SDS 385: Homework 2

G. Paulon

September 4, 2016

Problem 1. SGD for logistic regression

(A) In part (A) of the last homework we proved that the gradient of the negative log-likelihood can be expressed as

$$\nabla l(\beta) = -\sum_{i=1}^{n} \{x_i(y_i - m_i w_i)\} = \sum_{i=1}^{n} g_i(\beta)$$

where

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

and

$$\hat{y}_i = \mathbb{E}(y_i|\beta) = m_i w_i(\beta) = m_i \frac{1}{1 + \exp(-x_i^T \beta)}.$$

A nice interpretation can be given to this latter expression: the gradient is large when the data y_i's differ from their maximum likelihood estimates, i.e. the probabilities w_i's.

(B)

(C) The SGD exploits this fact in order to use an update step which is faster to compute. In fact, instead of using the gradient calculated from all n data points to choose the step direction, we use the gradient $g_i(\beta)$ calculated from a single data point, sampled randomly from the whole data set. In this version of the algorithm, sampling without replacement has been performed.

In order to asses the validity of the implementation, we run the algorithm on the real data for a reasonable number of iterations $N = 100000$. We tried different values for the step size γ, which was kept constant over the iterations. We report in Figures 1 - 3 the traceplots of all of the β parameters (i.e. the values of the parameters over the iterations). In order to facilitate the visualization, we thinned the samples by a factor 20 (we displayed one value of β every 20 iterations).

We can see that for small values of the step size γ, convergence is not reached for many of the components of β. When we increase the step size, on the other hand, convergence is reached more rapidly because the algorithm explores well the parameters space. However, the asymptotic variance of the parameters is higher, since at every iteration the weight attributed to the direction given by any new sampled data is high.

In this first implementation of the algorithm we do not worry about the choice of the tuning parameter, that is instead crucial for SGD. As a convergence diagnostics, we plot the iterates of the running average of $l_t(\beta)$, the individual log-likelihood contribution from the data point sampled at step t.

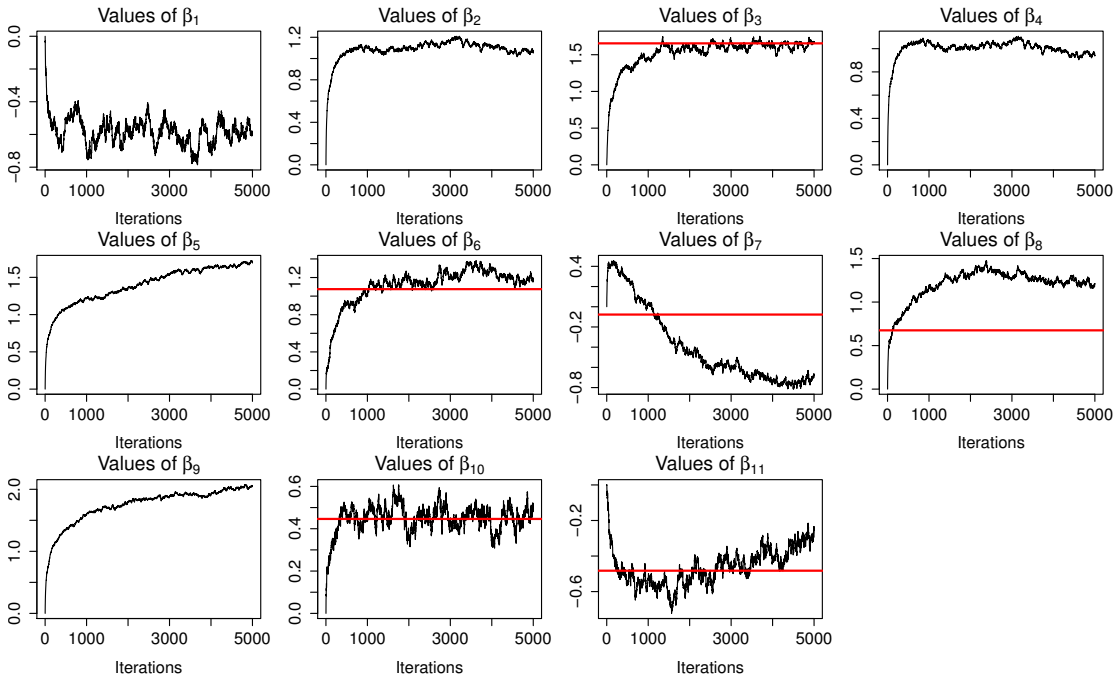

Figure 1: Traceplots of the β parameters when $\gamma = 0.01$.

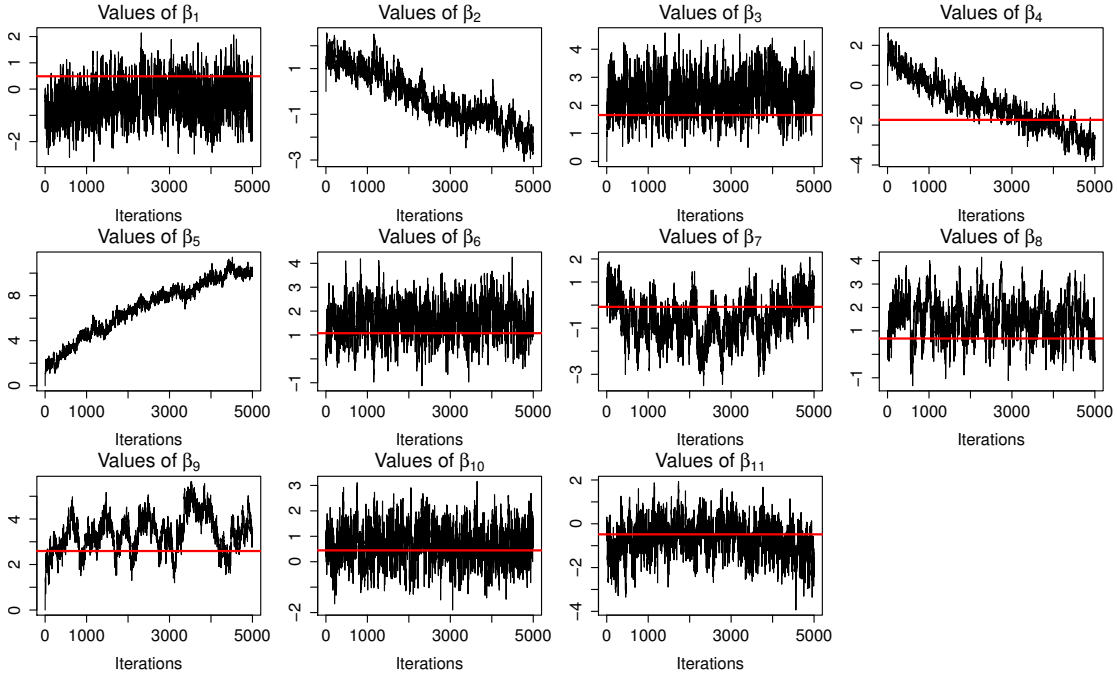

Figure 2: Traceplots of the β parameters when $\gamma = 0.5$.

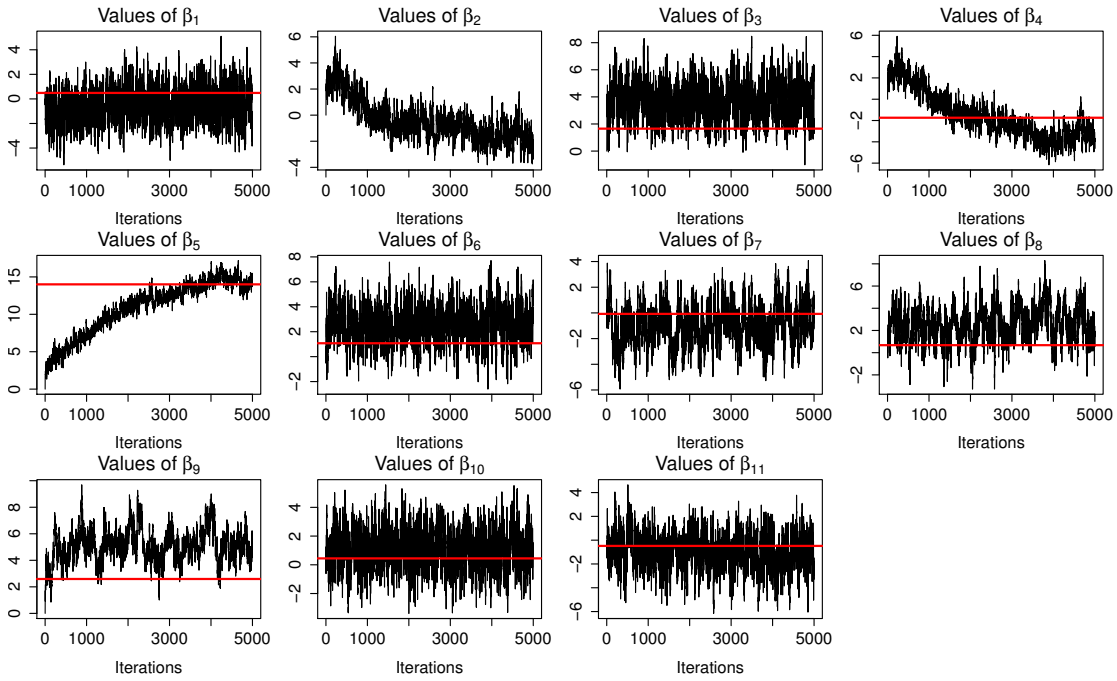

Figure 3: Traceplots of the β parameters when $\gamma = 1$.

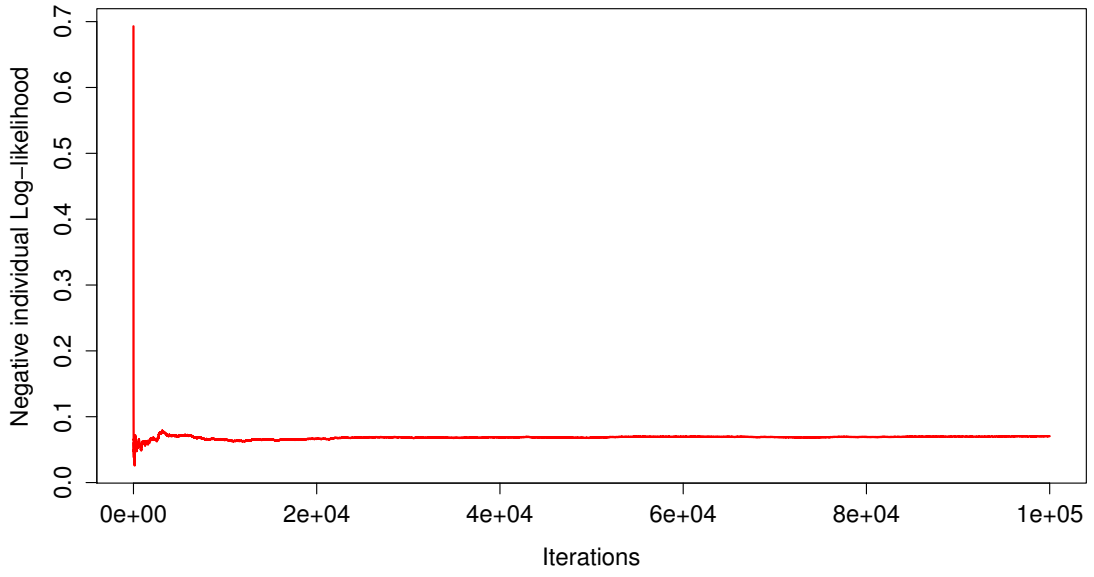

Figure 4: Running average of the single log-likelihood contributions.

(D) Now we try using a decaying step size. Specifically, we use the Robbins–Monro rule for step sizes:

$$\gamma(t) = C(t + t_0) - \alpha,$$

where $C > 0$, $\alpha \in [0.5, 1]$, and t_0 (the "prior number of steps") are constants. The exponent α is usually called the learning rate.

Ideally, we want to obtain large values of γ at the beginning of the algorithms, so that the "true" values of the parameters are rapidly reached. Afterwards, we want the step sizes to be reduced in order to diminish the variance of the estimates and to, eventually, converge.

(E)