\mathbf{Ua}$. In particular, if $\mathcal{U}$ is the standard basis, then $\mathbf{U} = \mathbf{I}$ and $\mathbf{x} = [a_1, a_2, \ldots, a_B]^t$, so $\mathbf{c} = \mathbf{V}^{-1}\mathbf{x}$. The matrix $\mathbf{W} = \mathbf{V}^{-1}\mathbf{U}$ is called the *change of basis matrix*.

Definition of **Pseudo-inverse**. The definition of a pseudo-inverse is motivated by linear systems of equations, $A\mathbf{x} = \mathbf{b}$. If A is not square, then A^{-1} is not defined so one cannot write $\mathbf{x} = A^{-1}\mathbf{b}$. However, one can write $(A^tA)\,\mathbf{x} = A^t\mathbf{b}$. Note that this is not equivalent to the original linear system of equations. The matrix, A^tA is $N \times N$ so, assuming that $N < M$, it is almost certainly invertible. Therefore, $\mathbf{x} = (A^tA)^{-1}\,A^t\mathbf{b}$ is a solution of $(A^tA)\,\mathbf{x} = A^t\mathbf{b}$. The matrix $(A^tA)^{-1}\,A^t\mathbf{b}$ is called the *pseudo-inverse* of A and is denoted by A^+.

A.1.3 Norms, Metrics, and Dissimilarities

Definition of **norm**. A norm produces a quantity computed from vectors that somehow measures the size of the vector. More formally, a norm is a function, $\mathcal{N} : \mathbb{R}^B \to \mathbb{R}_+$ with the properties that $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^B$ and $a \in \mathbb{R}$:

1. $\mathcal{N}(\mathbf{x}) \geq 0$ and $\mathcal{N}(\mathbf{x}) = 0$ if and only if $\mathbf{x} = 0$.
2. $\mathcal{N}(a\mathbf{x}) = |a|\mathcal{N}(\mathbf{x})$.
3. $\mathcal{N}(\mathbf{x}+\mathbf{y}) \leq \mathcal{N}(\mathbf{x}) + \mathcal{N}(\mathbf{y})$

Notation. The norm is usually written with the vertical bar notation: $\mathcal{N}(\mathbf{x}) = \|\mathbf{x}\|$.

Definition of $\mathcal{L}_p$ **norm, or just** p**-norm**.. Assume $p \geq 1$. The *p-norm* is defined by $\|\mathbf{x}\|_p = \left(\sum_{b=1}^{B} |x_b|^p\right)^{\frac{1}{p}}$.

Definition of **Euclidean Norm**.. The *Euclidean Norm* is the p-norm with $p = 2$, that is, the $\mathcal{L}_2$ norm. If p is not specified, then it is assumed that $\|\mathbf{x}\|$ denotes the Euclidean norm. Note that the Euclidean norm can be written as $\|\mathbf{x}\| = \sqrt{\mathbf{x}^t\mathbf{x}}$.

Definition of $\mathcal{L}_\infty$ **norm, or just** ∞**-norm**.. The ∞ *norm* is $\|\mathbf{x}\|_\infty = \max_{b=1}^{B} |x_b|$.

Definition of **Unit Norm Vector**. A vector $\mathbf{x}$ is said to have *unit norm* if $\|\mathbf{x}\| = 1$. Note that if $\mathbf{x}$ is any vector then $\frac{\mathbf{x}}{\|\mathbf{x}\|}$ has unit norm.

Example. The set of all vectors $\mathbf{x} \in \mathbb{R}^2$ with unit norm are equivalent to a circle, called the unit circle, since the set of all points with $x_1^2 + x_2^2 = 1$ is a circle of radius 1.

Definition of **Generalized Unit Circle**.. Let $\mathcal{N}$ be a norm. The set $\mathcal{U} = \{\mathbf{x}|\mathcal{N}(\mathbf{x}) = 1\}$ is called a *Generalized Unit Circle*. Often, for simplicity, the set $\mathcal{U}$is called a *Unit Circle* even though it is only a geometric circle of the Euclidean norm. It is sometimes useful to visualize the unit circle for other norms. Some examples are shown in A.2.

Fact.If $\mathbf{x}$ and $\mathbf{y}$ are vectors, then $\mathbf{x}^t\mathbf{y} = \|\mathbf{x}\|\|\mathbf{y}\|\cos\theta$, where θ is the angle between $\mathbf{x}$ and $\mathbf{y}$. If $\mathbf{x}$ and $\mathbf{y}$ are normalized, then the inner product of the vectors is equal to the cosine of the angle between the vectors.

$$\frac{\mathbf{x}^t\mathbf{y}}{\|\mathbf{x}\|\|\mathbf{y}\|} = \left(\frac{\mathbf{x}}{\|\mathbf{x}\|}\right)^t \left(\frac{\mathbf{y}}{\|\mathbf{y}\|}\right) = \cos\theta.$$

Definition of **Distance Function or Metric**. A *distance* or *metric* is a function defined on pairs of vectors,

$$d : \mathbb{R}^B \text{x} \mathbb{R}^B \to \mathbb{R}_+$$

and has the following properties:

1. $d(\mathbf{x}, \mathbf{y}) \geq 0$ and $d(\mathbf{x}, \mathbf{y}) = 0$ if and only if $\mathbf{x} = \mathbf{y}$
2. $d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$
3. For any $\mathbf{z}$, $d(\mathbf{x}, \mathbf{y}) \leq d(\mathbf{x}, \mathbf{z}) + d(\mathbf{z}, \mathbf{y})$

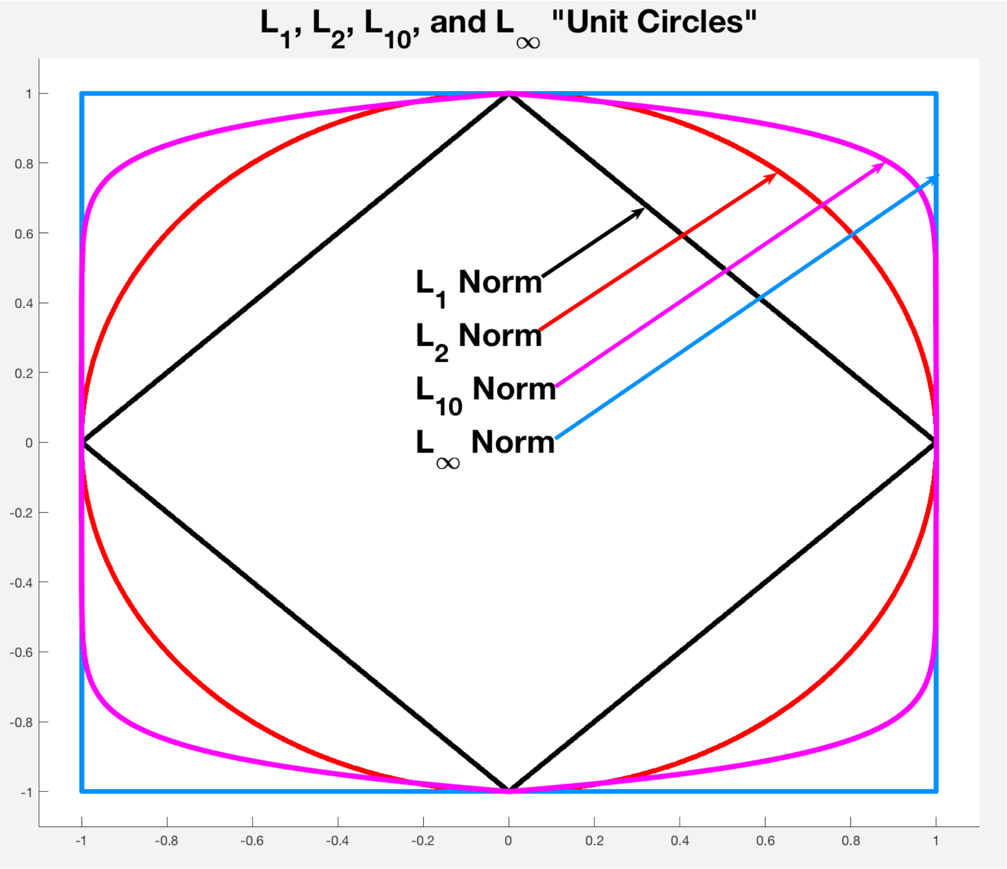

Figure A.2 Some Unit Circles. Note that as p increases, the $\mathcal{L}_p$ norm approaches the $\mathcal{L}_\infty$ norm.

Definition of $\mathcal{L}_p$ **distance**. $d_p(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_p$.

Note that the Euclidean distance squared can be written as

$$d_2(\mathbf{x}, \mathbf{y})^2 = \|\mathbf{x} - \mathbf{y}\|^2 = (\mathbf{x} - \mathbf{y})^t (\mathbf{x} - \mathbf{y}) = \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 - 2\mathbf{x}^t\mathbf{y}.$$

Therefore, if $\mathbf{x}$ and $\mathbf{y}$ are unit vectors, then measuring Euclidean distance between vectors is equivalent to measuring the cosine of the angle between the two vectors:

$$d_2(\mathbf{x}, \mathbf{y})^2 = \|\mathbf{x} - \mathbf{y}\|^2 = (\mathbf{x} - \mathbf{y})^t (\mathbf{x} - \mathbf{y}) = 1 + 1 - 2\mathbf{x}^t\mathbf{y} = 2 - 2\cos\theta$$

which implies that, $\cos\theta = 1 - 0.5\|\mathbf{x} - \mathbf{y}\|^2$.

Definition of **Orthogonality and Orthonormal Bases**. Two non-zero vectors $\mathbf{x}$ and $\mathbf{y}$ are called *orthogonal* if $\mathbf{x}^t\mathbf{y} = 0$. Note that orthogonal is another word for perpendicular since (for non-zero vectors) $\mathbf{x}^t\mathbf{y} = \|\mathbf{x}\|\|\mathbf{y}\|\cos\theta = 0$ only happens if $\cos\theta = 0$. An *orthonormal basis* is a basis $\mathcal{U} = \{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_B\}$ with the property that $\mathbf{u}_i\mathbf{u}_j = \delta_{ij}$ where δ_{ij} is the Kronecker delta function.

Example. Computing Coefficients of Orthonormal Bases. Suppose that $\mathcal{U} = \{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_B\}$ is an orthonormal basis of $\mathbb{R}^B$, and $\mathbf{x} \in \mathbb{R}^B$. Then there exists coefficients $c_1, c_2, \ldots, c_B$ such that $\mathbf{x} = c_1\mathbf{u}_1 + c_2\mathbf{u}_2 + \cdots + c_B\mathbf{u}_B$. In general, one must solve a linear system of equations to find the coefficients. However, with an orthonormal basis, it is much easier. For example, to find c_1, transpose $\mathbf{x}$

$$\mathbf{x}^t = (c_1\mathbf{u}_1 + c_2\mathbf{u}_2 + \cdots + c_B\mathbf{u}_B)^t = c_1\mathbf{u}_1^t + c_2\mathbf{u}_2^t + \cdots + c_B\mathbf{u}_B^t$$

and then multiply on the right by $\mathbf{u}_1$

$$\mathbf{x}^t\mathbf{u}_1 = c_1\mathbf{u}_1^t\mathbf{u}_1 + c_2\mathbf{u}_2^t\mathbf{u}_1 + \cdots + c_B\mathbf{u}_B^t\mathbf{u}_1 = c_1.$$

In fact, for every $b = 1, 2, \ldots, B$, $c_b = \mathbf{x}^t\mathbf{u}_b$.

Definition of **Projections**. The projection of a vector $\mathbf{x}$ onto a vector $\mathbf{y}$ is given by

$$Proj\,(\mathbf{x}, \mathbf{y}) = \mathbf{x}^t \frac{\mathbf{y}}{\|\mathbf{y}\|}.$$

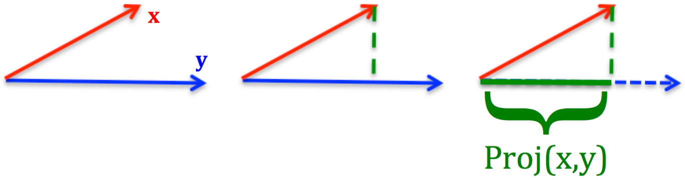

Figure A.3 Projection of a vector $\mathbf{x}$ onto a vector $\mathbf{y}$

A.1.4 Eigenanalysis and Symmetric Matrices

Eigenvalues, Eigenvectors, and Symmetric Matrices play a fundamental role in describing properties of spectral data obtained from imaging spectrometers. The definitions given here are not the most general but are sufficient for understanding and devising algorithms for imaging spectroscopy.

Definition of **Eigenvalues and Eigenvectors**. Let A be an $n \times n$ matrix. A number $\lambda \in \mathbb{R}$ and corresponding vector $\mathbf{v}_\lambda \in \mathbb{R}^n$ are called an *eigenvalue* and *eigenvector*, respectively, of A if $A\mathbf{v}_\lambda = \lambda\mathbf{v}_\lambda$. It is common to write $\mathbf{v}$ in place of $\mathbf{v}_\lambda$ if there is no ambiguity.

Definition of **Symmetric Matrix**. A square matrix A is called *symmetric* if $A = A^t$.

Definition of **Positive Definite and Semi-definite Matrices**. A square matrix A is called *positive definite* if, $\forall \mathbf{x} : \mathbf{x}^t A\mathbf{x} > 0$. A is *positive semi-definite* if $\forall \mathbf{x} : \mathbf{x}^t A\mathbf{x} \geq 0$. In these cases, if λ and $\mathbf{v}$ are an eigenvalue and corresponding eigenvector pair and $\mathbf{v} \neq \mathbf{0}$, then $\mathbf{v}^t A\mathbf{v} = \mathbf{v}^t \lambda \mathbf{v} = \lambda\|\mathbf{x}\|^2 > 0$. Since $\|\mathbf{x}\|^2 > 0$, it must be true that $\lambda > 0$. Thus, eigenvalues corresponding to nonzero eigenvectors of a positive definite matrix are positive. The same argument can be used for positive semi-definite matrices (in which case the eigenvalues are non-negative).

A.1.4.1 Orthonormal Basis Theorem If A is an $B \times B$ positive-definite, symmetric matrix , then the eigenvectors of A form an orthonormal basis for $\mathbb{R}^B$. This theorem is presented without proof; the interested reader can find it in many linear algebra texts.

Suppose $\mathcal{V} = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_B\}$ is an orthonormal basis of eigenvectors of A. Let $V = [\mathbf{v}_1|\mathbf{v}_2|\ldots|\mathbf{v}_B]$ be the matrix with columns equal to the elements of the basis $\mathcal{V}$. Then $AV = [\lambda_1\mathbf{v}_1|\lambda_1\mathbf{v}_2|\ldots|\lambda_B\mathbf{v}_B] = V\Lambda$ where $\Lambda = diag\,(\lambda_1, \lambda_2, \ldots, \lambda_B)$ is the diagonal $B \times B$ matrix with the eigenvalues along the diagonal and zeros elsewhere. Thus $A = V\Lambda V^t$. This relationship is called *diagonalization*. Often V is taken to be the matrix with *rows* equal to the transposes of the elements of $\mathcal{V}$. In this case, diagonalization is written as $A = V^t\Lambda V$.

Thus, a consequence of the Orthonormal Basis Theorem is that *any positive-definite, symmetric matrix A can be diagonalized by the matrix V whose columns are the orthonormal basis of eigenvectors of A and a diagonal matrix* Λ *with diagonal elements equal to the eigenvalues of A.*

Example. Suppose A is the matrix

$$\begin{bmatrix} 10 & 8 & 7 \\ 8 & 12 & 8 \\ 7 & 8 & 7 \end{bmatrix}$$

The reader can verify that the eigenvalues of A are $\lambda_1 = 25.2530, \lambda_2 = 2.9385, \lambda_3 = 0.8086$ and the corresponding eigenvectors are the columns of the matrix V given by

$$\begin{bmatrix} 0.5711 & 0.7565 & 0.3188 \\ 0.6485 & -0.6539 & 0.3897 \\ 0.5033 & -0.0158 & -0.8640 \end{bmatrix}$$

A.1.4.2 Singular Value Decomposition The Singular Value Decomposition (SVD) extends the concept of diagonalization to some non-square and non-invertible square matrices. There are very stable numerical algorithms for computing the SVD and is therefore often at the core of computational algorithms for tasks such as linear system solvers and computing inverses and pseudo-inverses of matrices.

Definition of **Singular Value Decomposition (SVD)**. Let A be an $M \times N$ with $M > N$. The *SVD* of A is given by $A = V \Sigma U^t$ where V is an $M \times M$ orthogonal matrix whose columns are eigenvectors of AA^t, U is an $M \times M$ orthogonal matrix of eigenvectors of $A^t A$, and Σ is an $M \times N$ matrix that is ”as diagonal as possible”, that is, Σ is of the form:

$$\Sigma = \begin{pmatrix} \sigma_1 & 0 & \dots & 0 & 0 \\ 0 & \sigma_2 & \dots & 0 & 0 \\ & & \ddots & & \\ 0 & 0 & \dots & \sigma_{N-1} & 0 \\ 0 & 0 & \dots & 0 & \sigma_N \\ 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & 0 & 0 \end{pmatrix}. \tag{A.1}$$

The first N rows and N columns of Σ consist of a diagonal matrix and the remaining $M - N$ rows are all zeros.

The pseudo-inverse of an $M \times N$ with $M > N$ matrix provides motivation for the definition, although the derivation is omitted here. Recall that the pseudo-inverse is given by $A^+ = \left(A^t A\right)^{-1} A^t$. It can be shown that $A^+ = U \Sigma^+ V^t$ where:

$$\Sigma^+ = \begin{pmatrix} \frac{1}{\sigma_1} & 0 & \dots & 0 & 0 & 0 \dots & 0 \\ 0 & \frac{1}{\sigma_2} & \dots & 0 & 0 & 0 \dots & 0 \\ & & \ddots & & & & \\ 0 & 0 & \dots & \frac{1}{\sigma_{N-1}} & 0 & 0 \dots & 0 \\ 0 & 0 & \dots & 0 & \frac{1}{\sigma_N} & 0 \dots & 0 \end{pmatrix}. \tag{A.2}$$

Note that Σ is $M \times N$ and Σ^+ is $N \times M$. Therefore, $\Sigma\Sigma^+$ is $M \times M$ and $\Sigma^+\Sigma$ is $N \times N$. Furthermore, $\Sigma^+\Sigma =$

$$\begin{pmatrix} \frac{1}{\sigma_1} & 0 & \dots & 0 & 0 & 0\dots & 0 \\ 0 & \frac{1}{\sigma_2} & \dots & 0 & 0 & 0\dots & 0 \\ & & \ddots & & & & \\ 0 & 0 & \dots & 0 & \frac{1}{\sigma_N} & 0\dots & 0 \end{pmatrix} \begin{pmatrix} \sigma_1 & 0 & \dots & 0 & 0 \\ 0 & \sigma_2 & \dots & 0 & 0 \\ & & \ddots & & \\ 0 & 0 & \dots & 0 & \sigma_N \\ 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & 0 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ & & \ddots & & \\ 0 & 0 & \dots & 0 & 1 \\ 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & 0 & 0 \end{pmatrix}$$

If $M < N$ and $rank\,(\mathbf{A})$ = M, then

$$\Sigma = \begin{pmatrix} \sigma_1 & 0 & \dots & 0 & 0 & 0\dots & 0 \\ 0 & \sigma_2 & \dots & 0 & 0 & 0\dots & 0 \\ & & \ddots & & & & \\ 0 & 0 & \dots & \sigma_{M-1} & 0 & 0\dots & 0 \\ 0 & 0 & \dots & 0 & \sigma_M & 0\dots & 0 \end{pmatrix}. \tag{A.3}$$

Calculating $(A^tA)^{-1}$ is often numerically unstable. The SVD is numerically stable.

Example. Suppose

$$A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{bmatrix}$$

Then the SVD of A is given by:

$$V = \begin{bmatrix} -0.2298 & 0.8835 & 0.4082 \\ -0.5247 & 0.2408 & -0.8165 \\ -0.8196 & -0.4019 & 0.4082 \end{bmatrix} \Sigma = \begin{bmatrix} 9.5255 & 0 \\ 0 & 0.5143 \\ 0 & 0 \end{bmatrix} U = \begin{bmatrix} -0.6196 & -0.7849 \\ -0.7849 & 0.6196 \end{bmatrix}$$

One can verify that $A^+ := (A^tA)^{-1}\,A^t$ = $U\Sigma^+V^t$.

A.1.4.3 Simultaneous Diagonalization Let C and C_n denote symmetric, positive semi-definite matrices. Then, there exists a matrix, W, that simultaneously diagonalizes C and Cn. The matrix W is constructed here. First, note that there exist matrices U and Λ such that

$$C = U^t\Lambda U.$$

Λ is of the form:

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 & \dots & 0 & 0 \\ 0 & \lambda_2 & \dots & 0 & 0 \\ & & \ddots & & \\ 0 & 0 & \cdots & \lambda_{B-1} & 0 \\ 0 & 0 & \cdots & 0 & \lambda_B \end{pmatrix}$$

where $\lambda_b \geq 0$. Hence, $\Lambda = \Lambda^{\frac{1}{2}}\Lambda^{\frac{1}{2}}$. Therefore

$$\Lambda^{-\frac{1}{2}}UCU^t\Lambda^{-\frac{1}{2}} = I.$$

Let

$$C_{n_t} = \Lambda^{-\frac{1}{2}}UC_nU^t\Lambda^{-\frac{1}{2}}.$$

Then $C_{n_t}^t = C_{n_t}$ and C_{n_t} is positive semi-definite. Hence, there exists V such that $V^tV = VV^t = I$ and

$$C_{n_t} = V^tD_{n_t}V \quad \text{or} \quad VC_{n_t}V^t = D_{n_t}$$

where $D_{n_t}$is a diagonal matrix with non-negative entries. Let $W = V\Lambda^{-\frac{1}{2}}U$. Then

$$WCW^t = V\Lambda^{-\frac{1}{2}}UCU^t\Lambda^{-\frac{1}{2}}V^t = VIV^t = VV^t = I$$

and

$$WC_nW^t = V\Lambda^{-\frac{1}{2}}UC_nU^t\Lambda^{-\frac{1}{2}}V^t = VC_{n_t}V^t = D_{n_t}.$$

Therefore, W simultaneously diagonalizes C and C_n.

A.1.5 Determinants, Ranks, and Numerical Issues

In this section, some quantities related to solutions of linear systems are defined and used to get a glimpse at some numerical issues that occur in Intelligent Systems all too frequently. A more thorough study requires a course in Numerical Linear Algebra. It is often assumed that solutions are estimated numerically and so there is error due to finite precision arithmetic.

The determinant is a number associated with a square matrix. The precise definition is recursive and requires quite a few subscripts so it is first motivated by the definition using a 2×2 matrix $\mathbf{A}$. The definition is then given and some important properties are discussed. Consider the problem of solving a linear system of two equations and two unknowns, $\mathbf{Ax} = \mathbf{b}$, using standard row reduction:

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} \quad \Rightarrow \quad \begin{bmatrix} -a_{21} & -\frac{a_{21}}{a_{11}} a_{12} \\ 0 & \frac{a_{11}a_{22} - a_{21}a_{12}}{a_{11}} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} -\frac{a_{21}b_1}{a_{11}} \\ -\frac{a_{21}b_1 + a_{11}b_2}{a_{11}} \end{bmatrix}$$

so, after a little more algebra, the solutions for x_1 and x_2 are found to be

$$x_2 = \frac{a_{11}b_2 - a_{21}b_1}{a_{11}a_{22} - a_{21}a_{12}} \quad \text{and} \quad x_1 = -\frac{a_{22}b_1 - a_{12}b_2}{a_{22}a_{11} - a_{12}a_{21}}$$

The denominator of the solutions are called the *determinant* of $\mathbf{A}$ which is often denoted by $det\,(\mathbf{A})$ or by $|\mathbf{A}|$. Notice that the numerators of the solutions for x_1 and x_2 are the determinants of the matrices

$$\mathbf{A}_1 = \begin{bmatrix} b_1 & a_{12} \\ b_2 & a_{22} \end{bmatrix} \text{ and } \mathbf{A}_2 = \begin{bmatrix} a_{11} & b_1 \\ a_{21} & b_2 \end{bmatrix}$$

Definition of **Rank of a matrix**. Let $\mathbf{A}$ denote an $M \times N$ matrix. It is a fact that the number of linearly independent rows of $\mathbf{A}$ is equal to the number of linearly independent columns of $\mathbf{A}$. The *Rank* of $\mathbf{A}$, often denoted by $rank\,(\mathbf{A})$, is the number of linearly independent rows (or columns). Note that, if $M > N$, then $rank\,(\mathbf{A}) \leq N$. It is almost always true that $rank\,(\mathbf{A}) = N$ but one must take care to make sure it is true. Similarly, if $\mathbf{A}$ is a square matrix of size $N \times N$, then it is almost always true that $rank\,(\mathbf{A}) = N$. In the latter case, $\mathbf{A}^{-1}$ exists and the matrix $\mathbf{A}$ is referred to as *Invertible* or *Nonsingular.* If $rank\,(\mathbf{A}) < N$, then $\mathbf{A}$ is non-invertible, also referred to as *Singular.*

Definition of **Null Space**. Assume $\mathbf{x}_1, \mathbf{x}_2 \neq \mathbf{0}$. Note that, if $\mathbf{Ax}_1$ =$\mathbf{Ax}_2$ = $\mathbf{0}$ then $\forall \alpha, \beta \in \mathbb{R}$, $\mathbf{A}\,(\alpha\mathbf{x}_1 + \beta\mathbf{x}_2)$ = $\alpha\mathbf{A}\,(\mathbf{x}_1) + \beta\mathbf{A}\,(\mathbf{x}_2)$= $\mathbf{0}$. Therefore, the set $\mathcal{N}_\mathbf{A}$ with the property that $\mathbf{n} \in \mathcal{N}_\mathbf{A} \Rightarrow \mathbf{An} = \mathbf{0}$ is a subspace of $\mathbb{R}^N$ called the *Null Space* of $\mathbf{A}$. Of course, $\mathbf{0} \in \mathcal{N}_\mathbf{A}$. In fact, $\mathbf{0}$ may be the only vector in $\mathcal{N}_\mathbf{A}$, particularly if $M > N$. However, if $M < N$, which is often the case for deep learning networks, then it is certain that there are nonzero vectors $\mathbf{x} \in \mathcal{N}_\mathbf{A} \subset \mathbb{R}^N$.

Example. Null Spaces and Singular Value Decomposition. If $M < N$ and $rank(A) = M$, then the SVD of $\mathbf{A}$ is of the form $\mathbf{A}$ = $\mathbf{V}\Sigma\mathbf{U}^t$ where Σ is defined in Eq. $A.3$. Since $\mathbf{V}$ and $\mathbf{U}$ are square, nonsingular matrices, $\mathcal{N}_\mathbf{V}$ = $\mathcal{N}_\mathbf{W}$ = $\{\mathbf{0}\}$. Therefore,

$$\mathcal{N}_{\mathbf{A}} = \left\{ \mathbf{x} | \mathbf{n} = \mathbf{U}^t \mathbf{x} \text{ and } \mathbf{n} = (0, 0, \ldots, 0, \mathbf{n}_{M+1}, \mathbf{n}_{M+2}, \ldots, \mathbf{n}_N)^t \right\}.$$

since some elements of the first M columns of Σ are nonzero and all the elements of the last $N - M + 1$ columns of Σ are zeros. If $M > N$ and $rank(A) = N$, then Σ is defined in Eq. $A.3$. Since there is a nonzero element in each column of Σ, $\mathcal{N}_{\mathbf{A}}$ = $\{\mathbf{0}\}$.

A.1.5.1 Condition Number Consider the following linear system:

$$\mathbf{A}\mathbf{x} = \mathbf{b} : \begin{bmatrix} -2 & 1 \\ -2.1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \tag{A.4}$$

Problems with this system are depicted in Fig. A.4

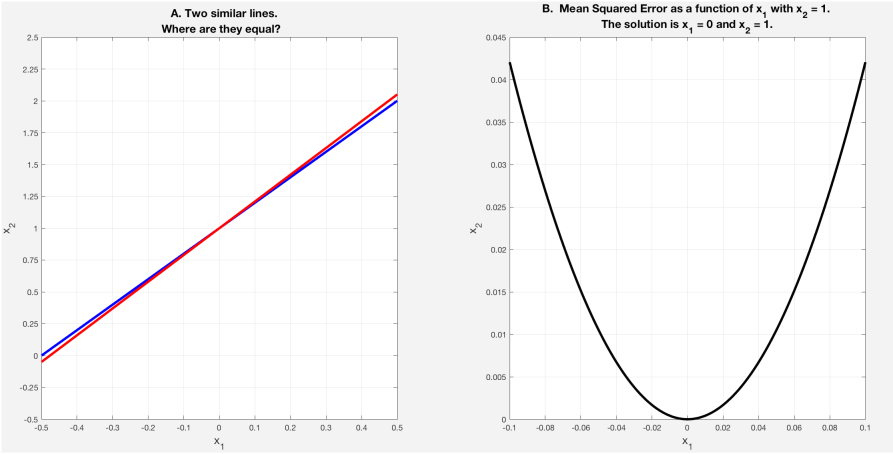

Figure A.4 A linear system associated with two lines (Red and Blue) that are almost parallel. The solution of the system is $x_1 = 0$ and $x_2 = 1$. Plot A depicts the lines. Plot B depicts the Mean Squared Error(MSE) of the system when x_1 ranges from -0.1 to 0.2 and x_2 is held constant at x_2=1. So, for example, if $|x_1| < 0.03$ then the MSE < 0.005. For some applications, 0.03 can be a very significant error.

Of course, We can make the MSE much smaller for a wider range of values of x_1 by letting a_{21} get close and close to 2, e.g. 2.01, 2.001, 2.0001, etc.

Error can be looked at relatively in addition to the MSE. For any square, nonsingular matrix $\mathbf{A}$ and linear system $\mathbf{A}\mathbf{x} = \mathbf{b}$ let $\mathbf{x}_t$ denote a solution of the system and $\mathbf{x}_f$ an estimated solution, which will have some error. Thus $\mathbf{A}\mathbf{x}_t = \mathbf{b}$ but $\mathbf{A}\mathbf{x}_f = \mathbf{b}_f$ and $\mathbf{b} \neq \mathbf{b}_f$. A couple of definitions are required. These will lead to the notion of the condition number of a matrix, which is a very useful quantity.

Definition of **Residual and Relative Residual for Linear Systems**. The *Residual* is $\mathbf{r} = \mathbf{b} - \mathbf{b}_f = \mathbf{b} - \mathbf{A}\mathbf{x}_f$ and the *Relative Residual* is $\frac{\|\mathbf{r}\|}{\|\mathbf{b}\|}$.

Definition of **Error and Relative Error for Linear Systems**. The *Error* is $\mathbf{e} = \mathbf{x}_t - \mathbf{x}_f$ and the *Relative Error* is $\frac{\|\mathbf{e}\|}{\|\mathbf{x}_t\|}$.

Definition of **Condition Number of a Matrix**. The *Condition Number* of a square matrix is $cond\,(\mathbf{A})$ = $\|\mathbf{A}\|\|\mathbf{A}^{-1}\|$. If $\mathbf{A}$ is not square, then bfA^{-1} is replaced by the pseudo-inverse $\mathbf{A}^{+}$. It is a fact that $cond\,(\mathbf{A}) \geq 1$.

The reader may be wondering how these definitions can be used since several assume the true solution is known but if the true solution is known, then there would be no need for an estimated solution. The reason is that bounds on these errors can be derived in terms of the condition number, which is a computable quantity. A brief overview is given here.

$$\frac{1}{cond\left(\mathbf{A}\right)}\frac{\|\mathbf{r}\|}{\|\mathbf{b}\|} \leq \frac{\|\mathbf{e}\|}{\|\mathbf{x}_t\|} \leq cond\left(\mathbf{A}\right)\frac{\|\mathbf{r}\|}{\|\mathbf{b}\|} \tag{A.5}$$

Eq. A.5 states that if $cond\left(\mathbf{A}\right)$ is small, then the relative error is also small compared to the relative residual. In the best case, if $cond\left(\mathbf{A}\right) = 1$, then the relative error and relative residual must be the same. On the other hand, if $cond\left(\mathbf{A}\right)$ is large, then the relative error can also be large. This is a situation one must guard against and, if one uses open-source software that is not documented and either compiled or difficult to read, then one runs the risk that the code does not check for large condition numbers. Although it is easy to visualize in two dimensions, it is harder to come up with ill-conditioned matrices. Intelligent systems often perform calculations with higher-dimensional data. Some examples of that are given.

Example. Here is a computer experiment you can do. Let $\mathbf{A}$ be a square 100×100 matrix of pseudo-random numbers generated by a normal, or Gaussian, distribution with mean 0 and standard deviation 1. The author did this 30 times and the condition numbers ranged from about 100 to about 20,000. Replace the 100^{th} row with the 99^{th} row + 10^{-8}. In the author's example, the condition number is now about 5×10^9. Take $\mathbf{b}$ = $\left(1, 2, 3, \ldots, 100\right)^t$ and solve the system $\mathbf{Ax} = \mathbf{b}$ using the matrix inverse (there are better ways). The relative residual in the author's case was about 2×10^{-7}. Therefore, the upper bound on the relative error is about $(5 \times 10^9)(2 \times 10^{-7}$ which is approximately 10^3, i.e. the relative error could be as high as 1000.

Example. There is a pattern classification technique called logistic regression. It is designed to distinguish one class of objects (Class 1) from another class of objects (Class 2), e.g. cats from dogs. There is a function $f\left(\mathbf{x}; \mathbf{w}\right)$ that takes as input a known vector calculated from an object from a known class, $\mathbf{x}$, and a vector of parameters to be estimated or learned, $\mathbf{w}$. The desired outcome is $f\left(\mathbf{x}; \mathbf{w}\right)$ = 1 if $\mathbf{x}$ was calculated from Class 1 and 0 for Class 2. This outcome is not generally achieved so the algorithm tries to achieve $f\left(\mathbf{x}; \mathbf{w}\right)$ = p where p is between 0 and 1 and p is large for Class 1 and low for Class 2. A method for attempting to find a solution is based on an iterative algorithm called *Iteratively Reweighted Least Squares*. The governing equations for the iteration have the form:

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \left(\mathbf{X}^t\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{X}^t\left(\mathbf{y} - \mathbf{p}\right)$$

It turns out that the matrix $\mathbf{X}^t\mathbf{W}\mathbf{X}$ can, and often does, have a very large condition number. The technique used to try to mitigate the effect of a very large condition number is called diagonal loading.

Definition of **Diagonal Loading**. If $\mathbf{A}$ is a square matrix and $\mathbf{I}$ is the identity matrix of the same size, then the calculation $\mathbf{B} = \mathbf{A} + \lambda\mathbf{I}$ where $\lambda \in \mathbb{R}$ is called *Diagonal Loading*. It is easy to see intuitively why diagonal loading can help with unstable numerical calculations because, if λ is much large than the other elements of $\mathbf{A}$, then $\mathbf{B}$ is almost diagonal so the condition number will be almost 1. Of course, $\mathbf{B}$ will be very different from $\mathbf{A}$ so there is a tradeoff between stability and accuracy. This is a difficult problem to solve and it is best to avoid algorithms that result in matrices with large condition numbers.