Assignment 1: CS 215

Mahant Sakhare - 24B0956, Hitansh Baria - 24B1075, Harshal Walke - 24B0954

1. **Plot with 30% corruption in sine wave**

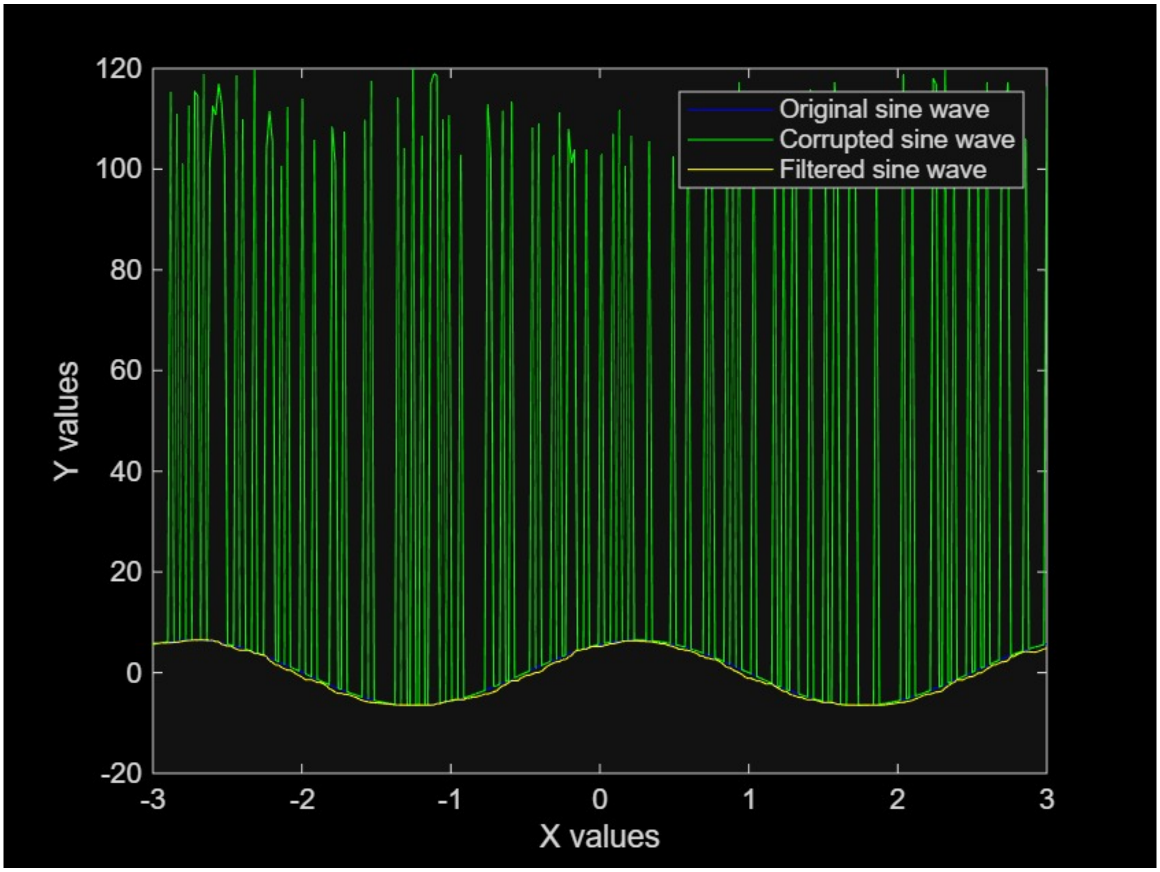

Figure 1: Plot with 30% corruption in sine wave.

Relative Mean Squared Errors:

- MSE between Original and Corrupted wave: 169.2203
- MSE between Original and Median Filtered wave: 23.7745
- MSE between Original and Mean Filtered wave: 57.1519
- MSE between Original and Quartile Median Filtered wave: 0.0138

Plot with 60% corruption in sine wave

Relative Mean Squared Errors:

- MSE between Original and Corrupted wave: 344.7812
- MSE between Original and Median Filtered wave: 388.8480
- MSE between Original and Mean Filtered wave: 213.2978
- MSE between Original and Quartile Median Filtered wave: 40.7685

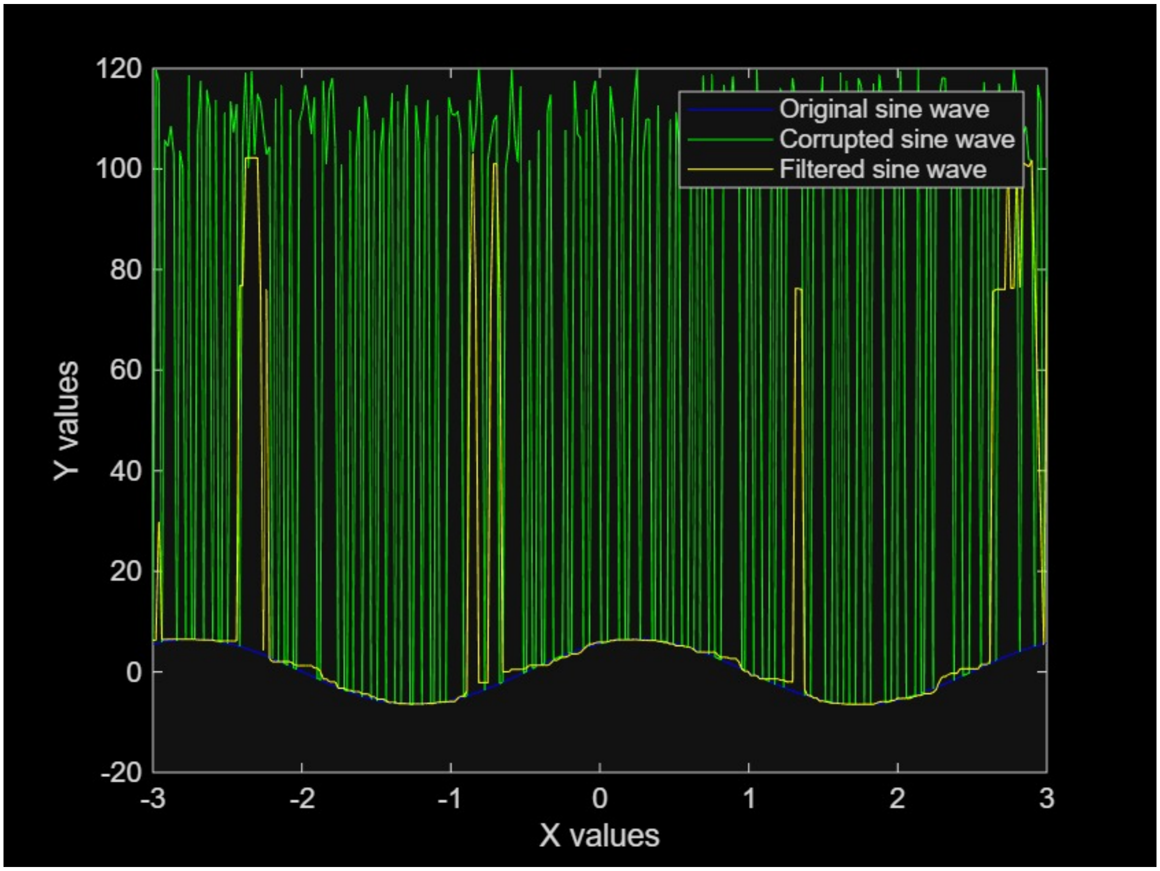

Figure 2: Plot with 60% corruption in sine wave.

Which of these method produces better relative mean squared error ?

Ans:

The quartile filter will produce the best result because we have added the positive noise to the wave function, making some samples artificially higher.

The mean filter, gets pulled upward by these high noisy values resulted in an elevated wave.

The median filter, which still sits in the middle of the sample window, leading to some unnatural spikes in the wave.

The 25% quartile filter deliberately picks a value from the lower end of the sorted neighbour. This allows it to ignore the upward noise spikes more effectively, preserving the true signal underneath and therefore producing the smallest relative mean squared error among the three filters.

2. **Mean :**

Given :
Old mean $= \mu$
Total no of points $= n$
New Data Point $= k$

Mean $= \frac{\text{Sum of Points}}{\text{Total no of points}}$
$\mu = \frac{\text{Sum of points}}{n}$
Sum of points $= \mu \times n$

Since we have a new data point, then
New Sum $= \mu \times n + k$
New Mean $= \frac{(\mu \times n)+k}{n+1}$

The formula for new mean is $\frac{(\mu \times n)+k}{n+1}$.

Median :

Given :
Old Median = M
New Data Point = k
Total no of points = n
Array of the Datapoints : A (A is sorted)

Case 1: When n is odd

The current median is:
$M = A\left(\frac{n+1}{2}\right)$

After insertion, the new size $n+1$ is **even** → the new median becomes the **average of the two middle values**:

If $k \leq A\left(\frac{n+1}{2}\right) = M$, insert k in the sorted array before or at the old median. Find its position. New median = average of elements at positions $\frac{n}{2}$ and $\frac{n}{2}+1$.

If $k > M$, insert k somewhere after the old median. New median = average of elements at positions $\frac{n}{2}+1$ and $\frac{n}{2}+2$.

Case 2: When n is even

The current median is the average of the two middle elements: $M = \frac{A\left(\frac{n}{2}\right)+A\left(\frac{n}{2}+1\right)}{2}$

After inserting the new data point k, the new total number of points $n+1$ is odd. The new median will be the single element at position $\frac{n+2}{2}$ in the newly sorted array. There are three possibilities for the new median:

- If $k \leq A\left(\frac{n}{2}\right)$: The new value k is inserted at or before the first of the two original middle elements. The element originally at position $\frac{n}{2}$ is shifted to the new median position of $\frac{n+2}{2}$. The new median is the original element $A\left(\frac{n}{2}\right)$.
- If $A\left(\frac{n}{2}\right) < k \leq A\left(\frac{n}{2}+1\right)$: The new value k is inserted between the two original middle elements. This places k directly at the new median position of $\frac{n+2}{2}$. The new median is k.
- If $k > A\left(\frac{n}{2}+1\right)$: The new value k is inserted after the second of the two original middle elements. The element at position $\frac{n}{2}+1$ is now at the new median position of $\frac{n+2}{2}$. The new median is the original element $A\left(\frac{n}{2}+1\right)$.

Standard Deviation :

Given:
Old mean: μ
Old standard deviation: σ
Old number of elements: n
New data value: k
New mean (after adding k): μ^*

Step 1: The variance (σ^2) of n values $A_1, A_2, \ldots, A_n$ is:

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

Step 2: After adding the new value k, the new mean is $\mu^* = \frac{n\mu+k}{n+1}$. The new variance is:

$$\sigma^{*2} = \frac{\sum_{i=1}^{n+1}(A_i - \mu^*)^2}{n} = \frac{\sum_{i=1}^{n}(A_i - \mu^*)^2 + (k - \mu^*)^2}{n}$$

Step 3: Using the identity $A_i - \mu^* = (A_i - \mu) + (\mu - \mu^*)$, we square it:

$$(A_i - \mu^*)^2 = (A_i - \mu)^2 + 2(A_i - \mu)(\mu - \mu^*) + (\mu - \mu^*)^2$$

Step 4: Summing the expansion from Step 3:

$$\sum_{i=1}^{n}(A_i - \mu^*)^2 = \sum_{i=1}^{n}(A_i - \mu)^2 + 2(\mu - \mu^*)\sum_{i=1}^{n}(A_i - \mu) + \sum_{i=1}^{n}(\mu - \mu^*)^2$$

Step 5: Since $\sum_{i=1}^{n}(A_i - \mu) = 0$ (property of the mean), the middle term vanishes:

$$\sum_{i=1}^{n}(A_i - \mu^*)^2 = \sum_{i=1}^{n}(A_i - \mu)^2 + n(\mu - \mu^*)^2$$

Step 6: The total sum of squares for all $n+1$ values is:

$$\sum_{i=1}^{n+1}(A_i - \mu^*)^2 = \sum_{i=1}^{n}(A_i - \mu)^2 + n(\mu - \mu^*)^2 + (k - \mu^*)^2$$

Step 7: Replace the old sum with σ^2. From Step 1, we know $\sum_{i=1}^{n}(A_i - \mu)^2 = (n-1)\sigma^2$. Substituting this in gives:

$$\sum_{i=1}^{n+1}(A_i - \mu^*)^2 = (n-1)\sigma^2 + n(\mu - \mu^*)^2 + (k - \mu^*)^2$$

Step 8: Divide by n to get the new variance:

$$\sigma^{*2} = \frac{(n-1)\sigma^2 + n(\mu - \mu^*)^2 + (k - \mu^*)^2}{n}$$

Step 9: First, we simplify the terms $(\mu - \mu^*)$ and $(k - \mu^*)$:

$$\begin{aligned}
\mu - \mu^* &= \mu - \frac{n\mu + k}{n+1} = \frac{(n+1)\mu - (n\mu + k)}{n+1} = \frac{\mu - k}{n+1} \\
k - \mu^* &= k - \frac{n\mu + k}{n+1} = \frac{(n+1)k - (n\mu + k)}{n+1} = \frac{nk - n\mu}{n+1} = \frac{-n(\mu - k)}{n+1}
\end{aligned}$$

Step 10: Next, we simplify the sum of their squared components from the numerator:

$$\begin{aligned}
n(\mu - \mu^*)^2 + (k - \mu^*)^2 &= n\left(\frac{\mu - k}{n+1}\right)^2 + \left(\frac{-n(\mu - k)}{n+1}\right)^2 \\
&= \frac{n(\mu - k)^2}{(n+1)^2} + \frac{n^2(\mu - k)^2}{(n+1)^2} \\
&= \frac{(n + n^2)(\mu - k)^2}{(n+1)^2} = \frac{n(n+1)(\mu - k)^2}{(n+1)^2} = \frac{n(\mu - k)^2}{n+1}
\end{aligned}$$

Step 11: Substitute the result from Step 10 back into the formula for σ^{*2}:

$$\sigma^{*2} = \frac{(n-1)\sigma^2 + \frac{n(\mu-k)^2}{n+1}}{n} = \frac{n-1}{n}\sigma^2 + \frac{(\mu - k)^2}{n+1}$$

The updated standard deviation is the square root of this variance:

$$\sigma^* = \sqrt{\frac{n-1}{n}\sigma^2 + \frac{(\mu - k)^2}{n+1}}$$

How to Update the Histogram of A, if new value is added to A

Ans:

A histogram divides the range of the data into intervals called bins. When a new value k arrives, you determine which bin it belongs to by checking which interval contains k.

This step ensures that the new data is properly categorized according to the histogram structure. After identifying the correct bin, you increase the count of that bin by 1.

This updates the histogram to reflect the presence of the new value. Incrementing only the relevant bin allows the histogram to be updated efficiently without recomputing the counts for all data points.

3. Given two events A and B such that

$$P(A) \geq 1 - q_1 \quad \text{and} \quad P(B) \geq 1 - q_2,$$

We want to prove:

$$P(A \cap B) \geq 1 - (q_1 + q_2).$$

We know that,

$$P(A^c \cup B^c) \leq P(A^c) + P(B^c).$$

Since $(A \cap B)^c = A^c \cup B^c$, we have

$$1 - P(A \cap B) \leq (1 - P(A)) + (1 - P(B)).$$

Rearranging:

$$\begin{aligned} P(A \cap B) &\geq P(A) + P(B) - 1 \\ &\geq (1 - q_1) + (1 - q_2) - 1 \\ &= 1 - (q_1 + q_2). \end{aligned}$$

$$\therefore \quad P(A \cap B) \geq 1 - (q_1 + q_2).$$

4. Let A be the event that the bus is red, and let E be the event that XYZ sees red objects as red.

We are given:

$$P(A) = 0.01, \quad P(A^c) = 0.99,$$
$$P(E \mid A) = 0.99, \quad P(E \mid A^c) = 0.02.$$

From Bayes' rule,

$$P(A \mid E) = \frac{P(E \mid A)\, P(A)}{P(E \mid A)\, P(A) + P(E \mid A^c)\, P(A^c)}$$

Substituting the values:

$$P(A \mid E) = \frac{0.99 \times 0.01}{0.99 \times 0.01 + 0.02 \times 0.99} = \frac{0.0099}{0.0297} = \frac{1}{3} \approx 33.3\%.$$

Conclusion: There is only about a 33% chance that the bus was actually red, and a 67% chance that it was blue, given the witness's statement. Thus, the defense lawyer can argue that the eyewitness testimony alone is insufficient, as it is more likely that the bus was blue than red.

5. Given: 95% of the residents favour candidate A over B and the remaining 5% favour B over A.

For A to win, at least 2 voters out of the 3 chosen voters should vote for A.

$$\begin{aligned}\text{Accuracy} &= \frac{\text{No. of ways A wins in exit poll}}{\text{No. of exit poll}} \times 100 \\ &= \frac{(95)^2 \cdot 5 \cdot {}^3C_2 + (95)^3 \cdot {}^3C_3}{(100)^3} \times 100 \\ &= \frac{15 \times (95)^2 + (95)^3}{(100)^3} \times 100 \\ &= 99.275\%\end{aligned}$$

If there are 10000 villagers, 9500 villagers would vote for A and remaining 500 for B.

$$\begin{aligned}\text{Accuracy} &= \frac{(9500)^2 \cdot (500) \cdot {}^3C_2 + (9500)^3 \cdot {}^3C_3}{(10000)^3} \times 100 \\ &= \frac{(95)^2 \cdot 5 \cdot {}^3C_2 + (95)^3 \cdot {}^3C_3}{(100)^3} \times 100 \\ &= 99.275\%\end{aligned}$$

6. Given: There are m voters in the village and probability that the voters prefer A over B, $p = \frac{k}{m}$. S is a randomly chosen subset (with replacement) containing n truthful voters and $q(S)$ is the proportion of voters from S who voted for A out of $n = |S|$.

For the given ordered sample S, let s denote the number of positions in S that correspond to voters who voted for $A(0 < s \leq n)$.

$$\therefore \quad q(S) = \frac{s}{n}. \tag{1}$$

Number of ordered samples that have exactly s A-voters $= \binom{n}{s} k^s (m-k)^{n-s}$.

(a)

$$\sum_S q(S) = \sum_{s=0}^{n} \frac{s}{n} \binom{n}{s} k^s (m-k)^{n-s} = \sum_{s=1}^{n} \frac{s}{n} \binom{n}{s} k^s (m-k)^{n-s} \tag{2}$$

We know,

$$(x+y)^n = \sum_{s=0}^{n} \binom{n}{s} x^s y^{n-s}$$

Differentiating both sides w.r.t x:

$$n(x+y)^{n-1} = \sum_{s=1}^{n} \binom{n}{s} s x^{s-1} y^{n-s}$$

Replacing x with k and y with $(m-k)$:

$$n(k+(m-k))^{n-1} = \sum_{s=1}^{n} \binom{n}{s} s k^{s-1} (m-k)^{n-s}$$

$$\Rightarrow nm^{n-1} = \sum_{s=1}^{n} \binom{n}{s} s k^{s-1} (m-k)^{n-s} \tag{3}$$

From Eq. (2) and Eq. (3):

$$\sum_S q(S) = km^{n-1} \quad \Rightarrow \quad \frac{\sum_S q(S)}{m^n} = \frac{k}{m}$$

$$\Rightarrow \frac{\sum_s q(s)}{m^n} = p$$

—

(b) From Eq. (1),

$$q(S) = \frac{s}{n} \Rightarrow q^2(S) = \frac{s^2}{n^2}.$$

$$\therefore \quad \sum_S q^2(S) = \sum_{s=1}^{n} \frac{s^2}{n^2} \binom{n}{s} k^s (m-k)^{n-s} \tag{4}$$

We know,

$$(x+y)^n = \sum_{\delta=0}^{n} \binom{n}{\delta} x^{\delta} y^{n-\delta}$$

Differentiating both sides w.r.t. x:

$$n(x+y)^{n-1} = \sum_{\delta=1}^{n} \delta \binom{n}{\delta} x^{\delta-1} y^{n-\delta} \tag{5}$$

Differentiating both sides again w.r.t. x:

$$n(n-1)(x+y)^{n-2} = \sum_{\delta=2}^{n} \delta(\delta-1) \binom{n}{\delta} x^{\delta-2} y^{n-\delta} \tag{6}$$

Replacing x with k and y with $(m-k)$ in Eq. (5) and (6):

$$nm^{n-1} = \sum_{\delta=1}^{n} \delta \binom{n}{\delta} k^{\delta-1} (m-k)^{n-\delta} \tag{7}$$

$$n(n-1)m^{n-2} = \sum_{\delta=2}^{n} \delta(\delta-1) \binom{n}{\delta} k^{\delta-2} (m-k)^{n-\delta} \tag{8}$$

From Eq. (1):

$$\sum_S q^2(S) = \sum_{s=1}^{n} \frac{s^2}{n^2} \binom{n}{s} k^s (m-k)^{n-s}$$

$$\Rightarrow \sum_S q^2(S) = \frac{1}{n^2} \left[\sum_{s=1}^{n} s^2 \binom{n}{s} k^s (m-k)^{n-s} \right]$$

$$= \frac{1}{n^2} \left[\sum_{s=2}^{n} s(s-1) \binom{n}{s} k^s (m-k)^{n-s} + \sum_{s=1}^{n} s \binom{n}{s} k^s (m-k)^{n-s} \right]$$

Using Eq.(7) and (8), we get,

$$\sum_S q^2(S) = \frac{k^2(n-1)m^{n-2}}{n} + \frac{km^{n-1}}{n}$$

$$= k^2 \cdot \frac{(n-1)m^{n-2}}{n} + \frac{km^{n-1}}{n}$$

$$= m^n \left[\frac{k^2}{m^2} \cdot \frac{(n-1)}{n} + \frac{1}{n} \cdot \frac{k}{m}\right]$$

$$= m^n \left[p^2 \cdot \frac{(n-1)}{n} + \frac{1}{n} \cdot p\right]$$

$$\Rightarrow \sum_S \frac{q^2(S)}{m^n} = \frac{p^2(n-1)}{n} + \frac{p}{n}$$

—

(c)

$$\sum_S (q(S) - p)^2 = \sum_S q^2(S) - 2p \sum_S q(S) + p^2 \sum_S 1$$

$$= m^n \left[\frac{p^2(n-1)}{n} + \frac{p}{n}\right] - 2p \cdot m^n \cdot p + p^2 \cdot m^n$$

$$= m^n \left[\frac{p^2(n-1)}{n} + \frac{p}{n} - 2p^2 + p^2\right]$$

$$\Longrightarrow \quad \frac{\sum_S (q(S) - p)^2}{m^n} = p^2 - \frac{p^2}{n} + \frac{p}{n} - 2p^2 + p^2$$

$$\Longrightarrow \sum_S \frac{(q(S) - p)^2}{m^n} = \frac{p(1-p)}{n}$$

—

(d) Let A denote the collection of subsets S of size n such that the sample proportion deviates from the true mean by more than δ:

$$A = \{S : |q(S) - p| > \delta\}.$$

From part (c), we know that

$$\text{Var}(q(S)) = \frac{1}{m^n} \sum_S (q(S) - p)^2 = \frac{p(1-p)}{n}.$$

By **Chebyshev's inequality**, for any random variable X with mean μ and variance σ^2:

$$\Pr\big(|X - \mu| > \delta\big) \le \frac{\sigma^2}{\delta^2}.$$

Here $X = q(S)$, $\mu = p$, and $\sigma^2 = \frac{p(1-p)}{n}$. Thus,

$$\Pr\big(|q(S) - p| > \delta\big) \le \frac{p(1-p)}{n\delta^2}.$$

Equivalently, since this probability is the proportion of subsets out of the total m^n,

$$\frac{|A|}{m^n} \le \frac{1}{\delta^2} \frac{p(1-p)}{n}.$$

Significance: This result means that the chance of the sample proportion $q(S)$ being far from the true proportion p becomes very small as the sample size n increases. Thus, larger polls give more reliable estimates of the true preference.

Running the Code for Question 1

1. Open MATLAB.
2. Ensure that the file `Question1.m` and `Question1b.m` is in the current MATLAB directory or path.
3. In the MATLAB command window, type `Question1` and press Enter to generate filtered sine wave for 30% noise.
4. Similarly, type `Question1b` and press Enter to generate filtered sine wave for 60% noise.
5. The script will execute, displaying the RMSE values for each filter in the command window.

Running the Code for Question 2

1. `Question2.m` contains only the functions specified in problem statement. We can't directly run this file.