HOMEWORK ASSIGNMENT #2
B555- MACHINE LEARNING

FNU ANIRUDH
(aanirudh@umail.iu.edu)

Question 1

Solution:-

This situation can be modeled as having data set D = $\{ x_i \}^{9}_{i=1}$, where each x_i = number of accidents in the plant in day i, for $1 \leq i \leq 9$.

We know that $\sum_{i=1}^{9} x_i = 79$

a) **Maximum Likelihood is given by**

$$\lambda_{ML} = \text{argmax}_{\lambda} \{p(D|\lambda)\}$$

Where probability of single observation x_i given λ is

$$p(x_i|\lambda) = \frac{\lambda^{x_i}\, e^{-\lambda}}{x_i\,!}$$

Since we assume a poisson distribution for number of accidents. From this it follows that the likelihood for a general data set D with n observation is

$p(D|\lambda) = \prod_{i=1}^{n} p(xi|\lambda)$ by independence of x_i

$$= \prod_{i=1}^{n} \frac{\lambda^{xi}\, e^{-\lambda}}{x_i\,!}$$ by assumption of poisson distribution

$$= \frac{\lambda \sum_{i=1}^{n} x_i\, e^{-n\lambda}}{\prod_{i=1}^{n} xi\,!}$$

This shows that the likelihood function is $l(\lambda) = \dfrac{\lambda \sum_{i=1}^{n} x_i\, e^{-n\lambda}}{\prod_{i=1}^{n} xi\,!}$

To find λ that maximizes the likelihood, we will first take a logarithm to simplify the calculation, then find its first derivative with respect to λ, and finally equate it with zero to find the maximum. Specifically, we express the log-likelihood $ɭɭ(D,\lambda) = \ln p(D|\lambda)$ as

$$ɭɭ(D, \lambda) = \ln \lambda \sum_{i=1}^{n} x_i - n\lambda - \sum_{i=1}^{n} \ln(x_i!)$$

and proceed with the first derivative as

$$\frac{d\,ɭɭ(D, \lambda)}{d\lambda} = 1/\lambda \sum_{i=1}^{n} x_i - n$$
$$= 0$$

Hence we can say that $n = 1/\lambda \sum_{i=1}^{n} x_i$

Or

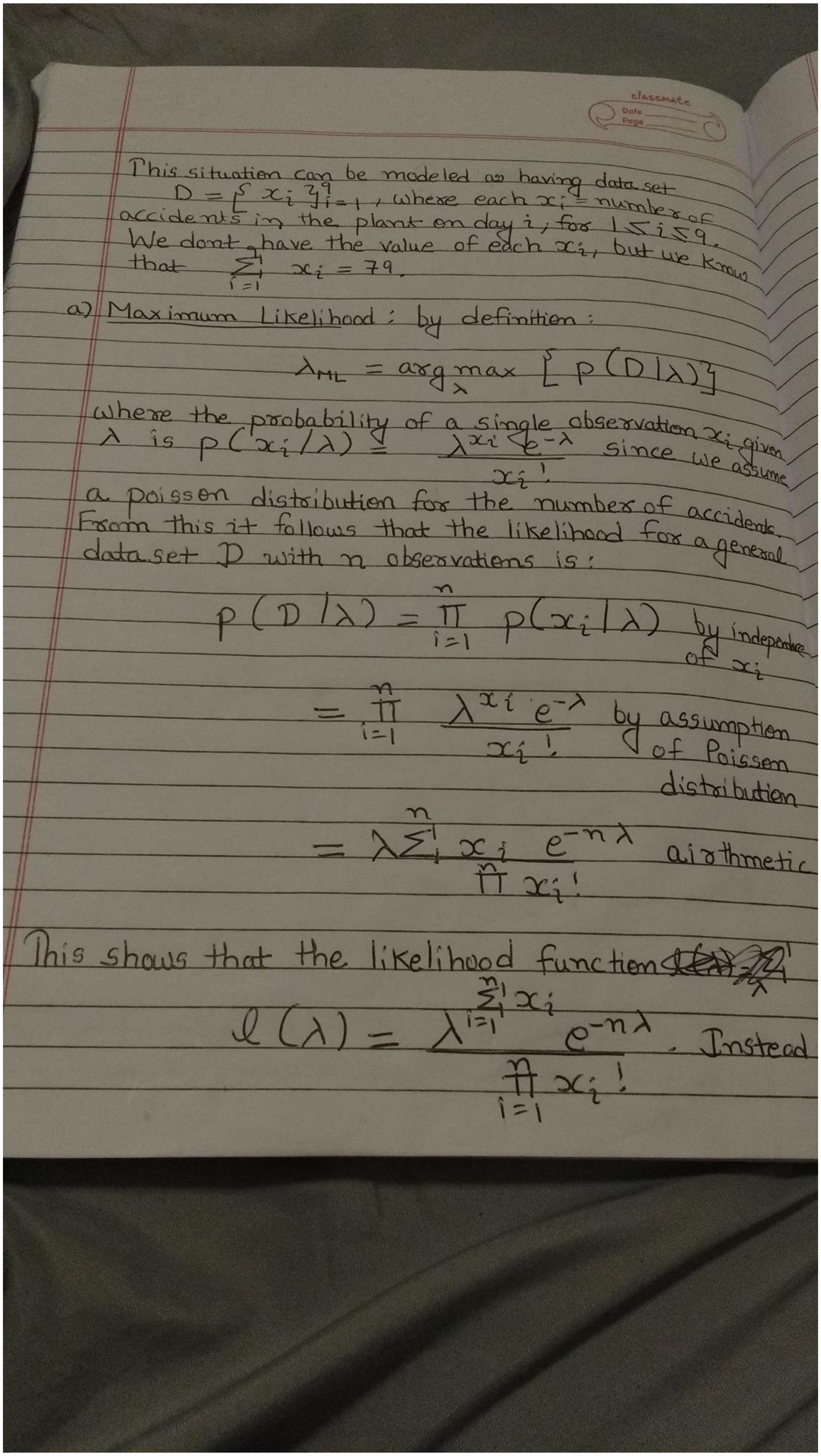
$$\lambda_{ML} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Since $x_i \in N$ for all i and $\lambda^2 > 0$. Hence we ignore the degenerate case where $x_i = 0$ for all i

Given that n=9 and $\sum_{i=1}^{9} x_i = 79$,

Hence,

$$\lambda_{ML} = \frac{79}{9} = 8.7777$$



This situation can be modeled as having data set $D = \{x_i\}_{i=1}^{9}$, where each x_i = number of accidents in the plant on day i; for $1 \leq i \leq 9$. We dont have the value of each x_i, but we know that $\sum_{i=1}^{9} x_i = 79$.

a) Maximum Likelihood : by definition :

$$\lambda_{ML} = \arg\max_{\lambda} \{p(D|\lambda)\}$$

where the probability of a single observation x_i given λ is $p(x_i/\lambda) \overset{\Delta}{=} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!}$ since we assume a poisson distribution for the number of accidents. From this it follows that the likelihood for a general data set D with n observations is :

$$p(D/\lambda) = \prod_{i=1}^{n} p(x_i|\lambda) \quad \text{by independence of } x_i$$

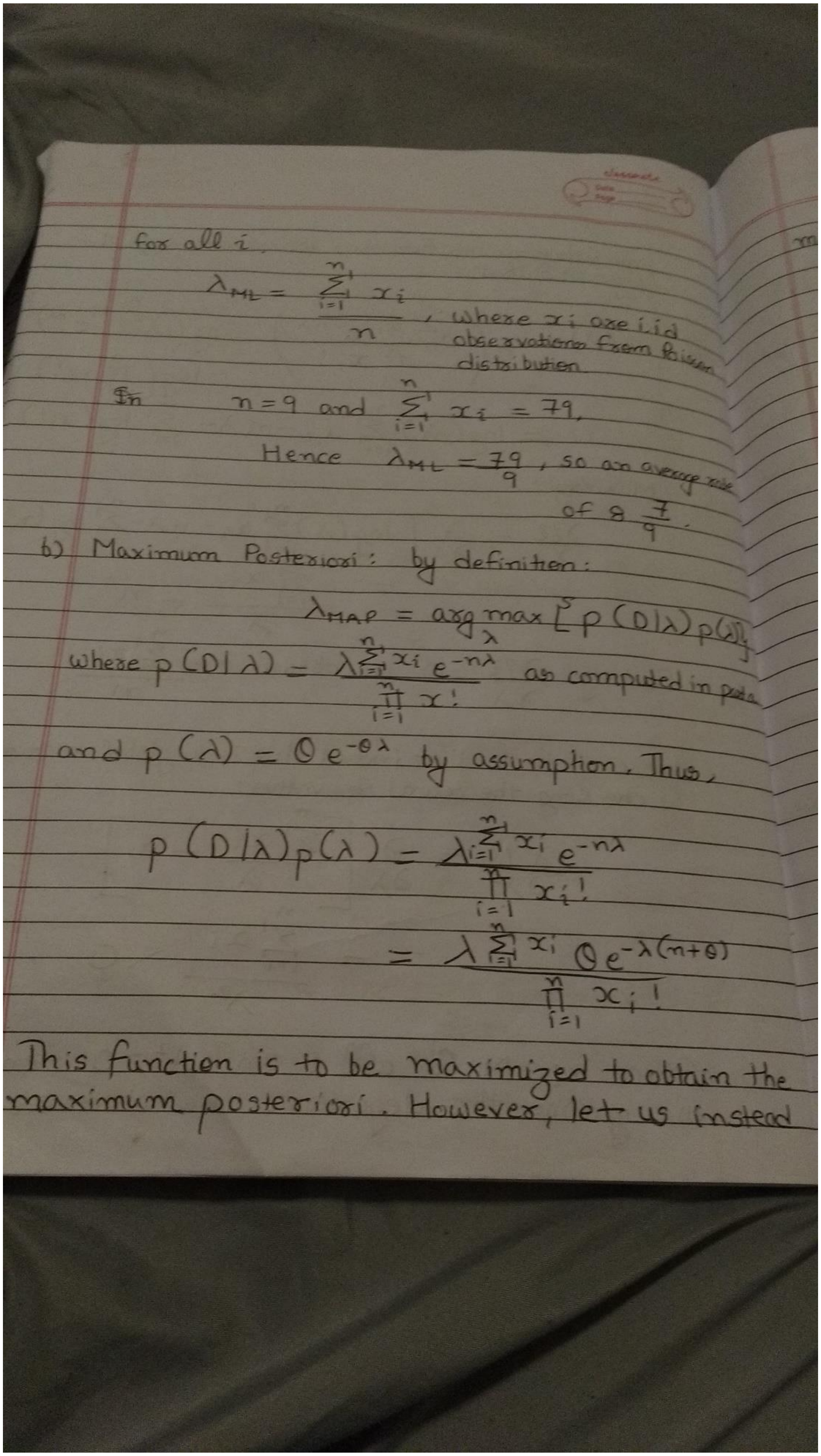
$$= \prod_{i=1}^{n} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!} \quad \text{by assumption of Poisson distribution}$$

$$= \frac{\lambda^{\sum_1^n x_i} e^{-n\lambda}}{\prod^n x_i!} \quad \text{airthmetic}$$

This shows that the likelihood function

$$l(\lambda) = \frac{\lambda^{\sum_{i=1}^{n} x_i} e^{-n\lambda}}{\prod_{i=1}^{n} x_i!}$$

. Instead



for all i,

$$\lambda_{ML} = \frac{\sum_{i=1}^{n} x_i}{n}$$, where x_i are i.i.d observations from Poisson distribution

In $n = 9$ and $\sum_{i=1}^{n} x_i = 79$.

Hence $\lambda_{ML} = \frac{79}{9}$, so an average rate of $8\frac{7}{9}$.

b) Maximum Posteriori: by definition:

$$\lambda_{MAP} = \arg\max_{\lambda} \left[p(D|\lambda)\, p(\lambda) \right]$$

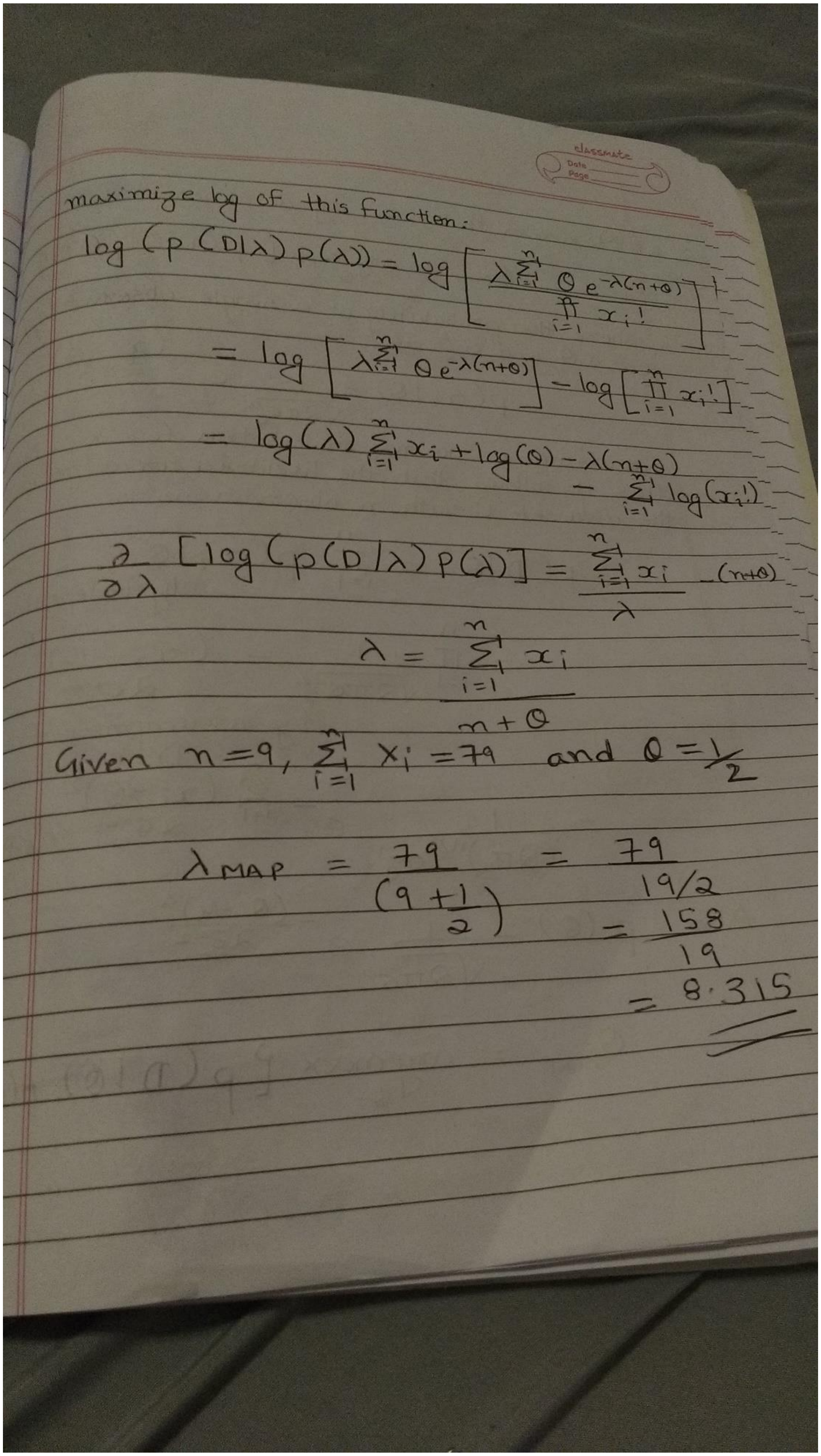
where $p(D|\lambda) = \frac{\lambda^{\sum_{i=1}^{n} x_i} e^{-n\lambda}}{\prod_{i=1}^{n} x!}$ as computed in part a

and $p(\lambda) = \theta e^{-\theta\lambda}$ by assumption. Thus,

$$p(D/\lambda)\, p(\lambda) = \frac{\lambda^{\sum_{i=1}^{n} x_i} e^{-n\lambda}}{\prod_{i=1}^{n} x_i!}$$

$$= \frac{\lambda^{\sum_{i=1}^{n} x_i}\, \theta e^{-\lambda(n+\theta)}}{\prod_{i=1}^{n} x_i!}$$

This function is to be maximized to obtain the maximum posteriori. However, let us instead



maximize log of this function:

$$\log(p(D|\lambda)p(\lambda)) = \log\left[\frac{\lambda^{\sum_{i=1}^{n}}\theta e^{-\lambda(n+\theta)}}{\prod_{i=1}^{n} x_i!}\right]$$

$$= \log\left[\lambda^{\sum_{i=1}^{n}}\theta e^{-\lambda(n+\theta)}\right] - \log\left[\prod_{i=1}^{n} x_i!\right]$$

$$= \log(\lambda)\sum_{i=1}^{n} x_i + \log(\theta) - \lambda(n+\theta) - \sum_{i=1}^{n}\log(x_i!)$$

$$\frac{\partial}{\partial\lambda}[\log(p(D|\lambda)p(\lambda)] = \frac{\sum_{i=1}^{n} x_i}{\lambda} - (n+\theta)$$

$$\lambda = \frac{\sum_{i=1}^{n} x_i}{n+\theta}$$

Given $n=9$, $\sum_{i=1}^{n} x_i = 79$ and $\theta = 1/2$

$$\lambda_{MAP} = \frac{79}{(9+\frac{1}{2})} = \frac{79}{19/2} = \frac{158}{19} = 8.315$$

b)

Maximum Posteriori by definition is

$\lambda_{MAP} = \text{argmax}_{\lambda} \{p(D|\lambda)\ p(\lambda)\}$

Where

$$p(D|\lambda) \quad = \quad \frac{\lambda \sum_{i=1}^{n} x_i\, e^{-n\lambda}}{\prod_{i=1}^{n} xi\,!} \qquad \text{from part a}$$

And $p(\lambda) = \Theta\ e^{-\Theta\lambda}$

$$p(D|\lambda)\ p(\lambda) = \frac{\lambda \sum_{i=1}^{n} x_i\, e^{-n\lambda}}{\prod_{i=1}^{n} xi\,!}$$

$$= \frac{\lambda \sum_{i=1}^{n} x_i\ \Theta\ e^{-\lambda\,(n+\Theta)}}{\prod_{i=1}^{n} xi\,!}$$

We can maximize this function by taking logarithm

$$\ln \{p(D|\lambda)\ p(\lambda)\} = \ln \frac{[\lambda \sum_{i=1}^{n} x_i\ \Theta\ e^{-\lambda\,(n+\Theta)}]}{\prod_{i=1}^{n} xi\,!}$$

$= \ln\,[\lambda \sum_{i=1}^{n} x_i\ \Theta\ e^{-\lambda\,(n+\Theta)}] - \ln\,[\prod_{i=1}^{n} xi\,!\,]$

$= \ln\,(\lambda\,) \sum_{i=1}^{n} x_i + \ln\,(\Theta) - \lambda\,(n+\Theta) - \sum_{i=1}^{n} \ln\,(x_i\,!)$

Differentiating and equating to zero we get

$$\frac{d \ln\,[p(D|\lambda)\ p(\lambda)]}{d\lambda} = \frac{\sum_{i=1}^{n} x_i}{\lambda} - (n+\Theta) \quad = 0$$

$$\lambda = \frac{\sum_{i=1}^{n} x_i}{n+\Theta}$$

Since $x_i \in N$ for all i and $\lambda^2 > 0$. Hence we ignore the degenerate case where $x_i = 0$ for all i.

$$\lambda_{MAP} = \frac{\sum_{i=1}^{n} x_i}{n+\Theta}$$

given that n=9 and $\sum_{i=1}^{n} x_i = 79$ and $\Theta = ½$

Hence

$$\lambda_{MAP} = \frac{79}{(9+1/2)} = \frac{79 \text{ x } 2}{19} = \frac{158}{19}$$

$$= 8.315$$

c) We know that

$$\lambda = \frac{\sum_{i=1}^{n} x_i}{n+\Theta}$$

$\lambda (n + \Theta) = \sum_{i=1}^{n} x_i$

Given that $\sum_{i=1}^{n} x_i < 4n$

$n + \Theta < 4n$

Or

$\Theta < 4 - \lambda$

2.

a) Maximum Posteriori:

$\Theta_{MAP} = \text{argmax}_{\Theta} \{p(D|\Theta)\ p(\Theta)\}$

Where the probability of a single observation x_i given Θ and σ_0^2 is

$$P(x_i\,|\Theta) = \frac{1}{\sqrt{2\pi\,\sigma_0^2}}\; e^{-(x_i - \Theta)\,2/2}\,\sigma_0^2$$

Since we assume a normal distribution for $X_{i,}$ From this it follows that the likelihood for a general data set D with n observation is

$$p(D|\Theta) = \prod_{i=1}^{n} \mathrm{p}(\mathrm{xi}|\Theta)$$ by independence of x_i

$$= {}^{n}\Pi_{i=1}\; \frac{1}{\sqrt{2\pi\,\sigma_0^2}}\; e^{-(x_i - \Theta)\,2/2}\sigma_0^2$$ (By Assumption of normal distribution)

$$= \frac{1}{(2\pi)^{n/2}\,\sigma_0^{n}} \qquad e^{-}\,({}^{n}\Sigma_{\,i=1}\,{}^{(x_i - \Theta)\,2/2}\,\sigma_0^2\,)$$

Also $$p(\Theta) = \frac{1}{\sqrt{2\pi\sigma}}\; e^{-(\Theta - M)\,2/2\,\sigma 2}$$

$$\Theta_{MAP} = \text{argmax}_{\Theta}\,\{p(D|\Theta)\;p(\Theta)\}$$

$$= \quad \text{argmax}_{\Theta} \quad \frac{1}{(2\pi)^{n/2}\,\sigma_0^{n}}\; e^{-}\,({}^{n}\Sigma_{\,i=1}\,{}^{(x_i - \Theta)\,2/2}\,\sigma_0^2\,) \qquad \frac{1}{\sqrt{2\pi\sigma}}\; e^{-(\Theta - M)\,2/2\,\sigma 2}$$

$$= \quad \text{argmax}_{\Theta} \quad \frac{1}{(2\pi)^{(n+1)/2}\,\sigma_0^{n}\,\sigma}\; e^{-}\,({}^{n}\Sigma_{\,i=1}\,{}^{(x_i - \Theta)\,2/2}\,\sigma_0^2\,) - {}^{(\Theta - M)2/2\,\sigma 2}$$

As usual, let us take the log

$$\ln\,(p(D|\Theta)\;p(\Theta)) = \ln\,(1/\,(2\pi)^{(n+1)/2}\,\sigma_0^{n}\,\sigma) - {}^{n}\Sigma_{\,i=1}\,{}^{(x_i - \Theta)\,2/2}\,\sigma_0^2 - {}^{(\Theta - M)2/2\,\sigma 2}$$

Maximize by differentiating

$$\frac{d}{d\Theta} [\ln (p(D|\Theta)\, p(\Theta)] = \frac{d}{d\Theta} [\ln (1/ (2\pi)^{(n+1)/2} \sigma_0^n \sigma) - {}^n\sum_{i=1} {}^{(x_i - \Theta)\, 2\, /\, 2} \sigma_0^2 - {}^{(\Theta - \mu) 2\, /\, 2\, \sigma 2}]$$

$$= \frac{{}^n\sum_{i=1} (x_i - \Theta)}{\sigma_0^2} - \frac{(\Theta - \mu)}{\sigma^2}$$

$$= \frac{({}^n\sum_{i=1} x_i) - n\Theta}{\sigma_0^2} - \frac{\Theta - \mu}{\sigma^2}$$

$$= \frac{\sigma^2 ({}^n\sum_{i=1} x_i) - \sigma^2 n\Theta - \sigma_0^2\, \Theta - \sigma_0^2\, \mu}{\sigma^2\, \sigma_0^2}$$

Equating Left hand side to zero, we get

$$0 = \sigma^2 ({}^n\sum_{i=1} x_i) - \sigma^2 n\Theta - \sigma_0^2\, \Theta - \sigma_0^2\, \mu$$

$$0 = -\Theta (\sigma^2 n + \sigma_0^2) + \sigma^2 ({}^n\sum_{i=1} x_i) + \sigma_0^2\, \mu$$

$$\Theta = \frac{\sigma^2 ({}^n\sum_{i=1} x_i) + \sigma_0^2\, \mu}{\sigma^2 n + \sigma_0^2}$$

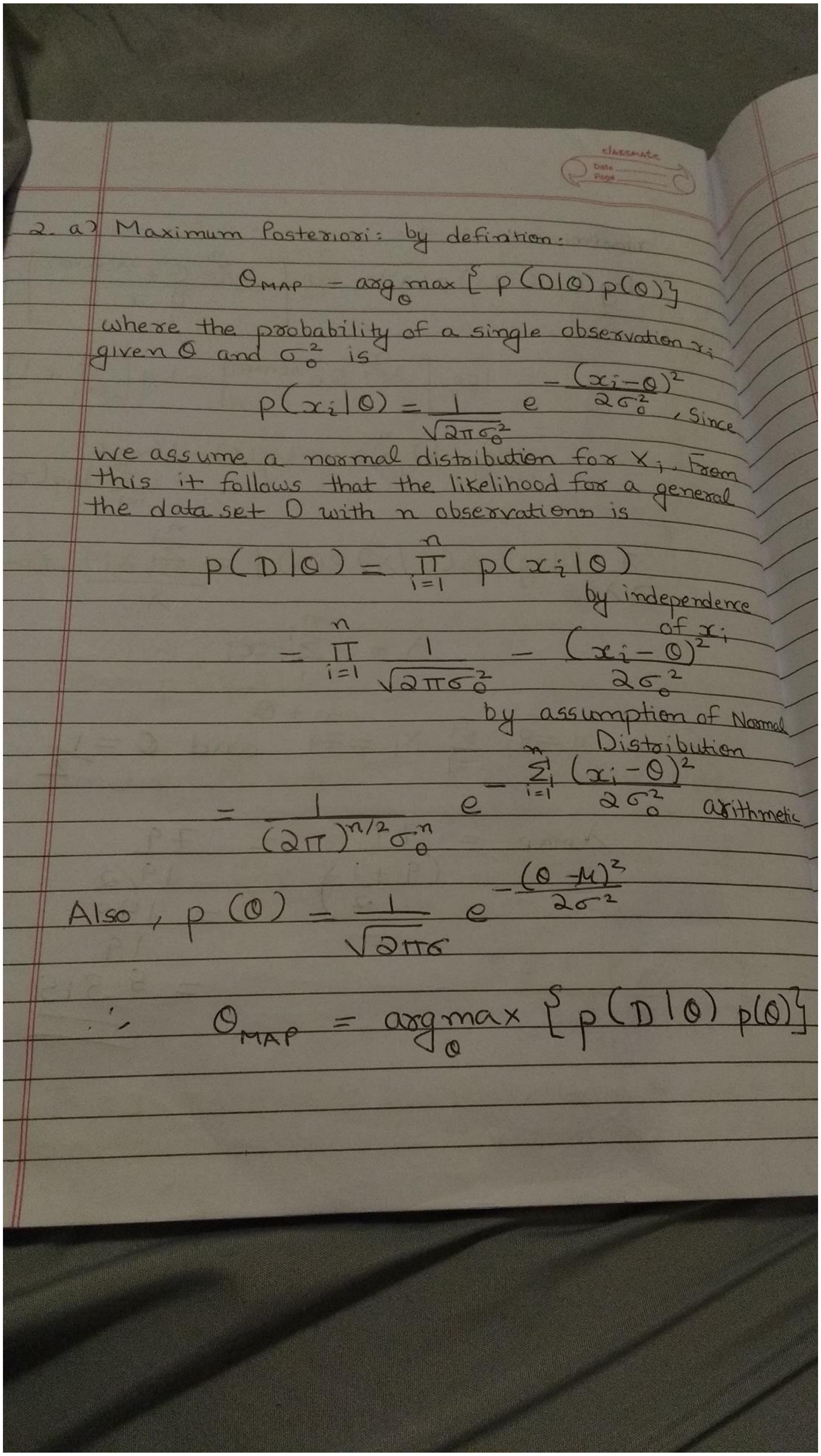
$$\frac{d^2}{d\Theta^2} [\ln p(D|\Theta)\, p(\Theta))] = \frac{d}{d\Theta} \left[\frac{({}^n\sum_{i=1} x_i) - n\Theta}{\sigma_0^2} - \frac{\Theta - \mu}{\sigma^2} \right]$$

$$= - \frac{n}{\sigma_0^2} - \frac{1}{\sigma^2} < 0$$

Since $n>0$ and $\sigma_0^2\ \sigma^2 > 0$

$$\Theta_{MAP} = \frac{\sigma^2 \left({}^{n}\sum_{i=1} x_i \right) - \sigma_0^2\, \mu}{\sigma^2 n - \sigma_0^2}$$

$$\Theta_{MAP} = \frac{\dfrac{\mu}{\sigma^2} - \dfrac{{}^{n}\sum_{i=1} x_i}{\sigma_0^2}}{\dfrac{1}{\sigma^2} - \dfrac{n}{\sigma_0^2}}$$



2. a) Maximum Posteriori: by definition:

$$\theta_{MAP} = \arg\max_{\theta} \{ p(D|\theta) p(\theta) \}$$

where the probability of a single observation x_i given θ and σ_0^2 is

$$p(x_i|\theta) = \frac{1}{\sqrt{2\pi\sigma_0^2}} e^{-\frac{(x_i-\theta)^2}{2\sigma_0^2}}$$

, Since we assume a normal distribution for X_i. From this it follows that the likelihood for a general the data set D with n observations is

$$p(D|\theta) = \prod_{i=1}^{n} p(x_i|\theta)$$

by independence of x_i

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_0^2}} - \frac{(x_i-\theta)^2}{2\sigma_0^2}$$

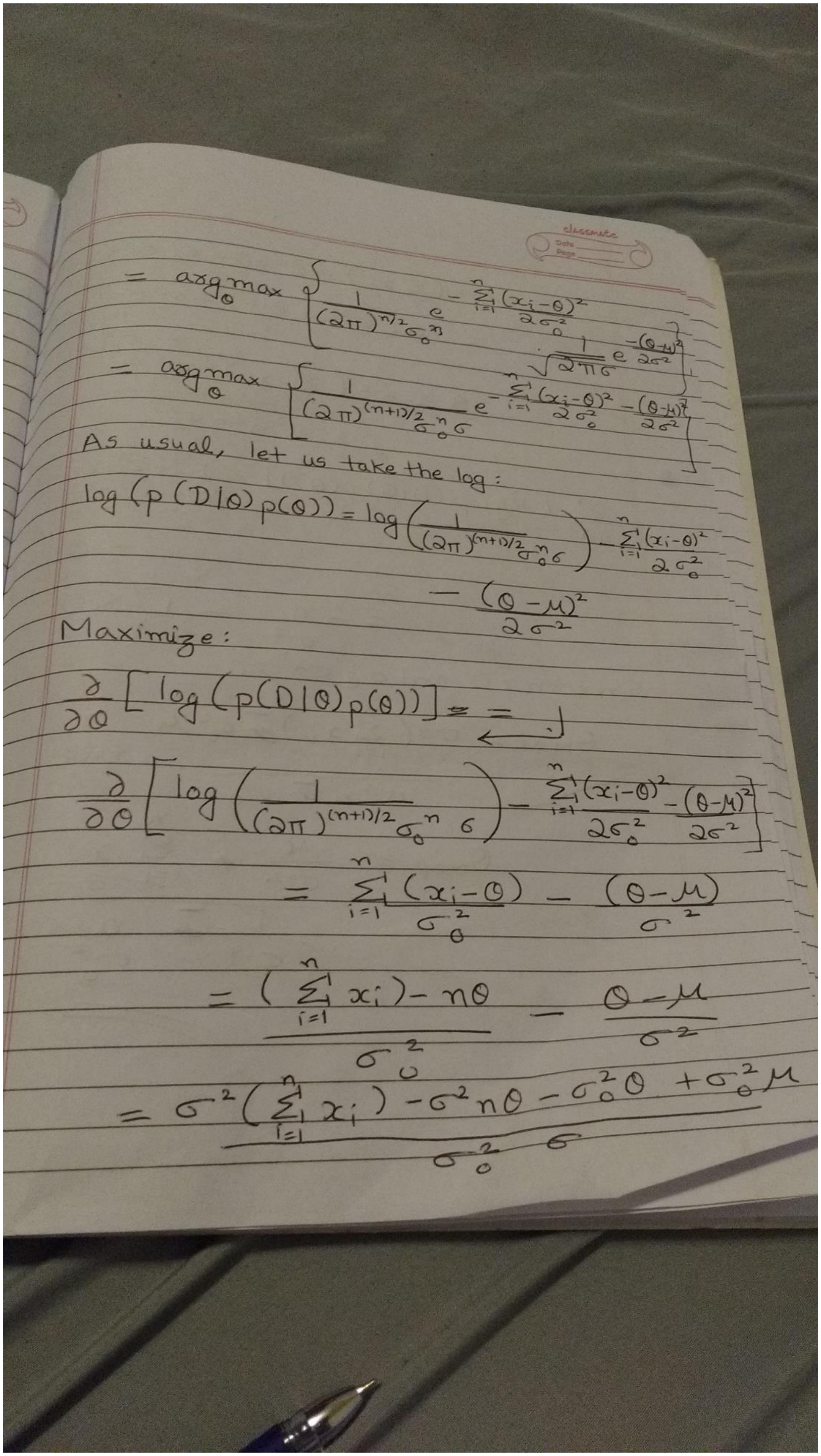
by assumption of Normal Distribution

$$= \frac{1}{(2\pi)^{n/2}\sigma_0^n} e^{-\frac{\sum_{i=1}^{n}(x_i-\theta)^2}{2\sigma_0^2}}$$

arithmetic

Also, $p(\theta) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(\theta-\mu)^2}{2\sigma^2}}$

$$\therefore \quad \theta_{MAP} = \arg\max_{\theta} \{ p(D|\theta) p(\theta) \}$$



$$= \underset{\theta}{\arg\max} \left\{ \frac{1}{(2\pi)^{n/2}\sigma_0^n} e^{-\sum_{i=1}^{n}\frac{(x_i-\theta)^2}{2\sigma_0^2}} \cdot \frac{1}{\sqrt{2\pi\sigma}} e^{\frac{-(\theta-\mu)^2}{2\sigma^2}} \right\}$$

$$= \underset{\theta}{\arg\max} \left[\frac{1}{(2\pi)^{(n+1)/2}\sigma_0^n \sigma} e^{-\sum_{i=1}^{n}\frac{(x_i-\theta)^2}{2\sigma_0^2} - \frac{(\theta-\mu)^2}{2\sigma^2}} \right]$$

As usual, let us take the log:

$$\log(p(D|\theta)p(\theta)) = \log\left(\frac{1}{(2\pi)^{(n+1)/2}\sigma_0^n \sigma}\right) - \sum_{i=1}^{n}\frac{(x_i-\theta)^2}{2\sigma_0^2} - \frac{(\theta-\mu)^2}{2\sigma^2}$$

Maximize:

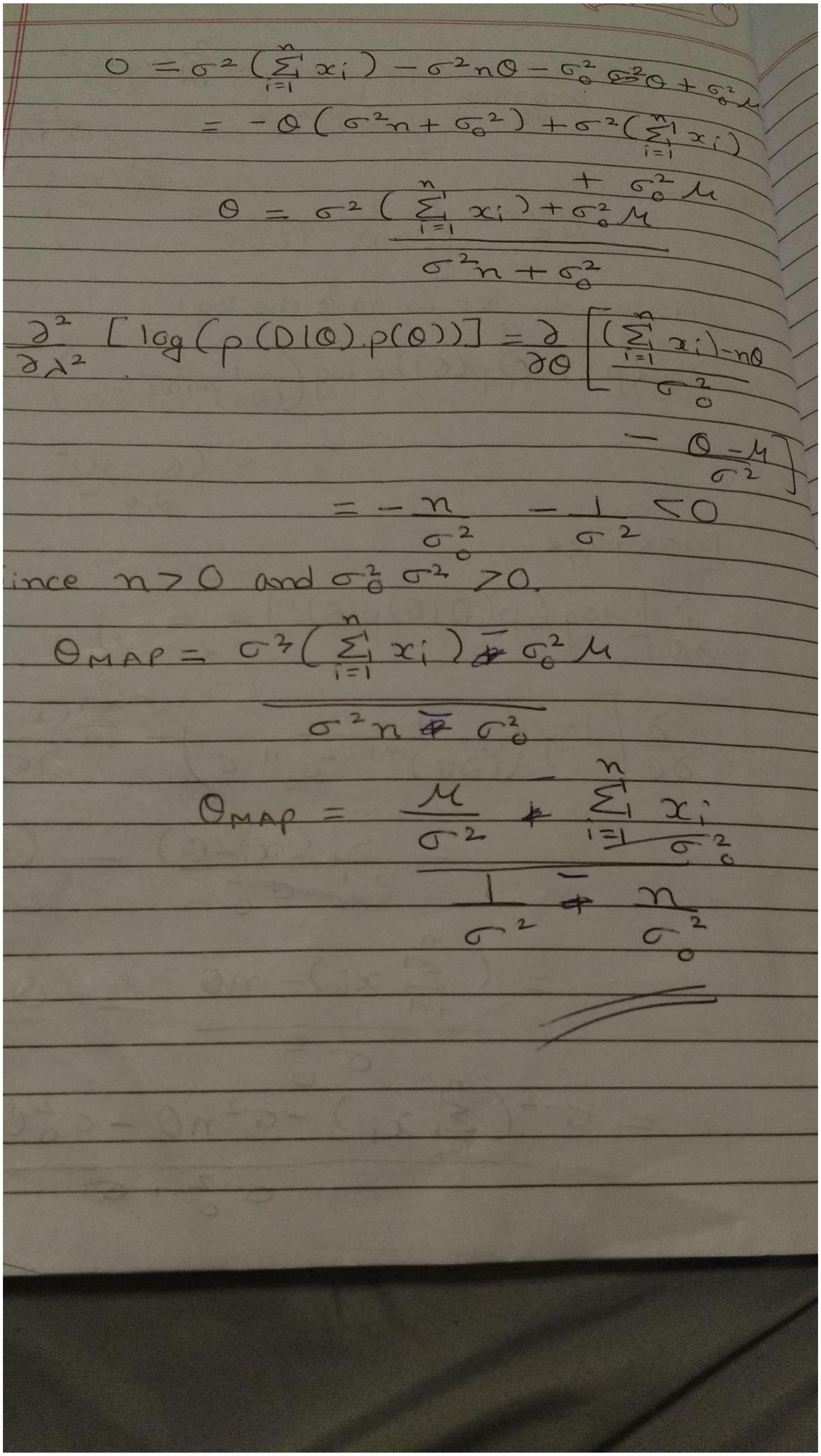
$$\frac{\partial}{\partial\theta}\left[\log(p(D|\theta)p(\theta))\right] = $$

$$\frac{\partial}{\partial\theta}\left[\log\left(\frac{1}{(2\pi)^{(n+1)/2}\sigma_0^n \sigma}\right) - \sum_{i=1}^{n}\frac{(x_i-\theta)^2}{2\sigma_0^2} - \frac{(\theta-\mu)^2}{2\sigma^2}\right]$$

$$= \sum_{i=1}^{n}\frac{(x_i-\theta)}{\sigma_\theta^2} - \frac{(\theta-\mu)}{\sigma^2}$$

$$= \frac{\left(\sum_{i=1}^{n} x_i\right) - n\theta}{\sigma_0^2} - \frac{\theta-\mu}{\sigma^2}$$

$$= \frac{\sigma^2\left(\sum_{i=1}^{n} x_i\right) - \sigma^2 n\theta - \sigma_0^2\theta + \sigma_0^2\mu}{\sigma_0^2 \sigma}$$



$$0 = \sigma^2 \left(\sum_{i=1}^{n} x_i\right) - \sigma^2 n\theta - \sigma_0^2 \cancel{\sigma^2}\theta + \sigma_0^2 \mu$$

$$= -\theta(\sigma^2 n + \sigma_0^2) + \sigma^2\left(\sum_{i=1}^{n} x_i\right) + \sigma_0^2 \mu$$

$$\theta = \frac{\sigma^2 \left(\sum_{i=1}^{n} x_i\right) + \sigma_0^2 \mu}{\sigma^2 n + \sigma_0^2}$$

$$\frac{\partial^2}{\partial \lambda^2}\left[\log(p(D|\theta).p(\theta))\right] = \frac{\partial}{\partial\theta}\left[\frac{\left(\sum_{i=1}^{n} x_i\right) - n\theta}{\sigma_0^2} - \frac{\theta - \mu}{\sigma^2}\right]$$

$$= -\frac{n}{\sigma_0^2} - \frac{1}{\sigma^2} < 0$$

ince $n > 0$ and σ_0^2 $\sigma^2 > 0$.

$$\theta_{MAP} = \frac{\sigma^3\left(\sum_{i=1}^{n} x_i\right) + \sigma_0^2 \mu}{\sigma^2 n + \sigma_0^2}$$

$$\theta_{MAP} = \frac{\frac{\mu}{\sigma^2} + \frac{\sum_{i=1}^{n} x_i}{\sigma_0^2}}{\frac{1}{\sigma^2} + \frac{n}{\sigma_0^2}}$$

b)

$$p(x)= \frac{1}{2b} \exp(- |x - \text{м} |/b)$$

We know that

$$\Theta_{MAP} = \text{argmax}_{\Theta} \{p(D|\Theta)\, p(\Theta)\}$$

$$P(\Theta)= \frac{1}{2b} \exp(- |\Theta - \text{м} | /b)$$

solving for MAP, by taking log and maximizing we get

$$\frac{\sum_{i=1}^{n}(x_i - \theta)}{\sigma_0^2} = -\frac{(\theta \ - \ \mu)}{|\theta \ - \ \mu|b}$$

Since μ = 0

$$\sum_{i=1}^{n}(x_i) - n\theta = (\frac{(sign\ of\ \theta)\sigma_0^2}{b})$$

$$n\theta = \left(\frac{sign\ of\ \theta}{b}\right)\sigma_0^2 + \sum_{i=1}^{n}(x_i)$$

$$\theta = \frac{\left(\frac{sign\ of\ \theta}{b}\right)\sigma_0^2 + \sum_{i=1}^{n}(x_i)}{n}$$

As Θ cannot be definite, we cannot have closed form solution hence we can solve this by using gradient descent.

c)

This is similar to part a) of the question except need to do in matrix, hence

$$\frac{\text{ɗ}}{\text{ɗ}\Theta} = \frac{-1}{2} \frac{\text{ɗ}}{\text{ɗ}\Theta} [\ \sigma^2\ \Theta^T\Theta + \sum (\ X^T X - X^T\Theta) - (\Theta^T X + \Theta^T\Theta)]$$

$$= (-1/2)\ [\ 2\ \sigma^2\ \Theta + \sum (0 - X_i - X_i + 2\Theta)]$$

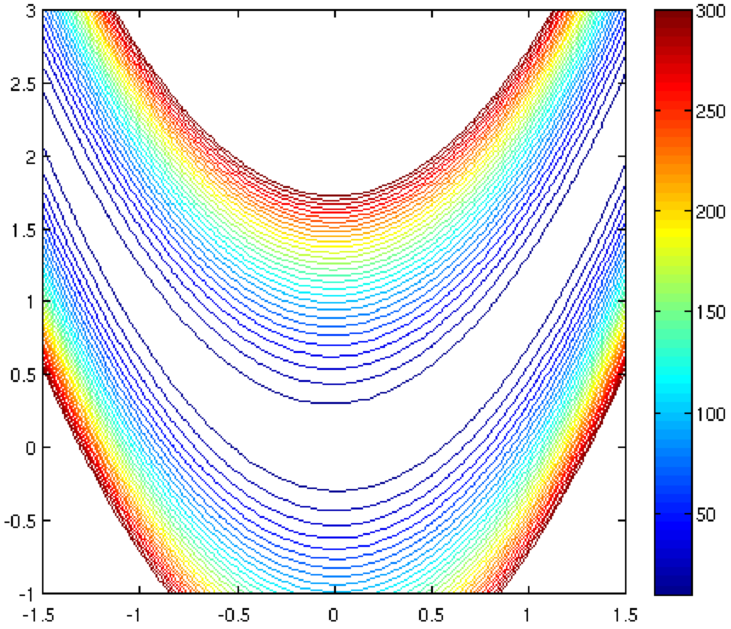
$$= (-1/2)\ [2\ \sigma^2\ \Theta + 2 \sum X_i + 2n\Theta]$$

$$= \sum X_i - (n + \sigma^2)\ \Theta$$

$$\Theta_{MAP} = \frac{\sum X_i}{(n + \sigma^2)}$$

3.

Contour plot to expression is given below and is called as rosenbrock function.



Gradient Descent

Case i) (1.2, 1.2)

1. Result too large is displayed on execution if start with learning rate =1 hence we need to reduce our learning rate. (failed to work even when I tried 0.5,0.25)

2. At learning rate=0.0005, I see that f (x) value kept on increasing.

3. At learning rate=0.00025, Gradient descent becomes stuck and follows zig zag pattern.

4. At learning rate=0.0001 and after 337734 iterations I can conclude that minimum value for the function is (1, 1).

Case ii) (-1.2, 1)

1. With starting point as (-1.2, 1) and learning rate =0.0001 Gradient descent takes more number of iterations to converge to minimum value (1, 1).

Newton' s Method

Case i) (1.2, 1.2)

```
Current function value: 0.000000
         Iterations: 12
         Function evaluations: 16
         Gradient evaluations: 27
         Hessian evaluations: 12
[ 0.99999998  0.99999997]
```

It took 12 iterations to converge to minimum value (1, 1) when our starting point is (1.2, 1.2) using Hessian in newton's method.

Case ii) (-1.2, 1)

```
Current function value: 0.000000
         Iterations: 85
         Function evaluations: 107
         Gradient evaluations: 191
         Hessian evaluations: 85
[ 1.  1.]
```

It took 85 iterations to converge to minimum value (1, 1) when our starting point is (-1.2, 1) using Hessian in newton's method.

Note: - Please refer attached programs to verify veracity of my solution.

4.

a) Once features are increased to 17 or greater, we get an error related to overfitting "**line 90, in _raise_linalgerror_singular raise LinAlgError("Singular matrix") numpy.linalg.linalg.LinAlgError: Singular matrix** "

This occurs when our $(X X^T)^{-1}$ is non-invertible or singular matrix which may be caused by identical, similar, or linearly dependent features. Determinant value for singular matrix is 0.

From section 4.5.1 in notes

This problem can be solved by singular value decomposition which can be solved by using $X = U\Sigma V^T$.

b)

In order to implement Ridge regression, we need to add λ and ||w|| in our algorithm.

After implementing ridge regression and after careful comparison I observe following

- Ridge regression provides better solution compared to FS Linear regression.
- Accuracy for Ridge regression is lower compared to FS Linear regression.
- Accuracy decreases after implementing regularization and choosing high value of lambda.
- Extremely high values of lambda can increase accuracy.

c)

There are different ways to do feature selection:-

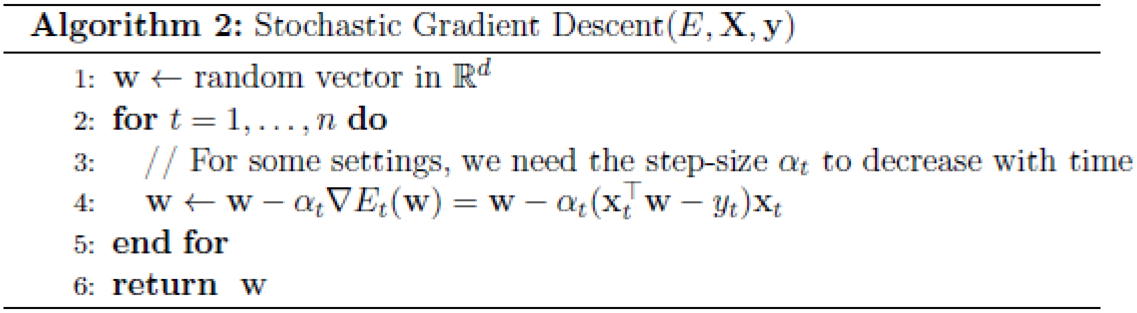
1. **Filter method**:- perform feature selection as a preprocessing step, independently of the learning algorithm used for model construction. An example of such a mechanism is variable ranking, using e.g. correlation coefficients between each feature and the dependent variable. Another filter approach selects features based on a linear mode and then constructs a non-linear model using the selected features.

2. **Wrapper method:-** are characterized as being a subset selection approach. The main idea of wrapper methods is to assess subsets of variables according to their usefulness to a given learning algorithm. Here, the learning algorithm is treated as a black box, and the best subset of features is determined according to the performance of the particular algorithm applied to build a regression mode.

3. **Embedded methods: -** incorporate feature selection as part of the training process, i.e., feature selection is done when building the predictive model. Such mechanisms usually involve changes in the objective function of the applied learning algorithm and therefore are commonly associated to a specific predictor. Examples of embedded methods are decision trees.

d) I implemented stochastic gradient descent in the program:-

```
self.weights = np.zeros(rows)
        for i in range(columns):
         self.weights = self.weights - np.dot(alpha,(np.dot(Xless.T,np.dot(Xless, self.weights) - ytrain)))
         print self.weights
```



Algorithm 2: Stochastic Gradient Descent$(E, \mathbf{X}, \mathbf{y})$

```
$\mathbf{w} \leftarrow$ random vector in $\mathbb{R}^d$
for $t = 1, \ldots, n$ do
    // For some settings, we need the step-size $\alpha_t$ to decrease with time
    $\mathbf{w} \leftarrow \mathbf{w} - \alpha_t \nabla E_t(\mathbf{w}) = \mathbf{w} - \alpha_t(\mathbf{x}_t^\top \mathbf{w} - y_t)\mathbf{x}_t$
end for
return $\mathbf{w}$
```

In stochastic approximation, we typically approximate the gradient with one sample.

e)

Both ordinary least square and logistic regression are derived from Generalized linear models.

In its simplest form, a linear model specifies the (linear) relationship between a dependent (or response) variable *Y*, and a set of predictor variables, the *X*'s, so that

$Y = b_0 + b_1X_1 + b_2X_2 + \ldots + b_kX_k$

In this equation *b_0* is the regression coefficient for the intercept and the *b_i* values are the regression coefficients (for variables 1 through *k*) computed from the data.

List of References

1. Pattern Recognition and Machine Learning :- Christopher M. Bishop
2. Machine Learning Notes:- Predrag Radivojac and Martha White
3. http://www.onmyphd.com/?p=gradient.descent
4. http://code.activestate.com/recipes/576762-newton-raphson-root-finding/
5. http://people.duke.edu/~ccc14/sta-663/BlackBoxOptimization.html
6. http://docs.scipy.org/doc/scipy/reference/generated/scipy.optimize.curve_fit.html
7. http://www.dcs.uchile.cl/images/dcs/publicaciones/jmirandap/Internacional/A%20Hybrid%20Forecasting%20Methodology%20using%20Feature%20Selection%20and%20Support%20Vector%20Regression.pdf
8. Parts of question 4 discussed with Rohit Nair, Sohail Jain.