CS109 Lecture 18 Worksheet: Information Theory

Surprise, Entropy, Information Gain, and KL Divergence

Summer 2026

Lecture 18 Worksheet: Information Theory

Topics: surprise (information content), entropy as expected surprise, choosing questions that maximize information gain, and measuring the distance between distributions with KL divergence.

How to use this sheet. Work in small groups. A TA and the instructor will circulate to help. We will go over selected solutions on the board. You do not need to finish every part; aim for understanding over speed. Keep fractions exact where you can. Unless stated otherwise, use $\log_2$ so answers are in *bits*.

Problem 1 (Review of Lecture 17): A Recursive Expectation

```python
def retry():
    success = bernoulli(p = 0.5)
    if success:
        return 4       # seconds of work
    else:
        return 2 + retry()
```

Let T be the value returned by `retry()`. Use the law of total expectation to compute $E[T]$.

Problem 2: Surprise

Recall the surprise (aka information content) of an event: $I(E) = \log_2 \frac{1}{P(E)}$.

(a) You draw a card and it is the ace of spades… twice in a row (with reshuffling). Each draw has probability $\frac{1}{52}$. What is the surprise of the first draw, in bits? (Leave as a log; ≈ 5.7.)

(b) What is the surprise of an event with probability 1? Why is that the right answer for a measure of surprise?

(c) Show that for independent events E and F, $I(E \cap F) = I(E) + I(F)$. In one sentence: why does this property make log the natural choice for measuring surprise?

Problem 3: Entropy

Entropy is expected surprise: $H(X) = \sum_x P(X{=}x) \log_2 \frac{1}{P(X{=}x)}$.

(a) Compute $H(X)$ for $X \sim \text{Bern}(0.5)$ and for $X \sim \text{Bern}(0.9)$. Which is more uncertain, and does the math match your intuition?

(b) Compute $H(X)$ for the PMF $\left(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8}\right)$.

(c) Among all distributions on 8 values, which has the largest entropy, and what is that entropy? Give an intuitive reason.

Problem 4: Which Diagnostic Test Is Better?

A production outage has one of four root causes, with prior probabilities: frontend $\frac{1}{2}$, backend $\frac{1}{4}$, database $\frac{1}{8}$, network $\frac{1}{8}$. Let X be the root cause; from Problem 3(b), $H(X) = 1.75$ bits.

You can run exactly one diagnostic test first:

- **Test A** determines whether the cause is the frontend or not: $\{$frontend$\}$ vs. $\{$backend, database, network$\}$.
- **Test B** determines whether the cause is client-side-or-backend or infrastructure: $\{$frontend, backend$\}$ vs. $\{$database, network$\}$.

(a) For Test A: compute the probability of each answer and the entropy of X remaining after each answer. Then compute the *expected* remaining entropy.

(b) Do the same for Test B.

(c) Which test gives more expected information gain? Why do we compare tests by *expected* remaining entropy rather than the remaining entropy of one particular answer?

Problem 5: Two Dice, Two Hints (pset5: info_theory_dice)

Consider calculating the sum of two dice. How much more uncertain would you be in the sum if you knew the first die showed a 1, compared to knowing only that the sum of the dice was less than or equal to 7?

Specifically, let X be the sum of two dice. Let $H_1(X)$ be the entropy of X if you were told that the first die was a 1. Let $H_2(X)$ be the entropy of X if you are told that $X \le 7$. What is $H_1(X) - H_2(X)$? Provide your answer to 5 decimal places.

Problem 6: KL Divergence by Hand

You believe a coin is fair, but its true probability of heads is 0.8. Let $P \sim \text{Bern}(0.8)$ be the true distribution and $Q \sim \text{Bern}(0.5)$ be your model. Recall

$$D(P \,\|\, Q) = \sum_x P(x) \log_2 \frac{P(x)}{Q(x)} \quad \text{(expected excess surprise from using model } Q \text{ when reality is } P).$$

(a) Compute $D(P \,\|\, Q)$.

(b) Compute $D(Q \,\|\, P)$. What do you notice?

(c) When is $D(P \,\|\, Q) = 0$? In one sentence: why does that make sense for "excess surprise"?

Challenge Problem (optional): KL Between Poissons (pset6: kl_poisson)

Calculate the KL divergence between $X_1 \sim \text{Poi}(\lambda_1 = 1)$ and $X_2 \sim \text{Poi}(\lambda_2 = 2)$. Give your answer to three decimal places. (Use natural log for this problem.)

Helpful identities: for $Y \sim \text{Poi}(\lambda)$, $\sum_{y=0}^{\infty} y \cdot P(Y{=}y) = \lambda$, and $\sum_{y=0}^{\infty} \frac{1}{y!} = e$.