Homework 7: Convex programs

Due date: 11:59pm on Wednesday April 5, 2023
See the course website for instructions and submission details.

1. 4 points

NFL Regression.

We examined the results of the first 10 weeks of the 2000 NFL season, and aim to use regression to determine a rating for each team. The file `nfl.inc` contains the margins of victory by the home team in each game. In each entry in the parameter margins, the visiting team is listed first, followed by the home team. For example, the entry

```
2.chiefs.titans 3
```

indicates that the game was played in week 2, the visiting team was the Chiefs, the home team was the Titans, and the home team (Titans) won by 3 points. Of course, a negative "margin of victory" indicates that the visiting team won the game.

(The starter code imports the file `nfl.inc` into a Julia array that may be easier to work with.)

(a) We aim to choose the team ratings to predict the results of all these games as well as possible, according to some loss function. We also aim to find the "average" home field advantage. If team i (the visiting team) has rating R_i, and team j (the home team) has rating R_j, and the home field advantage is H, then our prediction of the margin of victory for the home team is is $R_j - R_i + H$ Using a sum-of-squares objective, applied to the difference between our predicted outcomes and the actual outcomes from the data file, determine the ratings for each team and the home field advantage. Apply the following constraint to the ratings: They must sum to zero. Print the ratings for the teams (tabulated by team name).

(b) Repeat part (a) but with an ℓ_1 objective function in place of the sum-of-squares to obtain a different set of ratings. Print the ratings. Comment on the differences in the ratings obtained by these two methods.

(c) Now modify the question in part (a) as follows: Use a separate "home field advantage" variable H_i for each team. Specifically, if team ii (the visiting team) has rating R_i, and team j (the home team) has rating R_j, and the home field advantage for teach j is H_j, then our prediction of the margin of victory for the home team is is $R_j + H_j - R_i$. Use the sum-of-squares objective to find the best-fit ratings and home field advantages. Print the ratings and home field advantage for all teams.

(d) According to the ratings obtained in parts (a) and (b), who would you expect to win if the dolphins were to visit the seahawks, and what would be the expected margin?

2. 4 points The Huber loss. In statistics, we frequently encounter data sets containing *outliers*, which are bad data points arising from experimental error or abnormally high noise. Consider for example the following data set consisting of 15 pairs (x, y).

x	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
y	6.31	3.78	24	1.71	2.99	4.53	2.11	3.88	4.67	4.25	2.06	23	1.58	2.17	0.02

The y values corresponding to $x = 3$ and $x = 12$ are outliers because they are far outside the expected range of values for the experiment.

(a) Compute the best linear fit to the data using an ℓ_2 cost (least squares). In other words, we are looking for the a and b that minimize the expression:

$$\ell_2 \text{ cost:} \qquad \sum_{i=1}^{15} (y_i - ax_i - b)^2$$

Make a plot showing the data points and the fitted line.

(b) Alternative measures of the fit between the fitted line and the data may be less sensitive to the presence of outliers. Find the best linear fit again (including the outliers), but this time use the ℓ_1 cost function:

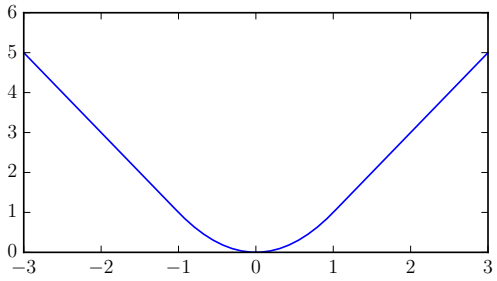
$$\ell_1 \text{ cost:} \qquad \sum_{i=1}^{15} | \, y_i - ax_i - b \, |$$

Make a plot containing the data and the best ℓ_1 linear fit. Does the ℓ_1 cost handle outliers better or worse than least squares? Explain why.

(c) A "hybrid" approach is to use an ℓ_2 penalty for points that are close to the line but an ℓ_1 penalty for points that are far away. Specifically, we'll use something called the *Huber loss*, defined as:

$$\phi(x) = \begin{cases} x^2 & \text{if } -M \le x \le M \\ 2M|x| - M^2 & \text{otherwise} \end{cases}$$

Here, M is a parameter that determines where the quadratic function transitions to a linear function. The plot on the right shows what the Huber loss function looks like for $M = 1$.



The formula above is simple, but not in a form that is useful for us. As it turns out, we can evaluate the Huber loss function at any point x by solving the following convex QP instead:

$$\phi(x) = \left\{ \begin{array}{ll} \underset{v,w}{\text{minimize}} & w^2 + 2Mv \\ \text{subject to:} & |x| \le w + v \\ & v \ge 0, \; w \le M \end{array} \right\}$$

Find the best linear fit to our data using a Huber loss with $M = 1$. Find the coefficients a and b by minimizing the cost function $\sum_{i=1}^{15} \phi(\, y_i - ax_i - b)$.

(d) Make one final plot (using starter code supplied) showing the data points and all three fitted lines. Which of the techniques gave the best fit to the non-outlier data?

3. 4 points Heat pipe design. A heated fluid at temperature T (degrees above ambient temperature) flows in a pipe with fixed length and circular cross section with radius r. A layer of insulation, with thickness w, surrounds the pipe to reduce heat loss through the pipe walls (w is much smaller than r). The design variables in this problem are T, r, and w.

The energy cost due to heat loss is roughly equal to $\alpha_1 Tr/w$. The cost of the pipe, which has a fixed wall thickness, is approximately proportional to the total material, i.e., it is given by $\alpha_2 r$. The cost of the insulation is also approximately proportional to the total insulation material, i.e., roughly $\alpha_3 rw$. The total cost is the sum of these three costs.

The heat flow down the pipe is entirely due to the flow of the fluid, which has a fixed velocity, i.e., it is given by $\alpha_4 Tr^2$. The constants α_i are all positive, as are the variables T, r, and w.

Now the problem: maximize the total heat flow down the pipe, subject to an upper limit $C_{\max}$ on total cost, and the constraints

$$T_{\min} \le T \le T_{\max}, \qquad r_{\min} \le r \le r_{\max} \qquad w_{\min} \le w \le w_{\max}, \qquad w \le 0.1r$$

a) Express this problem as a geometric program, and convert it into a convex optimization problem.

b) Consider a simple instance of this problem, where $C_{\max} = 500$ and $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 1$. Also assume for simplicity that each variable has a lower bound of zero and no upper bound. Solve this problem using JuMP. Use the `Ipopt` solver and the command `@NLconstraint(...)` to specify nonlinear constraints such as log-sum-exp functions. Have your code print the optimal values of T, r, and w, as well as the optimal objective value.