Appendix B

The Derivative

B.1 The Derivative of f

In this chapter, we give a short summary of the derivative. Specifically, we want to compare/contrast how the derivative appears for functions whose domain is $\mathbb{R}^n$ and whose range is $\mathbb{R}^m$, for any m, n. We begin by reviewing the definitions found in Calculus:

B.1.1 Mappings from $\mathbb{R}$ to $\mathbb{R}$

Definition: Let $f : \mathbb{R} \to \mathbb{R}$. Then define

$$f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$$

We interpret this quantity as the slope of the tangent line at a given x, or as the velocity at time x. Given this definition, we can give a local linear approximation of a nonlinear function f at $x = a$:

$$L(x) = f(a) + f'(a)(x - a)$$

which is simply the equation of the tangent line to f at $x = a$. For comparison purposes, note that the graph of this function is in $\mathbb{R}^2$, and if $u = x - a, v = f(x) - f(a)$, this function behaves as the linear function $v = f'(a)u$.

Furthermore, we know the basic Taylor series expansion about $x = a$ is an extension of the linearization:

$$f(x) = f(a) + f'(a)(x - a) + \frac{1}{2} f''(a)(x - a)^2 + \ldots + \frac{f^{(k)}(a)}{k!}(x - a)^k + \ldots$$

We have also seen the derivative when f has some different forms:

B.1.2 Mappings from $\mathbb{R}$ to $\mathbb{R}^n$ (Parametrized Curves)

Definition: Let $\boldsymbol{f} : \mathbb{R} \to \mathbb{R}^n$ via:

$$\boldsymbol{f}(t) = \begin{bmatrix} f_1(t) \\ f_2(t) \\ \vdots \\ f_n(t) \end{bmatrix} \quad \text{so that } \boldsymbol{f}'(t) = \begin{bmatrix} f_1'(t) \\ f_2'(t) \\ \vdots \\ f_n'(t) \end{bmatrix}$$

We normally think of the graph of f as a *parameterized curve*, and we differentiate (and integrate) component-wise. In this case, the linearization of f at $x = a$ is a matrix $(n \times 1)$ mapping:

$$L(x) = \boldsymbol{f}(a) + \boldsymbol{f}'(a)(x - a)$$

which takes a scalar x and maps it to a vector starting at $\boldsymbol{f}(a)$ and moves it in the direction of $\boldsymbol{f}'(a)$. The graph of this function lies in $\mathbb{R}^{n+1}$. If $u = x - a, v = \boldsymbol{f}(x) - \boldsymbol{f}(a)$, this function behaves like: $v = \boldsymbol{f}'(a)u$

In differential equations, we considered functions of this form when we looked at systems of differential equations. For example,

$$\dot{\boldsymbol{x}}(t) = A\boldsymbol{x}(t)$$

In this case, the origin is a critical point (fixed point), and we were able to classify the origin according to what the eigenvalues of A were (i.e., positive/negative, complex).

In the more general setting, we also considered the form: $\dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x})$ In this setting, $f : \mathbb{R}^n \to \mathbb{R}^n$, which we look at in the last section.

B.1.3 Mappings from $\mathbb{R}^n$ to $\mathbb{R}$: Surfaces

Definition: Let $f : \mathbb{R}^n \to \mathbb{R}$. Then the derivative in this case is the *gradient* of f:

$$\nabla f = \left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \dots, \frac{\partial f}{\partial x_n} \right)$$

where

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

and measures the rate of change in the direction of x_i. The linearization of f at $\boldsymbol{x} = \boldsymbol{a}$ is now a $1 \times n$ matrix mapping:

$$L(\boldsymbol{x}) = f(\boldsymbol{a}) + \nabla f(\boldsymbol{a})(\boldsymbol{x} - \boldsymbol{a})$$

The graph of this function lies in $\mathbb{R}^{n+1}$, and if $\boldsymbol{u} = \boldsymbol{x} - \boldsymbol{a}, v = f(\boldsymbol{x}) - f(\boldsymbol{a})$, then this function behaves like: $v = \nabla f(\boldsymbol{a})u$.

We use the gradient to measure the rate of change of f in the direction of a unit-length vector u by computing the directional derivative of f at $\boldsymbol{a}$:

$$D_u f = \nabla f(\boldsymbol{a}) \cdot \boldsymbol{u}$$

In the exercises, you are asked to verify that the direction u of fastest increase is in the direction of the gradient.

Geometrically, suppose we are looking at the contours of a function, $y = f(x_1, \ldots, x_n)$: That is, we plot $k = f(x_1, \ldots, x_n)$ for different values of k. Since a contour line is where f is constant, the gradient in the direction of the contour is zero. On the other hand, a vector in the direction of the gradient is orthogonal to the contour, and is the direction of fastest increase.

For example, consider $f(x, y) = x^2 + 2y^2$. Its gradient is $\nabla f = [2x, 4y]$. At the point $x = 0.5, y = 0.5$, the gradient vector is $\nabla f(0.5, 0.5) = [1, 2]$. In Figure B.1, we plot several contours of f, including the contour going through the point $(0.5, 0.5)$. Next, we plot several unit vectors emanating from that point, alongside of which we show the numerical values of the corresponding directional derivatives.

From this, we verify that the direction of maximum increase is in the direction of the gradient, the gradient is orthogonal to the direction tangent to the contour, and the direction of fastest decrease is the negative of the gradient.

This particular class of functions is especially important to us, since:

- Learning can be thought of as the process of minimizing error.
- All error functions can be cast as functions from $\mathbb{R}^n$ to $\mathbb{R}$.

But, before going into the details, let us finish our comparisons of the derivative.

B.1.4 Mappings from $\mathbb{R}^n$ to $\mathbb{R}^m$

The last, most general, class of function is the function that goes from $\mathbb{R}^n$ to $\mathbb{R}^m$. Such a function can always be defined coordinate-wise:

Definition: If $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^m$, then f can be written as:

$$\boldsymbol{f}(\boldsymbol{x}) = \begin{bmatrix} f^1(x_1, \ldots, x_n) \\ f^2(x_1, \ldots, x_n) \\ \vdots \\ f^m(x_1, \ldots, x_n) \end{bmatrix}$$

where each of the f^i are mapping $\mathbb{R}^n$ to $\mathbb{R}$. So, for example, a mapping of $\mathbb{R}^2$ to $\mathbb{R}^3$ might look like:

$$\boldsymbol{f}(\boldsymbol{x}) = \begin{bmatrix} x_1 + x_2 \\ \cos(x_1) \\ x_1 x_2 + \mathrm{e}^{x_1} \end{bmatrix}$$

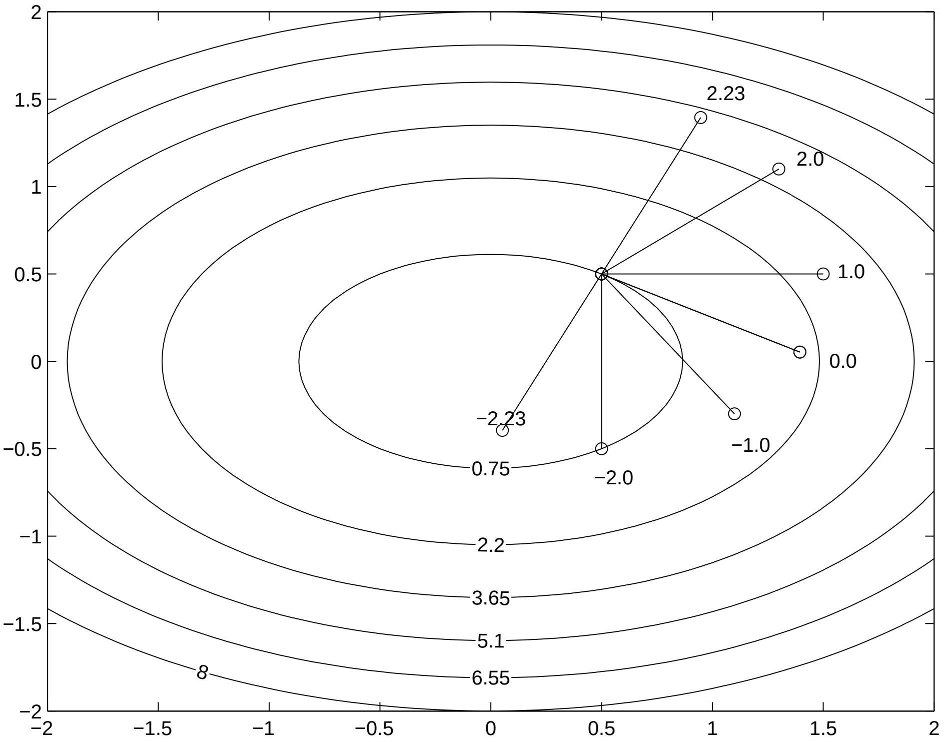

Figure B.1: The plot shows several contours of the function $f(x, y) = x^2 + 2y^2$, with the contour values listed vertically down the center of the plot. We also show several unit vectors emanating from the point $(0.5, 0.5)$, with their associated directional derivative values.

In this case, the derivative of $\boldsymbol{f}$ has a special name: The Jacobian of $\boldsymbol{f}$:

Definition: Let $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^m$. Then the Jacobian of $\boldsymbol{f}$ at $\boldsymbol{x}$ is the $m \times n$ matrix:

$$Df = \begin{bmatrix} \nabla f^1 \\ \nabla f^2 \\ \vdots \\ \nabla f^m \end{bmatrix} = \begin{bmatrix} f_1^1 & f_2^1 & \dots & f_n^1 \\ f_1^2 & f_2^2 & \dots & f_n^2 \\ \vdots & & & \vdots \\ f_1^m & f_2^m & \dots & f_n^m \end{bmatrix}$$

with $f_j^i = \frac{\partial f_i}{\partial x_j}$. You should look this over- it is consistent with all of our previous definitions of the derivative.

The linearization of $\boldsymbol{f}$ at $\boldsymbol{x} = \boldsymbol{a}$ is the affine map:

$$L(x) = \boldsymbol{f}(\boldsymbol{a}) + Df(\boldsymbol{a})(\boldsymbol{x} - \boldsymbol{a})$$

The graph of this function is in $\mathbb{R}^{n+m}$, and if $\boldsymbol{u} = \boldsymbol{x} - \boldsymbol{a}, v = \boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{a})$, then this function behaves like: $v = Df(\boldsymbol{a})u$.

B.2 Worked Examples:

Find the linearization of the given function at the given value of a:

1. $f(x) = 3x^2 + 4, x = 2$
2. $f(t) = (3t^2 + 4, \sin(t))^T, t = \pi$
3. $f(\boldsymbol{x}) = 3x_1x_2 + x_1^2, \boldsymbol{x} = (0, 1)^T$
4. $f(\boldsymbol{x}) = (x_1 + x_2, \cos(x_1), x_1x_2 + \mathrm{e}^{x_1})^T, \boldsymbol{x} = (0, 1)^T$

SOLUTIONS:

1. $f(2) = 16$, $f'(x) = 6x$, so $f'(2) = 12$. Thus,

   $$L(x) = 16 + 12(x - 2)$$

   Locally, this function is like: $v = 12u$.

2. $f(\pi) = (3\pi^2 + 4, 0)^T$, $f'(t) = (6t, \cos(t))^T$, $f'(\pi) = (6\pi, -1)^T$

   $$L(x) = \begin{bmatrix} 3\pi^2 + 4 \\ 0 \end{bmatrix} + \begin{bmatrix} 6\pi \\ -1 \end{bmatrix} (x - \pi)$$

   Locally, this means that f multiplies x by 6π in the first coordinate direction, and flips the second coordinate.

3. $f((0,1)^T) = 0$, $\nabla f = (3x_2 + 2x_1, 3x_1)$, $\nabla f((0,1)^T) = (3,0)$

$$L(x) = 0 + (3,0)\begin{bmatrix} x_1 - 0 \\ x_2 - 1 \end{bmatrix}$$

   Locally (at $(0,1)^T$), this means that f is increasing in the direction parallel to x_1, and is constant in the direction parallel to x_2.

4.

$$f((0,1)^T) = (1,1,1)^T, Df = \begin{bmatrix} 1 & 1 \\ -\sin(x_1) & 0 \\ x_2 + \mathrm{e}^{x_1} & x_1 \end{bmatrix}, Df((0,1)^T) = \begin{bmatrix} 1 & 1 \\ 0 & 0 \\ 2 & 0 \end{bmatrix}$$

   So the linearization is:

$$L(x) = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} + \begin{bmatrix} 1 & 1 \\ 0 & 0 \\ 2 & 0 \end{bmatrix}\begin{bmatrix} x_1 - 0 \\ x_2 - 1 \end{bmatrix}$$

B.3 Optimality

When we talk about "optimizing" a function, that function is necessarily a function from $\mathbb{R}^n$ to $\mathbb{R}$, since this is the only way that we can compare two outputs- that is, we need the range to be well-ordered. In this section, we will therefore concentrate on this class of functions.

In Calculus, we had a second derivative test for determining if we had a local maximum/minimum. That was:

> Let f be differentiable, and $f'(a) = 0$. If $f''(a) > 0$, then f has a local minimum at $x = a$. If $f''(a) < 0$, then f has a local maximum at $x = a$. If $f''(a) = 0$, this test is inconclusive.

We can do something similar for higher dimensions:

Definition: Let $f : \mathbb{R}^n \to R$. Then the Hessian of f is the $n \times n$ matrix:

$$Hf = \begin{bmatrix} f_{11} & f_{12} & \cdots & f_{1n} \\ f_{21} & f_{22} & \cdots & f_{2n} \\ \vdots & & & \vdots \\ f_{n1} & f_{n2} & \cdots & f_{nn} \end{bmatrix}$$

where $f_{ij} = \frac{\partial^2 f}{\partial x_i \partial x_j}$, and so Hf is the matrix of all the second partial derivatives, and takes the place of the second derivative.

We might recall Clairaut's Theorem: If f_{ij}, f_{ji} are continuous in a ball containing $\boldsymbol{x} = \boldsymbol{a}$, then $f_{ij}(\boldsymbol{a}) = f_{ji}(\boldsymbol{a})$. In this case, the Hessian is a symmetric matrix at $\boldsymbol{x} = \boldsymbol{a}$. This will have some consequences that we'll discuss after The Spectral Theorem.

Example: If $f(x_1, x_2) = 3x_1^2 + 2x_1x_2 + x_2^2 - 3x_2 + 4$,

$$\nabla f = [6x_1 + 2x_2, 2x_1 + 2x_2 - 3]$$

and

$$Hf = \begin{bmatrix} 6 & 2 \\ 2 & 2 \end{bmatrix}$$

As in the case of the directional derivative, we can also check the second derivative in the direction of u by computing:

$$\frac{\boldsymbol{u}^T Hf(\boldsymbol{a})\boldsymbol{u}}{\boldsymbol{u}^T\boldsymbol{u}} \text{ or, if we use a unit vector, } \boldsymbol{u}^T Hf(\boldsymbol{a})\boldsymbol{u}$$

In Calc I, we said that, if a is a stationary point, then we can check to see if $f''(a) > 0$ to see if $x = a$ is a local minimum. How does that translate to the Hessian?

Definition: The matrix A is said to be positive definite if $\boldsymbol{x}^T A\boldsymbol{x} > 0$ for all $\boldsymbol{x} \neq 0$. A weaker statement is to say that A is positive semidefinite, where in this case, $\boldsymbol{x}^T A\boldsymbol{x} \geq 0$ for $\boldsymbol{x} \neq 0$. Compare this computation to was we said was the second derivative in the direction of $\boldsymbol{u}$.

Example: Any diagonal matrix with positive values along the diagonal is a positive definite matrix. As a specific example, consider:

$$[x, y]\begin{bmatrix} 2 & 0 \\ 0 & 5 \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix} = [x, y]\begin{bmatrix} 2x \\ 5y \end{bmatrix} = 2x^2 + 5y^2$$

which is only zero if $x = y = 0$.

Example: If a matrix A has all positive eigenvalues, it is positive definite. We'll go through this in more detail in the section on linear algebra, but heuristically, you might imagine that the proof will use a diagonalization of A so that we're back to the previous example.

Let us consider the quadratic function in detail, as we will use this as a model for non-quadratic functions.

Let $F(x) = \frac{1}{2}\boldsymbol{x}^T A\boldsymbol{x} + \boldsymbol{b}^T\boldsymbol{x} + c$, where A is symmetric and $\boldsymbol{b}, c$ are constants. Then

$$\nabla F(\boldsymbol{x}) = A\boldsymbol{x} + \boldsymbol{b}$$

If we assume A is invertible, then the stationary point of F:

$$\boldsymbol{x} = -A^{-1}\boldsymbol{b}$$

and the Hessian does not depend on this value:

$$HF = A$$

So, if A is positive definite, the stationary point holds the unique minimum of F. Compare these computations to those we do in Calc I:

$$y = \frac{1}{2}ax^2 + bx + c$$

B.3.1 Necessary and Sufficient Conditions

Theorem: If f attains its local minimum (maximum) at $\boldsymbol{x} = \boldsymbol{a}$, then $\nabla f(\boldsymbol{a}) = \mathbf{0}$. The vector $\boldsymbol{a}$ is called a *stationary point* of f.

This is just the multidimensional version of Fermat's Theorem from Calculus I- we require all partial derivatives to vanish if we are at a local maximum or minimum.

Theorem: Necessary and Sufficient Conditions for Extrema

- First Order Condition: A necessary condition for $f : \mathbb{R}^n \to \mathbb{R}$ to have a local minimum (maximum) at $\boldsymbol{x} = \boldsymbol{a}$ is that $\boldsymbol{a}$ is a stationary point.
- Second Order Condition: Let $\boldsymbol{x} = \boldsymbol{a}$ be a stationary point. Then a necessary condition for f to have a local minimum (maximum) at $\boldsymbol{x} = \boldsymbol{a}$ is that $Hf(\boldsymbol{a})$ is positive semidefinite (negative semidefinite).

  A sufficient condition for a minimum (maximum) is that $Hf(\boldsymbol{a})$ is positive definite (negative definite).

We saw that the minimum was easy to compute if the function f was in a special form. We can force our function to have the quadratic form by appealing to the Taylor expansion.

Taylor's Theorem: Let $f : \mathbb{R}^n \to \mathbb{R}$ be analytic (That is, we'll assume all derivatives are continuous). Then the Taylor's series for f at $\boldsymbol{x} = \boldsymbol{a}$ is:

$$f(\boldsymbol{x}) = f(\boldsymbol{a}) + \nabla f(\boldsymbol{a})(\boldsymbol{x} - \boldsymbol{a}) + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{a})^T Hf(\boldsymbol{a})(\boldsymbol{x} - \boldsymbol{a}) + h.o.t.$$

where h.o.t. stands for "higher order terms". Recall that $\nabla f(\boldsymbol{a})$ is an $1 \times n$ vector, and $Hf(\boldsymbol{a})$ is an $n \times n$ symmetric matrix.

We will only compute Taylor's expansion up to second order, but for the sake of completeness (and your amusement!), we give the definition of the full Taylor expansion below.

In order to understand the notation, think about all the different partial derivatives we will have to compute. For example, if f depends on x, y, z, then there are 3 first partials, 9 second partials, 27 third partials, etc. Thus, we need a way of organizing all of these. Probably the best way to do this is with the following notation:

$$f_{i_1 i_2 \ldots i_k} = \frac{\partial^k f}{\partial x_{i_1} \partial x_{i_2} \ldots \partial x_{i_k}}$$

so the indices $i_1, i_2, \ldots, i_k$ are some permutation of the indices of $\boldsymbol{x}$ taken k at a time. Now we're ready to go!

The Taylor series expansion for f at $\boldsymbol{x} = \boldsymbol{a}$, where $\boldsymbol{x} \in \mathbb{R}^n$ is given by:

$$f(\boldsymbol{x}) = f(\boldsymbol{a}) + \sum_{i=1}^{n} f_i(\boldsymbol{a})(x_i - a_i) + \frac{1}{2!}\sum_{i_1=1}^{n}\sum_{i_2=1}^{n} f_{i_1 i_2}(\boldsymbol{a})(x_{i_1} - a_{i_1})(x_{i_2} - a_{i_2}) +$$

$$\ldots + \frac{1}{k!}\sum_{i_1=1}^{n}\cdots\sum_{i_k=1}^{n} f_{i_1 i_2 \ldots i_k}(\boldsymbol{a})(x_{i_1} - a_{i_1})\cdots(x_{i_k} - a_{i_k}) + \ldots$$

B.4 Worked Examples

1. Compute the (full) Taylor series expansion for

$$f(x, y) = x^2 + 2xy + 3y^2 + xy^2$$

   at $(1, 1)$.

2. Suppose we have a function f so that $f((0, 1)) = 3$, and $\nabla f((0, 1)) = [1, -2]$. Use the linear approximation to f to get an approximation to $f((0.5, 1.5))$.

3. Let $f(x, y) = xy + y^2$. At the point $(2, 1)$, in which direction is f increasing the fastest? How fast is it changing?

4. Suppose $A = \begin{bmatrix} 2 & 1 \\ 0 & -1 \end{bmatrix}$. Show that A is neither positive definite or negative definite by finding an $\boldsymbol{x}$ so that $\boldsymbol{x}^T A\boldsymbol{x} > 0$, and another $\boldsymbol{x}$ so that $\boldsymbol{x}^T A\boldsymbol{x} < 0$.

SOLUTIONS TO THE WORKED EXAMPLES:

1. First we see that f is a third degree polynomial (because of the xy^2 term). Therefore, we will need to compute the partial derivatives up to the third partials:

$$f_x = 2x + 2y + y^2, \quad f_y = 2x + 6y + 2xy$$

$$\text{so at } x = 1, y = 1 : f_x = 5, \quad f_y = 10$$

$$f_{xx} = 2, f_{xy} = f_{yx} = 2 + 2y = 4, f_{yy} = 6 + 2x = 8$$

$$f_{xxy} = 0 \quad f_{xxx} = 0 \quad f_{xyx} = 0 \quad f_{xyy} = 2$$

$$f_{yxx} = 0 \quad f_{yxy} = 2 \quad f_{yyx} = 2 \quad f_{yyy} = 0$$

   Now we get that:

$$f(x) = 7 + 5(x - 1) + 10(y - 1) + \frac{2}{2}(x - 1)^2 + \frac{8}{2}(x - 1)(y - 1) +$$

$$\frac{8}{2}(y-1)^2 + \frac{6}{6}(x-1)(y-1)^2$$

Note that the coefficient for $(x-1)(y-1)$ came from both f_{xy} and f_{yx}, and Note that the $(x-1)(y-1)^2$ coefficient came from all three partials: f_{xyy}, f_{yxy} and f_{yyx}.

2. The linearization of f:

$$f(x, y) = 3 + [1, -2] \begin{bmatrix} x \\ y \end{bmatrix} = 3 + x - 2y$$

so $f(0.5, 1.5)$ is approximately $3 + 0.5 - 3 = 0.5$.

3. The gradient is $[y, x + 2y]$, so at $(2, 1)$, we get $[1, 4]$. The direction in which f is increasing the fastest is in the direction of $[1, 4]^T$, and the rate of change in that direction is $\sqrt{17}$.

4. We see that $\boldsymbol{x}^T A\boldsymbol{x} = 2x^2 + xy - y^2$, so we can choose, for example, $x = 0, y = 1$ for a negative value, and $x = 1, y = 0$ for a positive.

B.5 Exercises

1. Let $f(x, y) = x^2y + 3y$. Compute the linearization of f at $x = 1, y = 1$.

2. Let

$$f(x, y, z, w) = \begin{bmatrix} x + zw - w^2 \\ 4 - 3xy - z^2 \\ \mathrm{e}^{xyzw} \\ x^2 + y^2 - 2xyw \end{bmatrix}$$

Compute the linearization of f at $(1, 2, 3, 0)^T$.

3. Definition: Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$. The Directional Derivative of f in the direction of the vector $\boldsymbol{u}$ is given by:

$$D_{\boldsymbol{u}} f = \nabla f \cdot \boldsymbol{u}$$

which is the dot product of the gradient with the vector $\boldsymbol{u}$. We interpret this quantity (which is a scalar) as "The rate of change of f in the direction of $\boldsymbol{u}$".

(a) What does the directional derivative give if $\boldsymbol{u} = \boldsymbol{e}^{(i)}$, the i^{th} standard basis vector?

(b) Consider f at some fixed value, $x = a$. Over all possible *unit* vectors $\boldsymbol{u}$, in which direction will f have it's maximum (minimum) rate of change? Hint: Think about how we can write the dot product in terms of $\cos(\theta)$, for an appropriate θ.

4. Let $\boldsymbol{x}$ be the variable, $A, \boldsymbol{b}$ be a constant matrix, vector respectively. Show that:

   (a) $D\boldsymbol{x} = I$

   (b) $\nabla\left(\boldsymbol{x}^T\boldsymbol{x}\right) = 2\boldsymbol{x}$

   (c) $\nabla\left(\boldsymbol{b}^T\boldsymbol{x}\right) = \boldsymbol{b}$

   (d) $D(A\boldsymbol{x}) = A$

5. Show that, if $\boldsymbol{u}, \boldsymbol{v}$ are vector-valued functions from $\mathbb{R}^n \to \mathbb{R}^n$, then:

$$\nabla\left(\boldsymbol{u}^T\boldsymbol{v}\right) = (D\boldsymbol{u})^T\,\boldsymbol{v} + (D\boldsymbol{v})^T\,\boldsymbol{u}$$

6. Use the previous exercises to show that, if G is a symmetric matrix, then $\nabla\left(\frac{1}{2}\boldsymbol{x}^T G\boldsymbol{x}\right) = G\boldsymbol{x}$.

7. Multiply the following expression out to its scalar representation:

$$[x, y]\begin{bmatrix} a & b \\ c & d \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix}$$

   Use this result to write: $5x^2 - xy + 6y^2$ as $\boldsymbol{x}^T A\boldsymbol{x}$, for a suitable matrix A.

8. Let $f(x, y) = 3x^2 + xy - y^2 + 3x - 5y$.

   (a) At the point $(1, 1)$, in which direction is f increasing the fastest? Decreasing the fastest? What is its rate of change in the fastest direction?

   (b) Compute the Hessian of f.

   (c) Rewrite f in the form: $\frac{1}{2}\boldsymbol{x}^T A\boldsymbol{x} + \boldsymbol{b}^T\boldsymbol{x}$, for a suitable matrix A and vector $\boldsymbol{b}$.

   (d) Find the stationary point of f using our previous formula, and verify that the gradient is zero there. Show that we do not have a minimum or a maximum at the stationary point.