MATHEMATICAL PRELIMINARIES

SCALAR CONCEPTS

The simplest object that we will be dealing with in this course is that of a scalar. A scalar α is an element of a field $\mathcal{F}$. A field is a type of algebraic structure that has certain processes defined on it. These processes manipulate elements of the field.

An example of a field is the set of real numbers $\mathcal{R}$. A field may also encompass the set of complex numbers $\mathcal{C}$. The set of integers $\mathcal{N}$ is a subset of either case. For each of these examples, scalars would be the elements of these sets, which are numbers. Arithmetic is used to manipulate these numbers to perform computations.

SCALAR OPERATIONS

SCALAR ADDITION

The first fundamental process is scalar addition. There are four main properties of scalar multi-plication:

Commutative Property: Given $x, y \in \mathcal{R}$, $x + y = y + x$, i.e., the result is independent of the order of scalars.

Associative Property: Given $x, y, z \in \mathcal{R}$, $(x + y) + z = x + (y + z)$, i.e., the grouping of scalars has no influence on the result.

Multiplicative Identity Property: Given $x \in \mathcal{R}$, $x + 0 = x$; the value 0 here is the identity element for the reals.

Inverse Element Property: For any $x \in \mathcal{R}$, there exists a $y \in \mathcal{R}$ such that $x + (-y) = 0$, where the value of 0 is the identity element for the reals.

SCALAR OPERATIONS

SCALAR MULTIPLICATION

Another necessary process is scalar multiplication. There are five main properties of scalar multiplication:

Commutative Property: Given $x, y \in \mathcal{R}$, $x * y = y * x$, i.e., the result is independent of the order of scalars.

Associative Property: Given $x, y, z \in \mathcal{R}$, $(x * y) * z = x * (y * z)$, i.e., the grouping of scalars has no influence on the result.

Multiplicative Identity Property: Given $x \in \mathcal{R}$, $x * 1 = x$; the value 1 here is the identity element for the reals.

Inverse Element Property: For any $x \in \mathcal{R}$, there exists a $y \in \mathcal{R}$ such that $x * y = 1$, where the value of 1 is the identity element for the reals.

Distributive Property: Multiplication distributes over addition: given $x, y, z \in \mathcal{R}$, $x * (y + z) = xy + xz$.

SCALAR OPERATIONS

SCALAR FUNCTIONS AND DERIVATIVES

Both scalar addition and multiplication are examples of functions. A function, which is also known as a mapping, is a relation between a set of inputs and a set of outputs with the property that each input is related to one or more outputs. Functions can be either bijective (one-to-one), surjective (many-to-one), or injective (one-to-some).

Functions typically change in response to alterations in their parameters. The measure of change is referred to as the derivative.

Let $f : \mathcal{R} \to \mathcal{R}$ be a real-valued function defined in an open neighborhood of a scalar $x \in \mathcal{R}$. The tangent line to the graph of a function f at x is the unique line through the point $(x, f(x))$ that does not meet the graph of f transversally; that is, the tangent line does not pass straight through the graph. The derivative of this function at x is the slope of the tangent line to the graph of f at $(x, f(x))$. To find the slope is to consider a point x_0 that is close to x, connect the points by a line segment, then find the segment's slope. As $x_0 \to x$, we recover the tangent slope:

$$\frac{d}{dx} f(x) = \lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0} = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}.$$

Example: Find the derivative of the following function: $f(x) = 2x^2 - 16x + 35$.

Solution: We can simply plug in this expression into the equation for the total derivative:

$$\frac{d}{dx} f(x) = \lim_{h \to 0} \frac{2(x+h)^2 - 16(x+h) + 35 - (2x^2 - 16x + 35)}{h}.$$

We then expand the quadratic expression and gather the terms. After some simplification, we find that

$$\begin{aligned} \frac{d}{dx} f(x) &= \lim_{h \to 0} \frac{2x^2 + 4xh + 2h^2 - 16x - 16h + 35 - 2x^2 + 16x - 35}{h} \\ &= \lim_{h \to 0} \frac{4xh + 2h^2 - 16h}{h} = \frac{h(4x + 2h - 16)}{h}. \end{aligned}$$

We can now cancel the common h in the numerator and the denominator. There will be only a single term in the numerator with an h in it. We can therefore take the limit as $h \to 0$ without

Example: Find the derivative of the following function: $f(x) = x/(x+1)$.

Solution: We can simply plug in this expression into the equation for the total derivative:

$$\frac{d}{dx}f(x) = \lim_{h\to 0} \frac{1}{h}\left(\frac{x+h}{x+h+1} - \frac{x}{x+1}\right).$$

As with the first example, we cannot simply evaluate the limit, due to a potential division by zero. We will need to simplify the expression somewhat, which can be done by combining the two terms in the numerator into a single rational expression

$$\begin{aligned}\frac{d}{dx}f(x) &= \lim_{h\to 0} \frac{1}{h}\left(\frac{(x+h)(x+1) - x(x+h+1)}{(x+h+1)(x+1)}\right) \\ &= \lim_{h\to 0} \frac{1}{h}\left(\frac{x^2+x+xh+h-(x^2+xh+x)}{(x+h+1)(x+1)}\right)\end{aligned}$$

SCALAR OPERATIONS

SCALAR FUNCTIONS AND PARTIAL DERIVATIVES

Often, a function may have more than one input that influences its output. We would still like to measure how much the function changes according to changes in one or more inputs. However, we cannot easily do this simultaneously across all inputs. Rather, we must consider the individual change across each input, which leads to the notion of the partial derivative.

Like ordinary derivatives, the partial derivative is defined as a limit. Let $f : \mathcal{R}^n \to \mathcal{R}$ be a real-valued function, defined on an open neighborhood, with m scalar variables $x \in \mathcal{R}^n$. The partial
derivative at a point $x = (x_1, x_2, \dots, x_n)$ with respect to the ith variable x_i is defined as

$$\frac{\partial}{\partial x_i} f(x) = \lim_{h \to 0} \frac{f(x_1, \dots, x_{i-1}, x_i + h, x_{i+1}, \dots, x_n) - f(x_1, x_2, \dots, x_n)}{h}.$$

SCALAR OPERATIONS

SCALAR FUNCTIONS AND PARTIAL DERIVATIVES: EXAMPLE

Example: Find the first-order partial derivatives of the following function:
$f(x, y) = x^4 + 6y^{1/2} - 10.$

Solution: We can first take the derivative of this function with respect to x. In such a case, all of the y variables will be treated as constants. To simplify the presentation, we will take the derivative using the power rule. We therefore have that

$$\frac{\partial}{\partial x} f(x, y) = \frac{\partial}{\partial x} x^4 = 4x^3.$$

Notice that the second and the third term of the original multi-variable function differentiate to zero in this case. It should be clear why these terms differentiated to zero: they are constants, and constants always differentiate to zero.

Now we can take the derivative of the function with respect to y. In such a case, all of the x variables will be treated as constants. To simplify the presentation, we will take the derivative using the power rule. We therefore have that

SCALAR OPERATIONS

SCALAR FUNCTIONS AND PARTIAL DERIVATIVES: EXAMPLE

Example: Find the first-order partial derivatives of the following function:
$f(x, y, z) = x^2 y - 10y^2 z^3 + 43x - 7\tan(4y)$.

Solution: We can first take the derivative of this function with respect to x. In such a case, all of the y and z variables will be treated as constants. We find that

$$\frac{\partial}{\partial x} f(x, y, x) = \frac{\partial}{\partial x}\left(x^2 y + 43x\right) = 2xy + 43.$$

Now we can take the derivative of the function with respect to y. In such a case, all of the x and z variables will be treated as constants.

$$\frac{\partial}{\partial y} f(x, y, x) = \frac{\partial}{\partial y}\left(x^2 y - 10y^2 z^3 - 7\tan(4y)\right) = x^2 - 20yz^3 - 28\sec^2(4y).$$

When taking the partial derivative of the tangent term, we can use the limit-based definition

Often, we will be considering a concatenation of multiple scalars. This leads to a structure called a vector. Each of these scalars might represent features about an object. For instance, if we are wanting to classify different types of flowers, we might measure their petal widths and lengths and estimate the average petal color. Each of these attributes would be features. Each new flower that is measured leads to a new vector.

More formally, a vector space over a field $\mathcal{F}$, such as the reals $\mathcal{R}$, is a set $\mathcal{V}$ together with two operations: addition and scalar multiplication. These two operations satisfy nine axioms. Elements of this vector field $v \in \mathcal{V}$ are referred to as vectors.

VECTOR OPERATIONS

VECTOR SCALING

There are a variety of processes that can be defined for vectors. The simplest is scaling. Given an n-dimensional, real vector $x \in \mathcal{R}^n$ with components $[x_1, x_2, \ldots, x_n]$, scaling this vector involves multiplying each component by a real scalar factor $\alpha \in \mathcal{R}$:

$$\alpha x = [\alpha x_1, \alpha x_2, \ldots, \alpha x_n] \in \mathcal{R}^n.$$

In the case of a Euclidean metric space, notice that scaling does not change the direction of the vector; it only impacts its magnitude. The same result is true in many other metric spaces.

VECTOR OPERATIONS
VECTOR SCALING

Many of the properties of scalar multiplication carry over to the vectorial case:

Scalar Additivity: Given $\alpha, \beta \in \mathcal{R}$ and $v \in \mathcal{R}^n$, $(\alpha + \beta)v = \alpha v + \beta v$.

Distributive Property: Multiplication distributes over vector addition: given $\alpha \in \mathcal{R}$ and $v, w \in \mathcal{R}^n$, $\alpha * (v + w) = \alpha v + \alpha w$.

Multiplicative Identity Property: Multiplication by the identity element doesn't change the vector: given $v \in \mathcal{R}^n$, $1 * v = v$.

Zero Element Property: Multiplication by the zero element yields the zero vector: given $v \in \mathcal{R}^n$, $0 * v = [0, 0, \dots, 0] \in \mathcal{R}^n$.

Inverse Property: Multiplication by the negative identity element gives the additive inverse: given $v \in \mathcal{R}^n$, $-1 * v = -v$.

VECTOR OPERATIONS

VECTOR ADDITION AND ELEMENT-WISE MULTIPLICATION

Addition and element-wise multiplication of vectors with the same dimensionality work component-wise. For vectors $x, y \in \mathcal{R}^n$ with components $[x_1, x_2, \ldots, x_n]$ and $[y_1, y_2, \ldots, y_n]$, we have that

$$x + y = [x_1 + y_1, x_2 + y_2, \ldots, x_n + y_n] \in \mathcal{R}^n.$$

Subtraction is a special case of addition where one vector is pre-scaled via a unit negative value. As well, for element-wise multiplication of two vectors, $x, y \in \mathcal{R}^n$, we have that

$$xy = [x_1 y_1, x_2 y_2, \ldots, x_n y_n] \in \mathcal{R}^n.$$

Element-wise division of vectors is a special case of multiplication. Both of these fundamental operations can change the direction and magnitude of vectors.

VECTOR OPERATIONS

VECTOR ADDITION AND ELEMENT-WISE MULTIPLICATION

Many of the properties of scalar addition carry over to the vectorial case:

Commutative Property: Given $v, w \in \mathcal{R}^n$, $v + w = w + v$.

Associative Property: Given $v, w, x \in \mathcal{R}^n$, $(v + w) + z = v + (w + z)$.

Zero Element Property: Addition by the zero vector gives back the original vector: given $v \in \mathcal{R}^n$, $0 + v = v$.

Inverse Element Property: For any $v \in \mathcal{R}^n$, there exists a $w \in \mathcal{R}^n$ such that $v + w = [0, 0, \dots, 0] \in \mathcal{R}^n$.

VECTOR OPERATIONS
INNER PRODUCT

△ △ △

Another type of vector multiplication is the inner product. The inner product of two vectors $x, y \in \mathcal{R}^d$ yields a scalar defined by

$$x^\top y = y^\top x = \sum_{i=1}^{d} x_i y_i.$$

The inner product is also sometimes referred to as the scalar product or dot product. Every inner product space defines a corresponding metric space. For a Euclidean metric space, vectors have both direction and magnitude. The inner product of two Euclidean vectors can thus be written in terms of vector length products and the cosine of the angle between vectors. There are two important cases of the Euclidean inner product to consider:

Orthogonal Vectors: When two vectors are orthogonal, that is, the angle between them is $\pi/2$ radians, the inner product $x^\top y = 0$.

Co-Directional Vectors: When two vectors are co-directional, that is, the angle between them is 0 radians, the inner product $x^\top y = \|x\|\|y\|$ is the product of the vector lengths. Here, the lengths of the vectors are with respect to the Euclidean norm.

▽ ▽ ▽

VECTOR OPERATIONS
INNER PRODUCT: EXAMPLE

Example: Find the inner product of the following vectors: $x = [1, -2, 5, 0]$ and $y = [5, 15, 2, 1]$.

Solution: We can simply multiply each vector component together and sum their result:
$x^\top y = 1 * 5 - 2 * 15 + 2 * 5 - 1 * 0 = -15.$

Example: Find the inner product of the following vectors: $x = [2, 0.7, 5, 1]$ and $y = [0.5, 4, 9, 18]$.

Solution: We can simply multiply each vector component together and sum their result:
$x^\top y = 2 * 0.5 + 0.7 * 4 + 5 * 9 + 18 * 1 = 66.8.$

VECTOR OPERATIONS
OUTER PRODUCT

The outer product of two vectors $x \in \mathcal{R}^d$ and $y \in \mathcal{R}^n$ yields a matrix whose entries $x_i y_i \in \mathcal{R}$ are scalar values

$$xy^\top = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}^\top = \begin{bmatrix} x_1 y_1 & x_1 y_2 & \cdots & x_1 y_n \\ x_2 y_1 & x_2 y_2 & \cdots & x_2 y_n \\ \vdots & \vdots & \ddots & \vdots \\ x_d y_1 & x_d y_2 & \cdots & x_d y_n \end{bmatrix} \in \mathcal{R}^{n \times m}.$$

The outer product is an operator that can act on other vectors. If we have a vector $w \in \mathcal{R}^k$, then $(xy^\top)w = x(y^\top w) = (x \cdot y)w$. Assuming that $x = y$, there is a simple geometric interpretation for this process. Recall that a vector describes a single direction in a vector space. All scalar multiples of that vector define a line lying in that direction. The exterior product of x with itself measures the projection of $x \cdot w$ onto x, then rescales x to to have that length.

VECTOR OPERATIONS

OUTER PRODUCT: EXAMPLE

Example: Find the outer product of the following vectors: $x = [1, -2, 5, 0]$ and $y = [5, 15, 2, 1]$.

Solution: We can simply multiply each vector component together to form a matrix

$$\begin{bmatrix} 1 \\ -2 \\ 5 \\ 0 \end{bmatrix} \begin{bmatrix} 5 \\ 15 \\ 2 \\ 1 \end{bmatrix}^{\top} = \begin{bmatrix} 5 & 15 & 2 & 1 \\ -10 & -30 & -4 & -2 \\ 25 & 75 & 10 & 5 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

Example: Find the outer product of the following vectors: $x = [2, 0.7, 5, 1]$ and $y = [0.5, 4, 9, 18]$.

Solution: We can simply multiply each vector component together to form a matrix

$$\begin{bmatrix} 2 \\ 0.7 \\ 5 \\ 1 \end{bmatrix} \begin{bmatrix} 0.5 \\ 4 \\ 9 \\ 18 \end{bmatrix}^{\top} = \begin{bmatrix} 1 & 8 & 18 & 361 \\ 0.35 & 2.8 & 6.3 & 12.6 \\ 2.5 & 20 & 45 & 90 \\ 0.5 & 4 & 9 & 18 \end{bmatrix}.$$

VECTOR OPERATIONS
NORM OF A VECTOR

Given $x \in \mathcal{R}^d$, the norm of x, $\|x\|$, is a non-negative function that assigns a non-negative length to each vector.

Norms have the following properties:

- Positivity: $\|x\| \geq 0$.
- Absolute Homogeneity: $\|kx\| = |k| \|x\|$.
- Triangle Inequality: for $y \in \mathcal{R}^d$, $\|x + y\| \leq \|x\| + \|y\|$.
- Point Separation: if $\|x\| = 0$, then x is the zero vector.

VECTOR OPERATIONS

NORM OF A VECTOR

One of the most well known vector norms is the L_p norm:

$$\|x\|_p = \left(\sum_{j=1}^{d} |x_j|^p \right)^{1/p}.$$

There are many important special cases of the L_p norm. We have, for example, the L_2 norm, which measures the magnitude of the vector: $\|x\|_2 = (x_1^2 + x_2^2 + \ldots + x_d^2)^{1/2}$.

Certain machine-learning models will require taking derivatives of or by vectors. There are three vector-based derivatives. The first is that of the derivative of a scalar by a vector. Given a scalar $\alpha \in \mathcal{R}$ and a vector $x \in \mathcal{R}^n$, the derivative is given by

$$\frac{\partial}{\partial x}\alpha = \left[\frac{\partial}{\partial x_1}\alpha, \frac{\partial}{\partial x_2}\alpha, \ldots, \frac{\partial}{\partial x_n}\alpha\right]^\top \in \mathcal{R}^{n\times 1}.$$

In vector calculus, the gradient of a scalar field in $\mathcal{R}^n$, whose independent coordinates are components of the vector, is the derivative of a scalar by a vector. We can also write the directional derivative $\nabla_u f(x) = \nabla f(x) \cdot u$ of a scalar function $f(x)$ of the space vector $x \in \mathcal{R}^n$ in direction $u \in \mathcal{R}^n$ in terms of a derivative of a scalar with respect to a vector: $\nabla_u f(x) = \partial f(x) u / \partial x$.

Another important derivative is of a vector by a scalar, which is given by

$$\frac{\partial}{\partial \alpha}x = \left[\frac{\partial}{\partial \alpha}x_1, \frac{\partial}{\partial \alpha}x_2, \ldots, \frac{\partial}{\partial \alpha}x_n\alpha\right]^\top \in \mathcal{R}^{n\times 1}.$$

The other construct is that of a matrix. A matrix over a field $\mathcal{F}$ is a rectangular array of scalars, each of which is a member of $\mathcal{F}$. The set of all matrices forms an algebraic ring $\mathcal{M}$ with two fundamental processes: entry-wise addition and matrix multiplication. Each of these processes satisfies a series of axioms, just like in the vector case.

MATRIX OPERATIONS
MATRIX SCALING

Addition and inner-product-based multiplication of vectors are degenerate instances of corresponding matrix operations. For example, suppose that we have a factor $\alpha \in \mathcal{R}$. The scaling of a matrix $X \in \mathcal{R}^{n \times m}$ is the entry-wise multiplication by that factor

$$\alpha X = \begin{bmatrix} \alpha x_{1,1} & \alpha x_{1,2} & \cdots & \alpha x_{1,m} \\ \alpha x_{2,1} & \alpha x_{2,2} & \cdots & \alpha x_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha x_{n,1} & \alpha x_{n,2} & \cdots & \alpha x_{n,m} \end{bmatrix} \in \mathcal{R}^{n \times m}.$$

Notice that scaling is commutative; hence, the pre- or post-multiplication of the matrix by the scalar does not change the result.

Example: Find the result of multiplying the following matrix by the scalar $\alpha = 0.2$

$$\begin{bmatrix} 9 & 4 & 10 & -2 \\ 6 & 8 & -4 & 0 \\ 18 & 5 & 10 & 19 \\ 0 & 1 & -8 & 4 \end{bmatrix}.$$

Solution: We simply scale each entry of the matrix by $\alpha = 0.2$, which leads to

$$\begin{bmatrix} 1.8 & 0.8 & 2 & -0.4 \\ 1.2 & 1.6 & -0.8 & 0 \\ 3.6 & 1 & 2 & 3.8 \\ 0 & 0.2 & -1.6 & 0.8 \end{bmatrix}.$$

MATRIX OPERATIONS

MATRIX ADDITION

Addition also works in an element-wise fashion. Given two equally sized matrices $X,\ Y \in \mathcal{R}^{n \times m}$, their addition is defined as

$$X + Y = Y + X = \begin{bmatrix} x_{1,1} + y_{1,1} & x_{1,2} + y_{1,2} & \cdots & x_{1,m} + y_{1,m} \\ x_{2,1} + y_{2,1} & x_{2,2} + y_{2,2} & \cdots & x_{2,m} + y_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} + y_{n,1} & x_{n,2} + y_{n,2} & \cdots & x_{n,m} + y_{n,m} \end{bmatrix} \in \mathcal{R}^{n \times m}.$$

Subtraction is again a special case of addition. Addition is commutative; hence, the order of the matrices is interchangeable without changing the results.

Example: Find the result for the following

$$\begin{bmatrix} 5 & 2 & 0 \\ 4 & 9 & 2 \\ 10 & -3 & -1 \\ 5 & 12 & 16 \end{bmatrix} + \begin{bmatrix} -11 & 0 & 6 \\ 7 & 1 & 0 \\ -6 & -8 & 12 \\ 2 & 10 & 4 \end{bmatrix}$$

Solution: We simply add the corresponding entries together

$$\begin{bmatrix} 5-11 & 2+0 & 0+6 \\ 4+7 & 9+1 & 2+0 \\ 10-6 & -3-8 & -1+12 \\ 5+2 & 12+10 & 16+4 \end{bmatrix} = \begin{bmatrix} -6 & 2 & 6 \\ 11 & 10 & 2 \\ 4 & -11 & 11 \\ 7 & 22 & 20 \end{bmatrix}.$$

MATRIX OPERATIONS

MATRIX TRANSPOSE

Matrices also have a transpose operation. A given matrix $X \in \mathcal{R}^{n\times m}$ is said to be transposed when its columns are converted into rows and vice versa, which leads to a matrix $X^\top \in \mathcal{R}^{m\times n}$ with

$$\begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,m} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,m} \end{bmatrix}^\top = \begin{bmatrix} x_{1,1} & x_{2,1} & \cdots & x_{n,1} \\ x_{1,2} & x_{2,2} & \cdots & x_{n,2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1,m} & x_{2,m} & \cdots & x_{n,m} \end{bmatrix} \in \mathcal{R}^{m\times n}.$$

MATRIX OPERATIONS
MATRIX TRANSPOSE: EXAMPLE

Example: Find the transpose of the following matrix

$$\begin{bmatrix} 2 & 0 & 3 & -8 \\ 1 & -5 & -2 & 6 \\ 7 & 4 & 0 & 9 \end{bmatrix}.$$

Solution: Each row of the original matrix becomes a column and each column of the original matrix becomes a row:

$$\begin{bmatrix} 2 & 1 & 7 \\ 0 & -5 & 4 \\ 3 & -2 & 0 \\ -8 & 6 & -9 \end{bmatrix}.$$

△ △ △

Sometimes we will need to invert a matrix. A matrix $X \in \mathcal{R}^{n \times n}$ is invertible if there exists another matrix $Y \in \mathcal{R}^{n \times n}$ such that $XY = I_{n \times n}$. The inverse can be obtained in a variety of ways, such as Gauss-Jordan elimination, the Cayley-Hamilton method, and Cramer's rule. Each of these methods has their own numerical advantages and disadvantages.

Gauss-Jordan elimination involves right adjoining a given matrix with the identity matrix. Row-based operations are applied to reduce the left-hand side to the identity matrix; the inverted matrix is given on the right

$$\text{step 1}: \left[\begin{array}{cccc|cccc} x_{1,1} & x_{1,2} & \cdots & x_{1,m} & 1 & 0 & \cdots & 0 \\ x_{2,1} & x_{2,2} & \cdots & x_{2,m} & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,m} & 0 & 0 & \cdots & 1 \end{array}\right]$$

$$\text{step 2}: \left[\begin{array}{cccc|cccc} x_{1,1}+\alpha_1 x_{2,1} & x_{1,2}+\alpha_1 x_{2,2} & \cdots & x_{1,m}+\alpha_1 x_{2,m} & 1 & \alpha_1 & \cdots & 0 \\ x_{2,1} & x_{2,2} & \cdots & x_{2,m} & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \end{array}\right]$$

▽ ▽ ▽

Example: Find the inverse of the following matrix

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}.$$

Solution: We append the identity matrix to the right-hand side of the original matrix. We then scale various rows and add them together until we get the identity matrix on the left-hand side

$$\text{step 1: append}: \left[\begin{array}{cc|cc} 1 & 2 & 1 & 0 \\ 3 & 4 & 0 & 1 \end{array}\right]$$

$$\text{step 2}: -3 * \text{row } 1 + \text{row } 2: \left[\begin{array}{cc|cc} 1 & 2 & 1 & 0 \\ 0 & -2 & -3 & 1 \end{array}\right]$$

$$\text{step 3}: -0.5 * \text{row } 2: \left[\begin{array}{cc|cc} 1 & 2 & 1 & 0 \\ 0 & 1 & 3/2 & -1/2 \end{array}\right]$$

Two matrices can be multiplied together, provided that the number of columns of the multiplier equals the number of rows of the multiplicand. Given $X \in \mathcal{R}^{n \times m}$ and $Y \in \mathcal{R}^{m \times k}$, the result is a matrix $XY \in \mathcal{R}^{n \times k}$. The actual computation proceeds as follows,

$$XY = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,m} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,m} \end{bmatrix} \begin{bmatrix} y_{1,1} & y_{1,2} & \cdots & y_{1,k} \\ y_{2,1} & y_{2,2} & \cdots & y_{2,k} \\ \vdots & \vdots & \ddots & \vdots \\ y_{m,1} & x_{m,2} & \cdots & y_{m,k} \end{bmatrix}$$

$$= \begin{bmatrix} \sum_{p=1}^{m} x_{1,p} y_{p,1} & \sum_{p=1}^{m} x_{1,p} y_{p,2} & \cdots & \sum_{p=1}^{m} x_{1,p} y_{p,k} \\ \sum_{p=1}^{m} x_{2,p} y_{p,1} & \sum_{p=1}^{m} x_{2,p} y_{p,2} & \cdots & \sum_{p=1}^{m} x_{2,p} y_{p,k} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{p=1}^{m} x_{n,p} y_{p,1} & \sum_{p=1}^{m} x_{n,p} y_{p,2} & \cdots & \sum_{p=1}^{m} x_{n,p} y_{p,k} \end{bmatrix}.$$

There are a few important properties of matrix multiplication:

Example: Perform the following multiplication

$$\begin{bmatrix} 3 & 4 & 2 \end{bmatrix} * \begin{bmatrix} 13 & 9 & 7 & 15 \\ 8 & 7 & 4 & 6 \\ 6 & 4 & 0 & 3 \end{bmatrix}$$

Solution: We simply compute a series of dot products

$$\begin{bmatrix} 3*13+4*8+2*6 \\ 3*9+4*7+2*4 \\ 3*7+4*4+3*0 \\ 3*15+4*6+2*3 \end{bmatrix}^\top = \begin{bmatrix} 83 \\ 63 \\ 37 \\ 75 \end{bmatrix}^\top.$$

There is an extension of the vector norm to matrices. For $X \in \mathcal{R}^{n\times d}$, a matrix norm $\|X\|$ is a vector norm on $\mathcal{R}^{n\times d}$ that assigns a non-negative value that corresponds to the magnitude.

Norms have the following properties:

- **Positivity**: $\|X\| \geq 0$.
- **Absolute Homogeneity**: $\|kX\| = |k|\|X\|$.
- **Triangle Inequality**: for $Y \in \mathcal{R}^{d}$, $\|X+Y\| \leq \|X\| + \|Y\|$.
- **Point Separation**: if $\|x\| = 0$, then x is the zero vector.
- **Submultiplicativity**: for $X, Y \in \mathcal{R}^{n\times n}$, $\|XY\| \leq \|X\|\|Y\|$.

MATRIX OPERATIONS
MATRIX DETERMINANT

For $X \in \mathcal{R}^{n \times n}$, the determinant $\det(X)$ is a useful value that can be computed from square matrices. For a 2×2 matrix, we have that

$$\det(X) = \begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - bc.$$

This extends to arbitrary dimensions. For example, when considering a 3×3 matrix, we have that

$$\det(X) = \begin{vmatrix} a & b & c \\ d & e & f \\ g & h & i \end{vmatrix} = a \begin{vmatrix} e & f \\ h & i \end{vmatrix} - b \begin{vmatrix} d & f \\ g & i \end{vmatrix} + c \begin{vmatrix} d & e \\ g & h \end{vmatrix}$$

which is simply $aei + bfg + cdh + ceg + bdi + afh$.

MATRIX OPERATIONS
MATRIX DETERMINANT

Matrix determinants have several properties:

Transpose Equivalency: $\det(X^\top) = \det(X)$.

Homogeneity: $\det(kX) = k^n \det(X)$, for $X \in \mathcal{R}^{n\times n}$.

Distributive Property: $\det(XY) = \det(X)\det(Y)$, for $X, Y \in \mathcal{R}^{n\times n}$.

Inverse Property: $\det(X^{-1}) = 1/\det(X)$.

Product Property: $\det(X) = \prod_{i=1}^{n} x_{i,i}$ if X is triangular. A matrix X is upper triangular if $x_{i,j} = 0$ whenever $i > j$. It is lower traingular if $x_{i,j} = 0$ whenever $i < j$.

An important use of the determinant is for matrix inverses: a matrix $X \in \mathcal{R}^{n\times n}$ has a unique inverse if $\det(A) \neq 0$.

There are multiple ways to factorize matrices. An important factorization is eigendecomposition or spectral decompositions, which represents a square matrix $X \in \mathcal{R}^{n \times n}$ in terms of its eigenvectors $v \in \mathcal{R}^n$ and its eigenvalues $\lambda \in \mathcal{R}$

$$Xv = \lambda v.$$

Recall that eigenvectors are non-zero vectors that have a fixed direction under a given linear transformation. The scaling of the eigenvectors under such linear transformations is determined by the eigenvalues.

When $Xv = \lambda v$ is re-written, we obtain a characteristic polynomial equation for determining the set of eigenvalues for X

$$\det(X - \lambda I_{n \times n}) = 0,$$

where $I_{n \times n} \in \mathcal{R}^{n \times n}$ is the identity matrix. This equation will have up to n distinct solutions. The set of eigenvalues is referred to as the spectrum of the matrix X. For each eigenvalue λ_i, we have a specific eigenvalue equation

Suppose that we have a matrix $X \in \mathcal{R}^{n \times d}$ whose entries depend on a scalar parameter $\alpha \in \mathcal{R}$. We can take the derivative of X component-by-component to get another matrix as

$$\frac{\partial}{\partial \alpha} X = \begin{bmatrix} \frac{\partial}{\partial \alpha} x_{1,1} & \frac{\partial}{\partial \alpha} x_{1,2} & \cdots & \frac{\partial}{\partial \alpha} x_{1,d} \\ \frac{\partial}{\partial \alpha} x_{2,1} & \frac{\partial}{\partial \alpha} x_{2,2} & \cdots & \frac{\partial}{\partial \alpha} x_{2,d} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial}{\partial \alpha} x_{n,1} & \frac{\partial}{\partial \alpha} x_{n,2} & \cdots & \frac{\partial}{\partial \alpha} x_{n,d} \end{bmatrix} \in \mathcal{R}^{n \times d}.$$

If we instead wish to take the derivative of $X^{-1} \in \mathcal{R}^{n \times n}$, then the derivative matrix must be pre- and post-multiplied by X^{-1},

$$\begin{bmatrix} \frac{\partial}{\partial \alpha} x_{1,1} & \frac{\partial}{\partial \alpha} x_{1,2} & \cdots & \frac{\partial}{\partial \alpha} x_{1,d} \end{bmatrix}$$

FUNCTION OPERATIONS

GRADIENTS OF VECTOR-BASED FUNCTIONS ON SCALAR DOMAINS

When attempting to uncover parameters for adaptive systems, we will need knowledge of the gradients of functions with several variables. The gradient of a real, differentiable function $f : \mathcal{R}^d \to \mathcal{R}$ with respect to its vector argument is defined in terms of partial derivatives

$$\nabla f(x) = \left[\frac{\partial}{\partial x_1} f(x),\ \frac{\partial}{\partial x_2} f(x),\ \cdots,\ \frac{\partial}{\partial x_d} f(x) \right]^\top .$$

Intuitively, the gradient points in the direction of the greatest rate of increase for the function. If we wish to find parameters that maximize a function, we would move in the direction of the gradient and proportional to its magnitude. If we seek to minimize a function, then we would search in the direction of the negative gradient to find good parameter values.

If this function $f : \mathcal{R}^d \to \mathcal{R}$ is at least twice differentiable, then the second-order gradient with respect to the vector argument yields the Hessian matrix

$$\begin{bmatrix} \frac{\partial^2}{\partial x_1 \partial x_1} f_1(x) & \frac{\partial^2}{\partial x_1 \partial x_2} f_1(x) & \cdots & \frac{\partial^2}{\partial x_1 \partial x_d} f_1(x) \end{bmatrix}$$

FUNCTION OPERATIONS

GRADIENTS OF VECTOR-BASED FUNCTIONS ON VECTOR DOMAINS

The preceding definitions are appropriate for scalar-valued domains. If we are dealing with vector-valued domains, then some changes are required. In particular, the gradient of a vector-valued function $h : \mathcal{R}^d \to \mathcal{R}^n$ on a vector domain is given by the Jacobian matrix

$$\nabla h(x) = \begin{bmatrix} \frac{\partial}{\partial x_1} h_1(x) & \frac{\partial}{\partial x_2} h_1(x) & \cdots & \frac{\partial}{\partial x_d} h_1(x) \\ \frac{\partial}{\partial x_1} h_2(x) & \frac{\partial}{\partial x_2} h_2(x) & \cdots & \frac{\partial}{\partial x_d} h_2(x) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial}{\partial x_1} h_n(x) & \frac{\partial}{\partial x_2} h_n(x) & \cdots & \frac{\partial}{\partial x_d} h_n(x) \end{bmatrix},$$

which can be written simply as a row vector $[\nabla h_1(x), \nabla h_2(x), \ldots, \nabla h_n(x)] \in \mathcal{R}^{d \times n}$. The second-order gradient with respect to this function becomes

FUNCTION OPERATIONS

GRADIENTS OF MATRIX-BASED FUNCTIONS ON SCALAR DOMAINS

We can also define notions of gradients for matrix-based functions on different domains. The gradient of a matrix-based function $g : \mathcal{R}^{d\times n} \to \mathcal{R}$ on a scalar domain is

$$\nabla g(X) = \begin{bmatrix} \frac{\partial}{\partial x_{1,1}} g(X) & \frac{\partial}{\partial x_{1,2}} g(X) & \cdots & \frac{\partial}{\partial x_{1,d}} g(X) \\ \frac{\partial}{\partial x_{2,1}} g(X) & \frac{\partial}{\partial x_{2,2}} g(X) & \cdots & \frac{\partial}{\partial x_{2,d}} g(X) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial}{\partial x_{n,1}} g(X) & \frac{\partial}{\partial x_{n,2}} g(X) & \cdots & \frac{\partial}{\partial x_{n,d}} g(X) \end{bmatrix},$$

or, rather, $[\nabla_{X(:,1)} g(X), \nabla_{X(:,2)} g(X), \ldots, \nabla_{X(:,n)} g(X)] \in \mathcal{R}^{d\times 1\times n}$. The second-order gradient of a matrix-based function $g : \mathcal{R}^{d\times n} \to \mathcal{R}$ on a scalar domain is given by

$$\begin{bmatrix} \nabla \frac{\partial}{\quad} g(X) & \nabla \frac{\partial}{\quad} g(X) & \cdots & \nabla \frac{\partial}{\quad} g(X) \end{bmatrix}$$

FUNCTION OPERATIONS

GRADIENTS OF MATRIX-BASED FUNCTIONS ON VECTOR AND MATRIX DOMAINS

△ △ △

The gradient of a function $g : \mathcal{R}^{d\times n} \to \mathcal{R}^k$ on a vector domain is the third-order tensor

$$\nabla g(X) = \begin{bmatrix} \nabla_{X(:,1)} g_1(X) & \nabla_{X(:,1)} g_2(X) & \cdots & \nabla_{X(:,1)} g_k(X) \\ \nabla_{X(:,2)} g_1(X) & \nabla_{X(:,2)} g_2(X) & \cdots & \nabla_{X(:,2)} g_k(X) \\ & & \ddots & \\ \nabla_{X(:,n)} g_1(X) & \nabla_{X(:,n)} g_2(X) & \cdots & \nabla_{X(:,n)} g_k(X) \end{bmatrix}.$$

This can be stated succinctly as $[\nabla g_1(X), \nabla g_2(X), \ldots, \nabla g_k(X)] \in \mathcal{R}^{d\times k\times n}$. The second-order gradient of a matrix-based function $g : \mathcal{R}^{d\times n} \to \mathcal{R}^k$ on a vector domain is

$$\nabla^2 g(X) = \begin{bmatrix} \nabla\nabla_{X(:,1)} g_1(X) & \nabla\nabla_{X(:,1)} g_2(X) & \cdots & \nabla\nabla_{X(:,1)} g_k(X) \\ \nabla\nabla_{X(:,2)} g_1(X) & \nabla\nabla_{X(:,2)} g_2(X) & \cdots & \nabla\nabla_{X(:,2)} g_k(X) \\ & & \ddots & \\ \nabla\nabla_{X(:,n)} g_1(X) & \nabla\nabla_{X(:,n)} g_2(X) & \cdots & \nabla\nabla_{X(:,n)} g_k(X) \end{bmatrix}.$$

▽ ▽ ▽