STAT 102B: Homework 2

George Michailidis

Due electronically through
BruinLearn/Gradescope on Monday April 28 at 11:00 pm

Problem 1: This homework focuses on logistic regression and optimizing the corresponding log-likelihood function.

The logistic regression model is discussed in detail in Lecture 3.2

The negative log-likelihood function we want to minimize with respect to the regression coefficient β is given by

$$\ell(\beta) = -\sum_{i=1}^{m} \big[y_i \log\big(\pi(x_i, \beta)\big) + (1 - y_i) \log(1 - \pi(x_i, \beta)) \big],$$

where $y_i \in \{0, 1\}$, $x_i \in \mathbb{R}^p$ and $\beta \in \mathbb{R}^p$.

Further,

$$\pi(x_i, \beta) = \frac{1}{1 + \exp(-x_i^\top \beta)}$$

Implement the following algorithms to obtain estimates of the regression coefficients β:

1. Gradient descent **with backtracking line search**.
2. Gradient descent with **backtracking line search and Nesterov momentum**.
3. Gradient descent with **AMSGrad-ADAM momentum** (no backtracking line search, since AMSGrad-ADAM adjusts step sizes per parameter using momentum and adaptive scaling).

4. Stochastic gradient descent with a **fixed schedule of decreasing step sizes**.
5. Stochastic gradient descent with **AMSGrad-ADAM-W momentum** (no backtracking line search, since the method adjusts step sizes per parameter using momentum and adaptive scaling).

To test your results use the `dataset-logistic-regression.csv`. The first column corresponds to the response $y \in \{0, 1\}$ and the remaining 100 columns to the 100 predictors.

To compare the quality of your results, use the following command in `R` that calculates the regression coefficient based on logistic regression (the gold standard for this data set)

```
glm_fit <- glm(y ~ 0 + X, family = binomial(link = "logit"))
```

Part (a): Discuss how you selected the various hyperparameters for **each** of the algorithms.

For example,

- for gradient descent with backtracking, what are the choices of ϵ (in the Armijo condition) and τ?
- for SGD, the decreasing step size schedule implemented.
- For AMSGrad-ADAM, the β_1 and β_2 parameters and initial step size η_0.
- For AMSGrad-ADAM-W, the β_1, β_2, η_0 and λ parameters.

Part (b)

For SGD and SGD with AMSGrad-ADAM-W, compare the following mini-batch sizes $s = \{100, 200, 500\}$.

For all five methods and the different step sizes for SGD and SGD with AMSGrad-ADAM-W (9 in total), report

1. The **estimation error** defined as:

$$\|\hat{\beta}_{\text{algo}} - \hat{\beta}_{\text{GLM}}\|_2^2$$

2. The number of iterations for the corresponding method

Organize your results in a **table** and write down in bullet form your main findings.