Problem Set 4

CS 331H

Due Thursday, April 6

1. Given a weighted, undirected graph, a source s, and a sink t, find the shortest path from s to t and back to s that uses each edge at most once. Aim for $O(m + n \log n)$ time. **Hint:** the idea is similar to the Ford-Fulkerson algorithm.

2. [Problem 372 of Brian Dean's book.] Given an undirected, unweighted graph, we would like to compute the subgraph of maximum edge density. The *edge density* of a subgraph is the number of edges divided by the number of vertices.

   Consider the following construction. For a given "guess" λ, construct a dummy source s and sink t. Draw an edge from s to each graph node u of capacity m; one from each graph node u to t of capacity $m + 2\lambda - d_u$, where d_u is the degree of u in the original graph; and give each edge (u, v) in the original graph capacity 1.

   (a) For a nonempty set S of vertices in the original graph, express the cost of cutting $S \cup \{s\}$ from the rest of the graph, in terms of the number of edges fully contained in S and the degrees in S.

   (b) Show that this value is less than mn if, and only if, the edge density of S is more than λ.

   (c) Show how a max-flow algorithm and binary search can narrow down on the maximum edge density of any subgraph. Show that after $O(\log n)$ steps of binary search, you can compute the maximum edge density exactly.

   (d) Show how to compute the set S^* of maximum edge density, not just its value.

3. 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

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

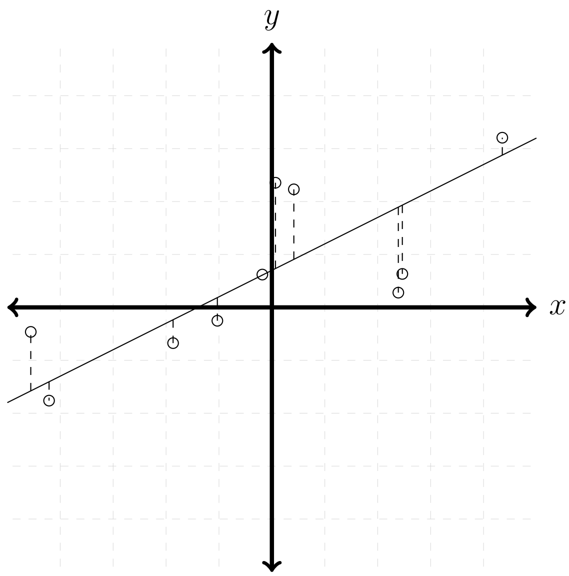

Figure 1: Illustration of regression. (The line wasn't actually found using regression, so it probably is not optimal.)

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 $T = O(n)$.

4. 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 ℓ_1 regression, one would like to minimize the ℓ_1 norm of the residuals:

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

The goal is to find α and β 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

$$\begin{aligned} &\text{Maximize } c^T x \\ &\text{Subject to } Ax \leq b \end{aligned}$$

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

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

(d) Prove that, in the optimal regression, at least half the points lie on the line of above it, and at least half lie on the line or below it.

(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 an optimal regression line. You may assume this, though I encourage you to figure out what happens in general.