

CS229 Fall 2017

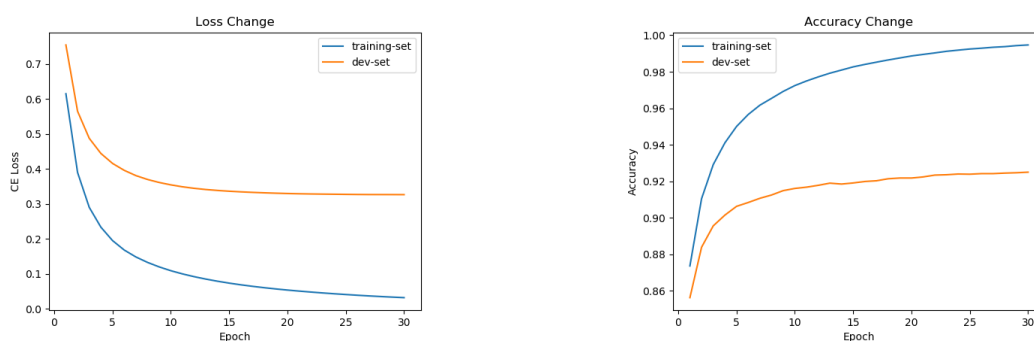
## Problem Set #4 Solutions: EM, DL & RL

Author: LFhase    rimemosa@163.com

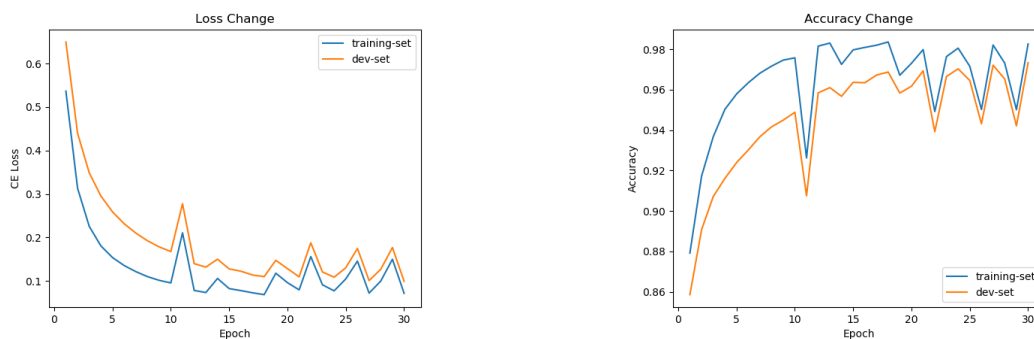
---

### Neural Networks: MNIST image classification

(a) Training the BP neural network version 1 without regularization.



(b) Training the BP neural network version 2 with regularization ( $\lambda = 0.25$ ). From the



result we can see, with regularization of  $W$ , the overfitting is prevented but the training process has more fluctuation.

(c) The accuracy of test set is 0.928700 (without regularization) and 0.972300 (with regularization).

## EM Convergence

Since EM algorithm has converged, the lower bound of  $l(\theta)$  is maximized. Let LB represent the lower bound and we have (drop down some scripts for convenience)

$$\begin{aligned}\nabla_{\theta} LB|_{\theta=\theta^*} &= \sum_i \sum_z Q_z \frac{Q_z}{p(x, z; \theta^*)} \frac{1}{Q_z} \nabla_{\theta} p(x, z; \theta)|_{\theta=\theta^*} \\ &= \sum_i \sum_z \frac{\nabla_{\theta} p(x, z; \theta)|_{\theta=\theta^*}}{p(x; \theta^*)} \\ &= \sum_i \frac{\nabla_{\theta} p(x; \theta)|_{\theta=\theta^*}}{p(x; \theta^*)} \\ &= \sum_i \nabla_{\theta} [\log p(x; \theta)]_{\theta=\theta^*} \\ &= \nabla_{\theta} l(\theta) \\ &= 0\end{aligned}$$

So when EM algorithm has converges, the  $l(\theta)$  acheives the maxima.

## PCA

Since the line between projection and  $x^{(i)}$  is perpendicular with the line between Origin point and projection, minimizing the distance can be written:

$$\operatorname{argmin}_{\mu: \mu^T \mu = 1} \sum_{i=1}^m \|x^{(i)} - x^{(i)T} \mu\|_2$$

Because  $x^{(i)2}$  is a constant, so the formula is equivalent to maximizing the variance.

## Independent components analysis

The W matrix is as below shows:

$$\begin{bmatrix} 72.15081922 & 28.62441682 & 25.91040458 & -17.2322227 & -21.191357 \\ 13.45886116 & 31.94398247 & -4.03003982 & -24.0095722 & 11.89906179 \\ 18.89688784 & -7.80435173 & 28.71469558 & 18.14356811 & -21.17474522 \\ -6.0119837 & -4.15743607 & -1.01692289 & 13.87321073 & -5.26252289 \\ -8.74061186 & 22.55821897 & 9.61289023 & 14.73637074 & 45.28841827 \end{bmatrix}$$

The S can be derived by  $XW^T$ .

# Markov decision processes

(a) gg