Rohan Chandra - hmwk7 Solutions

Question 1 to Question 3 plots

NOTE: I was getting unreliable results for $\epsilon = 1e^{-8}$. One of the methods kept getting stuck in an infinite loop. So I used a tolerance value of $1e^{-7}$ for which the gradient descent and Barzelai-Borwein method works perfectly although the Nesterov method is stil somewhat unreliable at this tolerance value. Hence a larger tolerance value for the Nesterov has been chosen (such as $1e^{-6}$) in our experiments.

Figure 1 shows the three convergence plots for $\kappa = 1$. Figure 2 shows the three convergence plots for $\kappa = 100$. You can see that gradient descent converges in about $90,000$ iterations whereas Barzelai-Borwein and Nesterov converges much faster in about 1800 and 3700 iterations respectively. Figure 3 shows a magnified view of the first $10,000$ iterations where the behaviour of the Nesterov and Barzelai-Borwein methods can be more clearly seen.

One observation I would like to highlight is that gradient descent method is the slowest and the Barzelai-Borwein method is the fastest as is proved by the plots below

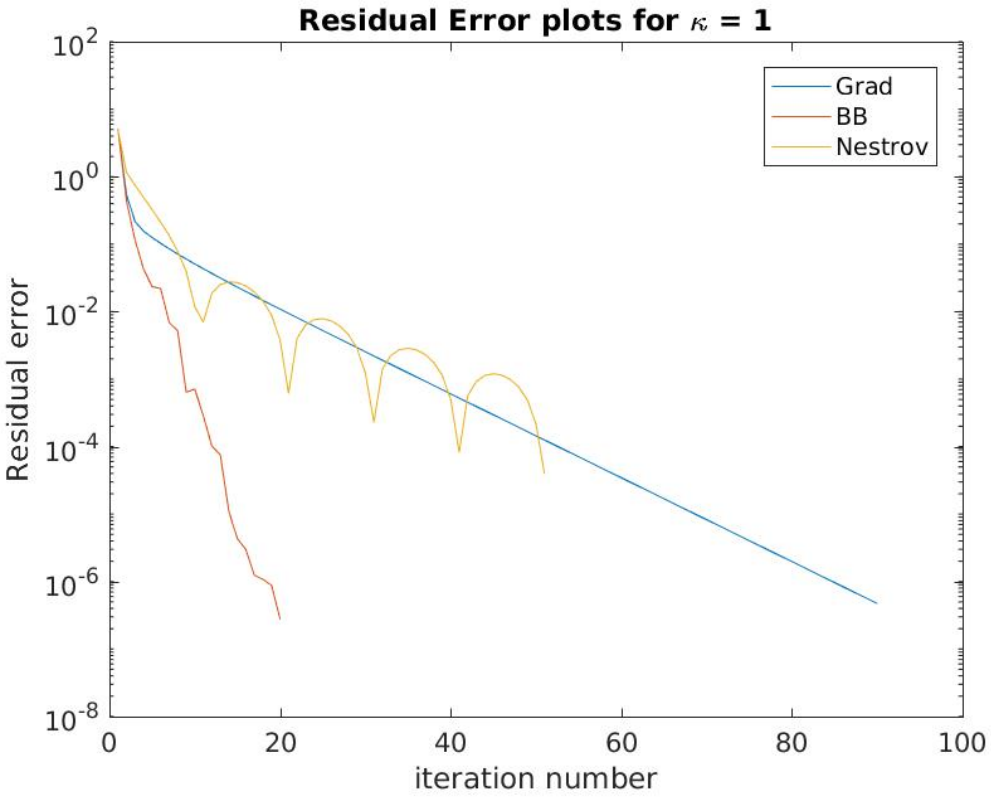

Figure 1: Convergence Plots for $\kappa = 1$. 50 iterations for Nesterov, 20 iterations for Barzelai-Borwein, and 90 iterations for Gradient Descent

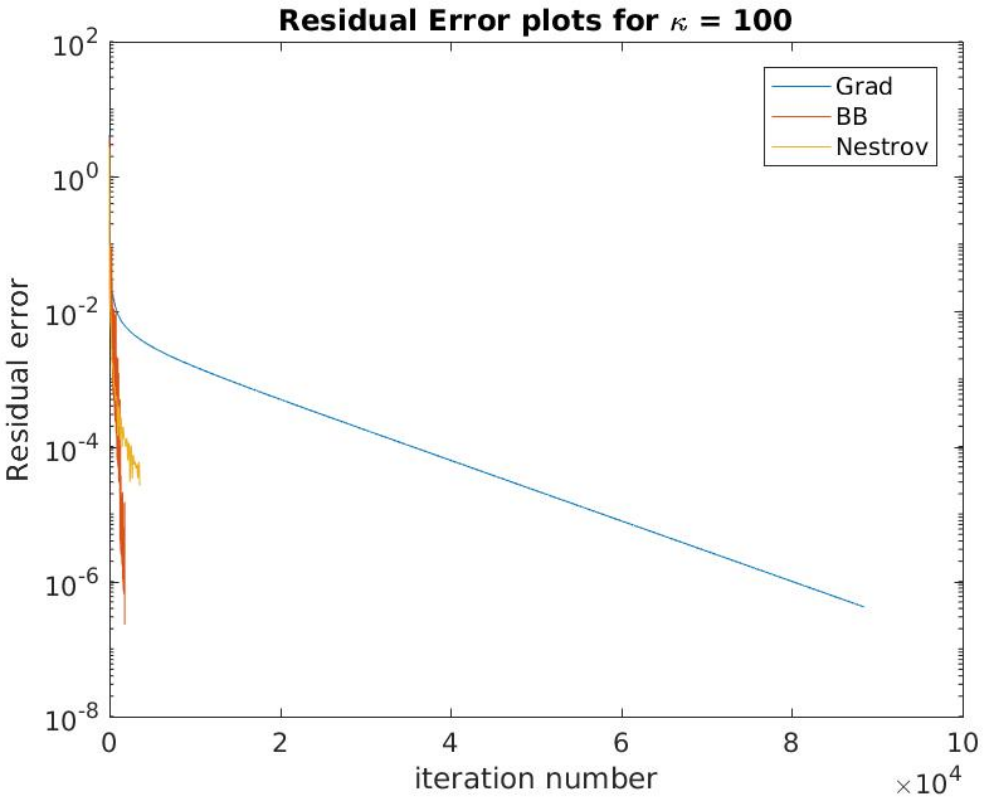

Figure 2: Convergence Plots for $\kappa = 100$. 3700 iterations for Nesterov, 1800 iterations for Barzelai-Borwein, and $90,000$ iterations for Gradient Descent

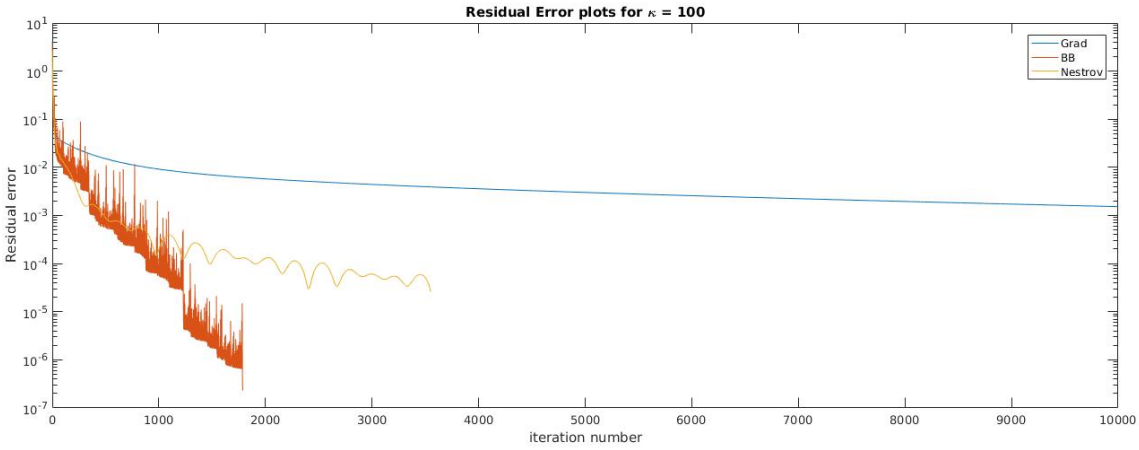

Figure 3: Magnified view for the convergence plots for Barzelai-Borwein and Nesterov methods for $\kappa = 100$

Question 5

5a

From strong convexity and the taylor approximation of the objective function, we impose an upper bound on the hessian which implies

$$f(y) \leq f(x) + \nabla f(x)^T (y - x) + \frac{M}{2}||y - x||^2$$

, where M is the libschitz constant.

From gradient descent, we have $y = x - \tau \nabla f(x)$. Substituting this in the above equation, and following the convergence proof we have,

$$f(x^{k+1}) \leq f(x^k) - \frac{2}{M}||\nabla f(x)||^2$$

.

Subtracting $f(x^*)$ from both sides, we have

$$f(x^{k+1}) - f(x^*) \leq f(x^k) - f(x^*) - \frac{2}{M}||\nabla f(x)||^2 \qquad (1)$$

From the lower bound on the hessian, we also have

$$||\nabla f(x)||^2 \geq 2m[f(x) - f(x^*)] \qquad (2)$$

where x^* is the optimal value and $f(x^{k+1}) - f(x^*)$ is the objective function.

Combining (1) and (2) and subtracting $f(x^*)$ from both sides, we have

$$f(x^{k+1}) - f(x^*) \leq (1 - \frac{m}{M})[f(x^k) - f(x^*)] \qquad (3)$$

where $f(x^{k+1}) - f(x^*)$ is the objective function.

From (3), we say the following:

The equality in (3) gives us a direct stopping condiiton in terms of objective functions if we have knowledge of m and M. However, m and M are known only in rare cases. This makes (3), and by extension, usage of objective functions impractical to use as as a stopping condition.

From (1), we infer that if $||\nabla f(x)||$ is sufficiently small, then $f(x^{k+1}) - f(x^*) \approx f(x^k) - f(x^*)$ which leads us to the objective function criterion.

Hence from (1),

$$||\nabla f(x)|| \to 0 \implies f(x^{k+1}) - f(x^*) < \epsilon$$

5b

From the previous question, we know that the value of tolerance depends on m and M. **In most cases, we do not know these values**. This means that we need an anchor point to steer our algorithm, otherwise we would be flying blind with no bearings. So we scale the tolerance with the gradient of the initial input $x0$.

5c

In terms of faster convergence, the Barzelai-Borwein method was definitely the fastest. Then came Nesterov and finally gradient descent

5d

High condition number $\implies$ tall, flat elliptical contours $\implies$ gradient direction bounces between (almost) parallel lines and **take longer to converge.**

Refer to figure 4

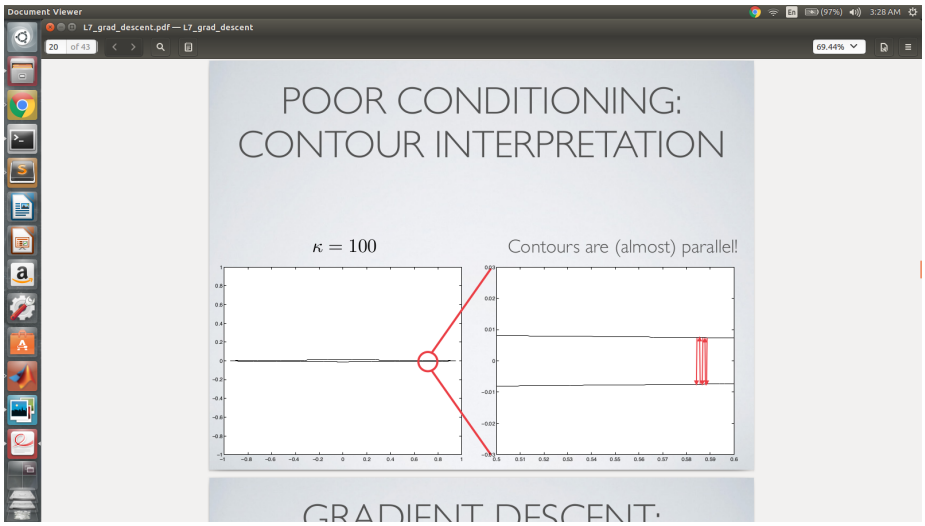

Figure 4: Large condition number leads to more iterations and slower convergence.

Question 6

The accuracy for training set is 99.5617% and the accuracy for test set is 95.29%. Kindly refer to figure 5 for the converence plot

```
Accuracy for Training Set = 99.5617
Accuracy for Test Set = 95.29
>>
```

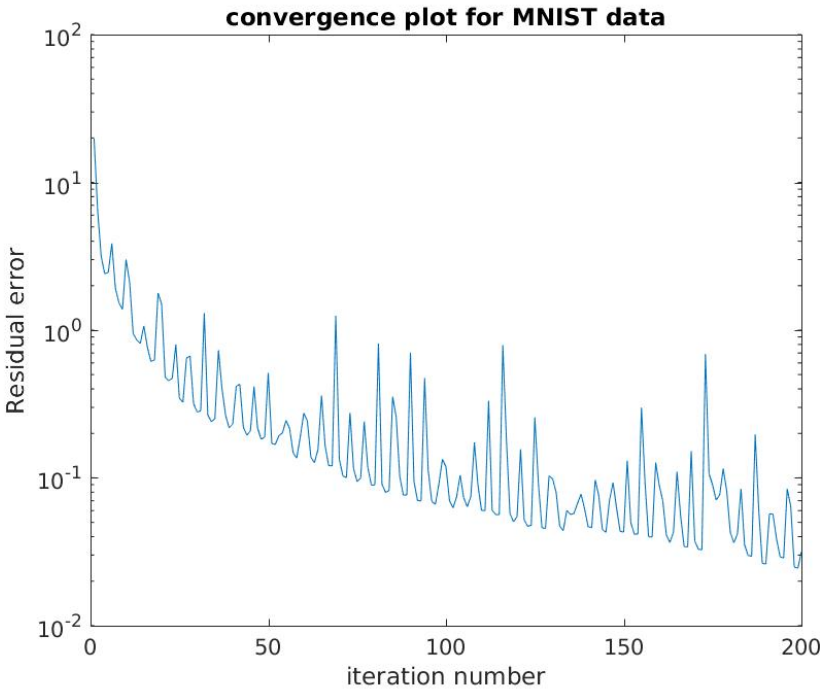

Figure 5: Convergence plot for MNIST data using the Barzelai Borwein Method

As is observed from the plot, we get 'zig-zag' lines that are characteristi of the Borzelai-Borwein method.