Notation:

Scalar: a

Vector: $\bar{v}$ (assumed to be a column unless otherwise stated)

Matrix: M

Tensors: $T = a \otimes b \otimes ... \otimes z$

Basic Operations:

1. Scalar Matrix Operations
   a. Addition
      Add the scalar to every element in the matrix. The sum of a matrix and a scalar is a matrix.

$$a = 3, M = \begin{bmatrix} 1 & 2 & 3 \\ 6 & 5 & 4 \\ 7 & 8 & 9 \end{bmatrix}$$

$$a + M = M + a = 3 + \begin{bmatrix} 1 & 2 & 3 \\ 6 & 5 & 4 \\ 7 & 8 & 9 \end{bmatrix} = \begin{bmatrix} 4 & 5 & 6 \\ 9 & 8 & 7 \\ 10 & 11 & 12 \end{bmatrix}$$

   b. Multiplication
      Multiply every element of the matrix by the scalar. The product of a matrix and a scalar is a matrix.

$$a = 3, M = \begin{bmatrix} 1 & 2 & 3 \\ 6 & 5 & 4 \\ 7 & 8 & 9 \end{bmatrix}$$

$$aM = Ma = 3 * \begin{bmatrix} 1 & 2 & 3 \\ 6 & 5 & 4 \\ 7 & 8 & 9 \end{bmatrix} = \begin{bmatrix} 3 & 6 & 9 \\ 18 & 15 & 12 \\ 21 & 24 & 27 \end{bmatrix}$$

2. Matrix Vector Addition
   Add the elements of the vector to each element in the corresponding row of the matrix. The sum of a vector and a matrix is a matrix.

$$\vec{v} = \begin{bmatrix} 2 \\ 4 \\ -5 \end{bmatrix}, \quad M = \begin{bmatrix} 1 & 2 & 3 \\ 6 & 5 & 4 \\ 7 & 8 & 9 \end{bmatrix}$$

$$\vec{v} + M = M + \vec{v} = \begin{bmatrix} 1 & 2 & 3 \\ 6 & 5 & 4 \\ 7 & 8 & 9 \end{bmatrix} + \begin{bmatrix} 2 \\ 4 \\ -5 \end{bmatrix} = \begin{bmatrix} 3 & 4 & 5 \\ 10 & 9 & 8 \\ 2 & 3 & 5 \end{bmatrix}$$

3. Matrix-Matrix Addition
   Add the corresponding elements in the matrices together. The matrices must be the same size. The sum of two matrices is a matrix of the same size.

$$A = \begin{bmatrix} 3 & -4 & 5 \\ 10 & 9 & -8 \\ -2 & 3 & 5 \end{bmatrix}, \quad B = \begin{bmatrix} 1 & 2 & 3 \\ 6 & 5 & 4 \\ 7 & 8 & 9 \end{bmatrix}$$

$$A + B = B + A = \begin{bmatrix} 1 & 2 & 3 \\ 6 & 5 & 4 \\ 7 & 8 & 9 \end{bmatrix} + \begin{bmatrix} 3 & -4 & 5 \\ 10 & 9 & -8 \\ -2 & 3 & 5 \end{bmatrix} = \begin{bmatrix} 4 & -2 & 8 \\ 16 & 14 & -4 \\ 5 & 22 & 14 \end{bmatrix}$$

4. Vector-Vector Multiplication
   The product of two vectors of the same length is either a scalar or a matrix, depending on how the vectors are multiplied.

$$\vec{v} = \begin{bmatrix} 2 \\ (-3) \\ 7 \end{bmatrix}, \quad \vec{w} = \begin{bmatrix} 9 \\ 8 \\ 1 \end{bmatrix}$$

$$\vec{v}^T \vec{w} = \begin{bmatrix} 2 & (-3) & 7 \end{bmatrix} \begin{bmatrix} 9 \\ 8 \\ 1 \end{bmatrix} = 18 - 24 + 7 = 1$$

$$\vec{v}\vec{w}^T = \begin{bmatrix} 2 \\ (-3) \\ 7 \end{bmatrix} \begin{bmatrix} 9 & 8 & 1 \end{bmatrix} = \begin{bmatrix} 18 & 16 & 2 \\ (-27) & (-24) & (-6) \\ 63 & 56 & 7 \end{bmatrix}$$

   The first product, $\vec{v}^T \vec{w}$, is called the dot product. It can also be written as

$$\vec{v}^T \vec{w} = ||\vec{v}|| * ||\vec{w}|| cos(\theta)$$

   If the two vectors are orthogonal, their dot product is 0.

5. Transpose of a Matrix
   The transpose of a matrix is found by flipping the matrix over its main diagonal. The main diagonal is the diagonal that begins at the element located at the first row and first column of the matrix. It extends through the second row and second column element all the way to the mth row and mth column element.

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 9 & 8 & 7 \end{bmatrix}, \quad A^T = \begin{bmatrix} 1 & 4 & 9 \\ 2 & 5 & 8 \\ 3 & 6 & 7 \end{bmatrix}$$

   The matrix does not have to be a square matrix to have a transpose. The main diagonal exists for all matrices.

$$B = \begin{bmatrix} 1 & 2 \\ 4 & 5 \\ 9 & 8 \end{bmatrix}, \quad B^T = \begin{bmatrix} 1 & 4 & 9 \\ 2 & 5 & 8 \end{bmatrix}$$

6. Matrix Matrix Multiplication
   The inner dimensions of the two matrices must be the same. The product matrix will have the same number of rows as the first matrix and the same number of columns as the second matrix.

$$A = \begin{bmatrix} 3 & -4 \\ 10 & 9 \\ -2 & 3 \end{bmatrix}, \quad B = \begin{bmatrix} 1 & 2 & 3 \\ 6 & 5 & 4 \end{bmatrix}$$

When performing matrix multiplication, take the sum of the products of the elements in the row of the first matrix and the column of the second matrix. The row in the first matrix and the column in the second matrix should correspond to the desired row and column in the resulting matrix.

$$AB = \begin{bmatrix} 3 & -4 \\ 10 & 9 \\ -2 & 3 \end{bmatrix} * \begin{bmatrix} 1 & 2 & 3 \\ 6 & 5 & 4 \end{bmatrix} = \begin{bmatrix} 3+(-24) & 6+(-20) & 9+(-16) \\ 10+54 & 20+45 & 30+36 \\ (-2)+18 & (-4)+15 & (-6)+12 \end{bmatrix}$$

$$= \begin{bmatrix} (-21) & (-14) & (-7) \\ 64 & 65 & 66 \\ 16 & 11 & 6 \end{bmatrix}$$

$$BA = \begin{bmatrix} 1 & 2 & 3 \\ 6 & 5 & 4 \end{bmatrix} * \begin{bmatrix} 3 & -4 \\ 10 & 9 \\ -2 & 3 \end{bmatrix} = \begin{bmatrix} 3+20+(-6) & (-4)+18+9 \\ 18+50+(-8) & (-24)+45+12 \end{bmatrix} = \begin{bmatrix} 17 & 23 \\ 60 & 33 \end{bmatrix}$$

Note that the product of A and B is not equal to the product of B and A. Why is this?

Properties of Matrix Multiplication

- Distributive

  $A(B+C) = AB + AC$
- Associative

  $A(BC) = (AB)C$
- Not Commutative (though a dot product is: $\overline{x}^T\overline{y} = \overline{y}^T\overline{x}$)

7. Matrix Vector Multiplication

   Multiplication of a matrix and a vector is similar to matrix-matrix multiplication. The inner dimensions of the matrix and the vector must match.

$$A = \begin{bmatrix} 3 & -4 \\ 10 & 9 \\ -2 & 3 \end{bmatrix}, \quad \vec{v_1} = \begin{bmatrix} 1 & 2 & 3 \end{bmatrix}, \quad \vec{v_2} = \begin{bmatrix} -1 \\ 4 \end{bmatrix}$$

Generally, a matrix can be multiplied by a vector if

- The vector is a row vector and is located on the left of the matrix

$$\vec{v_1}A = \begin{bmatrix} 1 & 2 & 3 \end{bmatrix} \begin{bmatrix} 3 & -4 \\ 10 & 9 \\ -2 & 3 \end{bmatrix} = \begin{bmatrix} 3+20+(-6) & (-4)+18+9 \end{bmatrix} = \begin{bmatrix} 17 & 23 \end{bmatrix}$$

- The vector is a column vector and is located on the right of the matrix

$$A\vec{v_2} = \begin{bmatrix} 3 & -4 \\ 10 & 9 \\ -2 & 3 \end{bmatrix} \begin{bmatrix} -1 \\ 4 \end{bmatrix} = \begin{bmatrix} (-3) + (-16) \\ (-10) + 36 \\ 2 + 12 \end{bmatrix} = \begin{bmatrix} -19 \\ 26 \\ 14 \end{bmatrix}$$

8. Determinant of a Matrix
   The determinant of a matrix maps matrices to real scalars. It is the product of all of the eigenvalues (see Eigenvectors and Eigenvalues section) of a matrix:

   $$det(A) = \Pi \lambda_i$$

   By hand,

   $$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}, \quad det(A) = ad - bc$$

   $$B = \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix}, \quad det(B) = a(ei - fh) - b(di - g) + c(dh - eg)$$

   The determinant is a measure of how much the space expands or contracts when multiplied by the matrix.

9. Inverse of a Matrix
   The product of a matrix and its inverse is the identity matrix.

   $$AA^{-1} = A^{-1}A = I$$

   The inverse of a matrix can be found by hand by augmenting the matrix with an identity matrix and using elementary row operations (additions or subtraction, multiplication by a constant, or swapping rows:

$$A = \begin{bmatrix} 3 & 0 & 2 \\ 2 & 0 & (-2) \\ 0 & 1 & 1 \end{bmatrix}$$

$$[A|I] = \left[\begin{array}{ccc|ccc} 3 & 0 & 2 & 1 & 0 & 0 \\ 2 & 0 & (-2) & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 \end{array}\right] = \left[\begin{array}{ccc|ccc} 5 & 0 & 0 & 1 & 1 & 0 \\ 2 & 0 & (-2) & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 \end{array}\right]$$

$$= \left[\begin{array}{ccc|ccc} 1 & 0 & 0 & 0.2 & 0.2 & 0 \\ 2 & 0 & (-2) & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 \end{array}\right] = \left[\begin{array}{ccc|ccc} 1 & 0 & 0 & 0.2 & 0.2 & 0 \\ 0 & 0 & (-2) & (-0.4) & 0.6 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 \end{array}\right]$$

$$= \left[\begin{array}{ccc|ccc} 1 & 0 & 0 & 0.2 & 0.2 & 0 \\ 0 & 0 & 1 & 0.2 & (-0.3) & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 \end{array}\right] = \left[\begin{array}{ccc|ccc} 1 & 0 & 0 & 0.2 & 0.2 & 0 \\ 0 & 1 & 0 & (-0.2) & 0.3 & 1 \\ 0 & 0 & 1 & 0.2 & (-0.3) & 0 \end{array}\right]$$

$$A^{-1} = \begin{bmatrix} 0.2 & 0.2 & 0 \\ (-0.2) & 0.3 & 1 \\ 0.2 & (-0.3) & 0 \end{bmatrix}$$

The inverse of a matrix is most commonly found using a program such as MATLAB/Octave.

In order to find the inverse of a matrix, the matrix must be a square matrix and its columns must be linearly independent.

The product of a matrix and its inverse is the identity matrix: $A^{-1}A = AA^{-1} = I$.

If a square matrix has mutually orthonormal rows and columns, $A^{-1} = A^T$

10. Trace of a Matrix

The trace of a matrix gives the sum of all diagonal entries of a matrix:

$$Tr(A) = \sum_i A_{i,i}$$

A scalar is its own trace: Tr(a) = a

The trace is invariant to the transpose operator. (Why?)

The trace of a square matrix composed of multiple factors is invariant to cyclic permutations of those factors (as long as the shapes of the factors allow it):

$$A \in \mathbb{R}^{m*n}, B \in \mathbb{R}^{n*p}, C \in \mathbb{R}^{p*m}$$

$$Tr(ABC) = Tr(BCA) = Tr(CAB)$$

The trace operator allows us to rewrite some operations without using summations. For example, the Frobenius norm (Norm section) becomes:

$$||A||_F = \sqrt{Tr(AA^T)}$$

Key Ideas:

- Linearly independent: no vector in a set of vectors is a linear combination of any other vectors in the set
- Rank: maximum number of linearly independent rows or columns in a matrix
- Span: the span of a set of vectors in a vector space is the intersection of all subspaces containing that set (ie, a collection of all linear combinations of the set of vectors)
- Range: span of the columns of a matrix, also known as the column space. If the columns of the matrix are linearly independent, the range is a basis of the matrix.
- Basis: a set of linearly independent vectors such that every vector in this vector space is a linear combination of this set
- Null space: also known as the kernel of a matrix. The null space of a matrix is the span of all vectors satisfying the equation $A\bar{x} = \bar{0}$. The columns of A are independent if the only solution to the previous equation is the 0 vector.

Norms

Often used in regularization of a regression.
A norm measures the size of a vector. It is a map from a vector to a non-negative scalar.

Properties of a norm:

1. $f(\overline{x}) = 0 \Rightarrow \overline{x} = 0$
2. $f(\overline{x}+\overline{y}) \le f(\overline{x})+f(\overline{y})$ (Triangle Inequality)
3. $\forall \alpha \in \mathbb{R}, f(\alpha\overline{x}) = |\alpha| f(\overline{x})$

Common Norms

- Euclidean norm (ℓ_2 norm)

$$||\vec{x}||_2 = \sqrt{\sum_i |x_i|^2}$$

- ℓ_2 - squared norm: useful when you want a derivative with respect to x to depend on the individual elements of x and not the entire vector

$$||\vec{x}||_2^2 = \sum_i |x_i|^2 = \vec{x}^T \vec{x}$$

- ℓ_1 norm: grows at the same rate everywhere, used when the difference between 0 and small, non-zero values is important

$$||\vec{x}||_1 = \sum_i |x_i|$$

- ℓ_∞ norm: also known as the max norm

$$||\vec{x}||_\infty = \max_i |x_i|$$

- Frobenius norm: intuitively similar to the ℓ_2 norm of a matrix

$$||A||_F = \sqrt{\sum_{i,j} A_{i,j}^2}$$

Note: a vector with a norm of 1 is called a unit vector.

Representing a system of equations

We can write a system of equations...

$$x_1 + x_2 + 2x_3 = 5$$

$$x_1 - x_2 = -2$$

$$x_2 + x_3 = 4$$

in terms of a matrix and a pair of vectors:

$$A = \begin{bmatrix} 1 & 1 & 2 \\ 1 & (-1) & 0 \\ 0 & 1 & 1 \end{bmatrix}, \quad \vec{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}, \quad \vec{b} = \begin{bmatrix} 5 \\ (-2) \\ 4 \end{bmatrix}$$

Performing the multiplication of $A\vec{x} = \vec{b}$, we get

$$A\vec{x} = \vec{b} \Rightarrow \begin{bmatrix} 1 & 1 & 2 \\ 1 & (-1) & 0 \\ 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 5 \\ (-2) \\ 4 \end{bmatrix}$$

$$= \begin{bmatrix} x_1 + x_2 + 2x_3 \\ x_1 - x_2 + 0x_3 \\ 0x_1 + x_2 + x_3 \end{bmatrix} = \begin{bmatrix} 5 \\ (-2) \\ 4 \end{bmatrix}$$

$$\begin{array}{c} x_1 + x_2 + 2x_3 = 5 \\ \Rightarrow \quad x_1 - x_2 = -2 \\ x_2 + x_3 = 4 \end{array}$$

To solve for $\bar{x}$, we can use the inverse of A (provided that A meets the requirements listed previously and that A has the same number of rows and columns as $\bar{b}$ has elements).

$$A\vec{x} = \vec{b}$$

$$A^{-1}A\vec{x} = A^{-1}\vec{b}$$

$$A^{-1}A = I, \qquad \vec{x} = A^{-1}\vec{b}$$

The textbook (http://www.deeplearningbook.org/contents/linear_algebra.html) states that determine whether the equation $A\vec{x} = \vec{b}$ has a solution is as simple as testing whether $\bar{b}$ is in the range of A. (see page 38)

Eigenvectors and Eigenvalues

Decomposition allows us to see functional properties of a matrix

The eigenvector of a square matrix is a nonzero vector such that, when multiplying the matrix by the eigenvector, only the scale of the eigenvector changes.

$$A\vec{v} = \lambda\vec{v}$$

The scale by which the eigenvector changes is the eigenvalue (λ).

Looking at the eigenvalues, we can learn a few facts about the original matrix:

- If there is an eigenvalue with the value of 0, the matrix is singular.
- If all eigenvalues are greater than or equal to 0, the matrix is positive semi-definite.
- If all eigenvalues are greater than 0, the matrix is positive definite.
- If all eigenvalues are less than 0, the matrix is negative definite.
- If all eigenvalues are less than or equal to 0, the matrix is negative semi-definite.

These properties are often leverages in machine learning techniques.

Note: any scaled eigenvector is also an eigenvector, so we usually limit the ourselves to unit eigenvectors.

The process of finding the eigenvectors and values is called eigendecomposition. We can start by rewriting the above equation as

$$A\vec{v} - \lambda I\vec{v} = 0$$

We choose to multiply the eigenvalues by the identity matrix so that we can factor out the vector:

$$A\vec{v} - \lambda I\vec{v} = 0 \Rightarrow (A - \lambda I)\vec{v} = 0$$

From here, we can look at the rows of the difference and use linear combinations to solve the augmented matrix $[A - \lambda I \mid \overline{0}]$ (not demonstrated here). For a more computer-friendly version of finding the eigenvalues (if a method for finding the eigenvalues was not built into the programming language),

$$det(A - \lambda I) = 0$$

(For an explanation of why we can use the determinant here, please see http://www.onmyphd.com/?p=eigen.decomposition.)

We can write a matrix in terms of its eigendecomposition:

$$A = V diag(\lambda) V^{-1} = V \Lambda V^{-1}$$

where V is a matrix of concatenated eigenvectors and Λ is a matrix containing only the eigenvalues on its diagonal.

For any matrix, not necessarily a square one:

$$A = Q \Lambda Q^T$$

where Q is an orthogonal matrix of the eigenvectors and Λ is a matrix containing only the eigenvalues on its diagonal.

Note: if any two eigenvectors have the same eigenvalue, then any set of orthogonal vectors in their span are also eigenvectors with that eigenvalue.

Singular Value Decomposition (SVD)

The SVD is a more general decomposition method than eigendecomposition. It results in a factorization of the form

$$A = UDV^T$$

where

$A \in \mathbb{R}^{m*n}$ is the original matrix

$U \in \mathbb{R}^{m*m}$ is the matrix of orthogonal left-singular vectors of A

$D \in \mathbb{R}^{m*n}$ is the diagonal matrix of singular values of A.

$V \in \mathbb{R}^{n*n}$ is the matrix of orthogonal right-singular vectors of A

Note: non-zero values of D are $\sqrt{eigenvalues(AA^T)}$

SVD is useful for generalizing matrix inversion to non-square matrices.

How are Eigendecomposition and Singular Value Decomposition different?

Recall: the eigendecomposition is $A = V diag(\lambda) V^{-1} = V \Lambda V^{-1}$ and the singular value decomposition is $A = UDV^T$

High-Level:

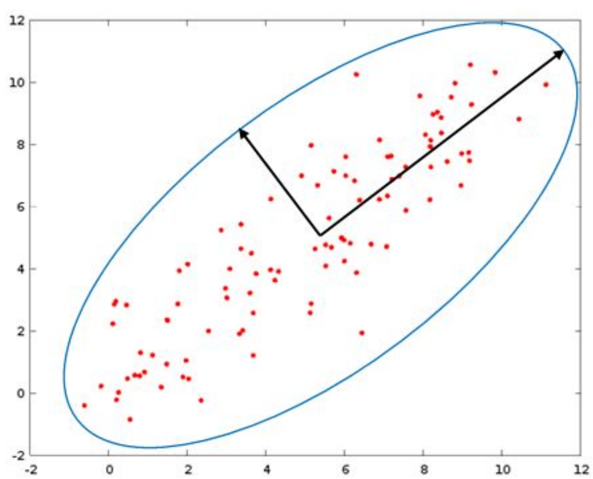
- Every real matrix has a singular value decomposition, but not necessarily an eigendecomposition.
- The matrices surrounding the matrix of eigenvalues are related, but the matrices surrounding the matrix of singular values are not necessarily related.
- The singular values of a matrix are all real and non-negative. The eigenvalues of a matrix can be any complex number.

Also, remember the singular values are the square roots of the eigenvalues of a matrix multiplied by its transpose: $\sqrt{eigenvalues(AA^T)}$

For further comparison, please see http://math.stackexchange.com/questions/320220/intuitively-what-is-the-difference-between-eigendecomposition-and-singular-valu

Principal Components Analysis (PCA)

Principal component analysis identifies orthogonal vectors that explain the largest amount of variance in the data. Intuitively, this process amounts to fitting an ellipsoid around the data and finding the axes of the ellipse:



The first component (major axis in the above image) explains the largest amount of variability in the data. The second component (minor axis) explains the next largest amount of variability possible with the constraint that it is orthogonal to the first.

The principal components come from eigendecomposition or SVD. The textbook shows how SVD can be applied to PCA in Section 2.12.

PCA is mostly used for exploratory data analysis and making predictive models.