G. Prashant
BS17B011

CH5115 - Parameter and State Estimation

Assignment 3

Question 1

Given,

$$f(y) = \begin{cases} e^{-(y-\theta)} & , x > \theta, \ -\infty < \theta < \infty \\ 0 & , \text{otherwise.} \end{cases}$$

Part (a)

$T_N = 2 f_{min}(y_N)$

Let random variable $Z = 2Y$, and the pdf becomes

$$f_Z(z) = \begin{cases} \frac{1}{2} e^{-\left(\frac{z}{2} - \theta\right)} & , z > 2\theta \\ 0 & , \text{otherwise} \end{cases}$$

And the statistic T_N becomes

$$T_N = \min(Z_N)$$

To determine the pdf of T_N, the 1st step is to find the cdf of Z. f_Z

$$F_Z(z) = \int_{2\theta}^{z} f_z(z)\, dz = \int_{2\theta}^{z} \frac{1}{2} e^{-\left(\frac{z}{2} - \theta\right)} dz$$

which gives,

$$Pr_Z(Z < z) = F_Z(z) = \begin{cases} 1 - e^{-\left(\frac{z}{2} - \theta\right)} & , \text{when } z > 2\theta \\ 0 & , \text{otherwise} \end{cases}$$

Considering z as a fixed point and given that we have N observations of Z_N, the cdf of T_N will be the probability that at least one observation is less than z.

$$F_{T_N}(z) = 1 - Pr_z(Z_i > z \ \forall i = 1, 2 \cdots N)$$

$$= 1 - (1 - F_Z(z))^N$$

$$= 1 - e^{-N\left(\frac{z}{2} - \theta\right)}, \quad \text{when } z > 2\theta$$

Pdf can be obtained by differentiating F_{T_N}.

$$f_{T_N}(z) = \begin{cases} \frac{N}{2} e^{-N\left(\frac{z}{2} - \theta\right)} & , z > 2\theta \\ 0 & , \text{otherwise.} \end{cases}$$

where $z = 2y$

Part 1b)

We have derived that, for $z \geq 2\theta$, the pdf of T_N is

$$f_{T_N}(z) = \frac{N}{2} e^{-N(\frac{z}{2} - \theta)}$$

$$E(T_N) = \int_{2\theta}^{\infty} z f(z;\theta)\, dz$$

$$= \frac{N}{2} \int_{2\theta}^{\infty} z\, e^{-N(z/2 - \theta)}\, dz$$

$$= \frac{N}{2} \left[z \int e^{-N(z/2-\theta)} dz - \int \left(\int e^{-N(z/2-\theta)} dz \right) dz \right]_{2\theta}^{\infty}$$

$$= \frac{N}{2} \left[\frac{-2z}{N} e^{-N(z/2-\theta)} + \frac{2}{N} \int e^{-N(z/2-\theta)} dz \right]_{2\theta}^{\infty}$$

$$= -z e^{-N(z/2-\theta)} \Big|_{2\theta}^{\infty} - \frac{2}{N} e^{-N(z/2-\theta)} \Big|_{2\theta}^{\infty}$$

$$= 2\theta + \frac{2}{N}.$$

Therefore bias $\Delta T_N = E(T_N) - \theta = \theta + \frac{2}{N}$.

To correct the bias, we obtain statistic such that

$$T_N' = \frac{T_N}{2} - \frac{1}{N} = \min(y_N) - \frac{1}{N}, \text{ where } E(T_N') = \theta$$

Part (c)

In order to prove T_N' converges to θ in probability, we have to show.

$$\lim_{N\to\infty} P_r\left(|T_N' - \theta| > \epsilon\right) = 0$$

Consider $T_N' > \theta$

$$P_r\left(T_N' - \theta > \epsilon\right) = P_r\left(T_N' > \theta + \epsilon\right)$$

$$= 1 - P_r\left(T_N' < \theta + \epsilon\right)$$

$$= 1 - F_{T_N'}\left(T_N' < \theta + \epsilon\right) = 1 - F_{T_N'}(\theta + \epsilon)$$

To derive the pdf and cdf of T_N',

$$f_{T_N'}(z;\theta) = f_{T_N}(z, \theta - 1/N) = N e^{-N\left(z - (\theta - \frac{1}{N})\right)}, \quad z$$

when $z > \theta - 1/N$

0, otherwise

$$F_{T_N'}(z;\theta) = \int_{\theta - 1/N}^{z} N e^{-N\left(z - (\theta - 1/N)\right)} dz$$

$$= -e^{-N\left(z - (\theta - 1/N)\right)} \Big|_{\theta - 1/N}^{z}$$

$$= 1 - e^{-N\left(z - (\theta - 1/N)\right)}$$

$$F_{T_N'}(\theta + \epsilon) = 1 - e^{-N(\epsilon + 1/N)}$$

$$Pr(T_N - \theta > \epsilon) = 1 - F_{T_N'}(\theta + \epsilon)$$

$$= e^{-N(\epsilon + 1/N)}$$

$$\lim_{N \to \infty} Pr(T_N' - \theta > \epsilon) = \lim_{N \to \infty} e^{-N(\epsilon + 1/N)} = 0$$

~~When T_N'~~

When $\theta - \frac{1}{N} < T_N' < \theta$

$$Pr(\theta - T_n > \epsilon) = Pr(T_n < \theta - \epsilon) = F_{T_N'}(\theta - \epsilon)$$

$$F_{T_N'}(\theta - \epsilon) = \begin{cases} 1 - e^{-N(-\epsilon + 1/N)}, & \epsilon < 1/N \\ 0, & \epsilon > 1/N. \end{cases}$$

As N increases to infinity, $1/N \longrightarrow 0$ and $\epsilon > 1/N$,

$\lim_{N \to \infty} Pr(\theta - T_n' > \epsilon) = 0$, when given $\theta - \frac{1}{N} < T_N' < \theta$

Hence, $\lim_{N \to \infty} Pr(|T_N' - \theta| > \epsilon) = 0$, this implies the statistic T_N' converges to the true value as N tends to infinity.

Question 2

Given

$$V[n] \triangleq V(f_n) = \sum_{K=0}^{N-1} v[K] \exp(-j2\pi f_n K) \quad , f_n = \frac{n}{N}$$

$$\frac{|V(f_n)|^2}{N} = \frac{a_n^2 + b_n^2}{N}$$

$$a_n = Re(V[n]) \quad , \quad b_n = Im(V[n])$$

This gives

$$a_n = \sum_{K=0}^{N-1} v[K] \cos(2\pi f_n K)$$

$$b_n = -\sum_{K=0}^{N-1} v[K] \sin(2\pi f_n K)$$

Part (a)

We invoke 2 properties of Gaussian distributions

① If $X \sim \mathcal{N}(\mu, \sigma^2) \Rightarrow (\alpha X) \sim \mathcal{N}(\mu, (\alpha\sigma)^2)$

② If $X \sim (\mu_x, \sigma_x^2)$ and $Y \sim (\mu_y, \sigma_y^2)$ and X, Y are independent, $(X+Y) \sim \mathcal{N}(\mu_x + \mu_y, \sigma_x^2 + \sigma_y^2)$

When

$$a_n = \sum_{K=0}^{N-1} v[K] \cos(2\pi f_n K)$$

Given $v[K] \sim GWN(0,1)$, let $u[K] = v[K] \cos(2\pi f_n K)$

where $u[K] \sim \mathcal{N}(0, \cos^2(2\pi f_n K))$

$$\Rightarrow a_n = \sum_{K=0}^{N-1} u[K]$$

III^{ly} $w[K] = v[K]\sin(2\pi f_n K)$, $w[K] \sim N(0, \sin^2(2\pi f_n K))$

$$b_n = \sum_{K=0}^{N-1} w[K]$$

As the processes $u[K]$ and $w[K]$ are Gaussian distributed with 0 mean, and each of the process has 0 autocovariance for non zero lags* (even though it is variance non stationary), we can say that both a_n and b_n are Gaussian distributed, as they are the sum of uncorrelated Gaussian random variables.

By. The mean and variance of a_n is given by

$$\mu_{a_n} = \sum_{K=0}^{N-1} \mu_{u[K]} = 0$$

$$\sigma_{a_n}^2 = \sum_{K=0}^{N-1} \sigma_{u[K]}^2 = \sum_{K=0}^{N-1} \cos^2 2\pi f_n K$$

III^{ly}

$$\mu_{b_n} = 0$$

$$\sigma_{b_n}^2 = \sum_{K=0}^{N-1} \sin^2(2\pi f_n K)$$

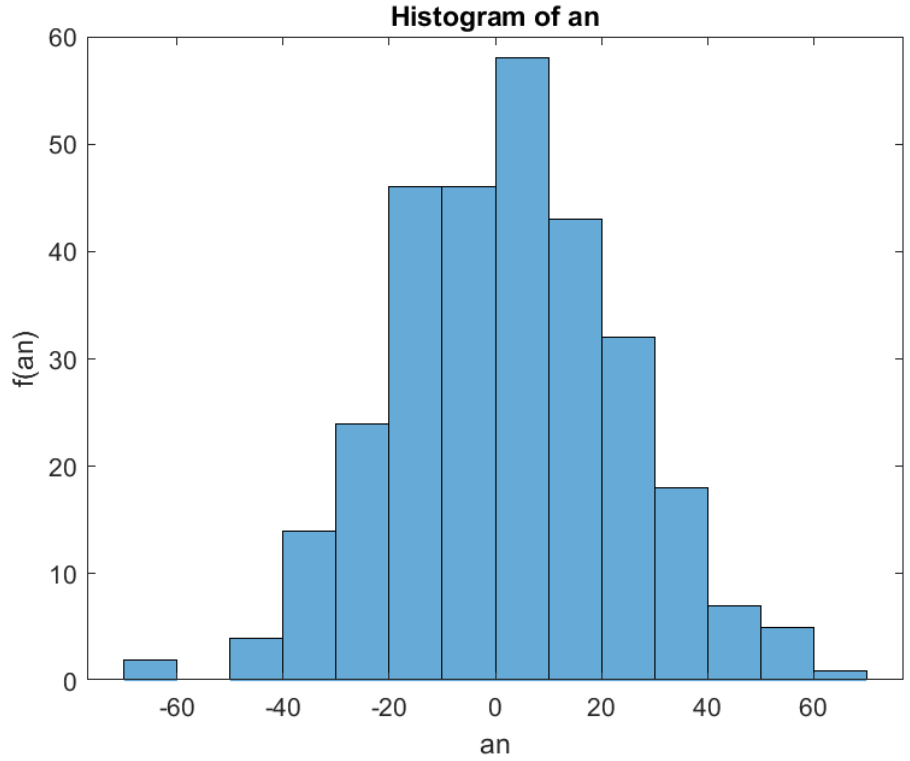
*Auto covariance of $u[K]$

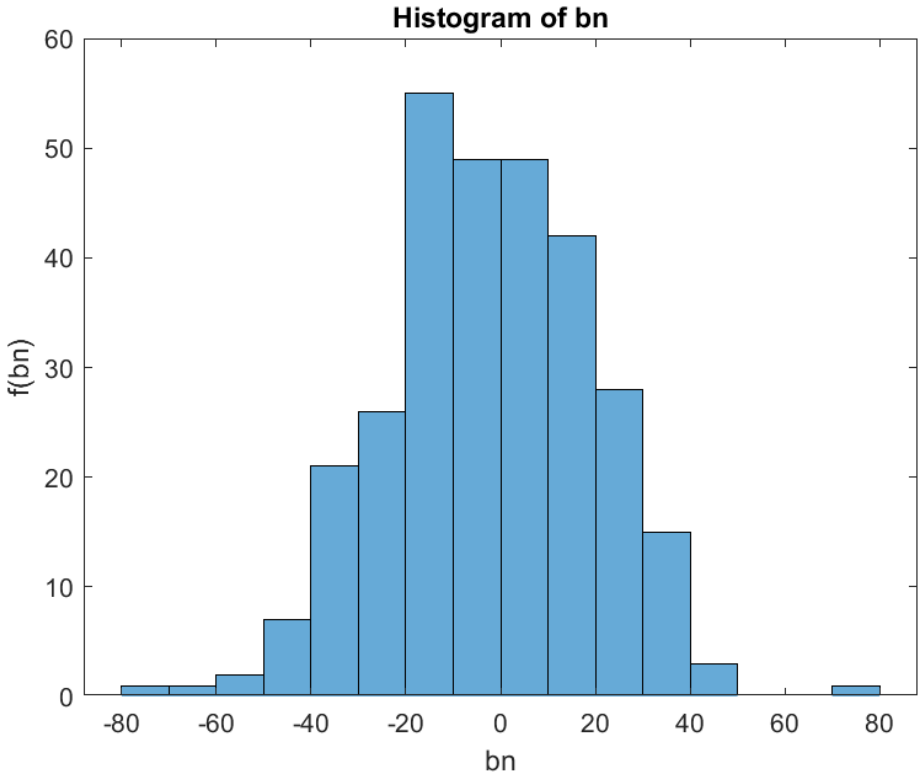
$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} u[K]\,u[K+l]\,f(u[K], u[K+l])\,du[K]\,du[K+l] = 0$$

III^{ly} for $w[K]$

Question 2(a)

The histograms of a_n and b_n is shown below, where n was chosen to be 12





These visualizations suggest that a_n and b_n are Gaussian distributed. This was further verified by performing Kolmogorov-Smirnov Test on the samples, where the null hypothesis (that the standardized version follows a standard normal distribution) was not rejected in both cases.

The details of the sample statistics are given below:

$$\overline{a_n} = 2.2657$$

$$\overline{b_n} = -2.0661$$

$$s_a^2 = 473.1549$$

$$s_b^2 = 464.7418$$

$$cov(a_n, b_n) = 4.9168$$

These estimates might not match with the theoretical values because of small number of samples and realizations.

Part (b)

$$\rho_{a_n b_n} = \frac{\text{cov}(a_n, b_n)}{\sigma_{a_n}\,\sigma_{b_n}} = \frac{E(a_n b_n) - E(a_n)E(b_n)}{\sigma_{a_n}\,\sigma_{b_n}}$$

We have seen, $E(a_n) = E(b_n) = 0$.

$$\Rightarrow \rho_{a_n b_n} = \frac{E(a_n b_n)}{\sigma_{a_n}\sigma_{b_n}}$$

$$a_n b_n = \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} V[i]V[j] \cos(2\pi f n i) \sin(2\pi f n j)$$

Again, $E\left(V[i]V[j] \cos(2\pi f n i) \sin(2\pi f n j)\right) = 0 \quad \forall\, i,j = 0 \text{ to } N-1$

$$\Rightarrow E(a_n b_n) = \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} E\left(V[i]V[j] \cos(2\pi f n i) \sin(2\pi f n j)\right)$$

$$= 0$$

Hence, the correlation between a_n and b_n

$$\rho_{a_n b_n} = 0$$

Part (c)

We are given

$$P(f_n) = \frac{a_n^2 + b_n^2}{N}$$

Since, a_n, b_n are normally distributed,
then a_n^2, b_n^2 are χ^2 distributed (when standardized)

As, $C_n = \frac{2P_{vv}(f_n)}{\gamma_{vv}(f_n)}$

C_n is also χ^2 distributed ~~(non-standard)~~

$$\gamma_{vv}(f_n) = \sum_{l=-\infty}^{\infty} \sigma_{vv}[l] \exp(-j2\pi f l)$$

Since $\sigma_{vv}[l] = 0 \quad \forall\, l \neq 0$ (VEK is GWN)

$$\gamma_{vv}(f_n) = 1$$

$$\Rightarrow C_n = 2P_{vv}(f_n)$$

$$E(C_n) = 2E(P_{vv}(f_n)) = 2E\left(\frac{a_n^2 + b_n^2}{N}\right)$$

$$= \frac{2}{N}\left(E(a_n^2) + E(b_n^2)\right)$$

$$= \frac{2}{N}\left(\text{var}(a_n) + \text{var}(b_n)\right)$$

$$= \frac{2}{N}\left(\sum_{K=0}^{N-1}\cos^2(2\pi f_n K) + \sum_{K=0}^{N-1}\sin^2(2\pi f_n K)\right)$$

$$= \frac{2}{N}(N) = 2$$

Hence, mean of $e_{1n} = E(e_{1n}) = 2$

For a χ^2 distribution, since $E(a_n) = E(b_n) = 0$

$$\text{var}(e_{1n}) = 2E(e_{1n}) = 4$$

Hence, the variance of $e_{1n} = E(e_{1n}^2) - E(e_{1n})^2 = 4$

Part (d)

We know that, the true value of PSD

$$\gamma_{vv}(f) = 1$$

$P(f_n)$ is mean square consistent iff

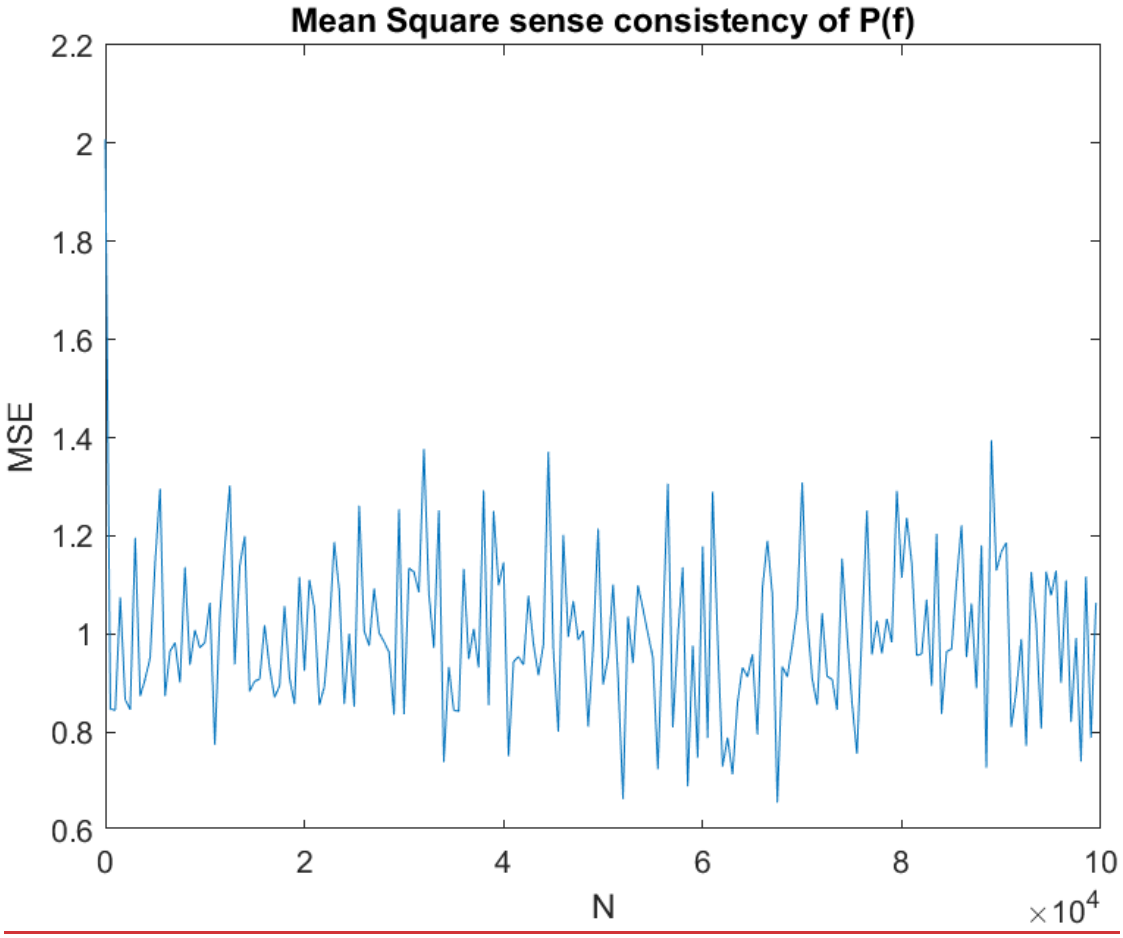
$$\lim_{N\to\infty} E\left((P(f_n) - 1)^2\right) = 0$$

$$E\left((P(f_n) - 1)^2\right) = E(P_n^2) - 1$$

$$= \text{var}(P_n) = \frac{\text{var}(e_{1n})}{4} = 1 \neq 0$$

Therefore, since the MSE is 1 irrespective of 1, $P(f_n)$ is not a consistent estimator of $\gamma_{vv}(f_n)$ in mean square sense.

Question 2(d)



It can be observed that the MSE of estimator does not converge to 0, and hence the estimator $P(f_n)$ is not consistent for estimating $\gamma_{vv}(f_n)$. This is in accordance with what was proved theoretically.

Question 3

Part (a)

The random process is given by,

$$y[k] = A\sin(2\pi f_0 k) + e[k], \quad e[k] \sim \mathcal{N}(0, \sigma_e^2)$$

where both A (amplitude) and frequency (f_0) are unknown.

First, we determine Fisher information of A and f_0.

Pdf of $y_N \rightarrow$ N observations

$$f(y_N; A, f_0) = \frac{1}{(2\pi\sigma_e^2)^{N/2}} \exp\left(-\frac{1}{2} \sum_{k=0}^{N-1} \frac{(y[k] - A\sin 2\pi f_0 k)^2}{\sigma_e^2}\right)$$

$$= l(A, f_0; y_N) \rightarrow \text{likelihood}.$$

Loglikelihood

$$L(A, f_0; y_N) = -\frac{N}{2} \ln(2\pi\sigma_e^2) - \frac{1}{2} \sum_{k=0}^{N-1} \frac{(y[k] - A\sin(2\pi f_0 k))^2}{\sigma_e^2}$$

Score of A

$$S_A = \frac{\partial L}{\partial A} = \frac{\sum_{k=0}^{N-1} (y[k] - A\sin(2\pi f_0 k))\sin(2\pi f_0 k)}{\sigma_e^2}$$

$$\frac{\partial S_A}{\partial A} = -\frac{\sum_{k=0}^{N-1} \sin^2(2\pi f_0 k)}{\sigma_e^2}$$

$$\Rightarrow I(A) = -E\left(\frac{\partial S_A}{\partial A}\right) = -E\left(-\frac{\sum_{k=0}^{N-1}\sin^2(2\pi f_0 k)}{\sigma_e^2}\right)$$

$$I(A) = \frac{\sum_{k=0}^{N-1}\sin^2(2\pi f_0 k)}{\sigma_e^2}$$

Score of f_0

$$S_{f_0} = \frac{\partial L}{\partial B}\frac{\partial L}{\partial f_0} = \frac{2\pi\sum_{k=0}^{N-1} Ak\cos(2\pi f_0 k)\left(y[k] - A\sin(2\pi f_0 k)\right)}{\sigma_e^2}$$

$$\frac{\partial S_{f_0}}{\partial f_0} = \frac{-4\pi^2 A}{\sigma_e^2}\sum_{k=0}^{N-1} k^2\left(y[k]\sin(2\pi f_0 k) + A\cos^2(2\pi f_0 k) - A\sin^2(2\pi f_0 k)\right)$$

$$I(f_0) = -E\left(\frac{\partial S_{f_0}}{\partial f_0}\right) = \frac{4\pi^2 A^2}{\sigma_e^2}\sum_{k=0}^{N-1} k^2\cos^2(2\pi f_0 k)$$

Score function corresponding to cross-derivative ~~of the parame~~.

$$S_{Af_0} = \frac{\partial L}{\partial A \partial f_0} = \frac{2\pi}{\sigma_e^2}\sum_{k=0}^{N-1} k\left(y[k]\cos(2\pi f_0 k) - 2A\sin(2\pi f_0 k)\cos(2\pi f_0 k)\right)$$

$$I(A, f_0) = -E\left(S_{Af_0}\right) = \frac{2\pi A}{\sigma_e^2}\sum_{k=0}^{N-1} k\sin(2\pi f_0 k)\cos(2\pi f_0 k)$$

Fisher Information matrix

$$I(\theta) = \begin{bmatrix} I(A) & I(A, f_0) \\ I(A, f_0) & I(f_0) \end{bmatrix}$$

$$\therefore (I(\theta))^{-1} = \frac{1}{I(A)\,I(f_0) - I(A,f_0)^2} \begin{bmatrix} I(f_0) & -I(A,f_0) \\ -I(A,f_0) & I(A) \end{bmatrix}$$

$$CRLB(A) = (I(\theta))^{-1}_{1,1}$$

$$= \frac{I(f_0)}{I(A)\,I(f_0) - I(A,f_0)^2}$$

$$= \frac{\sigma_e^2 \sum_{K=0}^{N-1} K^2 \cos^2(2\pi f_0 K)}{\sum_{K=0}^{N-1} \sin^2(2\pi f_0 K) \sum_{K=0}^{N-1} K^2 \cos^2(2\pi f_0 K) - \left(\sum_{K=0}^{N-1} K \sin(2\pi f_0 K)\cos(2\pi f_0 K)\right)^2}$$

Similarly,

$$CRLB(f_0) = (I(\theta)^{-1})_{2,2}$$

$$= \frac{I(A)}{I(A)\,I(f_0) - I(A,f_0)^2}$$

$$= \frac{\sigma_e^2}{4\pi^2 A^2} \frac{\sum_{K=0}^{N-1} \sin^2(2\pi f_0 K)}{\sum_{K=0}^{N-1} \sin^2(2\pi f_0 K) \sum_{K=0}^{N-1} K^2 \cos^2(2\pi f_0 K) - \left(\sum_{K=0}^{N-1} K \sin(2\pi f_0 K)\cos(2\pi f_0 K)\right)^2}$$

However, when ~~only~~ one of the parameters is known (fixed) and the the other is unknown, the CLRB of the estimator of the unknown parameter is only dependent on the Fisher Information of that parameter. Here,

$$CRLB(A) = (I(A))^{-1} \text{ and } CRLB(f_0) = (I(f_0))^{-1}$$

In the case when A is unknown and f_0 is known,

$$CRLB(A) = I(A)^{-1} = \frac{\sigma_e^2}{\sum_{K=0}^{N-1} \sin^2(2\pi f_0 K)}$$

When f_0 is unknown ~~& A~~ and A is known,

$$CRLB(f_0) = I(f_0)^{-1} = \frac{\sigma_e^2}{4\pi A^2 \sum_{K=0}^{N-1} K^2 \cos^2(2\pi f_0 K)}$$

It can be noticed that the CLRBs for amplitude and frequency estimates in the case when both the parameters are unknown, are different from that when derived in the case when one of the parameters is known. As both parameters are unknown, the corresponding to variance lower bound for each parameter is highly dependent on the other parameter as well.

Part (b)

Given GWN process

$$y[K] = e[K] \quad , \quad e[K] \sim \mathcal{N}(0, \sigma_e^2)$$

we can rewrite it as

$$y[K] = \sigma\, u[K], \quad u[K] \sim \mathcal{N}(0, 1)$$

In such a setting, the parameter σ which is to be estimated is associated with the noise and is hence not linear, i.e. it cannot be expressed as $y = l\sigma + v$.

This suggests that BLUE estimator cannot be designed directly for this process. However, we can log-transform the data to obtain a linear relationship in σ.

Let

$$z[k] = \log(|y[k]|) = \log(|\sigma u[k]|)$$

$$= \log\sigma + \log(|u[k]|)$$

Here, we take the absolute value and log transform the data.

$$z[k] = \sigma' + v[k]$$

where $\sigma' = \log(\sigma)$, $v[k] = \log(|u[k]|)$

Also,

$$E(v[k]) = E(\log(|u[k]|)) = \int_0^\infty \log(u[k]) \times N(0,1)\, du$$

$$= \int_0^\infty \log u \left(\frac{2}{\sqrt{2\pi}} \exp\left(\frac{-u^2}{2}\right)\right) du$$

$$= -0.6352 \neq 0$$

(integration performed using MATLAB).

To correct this nonzero expectation of noise, we define another noise process, $w[k] = v[k] + 0.6352$, such that $E(w[k]) = 0$. Therefore,

$$z[k] = \sigma'' + w[k]$$

where $\sigma'' = \sigma' - 0.6352$

Rewriting the process in vector notation,

$$\underset{N\times 1}{Z} = \underset{N\times 1}{l}\,\sigma'' + \underset{N\times 1}{\omega}$$

where $l = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}_{N\times 1}$

Constraint : $a^T l = 1$

Objective function: $\min_a a^T \Sigma_\omega a$,

$= \min_a a^T a$, as ~~$\Sigma_\omega = I$~~

Solving, we get.

$$\hat{a}^* = (l^T \Sigma_\omega^{-1} l)^{-1} l^T \Sigma_\omega^{-1}$$

and $$\hat{\sigma}''_{BLUE} = \hat{a}^{*T} Z = \left(\frac{N}{\sigma_\omega^2}\right)^{-1} (\sigma_\omega^2) \sum_{K=0}^{N-1} Z[K]$$

$$\hat{\sigma}''_{BLUE} = \frac{\sum_{K=0}^{N-1} Z[K]}{N}$$

$$\hat{\sigma}'_{BLUE} = \hat{\sigma}''_{BLUE} + 0.6352$$

$$= \frac{\sum_{K=0}^{N-1} Z[K]}{N} + 0.6352$$

$$\hat{\sigma}_{BLUE} = e^{\hat{\sigma}'_{BLUE}} = \exp\left(\frac{\sum_{K=0}^{N-1} \log(y[K])}{N} + 0.6352\right)$$

Question 4

Part (a)

Given $N = 100$ (number of samples)

$m = 14578$ (sample mean)

$S = 1845$ (sample standard deviation)

Let the true mean be μ

To determine $P_r(12000 < \mu < 16000)$, we can determine $P_r(\mu < 16000) - P_r(\mu < 12000)$.

To

We construct two t statistics t_1 and t_2, where

$$t_1 = \frac{12000 - m}{s/\sqrt{N}} = \frac{12000 - 14578}{1845/10} = -13.9729$$

$$t_2 = \frac{16000 - m}{s/\sqrt{N}} = \frac{16000 - 14578}{1845/10} = 7.7073$$

and we assume

$$T_t \sim t_{dist}(\nu), \text{ where } \nu = N - 1 = 99$$

(degrees of freedom)

$$Pr(12000 > M) = t\,cdf(t_1, \nu)$$
$$= t\,cdf(-13.9729, 99)$$
$$= 1.87 \times 10^{-25}$$
$$\approx 0.$$

$$Pr(16000 > M) = t\,cdf(t_2, \nu)$$
$$= t\,cdf(7.7073, 99)$$
$$= 99.999\ldots$$
$$\approx 1$$

Hence, $Pr(12000 < M < 16000) = 1$

Therefore, with 100% confidence, the we can say that the average molecular weight of the polymer lies between 12000 and 16000.

Part (b)

Sample mean of institution A = $\bar{x}_1 = 85.2$

$n_1 = 60$, $S_1 = 6.8$

Sample mean of institution B = $\bar{x}_2 = 87.2$

$n_2 = 55$, $S_2 = 8.8$

One tailed 2-Sample hypothesis test with unknown standard deviation.

We choose $\alpha = 0.05$, (95% confidence).

Let μ_1 and μ_2 be the true means of institutions A and B respectively.

$$H_0 : \mu_1 = \mu_2 \quad (\text{Precisely, } \mu_1 \leq \mu_2)$$

$$H_a : \mu_1 > \mu_2$$

$$t_s = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} = -1.355$$

We reject null hypothesis when

$$t_s > t_{1-\alpha, \nu}$$

where $\nu = \dfrac{(S_1^2/n_1 + S_2^2/n_2)^2}{(S_1^2/n_1)^2/(n_1-1) + (S_2^2/n_2)^2/(n_2-1)} = 101.469$

$t_{1-\alpha, \nu} = 1.66$

As $t_s < t_{1-\alpha, v}$, we do not reject the null hypothesis, and thereby with 95% confidence, we reject the claim that students from elite institution A perform better than students from elite institution B. In other words, we fail to reject the null hypothesis with 95% confidence.

Question 5

Part (a)

From Q1, we have

$$f(y) = \begin{cases} e^{-(y-\theta)} & , y > \theta \\ \\ 0 & , \text{otherwise} \end{cases}$$

For N observations y_N

$$f(y_N) = \begin{cases} e^{-\sum_{k=0}^{N-1}(y[k]-\theta)} & , y[k] > \theta \quad \forall k \\ \\ 0 & , \text{otherwise} \end{cases}$$

Consider step function,

$$H(x) = \begin{cases} 1 & , x \geq 0 \\ 0 & , \text{otherwise} \end{cases}$$

We can rewrite the pdf of y_N as

$$f(y_N;\theta) = H(\min(y_N) - \theta)\, e^{-\sum_{k=0}^{N-1}(y[k]-\theta)}$$

$$= H\left(\frac{T_N}{2} - \theta\right) e^{-\sum_{k=0}^{N-1}(y[k]-\theta)}$$

$$= \phi(T_N;\theta)$$

$$= H\left(\frac{T_N}{2} - \theta\right) e^{N\theta} \cdot e^{-\sum_{k=0}^{N-1} y[k]}$$

$$\phi(T_N;\theta) = H\left(\frac{T_N}{2}-\theta\right)e^{N\theta}$$

$$K(y_N) = e^{-\sum_{k=0}^{N-1} y[k]}$$

$$\Rightarrow f(y_N;\theta) = \phi(T_N;\theta)\,K(y_N)$$

Hence, by Neyman-Fisher factorization theorem, we can say T_N is a sufficient statistic for estimating θ

However, as we saw in Q1.b), the estimator T_N is biased with $E(T_N) = 2\theta + \frac{2}{N}$.

To correct the bias, we derived T_N' as

$$T_N' = \frac{T_N}{2} - \frac{1}{N} = \min(y_N) - \frac{1}{N}.$$

Therefore, we can say that the estimator T_N' is sufficient, unbiased and hence complete.

By Rao-Blackwell theorem, T_N' is the MVUE estimator of θ,

$$\hat{\theta}_{MVUE} = T_N' = \min(y_N) - \frac{1}{N}$$

Question 5 (b)

Part (b)

We derived that $T_N' = \min(y_N) - \frac{1}{N}$ is the minimum unbiased estimator for θ.

$$\text{Var}(\hat{\theta}^*) = \text{Var}(T_N') = E(T_N'^2) - E(T_N')^2$$

$$= \left[\left(\theta - \frac{1}{N}\right)^2 + \frac{2\theta}{N}\right] - \theta^2$$

$$= \theta^2 + \frac{1}{N^2} + \frac{2\theta}{N} - \theta^2$$

$$= \frac{1}{N}\left(2\theta + \frac{1}{N}\right) = \frac{1}{N^2}$$

$$= \left(\frac{1}{N} + 1\right)^2 - 1$$

$\Rightarrow \text{var}(\hat{\theta}^*) = 1/N^2$

For the purpose, $\text{var}(\hat{\theta}^*)$ was assigned to be the variance of T_N' across R realizations.

It was therefore shown that efficiency

$$\eta_{\hat{\theta}_1} = \frac{\text{var}(\hat{\theta}^*)}{\text{var}(\hat{\theta}_1)} = 9.2604 \times 10^{-4}$$

$$\eta_{\hat{\theta}_2} = \frac{\text{var}(\hat{\theta}^*)}{\text{var}(\hat{\theta}_2)} = 9.3133 \times 10^{-4}$$

Where $\hat{\theta}_1 = \bar{y} - 1$, $\hat{\theta}_2 = \tilde{y} - \ln 2$

Since $\eta_{\hat{\Theta}_2} > \eta_{\hat{\Theta}_1}$, $\hat{\Theta}_2$ is more efficient than $\hat{\Theta}_1$.

NOTE: This result differed each time the simulation was performed. A seed value of 42 was therefore set for this simulation.

Question 5(b)

$$\widehat{\theta_1} = \bar{y} - 1$$

$$\widehat{\theta_2} = \tilde{y} - \ln(2)$$

With a random state of 42, the efficiencies of both estimators were calculated as:

$$\eta_{\theta_1} = \frac{var(T_N')}{var(\widehat{\theta_1})} = 9.2604 \times 10^{-4}$$

$$\eta_{\theta_2} = \frac{var(T_N')}{var(\widehat{\theta_2})} = 9.3133 \times 10^{-4}$$

It can be observed that $\widehat{\theta_2}$ is a comparatively more efficient estimator than $\widehat{\theta_1}$