Lecture 7: Matrix Calculus

Mark Hasegawa-Johnson

ECE 417: Multimedia Signal Processing, Fall 2023
Much of this lecture is based on
`https://en.wikipedia.org/wiki/Matrix_calculus`

1. Notation

2. Types of Derivatives

3. Derivatives with Vectors

4. Derivatives with Matrices

5. Conclusions

Outline

1. Notation
2. Types of Derivatives
3. Derivatives with Vectors
4. Derivatives with Matrices
5. Conclusions

Notation

- x - a scalar
- $\mathbf{x} \in \Re^m$ - a column vector
- $\mathbf{x}^T \in \Re^n$ - a row vector
- $\mathbf{X} \in \Re^{m \times n}$ - a matrix

Trace and Determinant

- The trace of a matrix is the sum of its diagonal elements:

$$\operatorname{tr}(\mathbf{X}) = \sum_i x_{i,i}$$

- The determinant of an $n \times n$ matrix is

$$\begin{aligned} |\mathbf{X}| &= \sum_{j=1}^{n} (-1)^{i+j} x_{i,j} \left|\mathbf{X}_{\neg i,\neg j}\right| \quad \forall 1 \le i \le n \\ &= \sum_{i=1}^{n} (-1)^{i+j} x_{i,j} \left|\mathbf{X}_{\neg i,\neg j}\right| \quad \forall 1 \le j \le n \end{aligned}$$

where $\mathbf{X}_{\neg i,\neg j}$ is the submatrix computed by removing the i^{th} row and j^{th} column. The determinant may also be written as the product of the eigenvalues:

$$|\mathbf{X}| = \prod_{i=1}^{n} \lambda_i(\mathbf{X})$$

Norms of Vectors

- The Lp norm of a vector is

$$\|\mathbf{x}\|_p = \left(x_1^p + x_2^p + \cdots + x_n^p\right)^{1/p}$$

- If the subscript is omitted, you may assume the L2 norm, a.k.a. the Euclidean norm:

$$\|\mathbf{x}\| = \|\mathbf{x}\|_2 = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2}$$

- The Euclidean norm can also be calculated as the square root of the dot product of **x** with itself:

$$\|\mathbf{x}\| = \sqrt{\mathbf{x}^T\mathbf{x}}$$

Norms of Matrices

- The Frobenius norm of a matrix is the generalization of a Euclidean norm:

$$\|\mathbf{X}\|_F = \sqrt{\sum_i \sum_j |x_{i,j}|^2}$$

It can be written in an interesting way:

$$\|\mathbf{X}\|_F = \sqrt{\mathrm{tr}(\mathbf{X}^T\mathbf{X})}$$

- The L^p norm of a matrix is

$$\|\mathbf{X}\|_p = \sup_{\mathbf{v}} \frac{\|\mathbf{X}\mathbf{v}\|_p}{\|\mathbf{v}\|_p}$$

These norms have cool mathematical properties, but they are not as useful in practice, usually, as the Frobenius norm.

Outline

1. Notation
2. Types of Derivatives
3. Derivatives with Vectors
4. Derivatives with Matrices
5. Conclusions

Types of derivatives

Let's talk about six types of derivatives:

		Numerator		
		Scalar	Vector	Matrix
Denominator	Scalar	$\frac{\partial y}{\partial x}$	$\frac{\partial \mathbf{y}}{\partial x}$	$\frac{\partial \mathbf{Y}}{\partial x}$
	Vector	$\frac{\partial y}{\partial \mathbf{x}}$	$\frac{\partial \mathbf{y}}{\partial \mathbf{x}}$	
	Matrix	$\frac{\partial y}{\partial \mathbf{X}}$		

Outline

1. Notation
2. Types of Derivatives
3. Derivatives with Vectors
4. Derivatives with Matrices
5. Conclusions

Vector-by-Scalar: the Tangent Vector

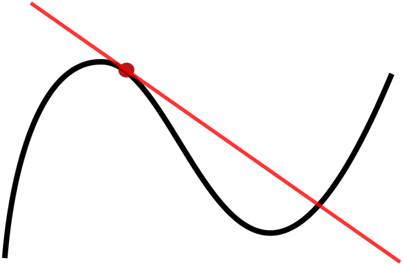
$\frac{\partial \mathbf{y}}{\partial x}$ is called the tangent vector:

$$\frac{\partial \mathbf{y}}{\partial x} = \begin{bmatrix} \frac{\partial y_1}{\partial x} \\ \vdots \\ \frac{\partial y_m}{\partial x} \end{bmatrix}$$

Tangent Vector

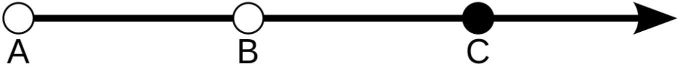
- Suppose that x is scalar, and $\mathbf{y} = [y_1, y_2]^T$ is a point in a vector space.
- The function $\mathbf{y}(x) = [y_1(x), y_2(x)]^T$ sketches a curve in that space.
- The tangent vector is

$$\frac{\partial \mathbf{y}}{\partial x} = \begin{bmatrix} \frac{\partial y_1}{\partial x} \\ \frac{\partial y_2}{\partial x} \end{bmatrix}$$



https://commons.wikimedia.org/wiki/File:Tangent_to_a_curve.svg

Tangent Vector, Example #1: Line



Here is the equation for a line:

$$\mathbf{y}(x) = \mathbf{a}x = \begin{bmatrix} a_1 x \\ a_2 x \end{bmatrix}$$

...and here is its tangent vector:

$$\frac{\partial \mathbf{y}}{\partial x} = \mathbf{a}$$

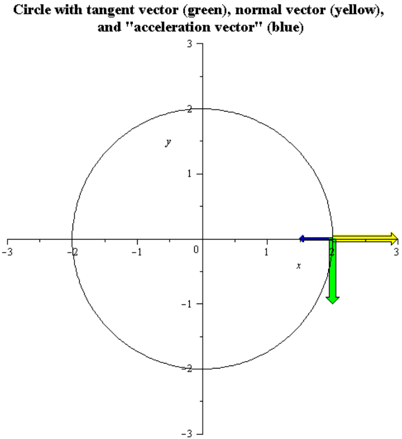
Tangent Vector, Example #2: Circle

Here is the equation for a circle:

$$\mathbf{y}(\theta) = \begin{bmatrix} r\cos\theta \\ r\sin\theta \end{bmatrix}$$

. . . and here is its tangent vector:

$$\frac{\partial \mathbf{y}}{\partial \theta} = \begin{bmatrix} -r\sin\theta \\ r\cos\theta \end{bmatrix}$$


Circle with tangent vector (green), normal vector (yellow), and "acceleration vector" (blue)

https://commons.wikimedia.org/wiki/File:Circle%2B3vectors_animated.gif

Rectifiable Curve

Suppose that x is scalar, and $\mathbf{y}(x) = [y_1(x), y_2(x)]^T$ is a curve. If $\frac{\partial \mathbf{y}}{\partial x}$ exists and is finite, then it is possible to calculate the length of the curve by integrating

$$\int \|\frac{\partial \mathbf{y}}{\partial x}\| dx$$

https://upload.wikimedia.org/wikipedia/commons/d/dc/Arc_length.gif

Tangent Vector, Example #2: Circle

If $\mathbf{y}(\theta) = [r\cos\theta, r\sin\theta]^T$, its tangent is

$$\frac{\partial \mathbf{y}}{\partial \theta} = \begin{bmatrix} -r\sin\theta \\ r\cos\theta \end{bmatrix}$$

The circumference of the circle is

$$\begin{aligned} c &= \int_{-\pi}^{\pi} \|\frac{\partial \mathbf{y}}{\partial \theta}\| d\theta \\ &= \int_{-\pi}^{\pi} \sqrt{(-r\sin\theta)^2 + (r\cos\theta)^2 d\theta} \\ &= \int_{-\pi}^{\pi} r d\theta = 2\pi r \end{aligned}$$

Scalar-by-Vector: the Gradient

If y is a scalar function of a vector $\mathbf{x}$, then ∇y is called the **gradient**:

$$\nabla y = \begin{bmatrix} \frac{\partial y}{\partial x_1} \\ \vdots \\ \frac{\partial y}{\partial x_m} \end{bmatrix}$$

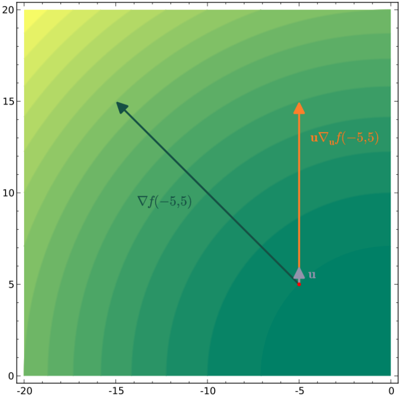
Since $\frac{\partial \mathbf{y}}{\partial x}$ was a column vector, we will define $\frac{\partial y}{\partial \mathbf{x}}$ to be a row vector. This is called "numerator layout notation," and will have some benefits later on.

$$\frac{\partial y}{\partial \mathbf{x}} = \nabla y^T = \left[\frac{\partial y}{\partial x_1}, \ldots, \frac{\partial y}{\partial x_m}\right]$$

Directional Derivative

Suppose $y(\mathbf{x})$ is a scalar function of a vector $\mathbf{x} = [x_1, x_2]^T$. If $\mathbf{u}$ is any unit vector, then the **directional derivative** of $y(\mathbf{x})$ in the $\mathbf{u}$ direction is written as

$$\begin{aligned}\nabla_{\mathbf{u}} y &= \nabla y^T \mathbf{u} \\ &= \frac{\partial y}{\partial \mathbf{x}} \mathbf{u} \\ &= \frac{\partial y}{\partial x_1} u_1 + \frac{\partial y}{\partial x_2} u_2\end{aligned}$$



https://commons.wikimedia.org/wiki/File:Directional_derivative_contour_plot.svg

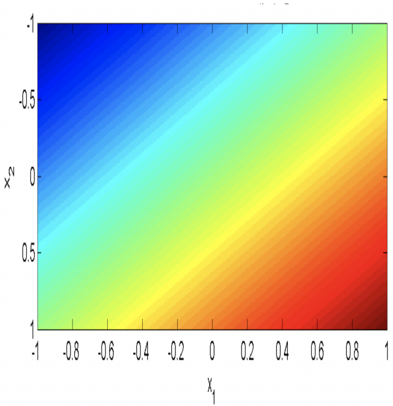
Gradient Example #1: Affine Function

An affine function of $\mathbf{x}$ is written:

$$
\begin{aligned}
y(\mathbf{x}) &= \mathbf{a}^T\mathbf{x} + b \\
&= a_1x_1 + a_2x_2 + b
\end{aligned}
$$

Its gradient is

$$
\frac{\partial y}{\partial \mathbf{x}} = \mathbf{a}^T
$$



Gradient Example #2: Euclidean Norm

Consider the Euclidean norm of a vector:

$$\|\mathbf{x}\|^2 = \mathbf{x}^T\mathbf{x} = x_1^2 + \ldots + x_n^2$$

By analogy to the affine function, setting $\mathbf{a}^T = \mathbf{x}^T$, we might think that $\frac{\partial \mathbf{x}^T\mathbf{x}}{\partial \mathbf{x}} = \mathbf{x}^T$. But from basic algebra, we know that

$$\begin{aligned}\frac{\partial(x_1^2 + \ldots + x_n^2)}{\partial \mathbf{x}} &= \left[\frac{\partial(x_1^2 + \ldots + x_n^2)}{\partial x_1}, \ldots, \frac{\partial(x_1^2 + \ldots + x_n^2)}{\partial x_n}\right] \\ &= [2x_1, \ldots, 2x_n] \\ &= 2\mathbf{x}^T\end{aligned}$$

What went wrong?

The Stop-Gradient Approach

One way to approach the problem of $\frac{\partial \mathbf{x}^T\mathbf{x}}{\partial \mathbf{x}}$ is the **stop-gradient** approach:

1. First, pretend that $\mathbf{x}^T$ is a constant vector that does not change when $\mathbf{x}$ changes — in other words, stop the gradient w.r.t. $\mathbf{x}^T$. We can write this as
$$\frac{\partial\, \text{sg}(\mathbf{x}^T)\mathbf{x}}{\partial \mathbf{x}} = \mathbf{x}^T$$
2. Second, take the derivative w.r.t. $\mathbf{x}^T$ while stopping the gradient w.r.t. $\mathbf{x}$:
$$\frac{\partial \mathbf{x}^T\, \text{sg}(\mathbf{x})}{\partial \mathbf{x}^T} = \mathbf{x}$$
3. Finally, realize that when $\mathbf{x}$ changes, $\mathbf{x}^T$ also changes, so we need to include both parts of the derivative:

$$\frac{\partial \mathbf{x}^T\mathbf{x}}{\partial \mathbf{x}} = \frac{\partial\, \text{sg}(\mathbf{x}^T)\mathbf{x}}{\partial \mathbf{x}} + \left(\frac{\partial \mathbf{x}^T\, \text{sg}(\mathbf{x})}{\partial \mathbf{x}^T}\right)^T = 2\mathbf{x}^T$$

Gradient Example #3: Linear Regression

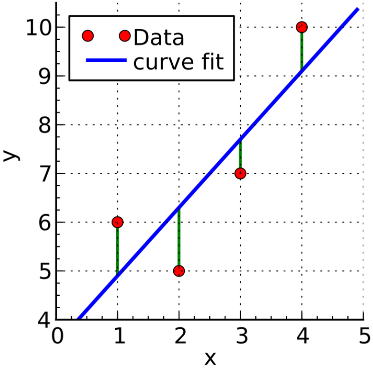
Linear regression is the problem of approximating y_i as a linear function of $\mathbf{x}_i$:

$$y_i \approx \mathbf{a}^T\mathbf{x}_i, \quad 1 \le i \le n$$

The approximation is computed by minimizing the mean-squared error:

$$\mathcal{L} = \sum_{i=1}^{n}\left(y_i - \mathbf{x}_i^T\mathbf{a}\right)^2$$

$$= \sum_{i=1}^{n}\left(y_i^2 - y_i\mathbf{a}^T\mathbf{x}_i - y_i\mathbf{x}_i^T\mathbf{a} + \mathbf{a}^T\mathbf{x}_i\mathbf{x}_i^T\mathbf{a}\right)$$



`https://commons.wikimedia.org/wiki/File:Linear_least_squares_example2.svg`

Gradient Example #3: Linear Regression

$$\mathcal{L} = \sum_{i=1}^{n} \left(y_i^2 - y_i \mathbf{a}^T \mathbf{x}_i - y_i \mathbf{x}_i^T \mathbf{a} + \mathbf{a}^T \mathbf{x}_i \mathbf{x}_i^T \mathbf{a} \right)$$

We can solve for $\mathbf{a}$ by finding the derivative, $\frac{\partial \mathcal{L}}{\partial \mathbf{a}}$, and setting it equal to zero:

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \mathbf{a}} &= \frac{\partial \mathcal{L}(\mathbf{a}, \mathrm{sg}(\mathbf{a}^T))}{\partial \mathbf{a}} + \left(\frac{\partial \mathcal{L}(\mathrm{sg}(\mathbf{a}), \mathbf{a}^T)}{\partial \mathbf{a}^T} \right)^T \\
&= \sum_{i=1}^{n} -2 y_i \mathbf{x}_i^T + 2 \mathbf{a}^T \mathbf{x}_i \mathbf{x}_i^T & = 0 \\
\mathbf{a} &= \left(\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T \right)^{-1} \left(\sum_{i=1}^{n} \mathbf{x}_i y_i \right)
\end{aligned}$$

Vector-by-Vector: the Jacobian

If $\mathbf{y}(\mathbf{x}) \in \Re^m$ is a vector function of a vector $\mathbf{x} \in \Re^n$, then $\frac{\partial \mathbf{y}}{\partial \mathbf{x}}$ is called the Jacobian.

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \cdots & \frac{\partial y_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial y_m}{\partial x_1} & \cdots & \frac{\partial y_m}{\partial x_n} \end{bmatrix}$$

Jacobian Example: Affine Transformation

An affine transformation of a vector $\mathbf{x}$ is written as:

$$\begin{aligned} \mathbf{y}(\mathbf{x}) &= \mathbf{A}\mathbf{x} + \mathbf{b} \\ &= \begin{bmatrix} a_{1,1}x_1 + a_{1,2}x_2 + b_1 \\ a_{2,1}x_1 + a_{2,2}x_2 + b_2 \end{bmatrix} \end{aligned}$$

Its Jacobian is:

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \mathbf{A}$$

Chain Rule

The Jacobian is most useful because we can use it in the chain rule. Suppose that $\mathbf{z} \in \Re^p$ is a function of $\mathbf{y} \in \Re^n$, and $\mathbf{y}$ is a function of $\mathbf{x} \in \Re^m$. Then

$$\frac{\partial \mathbf{z}}{\partial \mathbf{x}} = \frac{\partial \mathbf{z}}{\partial \mathbf{y}} \frac{\partial \mathbf{y}}{\partial \mathbf{x}}$$

$$= \left[\begin{array}{ccc} \frac{\partial z_1}{\partial y_1} & \cdots & \frac{\partial z_1}{\partial y_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial z_p}{\partial y_1} & \cdots & \frac{\partial z_p}{\partial y_n} \end{array}\right] \left[\begin{array}{ccc} \frac{\partial y_1}{\partial x_1} & \cdots & \frac{\partial y_1}{\partial x_m} \\ \vdots & \ddots & \vdots \\ \frac{\partial y_n}{\partial x_1} & \cdots & \frac{\partial y_n}{\partial x_m} \end{array}\right] = \left[\begin{array}{ccc} \frac{\partial z_1}{\partial x_1} & \cdots & \frac{\partial z_1}{\partial x_m} \\ \vdots & \ddots & \vdots \\ \frac{\partial z_p}{\partial x_1} & \cdots & \frac{\partial z_p}{\partial x_m} \end{array}\right]$$

Summary: Derivatives with Vectors

$$\frac{\partial \mathbf{a} x}{\partial x} = \mathbf{a}$$

$$\frac{\partial \mathbf{a}^T \mathbf{x}}{\partial \mathbf{x}} = \mathbf{a}^T$$

$$\frac{\partial \mathbf{x}^T \mathbf{x}}{\partial \mathbf{x}} = 2\mathbf{x}^T$$

$$\frac{\partial (y_i - \mathbf{a}^T \mathbf{x}_i)^2}{\partial \mathbf{a}} = 2(\mathbf{a}^T \mathbf{x}_i - y_i)\mathbf{x}_i^T$$

$$\frac{\partial \mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = \mathbf{A}$$

Outline

1. Notation
2. Types of Derivatives
3. Derivatives with Vectors
4. Derivatives with Matrices
5. Conclusions

Matrix-by-Scalar: the Tangent Matrix

If $\mathbf{Y}(x) \in \Re^{m \times n}$ is a matrix function of a scalar x, we can compute a tangent matrix, $\frac{\partial \mathbf{Y}}{\partial x}$:

$$\frac{\partial \mathbf{Y}}{\partial x} = \begin{bmatrix} \frac{\partial y_{1,1}}{\partial x} & \cdots & \frac{\partial y_{1,n}}{\partial x} \\ \vdots & \ddots & \vdots \\ \frac{\partial y_{m,1}}{\partial x} & \cdots & \frac{\partial y_{m,n}}{\partial x} \end{bmatrix}$$

Tangent Matrix Example #1: Linear Scaling

Suppose that

$$\begin{aligned}\mathbf{Y}(x) &= \mathbf{A}x \\ &= \begin{bmatrix} a_{1,1}x & \cdots & a_{1,n}x \\ \vdots & \ddots & \vdots \\ a_{m,1}x & \cdots & a_{m,n}x \end{bmatrix}\end{aligned}$$

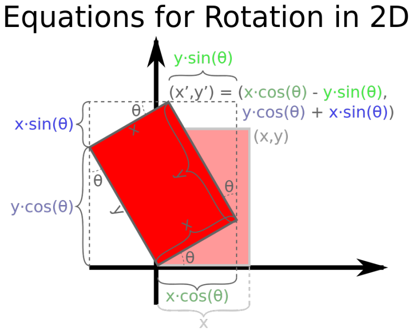
Then its tangent matrix is

$$\frac{\partial \mathbf{Y}}{\partial x} = \mathbf{A}$$

Tangent Matrix Example #2: Rotation Matrix

The rotation matrix, $\mathbf{T}(\theta)$, is the matrix that takes any input vector and rotates it by θ radians:

$$\mathbf{T}(\theta) = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$$



`https://commons.wikimedia.org/wiki/File:Visual_Derivation_of_Equations_For_Rotation_In_2D.svg`

Tangent Matrix Example #2: Rotation Matrix

The rotation matrix and its tangent are:

$$\mathbf{T}(\theta) = \left[\begin{array}{cc} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{array}\right], \qquad \frac{\partial \mathbf{T}}{\partial \theta} = \left[\begin{array}{cc} -\sin\theta & -\cos\theta \\ \cos\theta & -\sin\theta \end{array}\right]$$

One thing we can do with this is, by simple linearity, for any vector $\mathbf{v}(\theta) = \mathbf{T}(\theta)\mathbf{u}$, for a fixed initial $\mathbf{u}$, we can show that

$$\frac{\partial \mathbf{v}}{\partial \theta} = \frac{\partial \mathbf{T}(\theta)\mathbf{u}}{\partial \theta} = \left(\frac{\partial \mathbf{T}}{\partial \theta}\right)\mathbf{u}$$

Scalar-by-Matrix: the Gradient Matrix

$\frac{\partial y}{\partial \mathbf{X}}$ has no universally accepted name, but in machine learning it is often called the gradient matrix, in analogy with the gradient vector. In **numerator layout** order, $\mathbf{X}$ and $\partial y/\partial \mathbf{X}$ are defined as:

$$\mathbf{X} = \begin{bmatrix} x_{1,1} & \cdots & x_{1,n} \\ \vdots & \ddots & \vdots \\ x_{m,1} & \cdots & x_{m,n} \end{bmatrix}, \qquad \frac{\partial y}{\partial \mathbf{X}} = \begin{bmatrix} \frac{\partial y}{\partial x_{1,1}} & \cdots & \frac{\partial y}{\partial x_{m,1}} \\ \vdots & \ddots & \vdots \\ \frac{\partial y}{\partial x_{1,n}} & \cdots & \frac{\partial y}{\partial x_{m,n}} \end{bmatrix}$$

Gradient Matrix Example #1: Trace of X

For example, suppose $y = \text{tr}(\mathbf{X})$:

$$y(\mathbf{X}) = \text{tr}(\mathbf{X}) = x_{1,1} + x_{2,2} + \cdots$$

The gradient of the trace of a matrix is:

$$\frac{\partial y}{\partial \mathbf{X}} = \begin{bmatrix} \frac{\partial y}{\partial x_{1,1}} & \frac{\partial y}{\partial x_{2,1}} & \cdots \\ \frac{\partial y}{\partial x_{1,2}} & \frac{\partial y}{\partial x_{2,2}} & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix} = \begin{bmatrix} 1 & 0 & \cdots \\ 0 & 1 & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix}$$

In other words,

$$\frac{\partial\, \text{tr}(\mathbf{X})}{\partial \mathbf{X}} = I$$

Gradient Matrix Example #1: Trace of AX

Now suppose $y = \mathrm{tr}(\mathbf{AX})$, where $\mathbf{A} \in \Re^{n\times m}$ and $\mathbf{X} \in \Re^{m\times n}$. The $(i,j)^{\text{th}}$ element of the matrix $\mathbf{C} = \mathbf{AX}$ is $c_{i,j} = \sum_{k=1}^{m} a_{i,k}x_{k,j}$. The trace is the sum along the main diagonal, so

$$y(\mathbf{X}) = \mathrm{tr}(\mathbf{AX}) = \sum_{i=1}^{n} c_{i,i} = \sum_{i=1}^{n}\sum_{k=1}^{m} a_{i,k}x_{k,i}$$

The gradient of the trace of a matrix is:

$$\frac{\partial y}{\partial \mathbf{X}} = \left[\begin{array}{ccc} \frac{\partial y}{\partial x_{1,1}} & \cdots & \frac{\partial y}{\partial x_{m,1}} \\ \vdots & \ddots & \vdots \\ \frac{\partial y}{\partial x_{1,n}} & \cdots & \frac{\partial y}{\partial x_{m,n}} \end{array}\right] = \left[\begin{array}{ccc} a_{1,1} & \cdots & a_{1,m} \\ \vdots & \ddots & \vdots \\ a_{n,1} & \cdots & a_{n,m} \end{array}\right] = \mathbf{A}$$

In other words,

$$\frac{\partial\, \mathrm{tr}(\mathbf{AX})}{\partial \mathbf{X}} = \mathbf{A}$$

Gradient Matrix Example #2: Pre- and Post-Multiplication

Suppose we pre-multiply by some vector $\mathbf{u}$, and post-multiply by some other vector $\mathbf{v}$:

$$y(\mathbf{X}) = \mathbf{u}^T\mathbf{X}\mathbf{v} = \sum_{i=1}^{m}\sum_{j=1}^{n} u_i x_{i,j} v_j$$

Then the gradient is:

$$\frac{\partial y}{\partial \mathbf{X}} = \begin{bmatrix} \frac{\partial y}{\partial x_{1,1}} & \cdots & \frac{\partial y}{\partial x_{m,1}} \\ \vdots & \ddots & \vdots \\ \frac{\partial y}{\partial x_{1,n}} & \cdots & \frac{\partial y}{\partial x_{m,n}} \end{bmatrix} = \begin{bmatrix} u_1 v_1 & \cdots & u_m v_1 \\ \vdots & \ddots & \vdots \\ u_1 v_n & \cdots & u_m v_n \end{bmatrix} = \mathbf{v}\mathbf{u}^T$$

So the gradient of $\mathbf{u}^T\mathbf{X}\mathbf{v}$ is $\mathbf{v}\mathbf{u}^T$? Why does that make sense?

The Trace Equality

It's time to introduce one more fundamental fact about linear algebra, called **the trace equality**. For any compatibly-sized matrices $\mathbf{A} \in \Re^{m \times n}$ and $\mathbf{B} \in \Re^{n \times m}$,

$$\text{tr}(\mathbf{AB}) = \text{tr}(\mathbf{BA})$$

The Trace Equality

$$\mathrm{tr}(\mathbf{AB}) = \mathrm{tr}(\mathbf{BA})$$

Proof:

- The $(i,j)^{\text{th}}$ element of the matrix $\mathbf{C} = \mathbf{AB}$ is $c_{i,j} = \sum_{k=1}^{n} a_{i,k} b_{k,j}$. The trace sums the main diagonal, so

$$\mathrm{tr}(\mathbf{AB}) = \sum_{i=1}^{m} c_{i,i} = \sum_{i=1}^{m} \sum_{k=1}^{n} a_{i,k} b_{k,i}$$

- The $(j,k)^{\text{th}}$ element of the matrix $\mathbf{D} = \mathbf{BA}$ is $d_{j,k} = \sum_{i=1}^{m} b_{j,i} a_{i,k}$. The trace sums the main diagonal, so

$$\mathrm{tr}(\mathbf{BA}) = \sum_{k=1}^{m} d_{k,k} = \sum_{k=1}^{n} \sum_{i=1}^{m} b_{k,i} a_{i,k}$$

- Those two things are the same.

Gradient Matrix Example #2: Pre- and Post-Multiplication

- $\mathbf{u}^T\mathbf{X}\mathbf{v}$ is a scalar, so it is its own trace:

$$y(\mathbf{X}) = \mathbf{u}^T\mathbf{X}\mathbf{v} = \mathrm{tr}\left(\mathbf{u}^T\mathbf{X}\mathbf{v}\right)$$

- By the trace equality,

$$y(\mathbf{X}) = \mathrm{tr}\left(\mathbf{u}^T\mathbf{X}\mathbf{v}\right) = \mathrm{tr}\left(\mathbf{v}\mathbf{u}^T\mathbf{X}\right)$$

- So the gradient is:

$$\frac{\partial y}{\partial \mathbf{X}} = \frac{\partial\, \mathrm{tr}\left(\mathbf{v}\mathbf{u}^T\mathbf{X}\right)}{\partial \mathbf{X}} = \mathbf{v}\mathbf{u}^T$$

Gradient Matrix Example #3: Frobenius Norm Squared

There are several possible extensions of Euclidean norms to matrices, of which the Frobenius norm is the most useful. The Frobenius norm squared is just the sum of the squares of all elements of the matrix:

$$\|\mathbf{X}\|_F^2 = \sum_{i=1}^{m} \sum_{j=1}^{n} x_{i,j}^2$$

From the definition of a matrix gradient, it's pretty obvious that

$$\frac{\partial \|\mathbf{X}\|_F^2}{\partial \mathbf{X}} == \left[\begin{array}{ccc} 2x_{1,1} & \cdots & 2x_{m,1} \\ \vdots & \ddots & \vdots \\ 2x_{1,n} & \cdots & 2x_{m,n} \end{array} \right] = 2\mathbf{X}^T$$

Gradient Matrix Example #3: Frobenius Norm Squared

If you remember, the trace of $\mathbf{AX}$ is

$$\text{tr}(\mathbf{AX}) = \sum_{i=1}^{n}\sum_{k=1}^{m} a_{i,k}x_{k,i}$$

If we choose $\mathbf{A} = \mathbf{X}^T$, then $a_{i,k} = x_{k,i}$, and therefore

$$\text{tr}(\mathbf{X}\mathbf{X}^T) = \text{tr}(\mathbf{X}^T\mathbf{X}) = \sum_{i=1}^{n}\sum_{k=1}^{m} x_{k,i}^2 = \|\mathbf{X}\|_F^2$$

The gradient is:

$$\frac{\partial\,\text{tr}(\mathbf{X}^T\mathbf{X})}{\partial\mathbf{X}} = \frac{\partial\,\text{tr}(\text{sg}(\mathbf{X}^T)\mathbf{X})}{\partial\mathbf{X}} + \left(\frac{\partial\,\text{tr}(\mathbf{X}^T\,\text{sg}(\mathbf{X}))}{\partial\mathbf{X}^T}\right)^T = 2\mathbf{X}^T,$$

which is the same as the answer on the previous slide!

Gradient Matrix Example #4: Multiple Linear Regression

Multiple linear regression is the problem of approximating a vector output, $\mathbf{y}_i$, as a linear function of $\mathbf{x}_i$:

$$\mathbf{y}_i \approx \mathbf{A}^T \mathbf{x}_i, \quad 1 \le i \le n$$

It's useful to create **data matrices,** $\mathbf{X}$ and $\mathbf{Y}$, defined as

$$\mathbf{Y} = \begin{bmatrix} \mathbf{y}_1^T \\ \vdots \\ \mathbf{y}_n^T \end{bmatrix}, \qquad \mathbf{X} = \begin{bmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_n^T \end{bmatrix}$$

Then the multiple linear regression problem is to find $\mathbf{A}$ such that

$$\mathbf{Y} \approx \mathbf{X}\mathbf{A}$$

Gradient Matrix Example #4: Multiple Linear Regression

$$\mathbf{Y} \approx \mathbf{X}\mathbf{A}$$

The approximation is computed by minimizing the mean-squared error:

$$\begin{aligned}
\mathcal{L} &= \sum_{i=1}^{n} \|\mathbf{y}_i - \mathbf{A}^T\mathbf{x}_i\|_2^2 \\
&= \|\mathbf{Y} - \mathbf{X}\mathbf{A}\|_F^2 \\
&= \mathrm{tr}\left((\mathbf{Y} - \mathbf{X}\mathbf{A})^T(\mathbf{Y} - \mathbf{X}\mathbf{A})\right)
\end{aligned}$$

Gradient Matrix Example #4: Multiple Linear Regression

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \mathbf{A}} &= \frac{\partial\,\mathrm{tr}(\mathbf{Y}^T\mathbf{Y})}{\partial \mathbf{A}} - \left(\frac{\partial\,\mathrm{tr}(\mathbf{A}^T\mathbf{X}^T\mathbf{Y})}{\partial \mathbf{A}^T}\right)^T \\ &\quad - \frac{\partial\,\mathrm{tr}(\mathbf{Y}^T\mathbf{X}\mathbf{A})}{\partial \mathbf{A}} + \frac{\partial\,\mathrm{tr}(\mathbf{A}^T\mathbf{X}^T\mathbf{X}\mathbf{A})}{\partial \mathbf{A}} \\ &= 0 - \mathbf{Y}^T\mathbf{X} - \mathbf{Y}^T\mathbf{X} + 2\mathbf{A}^T\mathbf{X}^T\mathbf{X}\end{aligned}$$

Setting $\frac{\partial \mathcal{L}}{\partial \mathbf{A}} = \mathbf{0}$ (a matrix of all zeros) and solving for $\mathbf{A}$ gives

$$\begin{aligned}\mathbf{A} &= \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y} \\ &= \mathbf{X}^{\dagger}\mathbf{Y}\end{aligned}$$

Summary: Derivatives with Matrices

$$\begin{aligned}
\frac{\partial \mathbf{A}x}{\partial x} &= \mathbf{A} \\
\frac{\partial\, \mathrm{tr}(\mathbf{X})}{\partial \mathbf{X}} &= I \\
\frac{\partial\, \mathrm{tr}(\mathbf{A}\mathbf{X})}{\partial \mathbf{X}} &= \mathbf{A} \\
\frac{\partial\, \mathrm{tr}(\mathbf{X}^T\mathbf{X})}{\partial \mathbf{X}} &= 2\mathbf{X}^T \\
\frac{\partial\, \mathrm{tr}((\mathbf{Y}-\mathbf{X}\mathbf{A})^T(\mathbf{Y}-\mathbf{X}\mathbf{A}))}{\partial \mathbf{X}} &= -2\left(\mathbf{Y}-\mathbf{X}\mathbf{A}\right)^T\mathbf{X}
\end{aligned}$$

Outline

1. Notation
2. Types of Derivatives
3. Derivatives with Vectors
4. Derivatives with Matrices
5. Conclusions

Types of derivatives

		Numerator		
		Scalar	Vector	Matrix
Denominator	Scalar	$\frac{\partial y}{\partial x}$	$\frac{\partial \mathbf{y}}{\partial x}$	$\frac{\partial \mathbf{Y}}{\partial x}$
	Vector	$\frac{\partial y}{\partial \mathbf{x}}$	$\frac{\partial \mathbf{y}}{\partial \mathbf{x}}$	
	Matrix	$\frac{\partial y}{\partial \mathbf{X}}$		

Vector-by-Scalar: the Tangent Vector

$\frac{\partial \mathbf{y}}{\partial x}$ is called the tangent vector:

$$\frac{\partial \mathbf{y}}{\partial x} = \begin{bmatrix} \frac{\partial y_1}{\partial x} \\ \vdots \\ \frac{\partial y_m}{\partial x} \end{bmatrix}$$

Scalar-by-Vector: the Gradient

If y is a scalar function of a vector $\mathbf{x}$, then $\frac{\partial y}{\partial \mathbf{x}}^T$ is called the gradient.

$$\begin{aligned}\frac{\partial y}{\partial \mathbf{x}} &= \nabla y^T \\ &= \left[\frac{\partial y}{\partial x_1}, \ldots, \frac{\partial y}{\partial x_m}\right]\end{aligned}$$

Vector-by-Vector: the Jacobian

If $\mathbf{y}(\mathbf{x}) \in \Re^m$ is a vector function of a vector $\mathbf{x} \in \Re^n$, then $\frac{\partial \mathbf{y}}{\partial \mathbf{x}}$ is called the Jacobian.

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \cdots & \frac{\partial y_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial y_m}{\partial x_1} & \cdots & \frac{\partial y_m}{\partial x_n} \end{bmatrix}$$

Conclusions: Derivatives with Vectors

$$
\begin{aligned}
\frac{\partial \mathbf{a} x}{\partial x} &= \mathbf{a} \\
\frac{\partial \mathbf{a}^T \mathbf{x}}{\partial \mathbf{x}} &= \mathbf{a}^T \\
\frac{\partial \mathbf{x}^T \mathbf{x}}{\partial \mathbf{x}} &= 2\mathbf{x}^T \\
\frac{\partial (y_i - \mathbf{a}^T \mathbf{x}_i)^2}{\partial \mathbf{a}} &= 2(\mathbf{a}^T \mathbf{x}_i - y_i)\mathbf{x}_i^T \\
\frac{\partial \mathbf{A}\mathbf{x}}{\partial \mathbf{x}} &= \mathbf{A}
\end{aligned}
$$

Matrix-by-Scalar: the Tangent Matrix

If $\mathbf{Y}(x) \in \Re^{m\times n}$ is a matrix function of a scalar x, we can compute a tangent matrix, $\frac{\partial \mathbf{Y}}{\partial x}$:

$$\frac{\partial \mathbf{Y}}{\partial x} = \begin{bmatrix} \frac{\partial y_{1,1}}{\partial x} & \cdots & \frac{\partial y_{1,n}}{\partial x} \\ \vdots & \ddots & \vdots \\ \frac{\partial y_{m,1}}{\partial x} & \cdots & \frac{\partial y_{m,n}}{\partial x} \end{bmatrix}$$

Scalar-by-Matrix: the Gradient Matrix

$\frac{\partial y}{\partial \mathbf{X}}$ has no universally accepted name, but in machine learning it is often called the gradient matrix, in analogy with the gradient vector. In **numerator layout** order, $\mathbf{X}$ and $\partial y / \partial \mathbf{X}$ are defined as:

$$\mathbf{X} = \begin{bmatrix} x_{1,1} & \cdots & x_{1,n} \\ \vdots & \ddots & \vdots \\ x_{m,1} & \cdots & x_{m,n} \end{bmatrix}, \qquad \frac{\partial y}{\partial \mathbf{X}} = \begin{bmatrix} \frac{\partial y}{\partial x_{1,1}} & \cdots & \frac{\partial y}{\partial x_{m,1}} \\ \vdots & \ddots & \vdots \\ \frac{\partial y}{\partial x_{1,n}} & \cdots & \frac{\partial y}{\partial x_{m,n}} \end{bmatrix}$$

Conclusions: Derivatives with Matrices

$$\frac{\partial \mathbf{A}x}{\partial x} = \mathbf{A}$$

$$\frac{\partial \operatorname{tr}(\mathbf{X})}{\partial \mathbf{X}} = I$$

$$\frac{\partial \operatorname{tr}(\mathbf{AX})}{\partial \mathbf{X}} = \mathbf{A}$$

$$\frac{\partial \operatorname{tr}(\mathbf{X}^T\mathbf{X})}{\partial \mathbf{X}} = 2\mathbf{X}^T$$

$$\frac{\partial \operatorname{tr}((\mathbf{Y} - \mathbf{XA})^T(\mathbf{Y} - \mathbf{XA}))}{\partial \mathbf{X}} = -2\left(\mathbf{Y} - \mathbf{XA}\right)^T \mathbf{A}$$