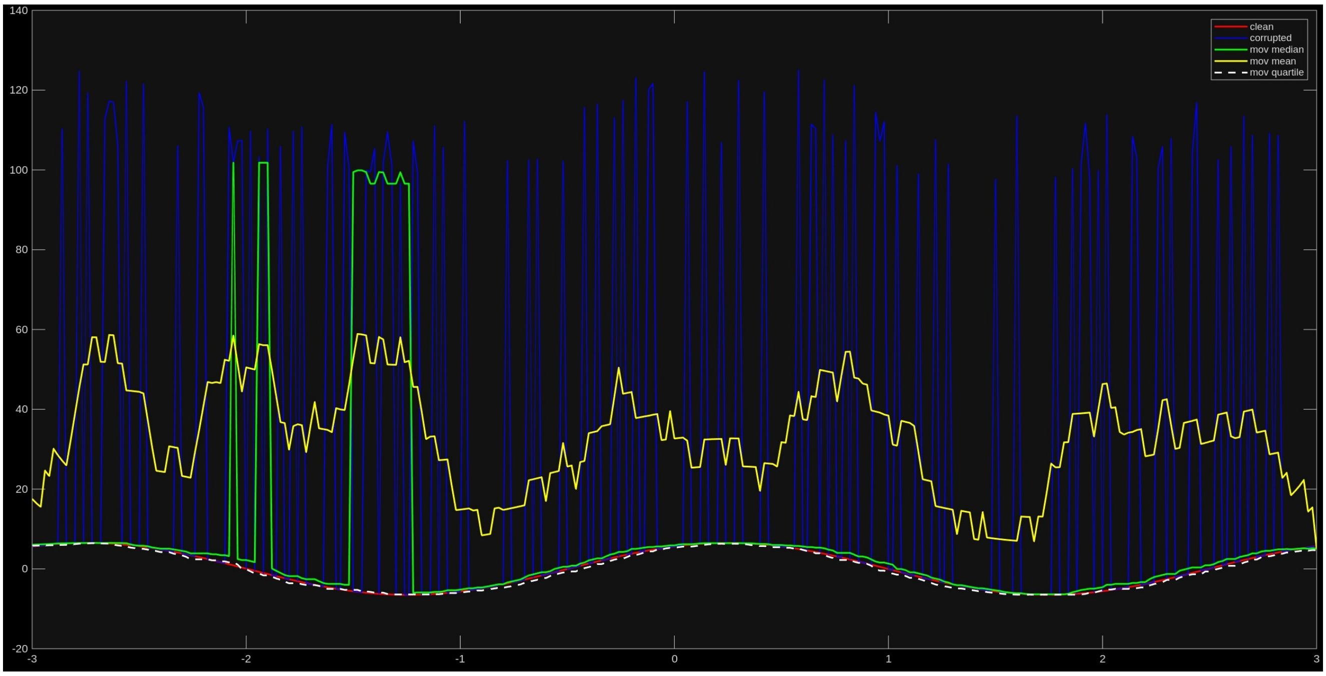
CS 215 — Assignment 1

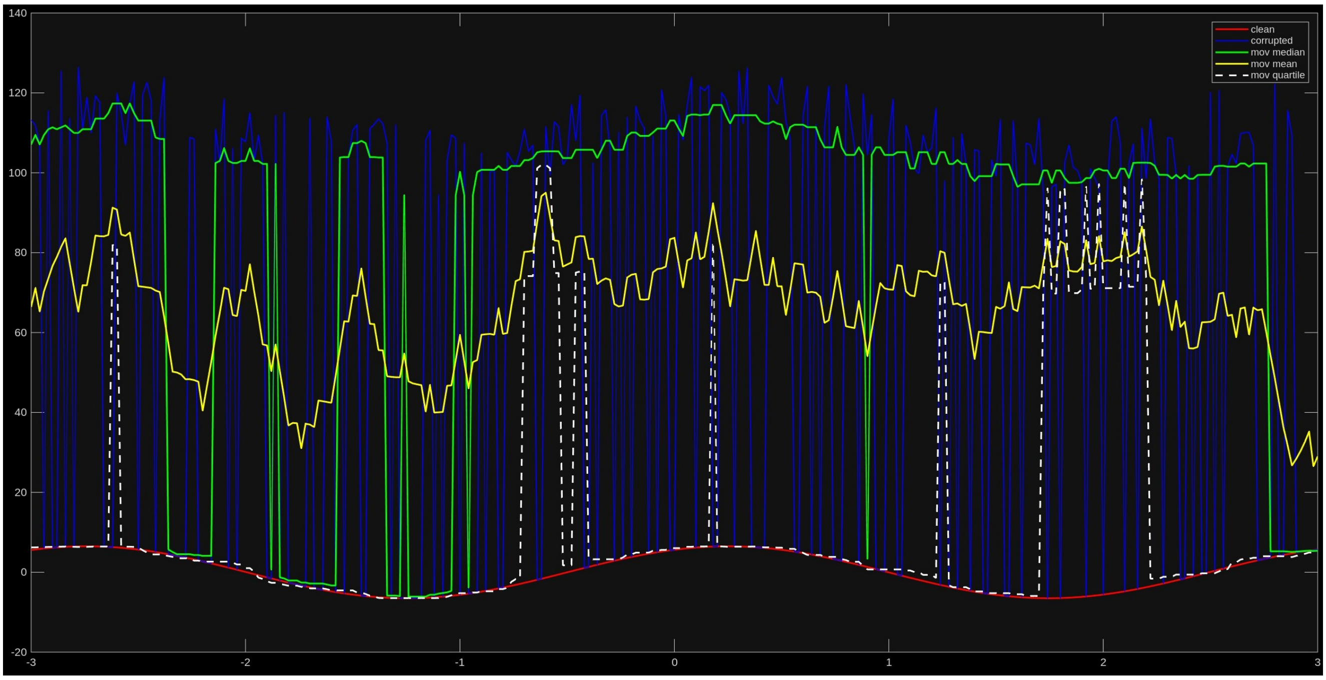
Aradhana R (24b1006)   Suhas Alladaboina (24b1009)

1: Filtering

Plot of 30% corrupted values in the array y (i.e $f = 30\%$) :



Plot of 60% corrupted values in the array y (i.e $f = 60\%$) :



How to run the code

The clean sine wave, the corrupted sine wave, and the three filtered sine waves corresponding to $f = 30\%$ and $f = 60\%$ are plotted in the same figure. The implementation is in the file `q1.m`, which can be executed to generate the plots and compute the three RMSE values for each case. The code can be executed by running `q1`.

Relative Mean Squared Error Values:

Fraction f	First Quartile	Median	Mean
30%	0.0122	30.4611	59.2474
60%	46.3620	426.4590	216.6413

The quartile filter performs better than both the median and mean filters because of how it selects the filtered value from the 17 neighboring values. For the first quartile filter, the values are sorted, and the element at the 25th percentile is chosen. Since the corrupted values are much higher than the original sine wave values, the first quartile will, in most cases for low corruption levels (e.g., f=30%), lie much closer to the true sine value. However, when the corruption increases to f=60%, more corrupted values appear within the 17-sample window, so the quartile estimate becomes less accurate and the RMSE increases.

The median filter also sorts the 17 neighboring values but takes the middle element. For low corruption, this mostly produces a value close to the sine wave, but as corruption increases, the median is more likely to be skewed by corrupted samples, reducing accuracy. The quartile filter generally performs better than the median filter because its position in the sorted list is less likely to be corrupted than the median position.

The mean filter is very sensitive to outliers, as every corrupted sample affects the average. For low corruption, this causes the mean filter to perform worse than both the quartile and median filters. In the high corruption case, the mean may sometimes produce slightly smaller errors than the median, since it averages both corrupted and uncorrupted values, but it will still be heavily biased due to large outliers.

2: Updating Mean/Median/Std

Formulas and Derivations

New Mean:

$$\text{New Mean} = \frac{\text{Old Mean} \times n + \text{New Data}}{n+1} = \frac{\text{Old Mean}(n+1)}{n+1} + \frac{\text{New Data} - \text{Old Mean}}{n+1}$$

$$\text{New Mean} = \text{Old Mean} + \frac{\text{New Data} - \text{Old Mean}}{n+1}$$

New Std Deviation:

$$\text{Old Std}^2 = \frac{\sum_{i=1}^{n} x_i^2 - n\,(\text{Old Mean}^2)}{n-1}$$

$$\sum_{i=1}^{n} x_i^2 = (n-1)\big(\text{Old Std}\big)^2 + n\big(\text{Old Mean}\big)^2$$

$$\text{New Std Deviation}^2 = \frac{\sum_{i=1}^{n} x_i^2 + (\text{New Data})^2 - (n+1)\big(\text{New Mean}\big)^2}{n}$$

$$= \frac{(n-1)\big(\text{Old Std}\big)^2 + n\big(\text{Old Mean}\big)^2 + (\text{New Data})^2 - (n+1)\big(\text{New Mean}\big)^2}{n}$$

New Median:

When $n == 0$: New Median = New Data

When $n == 1$: New Median = $\frac{\text{New Data+Old Median}}{2}$

When $n \bmod 2 == 0$: Let s and b denote the two middle values (each side has $n/2$ elements).

$$\text{New Median} = \text{Med}(s, b, x), \qquad \begin{cases} x \le s \le b & \Rightarrow \text{New Med} = s, \\ s \le x \le b & \Rightarrow \text{New Med} = x, \\ s \le b \le x & \Rightarrow \text{New Med} = b. \end{cases}$$

When $n \bmod 2 == 1$: Let the three central ordered values be $a < b < c$.

If $x < b$:
If $x < a$ then New Median = a.
If $a < x < b$ then New Median = x.

If $x > b$:
If $b < x < c$ then New Median = x.
If $c < x$ then New Median = c.

How to run/test the code

All three functions have been implemented in a single file, `q2.m`. Random test cases are generated using the `randn` function. The code can be executed by running `q2`.

Histogram Update

It is possible to update the histogram when a new data value is added without having to recalculate all bin counts. First, use the histogram's defined bin edges to identify which bin interval the new value falls into. After it has been located, add one to the count of that particular bin. Every other bin count stays the same.

3: Probability Bound

Bonferroni's inequality:

$P(A, B) \ge P(A) + P(B) - 1$

Given:

$P(A) \ge 1 - q_1,$

$P(B) \ge 1 - q_2$

Substituting in Bonferroni's inequality:

$P(A, B) \ge (1 - q_1) + (1 - q_2) - 1$

$P(A, B) \ge 1 - (q_1 + q_2)$

4: Lawyer's Argument

Probability that the bus was red:

$$P(R) = \frac{1}{100} = 0.01$$

Probability that the bus was blue:

$$P(B) = \frac{99}{100} = 0.99$$

Probability that a red bus is observed as red:

$$P(r \mid R) = \frac{99}{100} = 0.99$$

Probability that a blue bus is observed as red:

$$P(r \mid B) = \frac{2}{100} = 0.02$$

Bayes' Rule:

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

Probability that the bus was really a red one, given that XYZ observed it to be red:

$$P(R \mid r) = \frac{P(r \mid R)\, P(R)}{P(r \mid R)\, P(R) + P(r \mid B)\, P(B)}$$

Substituting the values,

$$P(R \mid r) = \frac{(0.99)(0.01)}{(0.99)(0.01) + (0.02)(0.99)}$$

$$P(R \mid r) = \frac{1}{3}$$

Even though the witness has a high accuracy (99%) in identifying red buses, red buses are rare in the town (only 1%). Using Bayes' theorem, the probability that the bus was actually red given that the witness said it was red is only about 33.3%. This means the XYZ's claim isn't reliable proof, since there's a much higher chance the bus was blue than red.

5: Specific Exit Poll

Setup: Each queried voter independently supports A with probability p (=95/100). For 3 draws (with replacement), the exit poll declares A to be the winner if $x \geq 2$ (Where x is the number of polled voters who support A).

$$P(\text{declare } A) = P(x \geq 2) = \binom{3}{2} p^2(1-p) + \binom{3}{3} p^3 = 3p^2(1-p) + p^3 = 3p^2 - 2p^3$$

$p = 0.95$, hence

$$P(\text{declare } A) = 3(0.95)^2 - 2(0.95)^3 = 0.99275$$

The probability that the exit poll declared a majority for A is 0.99275.

For 10,000 residents (still sampling 3 with replacement): The sampling distribution is unchanged (i.e probability that a draw results A's supporter is still p (=0.95)), so the accuracy is the same:

$$P(\text{declare } A) = 3p^2 - 2p^3 = 0.99275$$

6: General Exit Poll

Setup: Total number of voters = m. Label voters by the set $\{1, 2, \ldots, m\}$. Probability that a randomly chosen voter prefers A is $p = k/m$; hence the number of A–supporters is k. We choose an n-tuple S of voters *with replacement* (so $|S| = n$ and the total number of possible S is m^n).

Let $x_i = 1$ if the ith chosen voter in S voted for A and 0 otherwise, and define

$$q(S) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Then $q(S).\, n = \sum_{i=1}^{n} x_i$ is an integer between 0 and n.

Let $l(z)$ be the number of n-tuples S for which $q(S) = z/n$ (i.e. exactly z A supporters in S). Choosing the z A–positions and filling them from the k A–supporters, and the remaining $n - z$ positions from the $m - k$ B–supporters gives

$$l(z) = \binom{n}{z} k^z (m-k)^{n-z}$$

Sanity Check :

$$\sum_{z=0}^{n} l(z) = \sum_{z=0}^{n} \binom{n}{z} k^z (m-k)^{n-z} = (k + m - k)^n = m^n$$

(a) Mean of q(S): Show $\frac{1}{m^n}\sum_S q(S) = p$.

$$\frac{1}{m^n}\sum_S q(S) = \frac{1}{m^n}\sum_{z=0}^{n} \frac{z}{n}\, l(z) = \frac{1}{n\, m^n}\sum_{z=0}^{n} z\binom{n}{z} k^z (m-k)^{n-z}$$

$$\text{Using } \sum_{z=0}^{n} z\binom{n}{z} a^z b^{n-z} = \sum_{z=1}^{n} na\binom{n-1}{z-1} a^{z-1} b^{n-z} = n\, a\, (a+b)^{n-1} \qquad \text{with } a = k,\ b = m - k,$$

$$\frac{1}{m^n}\sum_S q(S) = \frac{1}{n\, m^n}\left(nk\, m^{n-1}\right) = \frac{k}{m} = p$$

(b) Mean of Squares of q(S): Show $\frac{1}{m^n}\sum_S q(S)^2 = \frac{p}{n} + \frac{p^2(n-1)}{n}$.

$$\frac{1}{m^n}\sum_S q(S)^2 = \frac{1}{m^n}\sum_{z=0}^{n}\left(\frac{z}{n}\right)^2 l(z) = \frac{1}{n^2 m^n}\sum_{z=0}^{n} z^2\binom{n}{z} k^z (m-k)^{n-z}$$

Write $z^2 = z + z(z-1)$

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

$$= \frac{1}{n^2 m^n}\left(\sum_{z=0}^{n} z\binom{n}{z} k^z (m-k)^{n-z} + \sum_{z=0}^{n} z(z-1)\binom{n}{z} k^z (m-k)^{n-z}\right)$$

Use the identities (with a=k, b=m-k):

$$\sum_{z=0}^{n} z\binom{n}{z}a^z b^{n-z} = \sum_{z=1}^{n} na\binom{n-1}{z-1}a^{\,z-1}b^{\,n-z} = n\,a\,(a+b)^{n-1},$$

$$\sum_{z=0}^{n} z(z-1)\binom{n}{z}a^z b^{n-z} = \sum_{z=2}^{n} n(n-1)a^2\binom{n-2}{z-2}a^{\,z-2}b^{\,n-z} = n(n-1)\,a^2\,(a+b)^{n-2}$$

Substitute a=k, b=m-k:

$$\sum_{z=0}^{n} z\binom{n}{z}k^z(m-k)^{n-z} = nk\,m^{\,n-1}, \qquad \sum_{z=0}^{n} z(z-1)\binom{n}{z}k^z(m-k)^{n-z} = n(n-1)k^2\,m^{\,n-2}$$

Plug back into the expression:

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

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

$$= \frac{k}{nm} + \frac{n-1}{n}\,\frac{k^2}{m^2}$$

$$= \frac{p}{n} + \frac{n-1}{n}\,p^2$$

(c) Mean Square deviation of q(S): Show $\frac{1}{m^n}\sum_S \big(q(S)-p\big)^2 = \frac{p(1-p)}{n}$.

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

$$= \Big(\frac{p}{n} + \frac{p^2(n-1)}{n}\Big) - p^2$$

$$= \frac{p(1-p)}{n}$$

(d) Proportion of n-tuples with $|q(S)-p| > \delta$.

Chebyshev's inequality: $\frac{|S_k|}{N} \le \frac{1}{k^2}, \qquad S_k = \{x : |x-\mu| > k\sigma\}$

Proportion of n-sized subsets S for which $|q(S)-p| > \delta$

$$= \frac{\big|\{S : |q(S)-p| > \delta\}\big|}{m^n}$$

$$= \frac{\big|\{S : |q(S)-p| > \sigma\,(\delta/\sigma)\}\big|}{m^n}$$

From Chebyshev inequality

$$\frac{\big|\{S : |q(S)-p| > \sigma\,(\delta/\sigma)\}\big|}{m^n} \le \frac{\sigma^2}{\delta^2} = \frac{1}{\delta^2}\,\frac{p(1-p)}{n}$$

Note: Minute observation

$$\sigma^2 = \frac{\sum(q(S)-p)^2}{m^n-1} \ne \frac{\sum(q(S)-p)^2}{m^n}$$

Fix: In proof of Chebyshev's inequality we show

$$1-\frac{|S_k|}{N} \geq 1-\frac{1}{k^2}+\frac{1}{nk^2}$$

$$\Rightarrow \quad \frac{|S_k|}{N} \leq \frac{n-1}{n\,k^2}$$

$$S_\delta = \{\, x_i : |x_i-\mu| > \delta \,\}.$$

$\dfrac{|S_\delta|}{N} \leq \dfrac{n-1}{n}\cdot\dfrac{\sigma^2}{\delta^2} = \dfrac{n-1}{n}\cdot\dfrac{1}{\delta^2}\cdot\dfrac{1}{n-1}\displaystyle\sum_{i=1}^{n}(x_i-\mu)^2 = \dfrac{1}{\delta^2}\cdot\dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(x_i-\mu)^2$. Hence the result is still valid.

Significance: The bound $\frac{\left|\{S:\,|q(S)-p|>\delta\}\right|}{m^n} \leq \frac{p(1-p)}{n\,\delta^2}$ says that the proportion of subsets whose $q(S)$ deviates from the true proportion p by more than δ is very small. This proportion can be interpreted as the probability that the exit poll will tell that the chance of A winning to be in the range $p-\delta$ to $p+\delta$.

Consequences:

- Increasing the sample size n decreases the error probability in inverse proportion to n.
- Allowing a larger tolerance δ decreases the error probability in inverse proportion to δ^2.
- The error probability is independent of m from this inequality which is opposite to our intuition.
- The variance term $p(1-p)$ is smallest when p is near 0 or 1; i.e., when one candidate has a clear majority, the exit poll is more likely to be correct.
- For winner prediction, we do not need an exact percentage—only the correct side of 1/2. If $p = 0.2$, tolerance can be upto 0.3; this implies $q(S) < 0.5$, so the winner is identified correctly. Since δ can be larger when p is farther from 1/2, the bound above gives an even higher probability of accurate winner prediction in cases when one candidate has a clear majority.