DAI Assignment 1

Ansh Garg - 24B1066
Kashyap Khandelwal - 24B0932
Pranav Patil - 24B1074

August 2025

1. Q1

SOLUTION: Graphs:

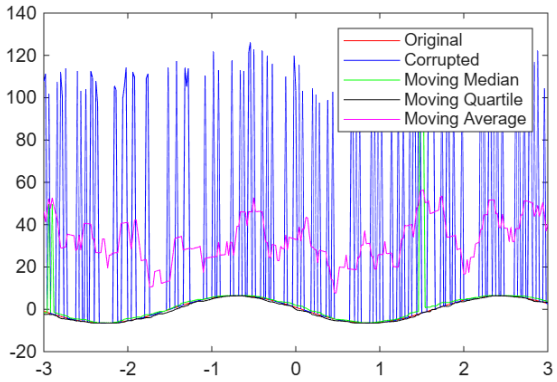

Figure 1: $f = 30\%$

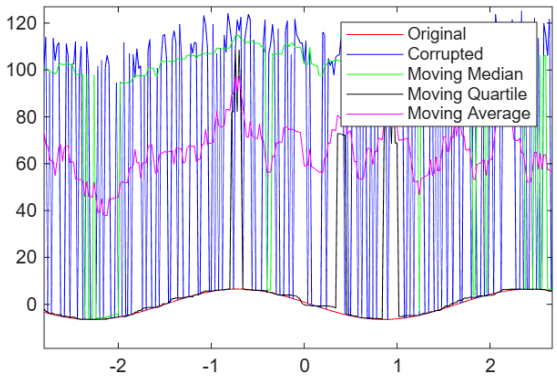

Figure 2: $f = 60\%$

Relative Mean Squared observed for different values of f and different methods of filtering:
f=30%: Moving median filtering: 5.612
f=30%: Moving quartile filtering: 0.010822
f=30%: Moving average filtering: 54.0188
f=60%: Moving median filtering: 433.358
f=60%: Moving quartile filtering: 21.9061
f=60%: Moving average filtering: 200.2803

Inference: The results clearly show that Moving Quartile Filtering performs best for both $f = 30\%$ and $f = 60\%$.

Since the noise added is always large and positive (100–120) compared to the sine wave's amplitude(6.5), the arithmetic mean is heavily biased upward, and even the median shifts higher when a large fraction of neighbors are corrupted.
The lower quartile (25th percentile), however, naturally selects from the lower range of the neighborhood values, effectively ignoring the high outliers and staying close to the true signal, which explains its consistently lower Relative Mean Squared Error.

2. Q2

SOLUTION:

- Proof for **newMean** As we know that

$$\text{mean} = \frac{\sum A_i}{n},$$

hence when we add the new element the new mean will be

$$\text{newMean} = \frac{\sum A_i + \text{NewData}}{n+1}.$$

Substituting $\sum A_i$ from above we get

$$\text{newMean} = \frac{\text{OldMean} \cdot n + \text{NewData}}{n+1}.$$

- Proof for **newMedian** We are assuming that the array A is sorted so now we take cases depending on the parity of n; also all the data values are distinct so we won't have to consider duplicates.

Case 1: n **even** We know

$$\text{oldMedian} = \frac{A(n/2) + A(n/2+1)}{2}.$$

Now when we add the new element we are trying to find the middle element of the new array.

  - If NewData $< A(n/2)$ then the middle element, i.e., the $(n/2+1)^{\text{th}}$ element is $A(n/2)$. It is the new median.
  - If NewData $> A(n/2+1)$ then the middle element, i.e., the $(n/2+1)^{\text{th}}$ element is $A(n/2+1)$.
  - If $A(n/2) \leq$ NewData $\leq A(n/2+1)$ then the median will be NewData itself because it is the $(n/2+1)^{\text{th}}$ element and that's what we are looking for.

Case 2: n **odd** The new median is the average of the $\frac{n+1}{2}^{\text{th}}$ and $\left(\frac{n+1}{2}+1\right)^{\text{th}}$ elements in the new array.
We first define two index variables:

$$\text{midIndex1} = \frac{n+1}{2} - 1, \quad \text{midIndex2} = \text{midIndex1} + 1$$

We use `else if` conditions to make the logic more compact. We consider the following four cases:

- **Case 1:** If NewDataValue $\leq A(\text{midIndex1})$, then

$$\text{newMedian} = \frac{A(\text{midIndex1}) + A(\text{midIndex1}+1)}{2}$$

  This can be easily interpreted by taking a small example array.
- **Case 2:** If NewDataValue $\geq A(\text{midIndex2}+1)$, then

$$\text{newMedian} = \frac{A(\text{midIndex2}) + A(\text{midIndex2}+1)}{2}$$

- **Case 3:** If NewDataValue $\geq A(\text{midIndex2})$, then

$$\text{newMedian} = \frac{A(\text{midIndex2}) + \text{NewDataValue}}{2}$$

- **Case 4:** If $A(\text{midIndex1}) < \text{NewDataValue} < A(\text{midIndex2})$, then

$$\text{newMedian} = \frac{A(\text{midIndex2}) + \text{NewDataValue}}{2}$$

- Proof for **new standard deviation** We know

$$\text{Standard Deviation} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}.$$

Now when we add the new value,

$$\text{newStd} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \text{newMean})^2 + (\text{NewData} - \text{newMean})^2}{n}}.$$

We have to simplify:

$$\sum_{i=1}^{n}\left[(x_i - \text{oldMean}) + (\text{oldMean} - \text{newMean})\right]^2$$

$$= \sum_{i=1}^{n}(x_i - \text{oldMean})^2 + n(\text{oldMean} - \text{newMean})^2 + 2(\text{oldMean} - \text{newMean})\sum_{i=1}^{n}(x_i - \text{oldMean}).$$

We know that the third term is zero, and substituting the first term gives the update formula for standard deviation directly. Thus, the formula for the updated variance becomes

$$\text{newVar} = \frac{(\text{oldStd})^2 \cdot (n-1) + n \cdot (\text{oldMean} - \text{newMean})^2 + (\text{NewData} - \text{newMean})^2}{n}$$

and hence the updated standard deviation is

$$\text{newStd} = \sqrt{\frac{(\text{oldStd})^2 \cdot (n-1) + n \cdot (\text{oldMean} - \text{newMean})^2 + (\text{NewData} - \text{newMean})^2}{n}}.$$

- **Updating the Histogram after adding a new Value** When a new value is added, we do not need to reconstruct the entire histogram from scratch. Instead, determine the bin in which the new value falls and increment the count of that bin.

3. Consider two events A and B such that $P(A) \geq 1-q_1$ and $P(B) \geq 1-q_2$. Show that $P(A,B) \geq 1-(q_1+q_2)$. [10 points]

   SOLUTION: We know that for any two events A and B:

   $$P(A \cup B) = P(A) + P(B) - P(A \cap B) \tag{1}$$

   And $A \cup B$ being an event in itself has a probability ≤ 1.

   $$P(A \cup B) \leq 1 \tag{2}$$
   $$P(A) + P(B) - P(A \cap B) \leq 1 \tag{3}$$
   $$P(A \cap B) \geq P(A) + P(B) - 1 \tag{4}$$
   $$P(A) + P(B) - 1 \geq (1-q_1) + (1-q_2) - 1 = 1 - q_1 - q_2 \tag{5}$$

   Now from (4) and (5):

   $$P(A,B) \geq 1 - (q_1 + q_2) \tag{6}$$

   We're interpreting $P(A,B)$ as $P(A \cap B)$.

4. Here is a simple example of probability in law. In a certain town, there exist 100 buses out of which 1 is red and 99 are blue. A person XYZ observes a serious accident caused by a bus at night and remembers that the bus was red in color. Hence, the police arrest the driver of the red bus. The driver pleads innocence. Now, a benevolent lawyer decides to defend the distressed bus driver in court. The lawyer ropes in an opthalmologist to test XYZ's ability to differentiate between the colors red and blue, under illumination conditions similar to those that existed that fateful night. The opthalmologist suggests that XYZ sees red objects as red 99% of the time and blue objects as red 2% of the time. What will be the main argument of the defense lawyer? (That is, what is the probability that the bus was really a red one, when XYZ observed it to be red?) **Show clearcut steps for your answer.** [15 points]

   SOLUTION: Let R be the event that the bus was actually red. R_O is the event that the color observed by XYZ was red. B and B_O are the counterparts for blue. Given Information:

   $$P(R_O/R) = 0.99$$
   $$P(R_O/B) = 0.02$$

   Now since the actual culprit is unknown, we assume the following, that every bus in the town is equally likely to be the culprit:

   $$P(R) = 0.01$$
   $$P(B) = 0.99$$

We calculate what is the probability that the bus was really a red one, when XYZ observed it to be red? i.e. $P(R/R_O)$:

$$P(R/R_O) = \frac{P(R_O/R)P(R)}{P(R_O \cap R) + P(R_O \cap B)} \quad \text{(Bayes' Theorem)} \tag{7}$$

$$P(R/R_O) = \frac{P(R_O/R)P(R)}{P(R_O/R)P(R) + P(R_O/B)P(B)} \tag{8}$$

$$P(R/R_O) = \frac{0.99 * 0.01}{0.99 * 0.01 + 0.02 * 0.99} \tag{9}$$

$$P(R/R_O) = \frac{1}{3} \tag{10}$$

Thus, the lawyer's argument would be that the red bus' driver was arrested and charged based upon the testimony of witness XYZ. But from our analysis, it is more likely than not that the witness mistook the bus as red when it was actually blue. There's only a 1 in 3 chance that the bus XYZ identified as red was actually red.

5. *In this question and the next one, we will understand the reason why exit polls make some statistical sense.* Consider a village with 100 residents, all of whom participate in an election contested by two candidates A and B. An exit poll is conducted after election day in which three residents (chosen uniformly at random with replacement) are asked as to whom they voted for. The candidate with the majority in the exit poll is declared to be the expected winner by the exit poll agency. If 95 percent of the residents favour candidate A over B and the remaining 5 percent favour B over A, what is the accuracy of this exit poll? (In other words, what is the probability that the exit poll that quizzed 3 voters, declared a majority for A?) Assume that the people give truthful answers in the exit poll. Now, suppose the village has 10,000 residents, all of whom are eligible to vote, and the exit poll again asked only 3 (truthful) voters (chosen uniformly at random with replacement). What is the accuracy of the exit poll now? [15 points]

   SOLUTION: In this situation, since we know for sure that 95 percent of the people voted for A, we can say with certainty that A won the election. Now choosing 3 people uniformly at random with replacement from the set of all villagers is equivalent to choosing 3 people one by one from the set of all villagers.

   For the exit poll to be accurate, either all 3 of the people chosen have to be A voters or 2 of them are A voters and 1 of them is B voters. In the latter case, the first or the second or the third person can be the B voter. Let E denote the event that the exit polls are accurate, A be the event that a person chosen at random voted for A, and B be the event that a person chosen at random voted for B, then:

$$P(E) = \binom{3}{1} P(A)^2 P(B) + P(A)^3$$

From given information $P(A) = 0.95$ and $P(B) = 0.05$. Plugging in the values gives $P(E) = 0.99275$. So the exit poll is highly accurate.

When we change the number of villagers to 10000, the answer remains exactly same because we are choosing villagers with replacement, thus the logic for calculating probability remains exactly the same. So, the probability of choosing an A voter or a B voter remains exactly the same as in the previous case.

6. Continuing the previous problem, the question to be asked and answered is what happens if the percentages for A and B are less obvious and unknown! Consider that there are m voters in the village, and there is a probability $p = k/m$ that the voters prefer A over B. blueEach voter has a unique index number from 1 to m. Let us suppose that the exit poll asks a randomly (with replacement) chosen subset $\mathcal{S}$ containing n (truthful) voters. Let us define the quantity $x_i = 1$ if the ith voter voted for A and 0 if he/she voted for B. Let $q(\mathcal{S})$ be the proportion of voters from $\mathcal{S}$ who voted for A out of $n = |\mathcal{S}|$. That is blue$q(S) = \sum_{i \in I(\mathcal{S})} x_i/n$ where $I(\mathcal{S})$ is a set containing the index (from 1 to m) of each voter in $\mathcal{S}$. Given this, do as directed:

   (a) Prove that $\sum_{\mathcal{S}} \dfrac{q(\mathcal{S})}{m^n} = p$. This is the average of the values of $q(\mathcal{S})$ across all subsets $\mathcal{S}$ of size n.

   (b) Prove that $\sum_{\mathcal{S}} \dfrac{q^2(\mathcal{S})}{m^n} = \dfrac{p}{n} + \dfrac{p^2(n-1)}{n}$.

   (c) Prove that $\sum_{\mathcal{S}} \dfrac{(q(\mathcal{S}) - p)^2}{m^n} = \dfrac{p(1-p)}{n}$.

   (d) Hence argue that the proportion of n-sized subsets $\mathcal{S}$ (out of m^n) for which $|q(S) - p| > \delta$, is less than or equal to $\dfrac{1}{\delta^2}\dfrac{p(1-p)}{n}$. Note that this proportion is quite small. This is also a nice application of one of the inequalities we studied in class (which one?). What is the significance of this result? [5+7+2+(3+3) = 20 points]

SOLUTION: Given, probability of A winning over B is k/m and choice of each voter is fixed, we have hat out of the m voters, k vote for A, and the remaining vote for B.
Consider a subset S, containing n voters. Let a of them be voters for A(Note that there can be repeated voters here). Now, to find how many such subsets S exist. We need to select a voters for A from k voters overall, with repetition. This gives k^a ways. Choosing the others: $(m-k)^{n-a}$. One can see that there exists an inherent ordering in the way we conduct the exit poll, considering the order we ask each voter in S whom they voted for. Thus, there are $\binom{n}{a}$ ways. Overall, number of subsets S with a voters for A are :

$$\binom{n}{a} k^a (m-k)^{n-a}$$

Now, $q(S) = \frac{\sum x_i}{n} = \frac{a}{n}$. For each S, $\sum x_i = a$.

(a)

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

Using the fact that $\frac{a}{n}\binom{n}{a} = \binom{n-1}{a-1}$ and then evaluating the binomial expansion $\sum \binom{n-1}{a-1}k^{a-1}(m-k)^{(n-1)-(a-1)} = m^{n-1}$, we get

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

(b)

$$q^2(S) = (\sum \frac{x_i}{n})^2 = \frac{\sum x_i^2}{n^2} + 2\frac{\sum_{i\neq j} x_i x_j}{n^2}$$

For subset S with a voters of A, $\sum x_i^2 = a$ and $\sum_{i\neq j} x_i x_j = \binom{a}{2}$. Thus,

$$\sum q^2(S) = \sum_{a=1}^{n} \frac{a}{n^2}\binom{n}{a}k^a(m-k)^{n-a} + \sum_{a=1}^{n} \frac{a(a-1)}{n^2}\binom{n}{a}k^a(m-k)^{n-a}$$

Using $a(a-1)\binom{n}{a} = n(n-1)\binom{n-2}{a-2}$ and solving the binomial expansion, we get,

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

Thus,

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

(c)

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

(d) This is in fact Chebyshev's Inequality! Consider the random variable given by $q(S)$. Its mean and variance are given by $\sum_S \frac{q(S)}{m^n} = p$ and $\sum_S \frac{(q(S)-p)^2}{m^n} = \frac{p(1-p)}{n}$ respectively. Thus, applying Chebyshev's inequality, we get

$$P\{|q(S)-p| > \delta\} \leq \frac{1}{\delta^2}\frac{p(1-p)}{n}$$

We can see that the proportion of data values far from the mean falls off quite quickly. This tells us a lot about the accuracy of the exit

poll and the results of elections.
Greater the value of n, greater the chances of getting accurate results. Another interesting observation is that closer p is to 50%, less is the accuracy of the exit poll ($p(1-p)$ increases, thereby increasing proportion of subsets S far from the mean.)