Problem definition

Mathematical model of the problem is as follow:

$$\min_{w \in \mathbb{R}^d} f(w) = \frac{1}{n} \sum_{i=1}^{n} \log(1 + \exp(-y_i w^T x_i)),$$

Data preparing:

1. Clean the data, examined the dataset and did some preliminary analysis of the features.
2. Standardized the data by removing the mean and scaling to the unit variance. Removed some of the features having correlation below than 0.02 to the response.
3. 569 samples with 26 features (n = 569, d = 26).

Algorithms:

Algorithms for providing predictions are including, gradient descent method, gradient descent with momentum (including heavy ball method and Nesterov's method), decomposition-type methods(stochastic gradient descent) and quasi-Newton algorithm(BFGS method) and BB (Barzilai-Borwein) Method (including LBB method and SBB method).

Gradient descent (GD) uses following recursion:

$$w_{i+1} = w_i - \alpha \nabla f(w_i)$$

Where α= 1/L, L is the max eigenvalue of A, and A=X'*X.

Heavy ball method and Nesterov's method:

HB:

$W^{k+1} = w^k - \alpha \nabla f(w^{k)} + beta*(w^k - w^{k-1})$;

Nesterov:

$$y^r = x^r + \beta_r(x^r - x^{r-1}), \quad \text{slip due to momentum}$$
$$x^{r+1} = y^r - \alpha \nabla f(y^r). \quad \text{move along gradient}$$

Stochastic gradient descent method (SGD):

$w^{t+1}=w^t-\alpha^t \nabla f(w^t)$

Quasi-Newton method(BFGS):

The algorithm is:

For $k=0,1,2\ldots.. n$,

initial w_0 and B_0

Obtain a direction pk by $B^k * p^k = -\nabla f(w^k)$; $w^{k+1}=w^k+\alpha^k p^k$; update B^k to B^{k+1}

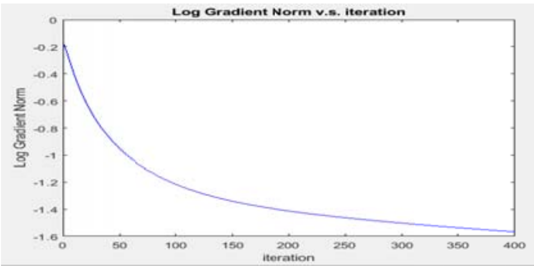
end.

Barzilai-Borwein(BB) method:

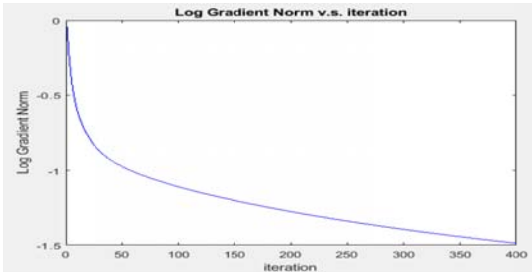
$w^{k+1}=w^k-H^k*\nabla f(w^k)$, where $H^k=I*\alpha k$ I denotes the identity matrix.

Experiments with real world data

GD method



n = 12, d = 5



n = 569, d = 26

HB method

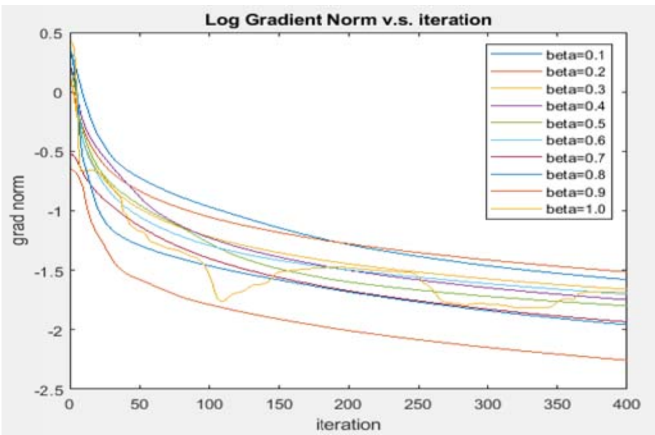
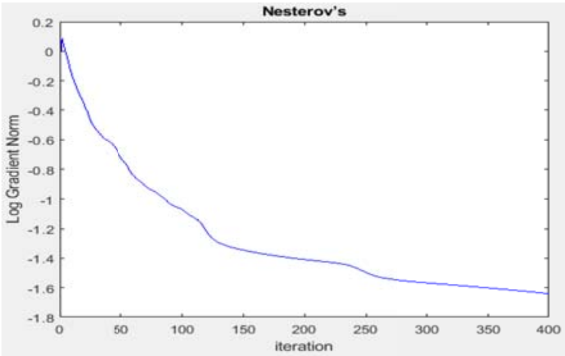
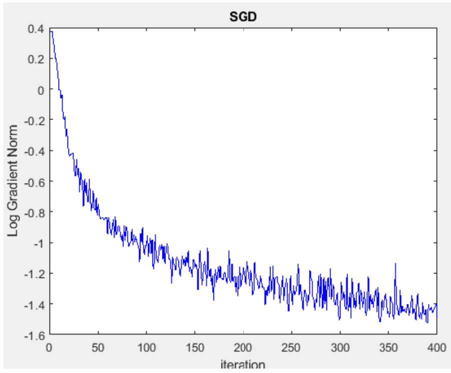


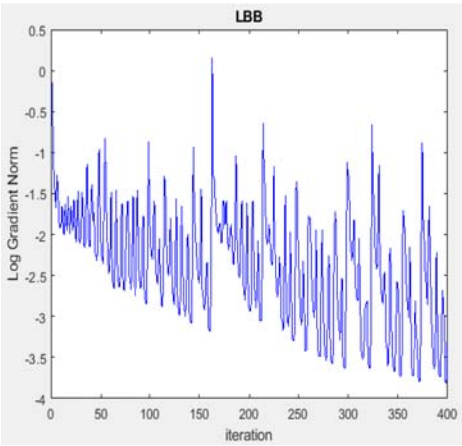
Figure 3

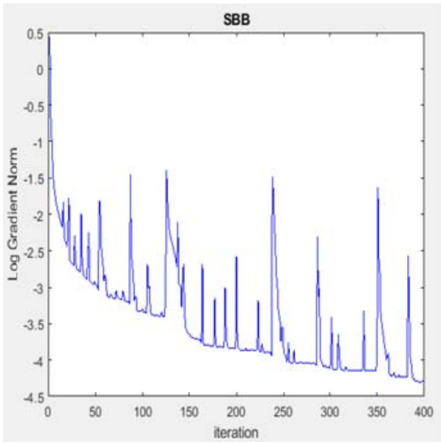
Nesterov's method and SGD

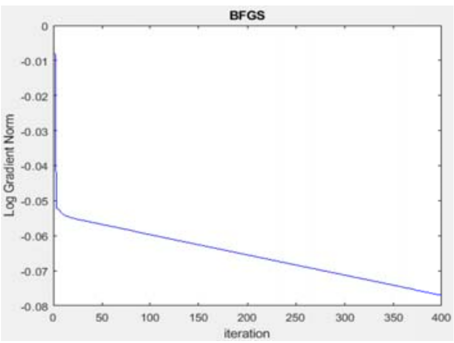




LBB, SBB and BFGS:



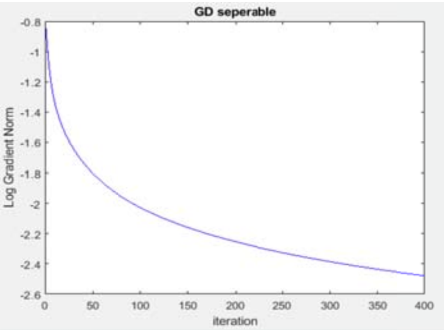


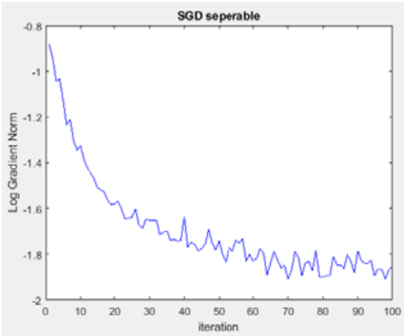


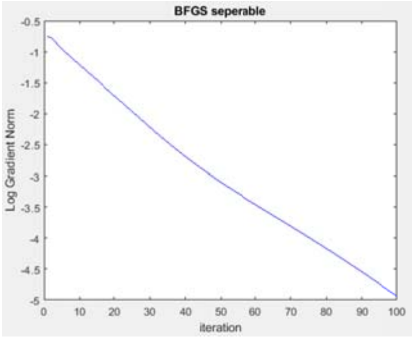
Experiments with Artificial Data

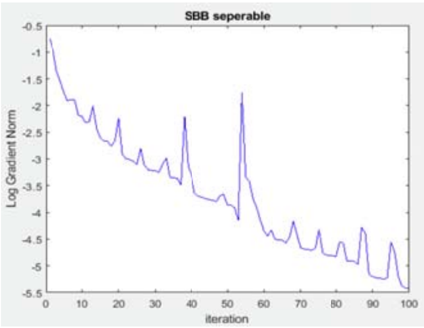
Define $w^* = [1; 1; \ldots; 1] \in \mathbb{R}^{dX1}$, and generate label $y_i = sign(x_i^T w^*), \forall i$. Here, $sign(z) = 1$ if $z \geq 0$ and $sign(z) = -1$ if $z < 0$.

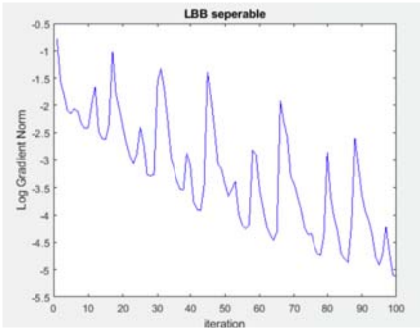
All settings are consistent with previous experiments. Results with artificial data as follow.











Reference

[1] Machine learning. Retrieved from: https://en.wikipedia.org/wiki/Machine_learning

[2] Vapnik, V. (1995). The natural of statistical Learning Theory. Springer, New York.

[3]dataset, available at：https://www.kaggle.com/uciml/breast-cancer-wisconsin-data/data

[4] Allison 2008, Convergence Failures in Logistic Regression. *SAS Global Forum* .

[5]Heinze, George and Michael Schemper (2002) “A Solution to the Problem of Separation in Logistic Regression.” *Statistics in Medicine 21*: 2409-2419.