Hints

1(a). Expand $H(M \mid C)$ using the key K.

1(b). First show that $P_{C|M=m}$ is the same for every m. Then change only the message distribution to uniform, and check that both secrecy and zero-error decoding survive.

1(c). Apply Fano to the observation (C, K), then reuse the chain-rule step from (a). For the last question, consider a very unlikely message.

2(b). Use $w_t = (1-t)w^* + te_j$ for $0 \le t \le 1$, where e_j puts all weight on j. If $\sum_x v(x) = 0$ and $Q(x) > 0$, you may use

$$\frac{d}{dt} H(Q + tv)\Big|_{t=0} = -\sum_x v(x) \log Q(x).$$

Compare the derivative of $F(w_t)$ with $D(P_j \| Q^*) - C$, then use the weighted average in (a) to obtain equality on the positive weights.

2(c). For the lower bound set $\kappa = \sum_x 2^{-\ell(x)}$ and $Q_\ell(x) = 2^{-\ell(x)}/\kappa$. Split the logarithm at Q^* to establish

$$\sum_j w_j^* D(P_j \| Q_\ell) = C + D(Q^* \| Q_\ell).$$

For the upper bound, use Shannon lengths for Q^* and Kraft's converse.

3(a). Separate the points where $P(x) - Q(x)$ is positive and negative.

3(b). Compare P_{ZJ} with $P_Z P_J$ using $g(z, j) = f_j(z)$. Compute the expectation under each distribution before applying Pinsker.

3(c). You may use $1 - h_2(1/2 + t) = 2t^2/\ln 2 + o(t^2)$ without proving it.

4(a). Expand $I(A; B) + I(A; C \mid B) + I(B; C \mid A)$.

4(b). Handle the empty join first. Otherwise choose a joined triple uniformly and bound the entropy of each pair by the logarithm of its table size. The pair itself need not be uniform on that table.

4(c). Use all pairs from three sets of size q.

Optional problems and extensions

Problem 2 extension. Use the logarithm-splitting identity from (c). Handle zero probabilities separately.

Problem 3 extension. Apply Pinsker to each posterior, average its squared bound, and apply Markov's inequality. Fix $N = 2$ when disproving the pointwise claim.

Problem 4 challenge. Equality must hold in every nonnegative term in (a) and every support-size bound in (b). Start by showing that equality in (a) forces mutual independence; pairwise independence alone is insufficient.

A(a). Compare P_{XY} with $P_X P_Y$ on the event $X \in L(Y)$. Apply KL data processing to the indicator of that event.

A(c). Choose Y uniformly on a cyclic alphabet and let $L(y)$ be a rotating window of L symbols. Given $Y = y$, put mass $1 - \delta$ uniformly inside the window and mass δ uniformly outside it.

Theorem reminders

All random variables and alphabets below are finite. Logarithms, entropies, and divergences are in bits; ln is the natural logarithm. Use $0 \log 0 = 0$.

Entropy bounds and chain rules

For a random variable X,

$$0 \le H(X) \le \log |\operatorname{supp} X|.$$

The upper bound is attained exactly when X is uniform on its support. $H(X \mid Y) = 0$ exactly when X is a function of Y almost surely. For any X, Y, Z,

$$H(X, Y \mid Z) = H(X \mid Z) + H(Y \mid X, Z), \qquad H(X \mid Y, Z) \le H(X \mid Y).$$

Conditioning reduces entropy *on average*; a particular observed value can increase conditional uncertainty.

Divergence and mutual information

For distributions P, Q on a common alphabet,

$$D(P\|Q) = \sum_{x : P(x) > 0} P(x) \log \frac{P(x)}{Q(x)} \ge 0.$$

This is Gibbs' inequality; equality holds exactly when $P = Q$. If $P(x) > 0$ and $Q(x) = 0$ for some x, then $D(P\|Q) = +\infty$. Mutual information satisfies

$$\begin{aligned} I(X;Y) &= D(P_{XY}\|P_X P_Y) = H(X) - H(X \mid Y), \\ I(X;Y) &= \sum_{y : P_Y(y) > 0} P_Y(y) D(P_{X|Y=y}\|P_X), \\ I(X;Y \mid Z) &= H(X \mid Z) - H(X \mid Y, Z) \ge 0, \\ I(X;Y,Z) &= I(X;Y) + I(X;Z \mid Y). \end{aligned}$$

$I(X;Y) = 0$ exactly when X, Y are independent. $I(X;Y \mid Z) = 0$ exactly when X, Y are conditionally independent given Z, for every positive-probability value of Z.

Fano's inequality

Let X take values in an alphabet of size $N \ge 2$, let $\widehat{X} = g(Y)$ be an estimate taking values in that alphabet, and set $p_e = \Pr[\widehat{X} \ne X]$. Then

$$H(X \mid Y) \le h_2(p_e) + p_e \log(N-1), \qquad h_2(t) = -t \log t - (1-t) \log(1-t).$$

Here $h_2(0) = h_2(1) = 0$. The function $h_2(t) + t \log(N-1)$ is nondecreasing for $0 \le t \le 1 - 1/N$. Thus an error bound $p_e \le \delta \le 1 - 1/N$ permits replacing p_e by δ in the displayed upper bound.

Theorem reminders (continued)

Kraft's inequality and prefix coding

Nonnegative integer lengths $\ell(x)$ on a finite alphabet can be realized by a binary prefix-free code if and only if

$$\sum_x 2^{-\ell(x)} \leq 1.$$

The sum need not be exactly one. Every binary prefix-free code for a source with distribution P satisfies

$$\mathbb{E}_{X\sim P}[\ell(X)] \geq H(P).$$

For a strictly positive distribution Q, the *Shannon lengths*

$$\ell_Q(x) = \lceil -\log Q(x) \rceil$$

satisfy Kraft's inequality and hence can be realized by a binary prefix-free code. Its expected length under Q obeys

$$H(Q) \leq \mathbb{E}_{X\sim Q}[\ell_Q(X)] < H(Q) + 1.$$

Data processing

If $X \to Y \to Z$ is a Markov chain, meaning that X, Z are conditionally independent given Y, then

$$I(X;Z) \leq I(X;Y).$$

For distributions P, Q and the *same* stochastic channel $K(y \mid x)$ applied to both, let $(PK)(y) = \sum_x P(x)K(y \mid x)$ and define QK similarly. Then

$$D(PK\|QK) \leq D(P\|Q).$$

In particular, applying the same deterministic function to both distributions cannot increase divergence.

Pinsker's inequality (in bits)

For distributions P, Q on a common finite set, define $\mathrm{TV}(P,Q) = \frac{1}{2}\sum_x |P(x) - Q(x)|$. Then

$$\mathrm{TV}(P,Q) \leq \sqrt{\frac{\ln 2}{2} D(P\|Q)}.$$

The factor $\ln 2$ is needed because D uses base-two logarithms.

Markov's inequality

If $V \geq 0$ and $u > 0$, then

$$\Pr[V \geq u] \leq \min\left\{1, \frac{\mathbb{E}[V]}{u}\right\}.$$