Advanced Algorithms

Homework 2

Collaborators

Instructions

Problems 1 and 2 ask you to utilize the powerful tools of linear and integer programming to model some interesting problems. Problems 3 and 4 touch on how linear relaxations are used in combinatorial optimization.

Some of these problems are hard! You are encouraged to work together on them or come to me if you are stuck. However, you should write up your solutions yourself. Please list the people you worked with at the top of your submission. Looking for answers on the internet is not allowed, nor is working with an AI-powered system for any part of this assignment. You must understand everything you submit and I reserve the right to ask you to orally explain your answer to me. You may write your solutions by hand or in latex. Either way, submit them on gradescope by 10:00pm on Tuesday, March 31. Good luck!

Problem 1 - Linear Regression via Linear Programming (20 points)

You may have encountered the problem of Linear Regression, or finding a line which best "fits" a set of data. In ordinary least squares regression (OLS), we fit a line $y \approx mx + b$ by minimizing the sum of *squared* differences between the line and the true value. Squaring makes large errors dominate, making the line more sensitive to outliers. An alternative is *least-absolute-deviations* regression (LAD), which minimizes the sum of the *absolute values* of the differences and is therefore more robust to outliers. Conveniently, LAD can be formulated and solved as a linear program, as you will show below.

We are given n data points $\{(x_i, y_i)\}_{i=1}^n \subseteq \mathbb{R}^2$. (This problem can of course be extended to higher dimensional data, but we will work in two dimensions for simplicity.) The goal is to fit a line $y = mx + b$ that minimizes the total absolute error

$$\sum_{i=1}^{n} \big| \, y_i - (mx_i + b) \, \big|.$$

(a) What are the decision variables in the LAD problem? Are there any constraints on these variables?

(b) What is the objective function in the LAD problem? Is it linear as written?

(c) Write down a linear program which you can use to solve the LAD problem. You may need to introduce additional decision variables and/or constraints, and modify the objective function. However, it should be clear (and you should explain) how an optimal solution to your linear program can be used to produce an optimal solution to LAD.

Problem 2 - Sudoku (20 points)

A Sudoku puzzle consists of a 9×9 grid subdivided into nine 3×3 boxes. The goal is to fill the grid with digits 1 through 9 so that each row, each column, and each 3×3 box contains every digit exactly once. Some digits are given in advance and must remain fixed.

2	4	7		8	1			6
	8			3	5		7	
5		9	4				2	
		6	3			9		4
	7	5		6		3		
	1		9	2				
	6	3			8			5
4			6			7		1
1			7		9		4	

Figure 1: A Sukodu puzzle

(a) Consider the first row of the above Sudoku grid. Formulate the condition that each digit 1 through 9 appears exactly once in row 1 as an integer linear program while respecting the given values.

(b) Formulate the problem of finding a solution to the entire Sudoku as an integer linear program. Clearly state the decision variables and constraints. You do not need to write every constraint explicitly, but you must convince me that you know how to solve a Sudoku puzzle using Integer Programming.

(c) Formulate the problem of certifying that a given Sudoku solution is unique as an integer program.

(d) (Bonus) Play the Sudoku shown in the figure :)

Problem 3 - Trees and LPs (20 points)

In this problem, we will compare two different formulations for the minimum spanning tree problem (MST). Let $G = (V, E)$ be a graph, and let $c_e \geq 0$ be the cost of edge e. Recall that the MST problem asks us to find a spanning tree of G with cheapest cost.

We will now write a linear program for this problem. Create a decision variable x_e for each edge $e \in E$ which indicates whether we choose e to be included in the spanning tree. Consider the following formulation, where $\delta(S)$ denotes the set of edges with exactly one endpoint in $S \subseteq V$.

$$\begin{aligned} \min \quad & \sum_{e \in E} c_e \, x_e \\ \text{s.t.} \quad & \sum_{e \in E} x_e \;=\; |V| - 1 \\ & \sum_{e \in \delta(S)} x_e \;\geq\; 1 \qquad \forall \, \emptyset \neq S \subset V, \\ & 0 \;\leq\; x_e \;\leq\; 1 \qquad \forall \, e \in E. \end{aligned}$$

(a) Explain why integer solutions to the above LP correspond to spanning trees of G with the same cost.

(b) Can a fractional solution "cheat"? The answer is yes: give an example of a graph G with costs on edges, and a fractional solution which has a cost lower than any minimum cost spanning tree of G. What does your example show about the integrality gap of this LP?

[Warning: this is somewhat tricky]

(c) Give a polynomial algorithm to determine whether a solution x is feasible to the above LP.

Now consider another formulation. Again, there is a variable x_e for each edge indicating whether we choose edge e into the MST. We denote by $E[S]$ the set of edges with both endpoints in $S \subseteq V$.

$$\begin{aligned} \min \quad & \sum_{e \in E} c_e \, x_e \\ \text{s.t.} \quad & \sum_{e \in E} x_e \;=\; |V| - 1, \\ & \sum_{e \in E[S]} x_e \;\leq\; |S| - 1 \qquad \forall \, \emptyset \neq S \subset V, \\ & 0 \;\leq\; x_e \;\leq\; 1 \qquad \forall \, e \in E. \end{aligned}$$

(e) Show that integer solutions to the above LP *also* correspond to spanning trees of G.

(f) Show that your example from part (c) is no longer feasible for this formulation. (In fact, this formulation turns out to be much stronger: it always exhibits an integer optimal solution).

Problem 4 - The Representative Committee Problem (Bonus 10 points)

We are given a set $P = \{p_1, \ldots, p_n\}$ of Computer Science professors. Each professor serves on one or more committees, and there are m committees total. To represent this, we will denote the ith committee by C_i, where C_i is a subset of P and indicates which professors serve on that committee. The set of committees is $S = \{C_1, \ldots, C_m\}$.

We would like to make a decision as a department. This means we need a representative set of professors - i.e. at least one professor should be present from each committee. But we also don't want to take up too many people's time (we're all busy grading papers). Thus, we seek to find a representative set of smallest size. First, take 2 minutes to think about how you might solve this problem.

Now, consider the following formulation. Introduce a binary variable x_p for each professor $p \in P$, where $x_p = 1$ means p is chosen for the representative set and $x_p = 0$ otherwise. The problem can be written as the following integer program:

$$\begin{aligned} \min \quad & \sum_{p \in P} x_p \\ \text{s.t.} \quad & \sum_{p \in C_i} x_p \ \geq \ 1 \quad \forall\, i \in [m], \\ & x_p \in \{0, 1\} \qquad \forall\, p \in P. \end{aligned}$$

(a) Show that there is a cost-preserving correspondence between feasible solutions to the above integer program and feasible solutions to the representative committee problem.

(b) Relax the integrality constraints on the variables x_p to obtain a linear program, then take the dual. Clearly write out the decision variables, objective function and constraints. How does the optimal value of the dual compare to the optimal value of the original linear program?

(c) Define a **disjoint committee set** to be a collection of committees with no professors in common. Using this, can you give an English interpretation of integer solutions to the dual program?

(d) Use duality to show that the size of any representative committee set is at least as big as any disjoint committee set.