

Solutions to CS 274A Homework 2

Probabilistic Learning: Theory and Algorithms, CS 274A, Winter 2009

February 3, 2009

Problem 1:

Let X be a Gaussian random variable with unknown parameters $\theta_1 = \mu$ and $\theta_2 = \sigma^2$. Given a data set $D = \{x(1), \dots, x(N)\}$, derive maximum likelihood estimators for θ_1 and θ_2 .

SOLUTION: *Using the pdf for a Gaussian, we have*

$$P(x(i)|\theta_1, \theta_2) = \frac{1}{\sqrt{2\pi\theta_2}} \exp\left(-\frac{(x(i) - \theta_1)^2}{2\theta_2}\right)$$

Therefore, under IID,

$$L(\underline{\theta}) = P(D|\theta_1, \theta_2) = \prod_i^n P(x(i)|\theta_1, \theta_2) = (2\pi\theta_2)^{-\frac{n}{2}} \exp\left(-\frac{\sum_i^n (x(i) - \theta_1)^2}{2\theta_2}\right),$$

$$l(\underline{\theta}) = \log(L(\underline{\theta})) = -\frac{n}{2} \log(2\pi\theta_2) - \frac{1}{2\theta_2} \sum_i^n (x(i) - \theta_1)^2$$

We can find $\hat{\theta}_{ML}$ by setting the partial derivatives to 0 and solving.

$$\frac{\partial l(\underline{\theta})}{\partial \theta_1} = \frac{1}{\theta_2} \sum_i^n (x(i) - \theta_1) = \frac{1}{\theta_2} \sum_i^n x(i) - n\theta_1 = 0 \Rightarrow \hat{\theta}_{1ML} = \frac{\sum x(i)}{n}$$

$$\frac{\partial l(\underline{\theta})}{\partial \theta_2} = \frac{n}{2\theta_2} + \frac{\sum_i^n (x(i) - \theta_1)^2}{2\theta_2^2} = 0 \Rightarrow \hat{\theta}_{2ML} = \frac{\sum (x(i) - \hat{\theta}_{1ML})^2}{n}$$

Problem 2:

(20 points)

Assume you are working for a large search engine company and you wish to build a probabilistic model for the number of search results that a user clicks on. Let X be a random variable taking values $x = 0, 1, 2, \dots$, where x represents the number of clicks. Assume that two different models are being considered for this problem:

- The geometric distribution, where

$$P(X = x) = (1 - p)p^x,$$

where p is the parameter of the model and $0 < p < 1$.

- The Poisson distribution, where

$$P(X = x) = \frac{e^{-\lambda} \lambda^x}{x!},$$

where $\lambda > 0$ is the parameter of the model.

Assume that you have a data set $D = \{x(1), \dots, x(N)\}$.

SOLUTION:

1. Write down expressions for the log-likelihood for both the geometric and Poisson distributions.

$$\begin{aligned} \text{(a) } L(\theta) &= P(D|\theta) = \prod_i^n P(X = x(i)) = \prod_i^n (1 - p)p^{x(i)} = (1 - p)^n p^{\sum_i^n x(i)} \\ l(\theta) &= \log(L(\theta)) = n \log(1 - p) + \log(p) \sum_i^n x(i) \end{aligned}$$

$$\begin{aligned} \text{(b) } L(\theta) &= P(D|\theta) = \prod_i^n \frac{\exp^{-\lambda} \lambda^{x(i)}}{x(i)!} = \frac{\exp^{-n\lambda} \lambda^{\sum_i^n x(i)}}{\prod_i^n x(i)!} \\ l(\theta) &= -n\lambda + \log(\lambda) \sum_i^n x(i) - \sum_i^n \log(x(i)!) \end{aligned}$$

2. Derive the maximum likelihood estimate of p for the geometric distribution.

$$\frac{\partial l(\theta)}{\partial p} = -n \frac{1}{(1 - p)} + \frac{1}{p} \sum_i^n x(i) = -np + (1 - p) \sum_i^n x(i) = 0 \Rightarrow \hat{p}_{ML} = \frac{\sum_i^n x(i)}{\sum_i^n x(i) + n}$$

3. Derive the maximum likelihood estimate of λ for the Poisson distribution.

$$\frac{\partial l(\theta)}{\partial \lambda} = -n + \frac{\sum_i^n x(i)}{\lambda} = 0 \Rightarrow \hat{\lambda}_{ML} = \frac{\sum_i^n x(i)}{n}$$

4. Let the data set D consist of the following table of values:

value	number of occurrences
0	15
1	30
2	25
3	16
4	7
5	3
6	2
7	0
8	1
9	1

Plot the log-likelihood for each of the geometric and Poisson models as a function of their respective parameters for this data set.

Note that the log-likelihood for the geometric peaks at $\hat{p}_{ML} = 2/3$ and the log-likelihood for the Poisson peaks at $\hat{\lambda}_{ML} = 2$. This is consistent with $n = 100$ and $\sum_{i=1}^n x(i) = 200$. Note that a common error in the solution is to use $n = 10$ instead of $n = 100$, i.e., to assume that n is the number of non-zero values in the table above. Note that this is not correct—the correct value of n is the number of individual Web users, e.g., there were 30 users who had a search query length of 1, 25 users with a search query length of 2, and so on.

- Using the maximum likelihood estimates of the parameters, on a single plot with the x-axis running from 0 to 10, plot the following:
 - The empirical probability (from the data) of each value
 - The probability distribution for the geometric model
 - The probability distribution for the Poisson model
- Is the geometric or Poisson model a better fit to this data? Explain the reasoning behind your answer

We can see from the plot of the empirical distribution and the 2 models that the Poisson model is a better fit, e.g., the geometric decays quickly and has its mode at 0, while the Poisson better reflects the fact that the empirical distribution is non-monotonic and has a mode away from 0.

We can get a more quantitative answer to “which model is better” by computing a goodness-of-fit statistic. One such statistic is the log-likelihood itself—if you compute the log-likelihood of the data, for each model, evaluated at the maximum likelihood value of their respective parameters, you will find that the Poisson has higher likelihood. One could also use other goodness-of-fit criteria such as squared-error (or the related Chi-square goodness-of-fit test). Note that if one model had more parameters than the other then we could not use the training

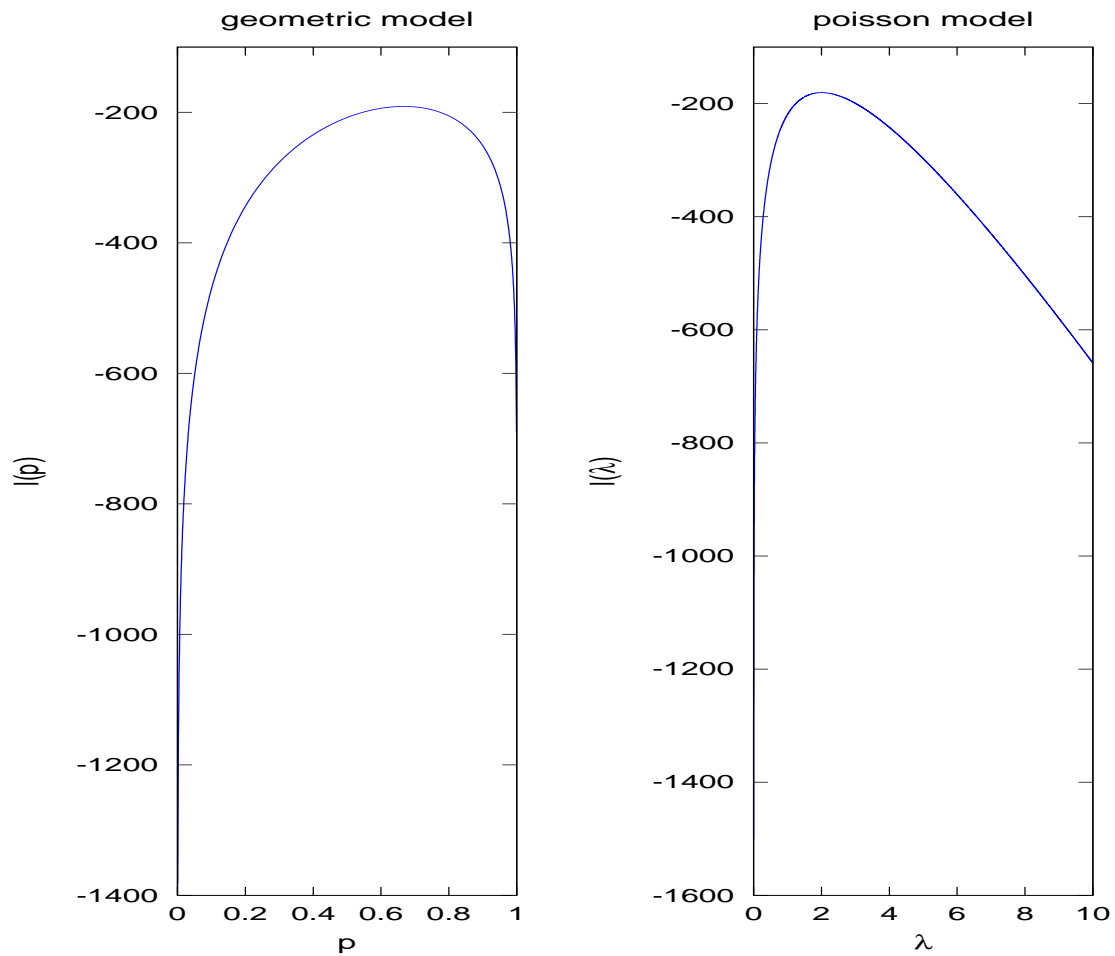


Figure 1: Log-likelihood for (a) geometric model and (b) Poisson model.

data to tell which one is “best” since the more complex model would have more degrees of freedom to fit the data. In such cases we can compute goodness-of-fit (e.g., log-likelihood) on *test data*, and select the model that has the best goodness-of-fit on the test data, where the parameters of each model have been fixed on a separate training data set.

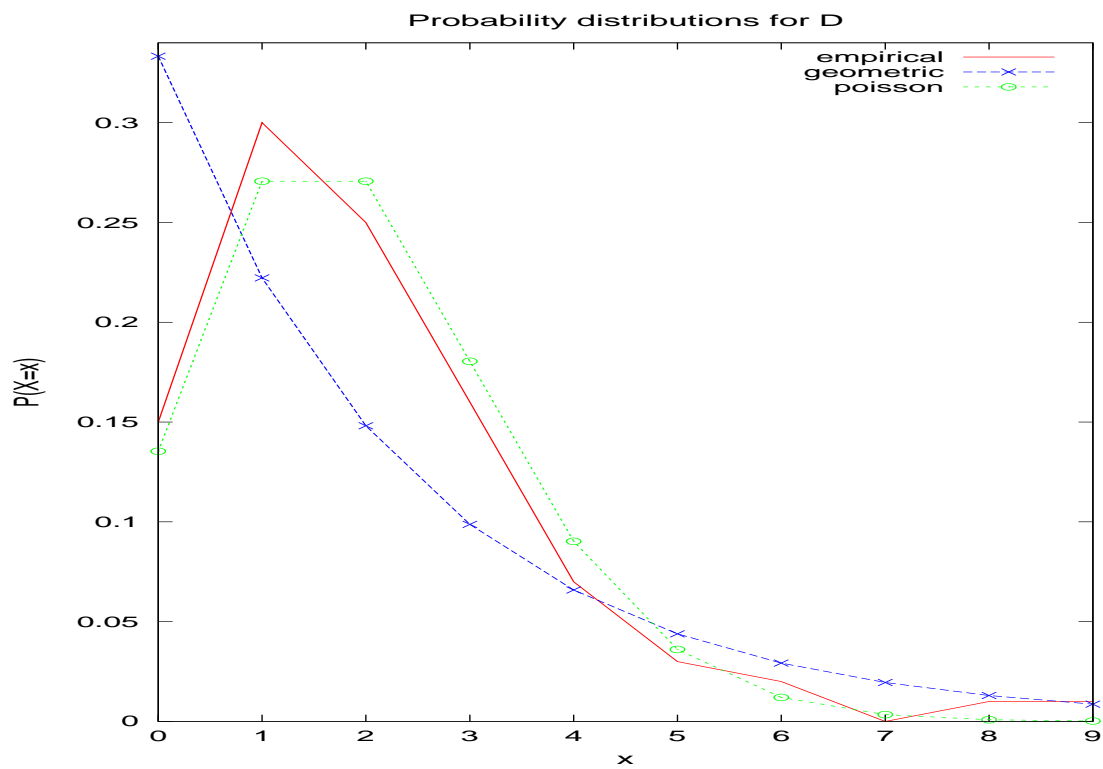


Figure 2: Empirical, geometric, and Poisson probability distributions using ML estimates.

Problem 3:

Consider building a probabilistic model for how often words occur in English. Let W be a random variable, taking values $w \in \{w_1, \dots, w_V\}$, where V is the number of words in the vocabulary. In practice V can be very large, e.g., $V = 100,000$ is not unusual (there are more words than this in English, but many rare words are not modeled).

The *multinomial model* for W is essentially the same as the binomial model for tossing coins, where we have independent trials, but instead of two possible outcomes there are now V possible outcomes for each “trial”. The parameters of the multinomial are $\theta = \{\theta_1, \dots, \theta_V\}$, where $\theta_k = P(W = w_k)$, and where $\sum_{k=1}^V \theta_k = 1$. Denote the observed data as $D = \{r_1, \dots, r_V\}$, where r_k is the number of times word k occurred in the data (these are the sufficient statistics for this model).

Derive the maximum likelihood estimates for each θ_k for this model.

SOLUTION:

We know that we can write the likelihood as follows, by definition. Note that this is proportional to the actual likelihood, since we’ve removed the coefficient containing the factorials from the front of the equation.

$$L(\underline{\theta}) \propto \prod_{i=1}^N \theta_i^{k_i}$$

Turning this into log likelihood, we get

$$l(\underline{\theta}) \propto \sum_{i=1}^N r_i \ln(\theta_i)$$

Now if we take the derivative here with respect to some θ_k , all other terms will drop out and we will get that the derivative of the log likelihood is zero when $\theta_k = \infty$. However, this ignores the constraint that all θ_i must sum to 1. To factor in this constraint, we add a Lagrange multiplier to our equation. This gives us

$$l(\underline{\theta}) \propto C(1 - \sum_{i=1}^N \theta_i) + \sum_{i=1}^N r_i \ln(\theta_i)$$

or equivalently,

$$l(\underline{\theta}) \propto -C(-1 + \sum_{i=1}^N \theta_i) + \sum_{i=1}^N r_i \ln(\theta_i)$$

Now we take the derivative with respect to some θ_k and solve for C .

$$\frac{dl(\theta_k)}{\theta_k} = 0 = \frac{r_k}{\theta_k} - C$$

$$\theta_k = \frac{r_k}{C}$$

We can now expand the constraint, using this expression for an arbitrary θ_i .

$$\sum_{i=1}^N \theta_i = 1$$

$$\sum_{i=1}^N \frac{r_i}{C} = 1$$

$$C = \sum_{i=1}^N r_i$$

Finally, substituting this back into our expression for θ_k , we get

$$\theta_k = \frac{r_k}{\sum_{i=1}^N r_i}$$

This ML estimate is as we expected, simply the observed frequency of each word, relative to the total number of observed words.

Problem 4:

Let X be uniformly distributed with lower limit a and upper limit b , where $b > a$, i.e.,

$$p(x) = \frac{1}{b-a}$$

for $a \leq x \leq b$ and $p(x) = 0$ otherwise.

SOLUTION:

1. Derive maximum likelihood estimators for a and b (think carefully about how to do this).

$L(\theta) = \prod_i^n p(x(i)) = \left(\frac{1}{b-a}\right)^n$ if $\forall_x x \in [a, b]$ else 0 therefore to maximize $L(\theta)$ we minimize $\frac{1}{b-a}$ with our constraint and find that $\hat{a}_{ML} = \min(D)$ and $\hat{b}_{ML} = \max(D)$.

2. Write 2 or 3 sentences suggesting why these maximum likelihood estimates might not necessarily be the best estimates.

This estimate is biased because we will always over-estimate a and under-estimate b , since the chance we will have seen a and b in our data is infinitesimally small.

3. There are multiple alternatives to maximum likelihood, including for example the method of moments. All of these methods essentially try to extend the estimated range of the uniform beyond the “minimal range” of the ML solution. Below is one solution suggested by a student. Suggest an alternative method for estimating the parameters.

An alternative estimate is to use $\hat{x} = \frac{\sum x(i)}{n}$ as our estimate of $\frac{b+a}{2}$ and $\hat{d} = \frac{\sum |x(i) - \hat{x}|}{n-1}$ as our estimate of $\frac{b-a}{4}$. From these estimates we can estimate a and b as $\hat{a} = \frac{b+a}{2} - \frac{b-a}{2} = \hat{x} - 2\hat{d}$ and $\hat{b} = \frac{b+a}{2} + \frac{b-a}{2} = \hat{x} + 2\hat{d}$.

Problem 5:

Consider two data sets D_1 and D_2 , each consisting of scalar measurements $x_i, i = 1, \dots, N_1$ for D_1 and $x_j, j = 1, \dots, N_2$ for D_2 . Assume that each set of measurements comes from a Gaussian distribution. The two Gaussian distributions share a common variance σ^2 and the mean μ_2 of the Gaussian for data set D_2 is known to be twice the value of the mean μ_1 for the first data set D_1 . μ_1, μ_2 and σ^2 are all assumed unknown.

SOLUTION:

- See the discussion in class for a definition of the graphical model.
- Define the log-likelihood for this problem.

The likelihood for this problem can be computed as follows:

$$\begin{aligned}
 p(D|\theta) &= \prod_{i=1}^{N_1} p(D_1|\theta_1) \prod_{i=1}^{N_2} p(D_2|\theta_2) \\
 &= \prod_{i=1}^{N_1} \left[\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x_{(i)}-\mu_1)^2} \right] \prod_{i=1}^{N_2} \left[\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x_{(i)}-\mu_2)^2} \right] \\
 \Rightarrow l(\theta) &= \sum_{i=1}^{N_1} \left[\frac{-1}{2} \log 2\pi\sigma^2 - \frac{1}{2\sigma^2}(x_{(i)}-\mu_1)^2 \right] + \sum_{i=1}^{N_2} \left[\frac{-1}{2} \log 2\pi\sigma^2 - \frac{1}{2\sigma^2}(x_{(i)}-\mu_2)^2 \right] \\
 &= \frac{-N_1}{2} \log 2\pi\sigma^2 - \frac{N_2}{2} \log 2\pi\sigma^2 - \frac{1}{2\sigma^2} \left[\sum_{i=1}^{N_1} (x_{(i)}-\mu_1)^2 + \sum_{i=1}^{N_2} (x_{(i)}-\mu_2)^2 \right]
 \end{aligned}$$

With the $\mu_2 = 2\mu_1$ assumption we can replace μ_2 in log-likelihood equation with $2\mu_1$ as follows :

$$= -\frac{N_1 + N_2}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \left[\sum_{i=1}^{N_1} [x_{(i)}^2 - 2x_{(i)}\mu_1] + N_1\mu_1^2 + \sum_{i=1}^{N_2} [x_{(i)}^2 - 4x_{(i)}\mu_1] + 4N_2\mu_1^2 \right]$$

- Derive the maximum likelihood estimators for the unknown parameters.

To solve the ML problem, we should set the derivative of log-likelihood function with respect to μ_1 and σ^2 to zero and solve the set of two equations:

$$\begin{aligned}
 \frac{\partial l(\theta)}{\partial \mu_1} &= \frac{-1}{2\sigma^2} \left[\sum_{i=1}^{N_1} (-2x_{(i)}) + 2N_1\mu_1 + \sum_{i=1}^{N_2} (-4x_{(i)}) + 8N_2\mu_1 \right] = 0 \\
 \Rightarrow 2N_1\mu_1 + 8N_2\mu_1 &= 2 \sum_{i=1}^{N_1} x_{(i)} + 4 \sum_{i=1}^{N_2} x_{(i)}
 \end{aligned}$$

$$\Rightarrow \mu_1 = \frac{\left[2 \sum_{i=1}^{N_1} x_{(i)} + 4 \sum_{i=1}^{N_2} x_{(i)} \right]}{2N_1\mu_1 + 8N_2\mu_1}$$

$$\Rightarrow \mu_2 = \frac{2 \left[2 \sum_{i=1}^{N_1} x_{(i)} + 4 \sum_{i=1}^{N_2} x_{(i)} \right]}{2N_1\mu_1 + 8N_2\mu_1}$$

$$\begin{aligned} \frac{\partial l(\underline{\theta})}{\partial \sigma^2} &= -\frac{N_1 + N_2}{2} \cdot \frac{2\pi}{2\pi\sigma^2} + \frac{1}{2(\sigma^2)^2} \left[\sum_{i=1}^{N_1} (x_{(i)} - \mu_1)^2 + \sum_{i=1}^{N_2} (x_{(i)} - \mu_2)^2 \right] = 0 \\ \Rightarrow \sigma^2 &= \frac{1}{N_1 + N_2} \left[\sum_{i=1}^{N_1} (x_{(i)} - \hat{\mu}_1)^2 + \sum_{i=1}^{N_2} (x_{(i)} - \hat{\mu}_2)^2 \right] \end{aligned}$$

Problem 6:

Consider a data set D consisting of N scalar measurements $x_i, 1 \leq i \leq N$, where each measurement is taken from a different Gaussian, such that each Gaussian has the same mean μ , and each Gaussian has a different variance $\sigma_i^2, 1 \leq i \leq N$, where these N variances are known.

SOLUTION:

- Define the log-likelihood for this problem

$$L(\theta) = \prod_i^N P(x_i|\mu, \sigma_i^2) = (2\pi)^{-\frac{N}{2}} \left(\prod_i^N \frac{1}{\sigma_i^2} \right) \exp\left(-\frac{1}{2} \sum_i^N \left(\frac{x_i - \mu}{\sigma_i}\right)^2\right)$$

$$l(\theta) = -\frac{N}{2} \log 2\pi - \sum_i^N \log \sigma_i - \frac{1}{2} \sum_i^N \left(\frac{x_i - \mu}{\sigma_i}\right)^2$$

- Derive the maximum likelihood estimator for μ

$$\frac{\partial l(\theta)}{\partial \mu} = \sum_i^N \frac{1}{\sigma_i} \left(\frac{x_i - \mu}{\sigma_i}\right) = \sum_i^N \frac{1}{\sigma_i^2} (x_i - \mu) = 0$$

$$\Rightarrow \hat{\mu}_{ML} = \frac{\sum_i^N \frac{x_i}{\sigma_i^2}}{\sum_i^N \frac{1}{\sigma_i^2}}$$

- Comment on the functional form of your solution: for example, can you interpret the result in the form of a weighted estimate? what are the weights?

The estimate can be interpreted as a weighted average of the data with weights $w_i = \frac{1}{\sigma_i^2}$

so that, $\hat{\mu}_{ML} = \frac{\sum_i w_i x_i}{\sum_i w_i}$

Problem 7:

Let \mathbf{X} be a d -dimensional real-valued random variable taking values \underline{x} . Assume \mathbf{X} has a multivariate Gaussian distribution with unknown mean $\underline{\mu}$ and a *diagonal* covariance matrix Σ .

SOLUTION:

- How many parameters are there for this model?

There are d μ and d σ^2 so there are $2d$ parameters.

- Given a data set $D = \{\underline{x}(1), \dots, \underline{x}(N)\}$ write down an expression for the log-likelihood that is expressed in terms of the parameters.

$$\begin{aligned} P(\underline{x}(i)|\underline{\mu}, \Sigma) &= \prod_j^d P(x_j(i)|\mu_j, \Sigma_{jj}) \\ L(\theta) &= \prod_i^N \prod_j^d \frac{1}{\sqrt{2\pi\Sigma_{jj}}} \exp\left(-\frac{(x_j(i)-\mu_j)^2}{2\Sigma_{jj}^2}\right) \\ l(\theta) &= \sum_i^N \sum_j^d \left(-\frac{1}{2} \log(2\pi\Sigma_{jj}) - \frac{1}{2\Sigma_{jj}}(x_j(i) - \mu_j)^2\right) \end{aligned}$$

- Derive maximum likelihood estimates for the unknown parameters.

$$\frac{\partial l(\theta)}{\partial \underline{\mu}} = \frac{\partial(\sum_i^N \sum_j^d \left(-\frac{1}{2\Sigma_{jj}}(x_j(i)-\mu_j)^2\right))}{\partial \underline{\mu}} = \sum_i^N \sum_j^d \frac{1}{\Sigma_{jj}} x_j(i) - N \sum_j^d \frac{\mu_j}{\Sigma_{jj}} = 0$$

We can solve for each μ independently to get: $\hat{\underline{\mu}}_{ML} = \frac{\sum \mathbf{X}}{N}$.

Similarly, we see that $\hat{\Sigma}_{ML} = \frac{\sum (\mathbf{X} - \underline{\mu})^2}{N} I$

Problem 8:

Write a MATLAB function called `binlikelihood.m` that takes in arguments r and n and plots the likelihood function for the binomial model. Use your function to generate plots for (a) $r = 5$, $n = 10$, (b) $r = 50$, $n = 100$, and (c) $r = 5$, $n = 50$, over the range $[0, 1]$. Submit copies of both your plots and your code.

```
function binlikelihood(r,n);
    theta = 0:.001:1;
    likelihood = theta.^r.*(1-theta).^(n-r);
    plot(theta,likelihood)
    axis([0 1 0 3/2*likelihood(find(theta==round(r/n*1000)/1000))]);
    xlabel('\theta');
    ylabel('L(\theta)');
    title(sprintf('L(\theta) for r = %d, n = %d', [r n]));
end;
```

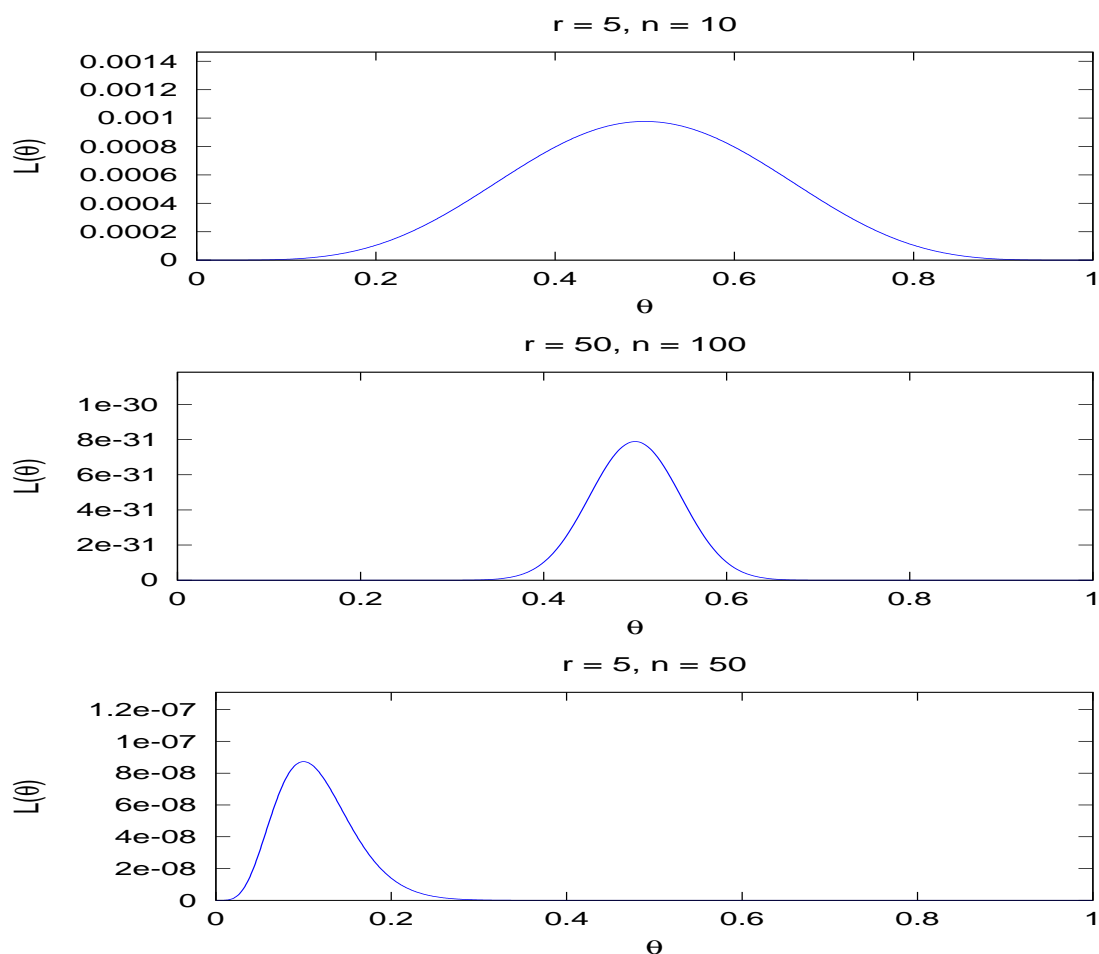


Figure 3: Output of binlikelihood for (a) $r = 5, n = 10$, (b) $r = 50, n = 100$, and (c) $r = 5, n = 50$ over the range $[0,1]$.

Problem 9:

Write a MATLAB function called `gausslogL.m` that takes in a data set in the form of an $n \times 1$ vector called `data`. The function should generate two plots:

1. the log-likelihood as a function of the mean μ , with σ fixed to its maximum likelihood value;
2. the log-likelihood as a function of σ , with μ fixed to its maximum likelihood value.

For μ , the range of the x-axis should correspond to the range of the data points supplied to the function. For σ , use a fixed range of 0.1 to 5 times the maximum likelihood estimate of σ .

Submit your code and your plots for the following two cases:

1. `data` is a vector of 10 draws from a Gaussian with mean $\mu = 10$ and $\sigma = 1$.
2. `data` is a vector of 1000 draws from the same distribution.

(If you wish, you can also plot the data points on the plot for the mean, as in Figure 2 in Note Set 3. This is optional).

```
function gausslogL(data);
    mu    = mean(data);
    sigma = std(data,1);

    theta_mu    = min(data):.01:max(data);
    theta_sigma = 0.1*sigma:.01:5*sigma;
    len = @(array) size(array)(1);
    l = @(data, mu, sigma) -len(data)/2*log(2*pi)      ...
                          -len(data)*log(sigma)      ...
                          -sumsq(data-mu)/(2*sigma^2);

    l_mu = theta_mu;
    for i = 1:length(theta_mu)
        l_mu(i) = l(data, theta_mu(i), sigma);
    end;

    l_sigma = theta_sigma;
    for i = 1:length(theta_sigma)
        l_sigma(i) = l(data, mu, theta_sigma(i));
    end;

    subplot(1,2,1)
    plot(theta_mu, l_mu)
    xlabel('\mu');
    ylabel('l(\mu)');
    title(sprintf('l(\mu) vs \mu'));

    subplot(1,2,2)
    plot(theta_sigma, l_sigma)
    xlabel('\sigma');
    ylabel('l(\sigma)');
    title(sprintf('l(\sigma) vs \sigma'));
end;
```

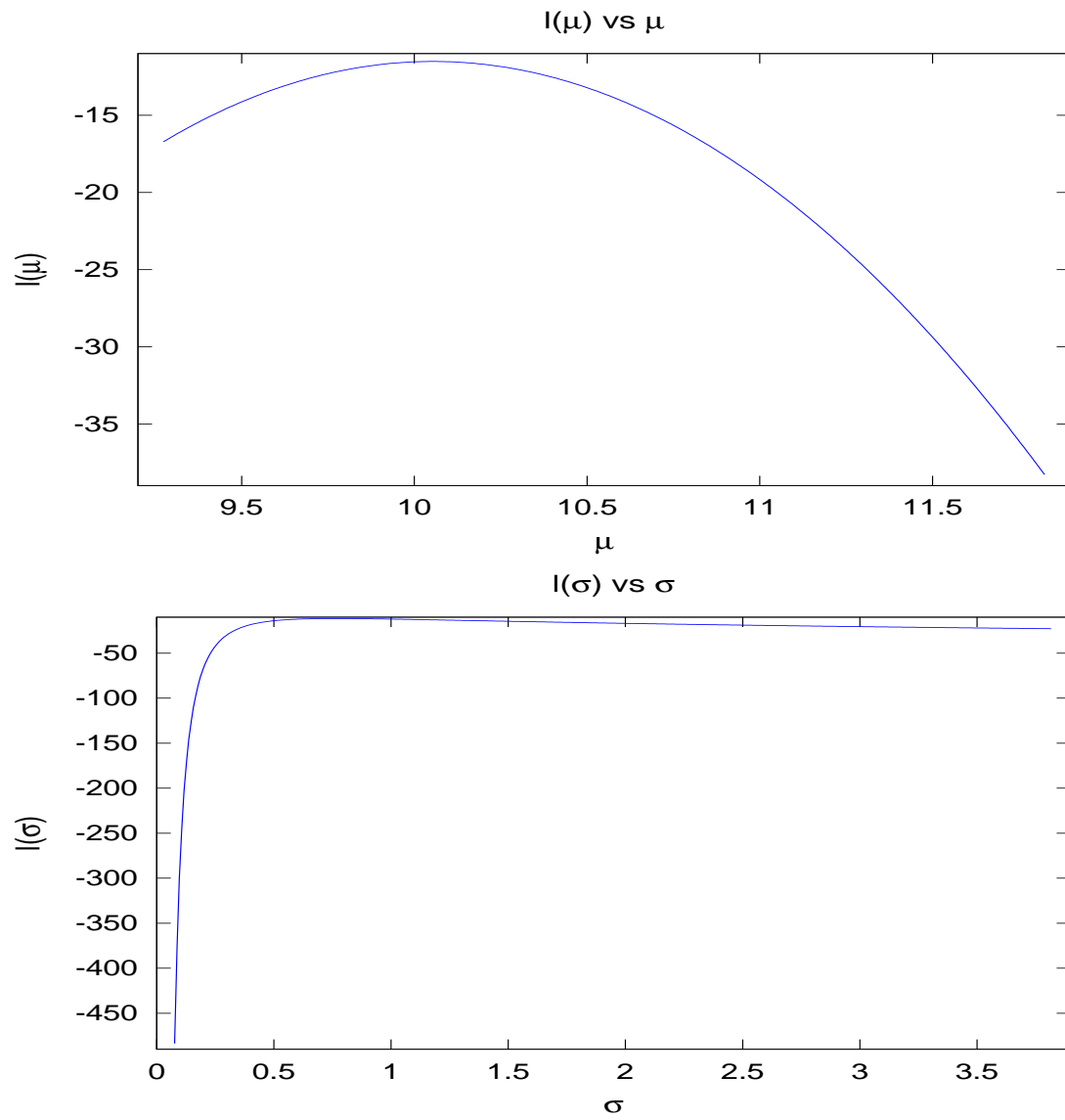


Figure 4: Output of `gausslogL` for 10 draws from a Gaussian with $\mu = 10$ and $\sigma = 1$.

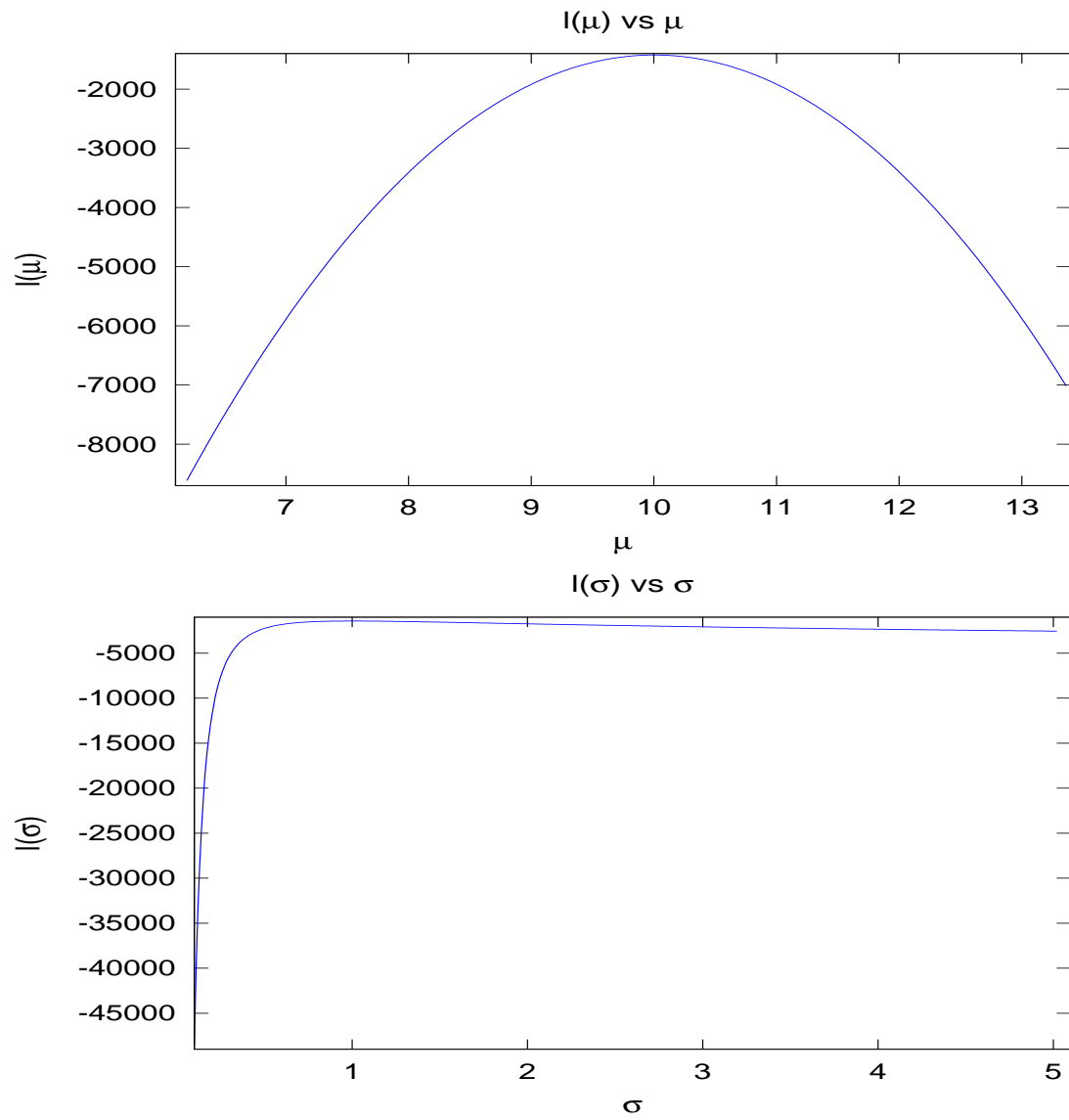


Figure 5: Output of gausslogL for 1000 draws from a Gaussian with $\mu = 10$ and $\sigma = 1$.