Homework 8 Solutions

In this homework you'll explore various optimization algorithms for forming statistical estimates in linear regression.

1. Run the following code block to create synthetic regression data, with 100 observations and 10 predictor variables:

```
n <- 100
p <- 10
s <- 3
set.seed(0)
x <- matrix(rnorm(n * p), n, p)
b <- c(-0.7, 0.7, 1, rep(0, p - s))
y <- x %*% b + rt(n, df = 2)
```

Notice that only 3 of the 10 predictor variables in total are actually relevant in predicting the response. (That is, only the first three coefficients in `b` are nonzero.) Examine the correlation coefficients between predictor variables `x` and the response `y`; would you be able to pick out each of the 3 relevant variables based on correlations alone?

```
cors <- apply(x, 2, cor, y)
cors
```

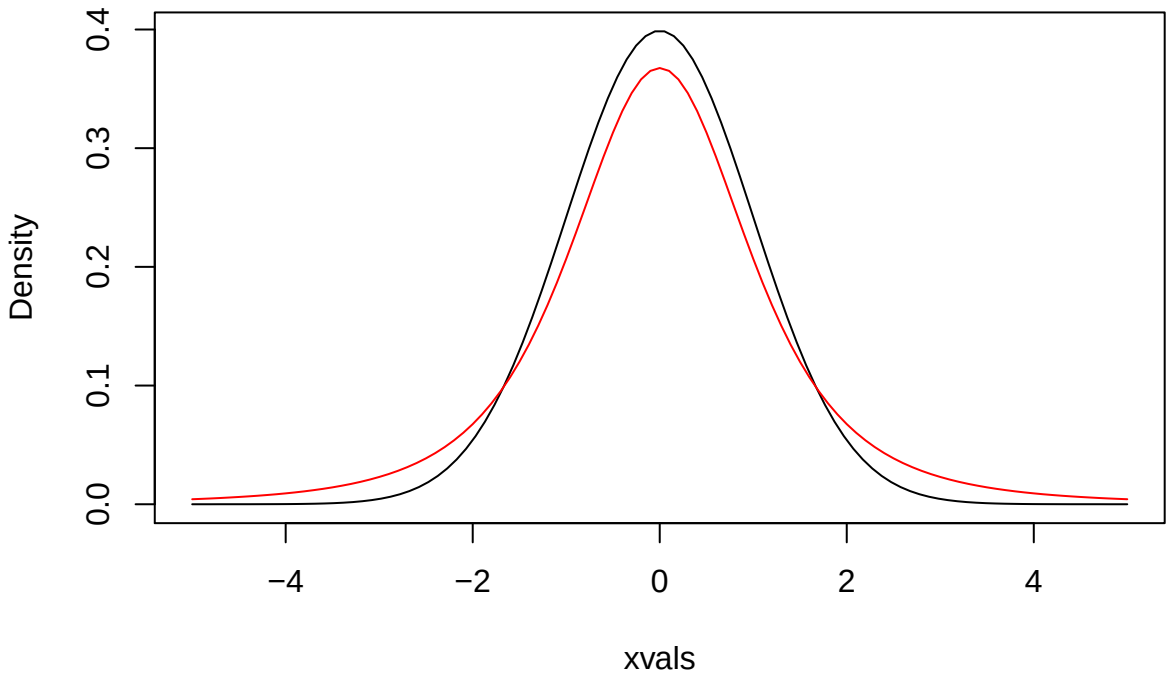
```
##  [1] -0.2526434175  0.1239284685  0.1673840288 -0.2522804417 -0.0371161818
##  [6]  0.1561141420 -0.1175268150 -0.0899681839 -0.0002104895  0.0506851086
```

```
order(abs(cors), decreasing = TRUE)
```

```
##  [1]  1  4  3  6  2  7  8 10  5  9
```

2. Note that the noise in the above simulation (the difference between `y` and `x %*% b`) was created from the `rt()` function, which draws t-distributed random variables. The t-distribution has thicker tails than the normal distribution, so we are more likely to see large noise terms than we would if we used a normal distribution. Verify this by plotting the normal density and the t-density on the same plot, with the latter having 3 degrees of freedom. Choose the plot ranges appropriately, and draw the densities in different colors, so that the plot is easy to read.

```
xvals <- seq(-5, 5, length.out = 100)
plot(xvals, dnorm(xvals), type = "l", ylab = "Density")
curve(dt(x, df = 3), add = TRUE, col = "red")
```



3. Because we know that the noise in our regression has thicker tails than the normal distribution, we are more likely to see outliers. Hence we're going to use the Huber loss function, which is more robust to outliers:

```
psi = function(r, c = 1) {
  return(ifelse(r^2 > c^2, 2*c*abs(r) - c^2, r^2))
}
```

Write a function called `huber.loss()` that takes in as an argument a coefficient vector `beta`, and returns the sum of `psi()` applied to the residuals (from regressing `y` on `x`). `x` and `y` should not be provided as arguments, but referred to directly in the function. You may stick with the default cutoff of `c=1`. This Huber loss is going to take the place of the usual (nonrobust) linear regression loss, i.e., the sum of squares of the residuals.

```
huber.loss <- function(beta) {
  sum(psi(y - x %*% beta))
}
```

4. Using the `grad.descent()` function from lecture, run gradient descent starting from `beta=rep(0,p)`, to get an estimate of the coefficients `beta` that minimize the Huber loss, when regressing `y` on `x`. Use the settings `max.iter=200`, `step.size=0.001`, and `stopping.deriv=0.1`. Store the output of `grad.descent()` in `gd`. How many iterations did it take to converge, and what are the final coefficient estimates?

```
library(numDeriv)
grad.descent <- function(f, x0, max.iter = 200, step.size = 0.05, stopping.deriv = 0.01, ...) {

  n    <- length(x0)
  xmat <- matrix(0, nrow = n, ncol = max.iter)
  xmat[,1] <- x0

  for (k in 2:max.iter) {
    # Calculate the gradient
    grad.cur <- grad(f, xmat[ ,k-1], ...)
```

```
    # Should we stop?
    if (all(abs(grad.cur) < stopping.deriv)) {
      k <- k-1; break
    }

    # Move in the opposite direction of the grad
    xmat[ ,k] <- xmat[ ,k-1] - step.size * grad.cur
  }

  xmat <- xmat[ ,1:k] # Trim
  return(list(x = xmat[,k], xmat = xmat, k = k))
}

gd <- grad.descent(huber.loss, x0 = rep(0,p), max.iter=200, step.size=0.001, stopping.deriv=0.1)
gd$x
```

```
##  [1] -0.87346579  0.61828938  0.87989797 -0.04910821  0.07277491
##  [6]  0.10229815 -0.12513246 -0.14559243 -0.11903666 -0.02250130
```

```
gd$k
```

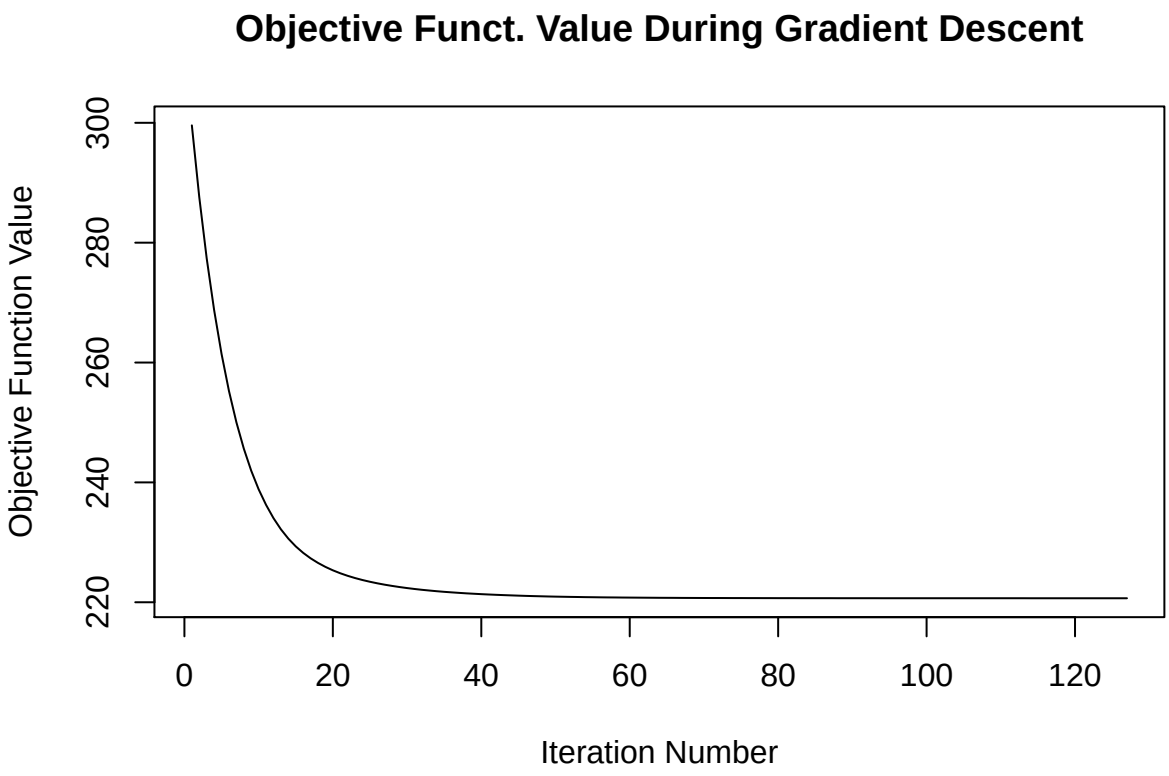
```
## [1] 127
```

The final estimate is given in `gd$x` and `gd$k` gives the number of iterations.

Note: you may need to run `install.packages("numDeriv")` in order to load the `numDeriv` library.

5. Using `gd`, construct a vector `obj` of the values objective function encountered at each step of gradient descent. Note: here the objective function for minimization is the Huber loss. Plot these values against the iteration number, to confirm that gradient descent is indeed making the objective function at each iteration. How does the progress of the algorithm compare at the start (early iterations) versus towards the end (later iterations)?

```
obj <- apply(gd$xmat[, 1:gd$k], 2, huber.loss)
plot(1:gd$k, obj, xlab = "Iteration Number", ylab = "Objective Function Value", type = "l", main = "Obj
```



The value of the objective function decreases sharply for the first 40 iterations or so, but then the progress slows down, with only small decreases for the final 44 iterations.

6. Rerun gradient descent as in question 4, but with `step.size=0.1`. Compute the new criterion values across iterations, and plot the last fifty criterion values. What do you notice now? Is the criterion decreasing at each step, and has gradient descent converged at the end (settled on a single criterion value)? What can you deduce from your plot is happening to the coefficient estimates (confirm this by looking at the `xmat` values in `gd`)?

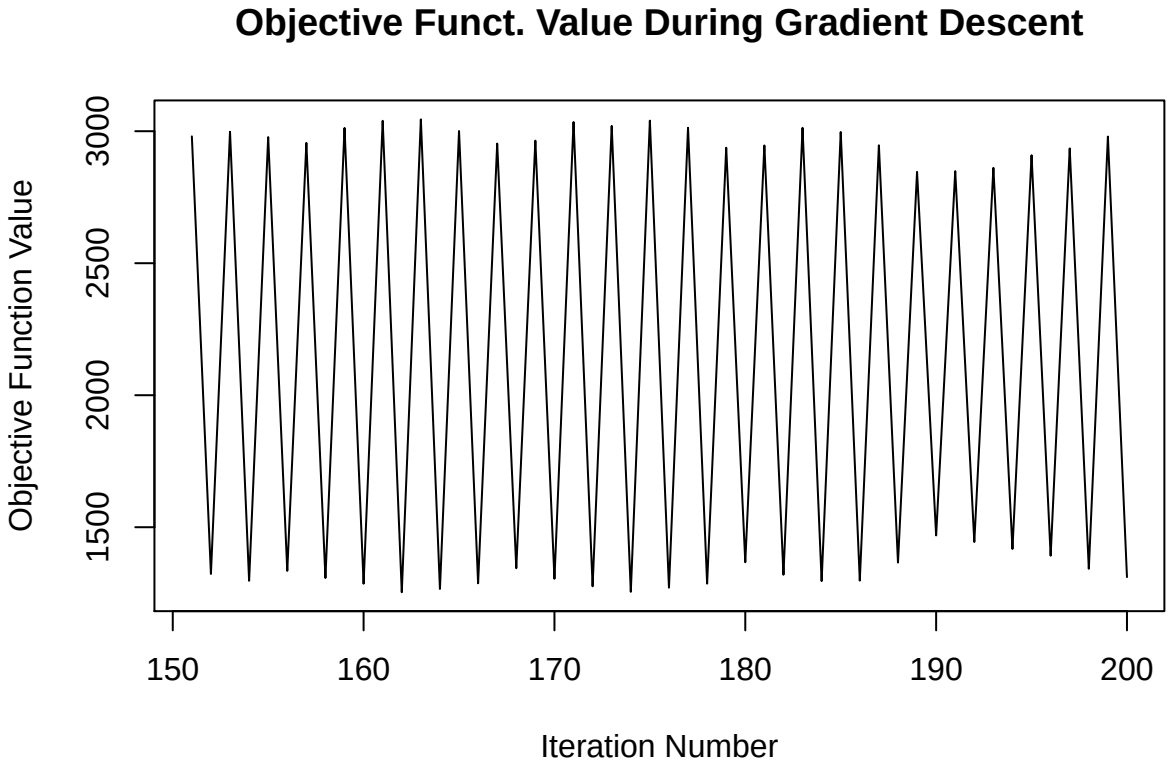
```
gd2 <- grad.descent(huber.loss, x0 = rep(0,p), max.iter=200, step.size=0.1, stopping.deriv=0.1)
gd2$x
```

```
##  [1]  1.0740298 -0.7971898  2.8860325 -1.8822687  2.1897562  0.8721260
##  [7] -1.0055026 -1.5049278  0.9241456  4.7508245
```

```
gd2$k
```

```
## [1] 200
```

```
obj <- apply(gd2$xmat[, 1:gd2$k], 2, huber.loss)
plot((gd2$k-49):gd2$k, obj[(gd2$k- 49):gd2$k], xlab = "Iteration Number", ylab = "Objective Function Va
```



The gradient descent algorithm has not converged, and it's not decreasing at each step but rather oscilatting between values of the objective function. The coeffcieint estimates seems to be bouncing back andforth between two points.

7. Inspect the coefficients from the first gradient descent run (stored in `gd`), and compare them to the true (unknown) underlying coefficients `b` constructed in question 1. They should be pretty close for the first 3 variables, but the next 7 are not very accurate—that is, they're not all close to 0, as they should be. In order to fix this, we're going to apply a **sparsified** version of gradient descent (formally known as proximal gradient descent). Modify the function `grad.descent()` so that at every iteration k, after taking a gradient step but before saving the new estimated coefficients, we threshold small values in these coefficients to zero. Here small means less than or equal to 0.05, in absolute value. Call the new function `sparse.grad.descent()` and rerun with the same settings as in question 4, in order to produce a sparse estimate of the regression coefficients. Stores the results in `gd.sparse`. What are the final coefficient estimates?

```
gd$x
```

```
##  [1] -0.87346579  0.61828938  0.87989797 -0.04910821  0.07277491
##  [6]  0.10229815 -0.12513246 -0.14559243 -0.11903666 -0.02250130
```

```
b
```

```
##  [1] -0.7  0.7  1.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0
```

Indeed, we see that the first three values of `gd$x` are close to the first three values of `b`, but the remaining values aren't that close to 0.

```
sparse.grad.descent <- function(f, x0, max.iter = 200, step.size = 0.05, stopping.deriv = 0.01, ...) {

  n    <- length(x0)
  xmat <- matrix(0, nrow = n, ncol = max.iter)
  xmat[,1] <- x0
```

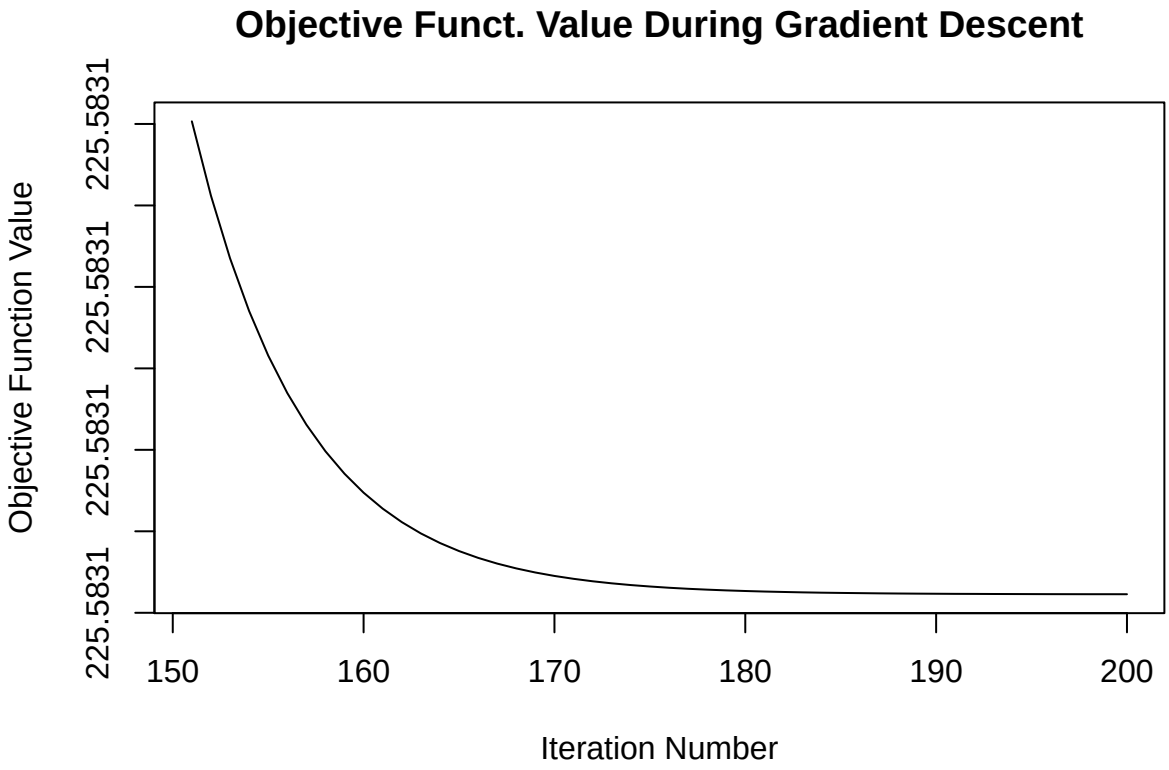
```
  for (k in 2:max.iter) {
    # Calculate the gradient
    grad.cur <- grad(f, xmat[ ,k-1], ...)

    # Should we stop?
    if (all(abs(grad.cur) < stopping.deriv)) {
      k <- k-1; break
    }

    # Move in the opposite direction of the grad and threshold
    update                      <- xmat[ ,k-1] - step.size * grad.cur
    update[abs(update) < 0.05] <- 0
    xmat[ ,k]                   <- update
  }

  xmat <- xmat[ ,1:k] # Trim
  return(list(x = xmat[,k], xmat = xmat, k = k))
}

gd.sparse <- sparse.grad.descent(huber.loss, x0 = rep(0,p), max.iter=200, step.size=0.001, stopping.deri
gd.sparse$x
```

```
##  [1] -0.8944804  0.6332991  0.8823860  0.0000000  0.0000000  0.0000000
##  [7]  0.0000000  0.0000000  0.0000000  0.0000000
```

```
gd.sparse$k
```

```
## [1] 200
```

```
# The below isn't necessary to be included in the homework, but it's interesting.
obj <- apply(gd.sparse$xmat[, 1:gd.sparse$k], 2, huber.loss)
plot((gd.sparse$k-49):gd.sparse$k, obj[(gd.sparse$k- 49):gd.sparse$k], xlab = "Iteration Number", ylab =
```



8. Now compute estimates of the regression coefficients in the usual manner, using `lm()`. How do these compare to those from question 4, from question 7? Compute the mean squared error between each of these three estimates of the coefficients and the true coefficients `b`. Which is best?

```
lm0 <- lm(y ~ -1 + x)
lm0$coef
```

```
##            x1            x2            x3            x4            x5
## -0.9477210986  0.4864220270  0.5875664655 -0.7416200316  0.0008874065
##            x6            x7            x8            x9           x10
##  0.3149846567 -0.3994729398 -0.2712937636 -0.1445449407  0.0788007924
```

```
gd$x
```

```
##  [1] -0.87346579  0.61828938  0.87989797 -0.04910821  0.07277491
##  [6]  0.10229815 -0.12513246 -0.14559243 -0.11903666 -0.02250130
```

```
gd.sparse$x
```

```
##  [1] -0.8944804  0.6332991  0.8823860  0.0000000  0.0000000  0.0000000
##  [7]  0.0000000  0.0000000  0.0000000  0.0000000
```

The linear model coefficients match closest to the solution found using regular gradient descent. This is not surprising, as there is no reason that least squares regression would provide a sparse solution.

```
mse.loss <- function(beta) {
  mean((b - beta)^2)
}

mse.loss(lm0$coef)
```

```
## [1] 0.1186581
```

```
mse.loss(gd$x)
```

```
## [1] 0.01208955
```

```
mse.loss(gd.sparse$x)
```

```
## [1] 0.005610471
```

Here the sparse estimate of the vector has the smallest MSE compared to the true coefficients `b`.

9. Rerun your Huber loss minimization in questions 4 and 7, but on different data. That is, just generate another copy of `y`, per the same formula as you used in question 1: `y = x %*% b + rt(n, df=2)`. How do the new coefficient estimates look from gradient descent, and sparsified gradient descent? Which has a better mean squared error when measured against the `b` used to generate data in question 1? What do you deduce about the sparse method (e.g., what does this suggest about the variability of its estimates)?

In order to ensure that your results are comparable to other students', please run the following before generating a new `y` vector:

```
set.seed(10)
y <- x %*% b + rt(n, df = 2)

gd <- grad.descent(huber.loss, x0 = rep(0,p), max.iter=200, step.size=0.001, stopping.deriv=0.1)
gd$x
```

```
##  [1] -0.46329748  0.92390614  0.92287242 -0.06526259  0.24633002
##  [6] -0.04406371  0.01858892 -0.18921630  0.19479185 -0.18395820
```

```
gd$k
```

```
## [1] 120
```

```
gd.sparse <- sparse.grad.descent(huber.loss, x0 = rep(0,p), max.iter=200, step.size=0.001, stopping.deri
gd.sparse$x
```

```
##  [1] 0.0000000 0.7850744 0.9398727 0.0000000 0.0000000 0.0000000 0.0000000
##  [8] 0.0000000 0.0000000 0.0000000
```

```
gd.sparse$k
```

```
## [1] 200
```

```
mse.loss(gd$x)
```

```
## [1] 0.02869228
```

```
mse.loss(gd.sparse$x)
```

```
## [1] 0.0500853
```

In this case, regular gradient descent estimate of the vector has the smallest MSE compared to the true coefficients `b`. This suggests high variability in the sparse estimate (comparing this result to question 8).

10. Repeat the experiment from question 9, generating 10 new copies of `y`, running gradient descent and sparse gradient descent, and recording each time the mean squared errors of each of their coefficient estimates to `b`. Report the average mean squared error, for gradient descent, and its sparse variant, over the 10 trials. Which average lower? Also report the minimum mean squared error, for the two methods, over the 10 trials. Which is lower? Is this in line with your interpretation of the variability associated with the sparse gradient descent method?

```
num <- 10
sparse.MSEs <- rep(NA, num)
reg.MSEs    <- rep(NA, num)

for (i in 1:10) {
  y <- x %*% b + rt(n, df = 2)

  gd <- grad.descent(huber.loss, x0 = rep(0,p), max.iter=200, step.size=0.001, stopping.deriv=0.1)

  gd.sparse <- sparse.grad.descent(huber.loss, x0 = rep(0,p), max.iter=200, step.size=0.001, stopping.d

  reg.MSEs[i]    <- mse.loss(gd$x)
  sparse.MSEs[i] <- mse.loss(gd.sparse$x)
}

mean(reg.MSEs)
```

```
## [1] 0.02495459
```

```
mean(sparse.MSEs)
```

```
## [1] 0.02650818
```

```
min(reg.MSEs)
```

```
## [1] 0.01430856
```

```
min(sparse.MSEs)
```

```
## [1] 0.0006265157
```

The minimum of the MSEs with the sparse estimates is much smaller, but the mean of the MSEs for these estimates is larger. This supports our interpretation – that the sparse estimate MSE has high variance.