Name: Zayd Hammoudeh

Assignment: CMPS218 Homework #1

Other Student Discussions: I discussed the problems in this homework with the following students:

- Hadley Black – He asked me about the first problem.
- Noujan Pashanasangi – We discussed second problem concerning the poisoned glass and in particular the third problem concerning soft K-Means stiffness.
- Will Bolden – We discussed the second and third problems. In particular, we discussed the fourth problem in depth.
- Konstantinos Zampetakis and Keller Jordan – We discussed the fourth problem.

Problem Assessments: All problems are complete to the best of my understanding. The answer to the fourth problem was the most challenging and is the one where I believe my answer may be the most suspect.

Chapter #2.9, Problem #33

Two points are selected at random on a straight line segment of length 1. What is the probability that a triangle can be constructed out of the three resulting segments?

For two randomly selected points, $0 \leq x_1 \leq 1$ and $0 \leq x_2 \leq 1$, define $L_{\max}$ as:

$$L_{\max} = \max\Big\{ \min\{x_1, x_2\},\ |x_2 - x_1|,\ 1 - \max\{x_1, x_2\} \Big\}.$$

By the triangle inequality theorem, it is clear that a triangle can only be formed from these three resulting segments if $L_{\max} < \frac{1}{2}$. Therefore if $x_1 < \frac{1}{2}$, then $\frac{1}{2} < x_2 < x_1 + \frac{1}{2}$. Similarly, if $x_1 > \frac{1}{2}$, then $x_1 - \frac{1}{2} < x_2 < \frac{1}{2}$.

Consider the tuple (x_1, x_2). Each point in the domain $[0, 1]^2$ is equally likely since x_1 and x_2 are selected (uniformly) at random. Figure 1 shows the portions of this domain (highlighted in light blue) where the lines segments induced by x_1 and x_2 can be used to form a triangle. Note that these shaded regions correspond to the valid ranges for x_1 and x_2 derived from the triangle inequality above.

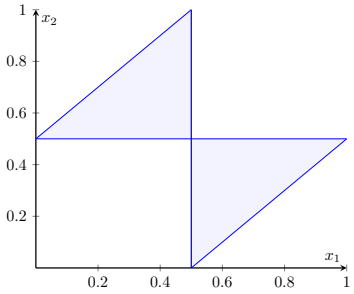

Figure 1: Regions of the domain where the line segments induced by (x_1, x_2) form a triangle.

Since the probability of selecting any point from the above domain is uniform, the probability of selecting points that form a triangle is found via:

$$\begin{aligned} \Pr(\text{Form a triangle}) &= \frac{\text{Shaded Area}}{\text{Total Area}} \\ &= \frac{2 \cdot \left(\frac{1}{2}\right)^3}{1} \\ &= \boxed{\frac{1}{4}}. \end{aligned}$$

Chapter #5.8, Problem #30

***Scientific American* carried the following puzzle in 1975.**

The poisoned glass: *'Mathematicians are curious birds,' the police commissioner said to his wife. 'You see, we had all those partly filled glasses lined up in rows on a table in the hotel kitchen. Only one contained poison, and we wanted to know which one before searching the glass for fingerprints. Our lab could test the liquid in each glass, but the tests take time and money, so we wanted to make as few of them as possible by simultaneously testing mixtures of small samples from groups of glasses. The university sent over a mathematics professor to help us. He counted the glasses, smiled and said:*
"Pick any glass you want, Commissioner. We'll test it first."
"But won't that waste a test?" I asked.
"No," he said. "it's part of the best procedure. We can test one glass first. It doesn't matter which one."
'How many glasses were there to start with?' the commissioner's wife asked.
'I don't remember. Somewhere between 100 and 200.'
What was the exact number of glasses?

Solve this puzzle and then explain why the professor is in fact wrong and the commissioner was right. What is in fact the optimal procedure for identifying the one poisoned glass? What is the expected waste relative to this optimum if one followed the professor's strategy? Explain the relationship to symbol coding.

The test for poison has a binary outcome, i.e., the sample either has poison or not. Therefore, assuming each cup has poison with equal probability, the size of the remaining set of glasses is, on average, cut in half with each test.

If the number of glasses, n, is a power of 2, then the number of tests required is $\lg n$, where lg is the base-2 logarithm. Note that the only power of 2 between 100 and 200 is 128. There was one extra glass that the professor tested separately. Therefore, there was $\boxed{\text{129 glasses}}$.

Glass ID	Probability of Poison	# Tests
1	1/129	1
2-129	128/129	1 + 7 = 8

Table 1: Number of tests required using the professor's strategy

Table 1 shows the number of tests required when using the professor's strategy. Glass #1 represents the first glass tested, i.e., the one selected at random. In the unlikely event that glass has the poison, no additional testing is required. In contrast, if the poison is in one of the other 128 glasses, seven tests (plus the additional one for the first glass) are required. Using the professor's strategy, the expected number of tests is:

$$\begin{aligned}\mathbb{E}(\text{Professor's Strategy}) &= \frac{1}{129} \cdot 1 + \frac{128}{129} \cdot 8 \\ &\approx \boxed{7.946}.\end{aligned}$$

In contrast, the optimal strategy is:

1. Select one glass at random and leave it off to the side.
2. Test a sample that combines wine from half of the remaining glasses (excluding the one off to the side).

3. If poison is observed in this tested sample, discard the untested glasses. Otherwise, discard the tested glasses.
4. Repeat steps #2 and #3 until only a single glass remains (excluding the one off to the side).
5. If poison was ever observed in any of the previous tests, then the remaining glass not off to the side has the poison, and no additional testing is required.
6. If poison was never observed in any test, then test just the remaining glass not off to the side. If poison is detected, then the answer is clear, and the tested glass has poison; otherwise, the glass off to the side has the poison.

Table 2 shows the number of tests required using this optimum strategy. Note that Glass #129 represents the remaining glass in step #6 where no sample tested positive for poison up to the last remaining glass not off to the side.

Glass ID	Probability of Poison	# Tests
1	1/129	7 + 1
2-128	127/129	7
129	1/129	7 + 1

Table 2: Number of tests required using the optimum strategy

Using this optimum strategy, the expected number of tests is:

$$\begin{aligned}\mathbb{E}(\text{Optimum Strategy}) &= \frac{2}{129} \cdot 8 + \frac{127}{129} \cdot 7 \\ &\approx \boxed{7.016}.\end{aligned}$$

It is clear then that the expected waste of the professor's strategy is $\boxed{0.93}$ tests.

Maximum compression of a symbol code is achieved by assigning shorter codes (i.e., those with less bits) to outcomes with higher probability. In contrast, the professor prioritized the least likely outcome by testing the randomly selected glass first. The optimal strategy described above always tests the most likely outcome (i.e., more glasses at once) similar to how symbol codes are encoded.

Chapter #20.2, Problem #2

Show that as the stiffness β goes to ∞, the soft K-means algorithm becomes identical to the original hard K-means algorithm except for the way in which means assigned no points behave. Describe what those means do instead of sitting still.

In the standard or "hard" K-means algorithm, each point is assigned to exactly one cluster. As such, each of a cluster's points have equal membership.

Soft K-means reduces the rigidity of the standard K-means by introducing a new "stiffness" hyper-parameter, β. Rather than each point being a member of exclusively one cluster, the *responsibility* for each point is shared (generally unevenly) among all K clusters. For cluster $k \in \{1, \ldots, K\}$ and point $\mathbf{x}^{(n)}$, the responsibility, $r_k^{(n)}$, is:

$$r_k^{(n)} = \frac{\exp(-\beta d(\mathbf{m}^k, \mathbf{x}^{(n)}))}{\sum_{k'} \exp(-\beta d(\mathbf{m}^{k'}, \mathbf{x}^{(n)}))}$$

where d is the distance metric, and $\mathbf{m}^k$ is the center of cluster k.

As β increases, then even small differences in d can cause massive changes in responsibility. As $\beta \to \infty$, all responsibility for a point will be assigned to its nearest cluster. This behavior is exactly the same as standard, "hard" K-means where points belong to only the cluster's whose centroid is closest.

Considering the second part of the question, the centroid update rule for cluster, k, is:

$$\mathbf{m}^{(k)} = \frac{\sum_n r_k^{(n)} \mathbf{x}^{(n)}}{R^{(k)}}$$

where the total responsibility, $R^{(k)}$, for cluster k is:

$$R^{(k)} = \sum_n r_k^{(n)}.$$

As mentioned previously, small differences in the distance, d, cause huge differences in the responsibility, r, when β approaches infinity. In hard K-means, it was possible for a cluster to be assigned no points. However, in soft K-Means, such clusters still have *some* responsibility for every point, albeit infinitesimally small when β approaches infinity. In that case, those means will move to the *location of the point with the highest responsibility, which may not necessarily be the closest.*

Chapter #22.5, Problem #15

The seven scientists. N **datapoints** $\{x_n\}$ **are drawn from** N **distributions, all of which are Gaussian with a common mean** μ **but with different unknown standard deviations** σ_n**. What are the maximum likelihood parameter** μ**,**$\{\sigma_n\}$ **given the data? For example, seven scientists (A, B, C, D, E, F, G) with wildly-differing experimental skills to measure** μ**. You expect some of them to do accurate work (i.e., to have small** σ_n**), and some of them to turn in wildly inaccurate results (i.e., to have enormous** σ_n**). Table 3 shows their seven results. What is the** μ**, and how reliable is each scientist?**

Scientist	x_n
A	-27.020
B	3.570
C	8.191
D	9.898
E	9.603
F	9.945
G	10.056

Table 3: Seven measurements $\{x_n\}$ of a parameter μ by seven scientists each having his own noise-level σ_n.

I hope that you agree that, intuitively, it looks pretty certain that A and B are both inept measurers, that D-G are better, and that the true value μ **is somewhere close to 10. But what does maximizing the likelihood tell you?**

Given μ, the probability that observer i with measurement standard deviation, σ_n, measures any single value, x_n, is:

$$\Pr\left(x_n|\mu,\sigma_n\right) = \frac{1}{\sqrt{2\pi\sigma_n^2}} \exp\left(\frac{-(x_n-\mu)^2}{2\sigma_n^2}\right).$$

If there are N (independent) observers, then the likelihood, L, of the measurements $\{x_n\}$ where $n \in \{1,\ldots,N\}$ is:

$$\begin{aligned} L &= \Pr\left(\{x_n\}|\mu,\{\sigma_n\}\right) \\ &= \prod_{n=1}^{N} \Pr\left(x_n|\mu,\sigma_n\right) \\ &= \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi\sigma_n^2}} \exp\left(\frac{-(x_n-\mu)^2}{2\sigma_n^2}\right). \end{aligned}$$

Taking the natural logarithm of this function yields:

$$\ln L = \sum_{n=1}^{N} \left(-\frac{1}{2}\ln\left(2\pi\sigma_n^2\right) - \frac{(x_n-\mu)^2}{2\sigma_n^2}\right).$$

Since the natural logarithm function is strictly increasing, we can take the derivative of $\ln L$ with respect to μ to find the maximum likelihood for μ. This yields:

$$0 = \sum_{n=1}^{N} \frac{(x_n - \mu)}{\sigma_n^2}$$
$$\mu = \frac{1}{\sum_{n=1}^{N} \sigma_n^{-2}} \sum_{n=1}^{N} \frac{x_n}{\sigma_n^2}.$$

For each σ_n, the maximum likelihood estimate is found by taking the partial derivative with respect to σ_n yielding:

$$0 = -\frac{2\pi\sigma_n}{4\pi\sigma_n^2} + \frac{(x_i - \mu)^2}{\sigma_n^3}$$
$$\sigma_n = \sqrt{(x_n - \mu)^2} = |x_n - \mu|.$$

To find the maximum likelihood, $L_{\max}$, substitute the definition for σ_n found above. This yields

$$\begin{aligned} L_{\max} &= \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi(x_n - \mu)^2}} \exp\left(\frac{-(x_n - \mu)^2}{2(x_n - \mu)^2}\right) \\ &= \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi(x_n - \mu)^2}} \exp\left(-\frac{1}{2}\right). \end{aligned}$$

Therefore, the likelihood is maximized when $\mu = x_n$. For this problem, x_n is any value in Table 3 meaning the maximum likelihood is when any of the seven scientists are perfectly correct, i.e., always report the true μ with no variation. This result is disquieting as it means the likelihood is maximized if μ is -27.020 or 3.570. As noted in the question itself, the visceral feeling is that the true μ is around ten give or take. Therefore, maximizing the likelihood may not always yield the best result.