INFORMATION THEORY AND APPLICATIONS IN COMPUTER SCIENCE

Homework 1: Descriptions and Decisions

Fall 2026

Release: September 30, 2026.
Initial submission due: Monday, October 12.
Final revision due: Monday, October 19.

Submission. Upload your initial submission and final revision to Gradescope. Follow the course policies for collaboration, attribution, and permitted resources, and explain your reasoning clearly.

Problem 1. Secret Communication [25 points]

A sender and receiver share a secret key. The sender uses it to transform a message into a *ciphertext*: an encoded representation that can be sent over a public channel. The receiver uses the key to try to recover the message from the ciphertext, while an observer sees only the ciphertext. We will ask how much uncertainty the key must contain to keep the message secret, and how the answer changes when some messages are much more likely than others.

A message M takes values in a set $\mathcal{M}$ of size $N \geq 2$. A key K is independent of M. An encoder, possibly randomized, produces a ciphertext C according to a fixed conditional distribution $P_{C|M,K}$. A deterministic decoder computes $\widehat{M} = d(C, K)$. Perfect secrecy means $I(M; C) = 0$; zero-error decoding means $\Pr[\widehat{M} \neq M] = 0$.

(a) **[6 points]** Under perfect secrecy and zero-error decoding, prove $H(K) \geq H(M)$. Your proof should identify where secrecy and correctness are used.

(b) **[9 points]** Suppose also that $P_M(m) > 0$ for every $m \in \mathcal{M}$. Prove the stronger conclusion $H(K) \geq \log N$, even if $H(M)$ is much smaller. The encoder and decoder must remain fixed in your argument. Explain why full support matters by giving a counterexample without it. Give an example attaining the bound for every full-support message distribution.

(c) **[10 points]** Now assume M is uniform on $\mathcal{M}$, but allow $I(M; C) \leq \eta$ and decoding error at most δ, where $0 \leq \delta \leq 1 - 1/N$. Write $h_2(t) = -t \log t - (1-t)\log(1-t)$ for binary entropy, with the usual endpoint values. Prove

$$H(K) \geq \log N - \eta - h_2(\delta) - \delta \log(N-1).$$

Deduce an explicit bound on $\log |\operatorname{supp} K|$, where $\operatorname{supp} K$ is the set of keys with positive probability. Why does this argument not justify replacing $H(M)$ by $\log N$ for a nonuniform full-support message under only these approximate assumptions?

Problem 2. Universal Encoding [25 points]

An encoder may know several plausible models for its data without knowing which one is correct. We want a single code whose extra expected length is small for every model, regardless of which is actually used. This problem relates that cost to how much one observation can reveal about the model that generated it.

Let $P_1, \dots, P_k$ be known distributions, each strictly positive on the same finite alphabet $\mathcal{A}$, with $|\mathcal{A}| \geq 2$. A single binary prefix code must serve all of them. For a probability vector $w = (w_1, \dots, w_k)$, let $J \sim w$ and, conditionally on $J = j$, sample $X \sim P_j$. Define

$$Q_w = \sum_j w_j P_j, \qquad F(w) = I(J; X), \qquad C = \max_w F(w).$$

Here C is just notation for this optimization, not an assumed coding theorem. Continuity on the compact probability simplex ensures that a maximizer w^* exists; write $Q^* = Q_{w^*}$. Zero coordinates of w^* are allowed.

(a) **[4 points]** Prove

$$F(w) = \sum_j w_j D(P_j \| Q_w) = H(Q_w) - \sum_j w_j H(P_j).$$

(b) **[7 points]** Prove that $D(P_j \| Q^*) \leq C$ for every j, with equality whenever $w_j^* > 0$. Justify the bound using a feasible perturbation of w^* and a one-sided derivative. You may differentiate finite sums and use the derivative formula in the separate hints handout.

(c) **[9 points]** For a prefix code with lengths ℓ, its redundancy under model j is $\mathbb{E}_{X \sim P_j}[\ell(X)] - H(P_j)$. Define

$$R_{\text{pref}} = \inf_{\text{binary prefix codes}} \max_j \left\{ \mathbb{E}_{X \sim P_j}[\ell(X)] - H(P_j) \right\}.$$

Prove $C \leq R_{\text{pref}} < C + 1$. In the lower bound, do not assume that a prefix code's Kraft sum is exactly one. In the upper bound, give an actual feasible length assignment.

(d) **[5 points]** Determine C and R_{pref} when the two models are Bernoulli(p) and Bernoulli($1 - p$), with $0 < p < 1/2$. Explain the meaning of C in terms of uncertainty about the model resolved by one observation; do not assume the model can be recovered exactly.

Optional extension (not graded). Prove $\min_Q \max_j D(P_j \| Q) = C$, where Q ranges over all distributions on $\mathcal{A}$, including distributions with a zero-probability symbol.

Problem 3. Selection bias and information [25 points]

If we inspect several statistics on the same data and then choose one to report, the chosen statistic can look unusually good simply because of that selection. We will quantify the difference from choosing the index before seeing the data with the same index frequencies, and ask how much information the selection needs to create that bias.

Let $Z \sim P_Z$ be a finite data set. A procedure observes Z and selects an index $J \in [m]$, possibly using independent random coins. For each j, a fixed function $f_j : \mathcal{Z} \to [0,1]$ is specified before Z is sampled. Write $\mu_j = \mathbb{E}[f_j(Z)]$, and define the expected selection bias

$$B = \mathbb{E}[f_J(Z)] - \mathbb{E}[\mu_J].$$

(a) [5 points] For distributions P, Q on a finite set and g valued in $[0,1]$, prove $|\mathbb{E}_P g - \mathbb{E}_Q g| \leq \mathrm{TV}(P,Q)$. Identify a function attaining the upper bound for the signed difference $\mathbb{E}_P g - \mathbb{E}_Q g$. Here $\mathrm{TV}(P,Q) = \frac{1}{2}\sum_x |P(x) - Q(x)|$.

(b) [10 points] Prove

$$|B| \leq \sqrt{\frac{\ln 2}{2} I(Z;J)}.$$

If the procedure first computes a finite summary W of the data and selects J using only W and fresh independent coins, deduce the same bound with $I(Z;W)$ in place of $I(Z;J)$.

(c) [10 points] To test sharpness, relabel two indices by $0, 1$. Let Z be a uniform bit, choose J so that $\Pr[J = Z \mid Z] = 1/2 + t$ for $0 < t < 1/2$, and set $f_j(z) = \mathbf{1}\{z = j\}$. Compute B and $I(Z;J)$. Show that $|B|/\sqrt{(\ln 2) I(Z;J)/2} \to 1$ as $t \downarrow 0$, and deduce that no universal constant c gives $|B| \leq c I(Z;J)$. You may use the binary-entropy expansion in the separate hints handout without proving it.

Optional extension (not graded). Let Z be uniform on $[N]$, $N \geq 2$. Independently, flip a flag with success probability $a \in (0,1)$; reveal $W = Z$ on success and set $W = \perp$ otherwise. Compute $I(Z;W)$ and $\mathrm{TV}(P_{Z|W=w}, P_Z)$ for every possible w. Use this example to disprove a universal pointwise bound

$$\mathrm{TV}(P_{Z|W=w}, P_Z) \leq c\sqrt{I(Z;W)} \quad \text{for all positive-probability } w.$$

Then prove the valid bound, for every finite joint distribution of (Z,W) and every $u > 0$,

$$\Pr_W\{\mathrm{TV}(P_{Z|W}, P_Z) \geq u\} \leq \min\left\{1, \frac{(\ln 2) I(Z;W)}{2u^2}\right\}.$$

Problem 4. How large can a database join be? [25 points]

Suppose three tables record which pairs of attributes are compatible. A full record is valid only if all three of its pairs appear in the corresponding tables. How many valid records can there be? We will use entropy to answer this counting question without enumerating all the records.

Three finite tables consist of pairs: $R \subseteq \mathcal{A} \times \mathcal{B}$, $S \subseteq \mathcal{B} \times \mathcal{C}$, and $T \subseteq \mathcal{A} \times \mathcal{C}$. Their join is

$$\mathcal{J} = \{(a, b, c) : (a, b) \in R,\ (b, c) \in S,\ (a, c) \in T\}.$$

(a) **[7 points]** For arbitrary finite random variables A, B, C, prove $2H(A, B, C) \leq H(A, B) + H(A, C) + H(B, C)$ directly from chain rules and nonnegativity of information. Find an exact expression for the difference as a sum of nonnegative mutual-information or conditional-mutual-information terms.

(b) **[7 points]** Deduce $|\mathcal{J}| \leq \sqrt{|R||S||T|}$.

(c) **[5 points]** For each positive integer q, construct three tables of size q^2 with a join of size q^3.

(d) **[6 points]** Let A, B be independent uniform bits and let $C = A \oplus B$, where $\oplus$ is addition modulo two. Check that every pair is independent, but that the inequality in (a) is strict. Explain which conditional-information terms can remain positive despite pairwise independence.

Optional challenge (not graded). When $\mathcal{J} \neq \varnothing$, characterize all equality cases in the square-root bound, including rows that participate in no joined triple. Also prove $|\mathcal{J}| \leq |R||S|$ and its two analogous bounds.

Problem A. List recovery and storage [Optional; not graded]

A search procedure need not return one exact answer to be useful: a short list containing the correct answer can sometimes be enough. We will ask how much information is needed to produce such a list. For a nonuniform source, a list's prior probability mass is often more informative than its number of entries.

Let X, Y have an arbitrary finite joint distribution. After observing $Y = y$, a procedure returns a list $L(y) \subseteq \mathcal{X}$. Define its success and its average prior mass by

$$s = \Pr[X \in L(Y)], \qquad q = \mathbb{E}_Y[P_X(L(Y))].$$

Assume $s \geq 1 - \delta$, $q \leq \alpha$, and $0 < \alpha \leq 1 - \delta \leq 1$, with $\alpha < 1$. Define binary divergence by $d_2(a\|b) = a \log(a/b) + (1 - a) \log((1 - a)/(1 - b))$, using continuous endpoint values when defined.

(a) Prove $I(X;Y) \geq d_2(1 - \delta\|\alpha)$, including justification for replacing the actual s, q by their bounds.

(b) Suppose X is uniform on $[N]$ and $\mathbb{E}|L(Y)| \leq L$, where $1 \leq L < N$ is an integer. Under $\delta \leq 1 - L/N$, show that, with h_2 denoting binary entropy,

$$I(X;Y) \geq \log N - h_2(\delta) - (1 - \delta) \log L - \delta \log(N - L).$$

Only the *average* list size is bounded; individual lists need not have size at most L.

(c) For every such N, L, δ, construct a joint distribution attaining equality in (b), with every list having size exactly L.

(d) Let $Y = (M, R)$, where R is independent of X and M is a stored message taking at most 2^b values. The message may depend on both X and R. Deduce a lower bound on b. Independent random coins used to form the output list may be included in R.

(e) Give a nonuniform source and a fixed singleton list with success 0.99 but zero revealed information. Explain why this does not contradict (a).