CSE 203B HW0

Oguz Paksoy [opaksoy@ucsd.edu]

January 2021

1 Matrix Properties

1.1 Linear System

$$\begin{bmatrix} 2 & 4 & 6 \\ 1 & -1 & 2 \\ 3 & 0 & 5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ -1 \\ 2 \end{bmatrix}$$

1.2 Range and Rank

Row reducing the matrix:

$$\begin{bmatrix} 2 & 4 & 6 \\ 1 & -1 & 2 \\ 3 & 0 & 5 \end{bmatrix} \xrightarrow{R_1 \leftarrow R_1/2} \begin{bmatrix} 1 & 2 & 3 \\ 1 & -1 & 2 \\ 3 & 0 & 5 \end{bmatrix} \xrightarrow{R_2 \leftarrow R_2 - R_1} \begin{bmatrix} 1 & 2 & 3 \\ 0 & -3 & -1 \\ 3 & 0 & 5 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 2 & 3 \\ 0 & -3 & -1 \\ 3 & 0 & 5 \end{bmatrix} \xrightarrow{R_3 \leftarrow R_3 - 3R_1} \begin{bmatrix} 1 & 2 & 3 \\ 0 & -3 & -1 \\ 0 & -6 & -4 \end{bmatrix} \xrightarrow{R_2 \leftarrow -R_2/3} \begin{bmatrix} 1 & 2 & 3 \\ 0 & 1 & 1/3 \\ 0 & -6 & -4 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 2 & 3 \\ 0 & 1 & 1/3 \\ 0 & -6 & -4 \end{bmatrix} \xrightarrow{R_3 \leftarrow R_3 + 6R_2} \begin{bmatrix} 1 & 2 & 3 \\ 0 & 1 & 1/3 \\ 0 & 0 & -2 \end{bmatrix} \xrightarrow{R_3 \leftarrow -R_3/2} \begin{bmatrix} 1 & 2 & 3 \\ 0 & 1 & 1/3 \\ 0 & 0 & 1 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 2 & 3 \\ 0 & 1 & 1/3 \\ 0 & 0 & 1 \end{bmatrix} \xrightarrow{R_2 \leftarrow R_2 - R_3/3} \begin{bmatrix} 1 & 2 & 3 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \xrightarrow{R_1 \leftarrow R_1 - 3R_3} \begin{bmatrix} 1 & 2 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 2 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \xrightarrow{R_1 \leftarrow R_1 - 2R_2} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = I_3$$

Since the reduced row echelon form of the matrix contains 3 pivots, its rank is 3. The range is the span of the column vectors that contain the pivot positions, which is onto $\mathbb{R}^3$ since the columns are independent of each other:

$$R = \text{span}\left\{ \begin{bmatrix} 2 \\ 1 \\ 3 \end{bmatrix}, \begin{bmatrix} 4 \\ -1 \\ 0 \end{bmatrix}, \begin{bmatrix} 6 \\ 2 \\ 5 \end{bmatrix} \right\} = \mathbb{R}^3$$

1.3 Nullspace

The nullspace is the set of points that satisfy the following equation:

$$\begin{bmatrix} 2 & 4 & 6 \\ 1 & -1 & 2 \\ 3 & 0 & 5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

Using row reduction on the augmented form, the same way it was done in question (1.2):

$$\left[\begin{array}{ccc|c} 2 & 4 & 6 & 0 \\ 1 & -1 & 2 & 0 \\ 3 & 0 & 5 & 0 \end{array}\right] \xrightarrow{\text{(1.2) } \textbf{row reductions}} \left[\begin{array}{ccc|c} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{array}\right] \implies \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

The nullspace is therefore a space which uses the 0 vector as its basis, which means that it only spans a single point.
In general, the number of dimensions of the range and the number of dimensions of the nullspace add up to the number of columns in a matrix. Here, we have a 3-dimensional range, and a 0-dimensional nullspace, which adds up to 3, the number of columns in our matrix. Furthermore, vectors in the nullspace (nullspace(A)) are orthogonal to the vectors in the range of the transpose (range $\left(A^T\right)$). Since our nullspace only consists of the 0 vector, then by definition, the dot product of any vector with the vectors in the nullspace is 0.

1.4 Trace, Determinant, Eigenvalues, Eigenvectors

The trace of a matrix is the sum of the elements along the main diagonal:

$$\text{trace}\left(\begin{bmatrix} 2 & 4 & 6 \\ 1 & -1 & 2 \\ 3 & 0 & 5 \end{bmatrix}\right) = 2 + (-1) + 5 = 6$$

The determinant for the matrix can be calculated as follows:

$$\begin{vmatrix} 2 & 4 & 6 \\ 1 & -1 & 2 \\ 3 & 0 & 5 \end{vmatrix} = 2\begin{vmatrix} -1 & 2 \\ 0 & 5 \end{vmatrix} - 4\begin{vmatrix} 1 & 2 \\ 3 & 5 \end{vmatrix} + 6\begin{vmatrix} 1 & -1 \\ 3 & 0 \end{vmatrix} = 2(-5) - 4(-1) + 6(3) = 12$$

The eigenvalues of a matrix A are λ values that satisfy $(A - \lambda I)x = 0$. We can figure out the eigenvalues by setting the determinant of this expression to 0:

$$\begin{aligned}
&\begin{vmatrix} 2-\lambda & 4 & 6 \\ 1 & -1-\lambda & 2 \\ 3 & 0 & 5-\lambda \end{vmatrix} \\
&= (2-\lambda)\begin{vmatrix} -1-\lambda & 2 \\ 0 & 5-\lambda \end{vmatrix} - 4\begin{vmatrix} 1 & 2 \\ 3 & 5-\lambda \end{vmatrix} + 6\begin{vmatrix} 1 & -1-\lambda \\ 3 & 0 \end{vmatrix} \\
&= (2-\lambda)(-1-\lambda)(5-\lambda) - 4(-1-\lambda) + 6(3+3\lambda) \\
&= (\lambda^2 - 7\lambda + 10)(-1-\lambda) - 4(-1-\lambda) - 18(-1-\lambda) \\
&= (\lambda^2 - 7\lambda - 12)(-1-\lambda) = 0 \\
&= -\left(\lambda - \frac{7-\sqrt{97}}{2}\right)\left(\lambda - \frac{7+\sqrt{97}}{2}\right)(\lambda+1) = 0 \\
&\implies \lambda_1 = \frac{7+\sqrt{97}}{2},\ \lambda_2 = -1,\ \lambda_3 = \frac{7-\sqrt{97}}{2}
\end{aligned}$$

Plugging these λ values into our equation $(A - \lambda I)x = 0$, we can figure out the corresponding eigenvectors from the general form of solutions to x:

$$\begin{bmatrix} 2-\lambda & 4 & 6 \\ 1 & -1-\lambda & 2 \\ 3 & 0 & 5-\lambda \end{bmatrix} \xrightarrow{R_1 \leftarrow \frac{1}{2-\lambda}R_1} \begin{bmatrix} 1 & \frac{4}{2-\lambda} & \frac{6}{2-\lambda} \\ 1 & -1-\lambda & 2 \\ 3 & 0 & 5-\lambda \end{bmatrix}$$

$$\begin{bmatrix} 1 & \frac{4}{2-\lambda} & \frac{6}{2-\lambda} \\ 1 & -1-\lambda & 2 \\ 3 & 0 & 5-\lambda \end{bmatrix} \xrightarrow{R_2 \leftarrow R_2 - R_1} \begin{bmatrix} 1 & \frac{4}{2-\lambda} & \frac{6}{2-\lambda} \\ 0 & \frac{\lambda^2-\lambda-6}{2-\lambda} & \frac{-2-2\lambda}{2-\lambda} \\ 3 & 0 & 5-\lambda \end{bmatrix}$$

$$\begin{bmatrix} 1 & \frac{4}{2-\lambda} & \frac{6}{2-\lambda} \\ 0 & \frac{\lambda^2-\lambda-6}{2-\lambda} & \frac{-2-2\lambda}{2-\lambda} \\ 3 & 0 & 5-\lambda \end{bmatrix} \xrightarrow{R_3 \leftarrow R_3 - 3R_1} \begin{bmatrix} 1 & \frac{4}{2-\lambda} & \frac{6}{2-\lambda} \\ 0 & \frac{\lambda^2-\lambda-6}{2-\lambda} & \frac{-2-2\lambda}{2-\lambda} \\ 0 & \frac{-12}{2-\lambda} & \frac{\lambda^2-7\lambda-8}{2-\lambda} \end{bmatrix}$$

$$\begin{bmatrix} 1 & \frac{4}{2-\lambda} & \frac{6}{2-\lambda} \\ 0 & \frac{\lambda^2-\lambda-6}{2-\lambda} & \frac{-2-2\lambda}{2-\lambda} \\ 0 & \frac{-12}{2-\lambda} & \frac{\lambda^2-7\lambda-8}{2-\lambda} \end{bmatrix} \xrightarrow{R_2 \leftarrow \frac{2-\lambda}{\lambda^2-\lambda-6} R_2} \begin{bmatrix} 1 & \frac{4}{2-\lambda} & \frac{6}{2-\lambda} \\ 0 & 1 & \frac{-2-2\lambda}{\lambda^2-\lambda-6} \\ 0 & \frac{-12}{2-\lambda} & \frac{\lambda^2-7\lambda-8}{2-\lambda} \end{bmatrix}$$

$$\begin{bmatrix} 1 & \frac{4}{2-\lambda} & \frac{6}{2-\lambda} \\ 0 & 1 & \frac{-2-2\lambda}{\lambda^2-\lambda-6} \\ 0 & \frac{-12}{2-\lambda} & \frac{\lambda^2-7\lambda-8}{2-\lambda} \end{bmatrix} \xrightarrow{R_3 \leftarrow R_3 + \frac{12}{2-\lambda} R_2} \begin{bmatrix} 1 & \frac{4}{2-\lambda} & \frac{6}{2-\lambda} \\ 0 & 1 & \frac{-2-2\lambda}{\lambda^2-\lambda-6} \\ 0 & 0 & \frac{(\lambda^2-7\lambda-12)(-\lambda-1)}{\lambda^2-\lambda-6} \end{bmatrix}$$

$$\begin{bmatrix} 1 & \frac{4}{2-\lambda} & \frac{6}{2-\lambda} \\ 0 & 1 & \frac{-2-2\lambda}{\lambda^2-\lambda-6} \\ 0 & 0 & \frac{(\lambda^2-7\lambda-12)(-\lambda-1)}{\lambda^2-\lambda-6} \end{bmatrix} \xrightarrow{R_1 \leftarrow R_1 - \frac{4}{2-\lambda} R_2} \begin{bmatrix} 1 & 0 & \frac{-6\lambda-14}{\lambda^2-\lambda-6} \\ 0 & 1 & \frac{-2-2\lambda}{\lambda^2-\lambda-6} \\ 0 & 0 & \frac{(\lambda^2-7\lambda-12)(-\lambda-1)}{\lambda^2-\lambda-6} \end{bmatrix}$$

Since we derived the eigenvalues from the equation $(\lambda^2 - 7\lambda - 12)(-1 - \lambda) = 0$, and the eigenvalues are not roots of the quadratic equation $\lambda^2 - \lambda - 6$, the third row of the matrix is a row of zeros for all eigenvalues. This means that the values for x_1 and x_2 can only be determined in terms of x_3, where $x_3 \in \mathbb{R} - \{0\}$:

$$x_1 = \frac{2(3\lambda + 7)}{(\lambda - 3)(\lambda + 2)} x_3, \quad x_2 = \frac{2(\lambda + 1)}{(\lambda - 3)(\lambda + 2)} x_3$$

Setting x_3 to an arbitrary nonzero value, we can derive the following expressions for each eigenvector:

$$\lambda_1 = \frac{7 + \sqrt{97}}{2} \implies \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = c \begin{bmatrix} \frac{-3+\sqrt{97}}{6} \\ \frac{1}{3} \\ 1 \end{bmatrix}, \ c \in \mathbb{R} - \{0\}$$

$$\lambda_2 = -1 \implies \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = c \begin{bmatrix} -2 \\ 0 \\ 1 \end{bmatrix}, \ c \in \mathbb{R} - \{0\}$$

$$\lambda_3 = \frac{7 - \sqrt{97}}{2} \implies \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = c \begin{bmatrix} \frac{-3-\sqrt{97}}{6} \\ \frac{1}{3} \\ 1 \end{bmatrix}, \ c \in \mathbb{R} - \{0\}$$

1.5 Matrix Proof

1.5.a Trace Commutativity

$$\begin{aligned}\text{trace}(AB) &= \sum_{i=1}^{m}(AB)_{ii} = \sum_{i=1}^{m}\sum_{j=1}^{n} A_{ij}B_{ji} \\ &= \sum_{j=1}^{n}\sum_{i=1}^{m} B_{ji}A_{ij} = \sum_{j=1}^{n}(BA)_{jj} = \text{trace}(BA)\end{aligned}$$

1.5.b Determinant Distributivity

Assume A is non-invertible. Since $(AB)^{-1}$ is equivalent to $B^{-1}A^{-1}$, AB is also non-invertible. The determinant of a non-invertible matrix is 0:

$$0 = 0 \cdot \det(B) \implies \det(AB) = \det(A)\det(B)$$

Now assume that A is invertible. This means there's a finite set of row reduction operations that reduce A to the identity matrix. Each of these operations can be represented via an elementary matrix product. Let $A = E_iE_{i-1}\dots E_2E_1I$. Since each matrix operation denotes a single row reduction operation, they each affect the determinant of the matrix in front of them by a constant multiplier. This multiplier is the scale itself for row scaling, -1 for row swapping, and 1 for row addition and subtraction operations. Let $e_i, e_{i-1}, \dots, e_2, e_1 \in \mathbb{R} - \{0\}$ be the corresponding multipliers. Then:

$$\begin{aligned}\det(A) &= \det(E_iE_{i-1}\dots E_2E_1I) \\ &= \det(e_ie_{i-1}\dots e_2e_1I) \\ &= e_ie_{i-1}\dots e_2e_1\det(I) \\ &= e_ie_{i-1}\dots e_2e_1\end{aligned}$$

When multiplying with another matrix B, the matrix $A = E_iE_{i-1}\dots E_2E_1I$ can be thought of row reduction operations that operate on B instead of the identity matrix. This permutes the determinant of B similarly:

$$\begin{aligned}\det(AB) &= \det(E_iE_{i-1}\dots E_2E_1B) \\ &= \det(e_ie_{i-1}\dots e_2e_1B) \\ &= e_ie_{i-1}\dots e_2e_1\det(B)\end{aligned}$$

From the previous equation, we know $\det(A) = e_ie_{i-1}\dots e_2e_1$. Therefore:

$$\det(A) = e_ie_{i-1}\dots e_2e_1 \implies \det(AB) = \det(A)\det(B)$$

Since A is either invertible or non-invertible, we can conclude that:

$$\det(AB) = \det(A)\det(B)$$

1.5.c Determinant-Eigenvalue Relation

Let $\lambda_1, \lambda_2, \ldots, \lambda_n$ be eigenvalues of A. From the definition of the characteristic polynomial of A, which is defined to be 0 at each eigenvalue:

$$p_A(t) = \det(A - tI) = (\lambda_1 - t)(\lambda_2 - t)\ldots(\lambda_n - t)$$

Then, for $t = 0$ we have:

$$p_A(0) = \det(A) = \lambda_1\lambda_2\ldots\lambda_n = \prod_{i=1}^{n}\lambda_i$$

1.6 Examples

1.6.a Trace

$$A = \begin{bmatrix} 1 & e & 0 \\ \pi & 2 & 0 \\ 0 & 0 & 2 \end{bmatrix},\ B = \begin{bmatrix} 1 & 0 & 0 \\ e & 2 & 0 \\ \pi & 0 & 1 \end{bmatrix}$$

$$AB = \begin{bmatrix} 1+e^2 & 2e & 0 \\ 2e+\pi & 4 & 0 \\ 2\pi & 0 & 2 \end{bmatrix},\ BA = \begin{bmatrix} 1 & e & 0 \\ 2e+\pi & 4+e^2 & 0 \\ \pi & e\pi & 2 \end{bmatrix}$$

$$\text{trace}(AB) = \text{trace}(BA) = 7 + e^2$$

1.6.b Determinant Distributivity

$$A = \begin{bmatrix} 1 & e & 0 \\ \pi & 2 & 0 \\ 0 & 0 & 2 \end{bmatrix},\ B = \begin{bmatrix} 1 & 0 & 0 \\ e & 2 & 0 \\ \pi & 0 & 1 \end{bmatrix},\ AB = \begin{bmatrix} 1+e^2 & 2e & 0 \\ 2e+\pi & 4 & 0 \\ 2\pi & 0 & 2 \end{bmatrix}$$

$$\det(A) = 4 - 2e\pi,\ \det(B) = 2,\ \det(AB) = 8 - 4e\pi = \det(A)\det(B)$$

1.6.c Determinant-Eigenvalue Relation

$$A = \begin{bmatrix} e+\pi & 1 & 1 \\ 0 & 5-2e & 1 \\ 0 & 0 & e-\pi \end{bmatrix}$$

The eigenvalues for A:

$$(\lambda-(e+\pi))(\lambda-(5-2e))(\lambda-(e-\pi)) = 0 \implies \lambda_1 = e+\pi,\ \lambda_2 = 5-2e,\ \lambda_3 = e-\pi$$

The determinant for A:

$$\det(A) = 5e^2 - 2e^3 - 5\pi^2 + 2e\pi^2 = (e+\pi)(5-2e)(e-\pi) = \lambda_1\lambda_2\lambda_3 = \prod_{i=1}^{3} \lambda_i$$

2 Matrix Operations

2.1 Gradient and Hessian of $f(x) = b^T x$

$$f(x) = b^T x = \sum_{i=1}^{n} b_i x_i$$

Gradient:

$$\nabla_x f(x) = \begin{bmatrix} \frac{\partial f(x)}{\partial x_1} \\ \frac{\partial f(x)}{\partial x_2} \\ \dots \\ \frac{\partial f(x)}{\partial x_n} \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \dots \\ b_n \end{bmatrix} = b$$

Since the gradient, the first order derivatives are constant, the Hessian is a matrix of zeroes:

$$\nabla_x^2 f(x) = \begin{bmatrix} \frac{\partial^2 f(x)}{\partial x_1^2} & \frac{\partial^2 f(x)}{\partial x_1 \partial x_2} & \dots & \frac{\partial^2 f(x)}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f(x)}{\partial x_1 \partial x_2} & \frac{\partial^2 f(x)}{\partial x_2^2} & \dots & \frac{\partial^2 f(x)}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f(x)}{\partial x_1 \partial x_n} & \frac{\partial^2 f(x)}{\partial x_2 \partial x_n} & \dots & \frac{\partial^2 f(x)}{\partial x_n^2} \end{bmatrix} = \begin{bmatrix} 0 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 \end{bmatrix}$$

2.2 Gradient and Hessian of $f(x) = x^T A x + b^T x + c$

Using the matrix-vector multiplication definitions:

$$f(x) = \sum_{j=1}^{n} \sum_{i=1}^{n} x_j A_{ji} x_i + \sum_{i=1}^{n} b_i x_i + c$$

Calculating the gradient using the sum definitions:

$$\nabla_x f(x) = \begin{bmatrix} \frac{\partial f(x)}{\partial x_1} \\ \frac{\partial f(x)}{\partial x_2} \\ \dots \\ \frac{\partial f(x)}{\partial x_n} \end{bmatrix} = \begin{bmatrix} (\sum_{i=1}^{n} A_{1i} x_i + \sum_{j=1}^{n} x_j A_{j1}) + b_1 \\ (\sum_{i=1}^{n} A_{2i} x_i + \sum_{j=1}^{n} x_j A_{j2}) + b_2 \\ \dots \\ (\sum_{i=1}^{n} A_{ni} x_i + \sum_{j=1}^{n} x_j A_{jn}) + b_n \end{bmatrix}$$

The derivative for $x^T A x$ double-counts the intersection of the row and column of A corresponding to the gradient index, however since $\frac{\partial (A_{kk} x_k^2)}{\partial x_k} = 2A_{kk} x_k$, this double counting coincides with the sums shown above. Then, by using the fact that A is symmetric:

$$\nabla_x f(x) = \begin{bmatrix} 2\sum_{i=1}^n A_{1i}x_i + b_1 \\ 2\sum_{i=1}^n A_{2i}x_i + b_2 \\ \dots \\ 2\sum_{i=1}^n A_{ni}x_i + b_n \end{bmatrix} = 2Ax + b$$

Differentiating the gradient with respect to x values once more, we can get the Hessian:

$$\begin{aligned} \nabla_x^2 f(x) &= \begin{bmatrix} \frac{\partial^2 f(x)}{\partial x_1^2} & \frac{\partial^2 f(x)}{\partial x_1 \partial x_2} & \dots & \frac{\partial^2 f(x)}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f(x)}{\partial x_1 \partial x_2} & \frac{\partial^2 f(x)}{\partial x_2^2} & \dots & \frac{\partial^2 f(x)}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f(x)}{\partial x_1 \partial x_n} & \frac{\partial^2 f(x)}{\partial x_2 \partial x_n} & \dots & \frac{\partial^2 f(x)}{\partial x_n^2} \end{bmatrix} \\ &= \begin{bmatrix} 2A_{11} & 2A_{12} & \dots & 2A_{1n} \\ 2A_{21} & 2A_{22} & \dots & 2A_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 2A_{n1} & 2A_{n2} & \dots & 2A_{nn} \end{bmatrix} = 2A \end{aligned}$$

2.3 Gradient of $f(X) = \log \det X$

Note that we can also derive via first-order approximation (e.g. Boyd & Vandenbergh Sec A.4), cofactor/Laplace expansion. Here, we give an informal derivation using an outer product reparameterization + trace trick.

Using the derivative of log:

$$\nabla_X f(X) = \frac{\partial f(X)}{\partial X} = \frac{1}{\det X} \frac{\partial (\det X)}{\partial X}$$

For the determinant's derivative, we first describe the adjugate matrix $\text{adj}(X)$. The adjugate matrix for an invertible matrix is the transpose of the cofactor matrix C_X, which can be used to invert a matrix as shown below:

$$X^{-1} = \frac{1}{\det X} C_X^T = \frac{1}{\det X} \text{adj}(X)$$

Since the derivative we're looking for yields a matrix, we should consider $\frac{\partial (\det X)}{\partial X_{ij}}$, for all i, j values, which will fill the ith row and jth column in the derivative matrix. This expression can be derived from the Jacobi formula. Furthermore, using the fact that $\frac{\partial X}{\partial X_{ij}}$ is a matrix of zeroes except at the ij, it follows that

$\frac{\partial X}{\partial X_{ij}}$ may be expressed as an outer product $u_i u_j^\top$, where u_i corresponds to the i-th row of the $n \times n$ identity. Thus,

$$\frac{\partial(\det X)}{\partial X_{ij}} = \text{trace}\left(\text{adj}(X)\frac{\partial X}{\partial X_{ij}}\right)$$

Finally,

$$\begin{aligned}
\frac{\partial(\log\det X)}{\partial X_{ij}} &= \frac{1}{\det X}\text{trace}\left(\text{adj}(X)\frac{\partial X}{\partial X_{ij}}\right) \\
&= \text{trace}\left(X^{-1}\frac{\partial X}{\partial X_{ij}}\right) \\
&= \text{trace}\left(X^{-1}u_i u_j^\top\right) \\
&= \text{trace}\left(u_j^\top X^{-1} u_i\right) = y_{ji}
\end{aligned}$$

where y_{ji} is the ji-th element of X^{-1} (i.e. the ij-th element of $(X^{-1})^\top$). So, we have that $\nabla_X f(X) = (X^{-1})^\top$. And since X is symmetric, we also have that $\nabla_X f(X) = X^{-1}$.

2.4 Examples

2.4.a Gradient and Hessian of $f(x) = x^T Ax + b^T x + c$

$$A = \begin{bmatrix} 2 & 5 \\ 5 & 3 \end{bmatrix}, b = \begin{bmatrix} 1 \\ 7 \end{bmatrix}, c = 11$$

$$f\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right) = (2x_1^2 + 10x_1x_2 + 3x_2^2) + (x_1 + 7x_2) + 11$$

$$\nabla_x f(x) = \begin{bmatrix} \frac{\partial f(x)}{\partial x_1} \\ \frac{\partial f(x)}{\partial x_1} \end{bmatrix} = \begin{bmatrix} 4x_1 + 10x_2 + 1 \\ 10x_1 + 6x_2 + 7 \end{bmatrix} = \begin{bmatrix} 4 & 10 \\ 10 & 6 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 1 \\ 7 \end{bmatrix} = 2Ax + b$$

$$\nabla_x^2 f(x) = \nabla_x\left(\begin{bmatrix} 4x_1 + 10x_2 + 1 \\ 10x_1 + 6x_2 + 7 \end{bmatrix}\right) = \begin{bmatrix} 4 & 10 \\ 10 & 6 \end{bmatrix} = 2A$$

2.5 Least Squares Problem (note: 2022 version has $m < n$)

The problem minimizes the following distance function $J : \mathbb{R}^n \to \mathbb{R}$, which corresponds to Euclidean distance squared:

$$\begin{aligned}
J(x) &= (Ax - b)^T(Ax - b) \\
&= (x^T A^T - b^T)(Ax - b) \\
&= x^T A^T Ax - x^T A^T b - b^T Ax + b^T b
\end{aligned}$$

The minimum point for this function can be found by setting its gradient to 0:

$$\min_x J(x) \rightarrow \frac{\partial J(x)}{\partial x} = 0$$

Note that $A^T A$ is symmetric. Using the derivations from (2.2), we can directly write the gradient as a vector:

$$\frac{\partial J(x)}{\partial x} = 2A^T Ax - A^T b - (b^T A)^T = 2A^T Ax - 2A^T b = 0$$

$$\implies A^T Ax = A^T b \implies \hat{x} = (A^T A)^{-1} A^T b$$

Note that since A has rank n, the $n \times n$ matrix $A^T A$ is invertible.