RGallery: A Package for 3 questions in Stochastic Average Gradient Project

Eric Xin Zhou

February 18, 2015

1 Easy level

Q1 : Use glmnet to fit an L2-regularized logistic regression model. Use the system.time function to record how much time it takes for several data set sizes, and make a plot that shows how execution time depends on the data set size.

1.1 Data simulation setup for L2

Given that test 1 need us provide different data set of different size to record how much time **glmnet** takes for different data size.

I generate Gaussian data with N observation and p predictors. with each pair of predictors $X_j, X_{j'}$ has the same population correlation ρ. If N and ρ are determined. We generate the observed data Y by adding several gaussian noise.

$$Y = \sum_{j=1}^{p} X_j \beta_j + kZ \tag{1}$$

If Y is a $N \times 1$ column vector, then $X_j, X_{j'}$ are all $N \times 1$ column vectors, so $\mathbf{X}$ is a $N \times p$ matrix and β is a $p \times 1$ column vector.

Z represents noise of observation, and k is chosen so that we can control signal-to-noise ratio to 3.0.

In generation model, we also should simulate the coefficient vector β, we define that

$$\beta_j = (-1)^j \exp\left(\frac{-2(j-1)}{20}\right) \tag{2}$$

This guarantee that the coefficients are constructed to have alternating signs and to be exponential descreasing.