INFORMATION THEORY AND APPLICATIONS IN COMPUTER SCIENCE

Lecture 5 | 14 October 2026

Student notes

5 Channels, capacity, and the random-coding proof

Overview and prerequisites. We prove the finite discrete memoryless channel coding theorem, including an explicit random-coding achievability argument. The prerequisites are entropy and information chain rules, data processing, Fano's inequality, and Chebyshev's inequality. Types and joint-typicality counting are not required. The main converse proof bounds error away from zero above capacity. Optional reading develops the strong converse, feedback and acknowledgments, and source–channel separation. All logarithms are base two, except ln, and information is measured in bits. We write $h_2(p) = -p \log p - (1-p)\log(1-p)$, with its continuous endpoint values.

Source coding removes redundancy; channel coding introduces redundancy to survive noise. A useful code must balance reliability against the number of messages represented by a block. We now show that the balance is governed exactly, to first order, by mutual information. The converse uses the information inequalities from Lectures 1–3; the construction uses concentration and the probabilistic method.

5.1 The model and the theorem

A *finite discrete memoryless channel* (DMC) has finite input and output alphabets $\mathcal{X}, \mathcal{Y}$ and transition probabilities $W(y \mid x)$. Its n-use transition distribution is

$$W^n(y^n \mid x^n) = \prod_{i=1}^{n} W(y_i \mid x_i).$$

The channel acts independently at different coordinates once the input word is fixed. This does not require the input coordinates to be independent.

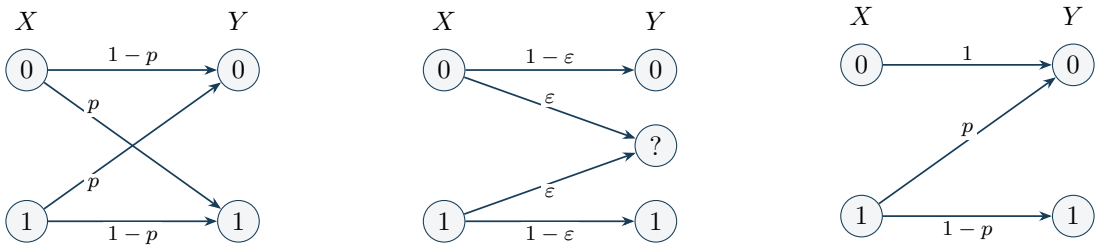
Examples to keep in mind. A binary symmetric channel BSC(p) flips each input bit with probability p, without marking the error. A binary erasure channel BEC(ε) replaces each bit by ? with probability ε; the receiver knows which positions were lost. We calculate both capacities in Section 5.2.

A *Z-channel* is asymmetric: zero always arrives as zero, while one becomes zero with probability p:

$$W(0 \mid 0) = 1, \qquad W(0 \mid 1) = p, \qquad W(1 \mid 1) = 1 - p.$$

This asymmetry motivates optimizing the input distribution: uniform input need not be best. These three channels are memoryless when fresh independent noise is used at each coordinate.

Binary symmetric channel **Binary erasure channel** **Z-channel**

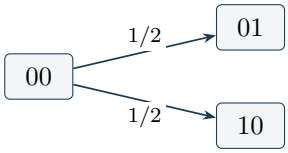


Each arrow is labeled by $W(y \mid x)$; omitted transitions have probability zero.

Noise with a fixed error budget. Bounded Hamming noise allows $y^n = x^n \oplus z^n$ with $\mathrm{wt}(z^n) \le t$: an adversary chooses the flip locations, and the decoder must succeed for every allowed choice. This is a relation between input and output words, with a global error budget. It does not specify probabilities for the possible outputs. Lecture 6 develops this model.

For a stochastic example that is not memoryless at the bit level, choose exactly t positions uniformly to flip, independently of the input. Each position has flip probability t/n, but for $0 < t < n$ the flips are dependent because their total is fixed. With $n = 2, t = 1$, the noise is 01 or 10, each with probability 1/2. A BSC(1/2) instead assigns probability 1/4 to each of $00, 01, 10, 11$. Thus the fixed-budget transition distribution does not factor into its one-bit marginals. The channel theorem below concerns the memoryless stochastic model.

Exactly one uniformly chosen flip in a two-bit block



For input 00, outputs 00 and 11 are impossible.
Under a BSC(1/2), all four output words have probability 1/4.

Binary message convention. An (n, k) code sends a k-bit message through n uses of the channel, using deterministic maps

$$\mathrm{Enc} : \{0,1\}^k \to \mathcal{X}^n, \qquad \mathrm{Dec} : \mathcal{Y}^n \to \{0,1\}^k, \qquad R_n = \frac{k}{n}.$$

For a binary-input channel, $\mathcal{X} = \{0, 1\}$, so the encoder maps k bits to n bits. For the BSC, $\mathcal{Y} = \{0, 1\}$; for the BEC, $\mathcal{Y} = \{0, 1, ?\}$. The decoder receives channel outputs, which need not be binary strings. The message alphabet is always $\{0,1\}^k$, even when the channel alphabets are more general. We allow $k = 0$, when there is a single message, the empty string.

The encoder chooses the entire word before transmission; there is no feedback. Let M be uniform on $\{0,1\}^k$. Alice transmits $X^n = \mathrm{Enc}(M)$, and Bob decodes $\hat{M} = \mathrm{Dec}(Y^n)$. The *average error* is

$$P_e = \Pr\{\hat{M} \neq M\}.$$

The probability averages over the message and the channel noise. The codebook is fixed and known to both parties; randomness used to prove its existence will not be a communication resource.

For an input distribution P, write $Q(y) = \sum_x P(x)W(y \mid x)$ and $I(P, W) = I(X; Y)$ under the joint distribution $P(x)W(y \mid x)$. Define

$$C = \max_P I(P, W).$$

The maximum exists because the probability simplex is compact and

$$I(P,W) = H(Q) - \sum_x P(x)H(W(\cdot \mid x))$$

is continuous. This includes zero probabilities, since $-u \log u$ extends continuously to $u = 0$. The input distribution is a design choice. Uniform messages need not produce uniform input coordinates; the maximization chooses the best distribution for one channel use.

Theorem 5.1 (Finite DMC coding theorem)**.** *For every $0 \le R < C$, there is a sequence of deterministic codes with $R_n \to R$ and $P_e \to 0$. Conversely, every (n,k) code satisfies*

$$\boxed{(1-P_e)k \le nC + 1.}$$

Consequently, any sequence with vanishing error has $\limsup_n R_n \le C$.

Thus C is the supremum of reliable asymptotic rates. The proof does not give an efficient method for constructing or decoding the codes, or sharp finite-blocklength guarantees. It establishes reliability at every fixed rate below C but does not settle reliability at a rate constrained to equal C at every blocklength. Rate zero is always possible using a single message, including when $C = 0$.

5.2 Two capacity calculations

For the binary symmetric channel BSC(p), $Y = X \oplus Z$ with $Z \sim$ Bernoulli(p) independent of X. If $X \sim$ Bernoulli(q), then $\Pr\{Y = 1\} = p + q(1-2p)$, so

$$I(X;Y) = H(Y) - H(Y \mid X) = h_2\big(p + q(1-2p)\big) - h_2(p) \le 1 - h_2(p).$$

With $q = 1/2$, the output is uniform and equality holds. Therefore

$$C\big(\mathrm{BSC}(p)\big) = 1 - h_2(p), \qquad 0 \le p \le 1.$$

At $p = 1/2$ the output is independent of the input and capacity is zero. At $p = 1$ every bit flips deterministically, so the receiver can invert it and capacity is one.

For the binary erasure channel BEC(ε), an input bit is replaced by ? with probability ε, independently of its value, and is otherwise delivered unchanged. An unerased output determines X; observing ? leaves the input distribution unchanged. Thus, for every input distribution,

$$H(X \mid Y) = \varepsilon H(X), \qquad I(X;Y) = (1-\varepsilon)H(X) \le 1 - \varepsilon.$$

A uniform input attains equality, giving $C(\mathrm{BEC}(\varepsilon)) = 1 - \varepsilon$. This includes the noiseless and all-erased endpoints.

Example 5.2 (Repetition and majority decoding)**.** Let $m \ge 1$ be odd. On a BSC(p) with $0 \le p < 1/2$, encode one message bit by

$$\mathrm{Enc}(0) = 0^m, \qquad \mathrm{Enc}(1) = 1^m,$$

and let Dec output the majority bit. This is an $(m,1)$ code, with rate $1/m$.

Let F_m be the number of flipped positions. For either transmitted bit, decoding fails exactly when at least $(m+1)/2$ positions flip. Since $F_m \sim \mathrm{Binomial}(m,p)$, the error probability is

$$\boxed{e_m = \Pr\{F_m \ge (m+1)/2\} = \sum_{j=(m+1)/2}^{m} \binom{m}{j} p^j (1-p)^{m-j}.}$$

The error is the same for both messages, so both the average error and the largest conditional error over messages equal e_m.

A short *Chernoff bound* shows that the error decays exponentially. For $0 < p < 1/2$, set $\lambda = \ln((1-p)/p) > 0$. On the event $F_m \geq m/2$, $e^{\lambda(F_m - m/2)} \geq 1$, so

$$\begin{aligned} e_m &\leq \Pr\{F_m \geq m/2\} \leq \mathbb{E}\big[e^{\lambda(F_m - m/2)}\big] \\ &= \big((1-p)e^{-\lambda/2} + pe^{\lambda/2}\big)^m = \big(2\sqrt{p(1-p)}\big)^m. \end{aligned}$$

The product formula uses independence of the m flips. Since $2\sqrt{p(1-p)} < 1$, the bound tends to zero exponentially in m. At $p = 0$, the error is zero. For each fixed $p < 1/2$, this proves reliable transmission of one bit, but the rate $1/m$ tends to zero. The channel theorem promises vanishing error while maintaining a positive rate below capacity.

Alternatively, encode k message bits in separate groups of m repetitions. This is an (mk, k) code, with rate $1/m$. Independent channel noise makes the probability of at least one incorrectly decoded message bit equal to

$$1 - (1 - e_m)^k.$$

For fixed m and $0 < p < 1/2$, this tends to one as $k \to \infty$. A small fixed error per group therefore does not ensure reliable recovery of a growing message.

Example 5.3 (Visible loss and hidden corruption)**.** An erasure probability of 0.1 leaves capacity 0.9 bits per use. A flip probability of 0.1 leaves only

$$1 - h_2(0.1) = 1 + 0.1 \log 0.1 + 0.9 \log 0.9 \approx 0.5310.$$

The receiver knows the positions of erasures; it must infer the positions of flips. On the BSC(0.1), rate 1/2 admits vanishing error, while rate 3/4 does not. Indeed, the converse gives the quantitative bound

$$P_e \geq 1 - \frac{C}{3/4} - \frac{1}{(3/4)n} \approx 0.2920 - \frac{4}{3n}$$

for codes with $k/n = 3/4$. Here P_e is the probability of an error in the entire decoded message.

5.3 Converse: bounding the information carried by a block

The converse starts with an arbitrary fixed code. Reliable decoding means that observing Y^n must resolve almost all k bits of message uncertainty. We compare that requirement with the at most nC bits that the channel can supply. The two ingredients are Fano's inequality and an entropy calculation that reduces the whole block to one channel use at a time.

Proof of the converse in Theorem 5.1. The case $k = 0$ is immediate, so assume $k \geq 1$.

Step 1: bound the uncertainty left after observing the output. Let $B = \mathbf{1}_{\{\hat{M} \neq M\}}$ indicate a decoding error. Since $\hat{M} = \mathrm{Dec}(Y^n)$, knowing M and Y^n determines B. The conditional entropy chain rule therefore gives

$$\begin{aligned} H(M \mid Y^n) &= H(B, M \mid Y^n) \\ &= H(B \mid Y^n) + H(M \mid B, Y^n). \end{aligned}$$

The first term is at most $H(B) = h_2(P_e)$. When $B = 0$, the output determines M exactly: it equals $\mathrm{Dec}(Y^n)$. When $B = 1$, the true message must be one of the other $2^k - 1$ messages. Entropy is at most the logarithm of the number of possibilities, so

$$H(M \mid Y^n) \leq h_2(P_e) + P_e \log(2^k - 1) \leq 1 + P_e k.$$

This is Fano's inequality in the form we need. Conditional terms on events of probability zero are omitted.

Because M is uniform on $\{0,1\}^k$, $H(M) = k$. Also $M \to X^n \to Y^n$ is a Markov chain: once the transmitted word is known, the channel output has distribution $W^n(\cdot \mid X^n)$ independently of which message produced it. By data processing,[1]

$$k = I(M;Y^n) + H(M \mid Y^n) \le I(X^n;Y^n) + 1 + P_e k.$$

Thus it remains to show $I(X^n;Y^n) \le nC$.

Step 2: bound how much information a block can carry. The input coordinates need not be independent: they are all functions of the same message. For the unconditional output entropy we use the fact that conditioning cannot increase entropy:

$$H(Y^n) = \sum_{i=1}^{n} H(Y_i \mid Y^{i-1}) \le \sum_{i=1}^{n} H(Y_i).$$

For the conditional output entropy, memorylessness gives an equality. For every fixed input word x^n of positive probability, $Y_1, \dots, Y_n$ are independent, with coordinate distributions $W(\cdot \mid x_i)$. Hence

$$H(Y^n \mid X^n = x^n) = \sum_{i=1}^{n} H(W(\cdot \mid x_i)).$$

Average this identity over the actual input distribution:

$$\begin{aligned} H(Y^n \mid X^n) &= \sum_{x^n} \Pr\{X^n = x^n\} \sum_{i=1}^{n} H(W(\cdot \mid x_i)) \\ &= \sum_{i=1}^{n} \sum_{x} \Pr\{X_i = x\} H(W(\cdot \mid x)) \\ &= \sum_{i=1}^{n} H(Y_i \mid X_i). \end{aligned}$$

Subtracting this equality from the unconditional entropy bound gives

$$\begin{aligned} I(X^n;Y^n) &= H(Y^n) - H(Y^n \mid X^n) \\ &\le \sum_{i=1}^{n} \big[H(Y_i) - H(Y_i \mid X_i)\big] \\ &= \sum_{i=1}^{n} I(X_i;Y_i). \end{aligned}$$

Let P_i denote the marginal distribution of X_i. Since the ith channel use still has transition probabilities W, $I(X_i;Y_i) = I(P_i, W) \le \max_P I(P,W) = C$. Summing proves $I(X^n;Y^n) \le nC$ for this arbitrary code.

Step 3: combine the requirement and the budget. Substitute the block information bound into Step 1 and rearrange:

$$k \le nC + 1 + P_e k, \qquad \text{so} \qquad (1 - P_e)k \le nC + 1.$$

[1] For any Markov chain $A \to B \to C$, conditional independence gives

$$P_{ABC}(a,b,c) = P_B(b) P_{A|B}(a \mid b) P_{C|B}(c \mid b).$$

Thus we can equivalently sample the middle variable B first, then sample A and C independently conditional on B. This reproduces the original joint distribution. Here, first sample X^n from the marginal induced by the fixed code, then draw M from $P_{M|X^n}$ and Y^n from $W^n(\cdot \mid X^n)$ independently conditional on X^n. In this equivalent experiment, M is a randomized post-processing of X^n, so data processing gives $I(Y^n;M) \le I(Y^n;X^n)$, which is the inequality used here by symmetry of mutual information. No deterministic inverse of the encoder is needed.

For a sequence with $P_e \to 0$, eventually $P_e < 1$, and division by n gives

$$R_n = \frac{k}{n} \le \frac{C + 1/n}{1 - P_e}.$$

The right side tends to C, proving $\limsup_n R_n \le C$. □

This reduction to one-use quantities is called *single-letterization*. The output-entropy inequality allows arbitrary dependence among input coordinates, so the bound applies to every code.

The Fano argument is a *weak* converse. If $R_n \to R > C$, then rearranging its bound gives

$$P_e \ge 1 - \frac{C + 1/n}{R_n}, \qquad \liminf_n P_e \ge 1 - \frac{C}{R} > 0.$$

Thus the error cannot vanish above capacity. The optional strong-converse argument after the main proof strengthens this to $P_e \to 1$.

5.4 Achievability: a likelihood test and a union bound

We now construct a code by choosing its codewords randomly and proving that at least one choice works. The decoder will test how well each candidate codeword explains the received output. There are two ways this can fail: the true word can score too low, or a competing word can score too high. The true word generates the output through the channel; a competing word is drawn independently of that output in the random experiment. The test distinguishes these two situations.

Information density: the score used by the decoder

Fix an input distribution P, and let $Q = PW$. For $P(x) > 0$ and $Q(y) > 0$, define the *information density*

$$\imath(x; y) = \log \frac{W(y \mid x)}{Q(y)},$$

using $\log 0 = -\infty$. At points with $P(x) = 0$ or $Q(y) = 0$, set it to $-\infty$; those points have zero probability under PW. Write $P^n(x^n) = \prod_{i=1}^n P(x_i)$ and $Q^n(y^n) = \prod_{i=1}^n Q(y_i)$ for the corresponding product distributions. For $P^n(x^n) > 0$ and $Q^n(y^n) > 0$, the block information density is

$$\imath_n(x^n; y^n) = \log \frac{W^n(y^n \mid x^n)}{Q^n(y^n)} = \sum_{i=1}^n \imath(x_i; y_i),$$

and is $-\infty$ otherwise. A large value means that the proposed input makes this output much more likely than it is under the background distribution Q^n. For a possible pair with positive joint probability, the one-use ratio is also

$$\frac{W(y \mid x)}{Q(y)} = \frac{P(x)W(y \mid x)}{P(x)Q(y)}.$$

The numerator is the probability of the pair when the output is generated from that input. The denominator is its probability when input and output are drawn independently with the same marginals. Information density measures evidence for the first experiment relative to the second.

Mutual information is the average information density:

$$\mathbb{E}[\imath(X; Y)] = \sum_{x,y:P(x)W(y|x)>0} P(x)W(y \mid x) \log \frac{W(y \mid x)}{Q(y)} = I(P, W).$$

An individual value can be negative. We need concentration of the sum, not positivity of each term. Dividing by $Q^n(y^n)$ lets one threshold compare outputs with different background probabilities.

The true word: concentration below its mean

We now reuse the averaging argument from Lecture 4's AEP proof. There the summands were $-\log P(X_i)$; here they are the single-use information densities

$$Z_i := \imath(X_i;Y_i) = \log \frac{W(Y_i \mid X_i)}{Q(Y_i)}, \qquad \imath_n(X^n;Y^n) = \sum_{i=1}^n Z_i.$$

Under the joint distribution P^nW^n of a transmitted pair,

$$\Pr\{X^n = x^n, Y^n = y^n\} = \prod_{i=1}^n P(x_i)W(y_i \mid x_i).$$

Thus the pairs (X_i, Y_i) are independent and identically distributed, and so are the Z_i. Within a pair, X_i and Y_i need not be independent. The expectation calculation above gives $\mathbb{E}Z_i = I(P,W)$.

Every pair that can occur has $P(x)W(y \mid x) > 0$, and hence $Q(y) \geq P(x)W(y \mid x) > 0$. Its logarithmic ratio is therefore finite. There are only finitely many such pairs, so Z_i ranges over a finite set of real numbers. In particular, $V := \mathrm{Var}(Z_i) < \infty$. Exactly as in Lecture 4, independence gives

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n Z_i\right] = I(P,W),$$
$$\mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^n Z_i\right) = \frac{1}{n^2}\sum_{i=1}^n \mathrm{Var}(Z_i) = \frac{V}{n}.$$

A downward deviation by at least γ is an absolute deviation by at least γ. Applying the same Chebyshev bound used in Lecture 4 therefore gives, for every $\gamma > 0$,

$$\Pr\{\imath_n(X^n;Y^n) \leq n(I(P,W) - \gamma)\} \leq \frac{V}{n\gamma^2}. \tag{5.1}$$

Thus the transmitted word passes any threshold below $nI(P,W)$ by a fixed positive amount per coordinate, with probability tending to one.

A competing word: a likelihood ratio has mean one

Fix any output word y^n with $Q^n(y^n) > 0$, and draw a candidate $\overline{X}^n \sim P^n$ independently of that output. Define the nonnegative random variable

$$A = \frac{W^n(y^n \mid \overline{X}^n)}{Q^n(y^n)} \geq 0.$$

Here only $\overline{X}^n$ is random. Averaging over it gives

$$\mathbb{E}A = \frac{\sum_{x^n} P^n(x^n)W^n(y^n \mid x^n)}{Q^n(y^n)}$$
$$= \frac{\prod_{i=1}^n \left(\sum_x P(x)W(y_i \mid x)\right)}{\prod_{i=1}^n Q(y_i)} = 1.$$

The numerator is precisely the output probability obtained by averaging the channel over an input drawn from P^n. This is why we normalized the test by $Q^n(y^n)$.

For any real threshold t, passing the test means $A > 2^t$. Pointwise,

$$\mathbf{1}_{\{A > 2^t\}} \leq 2^{-t}A.$$

If $A > 2^t$, the right side exceeds one; otherwise the left side is zero and the right side is nonnegative. Taking expectations proves

$$\boxed{\Pr\{\imath_n(\overline{X}^n; y^n) > t\} \le 2^{-t}\mathbb{E}A = 2^{-t}.} \tag{5.2}$$

This is Markov's inequality applied to A, not to its logarithm. If $A = 0$, the information density is $-\infty$ and the test fails, as it should. The estimate holds for every fixed output in the support of Q^n.

Assembling the random code

The concentration estimate puts the true score near $nI(P, W)$. The competing-word estimate bounds the total false-match probability by 2^{k-t} when there are 2^k messages. We therefore need a threshold strictly above the number of message bits and strictly below $nI(P, W)$. The gap $R < I(P, W)$ leaves room for both requirements.

Proof of achievability in Theorem 5.1. If $R = 0$, take a single message ($k_n = 0$) and let the decoder always return it. The error is zero. Now fix $0 < R < C$.

Step 1: choose a gap and a threshold. Choose one input distribution P such that $I := I(P, W) > R$, and keep it fixed as n grows. Such a choice exists because $C = \max_P I(P, W) > R$. Set

$$\gamma = \frac{I - R}{2} > 0, \qquad k_n = \lfloor nR \rfloor, \qquad t_n = n(I - \gamma) = n(R + \gamma).$$

Thus the threshold per coordinate lies halfway between R and I:

$$R < R + \gamma = I - \gamma < I, \qquad k_n \le nR.$$

The positive constant γ does not depend on n.

Step 2: specify the random encoder and its deterministic decoder. Draw a random codebook $\mathcal{C} = (C_u)_{u \in \{0,1\}^{k_n}}$, with all 2^{k_n} words drawn independently from P^n. Within each word, its n coordinates are also independent with distribution P. For any realized codebook $c = (c_u)_u$, set $\mathrm{Enc}_c(u) = c_u$.

Given y^n, form the set of passing messages

$$\mathcal{T}_c(y^n) = \{u \in \{0,1\}^{k_n} : \imath_n(c_u; y^n) > t_n\}.$$

If this set has exactly one element, Dec_c returns that message. Otherwise it returns a fixed default message, say 0^{k_n}. This defines the decoder for every possible output, including outputs with $Q^n(y^n) = 0$, on which every score is $-\infty$. The encoder and decoder are deterministic once c is fixed.

We do not condition the random words to be distinct. Repeated codewords are allowed by the definition of a code, and any errors they cause will be included in the analysis. Drawing the words independently is what makes the competing-word estimate available.

Step 3: bound the error for one message in the random experiment. Fix $u \in \{0,1\}^{k_n}$. First draw $\mathcal{C}$, then send C_u through the channel using fresh channel noise. Define

$$F_u = \{\imath_n(C_u; Y^n) \le t_n\}, \qquad G_{u,v} = \{\imath_n(C_v; Y^n) > t_n\} \quad (v \ne u).$$

If F_u does not occur and no $G_{u,v}$ occurs, the true message is the unique passing one, so decoding is correct. Therefore

$$\{\mathrm{Dec}_{\mathcal{C}}(Y^n) \ne u\} \subseteq F_u \cup \bigcup_{v \ne u} G_{u,v}.$$

This is an inclusion: the default guess can occasionally be correct even when one of the bad events occurs.

Because $C_u \sim P^n$ and the output is generated from C_u through W^n, the transmitted pair (C_u, Y^n) has distribution $P^n W^n$. Our choice $t_n = n(I - \gamma)$ lets us apply (5.1):

$$\Pr_{\mathcal{C},W}(F_u) \leq \frac{V}{n\gamma^2}.$$

The subscript emphasizes that this probability averages over the codebook as well as the channel; it is not yet the error of one fixed code.

Now fix a competitor $v \neq u$. The word C_v was sampled independently of C_u, and the channel noise used to form Y^n is independent of the whole codebook. Consequently C_v is independent of (C_u, Y^n). Even after conditioning on $Y^n = y^n$, the distribution of C_v is still P^n. Also $Y^n \sim Q^n$ in this experiment, so $Q^n(Y^n) > 0$ almost surely. For each output of positive probability, (5.2) yields

$$\Pr_{\mathcal{C},W}(G_{u,v} \mid Y^n = y^n) \leq 2^{-t_n}.$$

Averaging over the output yields $\Pr_{\mathcal{C},W}(G_{u,v}) \leq 2^{-t_n}$. This step fixes the output, not the entire codebook: after fixing the codebook, the competing word would no longer be a fresh draw from P^n.

There are $2^{k_n} - 1$ competitors. The union bound now gives

$$\begin{aligned}
\Pr_{\mathcal{C},W}\{\mathrm{Dec}_{\mathcal{C}}(Y^n) \neq u\} &\leq \Pr_{\mathcal{C},W}(F_u) + \sum_{v \neq u} \Pr_{\mathcal{C},W}(G_{u,v}) \\
&\leq \frac{V}{n\gamma^2} + (2^{k_n} - 1)2^{-t_n} \\
&\leq \frac{V}{n\gamma^2} + 2^{nR}2^{-n(R+\gamma)} \\
&= \frac{V}{n\gamma^2} + 2^{-n\gamma}.
\end{aligned}$$

The union bound does not require the competing events to be independent. They share the same output, so such independence should not be assumed.

Step 4: obtain one fixed codebook that works on average. For a fixed codebook c, let $e_u(c)$ be its channel error when message u is sent. Only channel noise is random in $e_u(c)$. The average error of that fixed code is

$$P_e(c) = 2^{-k_n} \sum_{u \in \{0,1\}^{k_n}} e_u(c).$$

Step 3 says that, for every u, $\mathbb{E}_{\mathcal{C}}[e_u(\mathcal{C})] \leq a_n$, where

$$a_n := \frac{V}{n\gamma^2} + 2^{-n\gamma}.$$

By linearity of expectation,

$$\mathbb{E}_{\mathcal{C}}[P_e(\mathcal{C})] = 2^{-k_n} \sum_u \mathbb{E}_{\mathcal{C}}[e_u(\mathcal{C})] \leq a_n.$$

The ensemble has finite support. At least one realized codebook c_n must therefore satisfy $P_e(c_n) \leq a_n$; if every supported codebook had larger error, their weighted average would be larger too. This selects one codebook for all messages. It guarantees average error, rather than a bound on each individual $e_u(c_n)$.

Choose such a codebook for every n. The resulting encoders and decoders need no fresh codebook randomness during communication. Since $V < \infty$ and $\gamma > 0$ are fixed, $a_n \to 0$. Finally,

$$0 \le R - \frac{k_n}{n} < \frac{1}{n},$$

so the rates tend to R. This proves both required conclusions for a sequence of deterministic (n, k_n) codes. □

There are fewer than 2^{nR} competitors, each passing with probability at most $2^{-n(R+\gamma)}$. Their total contribution vanishes, while the actual word passes because its mean information density is $R + 2\gamma$.

5.5 Reading the proof as a finite-block design rule

The same calculation works with other thresholds. For any input distribution P, integer $k \ge 0$, and $s > 0$, choose threshold $t = k+s$. The true word fails with probability $\Pr\{\imath_n(X^n; Y^n) \le k+s\}$; each of the $2^k - 1$ competitors passes with probability at most $2^{-(k+s)}$. Averaging and fixing a codebook therefore gives a deterministic (n, k) code satisfying

$$P_e \le \Pr\{\imath_n(X^n; Y^n) \le k + s\} + 2^{-s}, \qquad (X^n, Y^n) \sim P^n W^n. \tag{5.3}$$

Indeed, the competitor contribution is

$$(2^k - 1)2^{-(k+s)} = (1 - 2^{-k})2^{-s} \le 2^{-s}.$$

The existence statement refers to the codebook selected by averaging; the probability on the right is evaluated under the product-input experiment, not under an arbitrary fixed code's input distribution. A higher threshold makes false matches less likely, but also makes the true word more likely to fail. In the asymptotic proof, the threshold per use t_n/n lies strictly between the target rate R and $I(P, W)$, so both error terms vanish.

Example 5.4 (Information density on a BSC)**.** For a uniform input and $0 < p < 1/2$, the output distribution is uniform. The information density equals $\log(2(1-p))$ when the bit survives and $\log(2p)$ when it flips. Its mean is

$$(1-p)\log(2(1-p)) + p\log(2p) = 1 - h_2(p),$$

and the two-valued variance formula gives

$$V = p(1-p)\left(\log\frac{1-p}{p}\right)^2.$$

Thus a likely correct pair earns positive evidence and an unlikely flipped pair earns negative evidence. If K positions flip in a block, then

$$\imath_n = n\log(2(1-p)) - K\log\frac{1-p}{p}.$$

A likelihood threshold is therefore a Hamming-distance test in this symmetric case. For a general channel it compares likelihood ratios rather than merely counting differing coordinates. The proof allows these asymmetric channels without altering the false-match estimate.

Random codebooks and fixed codes. The converse applies to every fixed code, whose input coordinates may be dependent. In the construction, the transmitted word for any fixed message has distribution P^n while we average over codebooks. After we select a codebook, Alice and Bob use fixed maps with no shared randomness. This argument proves existence but gives no efficient method to find or decode the selected codebook. For efficient codes approaching capacity, see *Polar Codes* (more coming).

5.6 Optional reading: the strong converse

We now strengthen the converse by bounding the probability of successful decoding directly. We sketch the additional capacity-optimality fact needed for this argument.

Remark 5.5 (Strong converse for the same finite DMC). If $R_n \to R > C$, then every sequence of (n, k_n) codes has $P_e \to 1$. One additional capacity-optimality fact is needed. If P^* achieves capacity and $Q^* = P^*W$, then

$$D(W(\cdot \mid x)\|Q^*) \le C \qquad \text{for every } x \in \mathcal{X}.$$

Indeed, Q^* is positive on every output the channel can produce: otherwise mixing a suitable input into P^* would improve mutual information by a positive term of order $t\log(1/t)$, while all other changes are $O(t)$. For $P_t = (1-t)P^* + t\delta_x$, differentiating at $t = 0$ then gives $\frac{d}{dt} I(P_t, W)\Big|_{0+} = D(W(\cdot \mid x)\|Q^*) - C \le 0$, proving the fact. Here δ_x is the distribution that puts all its mass on input x.

For the actual transmitted word and received output, set

$$L = \log \frac{W^n(Y^n \mid X^n)}{(Q^*)^n(Y^n)}.$$

Given any fixed input word, L is a sum of independent terms. Its conditional mean is at most nC and its conditional variance is at most nV_*, where

$$V_* := \max_x \mathrm{Var}_{Y\sim W(\cdot|x)} \left(\log \frac{W(Y \mid x)}{Q^*(Y)} \right) < \infty.$$

Only outputs with positive $W(Y \mid x)$ enter this variance.

Let $D_u = \{y^n : \mathrm{Dec}(y^n) = u\}$ for $u \in \{0,1\}^k$. For any $s > 0$, the part of successful decoding with $L \le k - s$ satisfies

$$\Pr\{\hat{M} = M,\ L \le k - s\} \le 2^{-k} \sum_{u\in\{0,1\}^k} 2^{k-s}(Q^*)^n(D_u) = 2^{-s}.$$

The decoder regions partition the output alphabet. Splitting success at this likelihood threshold therefore gives

$$\boxed{1 - P_e \le \Pr\{L > k - s\} + 2^{-s}.}$$

If $k/n \ge C + \Delta$ for a fixed $\Delta > 0$, choose $s = n\Delta/2$. Apply Chebyshev conditional on each input word, then average, to obtain

$$1 - P_e \le \frac{4V_*}{n\Delta^2} + 2^{-n\Delta/2} \longrightarrow 0.$$

This proves the strong converse. The same conditional summands are bounded because the channel alphabets are finite, so a Chernoff/Hoeffding bound can replace the $O(1/n)$ term by an exponentially decaying term. Concentration thus sharpens the speed of convergence; the new likelihood argument is what strengthens the conclusion beyond Fano.

5.7 Optional reading: feedback and acknowledgments

Feedback lets the sender react to earlier channel outputs. We show that the converse still applies, and then use acknowledgments to construct a simple coding scheme for the erasure channel.

The information budget. The channel converse survives interaction. Even if the receiver can send arbitrary messages back for free, and the sender adapts every next input, each new use of a memoryless forward channel contributes at most C bits.

The feedback model. The sender holds a uniform message $M \in \{0,1\}^k$; the receiver has no initial information about it. There are a fixed n uses of a finite DMC W on a fixed schedule. Between these uses, allow arbitrary causal noiseless messages from receiver to sender, at no charge. The only sender-to-receiver communication is through W. Both parties may use finite private and shared random seeds. Their joint collection U is sampled independently of M and all forward-channel noise.

For the converse, give the complete collection of seeds U to both parties at the start. This only gives them more power. Once U is fixed, the receiver's feedback messages are functions of its past outputs and earlier feedback. Recursively, that entire feedback transcript is determined by (U, Y^{i-1}) before use i. Thus every permitted protocol is covered by

$$X_i = \mathrm{Enc}_i(M, U, Y^{i-1}), \qquad \hat{M} = \mathrm{Dec}(U, Y^n), \qquad P_e = \Pr\{\hat{M} \neq M\}.$$

The decoder can ignore the additional seeds. Even conditional on the message, the seeds, and the entire past, the current output has distribution $W(\cdot \mid X_i)$: the forward channel uses fresh noise.

Proposition 5.6 (The channel converse survives interaction)**.** *Let $C = \max_P I(P, W)$ for a finite discrete memoryless channel W. Every protocol in the model above satisfies*

$$\boxed{(1 - P_e)k \leq nC + 1.}$$

The supremum of reliable asymptotic rates remains C: the sender can ignore feedback and use the codes already proved in this lecture.

Proof. The idea is to charge each new output for the information it adds, given everything already available to the strengthened receiver. Write $Z_i = (U, Y^{i-1})$. Fix $Z_i = z$ of positive probability, and let P_z be the conditional distribution of X_i. Given M, z, the input is fixed; given X_i, z, the output uses the same channel W. Therefore

$$\begin{aligned} I(M; Y_i \mid Z_i = z) &= H(Y_i \mid z) - H(Y_i \mid M, z) \\ &= H(Y_i \mid z) - H(Y_i \mid X_i, z) \\ &= I(P_z, W) \leq C. \end{aligned}$$

The seeds start independent of M, so $I(M; U) = 0$. Average over z and apply the information chain rule:

$$I(M; U, Y^n) = \sum_{i=1}^{n} I(M; Y_i \mid U, Y^{i-1}) \leq nC.$$

For $k \geq 1$, Fano now finishes the proof exactly as before:

$$k = I(M; U, Y^n) + H(M \mid U, Y^n) \leq nC + 1 + P_e k.$$

The case $k = 0$ is immediate. □

Example 5.7 (Repeat until received on the erasure channel)**.** Send a k-bit message through BEC(ε). Repeat its first bit until it arrives unerased, then do the same with the second bit, and so on. The receiver's outputs tell both parties when to advance. After all k bits have arrived, send padding until the fixed budget of n uses is exhausted. For example, to send 101:

input	1	1	0	0	1
output	?	1	?	0	1

the receiver recovers 101.

Let S_n count the unerased uses among all n transmissions, including padding. At each use the erasure indicator is a fresh Bernoulli variable, independent of the message, the past, and the current input. Hence the erasure indicators are independent even though the inputs are chosen adaptively, and

$$S_n \sim \text{Binomial}(n, 1-\varepsilon), \qquad \mathbb{E}S_n = n(1-\varepsilon), \qquad \text{Var}(S_n) = n\varepsilon(1-\varepsilon).$$

If $S_n \geq k$, the entire message is recovered. If there are fewer successes, the receiver makes a fixed default guess. For any fixed $\delta > 0$ and integer $k \geq 1$ with $k/n \leq 1 - \varepsilon - \delta$, Chebyshev gives

$$P_e \leq \Pr\{S_n < k\} \leq \frac{n\varepsilon(1-\varepsilon)}{(n(1-\varepsilon)-k)^2} \leq \frac{\varepsilon(1-\varepsilon)}{n\delta^2} \longrightarrow 0.$$

Thus this elementary feedback scheme achieves every positive rate below $1-\varepsilon$, the capacity already calculated in this lecture. Zero rate uses a single message. At $\varepsilon = 1$ there is no positive rate to achieve.

For $\varepsilon = 1/4$, capacity is $3/4$ bit per forward use. Sending $k = n/2$ bits for even n gives $P_e \leq 3/n$. This illustrates vanishing block error at a positive rate; the finite-block bound need not be sharp.

Takeaway. With free receiver-to-sender feedback, the forward information budget remains at most nC. On the erasure channel, acknowledgments show where redundancy is still needed, giving a simple code at every rate below the same capacity.

5.8 Optional reading: expurgation and separation

From average error to maximal error

The channel theorem controls average error for a uniform message. A source compressor usually produces a nonuniform index. To combine the two, we first make the channel code reliable for every message.

Proposition 5.8 (Expurgation)**.** *An (n, k) code with $k \geq 1$ and average error at most ε contains, after relabeling retained messages, an $(n, k-1)$ code with maximal error at most 2ε. For a code with k message bits, maximal error means*

$$\max_{u \in \{0,1\}^k} \Pr\{\mathrm{Dec}(Y^n) \neq u \mid M = u\}.$$

Consequently every $0 \leq R < C$ admits deterministic codes with rates tending to R and vanishing maximal error.

Proof. Write e_u for the error conditional on message u, so $\sum_{u \in \{0,1\}^k} e_u \leq 2^k \varepsilon$. If $\varepsilon = 0$, all errors vanish. If $\varepsilon > 0$, fewer than 2^{k-1} messages can have $e_u > 2\varepsilon$, since their sum would otherwise exceed $2^k \varepsilon$. Thus in either case we can retain exactly 2^{k-1} messages with error at most 2ε. Label them bijectively by $\{0,1\}^{k-1}$ and keep their codewords. Use the old decoder, reassign any output naming a deleted message to one retained message, and then apply the new labels. For a retained message, a previously correct output remains correct, so its error cannot increase. The rate decreases by exactly $1/n$. For $R > 0$, apply this procedure to the codes from Theorem 5.1; eventually their message lengths are at least one. Rate zero uses a single message and has zero error. □

Combining source and channel codes

Theorem 5.9 (Separation with an explicit resource ratio)**.** *Let $S_1, S_2, \ldots$ be independent copies of a source S on a finite alphabet $\mathcal{S}$, and let W be a finite DMC of capacity C. Suppose $\ell_n / n \to \rho$ with $0 < \rho < \infty$. Transmission uses n channel uses to reconstruct the entire source block S^{ℓ_n}, with no feedback.*

(i) If $\rho H(S) < C$, deterministic encoders and decoders with source block error tending to zero exist. They can be obtained by first compressing the source and then channel coding its compressed index.

(ii) If any such sequence of encoders and decoders has block error tending to zero, then $\rho H(S) \leq C$.

Proof. For achievability choose real numbers r, R with

$$H(S) < r, \qquad \rho r < R < C.$$

The source coding theorem of Lecture 4 provides fixed-length codes for ℓ source symbols using L_ℓ indices, with $\log L_\ell \leq \ell r + 1$ and source block error $a_\ell \to 0$. Proposition 5.8 provides (n, k_n) channel codes with $k_n / n \to R$ and maximal errors $b_n \to 0$. Since

$$\limsup_n \frac{\log L_{\ell_n}}{n} \leq \rho r < R = \lim_n \frac{k_n}{n},$$

eventually $L_{\ell_n} \leq 2^{k_n}$. Assign the compressed indices to distinct binary messages in $\{0,1\}^{k_n}$. If the channel decoder returns an unused message, let the source decoder output a fixed default source word.

Compress the source, transmit its binary message through this channel code, and apply the source decoder to the recovered index. If source compression is correct and the channel message is recovered correctly, the whole source block is correct. Therefore the composite error is at most $a_{\ell_n} + b_n \to 0$. The maximal-error guarantee bounds the second term for the actual, possibly nonuniform, distribution of compressed indices.

For the converse, consider any encoder $S^{\ell_n} \mapsto X^n$ and decoder $Y^n \mapsto \widehat{S}^{\ell_n}$ with block error e_n. The one-symbol source alphabet case is immediate. Otherwise Fano and independence of the source coordinates give

$$H(S^{\ell_n}) = \ell_n H(S), \qquad H(S^{\ell_n} \mid Y^n) \le 1 + e_n \ell_n \log |\mathcal{S}|.$$

Data processing and the single-letterization proved in this lecture imply

$$I(S^{\ell_n}; Y^n) \le I(X^n; Y^n) \le nC.$$

Adding the last two information terms yields

$$\ell_n H(S) \le nC + 1 + e_n \ell_n \log |\mathcal{S}|.$$

Divide by n, use $\ell_n/n \to \rho < \infty$ and $e_n \to 0$, and obtain $\rho H(S) \le C$. □

For example, a Bernoulli(0.1) source has entropy about 0.4690 bits per symbol, and a BEC(0.2) supplies 0.8 bits per use. Sending one source symbol per channel use satisfies the strict condition. More generally, separation achieves every positive ratio $\rho < 0.8/0.4690 \approx 1.706$ source symbols per channel use; any larger fixed ratio precludes vanishing block error. The units matter: comparing $H(S)$ directly with C silently assumes one source symbol per channel use.

Exercises

1. An input bit passes through a BSC(p) and then a BSC(q) with independent noises. Derive the effective flip probability and capacity. For $p, q \in [0, 1/2]$, verify that the capacity is at most that of either component, both from the formula and from data processing.

2. For a fixed input distribution P, message length k, and $s > 0$, repeat the proof with threshold $k + s$ to prove that some deterministic (n, k) code has

   $$P_e \le \Pr\{\imath_n(X^n; Y^n) \le k + s\} + 2^{-s},$$

   where $(X^n, Y^n) \sim P^n W^n$. Explain how the two terms respond in opposite ways when s increases.

3. For uniform input to a BSC(p), $0 < p < 1/2$, compute the two possible values of $\imath(X; Y)$ and prove

   $$V = p(1-p)\left(\log \frac{1-p}{p}\right)^2.$$

   Use the random-coding error bound to give a numerical sufficient blocklength for $P_e \le 0.1$ at $p = 0.1$, $R = 1/2$. Explain why a sufficient blocklength from this bound need not be close to the smallest feasible one.