We will assume that vectors are column vectors:

\[
\mathbf{v} = \begin{bmatrix}
    v_1 \\
    v_2 \\
    v_3 \\
    \vdots \\
    v_n
\end{bmatrix}
\]

The transpose of this vector, \( \mathbf{v}^T \), is a row vector:

\[
\mathbf{v}^T = [v_1 \ v_2 \ v_3 \ \cdots \ v_n]
\]

The inner product, also known as the dot product, reduces two vectors of equal length to a scalar:

\[
\mathbf{x} \cdot \mathbf{y} = \mathbf{x}^T \mathbf{y} = \sum_i x_i y_i
\]

In contrast, the outer product takes two vectors and produces an \( n \times n \) matrix:

\[
\mathbf{x} \times \mathbf{y}^T = \mathbf{x} \mathbf{y}^T = \begin{bmatrix}
    x_1 y_1 & x_1 y_2 & \cdots & x_1 y_n \\
    x_2 y_1 & x_2 y_2 & \cdots & x_2 y_n \\
    \vdots & \vdots & \ddots & \vdots \\
    x_n y_1 & x_n y_2 & \cdots & x_n y_n
\end{bmatrix}
\]

A multivariable function takes a number of variables (or a vector) as parameters and returns a single value. In domain terms, a multivariable function \( f(x) \) maps from an \( n \)-dimensional vector space down to a scalar domain. The gradient is the basic vector derivative operation. Given a scalar function \( f(x) \),

\[
\nabla f(x) = \begin{bmatrix}
    \frac{\partial f}{\partial x_1} \\
    \frac{\partial f}{\partial x_2} \\
    \vdots \\
    \frac{\partial f}{\partial x_n}
\end{bmatrix}
\]
yields a vector representing the direction and the rate of change of the function \( f \) within \( \mathbb{R}^n \)-space.

**Example 1:** Let \( f(x) = 1^T x = \sum_i x_i \). Then the gradient of \( f \) is

\[
\nabla f(x) = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} = 1
\]

**Example 2:** Let \( f(x) = x^T x = \sum_i x_i^2 \). Then the gradient of \( f \) is

\[
\nabla f(x) = \begin{bmatrix} 2x_1 \\ 2x_2 \\ \vdots \\ 2x_n \end{bmatrix} = 2x
\]

If \( \nabla f(x') = 0 \) at a point \( x' \) than \( f(x') \) is either a minimum, maximum, or a saddle point of \( f \).

**Example 3:** Let \( f(x) = x_1 x_2 \). \( \nabla f([0, 0]) = 0 \). This is a saddle, curved up in the direction \([1, 1]\) and curved down in the direction \([1, -1]\).

1 Gradient Descent

For \( t = 1 \ldots \)

\[
x^{(t+1)} \leftarrow x^{(t)} - \eta \nabla f(x^{(t)})
\]

will converge to a local minimum if \( \eta \) is small enough, that is, less than some constant dependent on the function \( f \). Gradient descent will converge if \( \eta \) decreases over time as \( \eta_t = \frac{1}{t} \).

2 Convex Functions

A function \( f \) is **convex** if for all \( x_1, x_2 \) and \( \theta \) where \( 0 \leq \theta \leq 1 \),

\[
f(\theta x_1 + (1 - \theta) x_2) \leq \theta f(x_1) + (1 - \theta) f(x_2)
\]

The sum of convex functions is convex. The difference of convex functions is not necessarily convex. The max of convex functions is convex. If \( f(x) \) is convex and \( \nabla f(x') = 0 \), then \( f(x') \) is the global maximum of \( f \).

A function \( f \) is **concave** if its negative is convex.

The sum of concave functions is concave. The difference of concave functions is not necessarily concave. The min of concave functions is concave.
A convex optimization problem has the form

\[
\min_x f(x) \\
\text{s.t. } g_1(x) \leq 0 \\
g_2(x) \leq 0 \\
\vdots \\
g_m(x) \leq 0
\]

where \( f \) and each \( g_i \) are convex functions.

3 Newton’s Method

The Hessian \( \nabla^2 f \) is a matrix of a scalar-valued function \( f(x) \)'s second-order derivatives:

\[
\nabla^2 f = \begin{bmatrix}
\frac{\partial^2 f}{\partial x_1^2} & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_1} \\
\vdots & \ddots & \vdots \\
\frac{\partial^2 f}{\partial x_1 \partial x_n} & \cdots & \frac{\partial^2 f}{\partial x_n^2}
\end{bmatrix}
= \begin{bmatrix}
\frac{\partial^2 f}{\partial x_i \partial x_j}
\end{bmatrix}_{ij}
\]

Newton’s method (also known as Newton-Raphson)

For \( t = 1 \ldots \)

\[
x^{(t+1)} \leftarrow x^{(t)} - (\nabla^2 f(x^{(t)}))^{-1} \nabla f(x^{(t)})
\]

(1)

(2)

can be thought of as finding a quadratic approximation of the function \( f \) (by assuming a constant second derivative) and jumping directly to the minimum, maximum, or saddle point of that function.