STAT 341 - Assignment 2 - Fall 2022

Instructors: Nathaniel Stevens and Jack Davis

Due: Friday October 21 at 11:59pm EST

Your assignment must be submitted by the due date listed at the top of this document, and it must be submitted electronically in .pdf format via Crowdmark. This means that your responses for different question parts should begin on separate pages of your .pdf file. Note that your .pdf solution file must have been generated by R Markdown. Additionally:

- For mathematical questions: your solutions must be produced by LaTeX (from within R Markdown). Neither screenshots nor scanned/photographed handwritten solutions will be accepted – these will receive zero points.
- For computational questions: R code should always be included in your solution (via code chunks in R Markdown). If code is required and you provide none, you will receive zero points.
  - **Exception** any functions defined in the lecture notes can be loaded using `echo=FALSE` but any other code chunks should have `echo=TRUE`. e.g., the code chunk loading `gradientDescent` can use `echo=FALSE` but chunks that call `gradientDescent` should have `echo=TRUE`.
- For interpretation questions: plain text (within R Markdown) is required. Text responses embedded as comments within code chunks will not be accepted.

Organization and comprehensibility is part of a full solution. Consequently, points will be deducted for solutions that are not organized and incomprehensible. Furthermore, if you submit your assignment to Crowdmark, but you do so incorrectly in any way (e.g., you upload your Question 2 solution in the Question 1 box), you will receive a 5% deduction (i.e., 5% of the assignment's point total will be deducted from your point total).

Question One - 6 Marks - Gathering and Cleaning NBA data.

(a) [1 Mark] Go to the list of packages by name at the Comprehensive R Archive Network, or CRAN, at https://cran.r-project.org/web/packages/index.html and name a package for men's basketball and a package for women's basketball.

(b) [1 Mark] Using the information at https://hoopr.sportsdataverse.org/reference/index.html, what is the function name for getting the player box scores data for the most recent season of NBA data (if we didn't have this already as a .csv file on LEARN)?

(c) [2 Marks] Read in `NBA_Player_Boxscore_2021-22.csv`. Each row (observation) in the dataset provides information about a specific player within a specific game, i.e., a unique 'player-game' combination. For instance, in a given game we have information about each player's points (`pts`), rebounds (`reb`), blocks (`blk`), assists (`ast`), steals (`stl`), etc. In this question, determine the number of rows in the dataset devoted to each team by constructing a frequency `table()` of the number of times each `team_abbreviation` appears. Comment on the count from the typical team and any teams with unusually low counts that you might find.

(d) [2 Marks] Remove the unusual teams with a version of the following code. Replace `XXX` with the teams in question, and get the number of rows of the resulting dataset. How many rows does the new dataset `npb2` have?

(Note: It's always good practice to make a new copy of the dataset you're modifying when you do something like remove rows with `subset`, which is why it's `npb2 =` and not `npb =`.)

```r
npb = read.csv("filename here, watch your setwd().csv")
npb2 = subset(npb, team_abbreviation != "XXX" & team_abbreviation != "XXX")
```

Question Two - 18 Marks - Rebounds by Position

(a) [3 Marks] Construct a 1 × 3 figure containing a boxplot, histogram and quantile plot for the number of rebounds `reb` for all players. For the histogram pick a reasonable number of bins. For each plot, describe any features of the population.

(b) [1 Mark] Is a boxplot a good representation of this population? Why?

(c) [2 marks] Use the `by()` command to find the mean number of rebounds (`reb`) per player per game for different player positions using the `athlete_position_name` variable.

(d) [3 Marks] Construct a side-by-side boxplot of the number of rebounds by position. Describe any notable features.

(e) [1 Mark] Are the boxplots good representations of these seven populations? Why?

(f) [3 Marks] Since the number of rebounds in a game is always an integer value, points on a plot have an overdrawing issue where multiple data points are drawn exactly on top of each other, masking the actual count in plots like this question's boxplot. State how you would fix this and, re-plot the boxplot.

(g) [5 Marks] Find an appropriate power transformation for making the distribution of rebounds (for all positions together in one distribution) as symmetric as possible. Look in the lecture notes and the tutorial notes for helpful functions like `powerfun()`, an appropriate loss function, and an optimizer like `optim()`.

Question Three - 15 Marks - Raptors vs. Warriors. Influence on regression lines.

Use the following code to load the player box scores and keep only the box scores of the Toronto Raptors and some other guys out west.

```
npb_q3 = subset(npb, team_abbreviation %in% c("TOR", "GS"))
```

(a) [2 Marks] Create a scatterplot of rebounds and points scored (`pts`) in a game, shaping the points of the scatterplot by team. Only use the teams with `team_abbreviation = "TOR"` or `= "GS"`. Add two regression lines (one for each team) by treating rebounds as the covariate and points as the response. You may use `lm()` to fit these ordinary simple linear regressions.

(b) [5 Marks] For Toronto, calculate the influence of each player-game on each of the regression coefficients. Here, influence means the raw, scalar difference (*not* absolute difference) in the intercept or slope, respectively, after removing an observation.

- In a 1 × 2 figure, plot the influence on the intercept and on the slope by player-game.
- Comment on the plots.
- If there are any influential points determine if is there anything interesting about them.
- Are these influential points attributable to one or two players?

(c) [3 Marks] Use the `by()` function to get the average influence on the slope per game for each player. Comment on any unusual cases. (That is, take the influence of each player-game from part (b), and get the mean of the influence values (i.e., the raw differences) for all the games by a particular player. Repeat this process for every Raptor by using code similar to `by(dat$reb, dat$playername, mean)`)

(d) [3 Marks] Using the `gradientDescent` function from class, fit a regression line to points (`pts`) as a function of rebounds (`reb`) for Toronto, using the Huber loss function instead of the sum of squared loss. Note that you may use the `huber.fn()`, `createRobustHuberRho()`, `huber.fn.prime()`, and `createRobustHuberGradient()` functions from class. Use the default tuning parameter of $k = 1.345$. Print the slope and intercept of the regression line that you get using the Huber loss function. [*Hint*: consider using `rlm()` from the `MASS` package to check your answer.]

(e) [2 Marks] Plot a scatterplot of the Toronto Raptor's points (y) and rebounds (x), and plot both the least-squares and the Huber loss regression lines. Comment on the appropriateness of each one (i.e., Are they both good? Is one better?)

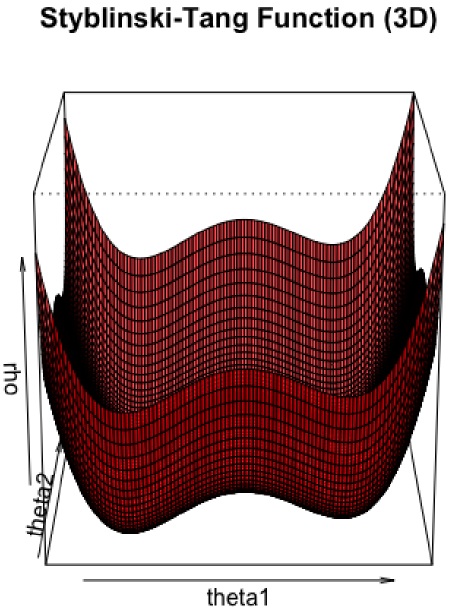
Question Four - 21 Marks - Optimizer Test Function.

Non-convex problems are ones where there might be more than one local minima, and as such there they are challenging to optimization methods to solve. For example, protein folding is a non-convex problem of many dimensions used for a wide range of disease research, and requires many computers working in parallel to find something acceptably close to a global minimum.

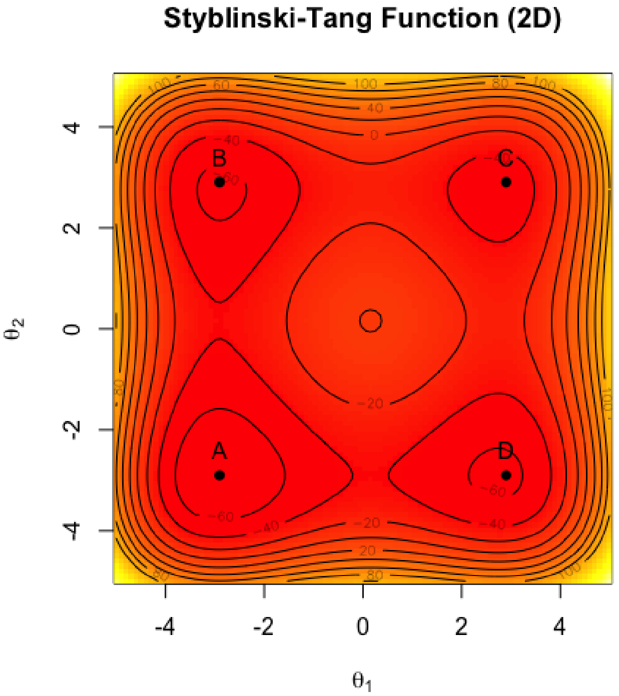
To make sure the optimization method being used for such difficult problems is up to the task, the method is first evaluated against a battery of test functions. Consider one of the harder, but possible, ones, the Styblinski-Tang function, for dimensionality $d = 2$. For $\boldsymbol{\theta} \in [-5, 5]^2$ the function is defined as follows

$$\rho(\boldsymbol{\theta}) = \frac{1}{2}(\theta_1^4 - 16\theta_1^2 + 5\theta_1 + \theta_2^4 - 16\theta_2^2 + 5\theta_2)$$

The figures below depict the function (as a 3-dimensional surface and with 2-dimensional contours) for $\theta_1 \in [-5, 5]$ and $\theta_2 \in [-5, 5]$.



Styblinski-Tang Function (3D)



Styblinski-Tang Function (2D)

As can be seen in the filled contour plot, this function has 4 local minima near $\theta_1 = \pm 3$, $\theta_2 = \pm 3$, only one of which is a global minimum.

(a) [4 Marks] Write `rho` and `gradient` functions for the Styblinski-Tang function which take a single vector-valued input `theta`. Note that you may use without proof or derivation the fact that

$$\frac{\partial \rho}{\partial \theta_1} = 4\theta_1^3 - 32\theta_1 + 5,$$

$$\frac{\partial \rho}{\partial \theta_2} = 4\theta_2^3 - 32\theta_2 + 5$$

(b) [5 Marks] In this question you will explore the surface of the Styblinski-Tang function using gradient descent. In particular you will consider 5 different starting values and explore the impact of changing one's starting location. Using the gradientDescent function (from class) together with the gridLineSearch and testConvergence functions (from class) as well as the rho and gradient functions from part (a), find the solution to

$$\underset{\boldsymbol{\theta}\in\mathbb{R}^2}{\operatorname{argmin}}\,\rho(\boldsymbol{\theta})$$

for each of the following five starting points $\widehat{\boldsymbol{\theta}}_0$.

$\widehat{\boldsymbol{\theta}}_0 = (0, 0),$

$\widehat{\boldsymbol{\theta}}_0 = (3, 3),$

$\widehat{\boldsymbol{\theta}}_0 = (-3, 3),$

$\widehat{\boldsymbol{\theta}}_0 = (3, -3),$

$\widehat{\boldsymbol{\theta}}_0 = (-3, -3).$

In each case state, to the nearest 0.001, the minimum that you converged to, and be sure to include the output from the gradientDescent function.

(c) [5 marks] Recreate the contour plot shown above. You may find the functions outer, image, and contour useful for this task.

(d) [2 points] Based on your investigations in parts (b) and (c) explain the importance of the starting value when performing non-convex optimization (when locating a global optimum is desired).

(e) [5 points] Repeat part (b) using the same five starting points and the generalized optimizer optim to check your work. Give the output and explain the differences (if any) and similarities.