CSE 543T Algorithms for Nonlinear Optimization: Homework 2

Due: April 10, 11:59pm

1**. Problem 3.1.1.a) and 3.1.1.b)** . For both problems, you need to study and use the Second Order Sufficiency Condition in Proposition 3.2.1 to verify that your solution is indeed a local minimum. (15%)

2. **Industrial design.** A cylindrical can is to hold 4 cubic inches of orange juice. The cost per square inch of constructing the metal top and bottom is twice the cost per square inch of constructing the cardboard side. What are the dimensions of the least expensive can? (15%)

3. **Duality.** Read Section 3.4 and study Example 3.4.2. Prove that the following two linear programs are dual to each other

Min c' x , subject to A' x $\geqslant$ b
Max b' μ , subject to A μ = c, $\mu \geqslant 0$ (15%)

4. **Problem 4.2.1 (a) (b) and (d)** (15%)
Hint: The augmented Lagrangian function with quadratic penalty is described in pages 398-404.

5. **Mathematical modeling for data mining.** (40%)

Linear regression is one of the fundamental models for data mining. The model describes a linear relationship between a number of numerical attributes **x**= (x_1, x_2, ..., x_n) and a predicted variable y in the form of

$$y = \boldsymbol{a}'\mathbf{x}+b,$$

where **a** $\in R^n$ and b $\in R$ are parameters to be determined by *training*. The training process takes a set of *K* training examples

$$(\mathbf{X}, Y) = \{(\mathbf{x^1}, y^1), (\mathbf{x^2}, y^2), ..., (\mathbf{x^K}, y^K)\},$$

where each $\mathbf{x^i} \in R^n$ is a vector of attributes. The parameters **a** and b are determined by minimizing the mean squared error (MSE):

$$\text{MSE} = \sum_{i=1}^{K} [y^i - (\mathbf{a}'\mathbf{x^i}+b)]^2$$

Build a linear regression for the following **program effort data.** Each training sample consists of an index of social setting, an index of family planning effort, and the percentage change in the crude birth rate (CBR) between 1965 and 1975, for 20 countries in Latin America. Here, we want to predict *change* (y) using *setting* (x_1) and *effort* (x_2). Therefore, we have that n = 2 and K =20.

```
                setting(x₁)  effort(x₂) change(y)

   Bolivia             46        0         1
   Brazil              74        0        10
   Chile               89       16        29
   Colombia            77       16        25
   CostaRica           84       21        29
   Cuba                89       15        40
   DominicanRep        68       14        21
   Ecuador             70        6         0
   ElSalvador          60       13        13
   Guatemala           55        9         4
   Haiti               35        3         0
   Honduras            51        7         7
   Jamaica             87       23        21
   Mexico              83        4         9
   Nicaragua           68        0         7
   Panama              84       19        22
   Paraguay            74        3         6
   Peru                73        0         2
   TrinidadTobago      84       15        29
   Venezuela           91        7        11
```

Write an AMPL model for the optimization problem, and submit it to NEOS to obtain the optimal parameters **a** and b in the linear regression model. You need to choose a suitable solver in NEOS. You cannot use any other existing software for linear regression. Submit the following:

1) The AMPL model file (and data file, if any)

2) A print-out of the solution from your NEOS solver.

3) A table listing the model error $y^i - (\mathbf{a}'\mathbf{x}^{\mathbf{i}}+b)$ for all the 20 countries.
4) Discuss the insights you gained from this analysis, such as: How does each attribute influence the change? Which attribute seems to have stronger correlation with the change? Does the linear regression model seem accurate to you?