# Problem Set 4

## CS 331H

## Due Tuesday, April 5

1. Consider the following variant of interval scheduling. You have $n$ intervals, each with a given integer start and end time $[s_i, t_i]$ and cost $c_i$, and would like to choose a subset $S$ that minimizes the cost

$$\text{cost}(S) = \sum_{i \in S} c_i$$

subject to the constraint that every integer time in $[0, T]$ is covered by at least at least $k$ different intervals in $S$.

Show how to reduce this problem to a minimum cost circulation problem. You may assume that $T = O(n)$.

2. [Problem 372 of Brian Dean's book: the maximum-density subgraph problem.] Given an undirected, unweighted graph, show how to compute the subgraph of maximum edge density using minimum cuts. Hint: for the $s$-$t$ graph construction provided in the book, and any cut $S$, express $|\text{cut}(S)|$ in terms of the number of edges in $S$ and the degrees of vertices in $S$. How does this relate to the edge density in $S$?

3. In regression, you are given a set of points $(x_i, y_i)$ and would like to find a line $y = mx + b$ such that the error is small by some measure. In $\ell_1$ regression, one would like to minimize the $\ell_1$ norm of the residuals:

$$\sum_{i=1}^{n} |(\alpha x_i + \beta) - y_i|.$$

The goal is to find $\alpha$ and $\beta$ minimizing this quantity.

   (a) Show how to express this problem as a linear program. Hint: the constraint $|a| \leq b$ is equivalent to the two constraints $a \leq b$ and $-a \leq b$.

   (b) Write your program in the primal form

$$\text{Maximize } c^T x$$
$$\text{Subject to } Ax \leq b$$

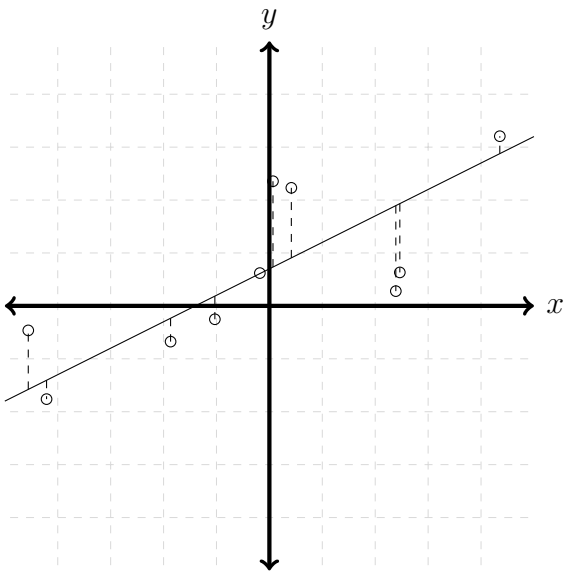


Figure 1: Illustration of regression. Not actually done via regression, so probably not optimal.

(c) Give the asymmetric dual form of your linear program.

$$\begin{aligned} \text{Minimize } & b^T y \\ \text{Subject to } & A^T y = c \\ & y \geq 0 \end{aligned}$$

(d) Prove that, in the optimal regression, half the points lie above the line and half the points lie below the line.

(e) Give a direct interpretation of the dual LP, explaining what each expression/variable signifies and why the result is correct. Hint: the dual variables correspond to whether $y_i$ is above or below the optimal regression line.

You may find it helpful to assume that no points lie on the optimal regression line. You may assume this, though I encourage you to figure out what happens in general.