Q1 :

$$-\frac{\partial E}{\partial V_h} = -\frac{\partial E}{\partial y^t} \cdot \frac{\partial y^t}{\partial V_h}.$$

$$= \sum_t \left[r^t \cdot \frac{1}{y^t} + (1-r^t) \frac{-1}{y^t} \right] \cdot (1-{y^t}^2) \cdot Z_h^t$$

$$= \sum_t \frac{(r^t - y^t)\,(1+y^t)}{y^t} \, Z_h^t.$$

$$\Delta V_h = -\eta \frac{\partial E}{\partial V_h}.$$

$$= -\eta \cdot \sum_t \frac{(r^t - y^t)(1+y^t)}{y^t} Z_h^t.$$

$$-\frac{\partial E}{\partial W_{h,j}} = -\frac{\partial E}{\partial y^t} \frac{\partial y^t}{\partial Z_h^t} \cdot \frac{\partial Z_h^t}{\partial W_{h,j}}$$

$$= \sum_t \left\{ \left[r^t \frac{1}{y^t} + (1-r^t) \frac{-1}{1-y^t} \right] (1-{y^t}^2) \cdot V_h \cdot (1-{Z_h^t}^2) x^t \right\} + 2W_{h,j}$$

$$= \sum_t \left\{ \frac{(r^t - y^t)(1+y^t)}{y^t} V_h (1-{Z_h^t}^2) x^t \right\} + 2W_{h,j}.$$

Q2 :

$f^t = P(r^t=1 \mid x^t) = P(r^t=1, k^t=1 \mid x^t) + P(r^t=1, k^t=0 \mid x^t)$

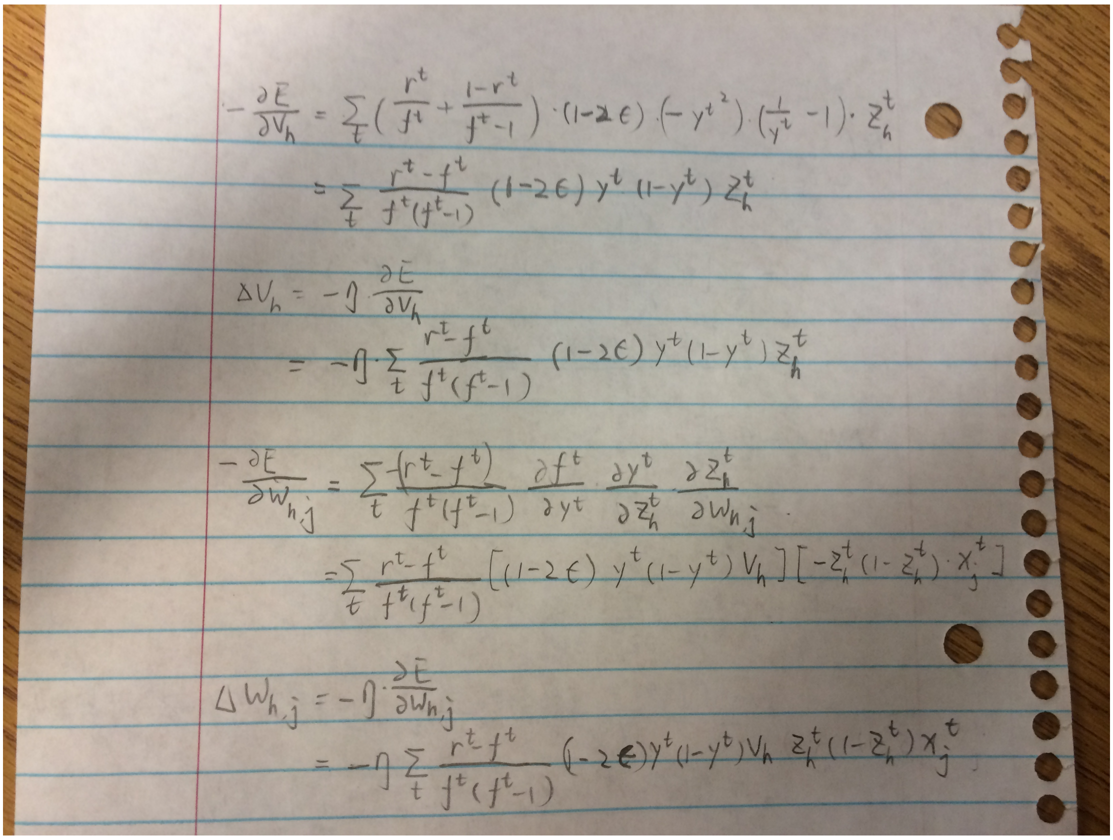
$y^t = P(k^t=1 \mid x^t)$

$\epsilon = P(r^t=1 / k^t=0, x^t) = P(r^t=0 / k^t=1, x^t)$

$\Rightarrow f^t = (1-\epsilon) \cdot y^t + \epsilon(1-y^t)$

$$-\frac{\partial E}{\partial V_h} = -\frac{\partial E}{\partial f^t} \cdot \frac{\partial f^t}{\partial y^t} \cdot \frac{\partial y^t}{\partial V_h}.$$

$$= \sum_t \left(\frac{r^t}{f^t} + \frac{1-r^t}{f^t - 1} \right) \cdot \frac{\partial f^t}{\partial y^t} \cdot \frac{\partial y^t}{\partial V_h}$$



$$-\frac{\partial E}{\partial V_h} = \sum_t \left(\frac{r^t}{f^t} + \frac{1-r^t}{f^t-1} \right) \cdot (1-2\epsilon) \cdot (-{y^t}^2) \cdot \left(\frac{1}{y^t} - 1\right) \cdot z_h^t$$

$$= \sum_t \frac{r^t - f^t}{f^t(f^t-1)} (1-2\epsilon)\, y^t (1-y^t)\, z_h^t$$

$$\Delta V_h = -\eta \cdot \frac{\partial E}{\partial V_h}$$

$$= -\eta \cdot \sum_t \frac{r^t - f^t}{f^t(f^t-1)} (1-2\epsilon)\, y^t (1-y^t)\, z_h^t$$

$$-\frac{\partial E}{\partial w_{hj}} = \sum_t \frac{-(r^t - f^t)}{f^t(f^t-1)} \frac{\partial f^t}{\partial y^t} \cdot \frac{\partial y^t}{\partial z_h^t} \cdot \frac{\partial z_h^t}{\partial w_{hj}}$$

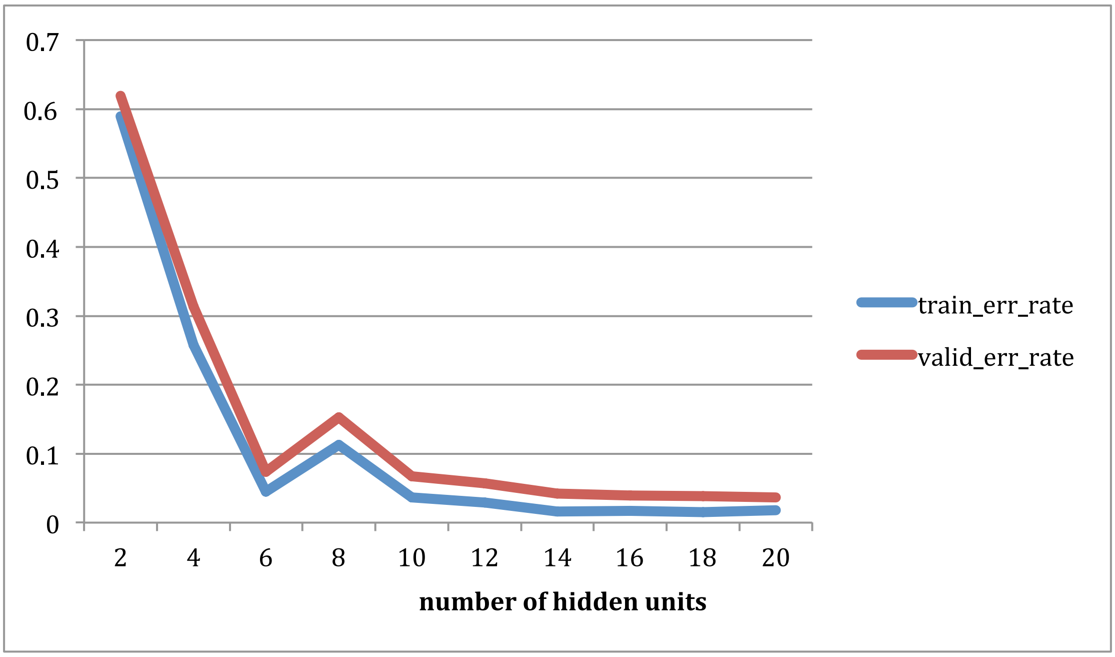
$$= \sum_t \frac{r^t - f^t}{f^t(f^t-1)} \left[(1-2\epsilon)\, y^t (1-y^t)\, V_h\right] \left[-z_h^t (1-z_h^t) \cdot x_j^t\right]$$

$$\Delta w_{h,j} = -\eta \cdot \frac{\partial E}{\partial w_{h,j}}$$

$$= -\eta \sum_t \frac{r^t - f^t}{f^t(f^t-1)} (1-2\epsilon) y^t (1-y^t) V_h\ z_h^t (1-z_h^t) x_j^t$$

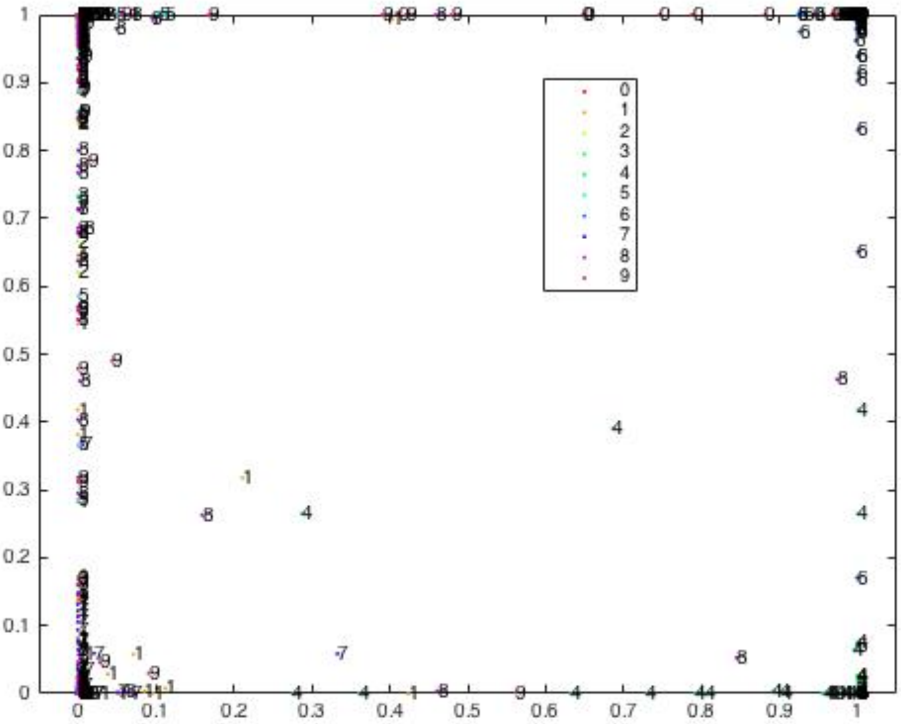
(a)

m	2	4	6	8	10	12	14	16	18	20
train_err_rate	0.590	0.258	0.045	0.113	0.037	0.029	0.0166	0.017	0.015	0.018
val_err_rate	0.619	0.314	0.074	0.153	0.067	0.0566	0.042	0.0395	0.0384	0.0363

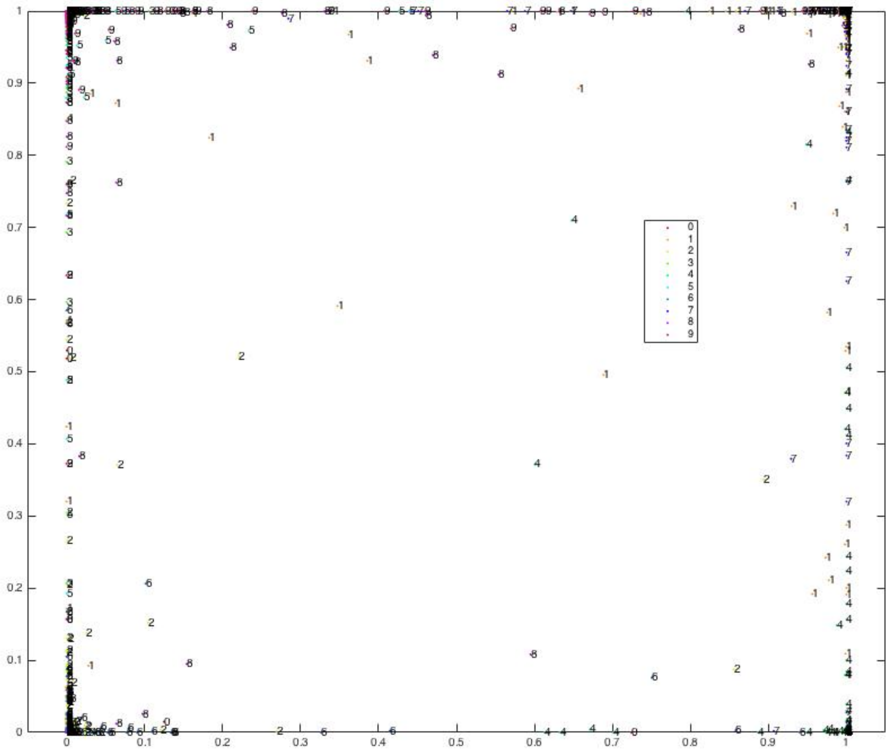


I should use 6 hidden units and using this I get the test_err_rate=0.0870 for test data set.

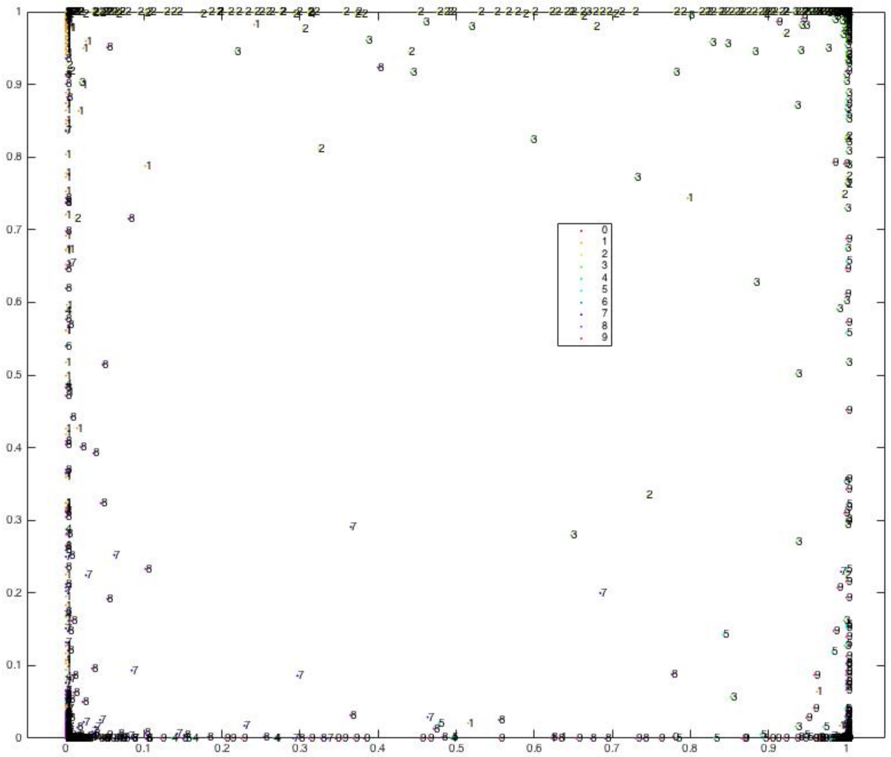
(b)
train:



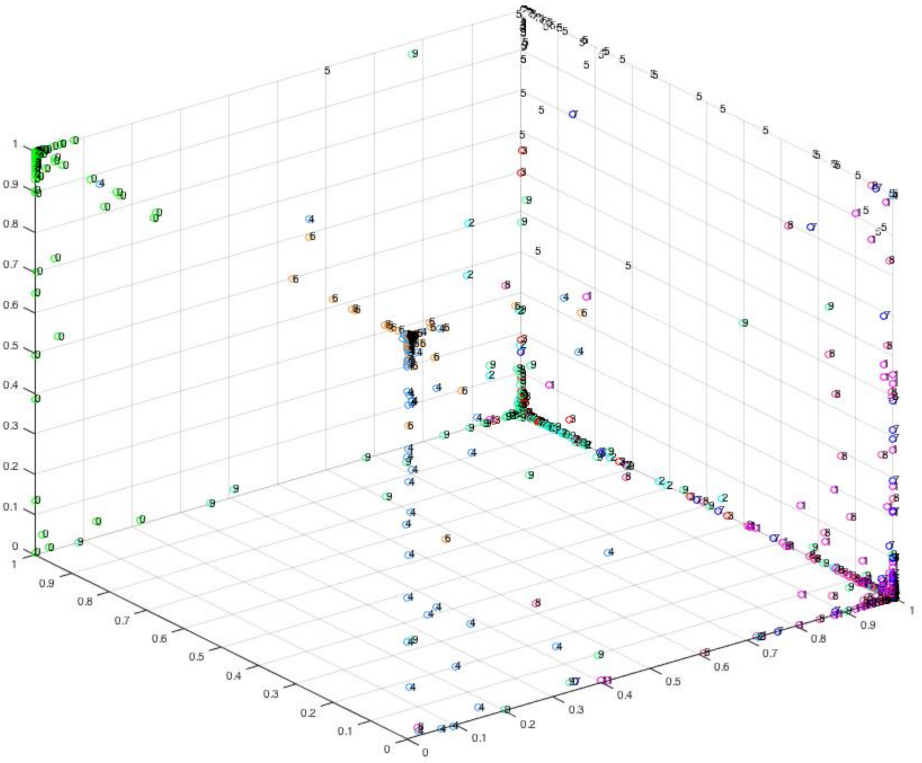
valid:


0
1
2
3
4
5
6
7
8
9
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

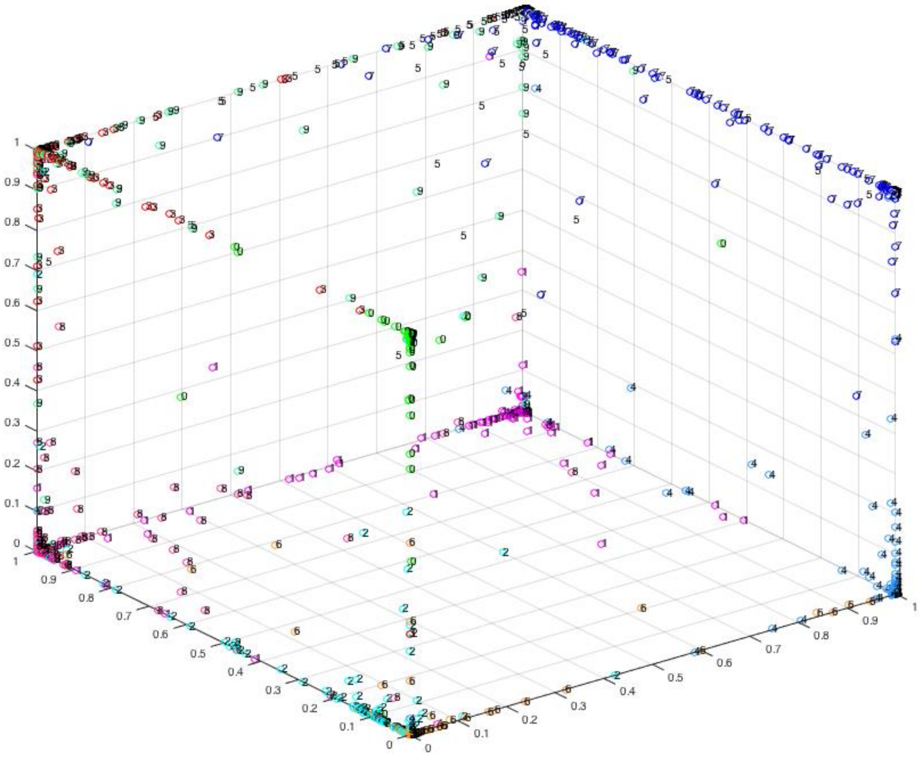
test:


0
1
2
3
4
5
6
7
8
9
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

(c)
train:

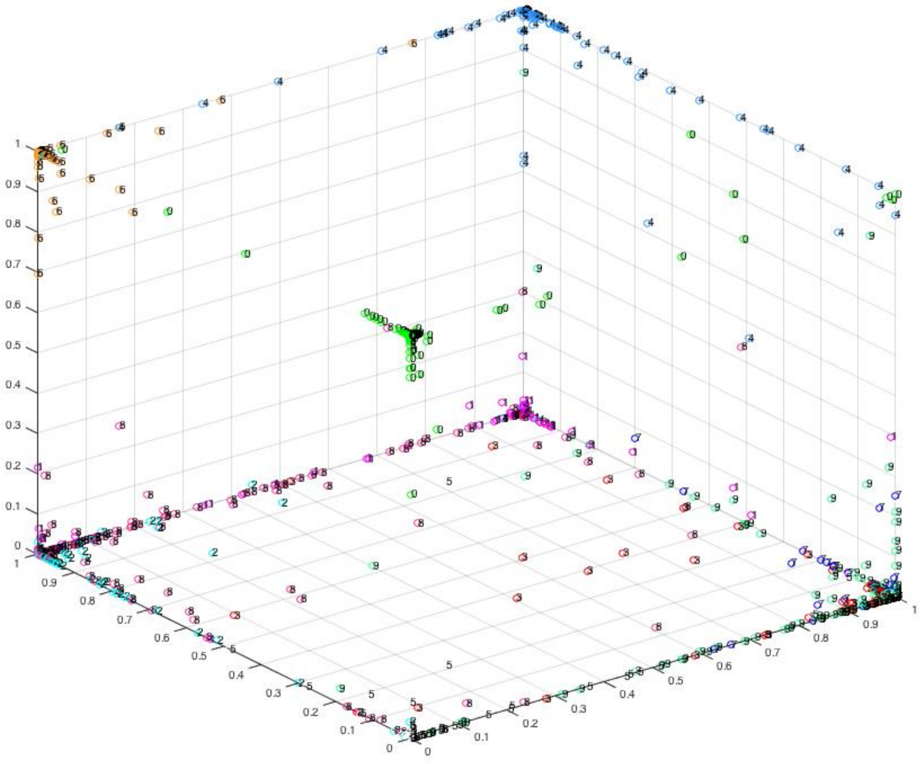


valid:



1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

test:



1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

We find that in 2-D plots, most points gather in four corners or at least allocate along the edges of square. In 3-D plots, most points gather in 8 corners or at least allocate along the edges of cube.

Because functionality of sigmoid function is like step function, output should be close to either zero or one with high probability. [0,0],[0,1],[1,0],[1,1] are the four corners in 2-D plot and points with at least one coordinate equals 0 or 1 should be on the edges. Similar phenomenon appears in 3-D plots.

(d)

structured MLP with overlapping

size of window	train_err_rate	valid_err_rate	test_err_rate
2*2	0.0085	0.0278	0.0331
4*4	0.0128	0.0368	0.0443

When we use smaller sliding window, it implies that we try to capture map characteristics with more hidden units and larger number of weights which means one pixel can contribute to more hidden units through larger number of weights resulting lower error rate.

structured MLP without overlapping

size of window	train_err_rate	valid_err_rate	test_err_rate
2*2	0.0160	0.0448	0.0416
4*4	0.0318	0.0330	0.0354

We can see when we use structured MLP without overlapping, we obtain larger error rate on training data than with overlapping. Because we lose the information of overlapping sliding window and ignore some map characteristics due to less hidden units and weights.