STAT 102B: Sample Exam I Questions

George Michailidis

1 Problems that require calculations

Problem 1:

Let $f(x) : (0, +\infty) \to \mathbb{R}$ with

$$f(x) = x - \log(x),$$

where $\log(\cdot)$ denote the **natural base** e logarithm

1. Show that $f(x)$ has a **unique global minimum** in $(0, +\infty)$. Justify your answer.
2. Let $x_0 = 2$ be the initial point used in the gradient descent algorithm. What will x_1 be based on the gradient descent algorithm, if the step size is set to $\eta = 0.5$?
3. Derive the range of values for the step size parameter η, so that gradient descent is convergent? Justify your answer.
4. Suppose that somebody that does not know how to derive the range of eligible step sizes η, decided to use an initial $\eta = 2$ for initial point $x_0 = 2$. Explain what calculations the backtracking line search algorithm will check to select and appropriate step size η_0 to proceed calculating the next update x_1.

Problem 2:

Let $n \geq 1$ be an integer and let $A \in \mathbb{R}^{n \times n}$ be a symmetric matrix (not necessarily positive definite) for which all of its eigenvalues are non-zero. Let $a\mathbb{R}^m$ be a given vector and we consider the function $f : \mathbb{R} \to \mathbb{R}$ given by

$$f(x) = \frac{1}{2}(x - a)^\top AA(x - a),$$

1. Does the function $f(x)$ have a minimizer? If yes, derive it, otherwise argue why not. Show your work.
2. If one decides to use the gradient descent algorithm, how should η be selected? Justify your answer.

Problem 3:

Consider the function $f : \mathbb{R} \to \mathbb{R}$ with

$$f(x) = (x + 2)^2$$

1. Write pseudo-code that implements the gradient descent algorithm, with **optimal selection** of the step size η.
2. If $x_0 = 1$ was selected as the initial point, what would the value of the optimal η be to calculate x_1 based on the gradient descent algorithm?
3. Based on the optimal step size derived in the previous part, calculate the value of x_1.
4. What are the admissible values of η that would lead to a convergent sequence of iterates x_k for the gradient descent algorithm with fixed step size? Justify your answer.

Problem 4:

Somebody run gradient descent for 20 iterations with $\eta = 0.3$ and computes the update x_k and the norm of the gradient $\nabla f(x_k)$; i.e, $||\nabla f(x_k)||_2$ after each iteration.

It is observed that the norm $||\nabla f(x_k)||_2$ decreases quickly and then levels off. Based on this, which of the following would be your recommendation:

- Use a larger value of η; e.g., $\eta = 0.5$.
- Use a larger value of η; e.g., $\eta = 0.1$.
- Keep $\eta = 0.3$.
- Additional information is needed to make a recommendation. What information would you need?

Justify your answer.

Problem 5:

Suppose you have a data set comprising of 1 million observations and 100,000 predictors. You want to use multivariate linear regression to estimate the 100,000 regression coefficients from the data.

- Should you prefer the closed form least squares solution or use the gradient descent algorithm? Justify your answer.
- If you decided to use the gradient descent algorithm, would you be able to find the optimal regression coefficients? Justify your answer.
- What strategy would you use to select the step size in the gradient descent algorithm? Justify your answer.

Problem 6: Write pseudo-code for a function $f : \mathbb{R}^m \to \mathbb{R}$ to perform gradient descent based on the backtracking line search algorithm.

2 Multiple choice Quiz Questions

Question 1: Which direction does the gradient descent algorithm move in each iteration?

- ☐ Random direction
- ☐ Direction of the gradient
- ☐ Opposite to the gradient
- ☐ Along the eigenvectors of the Hessian

Question 2: If the step size (learning rate) for the optimization problem of a statistical model is too large, gradient descent can:

- ☐ Converge slowly
- ☐ Not converge
- ☐ Overfit the data
- ☐ Always converge faster

Question 3: In the context of gradient descent, a "step size schedule" is used to:

- ☐ Randomly choose learning rates
- ☐ Ensure faster convergence
- ☐ Decrease learning rate over time
- ☐ Increase gradient magnitude

Question 4: In the heavy ball momentum method, the new direction is a combination of:

- ☐ Current gradient and noise

- ☐ Previous parameter and current value
- ☐ Previous update and current gradient
- ☐ Gradient norm and function value

Question 5: In AdaGrad, the learning rate is:

- ☐ Constant for all parameters
- ☐ Increasing over time
- ☐ Decreasing for frequently updated parameters
- ☐ Randomly sampled at each step

Question 6: Why is bias correction used in ADAM?

- ☐ To make convergence faster
- ☐ To compensate for initialization at zero
- ☐ To prevent exploding gradients
- ☐ To add noise to the gradient update

Question 7: In ADAM, the step size is:

- ☐ Fixed
- ☐ Increasing
- ☐ Scaled by gradient history
- ☐ Decreased linearly

Question 8: Mini-batch SGD uses:

- ☐ The entire data set

- ☐ Only one sample
- ☐ A subset of data points
- ☐ Data sorted by the objective function value

Question 9: What is a common downside of very small batch size SGD?

- ☐ Too slow
- ☐ Too stable
- ☐ High variance in updates
- ☐ Uses entire data set per iteration