1. Midterm

Solve all of the problems on the midterm again (including the ones you got correct).

Solution:
See midterm solutions.

2. Exponential: MLE & MAP

The random variable $X$ is exponentially distributed with mean 1. Given $X$, the random variable $Y$ is exponentially distributed with rate $X$.

(a) Find MLE[$X \mid Y$].
(b) Find MAP[$X \mid Y$].

Solution:

(a) The density of $Y$, given $X = x$, is $f(y) = x \exp(-xy)$ for $y > 0$, so $\ln f(y) = \ln x - xy$. To maximize this over $x$, we differentiate to obtain $1/x - y = 0$, so $x = 1/y$, that is, MLE[$X \mid Y$] = $1/Y$.

(b) The posterior density of $X$ is

$$f_{X\mid Y}(x \mid y) \propto f_{Y\mid X}(y \mid x)f_X(x) = x \exp(-xy)\exp(-x)$$

$$= x \exp(-x(1 + y))$$

so we can maximize $\ln x - x(1 + y)$ over $x$. Differentiating, we have $1/x - 1 - y = 0$, or $1/x = 1 + y$. Hence, MAP[$X \mid Y$] = $1/(1 + Y)$.

3. BSC: MLE & MAP

You are testing a digital link that corresponds to a BSC with some error probability $\epsilon \in [0, 0.5]$.

(a) Assume you observe the input and the output of the link. How do you find the MLE of $\epsilon$?
(b) You are told that the inputs are i.i.d. bits that are equal to 1 with probability 0.6 and to 0 with probability 0.4. You observe $n$ outputs ($n$ is a positive integer). How do you calculate the MLE of $\epsilon$?

(c) The situation is as in the previous case, but you are told that $\epsilon$ has PDF $4 - 8x$ on $[0, 0.5)$. How do you calculate the MAP of $\epsilon$ given $n$ outputs?

**Solution:**

(a) We observe the input $X$ and the output $Y$. Thus, if $P_\epsilon$ denotes the probability distribution when the error probability of the BSC is $\epsilon$, then for $(x, y) \in \{0, 1\}^2$,

$$
\epsilon_{\text{MLE}} = \arg \max_{\epsilon \in [0, 0.5]} P_\epsilon(X = x, Y = y) = \arg \max_{\epsilon \in [0, 0.5]} \epsilon^1 \{y \neq x\} (1 - \epsilon)^1 \{y = x\}.
$$

Now if $x \neq y$, the expression is clearly maximized on the largest possible value of $\epsilon$ which is $\epsilon = 0.5$. If $x = y$, the expression is maximized for smallest value of $\epsilon$ which is 0.

(b) Suppose that we observe the outputs $y_1, \ldots, y_n$. Thus,

$$
\epsilon_{\text{MLE}} = \arg \max_{\epsilon \in [0, 0.5]} P_\epsilon(Y_1 = y_1, \ldots, Y_n = y_n).
$$

Since every use of the channel is independent we have,

$$
P_\epsilon(Y_1 = y_1, \ldots, Y_n = y_n)
= \prod_{i=1}^n P_\epsilon(Y_i = y_i)
= \prod_{i=1}^n \left[ (0.6(1 - \epsilon) + 0.4\epsilon)^1 \{y_i = 1\} + (0.4(1 - \epsilon) + 0.6\epsilon)^1 \{y_i = 0\} \right]
= \prod_{i=1}^n (0.6 - 0.2\epsilon)^{y_i} (0.4 + 0.2\epsilon)^{1-y_i}
= (0.6 - 0.2\epsilon)^{\sum_{i=1}^n y_i} (0.4 + 0.2\epsilon)^{n - \sum_{i=1}^n y_i}.
$$

Let $t = \sum_{i=1}^n y_i$. As we can see, what matters for estimating $\epsilon$ is $t$. To find the maximizer of the expression, we first take the log and then set the derivative to 0. Thus,

$$
-0.2t
\quad + \quad 0.2(n - t)
\quad \overline{0.6 - 0.2\epsilon + 0.4 + 0.2\epsilon} = 0.
$$

Solving the equation, we get

$$
\epsilon_{\text{MLE}} = 3 - \frac{5t}{n}.
$$

Of course, since we know that $0 \leq \epsilon \leq 0.5$, if $\epsilon_{\text{MLE}}$ is not in the interval $[0, 0.5]$ we should pick the closest point to it which will be either 0 or 0.5.
(c) This time we want to maximize $P(Y_1 = y_1, \ldots, Y_n = y_n \mid \epsilon = \cdot) f_\epsilon(\cdot)$.

Similar to the calculations of previous part, we want to maximize,

$$(4 - 8\epsilon)(0.6 - 0.2\epsilon)^t (0.4 + 0.2\epsilon)^{(n-t)}.$$ 

Taking the log and setting the derivative equal to 0 we have

$$-8 \frac{t}{4 - 8\epsilon} + \frac{-0.2t}{0.6 - 0.2\epsilon} + \frac{0.2(n-t)}{0.4 + 0.2\epsilon} = 0.$$ 

Then, we get the following quadratic equation.

$$0 = -8(0.6 - 0.2\epsilon)(0.4 + 0.2\epsilon) - 0.2t(4 - 8\epsilon)(0.4 + 0.2\epsilon)$$

$$+ 0.2(n-t)(4 - 8\epsilon)(0.6 - 0.2\epsilon).$$

One can solve the long quadratic equation analytically, and find $\epsilon_{\text{MAP}}$.

We skip the painful algebra here. (You also get full credit, if you find the quadratic equation.)

4. Fun with Linear Regression

Suppose $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is an unknown linear function, i.e. it is of the form $f(x) = x^\top w = x_1 w_1 + \cdots + x_d w_d$, where $w \in \mathbb{R}^d$ is the unknown parameter of the linear function. We pick $n$ points $x^{(1)}, \ldots, x^{(n)} \in \mathbb{R}^d$, and we observe $y^{(1)}, \ldots, y^{(n)} \in \mathbb{R}$ that are generated according to the model

$$y^{(i)} = f(x^{(i)}) + \epsilon_i, \text{ for } i = 1, \ldots, n,$$

where $\epsilon_1, \ldots, \epsilon_n$ are i.i.d. $\mathcal{N}(0, \sigma^2)$ random variables.

Let us first estimate $w$ when we have no prior information about it.

(a) Compute the likelihood of the parameter $w$ given the data $\{(x^{(i)}, y^{(i)})\}_{i=1}^n$

$$L(w \mid \{(x^{(i)}, y^{(i)})\}_{i=1}^n) := \prod_{i=1}^n p(y^{(i)} \mid x^{(i)}; w).$$

(b) Explicitly define a matrix $X \in \mathbb{R}^{n \times d}$ and a vector $y \in \mathbb{R}^n$ such that the optimal points of the problem

$$\min_{w \in \mathbb{R}^d} ||Xw - y||_2^2,$$

correspond to the maximizers of the likelihood.

Now assume a zero-mean Gaussian prior for each $w_i$, $i = 1, \ldots, d$. In particular assume that $w_1, \ldots, w_d$ are i.i.d. $\mathcal{N}(0, \tau^2)$, and they are also independent of the data.

(c) Compute, up to a normalization constant, the posterior distribution of $w$ given the data $\{(x^{(i)}, y^{(i)})\}_{i=1}^n$. 

3
(d) Explicitly define a matrix $X \in \mathbb{R}^{n \times d}$, a vector $y \in \mathbb{R}^n$ and a positive scalar $\lambda \in \mathbb{R}$ such that the optimal point of the problem

$$\min_{w \in \mathbb{R}^d} \|Xw - y\|_2^2 + \lambda\|w\|_2^2,$$


correspond to the maximizer of the posterior distribution of $w$.

**Solution:**

(a)

$$L(w \mid \{(x^{(i)}, y^{(i)})\}_{i=1}^n) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{(x^{(i)\top} w - y^{(i)})^2}{2\sigma^2}\right)$$

$$= \left(\frac{1}{\sqrt{2\pi\sigma}}\right)^n \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (x^{(i)\top} w - y^{(i)})^2\right).$$

(b) If we define

$$X = \begin{bmatrix} x^{(1)\top} \\ \vdots \\ x^{(n)\top} \end{bmatrix} \text{ and } y = \begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(n)} \end{bmatrix},$$

then the likelihood function can be written as

$$L(w \mid \{(x^{(i)}, y^{(i)})\}_{i=1}^n) = \left(\frac{1}{\sqrt{2\pi\sigma}}\right)^n \exp\left(-\frac{1}{2\sigma^2} \|Xw - y\|_2^2\right),$$

and since the exponential is an increasing function we can clearly see that

$$\arg\max_{w \in \mathbb{R}^d} L(w \mid \{(x^{(i)}, y^{(i)})\}_{i=1}^n) = \arg\min_{w \in \mathbb{R}^d} \|Xw - y\|_2^2.$$
then the posterior distribution is proportional to
\[ p(w \mid \{(x^{(i)}, y^{(i)})\}_{i=1}^{n}) \propto \exp\left(-\frac{1}{2} \|Xw - y\|_2^2 - \frac{1}{2} \lambda \|w\|_2^2\right), \]
and since the exponential is an increasing function we can clearly see that
\[ \arg \max_{w \in \mathbb{R}^d} p(w \mid \{(x^{(i)}, y^{(i)})\}_{i=1}^{n}) = \arg \min_{w \in \mathbb{R}^d} \{\|Xw - y\|_2^2 + \lambda \|w\|_2^2\}. \]

5. Community Detection Using MAP

It will be useful to work on this problem in conjunction with Q3 of Lab 6. The stochastic block model (SBM), as defined in Lab 6 is a random graph \( G(n, p, q) \) consisting of two communities of size \( n/2 \) each such that the probability an edge exists between two nodes of the same community is \( p \) and the probability an edge exists between two nodes in different communities is \( q \), where \( p > q \). The goal of the problem is to exactly determine the two communities given only the graph. Show that the MAP estimate of the two communities is equivalent to finding the min-bisection of the graph (i.e. the split of \( G \) into two groups of size \( n/2 \) that has the minimum edge weight across the partition).

Solution:

Let \( G \) be a random variable whose values are realizations of the graph and let \( A \) be a random variable representing the labeling of the two communities. We are interested in \( \text{MAP}[A \mid G] = \arg \max_A \mathbb{P}(G \mid A) \mathbb{P}(A) \). Note that since each assignment of labels is equally likely, the MAP rule is equivalent to the MLE, and we are simply interested in \( \arg \max_A \mathbb{P}(G \mid A) \). Let \(|E|\) be the number of edges across the partition in assignment \( A \) and let \(|T|\) be the number of edges in \( G \). We see that
\[
\mathbb{P}(G \mid A) = q^{|E|} (1 - q)^{(n/2)^2 - |E|} p^{|T| - |E|} (1 - p)^{2(n/2) - (|T| - |E|)}
\]
\[= \left( \frac{q}{1 - q} \cdot \frac{1 - p}{p} \right)^{|E|} \cdot \left( \frac{p}{1 - p} \right)^{|T|} \cdot (1 - p)^{2(n/2)} \cdot (1 - q)^{n^2/4}. \]
Now, note that the last three terms are constant for any assignment of labels and do not affect the likelihood. Also note that
\[ \left( \frac{q}{1 - q} \cdot \frac{1 - p}{p} \right) < 1, \]
since \( p > q \) so we can see that increasing \(|E|\) corresponds to decreasing the likelihood, and the MAP rule is to select the partition with the smallest number of edges across it, which is exactly the min-bisection of the graph.

6. [Bonus] Bayesian Estimation of Poisson Distribution

The bonus question is just for fun. You are not required to submit the bonus question, but do give it a try and write down your progress.

We have already learned about MLE (non-Bayesian perspective) and MAP (Bayesian perspective). In this problem, we will introduce the fully Bayesian approach to statistical estimation.

Suppose that \( X_i, i = 1, \ldots, n \), are i.i.d. drawn from a Poisson distribution of unknown mean \( M \) (\( M \) is a random variable). As a Bayesian practitioner, you have a prior belief that \( M \) is Erlang of order \( k \) and rate \( \alpha \).
(a) Find the posterior distribution \( f_{M \mid X}(\mu \mid x_1, \ldots, x_n) \).

(b) If we were using the MLE or MAP rule, then we would choose a single value \( \mu \) for \( M \); this is sometimes called a point estimate. This amounts to saying \( X \) has the Poisson distribution with mean \( \mu \).

In the Bayesian approach, we will not use a point estimate. Instead, we will keep the full information of the posterior distribution of \( \mu \), and we compute the distribution of \( X \) as

\[
P(X = x) = \int_{0}^{\infty} P(X = x \mid M = \mu) f_{M \mid X}(\mu \mid x_1, \ldots, x_n) \, d\mu.
\]

Notice that in the Bayesian approach, we do not necessarily have a Poisson distribution for \( X \) anymore. Compute \( P(X = x) \) in closed-form.

(c) You may have noticed from the previous part that the fully Bayesian approach is often computationally intractable. This is one of the main reasons why point estimates are common in practice.

Compute the MAP estimate for \( M \) and calculate \( P(X = x) \) again using the MAP rule.

\textbf{Solution:}

(a) The likelihood of the data is

\[
P(X_1 = x_1, \ldots, X_n = x_n \mid M = \mu) = \prod_{i=1}^{n} \frac{e^{-\mu x_i}}{x_i!}.
\]

The prior distribution for \( M \) is

\[
f_M(\mu) = \frac{\alpha^k \mu^{k-1} e^{-\alpha \mu}}{(k-1)!}, \quad \mu \geq 0.
\]

Therefore, the posterior distribution is

\[
f_{M \mid X}(\mu \mid x_1, \ldots, x_n) \propto \mu^{k-1} e^{-\alpha \mu} \prod_{i=1}^{n} e^{-\mu x_i} = \mu^{\sum_{i=1}^{n} x_i + k - 1} e^{-(\alpha+n)\mu}.
\]

Therefore, we see that \( f_{M \mid X} \) must be the Erlang density with rate \( \alpha + n \) and order \( \sum_{i=1}^{n} x_i + k \). Therefore,

\[
f_{M \mid X}(\mu \mid x_1, \ldots, x_n) = \frac{(\alpha + n)^{\sum_{i=1}^{n} x_i + k} \mu^{\sum_{i=1}^{n} x_i + k - 1} e^{-(\alpha+n)\mu}}{(\sum_{i=1}^{n} x_i + k - 1)!}, \quad \mu \geq 0.
\]

(b) We compute

\[
P(X = x) = \int_{0}^{\infty} \frac{e^{-\mu x} (\alpha + n)^{\sum_{i=1}^{n} x_i + k} \mu^{\sum_{i=1}^{n} x_i + k - 1} e^{-(\alpha+n)\mu}}{(\sum_{i=1}^{n} x_i + k - 1)!} \, d\mu
\]

\[
= \frac{(\alpha + n)^{\sum_{i=1}^{n} x_i + k}}{x! (\sum_{i=1}^{n} x_i + k - 1)!} \int_{0}^{\infty} \mu^{x + \sum_{i=1}^{n} x_i + k - 1} e^{-(\alpha+n+1)\mu} \, d\mu.
\]
Recognizing the integrand as a un-normalized Erlang distribution with rate $\alpha + n + 1$ and order $x + \sum_{i=1}^{n} x_i + k$, we can evaluate the integral by recalling that a probability density function integrates to 1.

$$P(X = x) = \frac{(\alpha + n)^{\sum_{i=1}^{n} x_i + k}}{x!^{\sum_{i=1}^{n} x_i + k - 1}} \frac{(x + \sum_{i=1}^{n} x_i + k - 1)!}{(\alpha + n + 1)^{x + \sum_{i=1}^{n} x_i + k - 1}}.$$ 

Let $m = \sum_{i=1}^{n} x_i + k$. We can write the probability as

$$P(X = x) = \left(\frac{x + m - 1}{m - 1}\right)^m \left(\frac{\alpha + n}{\alpha + n + 1}\right)^x.$$ 

You may recognize this distribution. $X + m$ has the Pascal distribution with $p = (\alpha + n)/(\alpha + n + 1)$ and order $m$.

(c) Since we already have the posterior distribution, we simply need to take logarithms and differentiate.

$$\frac{d}{d\mu} \left(\sum_{i=1}^{n} x_i + k - 1\right) \ln \mu - (\alpha + n)\mu = \frac{\sum_{i=1}^{n} x_i + k - 1}{\mu} - (\alpha + n),$$

so our MAP estimate is

$$\text{MAP}[M \mid X] = \frac{\sum_{i=1}^{n} x_i + k - 1}{\alpha + n}.$$ 

Let $\beta = (\sum_{i=1}^{n} x_i + k - 1)/(\alpha + n)$. Then, $X$ is Poisson with mean $\beta$, or

$$P(X = x) = \frac{e^{-\beta} \beta^x}{x!}.$$