Quantifying Information

2026-27

Mathematics Club

Centre for Innovation
IIT Madras

Instructions

General

- This is a practice sheet designed to make you explore and learn the topic inside out to your maximum capacity. Don't forget to have fun while doing the problems.
- While you are free to use any of the plethora of resources on the internet, we would suggest **refraining from using LLMs** in any way to solve the problems. You may cite your reference(s) that you use.
- For some questions, there is no "correct" answer, and it depends on your understanding and overview of the topic. Feel free to express your opinions.
- Attempt the most questions you can. We prefer in-depth answers over shallow ones, and show how you thought about everything.
- The goal of both the session and assignment is for you to have fun and learn as much as possible. We have provided some resources that can help you in the end. You may refer other sources as well. All the best!

Assignment Submission Format

- Your solutions can be done in any manner, written manually, digitally, using documents, or even LaTeX. In the Google Form given below, please submit all your scanned copies as a PDF file with the given format <FullName_SS_Info_Theory.pdf>. If I am John Doe, I would submit <John_Doe_SS_Info_Theory.pdf>.

Submission

- Please attach your final PDF in this Google Form:- Submission Link
- The deadline for submission of your assignments is **31st July 2026, 23:59**.
- The bonus questions are for provoking curiosity. They are not compulsory to attend, and if you do, there is no need to submit the code.

Contact Us

- Harish Murugaperumal - +91 73059 78155
- Anirudh Pai - +91 80739 14299

Quantifying Information

Mathematics Club

§1 Information and Entropy

1. Find the information content of a statement that says the next digit of pi is 7 (assuming all digits are equally likely).

2. What is the information obtained from an event $i = 6$, where the probability of an event is such:

$$P(x) = \frac{x^2}{140}$$

   Given that $1 \leq i \leq 7$.

3. A fisherman says that there is a $\frac{1}{4}$ chance of a fish getting caught, and if there is a $\frac{1}{5}$ chance of raining and catching a fish, find the information given in the statement "There is no rain today.", assuming that the events are independent.

4. Which discrete probability distribution has the highest entropy? Could you prove it? What is the entropy?
   Hint: *Use Lagrange multipliers.*

5. Compute the entropy of a fair 8-sided die two ways:
   a. using the summation formula directly
   b. using the shortcut.

   Confirm they match.

6. A loaded 4-sided die has probabilities $0.5, 0.25, 0.125, 0.125$ for faces 1, 2, 3, 4. Compute $H(X)$. Is it more or less than $\log_2(4) = 2\log_2(2) = 2$? Why does that make sense?

§2 Applications - Huffman Encoding and Shannon Theorems

1. Build a Huffman tree for the sentences (use coding if you can), and assign codes to each character.
   a. "Life is like riding a bicycle. To keep your balance, you must keep moving."
   b. "Success is not final; failure is not fatal: It is the courage to continue that counts."

2. a. Build a Huffman tree for the string "the quick fox outruns the hare". Give the codewords and compute $L = \sum_i P(x_i) l_i$ (average codeword length).
   b. Without a fixed-length code, the string $\lceil \log_2(16) \rceil = 4$ bits per character, or 120 bits total. Using your Huffman code from the previous question. compute the total number of bits needed to encode the whole string, and find the percentage savings versus the fixed-length version.
   c. Verify that L satisfies the Source Coding Theorem.

d. Using the Shannon-Hartley formula, find the capacity of a channel C with bandwidth $B = 200Hz$, and the signal-to-noise ratio $S/N = 16$. What is the minimum time needed to transfer the encoded string above in this channel, ensuring that no data is lost?

§3 Cross-Entropy, KL Divergence

This is some data taken from two fictional LLMs trained on a primitive fictional language **Abilu**. The data says that after a given word *malus*, these 10 words follow that word in common usage. These are the probability distributions assigned to those 10 words, taken from common usage of these words in daily speech and writing. $P(x)$ is the true language data from real-world usage, and $Q_1(x)$ and $Q_2(x)$ are data learnt by LLMs 1 and 2.

Word	**P(x)**	$\mathbf{Q_1(x)}$ **(LLM 1)**	$\mathbf{Q_2(x)}$ **(LLM 2)**
ahjud	0.15	0.14	0.11
amediko	0.10	0.11	0.14
hulusari	0.01	0.02	0.05
polf	0.20	0.19	0.15
meows	0.19	0.20	0.24
opgnat	0.03	0.04	0.02
molskiahjud	0.02	0.01	0.01
hulupolos	0.01	0.01	0.02
pofa	0.20	0.21	0.17
poersii	0.19	0.17	0.09

Answer the given questions below from this distribution.

1. Calculate the cross-entropy $H(P, Q_1)$ and $H(P, Q_2)$.
2. What do you think this represents? Which model has higher cross-entropy, and what does it indicate?
3. *(Bonus)* How do you "fix" this error? What do you think are the error correcting mechanisms for cross-entropy?
4. Find the **perplexity** (Perplexity $= 2^{H(P,Q)}$) of LLM 1. What does it tell you about the LLM?
5. Using the formulae, prove that $H(P, Q) = H(P) + D_{KL}(P||Q)$.
6. Find $H(P)$. Using the result from question 1, what does $H(P, Q_i) - H(P)$ indicate for $i = 1, 2$.
7. Find $D_{KL}(P||Q_1)$ and $D_{KL}(P||Q_2)$.
8. Find $H(Q_1, P)$. Are $H(P, Q_1)$ and $H(Q_1, P)$ equal? What do you think $H(Q_1, P)$ indicates?
9. Find $D_{KL}(Q_1||P)$. Are $D_{KL}(P||Q_1)$ and $D_{KL}(Q_1||P)$ equal? Can this metric be used as a suitable error function?
10. If for the word *molskiahjud*, $Q_1(\textit{molskiahjud}) = 0$, what would happen to $D_{KL}(P||Q_1)$ and $H(P, Q_1)$? What if instead $P(\textit{molskiahjud}) = 0$?

§4 Open Ended

These questions are optional, and only exist for you to think about this topic and build curiosity. They do not have set answers.

1. What do you understand by the "information content" of an event? Why do you think the word "surprise" is used?
2. In a 4 coin toss, if there is a configuration of 2 heads and 2 tails (order doesn't matter), would you be surprised? Would you say the same if you got all tails? Why? How does information theory quantify this?
3. Adding on to the previous question, which occurrences do you consider "lucky" if they occur? Why?
4. (*Bonus, for brownie points*) If you are proficient in data algorithms, the optimal strategy of the first game we played must ring a bell. Why do you think that's the optimal strategy? Why is $\log_2$ used in every formula?
5. Why do you think Shannon entropy defines the absolute limit of how much data can be compressed?
6. With Huffman encoding, you can go as close to capacity, but you cannot transmit exactly at the capacity. Why do you think the Source Coding Theorem sets an exact boundary $H(X) \leq L$, while the Noisy-Channel theorem only provides an upper $<$ (less than) bound? What is fundamentally different about "compressing a message" versus "surviving noise" that might explain why one theorem tolerates touching its limit and the other doesn't?
7. Read about how cross-entropy is used in LLM generation. Why do you really think cross-entropy is used here? What is the difference in significance between $H(P, Q)$ and $D_{KL}(P||Q)$ as error functions in LLMs?
8. *(Bonus)* You have a biased coin. You want to find out its bias. To do so, you flip the coin n times, and you get h heads and t tails. Given this, which is a better guess, the maximising likelihood (MLE) estimate $p = \frac{h}{n}$, or the Bayesian estimate $p = \frac{h+1}{n+2}$?

§5 Resources

I have compiled a list of resources for your usage for the assignment, including the Google Slides presentation.

1. Google Slides - Link to Slides
2. Resources Document - Resources