1 Shalev 24.2

The situation described is MLE being performed on m samples that have supposedly been picked from a bernoulli distribution basically the regularized version adds one positive example($x_i = 1$) and one negtive example ($x_i = 0$) this is done to prevent overfitting in the case that all the results in the samples are ($x_i = 1$), where normal MLE would have yielded $\theta = 1$. Therefre in-order to avoid this scenario,(where $\theta_{MLE} = 0, 1$) we add two extra terms supporting each class. we already know that the solution for MLE is the fraction of positive samples within the set of samples, with one extra positive sample added.

$$\text{Num of positive samples} = \sum_{i=1}^{i=m} x_i \; + 1$$

$$\text{Total samples} = m + 1 + 1$$

$$\theta_{MLE} = \text{Fraction of 'heads'} = \frac{\sum x_i + 1}{m + 2}$$

2 Murphy 3.5

It is better to deal with the log likelihood in this case, once we multiply all the likelihoods and take the logarithm, we obtain the following expression

$$\frac{\partial}{\partial \lambda} \log\Big(\prod \text{Poi}(x_i|\lambda_i)\Big) = 0 \tag{1}$$

$$\frac{\partial}{\partial \lambda}(-\lambda m + \sum x_i \log \lambda - \sum \log x_i!) = 0 \tag{2}$$

$$\lambda = \frac{\sum x_i}{m} \tag{3}$$

This also makes sense , the poisson is an approximated version of the binomial, the average number of successes, is just the average of the success rate.

3 Murphy 3.8

The uniform distribution will try to keep all the points 'just' inside the rectangular bump, i.e. we must minimize a such that all the points $x_i \in [-a, a]$.(because the probability of a sample being inside is non zero and $\propto 1/a$)

$$\hat{a} = max(|x_i|) \tag{4}$$

$$p(x_{n+1}|\hat{a}) = \frac{1}{2\hat{a}} I(|x_{n+1}| \leq \hat{a})) \tag{5}$$

There are 2 problems with this approach, one is that it does not consider anything that is even slightly outside the maximum absolute value seen in the training data. The model by selection itself has become prone to a large generalization error. Secondly, think about the scenario where we have a single misclassified sample. This misclassified sample with a large value of x will offset the entire training procedure , and $\hat{a} = |x_{misclassified}|$. The model has no flexibility, it only cares about a single training sample

4 Murphy 3.11

Multiplying all the probabilities and taking the log likelihood we get

$$\frac{\partial}{\partial\theta}[mlog(\theta) - \theta(\sum x_i)] = 0 \tag{6}$$

$$\hat{\theta} = \frac{m}{\sum x_i} \tag{7}$$

For the example given

$$\hat{\theta} = \frac{3}{5+6+4} = 0.2 \tag{8}$$

5 Murphy - 4.5

The posterior has been given to us, and we are to construct the joint.

	$y=0$	$y=1$
$x=0$	$\theta_2(1-\theta_1)$	$(1-\theta_2)(1-\theta_1)$
$x=1$	$(1-\theta_2)\theta_1$	$(\theta_1)(\theta_2)$

Samples are i.i.d form the joint. What is to notice is that the distributions $p(y|x=0)$ and $p(y|x=1)$ are tied using Bernoulli's that have complementary parameters. On Multiplication of the joint probabilities from the dataset we get

$$p(D|\theta) = \theta_1^4\theta_2^4(1-\theta_1)^3(1-\theta_2)^3 \tag{9}$$

$$\text{Maximizing we get } \theta_1 = \theta_2 = \frac{4}{7} \tag{10}$$

$$p(D|\theta) = \frac{4^8 3^6}{7^{14}} \tag{11}$$

6 Murphy 10.1

We must write this neatly, **w** is a matrix and x_i is a vector. $\mu_{ik} = \text{Softmax}(w^T x_i)_k$ SInce we are only dealing wit the i'th elements here we can drop the subscript and note that $\eta_{ij} = (w^T x_i)_j$

$$
\begin{align}
\frac{\partial \mu_{ik}}{\partial \eta_{ij}} &= \frac{\partial}{\partial (w^T x_i)_j} \left(\frac{e^{(}w^T x_i)_k}{\sum_l (w^T x_i)_l} \right) \tag{12} \\
& \tag{13} \\
= \frac{I(k=j) e^{(w^T x_i)_k} \sum_l e^{(w^T x_i)_l} - e^{(w^T x_i)_j} e^{(w^T x_i)_k}}{(\sum_l e^{(w^T x_i)_l})^2} & \tag{14} \\
&= \mu_{ik}(I(k=j) - \mu_{ij}) \tag{15}
\end{align}
$$

Calculating the Hessian.

7 Murphy 10.3

LDA models linear parameter scores using MLE on the joint distribution, QDA is more general than LDA , ie QDA contains LDA as a apecial case, so the error that is given (in terms of posterior likelihood over the training set) for QDA will be always better than that of LDA. Similar arguments can be given for QuadLog and LinLog. QuadLog will contain LinLog and hence the log likelihood on the dataset for QuadLog will atleast be as good as that of LinLog.

$$
\begin{align}
L(\text{QuadLog}) &\leq L(\text{LinLog}) \tag{16} \\
L(\text{QuadLog}) &\leq L(\text{LDA(GaussI)}) \tag{17} \\
L(\text{QDA(GuassX)}) &\leq ? \geq L(\text{QuadLog}) \tag{18} \\
& \tag{19}
\end{align}
$$