,

Important: This homework tests basic mathematical skills required for the course. Use it as a self-test of your background knowledge and try to do it **without looking anything up**. If you are having to look things up a lot, or if more than 20% of concepts are unfamiliar to you, you will struggle in the course. Note that the order/points of questions do not always imply difficulty.

Grading: The homework assignments in this course are self-graded. Please complete the homework and submit a PDF file (**the only accepted file format**). Typed or scanned handwritten solutions are fine. Once the solutions are available, please use them to assign points to your own **submitted** solution and submit the points. Refer to the course website for due dates and submission links. A subset of the assignments will be chosen at random and double checked by the course staff. It is in your best interest to complete the assignments as they will prepare you for the exams.

Total: 50 points.

Self-Grading Guidelines: Follow the solutions below to assign points to the solutions you handed in. In general, if your solution was correct and included all derivation steps, you should give it full points. If your solution was "on the right track" but you made one or two minor mathematical mistakes or omitted one or two derivation steps, then subtract 1 point. If you attempted to solve the question and gave detailed steps that nonetheless lead to an incorrect or incomplete solution, or just gave the solution without derivation or explanation of how you arrived at it, then give yourself half of the points (rounded up). Otherwise, assign zero points.

1 Basic Calculus [10 pts]

The following questions test your basic skills in computing the derivatives of univariate functions, as well as applying the concept of *convexity* to determine the properties of the functions.

(a) (3 pts) Find all extrema of the function $f(x) = ln(1 - x^2)$. For each extremum, state if it is a maximum or a minimum.

SOLUTION:
Extrema points of a function are either its maxima or minima, and must occur at critical points. critical points are points withing the domain of the function, which either have a gradient $f'(x) = 0$ or are non-differentiable. If the points is differentiable one can distinguish whether a critical point is a local maximum or local minimum second derivative test or higher-order derivative test, given sufficient differentiability. For example to find extrema of a twice-differentiable function using the second derivative test, we find all points where $f'(x) = 0$, then for each extremum point x, if $f''(x) > 0$ then it's a minimum, and if $f'' < 0$ then it is a maximum (otherwise the test is inconclusive). See https://en.wikipedia.org/wiki/Maxima_and_minima for more details.
For this example, $f'(x) = 2x/(x^2 - 1)$ with critical point at $x = 0$ and $f''(x) = -\frac{2(x^2+1)}{(x^2-1)^2}$ with $f''(0) < 0$ so $x = 0$ is a local (and global) maximum.

(b) (3 pts) Show that $f(x) = \ln \frac{1}{1+e^x}$ is concave.

SOLUTION:
A differentiable function f is concave on an interval if and only if its derivative function f' is monotonically decreasing on that interval, that is $f'' < 0$. See https://en.wikipedia.org/wiki/Concave_function for more details.
In this case,
$f'(x) = -e^x/(e^x + 1)$;
$f''(x) = -e^x/(e^x + 1)^2$
Since f'' is always negative, the function is concave.

(c) (4 pts) Show that $f(x) = e^{-x^2/2}$ is neither convex nor concave.

SOLUTION:
$f'(x) = -xe^{-x^2/2}$; $f''(x) = (x^2 - 1)e^{-x^2/2}$;
$f''(0) = (-1)(1) < 0$, $f''(2) = (3)(1/e^2) > 0$,
therefore f'' is neither positive everywhere nor negative everywhere, and thus f is neither concave nor convex.

2 Continuous Random Variables [10 pts]

(a) (2 pts) Given a continuous random variable X with probability density function $f(X)$, what are the expressions for the mean and variance of this variable?

SOLUTION:
This question is just asking for the definitions. The mean or expected value is

$$E(X) = \int_{-\infty}^{\infty} X f(X) dX$$

The variance is

$$E((X - E(X))^2)$$

(b) (2 pts) Can the value of the probability density function (PDF) $f(X)$ exceed 1? Why or why not?

SOLUTION: Yes. Remember that we cannot think of $f(X)$ as the probability of observing any specific value of X. Unlike a probability mass function *defined over a discrete r.v., the PDF does not have to fall between $[0, 1]$. See https://en.wikipedia.org/wiki/Probability_mass_function As an example, think of the PDF in question (c) below, if the interval where X is defined was less than 1. Also see this video for an intuitive explanation of PDFs.*

(c) (2 pts) Consider a random variable X that follows the *uniform distribution* between a and 1, i.e. its PDF is equal to a constant c on this interval, and 0 otherwise. Derive c in terms of a.

SOLUTION: The PDF is given by $p(x) = c$. Since the PDF should integrate to 1, $\int_a^1 c\, dx = 1$. Solving for c we get $c = 1/(1 - a)$.

(d) (2 pts) Derive the expected value of X in terms of a and b. Show all your steps.

SOLUTION: This should be $\int_a^1 x dx/(1 - a)$, or $(1 + a)/2$.

(e) (2 pts) Derive the cumulative distribution function $F(X)$ on the interval $a \leq X \leq 1$.

SOLUTION: The cumulative distribution function, or CDF, of a probability density function $f(t)$ is defined as

$$F(X) = \int_{-\infty}^{X} f(t) dt$$

In this case, we get

$$F(X) = \int_{a}^{X} 1/(1 - a) dt = \frac{X}{1 - a} - \frac{a}{1 - a} = \frac{X - a}{1 - a}$$

See https://en.wikipedia.org/wiki/Cumulative_distribution_function

3 Discrete Random Variables [10 pts]

(a) (2 pts) Two students taking a Machine Learning class became project partners. They are trying to decide what operating system to use for the project. Suppose each student has a laptop, which could be one of three types: Mac OS, Windows, or Linux. If the distribution of laptops among students follows the PDF shown below, what is the probability that the two teammates have **different** laptops?

Mac OS	0.5
Windows	0.3
Linux	0.2

SOLUTION: One way to solve this is to first compute the probability that the two students, call them X and Y, have the same laptops (m, w or l), i.e. p(same):

$$p(X = m, Y = m) + p(X = w, y = w) + p(X = l, Y = l) = (.5)(.5) + (.3)(.3) + (.2)(.2) = .38$$

Then the probability the have different laptops p(different) = 1 - p(same) = .62.

Suppose we have three discrete random variables x, y and z that take values 0 or 1 according to the distribution below.

		$z = 0$	$z = 1$
$x = 0$	$y = 0$	$\frac{1}{12}$	$\frac{1}{12}$
	$y = 1$	0	$\frac{1}{4}$

		$z = 0$	$z = 1$
$x = 1$	$y = 0$	0	$\frac{1}{12}$
	$y = 1$	$\frac{1}{4}$	$\frac{1}{4}$

(b) (2 pts) Find the joint distribution of y and z

SOLUTION: The key here is to understand that the above is the joint probability (not conditional) over x, y, z. Therefore, using the sum rule, $p(y, z) = p(x = 1, y, z) + p(x = 0, y, z)$, and the corresponding table is obtained by summing the two tables above.

	$z = 0$	$z = 1$
$y = 0$	$\frac{1}{12}$	$\frac{1}{6}$
$y = 1$	$\frac{1}{4}$	$\frac{1}{2}$

(c) (2 pts) Find the marginal distributions of y and z

SOLUTION: Again, using the sum rule, we get $p(y) = p(y, z = 0) + p(y, z = 1)$ and plugging in the values we found in (b), and similarly for $p(z)$. This process of summing over a variable is called marginalization.

	$y = 0$	$y = 1$
$p(y)$	$\frac{1}{4}$	$\frac{3}{4}$

	$z = 0$	$z = 1$
$p(z)$	$\frac{1}{3}$	$\frac{2}{3}$

(d) (2 pts) Find the conditional distribution of x given that $y = 0$.

SOLUTION: Here we can use the product rule, $p(x, y = 0) = p(x|y = 0)p(y = 0)$, so

$p(x|y=0) = p(x, y=0)/p(y=0)$, *and plug in the values from previous steps.*

	$x=0$	$x=1$
$p(x\|y=0)$	$\frac{2}{3}$	$\frac{1}{3}$

(e) (2 pts) Are y and z independent? Explain.

SOLUTION: Yes. We can show this by proving that $p(y,z) = p(y)(p(z)$, *which happens to be true in this case.*

4 Basic Linear Algebra [10 pts]

(a) (3 pts) Let $\boldsymbol{A}$ be a 4x3 matrix, $\boldsymbol{B}$ be a 5x4 matrix, and $\boldsymbol{C}$ be a 4x4 matrix. Determine which of the following products are defined and find the size of those that are defined. Note, X^T refers to the transpose of X.

SOLUTION:

$\boldsymbol{AB}$: undefined

$\boldsymbol{BA}$: yes, 5x3

$\boldsymbol{A^T C}$: yes, 3x4

$\boldsymbol{CA^T}$: undefined

$\boldsymbol{BC^T}$: yes, 5x4

$\boldsymbol{CB}$: undefined

(b) (3 pts) Suppose we would like to predict the profits of "Sunny Coffee", a bakery chain with locations in three different cities. Given the price of flour x, price of sugar y and price of oil z, the profit can be modelled as a linear function of these variables. That is, for each of the locations $i = 1, ..., 3$, the profit is $p_i = a_i + b_i x + c_i y + d_i z$.

Write down the matrix-vector product that produces the 3-dimensional vector of profits for the three locations.

SOLUTION:

Let $A = [a_1\ b_1\ c_1\ d_1;\ a_2\ b_2\ ...]$, $b = [1\ x\ y\ z]'$, and $p = [p_1\ p_2\ p_3]'$, then the required matrix-vector product is

$$A * b = p$$

(c) (4 pts) Let $\boldsymbol{A}$ and $\boldsymbol{B}$ be two $\mathbb{R}^{\mathsf{D}\times\mathsf{D}}$ symmetric matrices. Suppose $\boldsymbol{A}$ and $\boldsymbol{B}$ have the exact same set of eigenvectors $\boldsymbol{u}_1, \boldsymbol{u}_2, \cdots, \boldsymbol{u}_{\mathsf{D}}$ with the corresponding eigenvalues $\alpha_1, \alpha_2, \cdots, \alpha_{\mathsf{D}}$ for $\boldsymbol{A}$, and $\beta_1, \beta_2, \cdots, \beta_{\mathsf{D}}$ for $\boldsymbol{B}$. Write down the eigenvectors and their corresponding eigenvalues for the following matrices. (*Hint.* Represent $\boldsymbol{A}$, $\boldsymbol{B}$ using the eigenvectors, e.g., $\boldsymbol{A} = \sum_d \alpha_d \boldsymbol{u}_d \boldsymbol{u}_d^{\mathrm{T}}$.)

SOLUTION:

- $\boldsymbol{C} = \boldsymbol{A} + \boldsymbol{B}$

  Re-arranging terms in the sums, $C = \sum_d \alpha_d u_d u_d^T + \sum_d \beta_d u_d u_d^T = \sum_d (\alpha_d + \beta_d) u_d u_d^T$. Therefore the eigenvalues are $(\alpha_d + \beta_d)$ and eigenvectors are u_d.

- $\boldsymbol{D} = \boldsymbol{A} - \boldsymbol{B}$.

  Similarly, the eigenvalues are $(\alpha_d - \beta_d)$ and eigenvectors u_d.

- $\boldsymbol{E} = \boldsymbol{AB}$

  $E = (\sum_i \alpha_i u_i u_i^T)(\sum_j \beta_j u_j u_j^T) = \sum_{i,j} \alpha_i \beta_j u_i u_i^T u_j u_j^T$.

  Using the orthogonality properties of eigenvectors of symmetric matrices (https://en.wikipedia.org/wiki/Symmetric_matrix), we know that $u_i^T u_i = 1$ and $u_j^T u_i = 0$ for $i \neq j$, so

  $$E = \sum_{i=j} \alpha_i \beta_i u_i u_i^T + \sum_{i \neq j} \alpha_i \beta_j 0$$

  and so the eigenvalues are $(\alpha_d \beta_d)$ and eigenvectors are u_d.

- $\boldsymbol{F} = \boldsymbol{A}^{-1}\boldsymbol{B}$ (assume $\boldsymbol{A}$ is invertible)

  By taking the inverse of A and multiplying, and using the same properties as above, we get that the eigenvalues are (β_d/α_d) and the eigenvectors are u_d.

5 Vector Calculus [10 pts]

Consider the quadratic function $f(x) = x^T Ax$ where x is a column vector and A is an nxn constant matrix.

(a) (1 pts) Express $f(x)$ as a sum of terms (hint: use Σ).

SOLUTION: $f(x) = \sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij} x_i x_j$

(b) (4 pts) Compute the partial derivative of the function with respect to the kth element of x, i.e. $\dfrac{\partial f(x)}{\partial x_k}$, using the expression from (a). Express your answer as a sum of terms.

SOLUTION

$$\frac{\partial f(x)}{\partial x_k} = \frac{\partial}{\partial x_k} \sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij} x_i x_j$$

$$= \frac{\partial}{\partial x_k} [\sum_{i \neq k} \sum_{j \neq k} A_{ij} x_i x_j + \sum_{i \neq k} A_{ik} x_i x_k + \sum_{j \neq k} A_{kj} x_k x_j + A_{kk} x_k^2]$$

$$= \sum_{i \neq k} A_{ik} x_i + \sum_{j \neq k} A_{kj} x_j + 2A_{kk} x_k = \sum_{i} A_{ik} x_i + \sum_{j} A_{kj} x_j .$$

(c) (2 pts) Now write down the gradient vector $\bigtriangledown_x f(x)$ in matrix/vector notation, using the answer from (b). What is its dimension and meaning?

SOLUTION: Note that we can re-write the answer in (b) as

$$\sum_{i} A_{ik} x_i + \sum_{j} A_{kj} x_j = A_k^T x + A_k x$$

where A_k refers to the kth row of A. Note that if A is symmetric, then the answer is $2A_k x$. Then $\bigtriangledown_x f(x) = A^T x + Ax$ is a vector with dimension $nx1$, where each element is the partial derivative computed in (b). If A is symmetric, then the answer is $2Ax$.

(d) (3 pts) Compute the second derivative of $f(x)$, $\bigtriangledown_x^2 f(x)$, in matrix form.

SOLUTION:

$$\bigtriangledown_x^2 f(x) = \bigtriangledown_x [A^T x + Ax] = \bigtriangledown_x [A^T x] + \bigtriangledown_x [Ax]$$

The term Ax is a vector y where each element is a dot product of the kth row of A and x, or $y_k = A_k x$; taking the gradient of each element y_k with respect to x gives the row A_k. Thus the result is a vector of rows of A, which is just A. Similarly treating the first term, we get

$$\bigtriangledown_x^2 f(x) = A^T + A$$

or $2A$ if A is symmetric.