Homework 9 Solutions

Math 198: Math for Machine Learning

Due Date:
Name:
Student ID:

Instructions for Submission

Please include your name and student ID at the top of your homework submission. You may submit handwritten solutions or typed ones (LaTeX preferred). If you at any point write code to help you solve a problem, please include your code at the end of the homework assignment, and mark which code goes with which problem. Homework is due by start of lecture on the due date; it may be submitted in-person at lecture or by emailing a PDF to both facilitators.

1 Practice with Newton's Method

1. Let $f(x) = x^4$. For any given x_0, what will the update rule given by Newton's method be? Will it lead us to the function's minimum? Why or why not?
   We have $f'(x) = 4x^3$ and $f''(x) = 12x^2$, so the update rule is $x_{k+1} = \frac{2}{3}x_k$. By inspection, the minimum of $f(x)$ is $x = 0$; Newton's method will converge towards this value but never reach it.

2. Let $f(x) = x^3$. For any given x_0, what will the update rule given by Newton's method be? Will it lead us to the function's minimum? Why or why not?
   We have $f'(x) = 3x^2$ and $f''(x) = 6x$, so the update rule is $x_{k+1} = \frac{1}{2}x_k$. This function does not have a minimum, and so Newton's method will get stuck near the saddle point $x = 0$.

2 Gauss-Newton Algorithm Proofs

1. Let $f(\mathbf{x}; \beta)$ be a nonlinear function from $\mathbb{R}^n \to \mathbb{R}$ parameterized by an m-dimensional vector β. Define $L(\beta) = \sum_{i=1}^{n}(y_i - f(\mathbf{x_i}; \beta))^2 = ||\mathbf{r}(\beta)||_2^2$ be the loss function we wish to minimize. (Note that this is an equivalent formulation to how we present the Gauss-Newton algorithm in note 9.)

   (a) Show that $\nabla L(\beta) = 2\mathbf{J_r}^\top(\beta)\mathbf{r}(\beta)$.
   First, we note that

$$\nabla L(\beta)_j = \frac{\partial L}{\partial \beta_j} = \frac{\partial}{\partial \beta_j}\Big[\sum_{i=1}^{n} r_i(\beta)^2\Big] = \sum_{i=1}^{n}\frac{\partial}{\partial \beta_j}\big[r_i(\beta)^2\big] = 2\sum_{i=1}^{n} r_i(\beta)\frac{\partial r_i}{\partial \beta_j}$$

   We recall that $\mathbf{r}$ is composed of residual functions $r_i(\beta) = y_i - f(\mathbf{x_i}; \beta)$, and so

$$\mathbf{J_r}(\beta) = \begin{bmatrix} \nabla r_1(\beta)^\top \\ \vdots \\ \nabla r_n(\beta)^\top \end{bmatrix} = \begin{bmatrix} \frac{\partial r_1}{\partial \beta_1} & \cdots & \frac{\partial r_1}{\partial \beta_m} \\ \vdots & \ddots & \vdots \\ \frac{\partial r_n}{\partial \beta_1} & \cdots & \frac{\partial r_n}{\partial \beta_m} \end{bmatrix}$$

We therefore have

$$\mathbf{J_r}^\top(\beta)\mathbf{r}(\beta) = \begin{bmatrix} \frac{\partial r_1}{\partial \beta_1} & \cdots & \frac{\partial r_n}{\partial \beta_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial r_1}{\partial \beta_m} & \cdots & \frac{\partial r_n}{\partial \beta_m} \end{bmatrix} \begin{bmatrix} r_1(\beta) \\ \vdots \\ r_n(\beta) \end{bmatrix}$$

and so $2[\mathbf{J_r}^\top(\beta)\mathbf{r}(\beta)]_j = 2\sum_{i=1}^n r_i(\beta)\frac{\partial r_i}{\partial \beta_j} = \nabla L(\beta)_j$ as desired.

(b) Show that $\nabla^2 L(\beta) = 2(\mathbf{J_r}(\beta)^\top \mathbf{J_r}(\beta) + \sum_{i=1}^n r_i(\beta)\nabla^2 r_i(\beta))$.
We again proceed by noting

$$\nabla^2 L(\beta)_{jk} = \frac{\partial^2 L}{\partial \beta_j \partial \beta_k} = \frac{\partial}{\partial \beta_j}\big[2\sum_{i=1}^n r_i(\beta)\frac{\partial r_i}{\partial \beta_k}\big] = 2\sum_{i=1}^n \frac{\partial}{\partial \beta_j}\big[r_i(\beta)\frac{\partial r_i}{\partial \beta_k}\big]$$

Using the product rule, we then derive

$$\frac{\partial}{\partial \beta_j}\big[r_i(\beta)\frac{\partial r_i}{\partial \beta_k}\big] = \frac{\partial r_i}{\partial \beta_j}\frac{\partial r_i}{\partial \beta_k} + r_i(\beta)\frac{\partial^2 r_i}{\partial \beta_j \partial \beta_k}$$

and so

$$\nabla^2 L(\beta)_{jk} = 2\sum_{i=1}^n \big[\frac{\partial r_i}{\partial \beta_j}\frac{\partial r_i}{\partial \beta_k} + r_i(\beta)\frac{\partial^2 r_i}{\partial \beta_j \partial \beta_k}\big]$$

We then have

$$\big[\mathbf{J_r}(\beta)^\top \mathbf{J_r}(\beta)\big]_{jk} = \sum_{i=1}^n \frac{\partial r_i}{\partial \beta_j}\frac{\partial r_i}{\partial \beta_k}$$

and of course $\nabla^2 r_i(\beta)_{jk} = \frac{\partial^2 r_i}{\partial \beta_j \partial \beta_k}$, so

$$\nabla^2 L(\beta)_{jk} = 2(\big[\mathbf{J_r}(\beta)^\top \mathbf{J_r}(\beta)\big]_{jk} + \sum_{i=1}^n r_i(\beta)\nabla^2 r_i(\beta)_{jk})$$

as desired.