MATH4ML WORKSHEET

1. FUNCTIONS OF A SINGLE VARIABLE

(1) Is the function $\cos(\pi - x)$ increasing or decreasing at $x = \frac{\pi}{2}$? In other words, if the current input into the function is $\frac{\pi}{2}$, should I increase or decrease the input slightly if I want the output of the function to increase?

(2) Suppose two functions $f(x)$ and $g(x)$ are both increasing at $x = 1$. Can we say whether $f(g(x))$ is increasing, decreasing or flat at $x = 1$?

(3) List all points where the functions $|x|$ and $|x|^2$ are not differentiable.

(4) In this question we will determine the value of the derivative of the function

$$K(x) = \Big(\big((x^2+1)^2 + 2 \big)^2 + 3 \Big)^2 + 4$$

with respect to x when $x = -1$.

(a) Find the derivatives of each of the functions provided below with respect to the input variable x:
- $f(x) = x^2 + 1$
- $g(x) = x^2 + 2$
- $h(x) = x^2 + 3$
- $k(x) = x^2 + 4$.

(b) Confirm that $K(x) = k\big(h\big(g(f(x))\big)\big)$.

(c) We will now begin building an expression for the derivative piece-by-piece. First, write $G(x) = g(f(x))$ and determine an expression for $G'(x)$ in terms of g', f' and f by applying the chain rule. Then substitute in the expressions for g', f' and f to get an expression in terms of x.

(d) Next, write $H(x) = h(G(x)) = h(g(f(x))$ and determine an expression for $H'(x)$ in terms of h', G' and G by applying the chain rule. Reduce to an expression in x as before.

(e) Finally, notice $K(x) = k(H(x)) = k(h(g(f(x)))$ and determine an expression for $K'(x)$ in terms of k', H' and H by applying the chain rule. Reduce to an expression in x as before.

(f) The expression you found above gives the derivative for all values of x, so plug in $x = -1$ to determine the value we set out to find.

(g) Give yourself a pat on the back for completing a rather tedious computation.

(5) In the previous question, we determined a symbolic expression for the derivative of a complicated function. But we only cared about the value of this derivative at a single point (namely $x = -1$). We will now determine this value without computing a full expression for the derivative.

Let f, g, h, k and K be as in Problem 4 and answer the questions below.

Date: 26 August 2025.

(a) Fill out the diagram below as the input value $x = -1$ gets successively transformed into the output by applying f, g, h and k in that order. The entries should all be (real) numbers and not symbolic expressions involving 'x'.

$$\begin{array}{ccccccccc} \boxed{-1} & \xrightarrow{\quad f \quad} & \boxed{} & \xrightarrow{\quad h \quad} & \boxed{} & \xrightarrow{\quad g \quad} & \boxed{} & \xrightarrow{\quad k \quad} & \boxed{} \\ x & & f(x) & & g(f(x)) & & h(g(f(x)) & & k(h(g(f(x)))) \end{array}$$

(b) Now fill out this next diagram from right to left, using the values from above evaluate the expressions for the derivatives you know for f, h, g and k from Problem 4a. Once again, no symbolic expressions!

$$\begin{array}{ccccccccc} \boxed{} & \xleftarrow{\quad f' \quad} & \boxed{} & \xleftarrow{\quad h' \quad} & \boxed{} & \xleftarrow{\quad g' \quad} & \boxed{} & \xleftarrow{\quad k' \quad} & \boxed{1} \\ f'(x) & & g'(f(x)) & & h'(g(f(x)) & & k'(h(g(f(x))) & & \end{array}$$

(c) Obtain the final value for the derivative by multiplying the values in each individual box in the second diagram.

You have just implemented a simple version of ***backpropagation****, which is how modern neural networks keep track of derivatives to help their learning. We will revisit this in more detail in DSCI 572.*

(6) In machine learning, we often minimize a loss function. For **convex** loss functions (e.g. mean squared error in linear regression), we are guaranteed that gradient descent will reach the global minimum, however for non-convex loss functions (e.g. deep neural networks), gradient descent may only find a local minimum. In this exercise, we will explore why this might be the case.

Let $f(x)$ be a convex function of one variable. Answer the following:

(a) Suppose x_0 is a local minimum of f and $x_1 \neq x_0$ is some other point not equal to x_0. What can we say about the line joining the points $(x_0, f(x_0))$ and $(x_1, f(x_1))$? Where does this line sit in relation to the graph of $f(x)$?

(b) Show that if $f(x)$ attains a local minimum at some point x_0, then it actually attains a **global** minimum at x_0. *Hint: Let x_1 be a global minimum, and make use of part 6a.*

(7) Classify the functions given below as convex or non-convex, and discuss whether gradient descent would always succeed in finding the global minimum if these were loss functions:

(a) $\sin(\pi x) + 1 - e^{-x^2}$ (b) x^3 (c) $|x|^3$
(d) $\log(1 + e^x)$ (e) $x^2 - 3x$ (f) $|x^2 - 3x|$
(g) $f(x) = \max\{x^2 - 2x + 7, e^{\sqrt{x}}\}$

Hint: Drawing graphs may help you visualize some of these functions!

(8) **Bonus:** Suppose you have a dataset with 300 data points. How many different ways can these data points be divided into two classes 'A' and 'B'? In other words, how many functions are there from a set with 300 elements to a set with 2 elements?

2. FUNCTIONS OF MULTIPLE VARIABLES

(1) Consider the following three plots:

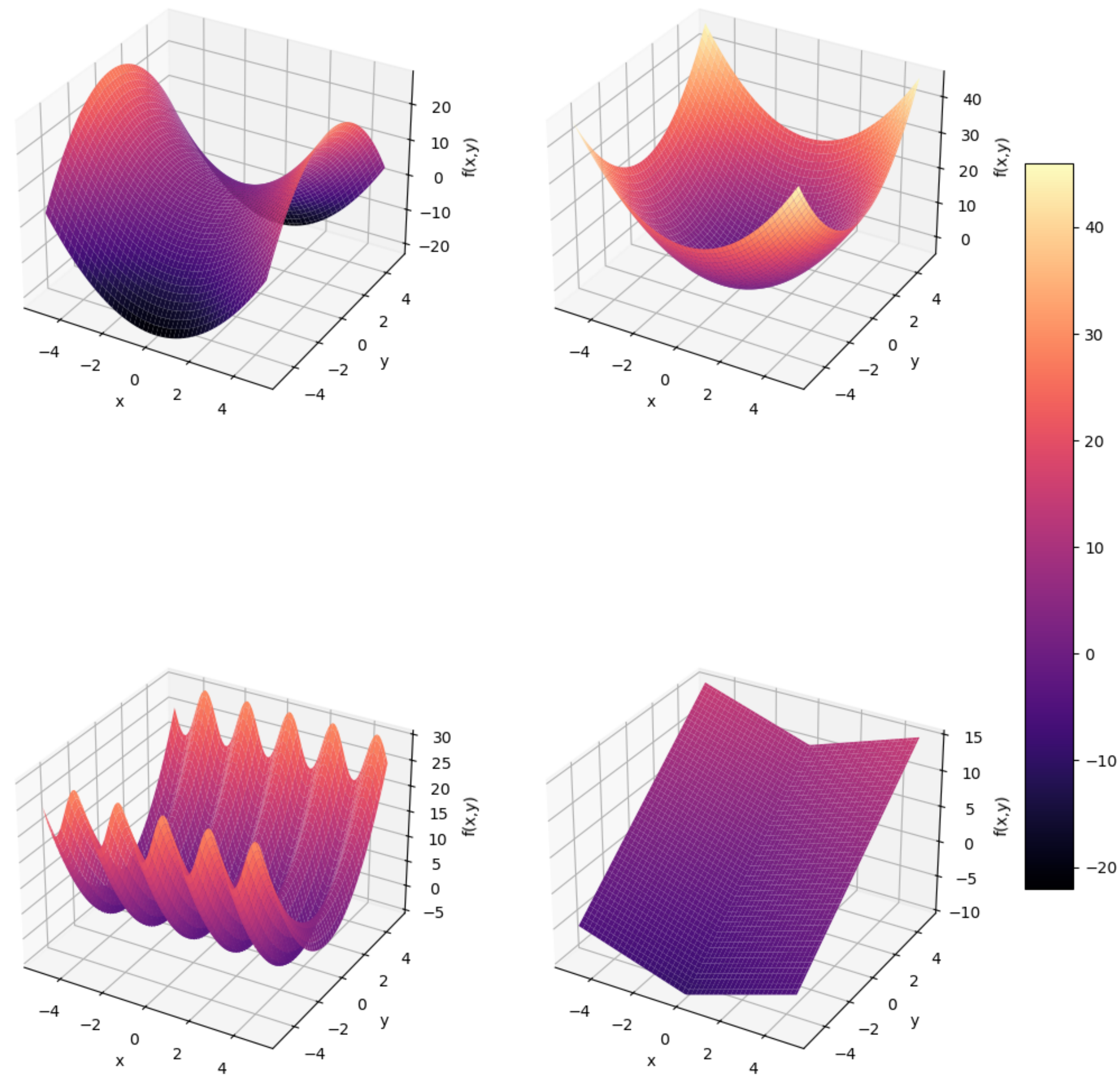


FIGURE 1. Some plots

For each plot, select the function from the following list whose graph that plot represents.

(a) $f(x, y) = x^2 + y^2 - 4$ (b) $f(x, y) = x - 4y + 7$ (c) $f(x, y) = x^2 - y^2 + 3$
(d) $f(x, y) = |x| + 3y$ (e) $f(x, y) = y^2 + \sin x$

(2) Find a continuous function of two variables that:

(a) has a single global minimum.

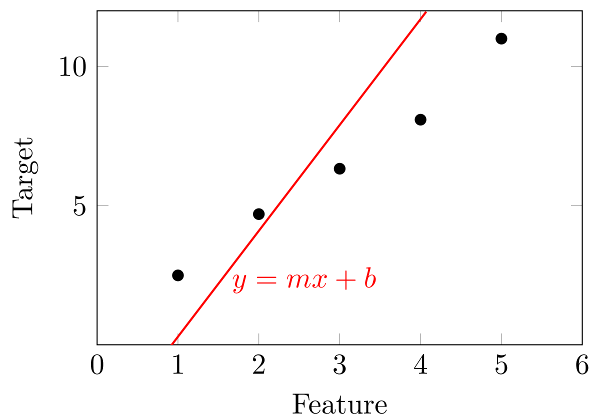
(b) has a global minimum at every point with $x = 0$.

(c) **Bonus:** has a global minimum at every point along the line $3x - 4y = 0$.

(3) You are trying to fit a least-squares regression model to the following data :

Feature	Target
1	2.5
2	4.7
3	6.33
4	8.09
5	11

The data is plotted in the figure below with "Feature" on the x-axis and "Target" on the y-axis, with an initial 'guess' for a line with slope m and y-intercept b in red.



Answer the following:

(a) What is the value predicted by our linear regression (red line) when $x = 2$? Write your answer in terms of m and b.

(b) What is the *squared* error for the data point $x = 4$, i.e. the difference between predicted and actual values? Again, your answer should involve m and b

(c) Write a complete expression for the *total squared error* between predicted and actual values for the target, i.e. the sum of individual errors for each data point.

(d) How many "inputs" does the total squared error function accept? In other words, the total squared error is a function of how many variables?

(e) Find the gradient vector for the total squared error function (in terms of m and b).

(f) Find the explicit gradient vector at $m = b = 0$.

(g) Find a linear function of m and b that has the same gradient at $m = b = 0$ as your total squared error function above.

(h) **Bonus:** Find the linear function that best approximates your total squared error function near $m = b = 1$. Hint: what would the gradient of this linear function need to be? And what must be its value at $m = b = 1$?

(4) Let A be the matrix given by

$$A = \begin{bmatrix} 4 & 0 \\ 0 & \frac{1}{2} \end{bmatrix}.$$

For a 2D vector $\bar{\mathbf{x}} = [x_1, x_2]^T$, define

$$f(\bar{\mathbf{x}}) = e^{-\bar{\mathbf{x}}^T A \bar{\mathbf{x}}}$$

(a) Set $x_2 = 0$ and draw a graph of the resulting (univariate) function with repsect to x_1.

(b) Draw a similar graph with respect to x_2 after setting $x_1 = 0$.

(c) Is $f([x_1, 0]^T)$ bigger or smaller than $f[x_1, 10]^T)$? Where does the function f attain its maximum value?

(d) Draw a graph of the overall function f as a function of 2 variables.

(e) **Bonus:** Now set

$$A = \begin{bmatrix} 9 & 7 \\ 7 & 9 \end{bmatrix} \qquad \text{and} \qquad \mu = \begin{bmatrix} 10 \\ 10 \end{bmatrix}$$

and define

$$f(\bar{\mathbf{x}}) = e^{-(\bar{\mathbf{x}}-\mu)^T A(\bar{\mathbf{x}}-\mu)} .$$

What does the graph of $f(\bar{x})$ look like in this case?

Note: The functions you graphed in this question are closely related to ***multivariate Gaussians****, which are like normal distributions in higher-dimensional space. The matrix A controls the shape and orientation of the 'bell' curve.*

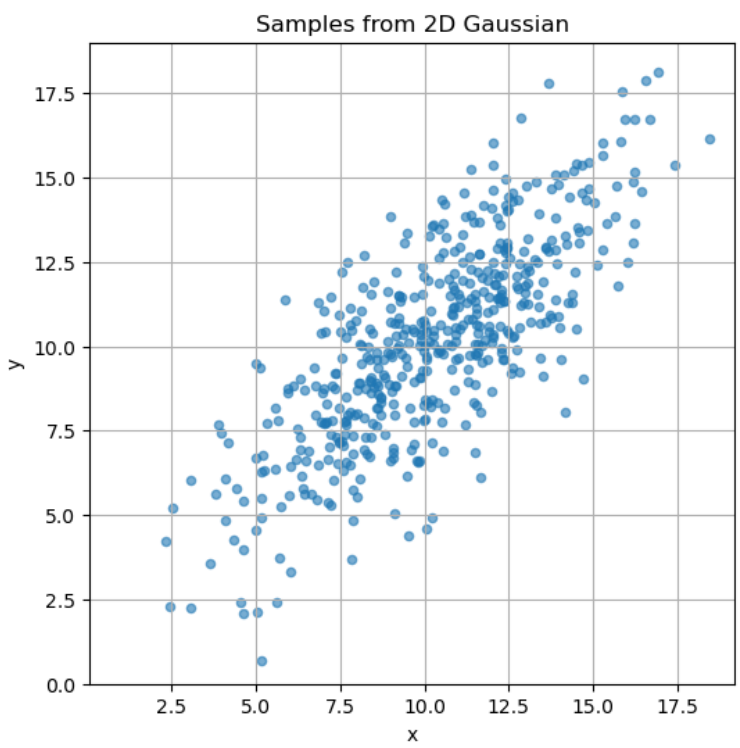


FIGURE 2. A collection of 500 data points randomly sampled from a two-variable Gaussian with mean and covariance corresponding to μ and A from Problem 4e.

3. BASIC MATRIX OPERATIONS

(1) Let $v_1 = (2, 3, -1)$, $v_2 = (1, 0, 1)$ and $v_3 = (0, 0, 4)$. Calculate the sum $3v_1 - 2v_2 + v_3$.

(2) Calculate the value of the matrix product

$$\begin{bmatrix} 2 & 1 & 0 \\ 3 & 0 & 0 \\ -1 & 1 & 4 \end{bmatrix} \cdot \begin{bmatrix} 3 \\ -2 \\ 1 \end{bmatrix}$$

Explain in one or two sentences how this calculation relates to the sum from the previous section.

(3) Find a 5×1 vector v such that for any 5×5 matrix A, the product Av returns the last column of A.

(4) Let A be a 4×4 matrix. Find a matrix B such that the product AB evaluates to a matrix that
   (a) contains the columns of A in reverse order (from last to first).
   (b) contains only the second column of A.
   (c) is identical to A.
   (d) is such that its ith column is the sum of the first i columns of A.

(5) Let A be a 7×7 matrix. Can you find 7×7 matrices U and V such that the product UAV contains zeros everywhere except the one entry in row 3 and column 4?

(6) Consider the matrix

$$A = \begin{bmatrix} \frac{1}{3} & -4 \\ 2 & \frac{1}{2} \end{bmatrix}$$

and define a function f (with input and output 2-dimensional) using the formula

$$f(x, y) = f\left(\begin{bmatrix} x \\ y \end{bmatrix}\right) = A \cdot \begin{bmatrix} x \\ y \end{bmatrix}$$

.

   (a) What is $f\left(\begin{bmatrix} 1 \\ 0 \end{bmatrix}\right)$? How about $f\left(\begin{bmatrix} 0 \\ 1 \end{bmatrix}\right)$?
   (b) How do these values relate to the matrix A?
   (c) Can you visualize what the function f is doing in the xy plane by visualizing its action on the coordinate axes? Does it appear to 'stretch' or 'shrink' vectors in the xy plane? Does it rotate them?
   Hint: Plot the coordinate vectors in the 2D plane before and after applying this transformation.

(7) Find a 3×3 matrix A such that for any 3×1 vector v, the product Av returns a vector that points along the direction $\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$.

(8) Consider the matrix A and vectors v and w given by

$$A = \begin{bmatrix} 1 & 0 & 3 \\ -0 & -1 & 1 \\ -4 & 4 & 2 \end{bmatrix}, \qquad v = \begin{bmatrix} 0 \\ 1 \\ -1 \end{bmatrix}, \qquad w = \begin{bmatrix} 3 \\ 1 \\ -4 \end{bmatrix}.$$

Show that the dot product of w with Av is the same as the dot product of $A^T w$ with v. Would this observation generalize to other matrices and vectors?

(9) If v is an $n \times 1$ vector and A is an $n \times n$ matrix, then the function f defined by

$$f(v) = v^T A v$$

is a *real-valued* function of n variables i.e. it takes an n-dimensional vector as input and outputs a real-number. Thus, we can ask about the gradient of f. In this question, we will compute this gradient in the case that A is a symmetric matrix.

(a) The first entry in Av is formed by moving along the first row of A, and taking the product of the jth entry in that row with the jth entry in v, and summing the n individual products (i.e. the n columns of A). Similarly the second entry of Av is formed by repeating this process with the second row of A.

Derive an expression for the ith entry of the vector Av and call it a_i. Write this expression below, it should involve a sum across n different terms:

$$a_i = \sum_{j=1}^{n} \boxed{} .$$

(b) Now derive an expression for the value $v^T A v$ involving the a_i. This should also involve a sum of n terms (remember Av is a column vector, just like v!).

$$v^T A v = \sum_{i=1}^{n} \boxed{} \cdot a_i .$$

(c) Substitute in the expression for a_i you found earlier to write this answer as a double sum:

$$v^T A v = \sum_{i=1}^{n} \sum_{j=1}^{n} \boxed{}$$

(d) Notice that the answer you wrote above is a one-line expression for the value of the quantity $v^T A v$. Differentiate this expression with respect to the kth entry in v (i.e. the kth input variable for the function f):

$$\frac{\partial f}{\partial v_k} = \boxed{}$$

(e) Use your answer to the previous question along with the assumption that A is **symmetric** to conclude that

$$\nabla f = 2A\beta$$

(f) We now return to the linear regression problem from the previous section of this worksheet. Consider the matrix

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

along with the vectors

$$w = \begin{bmatrix} b \\ m \end{bmatrix} \qquad \text{and} \qquad y = \begin{bmatrix} 2.5 \\ 4.7 \\ 6.33 \\ 8.09 \\ 11 \end{bmatrix}$$

Answer the following:

(i) Show that the expression $\|Xw - y\|^2$ evaluates precisely to the total squared error you calculated in Problem 3 from Section 2. (Here $\|\cdot\|$ denotes the norm of a vector.)

(ii) Check that

$$\begin{aligned} \|Xw - y\|^2 &= (Xw - y)^T \cdot (Xw - y) \\ &= w^T X^T Xw - w^T X^T y - y^T Xw - y^T y. \end{aligned}$$

(iii) Explain why $\nabla(y^T y) = 0$ and

$$\nabla(w^T X^T y) = \nabla(y^T Xw) = X^T y$$

Hint: What are the variables here that we want to find derivatives with respect to?

(iv) Determine $\nabla w^T X^T Xw$ using the fact that $X^T X$ is symmetric and recalling Problem 9 above.

(v) Write an expression for the gradient of $\|Xw - y\|^2$ in terms of X, w and y.