CS109 Lecture 18 Worksheet: ANSWER KEY

Information Theory

Summer 2026

Lecture 18 Worksheet: Answer Key

Instructor / TA copy. Problem-set solutions are hidden in the standard build; run `./convert-to-pdf.sh 18 solutions` for the full instructor/TA edition.

Problem 1 (Review): A Recursive Expectation

Condition on the Bernoulli:

$$E[T] = \tfrac{1}{2}(4) + \tfrac{1}{2}\big(2 + E[T]\big) = 3 + \tfrac{1}{2}E[T] \implies E[T] = 6.$$

Problem 2: Surprise

(a) $I = \log_2 52 \approx 5.70$ bits. (The second identical draw adds another $\log_2 52$; see part (c).)

(b) $I = \log_2 \frac{1}{1} = 0$ bits: a certain event carries no surprise, exactly as a measure of surprise should.

(c) By independence, $P(E \cap F) = P(E)P(F)$, so

$$I(E \cap F) = \log_2 \frac{1}{P(E)P(F)} = \log_2 \frac{1}{P(E)} + \log_2 \frac{1}{P(F)} = I(E) + I(F).$$

The log is the (essentially unique) monotonic choice that makes surprise *add* across independent events, which matches intuition: two independent shocks are twice the news.

Problem 3: Entropy

(a) $H(\text{Bern}(0.5)) = 0.5 \log_2 2 + 0.5 \log_2 2 = 1$ bit.

$$H(\text{Bern}(0.9)) = -\big(0.9 \log_2 0.9 + 0.1 \log_2 0.1\big) \approx 0.469 \text{ bits.}$$

The fair coin is more uncertain — matching intuition, since the biased coin's outcome is fairly predictable.

(b)

$$H = \tfrac{1}{2} \log_2 2 + \tfrac{1}{4} \log_2 4 + \tfrac{1}{8} \log_2 8 + \tfrac{1}{8} \log_2 8 = \tfrac{1}{2} + \tfrac{1}{2} + \tfrac{3}{8} + \tfrac{3}{8} = 1.75 \text{ bits.}$$

(c) The uniform distribution: $H = \log_2 8 = 3$ bits. Uncertainty is maximized when no outcome is favored — every observation is equally surprising, so expected surprise is as large as possible.

Problem 4: Which Diagnostic Test Is Better?

(a) **Test A.** $P(\text{yes}) = \frac{1}{2}$: cause identified, remaining entropy 0. $P(\text{no}) = \frac{1}{2}$: conditional distribution over {backend, db, network} is $\left(\frac{1}{2}, \frac{1}{4}, \frac{1}{4}\right)$, with entropy $\frac{1}{2}(1) + \frac{1}{4}(2) + \frac{1}{4}(2) = 1.5$ bits.

$$\text{Expected remaining entropy} = \tfrac{1}{2}(0) + \tfrac{1}{2}(1.5) = 0.75 \text{ bits} \quad (\text{gain } 1.75 - 0.75 = 1.0 \text{ bit}).$$

(b) **Test B.** $P(\text{yes}) = \frac{1}{2} + \frac{1}{4} = \frac{3}{4}$: conditional distribution $\left(\frac{2}{3}, \frac{1}{3}\right)$, entropy ≈ 0.918 bits. $P(\text{no}) = \frac{1}{4}$: conditional distribution $\left(\frac{1}{2}, \frac{1}{2}\right)$, entropy 1 bit.

$$\text{Expected remaining entropy} = \tfrac{3}{4}(0.918) + \tfrac{1}{4}(1) \approx 0.939 \text{ bits} \quad (\text{gain} \approx 0.811 \text{ bits}).$$

(c) Test A is better (expected remaining entropy $0.75 < 0.939$), even though its "no" answer leaves more uncertainty than nothing — note the contrast with the tempting "split the outcomes in half" heuristic, which ignores the prior. Before running the test the answer is random, so the remaining entropy is a random variable; the principled score for a test is its *expectation*.

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

(Problem set problem)

Problem 6: KL Divergence by Hand

(a)
$$D(P\|Q) = 0.8 \log_2 \frac{0.8}{0.5} + 0.2 \log_2 \frac{0.2}{0.5} = 0.8(0.678) + 0.2(-1.322) \approx 0.278 \text{ bits}.$$

(b)
$$D(Q\|P) = 0.5 \log_2 \frac{0.5}{0.8} + 0.5 \log_2 \frac{0.5}{0.2} = 0.5 \log_2(0.625) + 0.5 \log_2(2.5) \approx 0.322 \text{ bits}.$$

KL divergence is *not symmetric*: $D(P\|Q) \neq D(Q\|P)$. It is not a true "distance" — the direction (which distribution is reality vs. model) matters.

(c) $D(P\|Q) = 0$ exactly when $P = Q$. If your model is exactly right, there is no *excess* surprise: every outcome is exactly as surprising under the model as under reality.

Challenge Problem: KL Between Poissons (pset6: kl_poisson)

(Problem set problem)