Some General Comments

- Most of you have finished a reasonable comparison.
- If you use a subset (e.g., 4 or 6 classes), you should indicate that in your report.
Results of SG I

- Most of you found that SG diverges if the learning rate is too large
- This is right
- Selecting the initial learning rate is a painful issue in using SG
Newton: With/Without LM I

- What I expected was
  - Without LM: faster in terms of iterations
  - With LM: faster in terms of time
- See the earlier comparison on fully-connected networks
- Reasons are
  - Without LM, Newton directions are used, so fewer iterations
With LM, each Newton iteration is cheaper (less CG steps)

But your results show that

With LM: faster on both iterations and time

One explanation is that Newton directions in early iterations are not useful. So LM leads to no advantages on iterations?

Another point I expected is:

With/without LM achieves same final test accuracy
Newton: With/Without LM III

- We can clearly see that from results of fully-connected networks.
- But for CIFAR10, without LM the accuracy is lower.
- This isn’t very good and makes Newton methods less robust.
- Overall the use of LM in Newton for NN is still a research issue.
- More investigation is needed.