

INFORMATION THEORY AND APPLICATIONS IN COMPUTER SCIENCE

Lecture 4 | 5 October 2026

Student notes

4 Average encoding and the compression threshold

Overview and prerequisites. Lecture 3 ended after Section 3.6, including Pinsker. We first use that inequality to prove average encoding, then prove the almost-lossless fixed-length compression threshold for a fixed finite-alphabet i.i.d. source. Above entropy the block error tends to zero; below entropy it tends to one for every code. We use the entropy/KL identities from Lectures 1–3, independence, and a second-moment inequality. Expected-length prefix coding is recalled for comparison with the new fixed-budget model.

4.1 Average encoding: completing the inference toolkit

We next compare dependence with independence. Let X be an input and W a message or observation correlated with it. If their mutual information is small, the joint experiment cannot be very different from one that draws X and W independently with the same marginals.

Theorem 4.1 (Average encoding). *For any finite joint distribution of (X, W) ,*

$$\begin{aligned} \text{TV}(P_{XW}, P_X P_W) &= \mathbb{E}_X [\text{TV}(P_{W|X}, P_W)] \\ &\leq \sqrt{\frac{\ln 2}{2} I(X; W)}. \end{aligned} \quad (4.1)$$

Moreover,

$$\mathbb{E}_X [\text{TV}(P_{W|X}, P_W)^2] \leq \frac{\ln 2}{2} I(X; W). \quad (4.2)$$

Proof. The direct argument for the mean. Apply Pinsker to the joint distribution and the product of its marginals. The identity from Lecture 2 gives

$$\text{TV}(P_{XW}, P_X P_W) \leq \sqrt{\frac{\ln 2}{2} D(P_{XW} \| P_X P_W)} = \sqrt{\frac{\ln 2}{2} I(X; W)}.$$

To identify this joint distance as a mean conditional distance, expand it:

$$\begin{aligned} \text{TV}(P_{XW}, P_X P_W) &= \frac{1}{2} \sum_{x,w} |P_X(x)P_{W|x}(w) - P_X(x)P_W(w)| \\ &= \sum_{x: P_X(x) > 0} P_X(x) \left(\frac{1}{2} \sum_w |P_{W|x}(w) - P_W(w)| \right) \\ &= \mathbb{E}_X [\text{TV}(P_{W|X}, P_W)]. \end{aligned}$$

The terms at zero-probability x contribute zero and require no conditional distribution. This is the short board proof of (4.1).

The conditional argument, which also gives the square bound. For every x with positive probability, conditional Pinsker says

$$\text{TV}(P_{W|x}, P_W)^2 \leq \frac{\ln 2}{2} D(P_{W|x} \| P_W).$$

If $P_{W|x}(w) > 0$, then $P_W(w) \geq P_X(x)P_{W|x}(w) > 0$, so this divergence is finite. Multiply by $P_X(x)$ and sum. The resulting KL average is

$$\begin{aligned} \sum_x P_X(x) D(P_{W|x} \| P_W) &= \sum_{x,w} P_{XW}(x,w) \log \frac{P_{W|x}(w)}{P_W(w)} \\ &= \sum_{x,w} P_{XW}(x,w) \log \frac{P_{XW}(x,w)}{P_X(x)P_W(w)} = I(X; W). \end{aligned}$$

This proves (4.2). It also proves the mean bound by Cauchy–Schwarz: with $T(x) = \text{TV}(P_{W|x}, P_W)$,

$$(\mathbb{E}T)^2 = \left(\sum_x \sqrt{P_X(x)} \sqrt{P_X(x)} T(x) \right)^2 \leq \left(\sum_x P_X(x) \right) \left(\sum_x P_X(x) T(x)^2 \right) = \mathbb{E}[T^2].$$

Taking square roots gives (4.1) again. Notice the distinction between the mean of the squares and the square of the mean. $\square$

What an independent replacement guarantees. In the real experiment, draw (X, W) according to its joint distribution. In the replacement experiment, draw $X \sim P_X$ and independently draw $\widetilde{W} \sim P_W$. In both experiments use the same test or algorithm and the same definition of success. For a deterministic success event B ,

$$|\Pr[(X, W) \in B] - \Pr[(X, \widetilde{W}) \in B]| \leq \text{TV}(P_{XW}, P_X P_W) \leq \sqrt{\frac{\ln 2}{2} I(X; W)}.$$

If the algorithm is randomized, let $a(x, w) \in [0, 1]$ be its conditional success probability and apply the bounded-score TV inequality from Lecture 3. Thus the same bound holds for randomized tests and any fixed score taking values in $[0, 1]$. For example, $I(X; W) \leq 0.01$ bits makes the change in success probability less than 0.059. This says nothing about whether the marginal P_W is known or efficiently samplable. Nor does it assert that W is useless on every fixed input.

Exceptional inputs and the role of averaging. Let $T(X) = \text{TV}(P_{W|X}, P_W)$. On the event $T(X) \geq t > 0$ its square is at least t^2 . Consequently

$$t^2 \Pr[T(X) \geq t] \leq \mathbb{E}[T(X)^2] \leq \frac{\ln 2}{2} I(X; W),$$

which gives

$$\Pr_X[\text{TV}(P_{W|X}, P_W) \geq t] \leq \min \left\{ 1, \frac{(\ln 2) I(X; W)}{2t^2} \right\}. \quad (4.3)$$

The probabilities here use the actual distribution of X . For a concrete caution, let X be Bernoulli(a) and let $W = X$. Then $I(X; W) = h_2(a) \rightarrow 0$ as $a \downarrow 0$, but on the rare input $X = 1$, $\text{TV}(P_{W|X=1}, P_W) = 1 - a \rightarrow 1$. On input 0 the distance is a , so the mean distance is $2a(1 - a) \rightarrow 0$. Small average information is fully consistent with a distinctive observation on a rare input.

From inference back to compression. The one-shot coding theorem bounds an *expected* description length. Applying it to a block of independent samples immediately gives a prefix code with expected length between $nH(P)$ and $nH(P) + 1$. But a storage device with a fixed budget asks a different question: how many bits suffice to recover the block with high probability? The answer is again entropy. The proof explains why: most of the probability lies on a set whose size has exponential rate $H(P)$.

Throughout, $X_1, \dots, X_n$ are independent with common distribution P on a finite alphabet $\mathcal{X}$. Write $X^n = (X_1, \dots, X_n)$ and $P^n(x^n) = \prod_{i=1}^n P(x_i)$. In expressions involving surprise, restrict to $\text{supp}(P)^n$; strings outside this support have probability zero.

The blocklength counts source symbols, whereas the description length counts bits; a rate is measured in bits per source symbol. We keep the distribution and alphabet fixed while the blocklength grows. The independence assumption is substantive. If every coordinate were a copy of a single random draw, one description of that draw would describe the whole block, and its entropy would not grow linearly with the number of coordinates.

4.2 A law of large numbers for surprise

Define the surprise $Z_i = \log(1/P(X_i))$. Its mean is $H(P)$, and its variance

$$V(P) = \mathbb{E}[(Z_i - H(P))^2]$$

is finite: the positive probabilities of a fixed finite distribution are bounded away from zero. Independence gives the key identity

$$-\frac{1}{n} \log P^n(X^n) = \frac{1}{n} \sum_{i=1}^n Z_i. \quad (4.4)$$

For $\varepsilon > 0$, define the *typical set*

$$A_{n,\varepsilon} = \left\{ x^n \in \text{supp}(P)^n : \left| -\frac{1}{n} \log P^n(x^n) - H(P) \right| \leq \varepsilon \right\}.$$

Thus every member satisfies

$$2^{-n(H(P)+\varepsilon)} \leq P^n(x^n) \leq 2^{-n(H(P)-\varepsilon)}. \quad (4.5)$$

Theorem 4.2 (Finite-alphabet asymptotic equipartition property). *For every $n \geq 1$ and $\varepsilon > 0$,*

$$\delta_n(\varepsilon) := P^n(A_{n,\varepsilon}^c) \leq \min \left\{ 1, \frac{V(P)}{n\varepsilon^2} \right\}.$$

Consequently the surprise per symbol converges in probability to $H(P)$.

Proof. The average in (4.4) has mean $H(P)$. Its variance is $V(P)/n$, because the cross terms in the variance of the sum vanish by independence. For any real random variable U ,

$$\mathbb{E}[(U - \mathbb{E}U)^2] \geq \varepsilon^2 \Pr\{|U - \mathbb{E}U| > \varepsilon\}.$$

Apply this inequality to the average of the Z_i . This proves the stated bound, which tends to zero for each fixed positive ε . $\square$

Only this elementary law of large numbers is needed here. The tolerance can also shrink: if $\varepsilon_n \rightarrow 0$ and $n\varepsilon_n^2 \rightarrow \infty$, then $\delta_n(\varepsilon_n) \rightarrow 0$. For example, $\varepsilon_n = n^{-1/4}$ works. The Chebyshev bound suffices to locate the asymptotic threshold; it need not give sharp blocklength estimates. When $V(P) = 0$, every possible block is typical.

Corollary 4.3 (Size of the typical set). *For every n and $\varepsilon > 0$,*

$$(1 - \delta_n(\varepsilon))2^{n(H(P) - \varepsilon)} \leq |A_{n,\varepsilon}| \leq 2^{n(H(P) + \varepsilon)}. \quad (4.6)$$

Proof. Summing the lower probability bound in (4.5) gives $1 \geq P^n(A_{n,\varepsilon}) \geq |A_{n,\varepsilon}|2^{-n(H(P) + \varepsilon)}$. Summing the upper probability bound gives $1 - \delta_n(\varepsilon) \leq |A_{n,\varepsilon}|2^{-n(H(P) - \varepsilon)}$. Rearrange both inequalities. $\square$

With a shrinking tolerance as above, taking logarithms and dividing by n in (4.6) shows $n^{-1} \log |A_{n,\varepsilon}| \rightarrow H(P)$. This is the precise meaning of “about $2^{nH(P)}$ strings.” It does not assert that the typical set is uniform: probabilities within it may differ by a factor $2^{2n\varepsilon}$.

4.3 Fixed-length, almost-lossless source coding

A block code using ℓ_n bits consists of maps

$$f_n : \mathcal{X}^n \longrightarrow \{0, 1\}^{\ell_n}, \quad g_n : \{0, 1\}^{\ell_n} \longrightarrow \mathcal{X}^n.$$

Its rate is ℓ_n/n , and its error probability is $P_e^{(n)} = \Pr\{g_n(f_n(X^n)) \neq X^n\}$. The decoder receives the entire fixed-length block, so no additional delimiting information is needed. The encoder need not be injective; collisions are allowed, but at most one source string in each collision can be reconstructed correctly.

If zero error were required on every positive-probability string, the encoder would have to be injective on $\text{supp}(P)^n$. Writing $s = |\text{supp}(P)|$, this would require $2^{\ell_n} \geq s^n$, or $\ell_n \geq \lceil n \log s \rceil$, and simple enumeration achieves this length. Allowing a vanishing error probability permits us to omit rare strings and replace the support-size rate $\log s$ by $H(P)$. For a nonuniform source, these rates are different even though recovery remains exact whenever it succeeds.

Proposition 4.4 (The best finite-block code). *If a code has M possible descriptions, its largest possible success probability is the sum of the $\min\{M, |\mathcal{X}|^n\}$ largest probabilities in P^n .*

Proof. The set of strings recovered correctly has at most M elements, since each description has only one decoded value. Its total probability is at most the sum of the M largest masses, with all masses included when there are fewer than M strings. Conversely, choose the strings with these largest masses, give them distinct descriptions, and map every remaining string to any description. Decode the chosen descriptions to their assigned strings. This achieves the stated probability. $\square$

Typicality identifies the asymptotic optimum without sorting exponentially many probabilities. Its code need not be optimal at each finite blocklength.

Theorem 4.5 (Source coding, with a strong converse). *Let $R \geq 0$ and $\ell_n = \lceil nR \rceil$. If $R > H(P)$, there are codes with these lengths and $P_e^{(n)} \rightarrow 0$. If $R < H(P)$, every sequence of such codes satisfies $P_e^{(n)} \rightarrow 1$.*

Proof. For achievability choose $\varepsilon > 0$ with $H(P) + \varepsilon < R$. By Corollary 4.3,

$$|A_{n,\varepsilon}| \leq 2^{n(H(P) + \varepsilon)} \leq 2^{\ell_n}.$$

Assign distinct ℓ_n -bit names to the typical strings, and define the decoder to invert these assignments. Map all remaining strings to any name; define the decoder arbitrarily on unused names. If the typical set is empty, use any code. Every typical string is recovered, so in either case

$$P_e^{(n)} \leq \delta_n(\varepsilon) \leq \frac{V(P)}{n\varepsilon^2} \rightarrow 0.$$

This is an existence construction; enumerating the typical set can require exponential computation. For the converse let $S_n = \{x^n : g_n(f_n(x^n)) = x^n\}$ be the correctly decoded set. The encoder is injective on S_n : two distinct strings with the same name cannot both equal its decoded value. Hence $|S_n| \leq 2^{\ell_n}$. Choose $0 < \varepsilon < H(P) - R$. Split the success probability according to whether the source string is typical:

$$\begin{aligned} 1 - P_e^{(n)} = P^n(S_n) &\leq P^n(A_{n,\varepsilon}^c) + P^n(S_n \cap A_{n,\varepsilon}) \\ &\leq \delta_n(\varepsilon) + |S_n| 2^{-n(H(P)-\varepsilon)} \\ &\leq \frac{V(P)}{n\varepsilon^2} + 2^{1-n(H(P)-R-\varepsilon)} \longrightarrow 0. \end{aligned} \quad (4.7)$$

The last step uses $\ell_n \leq nR + 1$. Both terms vanish. $\square$

Counting codewords alone would not prove the converse: a small set can carry large probability. The typical probability bound is what converts a count into a bound on success. The theorem leaves the boundary $R = H(P)$ open; its behavior requires finer information than this argument provides.

Corollary 4.6 (The smallest set carrying a fixed probability). *For fixed $0 < \eta < 1$, put $N_n(\eta) = \min\{|S| : S \subseteq \mathcal{X}^n, P^n(S) \geq 1 - \eta\}$. Then $n^{-1} \log N_n(\eta) \rightarrow H(P)$.*

Proof. Fix $\varepsilon > 0$. For all sufficiently large n , the typical set has probability at least $1 - \eta$, so $N_n(\eta) \leq 2^{n(H(P)+\varepsilon)}$. In the other direction, any qualifying set S satisfies

$$1 - \eta \leq P^n(S) \leq \delta_n(\varepsilon) + |S| 2^{-n(H(P)-\varepsilon)}.$$

Eventually $\delta_n(\varepsilon) \leq (1 - \eta)/2$, and therefore

$$|S| \geq \frac{1 - \eta}{2} 2^{n(H(P)-\varepsilon)}.$$

Taking logarithms, dividing by n , and taking lower and upper limits places both limits between $H(P) - \varepsilon$ and $H(P) + \varepsilon$. Now let ε decrease to zero. $\square$

Capturing half the probability or ninety-nine percent requires the same first-order rate, although their finite-blocklength requirements can differ.

4.4 Worked comparisons: the same source, three coding tasks

The following comparison applies the source-coding results above to a biased binary source, keeping expected length, fixed length, and error probability distinct.

Example 4.7 (A biased binary source). Take $P(1) = 1/4$. A string with k ones has empirical fraction $\alpha = k/n$. Let T_α denote all such strings. Then

$$|T_\alpha| = \binom{n}{k}, \quad P^n(x^n) = (1/4)^k (3/4)^{n-k}.$$

Its surprise per symbol is

$$-\frac{1}{n} \log P^n(x^n) = \log(4/3) + \alpha \log 3, \quad H(P) = h_2(1/4) \approx 0.8113.$$

Thus typicality is exactly $|\alpha - 1/4| \leq \varepsilon / \log 3$. At $n = 20$ and $\varepsilon = 0.1$, this means $k \in \{4, 5, 6\}$. The class with five ones has

$$|T_\alpha| = \binom{20}{5} = 15504, \quad P^{20}(T_{1/4}) = 15504 (1/4)^5 (3/4)^{15} \approx 0.2023.$$

Every string in that class has probability $2^{-20h_2(1/4)}$, but the class itself does not have probability close to one. A typical set includes a range of empirical counts. Also, the most likely string is all zeros; its surprise per symbol is only $\log(4/3) \approx 0.4150$, so it is outside this typical set. Its probability $(3/4)^n$ nevertheless tends to zero.

For the fair binary source, every string has surprise n , so every string is typical under our definition, even the all-zero string. Typicality here controls surprise, not every empirical frequency.

Return now to the Bernoulli(1/4) source. Every binary block has positive probability. Zero-error fixed-length coding therefore uses at least n bits, despite the strong bias. A zero-error prefix code can instead have expected length between $nh_2(1/4)$ and $nh_2(1/4) + 1$. A fixed-length code with vanishing error uses $\lceil nR \rceil$ bits for any fixed $R > h_2(1/4)$. These are three different performance contracts; the near equality of two asymptotic rates does not make their finite-block guarantees interchangeable.

The surprise takes the two values 2 and $\log(4/3)$, with probabilities 1/4 and 3/4. A two-valued variable taking a, b with probabilities $p, 1 - p$ has variance $p(1 - p)(a - b)^2$: subtract its mean to get deviations $(1 - p)(a - b)$ and $-p(a - b)$, then square and average. Thus

$$V(P) = \frac{3}{16}(\log 3)^2 \approx 0.4710.$$

At $R = 0.9$, choose $\varepsilon = 0.04$, so $H(P) + \varepsilon < 0.9$. Our encoder has

$$P_e^{(n)} \leq \frac{V(P)}{0.0016n} < \frac{295}{n}.$$

In particular $n \geq 29500$ suffices for error at most 0.01. This deliberately loose Chebyshev calculation proves a sufficient blocklength. It does not suggest that shorter codes cannot work. At rate $R = 0.7$, the choice $\varepsilon = 0.05$ instead gives the converse

$$1 - P_e^{(n)} \leq \frac{V(P)}{0.0025n} + 2^{1-n(h_2(1/4)-0.75)}.$$

The first term bounds exceptional source blocks; the second bounds all correctly described typical blocks together. Both decay to zero.

A small optimal code. For $n = 2$ the probabilities of 00, 01, 10, 11 are 9/16, 3/16, 3/16, 1/16. A single description bit can name only two strings. Its optimal success is therefore $(9+3)/16 = 3/4$, regardless of which of 01 or 10 wins the tie. By comparison the prefix code $00 \mapsto 0$, $01 \mapsto 10$, $10 \mapsto 110$, $11 \mapsto 111$ is error-free and has expected length

$$\frac{9}{16} \cdot 1 + \frac{3}{16} \cdot 2 + \frac{3}{16} \cdot 3 + \frac{1}{16} \cdot 3 = \frac{27}{16}.$$

The first construction pays with error; the second pays with variable length.

Exercises

1. Let X be Bernoulli(1/4) and set $X_i = X$ for every i . Compute the entropy and surprise of X^n , and construct a zero-error one-bit description of the block. Identify both uses of independence in the AEP proof that fail for this source. Explain why replacing the true block distribution by the product of its marginals would give the wrong answer.
2. For $P(1) = 1/4$, compute $V(P)$. At rate $R = 0.9$, choose a specific ε and use the achievability proof to give an explicit blocklength sufficient for error at most 0.01. Explain why this is a sufficient blocklength rather than a claim of optimality.