Linear Algebra Review

Ryan Miller

Grinnell College
Statistics

Definitions and Notation

Let i index rows and j index columns. A matrix is a rectangular array of numbers:

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

- We'll denote matrices using bold uppercase (ie: $\mathbf{X}$), vectors using bold lowercase (ie: $\mathbf{y}$), and scalars using unbolded lowercase (ie: c)
- We'll often describe the dimensions of matrices, here $\mathbf{A}$ has dimensions: 2×3
- Unless stated otherwise, vectors are matrices with a single column, for example $\mathbf{y}$ has dimensions $n \times 1$

Addition and Subtraction

There are two types of addition and subtraction to consider:

1. Scalar - Matrix: The scalar value is added/subtracted from *every element*

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

2. Matrix - Matrix: Values are added/subtracted *elementwise*

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

Multiplication

There are two types of multiplication:

1. Scalar - Matrix: Multiply every element by the scalar

$$3\begin{bmatrix}1 & 2\\ 3 & 4\end{bmatrix} = \begin{bmatrix}3 & 6\\ 9 & 12\end{bmatrix}$$

2. Matrix - Matrix: Rows of the left matrix are multiplied by columns of the right matrix

$$(\mathbf{AB})_{ij} = (\text{row } i \text{ of } \mathbf{A}) \cdot (\text{column } j \text{ of } \mathbf{B}) = \sum_{k=1}^{n} \mathbf{A}_{ik}\mathbf{B}_{kj}$$

Note: $(m \times n)(n \times p) = (m \times p)$, so the "inside" dimensions must match for multiplication to be possible.

Practice

Consider the data below:

$$\mathbf{A} = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}, \qquad \mathbf{B} = \begin{bmatrix} 5 & 6 \\ 7 & 8 \end{bmatrix}$$

1. What are the *dimensions* (number of rows and columns) of **AB**?
2. What is the element in position (2,1) of **AB**?

Transposes

Transposition takes the rows of a matrix and represents them as columns. Consider:

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

1. What are the dimensions of $\mathbf{A}^T$ and $\mathbf{A}^T\mathbf{A}$?
2. What is the element in position (2,1) of $\mathbf{A}^T$? Where was this element in $\mathbf{A}$?

Special Cases

1. Symmetric Matrices - $\mathbf{A} = \mathbf{A}^T$, so transposition doesn't change the matrix
2. Diagonal - a symmetric matrix with zeroes in all off-diagonal elements
3. Identity - a diagonal matrix with ones in all on-diagonal elements, denoted using $\mathbf{I}$. Note that $\mathbf{I} \cdot \mathbf{A} = \mathbf{A} \cdot \mathbf{I} = \mathbf{A}$

Inverses

In algebra we use division to undo multiplication, the matrix equivalent is **inversion**. Consider:

$$\mathbf{A} \cdot \mathbf{x} = \mathbf{b}$$

- To solve for $\mathbf{x}$ we can multiply both sides of the equation by $\mathbf{A}^{-1}$, the inverse of $\mathbf{A}$
- Not every matrix has an inverse
  - Only square matrices with linearly independent columns are invertible

Useful Matrix Properties

For compatible matrices $\mathbf{A}$ and $\mathbf{B}$:

$$(\mathbf{A}+\mathbf{B})^T = \mathbf{A}^T + \mathbf{B}^T$$

$$(\mathbf{AB})^T = \mathbf{B}^T\mathbf{A}^T$$

$$(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$$

$$(\mathbf{A}^T)^{-1} = (\mathbf{A}^{-1})^T$$

Transposes and inverses reverse the order of multiplication.

Practice

Suppose $\mathbf{M}$ is an $n \times p$ matrix and $\mathbf{a}$ is an $n \times 1$ vector consider.

1. Is $\mathbf{M}^T\mathbf{M}$ symmetric? Can you justify your answer?
2. What are the dimensions of $(\mathbf{M}^T\mathbf{M})^{-1}\mathbf{M}^T\mathbf{a}$?

More Practice

Suppose:

$$\mathbf{M}^T(\mathbf{a} - \mathbf{M}\mathbf{b}) = \mathbf{0}$$

Solve for $\mathbf{b}$

Gradients

Consider the parameter vector $\mathbf{w} = (w_1, w_2, w_3)$ and the function $f(\mathbf{w}) = 2\mathbf{w}^T\mathbf{w}$. Expanding:

$$f(\mathbf{w}) = 2(w_1^2 + w_2^2 + w_3^2).$$

The **gradient** is the vector of partial derivatives:

$$\nabla f(\mathbf{w}) = \begin{bmatrix} \frac{\partial f}{\partial w_1} \\ \frac{\partial f}{\partial w_2} \\ \frac{\partial f}{\partial w_3} \end{bmatrix}.$$

For this function:

$$\nabla f(\mathbf{w}) = \begin{bmatrix} 4w_1 \\ 4w_2 \\ 4w_3 \end{bmatrix} = 4\mathbf{w}.$$

Comments on Gradients

- The gradient provides information on how a function is changing, which we can use to find its minimum/maximum values
- If the gradient is positive, the tangent to the function is sloping upward, so we might travel in the negative direction to find lower values of the function
- Many machine learning algorithms repeatedly move in the direction $-\nabla f(\mathbf{w})$ to minimize a *cost function*

Practice

Let $\mathbf{w} = (w_1, w_2, w_3)$ and consider the function:

$$f(\mathbf{w}) = 3\mathbf{w}^T\mathbf{w} + 2\mathbf{w}^T \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$$

Find the gradient $\nabla f(\mathbf{w})$

Matrix Derivative Shortcuts

Several derivative results appear repeatedly in machine learning:

$$f(\mathbf{x}) = c\mathbf{x}^T\mathbf{x} \qquad \Longrightarrow \qquad \frac{df}{d\mathbf{x}} = 2c\mathbf{x}$$

$$f(\mathbf{x}) = c\mathbf{x}^T\mathbf{A}\mathbf{x} \qquad \Longrightarrow \qquad \frac{df}{d\mathbf{x}} = 2c\mathbf{A}\mathbf{x}$$

$$f(\mathbf{x}) = c\mathbf{x}^T\mathbf{a} \qquad \Longrightarrow \qquad \frac{df}{d\mathbf{x}} = c\mathbf{a}$$

We will use these results repeatedly rather than deriving gradients from scratch every time.

Jacobians

When $f()$ outputs a single number the gradient is a vector with one element per parameter. A **Jacobian** is used when differentiating a function whose output is a vector:

$$f(\mathbf{w}) = 2\mathbf{w} = \begin{bmatrix} 2w_1 \\ 2w_2 \\ 2w_3 \end{bmatrix}$$

The Jacobian is the matrix of all partial derivatives:

$$\frac{\partial f}{\partial \mathbf{w}} = \begin{bmatrix} \frac{\partial f_1}{\partial w_1} & \frac{\partial f_1}{\partial w_2} & \frac{\partial f_1}{\partial w_3} \\ \frac{\partial f_2}{\partial w_1} & \frac{\partial f_2}{\partial w_2} & \frac{\partial f_2}{\partial w_3} \\ \frac{\partial f_3}{\partial w_1} & \frac{\partial f_3}{\partial w_2} & \frac{\partial f_3}{\partial w_3} \end{bmatrix} = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{bmatrix} = 2\mathbf{I}$$

Practice

Consider:

$$f(\mathbf{v}) = \exp(\mathbf{v}) = \begin{bmatrix} e^{v_1} \\ e^{v_2} \\ \vdots \\ e^{v_n} \end{bmatrix}$$

Show the Jacobian is:

$$\frac{\partial f}{\partial \mathbf{v}} = \begin{bmatrix} e^{v_1} & 0 & \cdots & 0 \\ 0 & e^{v_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & e^{v_n} \end{bmatrix} = \text{diag}(e^{\mathbf{v}}).$$

Chain Rule

Now consider:

$$\mathbf{v} = \mathbf{X}\mathbf{w} \text{ and } f(\mathbf{w}) = \exp(\mathbf{v})$$

Rather than expanding everything, we can differentiate in stages:

$$\frac{\partial f}{\partial \mathbf{w}} = \frac{\partial f}{\partial \mathbf{v}} \frac{\partial \mathbf{v}}{\partial \mathbf{w}}$$

This is the matrix version of *chain rule*.

Chain Rule (cont.)

In this example:

$$\frac{\partial \mathbf{v}}{\partial \mathbf{w}} = \mathbf{X} \text{ and } \frac{\partial f}{\partial \mathbf{v}} = \text{diag}(e^{\mathbf{v}})$$

Thus:

$$\frac{\partial f}{\partial \mathbf{w}} = \text{diag}(e^{\mathbf{v}})\mathbf{X}$$

Practice

Suppose:

$$f(\mathbf{w}) = (\mathbf{y} - \mathbf{X}\mathbf{w})^T (\mathbf{y} - \mathbf{X}\mathbf{w})$$

where $\mathbf{y}$ is an $n \times 1$ vector, $\mathbf{X}$ is an $n \times p$ matrix, and $\mathbf{w}$ is a $p \times 1$ vector.

Find $\frac{\partial f}{\partial \mathbf{w}}$ using chain rule. *Hint*: Introduce a new variable $\mathbf{z}$.