ECE-4973: Computer Vision
Samuel Cheng

1 Some math background

1.1 Linear algebra

1.1.1 Vector

- A vector is an array of numbers. An n-dimensional vector contains n elements. n is also sometimes known as the size of the vector
- A vector is usually denoted by a bold lower case letter
- Only two vectors of same size can be added or subtracted from one another. The sum of two vectors is simply equal to the vector of the sum of their elements. For example, $\mathbf{u} = \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix}$ and $\mathbf{v} = \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix}$, then $\mathbf{u} + \mathbf{v} = \begin{pmatrix} u_1 + v_1 \\ u_2 + v_2 \\ u_3 + v_3 \end{pmatrix}$
- If a is scalar, then $a \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} = \begin{pmatrix} au_1 \\ au_2 \\ au_3 \end{pmatrix}$
- It is easy to verify that scalar product and vector addition satisfy distributive law. That is, $a(\mathbf{u} + \mathbf{v}) = a\mathbf{u} + a\mathbf{v}$
- $\mathbf{u} = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}$ is a column vector and $\mathbf{v} = \begin{pmatrix} v_1, v_2 \end{pmatrix}$ is a row vector. By convention, it is common to assume all vectors are column vectors unless it is specified

Definition 1.1 (Vector space)**.** An n-dimensional vector space consists of all length-n vectors.

Definition 1.2 (Linear independence)**.** A set of vectors $\mathbf{v}_1, \cdots, \mathbf{v}_n$ are (linearly) independent if for any scalar $a_1, \cdots, a_n$, $a_1\mathbf{v}_1 + a_2\mathbf{v}_2 + \cdots + a_n\mathbf{v}_n = 0$ if and only if $a_1 = a_2 = \cdots = a_n$.

- If $\mathbf{v}_1, \cdots, \mathbf{v}_n$ are not linearly independent, they are linearly independent.

Definition 1.3 (Basis)**.** A set of vector forms a basis of the length-n vector space if any vector in the vector space can be represented as a linear combination of the vectors.

Definition 1.4 (Inner product)**.** The inner product or dot product of $\boldsymbol{u}$ and $\boldsymbol{v}$, denoted by $\langle \boldsymbol{u}, \boldsymbol{v} \rangle$, is the sum of product of their elements. That is, $\sum_i \boldsymbol{u}_i \boldsymbol{v}_i$

Definition 1.5 (Orthogonal)**.** $\boldsymbol{u}$ and $\boldsymbol{v}$ are orthogonal, denoted by $\boldsymbol{u} \perp \boldsymbol{v}$, if $\langle \boldsymbol{u}, \boldsymbol{v} \rangle = 0$

- A matrix is a table of numbers. An $m \times n$ matrix has m rows and n columns. We often denote a matrix with a upper case letter. For example, $A = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix}$ is a matrix of 2 rows and 3 columns
- The dimensions of a matrix are the number of rows and the number of columns. For example, A is a 2×3 matrix
- We can also view a matrix as a column "vector" of row vectors. For example, $A = \begin{pmatrix} \begin{pmatrix} 1 & 2 & 3 \end{pmatrix} \\ \begin{pmatrix} 4 & 5 & 6 \end{pmatrix} \end{pmatrix}$. We can also view the matrix as a row "vector of column vectors. That is, $A = \left(\begin{pmatrix} 1 \\ 4 \end{pmatrix} \begin{pmatrix} 2 \\ 5 \end{pmatrix} \begin{pmatrix} 3 \\ 6 \end{pmatrix} \right)$.
- A **square** matrix is a matrix with the same number of rows and columns ($m = n$)

- A **transpose** of an $n \times m$ matrix C, denoted as $C^\top$, is an $m \times n$ matrix with the (i, j) element as the (j, i) element of C. For example, $A^\top = \begin{pmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{pmatrix}$
  - Note that $(A^\top)^\top = A$

Definition 1.6 (Rank)**.** The column rank of a matrix is the maximum number of arbitrary columns that are independent. And the row rank of a matrix is the maximum number of arbitrary rows that are independent. One can show that the row rank and the column rank of a matrix are the same.

Definition 1.7 (Full rank)**.** When the rank of the matrix equals to the minimum of the number of columns and the number of rows, we said the matrix is full rank.

Example 1.1. $A = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 2 & 0 \\ 1 & 3 & 0 \end{pmatrix}$ has rank 2. The column rank is obviously 2 as $\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$ and $\begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}$ are linearly independent. Note that an all-zero vector (column 3) linearly depends on any set of vectors since $a_1\mathbf{v}_1 + a_2\mathbf{v}_2 + \cdots + a_n\mathbf{v}_n + a_0 0 = 0$ for $a_0 = 1$ and $a_1 = a_2 = \cdots = a_n = 0$. So not all coefficients are zeros.

The row rank is also 2 as $(1, 1)$ and $(1, 2)$ are obviously linearly independent. But $(1, 1), (1, 2)$, and $(1, 3)$ are not linearly independent. We have $(1, 2) - (1, 1) = (1, 3) - (1, 2) \Rightarrow (1, 1) - 2(1, 2) + (1, 3) = 0$ for non-zero coefficients 1,-2, and 1.

- The conjugate transpose of a matrix is known as the **Hermitian** of a matrix. Usually denoted as $A^\dagger$
- A matrix that satisfies $A^\top = A$ is **symmetric**
- A matrix that satisfies $A^\dagger = A$ is **Hermitian**
  - Note that ***real symmetric matrix is Hermitian***
- A matrix is **unitary** if $A^\dagger A = I$
- A *real* matrix is **orthogonal** if $A^\top A = I$

1.1.2 Matrix multiplication

Definition 1.8 (Matrix multiplication)**.** Consider an $n \times r$ matrix A and an $r \times m$ matrix B, the product AB is a $n \times m$ matrix such that the (i, j) element of AB is

$$(AB)_{i,j} = \sum_{k=1}^{r} A_{i,k} B_{k,j}$$

- Note that the number of columns of A and the number of rows of B have to be the same, otherwise, the product AB is invalid. And apparently, $AB \neq BA$!
- Associative law:$(AB)C = A(BC)$
- Distributive law: $(A + B)C = AC + BC$
- Vectors a matrices too. Consider two length-n column vectors $\mathbf{u}$ and $\mathbf{v}$. $\mathbf{u}^\top\mathbf{v} = \sum_{k=1}^{n} u_k v_k$ is known as the dot product or inner product of $\mathbf{u}$ and $\mathbf{v}$
- $(AB)^\top = B^\top A^\top$
- Let $\mathbf{0}$ be a matrix with all elements equal to 0. For any matrix A that $A\mathbf{0} = \mathbf{0}$
  - Note that $AB = \mathbf{0}$ and $A \neq \mathbf{0}$ does not imply $B = 0$
- A square matrix with all diagonal elements equal to 1 and all other elements equal to 0 is call an identity matrix. Usually denoted by I (or I_n). For any matrix A with the number of columns equal to n, $AI = A$. And for any matrix B with the number of rows equal to n, $IB = B$.

- A square matrix B is the inverse matrix of a square A if $BA = I$. We denote the inverse as A^{-1}
  - $(AB)^{-1} = B^{-1}A^{-1}$
  - If $AB = I$ and $DA = I$, $D = D(AB) = (DA)B = B$

Definition 1.9 (Invertible matrix)**.** A matrix is invertible if its inverse exists.

Lemma 1.1 (Invertible)**.** *Note that a matrix is invertible if and only if it is full rank.*

Proof. Consider the problem $A\mathbf{x} = \mathbf{0}$. If A is invertible, then $A^{-1}(A\mathbf{x}) = A^{-1}\mathbf{0} \Rightarrow \mathbf{x} = \mathbf{0}$. So all columns have to be linearly independent, so A is full-rank.

Consider the matrix A and apply the following row operations on the matrix.

- scaling
- exchange one row with another row
- adding a weighed row to another row

The above operations are the equivalent of multiplying A with elementary matrices of forms. If A is full-rank, it is always possible to convert the matrix to I with the row operations. Thus an inverse exists, and that is just equal to the product the elementary matrices. □

1.1.3 Trace and determinant

Definition 1.10 (Trace)**.** The trace of a square matrix is the sum of its diagonal. That is, $tr(A) = \sum_{i=1}^{n} A_{i,i}$ for an $n \times n$ matrix A.

- $tr(AB) = tr(BA)$

Definition 1.11 (Determinant)**.** The determinant of an $n \times n$ matrix A, denoted by $det(A)$, is defined as

$$\sum_{\sigma \in P_n} \epsilon_\sigma A_{1,\sigma(1)} A_{2,\sigma(2)} \cdots A_{n,\sigma(n)},$$

where P_n is the set of all permutations σ and $\epsilon_\sigma = -1^{N(\sigma)}$ (with $N(\sigma)$ being the number of inversions) is the parity of the permutation

Example 1.2 (Determinant of 3×3 matrix)**.** For a matrix $A = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}$,

$$\begin{aligned} \det(A) &= \sum_{\sigma \in P_3} \epsilon_\sigma a_{1,\sigma(1)} a_{2,\sigma(2)} a_{3,\sigma(3)} \\ &= a_{1,1}a_{2,2}a_{3,3} - a_{1,1}a_{2,3}a_{3,2} - a_{1,2}a_{2,1}a_{3,3} + a_{1,2}a_{2,3}a_{3,1} + a_{1,3}a_{2,1}a_{3,2} - a_{1,3}a_{2,2}a_{3,1} \end{aligned}$$

- $\det(A^\top) = \det(A)$
- $\det([\lambda \mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_n]) = \lambda \det([\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_n])$
- $\det([\mathbf{b} + \mathbf{c}, \mathbf{a}_2, \cdots, \mathbf{a}_n]) = \det([\mathbf{b}, \mathbf{a}_2, \cdots, \mathbf{a}_n]) + \det([\mathbf{c}, \mathbf{a}_2, \cdots, \mathbf{a}_n])$
- $\det([\mathbf{a}_2, \mathbf{a}_1, \cdots, \mathbf{a}_n]) = -\det([\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_n])$
- $\det([\mathbf{a}, \mathbf{a}, \cdots, \mathbf{a}_n]) = 0$
- $\det(AB) = \det(A)\det(B)$

Proof.

$$\begin{aligned}
\det(AB) &= \det([b_{11}\mathbf{a}_1 + \cdots + b_{n1}\mathbf{a}_n, b_{12}\mathbf{a}_1 + \cdots + b_{n2}\mathbf{a}_n, \cdots, b_{1n}\mathbf{a}_1 + \cdots + b_{nn}\mathbf{a}_n]) \\
&= \sum_{\sigma \in P_n} \det([b_{\sigma(1),1}\mathbf{a}_{\sigma(1)}, b_{\sigma(2),2}\mathbf{a}_{\sigma(2)}, \cdots, b_{\sigma(n),n}\mathbf{a}_{\sigma(n)}]) \\
&= \sum_{\sigma \in P_n} b_{\sigma(1),1} b_{\sigma(2),2} \cdots b_{\sigma(n),n} \det([\mathbf{a}_{\sigma(1)}, \mathbf{a}_{\sigma(2)}, \cdots, \mathbf{a}_{\sigma(n)}]) \\
&= \sum_{\sigma \in P_n} b_{\sigma(1),1} b_{\sigma(2),2} \cdots b_{\sigma(n),n} \epsilon_\sigma \det([\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_n]) \\
&= \det(B)\det(A)
\end{aligned}$$

□

- $\det(A^{-1}) = \det(A)^{-1}$
- Sylvester's determinant theorem: $\det(I + AB) = \det(I + BA)$

1.1.4 Eigenvalues and eigenvectors

Definition 1.12 (Eigenvector and eigenvalue)**.** For a square matrix A, λ and $\mathbf{v}$ are an eigenvalue and an eigenvector of A, respectively, if $A\mathbf{v} = \lambda\mathbf{v}$.

- If $\boldsymbol{v}$ is an eigenvector, so thus $a\boldsymbol{v}$.
- Consider all linearly independent eigenvector $v_1, \cdots, v_n$ and their corresponding eigenvalues $\lambda_1, \cdots, \lambda_n$. Note that

$$A\underbrace{[v_1, v_2, \cdots, v_n]}_{V} = [\lambda_1 v_1, \lambda_2 v_2, \cdots, \lambda_n v_n] = \underbrace{[v_1, v_2, \cdots, v_n]}_{V} \underbrace{\begin{pmatrix} \lambda_1 & 0 & 0 & \cdots & 0 \\ 0 & \lambda_2 & 0 & \cdots & 0 \\ 0 & 0 & \ddots & & \vdots \\ \vdots & \vdots & & & \lambda_n \end{pmatrix}}_{\Lambda}$$

  If V is full-rank, then we can diagonalize A as $\Lambda = V^{-1}AV$.
- Note that $\det(A) = \det(V)\det(\Lambda)\det(V^{-1}) = \det(V)\det(\Lambda)\det(V)^{-1} = \det(\Lambda) = \prod_{i=1}^n \lambda_i$
- Similarly, $tr(A) = tr(V\Lambda V^{-1}) = tr(V(\Lambda V^{-1})) = tr(\Lambda V^{-1}V) = tr(\Lambda) = \sum_{i=1}^n \lambda_i$

Lemma 1.2 (Hermitian matrix)**.** *Hermitian matrix H has real eigenvalues*

Proof. If $H\boldsymbol{v} = \lambda\boldsymbol{v}$, $\lambda\boldsymbol{v}^\dagger\boldsymbol{v} = \boldsymbol{v}^\dagger H\boldsymbol{v} = \boldsymbol{v}^\dagger H^\dagger\boldsymbol{v} = (H\boldsymbol{v})^\dagger\boldsymbol{v} = (\lambda\boldsymbol{v})^\dagger\boldsymbol{v} = \lambda^*\boldsymbol{v}^\dagger\boldsymbol{v} \Rightarrow (\lambda - \lambda^*)\boldsymbol{v}^\dagger\boldsymbol{v} = 0$, where λ^* is the complex conjugate of λ. Since $\boldsymbol{v}^\dagger\boldsymbol{v} > 0$ unless $\boldsymbol{v}$ is all zero, $\lambda = \lambda^*$ and so λ is real. □

Lemma 1.3 (Hermitian matrix)**.** *Eigenvectors of different eigenvalues are orthogonal for Hermitian matrices.*

Proof. If for $\lambda_1 \neq \lambda_2$, $H\mathbf{v}_1 = \lambda_1\mathbf{v}_1$ and $H\mathbf{v}_2 = \lambda_2\mathbf{v}_2$, $\lambda_2\mathbf{v}_1^\dagger\mathbf{v}_2 = \mathbf{v}_1^\dagger H\mathbf{v}_2 = \mathbf{v}_1^\dagger H^\dagger\mathbf{v}_2 = (H\mathbf{v}_1)^\dagger\mathbf{v}_2 = (\lambda_1\mathbf{v}_1)^\dagger\mathbf{v}_2 = \lambda_1^\dagger\mathbf{v}_1^\dagger\mathbf{v}_2 = \lambda_1\mathbf{v}_1^\dagger\mathbf{v}_2$. So if $\lambda_1 \neq \lambda_2$, then $\mathbf{v}_1^\dagger\mathbf{v}_2 = 0$. That is, $\mathbf{v}_1 \perp \mathbf{v}_2$. □

Lemma 1.4 (Hermitian matrix)**.** *Any $n \times n$ Hermitian matrix has a complete set of orthogonal eigenvectors that form a basis of n-dimensional vector space.*

Corollary 1.1 (Hermitian matrix)**.** *From Lemma 1.4, any Hermitian matrix can be diagonalized by a unitary matrix (its eigenvector matrix). That $H = U\Lambda U^\dagger$, where $UU^\dagger = I$. As a degenerated case, any real symmetric matrix is Hermitian and its eigenvector matrix is orthogonal (both unitary and real). Thus H can be decomposed as $H = O\Lambda O^\top$, where O is real and $OO^\top = I$*

1.1.5 Positive definite matrices

- A Hermitian matrix S is positive definite if for any vector u, $u^\dagger S u > 0$
- A Hermitian matrix S is positive semi-definite if for any vector u, $u^\dagger S u \geq 0$
- A real symmetric matrix S is positive definite if for any real vector u, $u^\top S u > 0$
- A real symmetric matrix S is positive semi-definite if for any real vector u, $u^\top S u \geq 0$

Example 1.3 (Complex positive semi-definite)**.** For any complex matrix A, $S = A^\dagger A$ is positive semi-definite since $S^\dagger = (A^\dagger A)^\dagger = A^\dagger A = S$ and for any complex vector u, $u^\dagger S u = u^\dagger A^\dagger A u = (Au)^\dagger (Au) \geq 0$

Example 1.4 (Real positive semi-definite)**.** For any real matrix A, $S = A^\dagger A = A^\top A$ is positive semi-definite since $S^\top = (A^\top A)^\top = A^\top A = S$ and for any real vector u, $u^\top S u = u^\top A^\top A u = (Au)^\top (Au) \geq 0$

1.1.6 SVD

- Any matrix A has a singular value decomposition of a form of $UDV^\dagger$, where U and V are unitary and D is diagonal
- If A is real, the unitary matrices are real orthogonal instead. And thus $A = UDV^\top$
- For both complex and real case, the diagonal elements of D are called the singular values of A and the columns of U and V are called the left and right singular vectors of A. And U and V themselves are the left and right singular vector matrices.
- Consider $S = AA^\dagger$. Note that S is positive semi-definite and $S = UDV^\dagger VD^\dagger U^\dagger = U\tilde{D}U^\dagger$, where $\tilde{D} = DD^\dagger$ is a real diagonal matrix with non-negative elements. Apparently, U is also the eigenvector matrix of $AA^\dagger$.
- Similarly, consider $S = A^\dagger A$. Note that S is positive semi-definite and $S = VD^\dagger U^\dagger UDV^\dagger = V\tilde{D}V^\dagger$. Apparently, V is also the eigenvector matrix of $A^\dagger A$.
- Note that eigenvalues of $A^\dagger A$ and $AA^\dagger$ are magnitudes squared of the singular values of A

1.2 Matrix calculus

- We may take a vector or matrix as input variables of a function. Consequently, we can compute of "derivative" w.r.t. to the input vector of matrix
  - We only consider scalar function here for simplicity. For generalization, one can consider a function with more outputs simply as concatenation of scalar functions.
- Consider $f(\boldsymbol{u})$ as a function of vector $\boldsymbol{u}$. The derivative $\frac{df}{d\boldsymbol{u}}$ (essentially the gradient $\nabla f(\boldsymbol{u})$) is defined as a vector of same shape of $\boldsymbol{u}$ and with the ith element, $\left(\frac{df}{d\boldsymbol{u}}\right)_i = \frac{\partial f}{\partial u_i}$
  - At the extremum of f w.r.t. $\boldsymbol{u}$, we should have $\frac{df}{d\boldsymbol{u}} = \boldsymbol{0}$
- Consider $f(A)$ as a function of matrix A. The derivative $\frac{df}{dA}$ (essentially the gradient $\nabla f(A)$) is defined as a matrix of same shape of A and with the (i,j)-element, $\left(\frac{df}{dA}\right)_{i,j} = \frac{\partial f}{\partial A_{i,j}}$
  - At the extremum of f w.r.t. A, we should have $\frac{df}{dA} = 0$

Example 1.5. $f(\boldsymbol{v}) = \boldsymbol{u}^\top \boldsymbol{v}$, $\left(\frac{df}{d\boldsymbol{v}}\right)_i = \frac{\partial \boldsymbol{u}^\top \boldsymbol{v}}{\partial v_i} = \frac{\partial}{\partial v_i}\sum_j u_j v_j = u_i$. Therefore, $\frac{df}{d\boldsymbol{v}} = u$

Example 1.6. $f(A) = \boldsymbol{u}^T A \boldsymbol{v}$, $\left(\frac{df}{dA}\right)_{i,j} = \frac{\partial \boldsymbol{u}^\top A \boldsymbol{v}}{\partial A_{i,j}} = \frac{\partial}{\partial A_{i,j}}\sum_l \sum_k u_k A_{k,l} v_l = u_i v_j$. Therefore, $\frac{df}{dA} = \boldsymbol{u}\boldsymbol{v}^\top$

1.2.1 Lagrange multiplier

An important trick of optimization with equality or inequality constraints is the Lagrange multiplier.

Equality constraint Consider the problem that

$$\min_x f(x) \qquad \text{s.t.} \qquad g(x) = c$$

We can introduce a Lagrange multiplier λ and rewrite the problem as an unconstrained optimization problem

$$\min_x \max_\lambda \underbrace{f(x) + \lambda(g(x) - c)}_{L(x,\lambda)} \tag{1}$$

Note that

$$\max_\lambda L(x, \lambda) = \begin{cases} f(x) & \text{if } g(x) = c \\ \infty & \text{otherwise} \end{cases}$$

as desired. We may swap the order of the optimization in (1) and consider instead

$$\max_\lambda \min_x \underbrace{f(x) + \lambda(g(x) - c)}_{L(x,\lambda)}$$

And $\min_x f(x) + \lambda(g(x) - c)$ gives us $\nabla f(x) + \lambda \nabla g(x) = 0$

Inequality constraint Similarly, consider instead a problem with inequality constraint that

$$\min_x f(x) \qquad \text{s.t.} \qquad g(x) \geq c$$

Again, we can introduce a Lagrange multiplier λ and rewrite the problem as an unconstrained optimization problem

$$\min_x \max_{\lambda \geq 0} \underbrace{f(x) - \lambda(g(x) - c)}_{L(x,\lambda)}$$

Note that we restrict λ be non-negative this time and again

$$\max_{\lambda \geq 0} L(x, \lambda) = \begin{cases} f(x) & \text{if } g(x) \geq c \\ \infty & \text{otherwise} \end{cases}$$

as desired. We may swap the order of the optimization as before to continue solving the problem.

N.B. For each constraint, we have one Lagrange multiplier. For example, for "a" constraint $A\mathbf{x} = \mathbf{c}$, we can include the regularization term as $\lambda^\top (A\mathbf{x} - \mathbf{c})$, where λ has the same shape as $\mathbf{c}$.

1.2.2 Solving sets of linear equations

Consider a set of linear equations

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \cdots + a_{1n}x_n = c_1 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 + \cdots + a_{2n}x_n = c_2 \\ \vdots \\ a_{m1}x_1 + a_{m2}x_2 + a_{m3}x_3 + \cdots + a_{mn}x_n = c_m \end{cases}$$

with m equations and n unknowns $x_1, x_2, \cdots, x_n$. We can rewrite the set of equations in matrix form as

$$\underbrace{\begin{pmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ & & \vdots & & \\ a_{m1} & a_{m2} & a_{m3} & \cdots & a_{mn} \end{pmatrix}}_{A} \underbrace{\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}}_{\mathbf{x}} = \underbrace{\begin{pmatrix} c_1 \\ \vdots \\ c_m \end{pmatrix}}_{\mathbf{c}} \tag{2}$$

$m = n$**:** If A is square ($m = n$), and assume that A is full-rank[1], we can multiply A^{-1} on both sides and so $\mathbf{x} = A^{-1}\mathbf{c}$.

$m > n$**:** When $m > n$ and assume that A has rank n, there are more equations than unknown. In general, we won't be able to find $\mathbf{x}$ that satisfies (2) exactly. Instead reformulate the problem[2] as

$$\min_{\mathbf{x}} \underbrace{(A\mathbf{x} - \mathbf{c})^\top (A\mathbf{x} - \mathbf{c})}_{f(\mathbf{x})} \tag{3}$$

So we want $\frac{df(\mathbf{x})}{d\mathbf{x}} = \frac{d}{d\mathbf{x}}(A\mathbf{x} - \mathbf{c})^\top (A\mathbf{x} - \mathbf{c}) = \frac{d}{d\mathbf{x}}[\mathbf{x}^\top A^\top A\mathbf{x} - \mathbf{c}^\top A\mathbf{x} - \mathbf{x}^\top A^\top \mathbf{c} + \mathbf{c}^\top \mathbf{c}] = 0$. This gives us[3]

$$\frac{df(\mathbf{x})}{d\mathbf{x}} = 2A^\top A\mathbf{x} - 2A^\top \mathbf{c} = 0 \tag{4}$$

$$\Rightarrow A^\top A\mathbf{x} = A^\top \mathbf{c} \tag{5}$$

$$\Rightarrow \mathbf{x} = (A^\top A)^{-1} A^\top \mathbf{c}, \tag{6}$$

where $(A^\top A)^{-1} A^\top$ is known as the pseudo-inverse of A

$m < n$**:** When the $m < n$, we have more unknown than equations and so in general there are infinite number of solutions. In that case, it is often that we want to impose some "regularization" to favor a less complex $\mathbf{x}$. For example, we may add the l_2-norm of $\mathbf{x}$ as a model cost. So instead, we may formulate the problem as

$$\min_{\mathbf{x}} (A\mathbf{x} - \mathbf{c})^\top (A\mathbf{x} - \mathbf{c}) + \lambda \mathbf{x}^\top \mathbf{x} \tag{7}$$

This gives a solution $\mathbf{x} = (A^\top A + \lambda I)^{-1} A^\top \mathbf{c}$

Alternatively[4], we may formulate the problem as

$$\min_{\mathbf{x}} \mathbf{x}^\top \mathbf{x} \qquad \text{s.t.} \qquad A\mathbf{x} = \mathbf{c}. \tag{8}$$

Using Lagrange multiplier, rewrite the problem as

$$\min_{\mathbf{x}} \mathbf{x}^\top \mathbf{x} + \lambda^\top (A\mathbf{x} - \mathbf{c}) \tag{9}$$

This gives $2\mathbf{x} + A^\top \lambda = 0 \Rightarrow \mathbf{x} = A^\top \tilde{\lambda}$. From the constraint, $A\mathbf{x} = \mathbf{c}$, we have $AA^\top \tilde{\lambda} = \mathbf{c}$ and so $\mathbf{x} = A^\top \tilde{\lambda} = A^\top (AA^\top)^{-1}\mathbf{c}$.

1.2.3 Finding the minimum of a quadratic form

Another common problem that we may encounter is to find the minimum of $\|A\mathbf{x}\|$ for some *real* matrix A. Note that the problem is not well-defined if we don't put any constraint on $\mathbf{x}$. Otherwise, the minimum is simply the trivial all-zero vector. Instead we need to constrain $\|\mathbf{x}\|$ to be some constant, and WLOG we can pick $\|\mathbf{x}\| = 1$. Rewrite the problem as an optimization problem

$$\min_{\mathbf{x}} \|A\mathbf{x}\| \qquad \text{s.t.} \qquad \|\mathbf{x}\| = 1 \tag{10}$$

$$\Rightarrow \min_{\mathbf{x}} \mathbf{x}^\top A^\top A\mathbf{x} \qquad \text{s.t.} \qquad \mathbf{x}^\top \mathbf{x} = 1 \tag{11}$$

Using Lagrange multiplier, we can rewrite the problem as

$$\min_{\mathbf{x}} \mathbf{x}^\top A^\top A\mathbf{x} - \lambda(\mathbf{x}^\top \mathbf{x} - 1) \tag{12}$$

This gives us

$$A^\top A\mathbf{x} = \lambda \mathbf{x} \tag{13}$$

This suggests that λ to be eigenvalue of $A^\top A$ and $\mathbf{x}$ is the eigenvector. However, *which eigenvalue and eigenvector*?

Since $A^\top A$ is positive semi-definite, the eigenvalues of $A^\top A$ are all real and non-negative[5] Let

[1] Otherwise, it is the same as $m < n$.

[2] Assume real matrices and vectors here.

[3] Note that $\left(\frac{d}{d\mathbf{x}}\mathbf{x}^\top B\mathbf{x}\right)_i = \frac{\partial}{\partial x_i}\mathbf{x}^\top B\mathbf{x} = \frac{\partial}{\partial x_i}\sum_j \sum_k x_j B_{j,k} x_k = \sum_k B_{i,k}x_k + \sum_j B_{j,i}x_j = \sum_k (B_{i,k} + B^\top_{i,k})x_k = ((B + B^\top)\mathbf{x})_i$, thus $\frac{d}{d\mathbf{x}}\mathbf{x}^\top B\mathbf{x} = (B + B^\top)\mathbf{x}$

[4] We may also use other norm such as l_1-norm for the model cost. But generally there is no closed form solution for those cases.

[5]. Note that we also have a set of orthogonal eigenvectors that span the entire vector space (c.f. Lemma 1.4).

$\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n$ be the normalized[6] orthogonal[7] eigenvectors of $A^\top A$ with eigenvalues $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$. Note that

$$\mathbf{x}_i^\top A^\top A \mathbf{x}_i = \mathbf{x}_i^\top \lambda_i \mathbf{x}_i = \lambda_i \mathbf{x}_i^\top \mathbf{x}_i = \lambda_i \|\mathbf{x}_i\|^2 = \lambda_i.$$

Therefore, the optimum $\mathbf{x}$ should simply be $\mathbf{x}_1$, the eigenvector of $A^\top A$ with the minimum eigenvalue[8].

Principal component analysis (PCA)

The above discussion is closely related to PCA. Consider a dataset of N length-M vectors, often $M \gg N$. Let's pack all data vectors into a matrix A such that each row is a data vector. So that $A \in \mathbb{R}^{N \times M}$. The goal of PCA is to reduce the dimension of the features to $m \ll M$.
For simplicity, let's first consider $m = 1$, we want to find a projection $\mathbf{p} \in \mathbb{R}^{M \times 1}$ such that $\mathbf{p}^\top \mathbf{p} = 1$ and the signal power after projection $\|A\mathbf{p}\|^2 = \mathbf{p}^\top A^\top A \mathbf{p}$ is maximized. That is,

$$\max \mathbf{p}^\top A^\top A \mathbf{p} \qquad \text{s.t.} \qquad \mathbf{p}^\top \mathbf{p} = 1 \tag{14}$$

By the same argument as before, the optimum $\mathbf{p}$ will be the eigenvector of $A^\top A$ with the maximum eigenvector.
For $m > 1$, we should just pick the projection matrix who columns are the eigenvectors of $A^\top A$ of the first m largest eigenvalues. Note that we can reduce amount of computation by using SVD on A instead.

1.3 Homogeneous coordinate

Homogenous coordinate is an extension of the Cartesian coordinate. A dummy coefficient 1 is appended to the regular Cartesian coordinate to form the Homogenous coordinate. For example, a 2-D point (x, y) in the Cartesian coordinate will be represented as $(x, y, 1)$ in the homogenous coordinate. For a Homogenous coordinate with the last coefficient being non-zero, we can renormalize it by dividing all coefficients with the last coefficient. Note that the physical location represented by a Homogenous is unchanged w.r.t. scaling. For example, $(u, v, w) = (u/w, v/w, 1)$ and thus the physical point that (u, v, w) represents is actually $(u/w, v/w)$. It will become clear why homogeneous coordinate is useful soon in the next sub-section.

1.4 Coordinate transformation

1.4.1 2D

Note that while rotation can be represented as a matrix operation (multiplication of a matrix) on a point for the Cartesian coordinate, it is not possible for translation. But everything can be represented as matrix multiplications using the homogeneous coordinate system

	Coordinate	Input	Operation	Output
Translation	Cartesian	$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$	shift by $\begin{pmatrix} p_1 \\ p_2 \end{pmatrix}$	$\begin{pmatrix} x_1 + p_1 \\ x_2 + p_2 \end{pmatrix}$
	Homogeneous	$\begin{pmatrix} x_1 \\ x_2 \\ 1 \end{pmatrix}$	shift by $\begin{pmatrix} p_1 \\ p_2 \\ 1 \end{pmatrix}$	$\begin{pmatrix} x_1 + p_1 \\ x_2 + p_2 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & p_1 \\ 0 & 1 & p_2 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ 1 \end{pmatrix}$
Rotation	Cartesian	$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$	rotate by θ	$\begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$
	Homogeneous	$\begin{pmatrix} x_1 \\ x_2 \\ 1 \end{pmatrix}$		$\begin{pmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ 1 \end{pmatrix}$

[6] $\|\mathbf{x}_i\| = 1, \forall i$.
[7] $\mathbf{x}_i^\top \mathbf{x}_j = 0$ for $i \neq j$.
[8] We may also use SVD and find the right singular vector with the minimum singular value instead.

1.4.2 3D

	Coordinate	Input	Operation	Output
Translation	Cartesian	$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$	shift by $\begin{pmatrix} p_1 \\ p_2 \\ p_3 \end{pmatrix}$	$\begin{pmatrix} x_1 + p_1 \\ x_2 + p_2 \\ x_3 + p_3 \end{pmatrix}$
	Homogeneous	$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ 1 \end{pmatrix}$	shift by $\begin{pmatrix} p_1 \\ p_2 \\ p_3 \\ 1 \end{pmatrix}$	$\begin{pmatrix} x_1 + p_1 \\ x_2 + p_2 \\ x_3 + p_3 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & p_1 \\ 0 & 1 & 0 & p_2 \\ 0 & 0 & 1 & p_3 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ 1 \end{pmatrix}$
Rotation	Cartesian	$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$	rotate along z-axis by θ	$\begin{pmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$
	Homogeneous	$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ 1 \end{pmatrix}$		$\begin{pmatrix} \cos\theta & -\sin\theta & 0 & 0 \\ \sin\theta & \cos\theta & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ 1 \end{pmatrix}$
	Cartesian	$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$	rotate along x-axis by θ	$\begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & -\sin\theta \\ 0 & \sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$
	Homogeneous	$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ 1 \end{pmatrix}$		$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos\theta & -\sin\theta & 0 \\ 0 & \sin\theta & \cos\theta & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ 1 \end{pmatrix}$
	Cartesian	$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$	rotate along y-axis by θ	$\begin{pmatrix} \cos\theta & 0 & \sin\theta \\ 0 & 1 & 0 \\ -\sin\theta & 0 & \cos\theta \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$
	Homogeneous	$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ 1 \end{pmatrix}$		$\begin{pmatrix} \cos\theta & 0 & \sin\theta & 0 \\ 0 & 1 & 0 & 0 \\ -\sin\theta & 0 & \cos\theta & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ 1 \end{pmatrix}$

Rotating from axes $\mathbf{x}, \mathbf{y}, \mathbf{z}$ to $\mathbf{u}, \mathbf{v}, \mathbf{w}$

Note that the rotation matrix is simply $\begin{pmatrix} \mathbf{u}^\top \\ \mathbf{v}^\top \\ \mathbf{w}^\top \end{pmatrix}$. One can easily verify as follows.

- Under the original frame of reference, if we rotate $\mathbf{u}$ with the given matrix, we have $\begin{pmatrix} \mathbf{u}^\top \\ \mathbf{v}^\top \\ \mathbf{w}^\top \end{pmatrix} \mathbf{u} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$ as desired since the point is precisely $\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$ under the new frame of reference.
- Similarly, we have $\begin{pmatrix} \mathbf{u}^\top \\ \mathbf{v}^\top \\ \mathbf{w}^\top \end{pmatrix} \mathbf{v} = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$ and $\begin{pmatrix} \mathbf{u}^\top \\ \mathbf{v}^\top \\ \mathbf{w}^\top \end{pmatrix} \mathbf{w} = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$ as desired.