SGD for logistic regression

(A) From exercise01, the gradient descent of β is $\nabla l(\beta)=\sum_{i=1}^{N}(m_i w_i - y_i)x_i$

View as $\hat{y}_i = E(y_i \mid \beta) = m_i w_i$ (expectation of Binomial distribution) the fitted value of y_i,

when given β, the gradient descent is $\nabla l(\beta)=\sum_{i=1}^{N}(\hat{y}_i - y_i)x_i = \sum_{i=1}^{N} g_i(\beta)$

Where $g_i(\beta) = (\hat{y}_i - y_i)x_i$

(B) Since we draw a single data point from the sample, *i* is the only random variable which follows a discrete uniform distribution, i.e.,

$$P(i=k) = \begin{cases} \frac{1}{n} & k \in \{1,2,...,n\} \\ 0 & otherwise \end{cases}$$

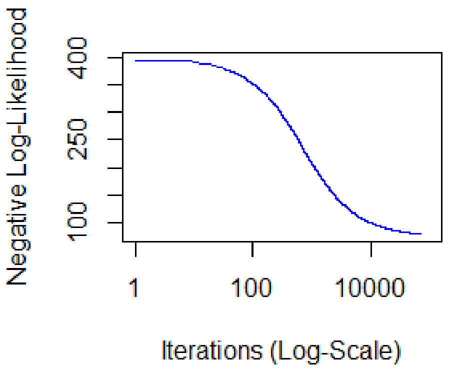
Then the expectation of $ng_i(\beta)$ is

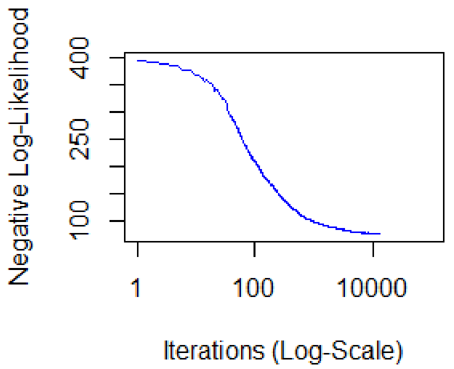
$$E[ng_i(\beta)] = nE[g_i(\beta)] = n\sum_{j=1}^{n} g_j(\beta)P(i=j) = n\sum_{j=1}^{n} g_j(\beta)\frac{1}{n} = \sum_{j=1}^{n} g_j(\beta) = \nabla l(\beta)$$

Therefore, $ng_i(\beta)$ is an unbiased estimator of $\nabla l(\beta)$.

(C) See the attached R code. The convergence condition is

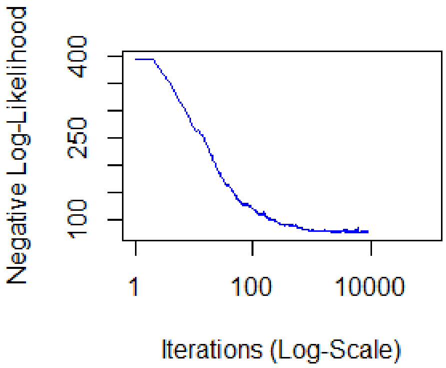
$$\frac{|\, l(\beta^{(t)}) - l(\beta^{(t-1)}) \,|}{|\, l(\beta^{(t-1)}) \,| + \varepsilon} < \varepsilon, \ where \ \varepsilon = 1e-10$$





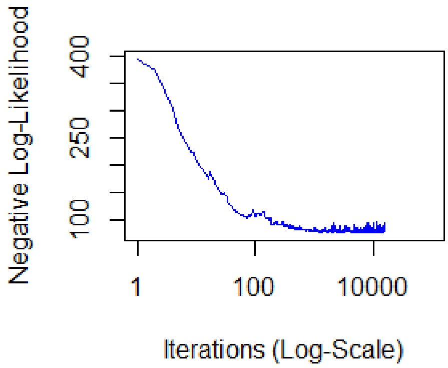
step size γ = 0.001
i) SGD converged in iterations: 68250

step size γ = 0.01
ii) SGD converged in iterations: 13034



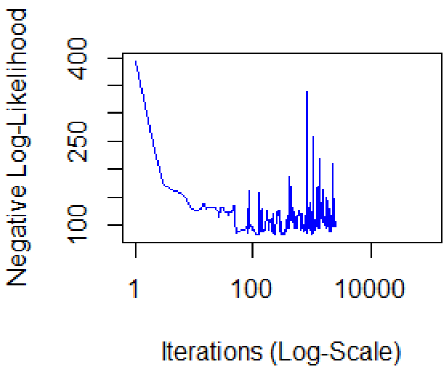
step size $\gamma = 0.05$

iii) SGD converged in iterations: 8699



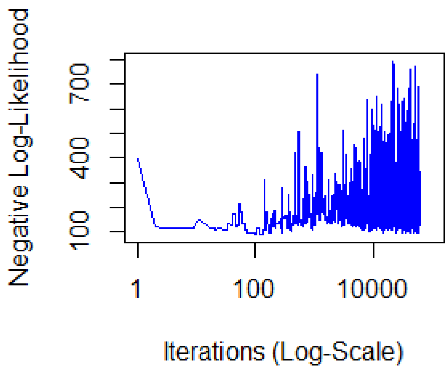
step size $\gamma = 0.1$

iv) SGD converged in iterations: 15317



step size $\gamma = 0.5$

v) SGD converged in iterations: 2504

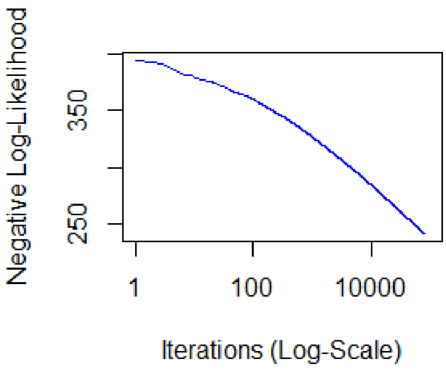


step size $\gamma = 1$

vi) SGD converged in iterations: 59862

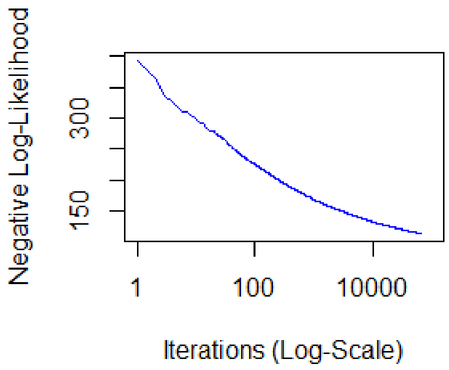
Here gradient is calculated from a single data point, which is sampled randomly from the whole data set. Since this single-data-point is an unbiased estimate of the full-data gradient, we move in the right direction toward the minimum. We choose step size $\gamma = 0.001$, 0.1, 0.05, 0.1, 0.5, and 1 respectively and loop through 100,000 iterations. From the above figures, the negative log likelihood has converged in less than 100,000 iterations with each of the step sizes.

(D) Try a decaying step size using Robbins-Monro rule for step sizes: $\gamma^{(t)} = C(t + t_0)^{-\alpha}$, where $C > 0$, $0.5 \le \alpha \le 1$, and t_0 (the prior number of steps) are constants. The exponent α is usually called the learning rate. Clearly the closer α is to 1, the more rapidly the step sizes decay. Here we pick $\alpha = 0.75$ and $t_0 = 1$, with a range of C = 0.01, 0.1, 1, 10, 50, 100. The results are shown in the figures below and we obtain good approximations of β with different C. The figures show that with larger C like 50 or 100 and a fixed $\alpha = 0.75$, the estimates are obtained really rapidly.



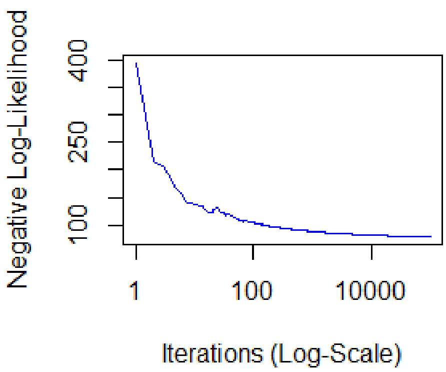
C = 0.01

i) SGD converged in iterations: 79203



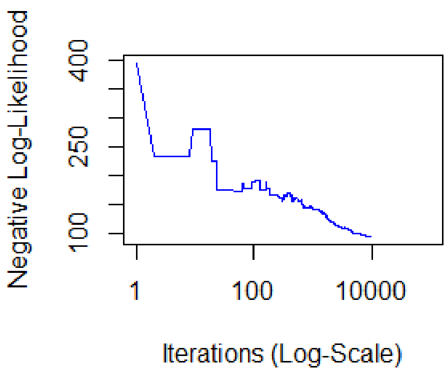
C = 0.1

ii) SGD converged in iterations: 66500



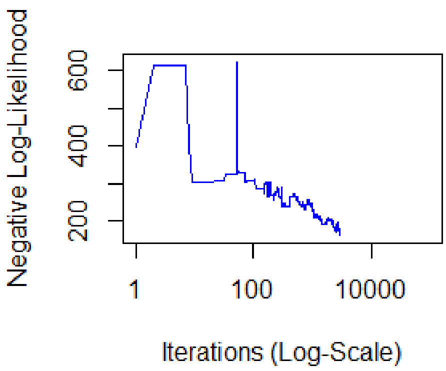
C = 1

iii) SGD did not converge



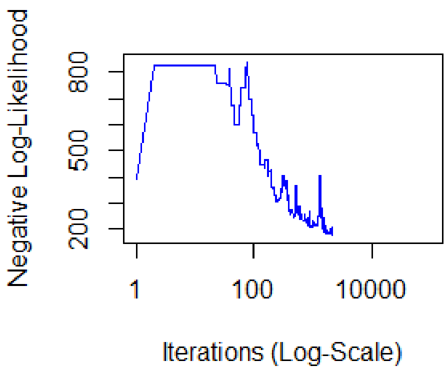
C = 10

iv) SGD converged in iterations: 9330



C = 50

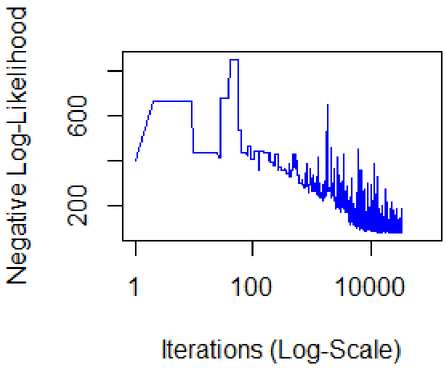
v) SGD converged in iterations: 2939



C = 100

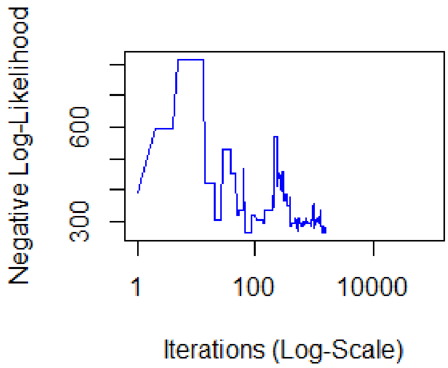
vi) SGD converged in iterations: 2148

Then, we choose varying α = 0.5, 0.6, 0.7, 0.8, 0.9, and 1 with a fixed C = 50. From the figures below, the negative log likelihood converged rather quickly with C = 50 and α = 0.8 (only 81 iterations). This shows that good estimates are obtained rapidly with Robbins-Monro rules.



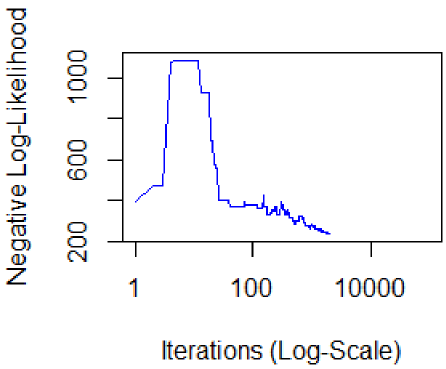
α = 0.5

i) SGD converged in iterations: 33737



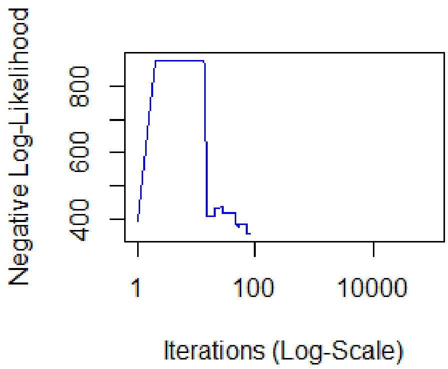
α = 0.6

ii) SGD converged in iterations: 1524



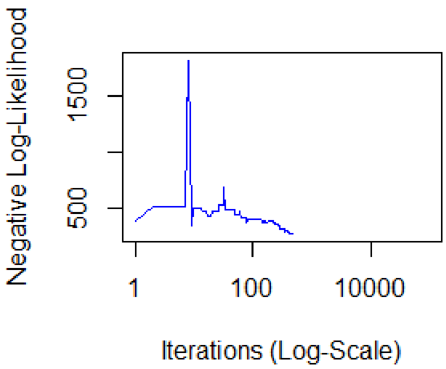
α = 0.7

iii) SGD converged in iterations: 2023



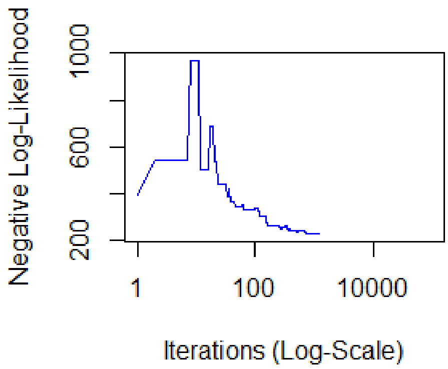
α = 0.8

iv) SGD converged in iterations: 81



α = 0.9

v) SGD converged in iterations: 474



α = 1

vi) SGD converged in iterations: 1190