Homework 2

Part A: Conceptual Exercises

Show all steps in your derivations. You may use any identities derived in class.

1. **Linear Programs**
   GreenGrid Energy is a renewable energy provider that generates power using two sources: **Solar Panels** (x_1, in MWh) and **Wind Turbines** (x_2, in MWh).

   The company wants to **maximize its daily revenue**.

   - Solar power yields a revenue of **$40** per MWh.
   - Wind power yields a revenue of **$60** per MWh.

   The generation is subject to the following technical and contractual constraints:

   - **Grid Capacity:** The transmission lines can handle a maximum of **100 MWh** of total power combined.
   - **Maintenance Limits:** Due to crew availability, the operation involves maintenance hours. Solar requires 2 hours per MWh, and Wind requires 5 hours per MWh. There are at most **300 maintenance hours** available daily.
   - **Green Mandate:** To meet a government contract, the company *must* generate **at least 20 MWh** from Wind.
   - **Non-negativity:** $x_1 \geq 0, x_2 \geq 0$.

   Questions:

   (a) Write down the optimization problem as a Linear Program.

   (b) Derive the dual form of your program in part (a). Let the dual variables be y_1 (Grid), y_2 (Maintenance), and y_3 (Mandate).

   (c) Suppose that we have solved the Primal problem and found that the optimal production plan is:

   $$x_1^* = 66.67 \text{ MWh}, \quad x_2^* = 33.33 \text{ MWh}.$$

   At this solution, the total power generated is 100 MWh, and the total maintenance used is 300 hours.

   (i) Based on **Complementary Slackness**, determine which of the dual variables (y_1, y_2, y_3) must be exactly zero and which *could* be non-zero. Explain why.

(ii) Interpret the meaning of the dual variable y_1 in plain English. If the Grid Capacity increased from 100 to 101 MWh, how would the daily revenue change in terms of y_1?

2. **Convex Functions #1**
Let $f_1 : \mathbb{R}^n \to \mathbb{R}$ and $f_2 : \mathbb{R}^n \to \mathbb{R}$ be two convex functions. Prove that their sum $f_1 + f_2$ is also a convex function.

3. **Convex Functions #2**
Consider the function $f(x, y) = e^x + e^y$. Determine the Gradient vector ∇f and the Hessian matrix $\nabla^2 f$. Prove that this function is convex over $\mathbb{R}^2$ by showing the Hessian is positive semidefinite.

4. **KKT Conditions**
Consider the problem:
$$\min_{x\in\mathbb{R}} \frac{1}{2}(x-3)^2 \quad \text{subject to} \quad x \leq 1.$$
   (a) Solve the problem geometrically/intuitively.
   (b) Write down the Lagrangian $L(x, \lambda)$.
   (c) Write the four KKT conditions for this problem.
   (d) Solve the KKT system to find x^* and λ^*.

5. **Support Vector Machines**
In the Hard-Margin SVM, the dual variables α_i correspond to the constraints $y_i(\beta^T x_i + \beta_0) \geq 1$. If a specific data point x_k has a corresponding optimal dual variable $\alpha_k^* = 0$, where is this point located relative to the margin? If $\alpha_k^* > 0$, where is it located?

Part B: Programming Exercises

Note: You may use Python with NumPy, SciPy (`scipy.optimize.linprog`, `scipy.optimize.milp`), or CVXPY for these exercises.

1. **Linear Programs**
   A university cafeteria aims to plan a meal consisting of three staple foods: **Rice**, **Chicken**, and **Vegetables**. The goal is to **minimize the total cost** while meeting specific nutritional requirements.

   Data Table:

Food	Cost ($)	Calories	Protein (g)	Vit C (mg)
Rice (1 serving)	0.50	200	4	0
Chicken (1 serving)	2.50	300	30	0
Vegetables (1 serving)	1.00	50	2	40

   Nutritional Constraints:

   - Minimum Calories: 600
   - Minimum Protein: 20g
   - Minimum Vitamin C: 40mg

   Tasks:

   (a) Formulate and solve this as a Linear Program. Report the optimal number of servings for each food and the total cost.

   (b) We now analyze the Dual Problem:

   - Analytically derive the Dual LP. (The variables will be the "implicit prices" of Calories, Protein, and Vitamin C).
   - Solve this Dual LP using Python.
   - Verify **Strong Duality** by checking if $Objective_{primal} = Objective_{dual}$.
   - **Interpretation:** Based on the dual solution, which nutrient is the most expensive "bottleneck"? (i.e., which has the highest shadow price?). If the requirement for Vitamin C increased by 1 mg, how much would the meal cost increase?

2. **Mixed-Integer Programs**
   A logistics company considers opening warehouses in three possible locations (New York, Dallas, Chicago) to serve four regional markets (Cities 1, 2, 3, and 4).

   The decision involves a trade-off between the **Fixed Construction Cost** of opening a warehouse and the **Variable Shipping Costs** to serve the cities. Each city has a demand of exactly 1 unit.

   Data:

   - **Fixed Costs (f_i):**
     - Warehouse 1 (New York): $400
     - Warehouse 2 (Dallas): $200
     - Warehouse 3 (Chicago): $300

- **Shipping Costs (c_{ij}) per unit:**

	City 1	City 2	City 3	City 4
New York	\$20	\$500	\$500	\$500
Dallas	\$250	\$20	\$400	\$20
Chicago	\$100	\$500	\$20	\$100

Tasks:

(a) Write down the Mixed-Integer Linear Program.

- Define binary variables $y_i \in \{0, 1\}$ for opening warehouse i.
- Define continuous (or binary) variables $x_{ij} \in [0, 1]$ representing the fraction of City j's demand served by Warehouse i.
- Constraints: (1) Each city must be fully served ($\sum_i x_{ij} = 1$), (2) You cannot ship from a warehouse unless it is open ($x_{ij} \leq y_i$).

(b) Solve the MIP using Python (`scipy.optimize.milp` or `cvxpy`).

(c) Answer the following questions:

- Which warehouses should be opened?
- Which warehouse serves which city?
- What is the optimal total cost?

3. **Best Subset Selection on Real Data (Diabetes Dataset)**
In this exercise, you will compare the performance of a heuristic feature selection method (Lasso) against an exact method (MIP Best Subset Selection) using the real-world **Diabetes dataset** provided by Scikit-Learn.

The dataset contains $n = 10$ baseline variables (age, sex, bmi, bp, and six blood serum measurements) and a quantitative measure of disease progression (y).

Code Setup: Use the following snippet to load and prepare the data.

```python
!pip install "cvxpy[SCIP]"
import numpy as np
from sklearn.datasets import load_diabetes
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler

# 1. Load Data
data = load_diabetes()
X, y = data.data, data.target

# 2. Scale features (Important for numerical stability)
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)

# 3. Split into Train and Test sets
X_train, X_test, y_train, y_test = train_test_split(
    X_scaled, y, test_size=0.3, random_state=42
)
```

Tasks:

(a) **Lasso Regression:**

- Train a Lasso regression model on the training set.
- Tune the regularization parameter α so that the model selects **exactly 4 non-zero features**.
- Record the Mean Squared Error (MSE) on the **Test set**.
- List the names of the 4 features selected by Lasso.

(b) **Best Subset Selection (MIP Exact Selection):**

- Formulate the regression problem as a Mixed-Integer Quadratic Program (MIQP).
- **Requirement:** The model must include an **intercept term** β_0 that is **unpenalized**.
- The objective is to minimize $||y - (\beta_0 \mathbf{1} + X^\top \beta)||_2^2$ subject to $\sum_{j=1}^{10} z_j \leq 4$ and Big-M constraints linking β_j to binary variables z_j.
- Solve the MIP.
- Calculate the MSE on the **Test set**.
- List the names of the 4 features selected by the MIP.

(c) **Comparison:**

- Did Lasso and MIP select the same subset of features?
- Which method achieved a lower Test MSE?