In general your reports are now more well written.
We will focus on discussing the accuracy difference.
Test Accuracy I

- Your MATLAB results should be almost the same because you don’t change the code except removing the regularization term.
- The MATLAB results of 10 runs should be around 44.xx%.
- Some report that MATLAB gives exact 44.9% accuracy for all 10 runs.
- This is incorrect. You should have changed the seed.
- We said in the project description: “To handle this situation, let’s do 10 runs under different seeds and check the mean accuracy.”
About the comparison with PyTorch, if your implementation is right, the results should be very close.
In the project description we mentioned an issue about PyTorch’s MSE calculation.

For $\mathbf{v}_1, \mathbf{v}_2 \in \text{batch\_size} \times \#\text{classes}$

they divide the sum by

$\text{batch\_size} \times \#\text{classes}$

Therefore, in our code we need a statement like

$\text{loss} = \text{F.mse\_loss(output, target\_one\_hot)}*10$
If we don’t do that, from the update rule (no regularization)

\[
v \leftarrow \alpha v - \eta \frac{1}{|S|} \nabla_\theta \sum_{i : i \in S} \xi(\theta; y^i, Z^{1,i}) \]

\[
\theta \leftarrow \theta + v
\]

what we actually do is

\[
v \leftarrow \alpha v - \eta \left( \frac{1}{|S| \times 10} \nabla_\theta \sum_{i : i \in S} \xi(\theta; y^i, Z^{1,i}) \right)
\]

\[
\theta \leftarrow \theta + v
\]
Thus results are different

In other words, you need to enlarge the learning rate from 0.003 to 0.03 in order to still get the 44% result

Thank my student Cheng-Hung for experimentally confirming this
Other Issues I

- Now we see almost the same curve of the loss values.
- Before we got that, in fact we worried that two minor issues might cause troubles.
- First, their subset selections at each epoch are different.
- Ours is randomly select 128 elements $\left\lceil \frac{l}{128} \right\rceil$ times.
- Theirs is randomly permute data and split them to $128, 128, \ldots, l - \left\lfloor \frac{l}{128} \right\rfloor \times 128$. 
Thus in ours, in one epoch, the same instance may be selected several times.

But this doesn’t happen in theirs.

The second issue is float versus double.

Results of using float are less accurate.

If we run a lot more epochs, eventually the two curves may become different.