

INFORMATION THEORY AND APPLICATIONS IN COMPUTER SCIENCE

Lecture 3: Information, Inference, and Indistinguishability

Draft

28 September 2026 · Revised 30 September 2026

Conventions. Alphabets are finite. All logarithms are base two except $\ln$, and information is in bits. Conditional distributions are used only at positive-probability events; zero terms in entropy sums contribute zero. If $P(x) > 0 = Q(x)$, then $D(P\|Q) = \infty$. Write $h_2(t) = -t \log t - (1-t) \log(1-t)$, with $h_2(0) = h_2(1) = 0$, $X^n = (X_1, \dots, X_n)$, and $X_{<i} = (X_1, \dots, X_{i-1})$. Estimators and encoders have no computational restrictions. Any private randomness they use is independent of the hidden source before the observation is supplied.

3 From information to limits on inference

An observation can have many possible values and still tell us little about what we want to learn. We will quantify two consequences. First, accurate recovery of an unknown object requires information about it. Second, a small amount of information limits how well an observation can distinguish two experiments.

The first two lectures established the identities

$$H(X, Y) = H(Y) + H(X | Y), \quad I(X; Y) = H(X) - H(X | Y) = D(P_{XY} \| P_X P_Y),$$

and nonnegativity of conditional information and divergence. We will use these as tools. Fano translates an estimator's error into residual uncertainty. The KL chain rule accounts for evidence acquired in stages. Total variation then gives testing its operational meaning, and Pinsker converts a KL bound into a testing bound.

These notes develop Fano's inequality, coordinate-recovery lower bounds, the KL chain rule and data processing, total variation and optimal testing, and Pinsker's inequality. The proofs include the intermediate calculations and the support and endpoint cases needed to apply the results precisely.

3.1 Fano's inequality: accurate recovery requires information

Compression asks how much must be sent so a decoder can recover a source. Inference starts with an observation and asks whether recovery is possible from that observation at all. The obstacle can remain even with unlimited computation.

Let X take values in an alphabet of size $m \geq 2$, let Y be an observation, and let $\hat{X} = g(Y)$ be an estimate in the same alphabet. Write $e = \Pr[\hat{X} \neq X]$. If the guess is usually right, the remaining uncertainty has a simple description: say whether the guess was wrong and, only when it was, which of the other $m - 1$ values was correct.

Theorem 3.1 (Fano's inequality). *Under this model,*

$$\boxed{H(X | Y) \leq h_2(e) + e \log(m-1) \leq 1 + e \log m.} \quad (3.1)$$

Proof. Formalize the correction-message idea with the error indicator $E = \mathbf{1}\{g(Y) \neq X\}$. Because E is determined by (X, Y) , $H(E | X, Y) = 0$. Expanding the same joint conditional entropy in two orders,

$$\begin{aligned} H(X | Y) &= H(X, E | Y) \\ &= H(E | Y) + H(X | E, Y). \end{aligned}$$

The first term is at most $H(E) = h_2(e)$. For the second term, condition on whether an error occurred. If $E = 0$ and $Y = y$, then $X = g(y)$ and there is no uncertainty. If $E = 1$ and $Y = y$, there are at most $m-1$ possible values of X , so entropy is at most $\log(m-1)$. Averaging gives $H(X | E, Y) \leq e \log(m-1)$. Zero-probability cases contribute nothing. Finally, $h_2(e) \leq 1$ and $\log(m-1) \leq \log m$ give the coarser form. $\square$

The bound is on entropy; it does not claim an integer correction code of exactly $h_2(e) + e \log(m-1)$ bits. The flag interpretation explains the chain-rule calculation without introducing uncharged rounding costs.

The same inequality holds for a randomized estimator whose output uses no additional information about X beyond Y . In this case $X \rightarrow Y \rightarrow \hat{X}$, so $H(X | Y, \hat{X}) = H(X | Y)$. Apply the proof to the augmented observation $(Y, \hat{X})$, from which the guess is deterministic.

Reading the direction correctly. Fano upper-bounds residual entropy in terms of an estimator's error. To obtain a lower bound on error, write $H(X | Y) = H(X) - I(X; Y)$ and rearrange. If X is uniform on m values,

$$I(X; Y) \geq \log m - h_2(e) - e \log(m-1), \quad (3.2)$$

$$e \geq 1 - \frac{I(X; Y) + 1}{\log m}. \quad (3.3)$$

If the right side of the second bound is negative, it gives no information beyond $e \geq 0$. For nonuniform X , replace the initial $\log m$ in the first line by $H(X)$. A large alphabet by itself is not enough: a distribution concentrated on one likely value can be guessed accurately with no observation.

The precise form matters for a binary unknown: when $m = 2$,

$$H(X | Y) \leq h_2(e),$$

because on an error the other bit is determined. The coarser inequality would often lose the entire useful bound. More generally, do not replace e by an arbitrary upper bound ε without checking monotonicity: $h_2(t) + t \log(m-1)$ is increasing on $0 \leq t \leq 1 - 1/m$.

Example 3.2 (Many possible observations, limited information). Let X be uniform on $\{0, 1\}^{20}$. Independently of X , with probability 0.6 an observation reveals X exactly; with probability 0.4 it returns an erasure symbol $\perp$. An erasure leaves all twenty bits unknown, so

$$H(X | Y) = 0.4 \cdot 20 = 8, \quad I(X; Y) = 12.$$

Fano gives $e \geq 1 - (12 + 1)/20 = 0.35$. The optimal error is actually $0.4(1 - 2^{-20})$, approximately 0.4: after an erasure any guess succeeds with probability 2^{-20} . The observation alphabet has $2^{20} + 1$ values; counting its possible values alone gives no obstacle to recovery. Mutual information captures how often the observation fails to resolve the source.

A memory or message budget. If an observation is a message W taking at most 2^b values, then

$$I(X; W) \leq H(W) \leq b.$$

For a uniform hundred-bit record and a message with at most 2^{60} values, Fano therefore gives $e \geq 1 - 61/100 = 0.39$, even if the encoder is randomized. In this particular noiseless message problem, direct counting is much stronger: for any deterministic decoder g ,

$$\Pr[g(W) = X] = 2^{-100} \sum_w \Pr[W = w \mid X = g(w)] \leq 2^{-40}.$$

A randomized decoder cannot improve its success at a fixed w beyond the largest posterior probability. Fano is useful because the same argument works with *any* bound on mutual information, including observations from a noisy channel for which counting output values is uninformative.

Here “ b bits” means at most 2^b possible messages, not an average transmission length of b . If encoder and decoder share randomness R independent of X , and W has at most 2^b possible values for each r , the observation is (W, R) and the correct accounting is

$$I(X; W, R) = I(X; R) + I(X; W \mid R) \leq H(W \mid R) \leq b.$$

The message may depend on both X and R ; we do not assume that R is independent of the pair (X, W) .

3.2 Storing a record for later coordinate queries

Suppose we do not need to recover a whole record at once. We store a short message and later ask for just one of its bits. It might appear that one-bit answers need very little storage. The issue is that the same stored message must support every possible query, and it is created *before* the query index is known.

Proposition 3.3 (Coordinate-recovery lower bound). *Let X^n be uniform on $\{0, 1\}^n$. A possibly randomized encoder produces W with at most 2^b possible values. For each $i \in \{1, \dots, n\}$, suppose there is a decoder $g_i(W)$ with error $e_i = \Pr[g_i(W) \neq X_i]$. If the average error $\bar{e} = \frac{1}{n} \sum_i e_i$ is at most $\varepsilon \leq 1/2$, then*

$$\boxed{b \geq n(1 - h_2(\varepsilon))}. \quad (3.4)$$

Proof. The idea is to apply binary Fano to each answer and add the residual uncertainties using the chain rule. Independence of the *prior* bits gives $H(X^n) = n$; no conditional independence given W is assumed. Specifically,

$$\begin{aligned} H(X^n \mid W) &= \sum_{i=1}^n H(X_i \mid W, X_{<i}) \\ &\leq \sum_{i=1}^n H(X_i \mid W) \leq \sum_{i=1}^n h_2(e_i) \leq n h_2(\bar{e}) \leq n h_2(\varepsilon). \end{aligned}$$

The first inequality drops $X_{<i}$ from the conditioning; it does not assume that the bits stay independent after W is observed. Binary Fano gives the next inequality. For the averaging step, $h_2''(t) = -1/(t(1-t) \ln 2) < 0$ on $(0, 1)$, so concavity, with weights $1/n$, gives $\frac{1}{n} \sum_i h_2(e_i) \leq h_2(\frac{1}{n} \sum_i e_i) = h_2(\bar{e})$. Finally, $\bar{e} \leq \varepsilon \leq 1/2$ and h_2 is increasing on $[0, 1/2]$. Only the average error must be at most $1/2$; individual errors need not be. Consequently

$$b \geq H(W) \geq I(X^n; W) = n - H(X^n \mid W) \geq n(1 - h_2(\varepsilon)).$$

□

At error $\varepsilon = 0.1$, this requires approximately $0.5310n$ bits. The hypothesis permits different errors on different coordinates; only their average is bounded. It is equivalent to an error guarantee for a uniformly random independent query index. If the encoder were told the requested index first, it could simply send the requested bit. That is a different task.

Shared randomness independent of the source can be included in the observation as above, and randomized query decoders also satisfy binary Fano. This argument is the basic information bound behind one-way communication for the indexing problem and memory lower bounds for recovering stored data.

3.3 Accumulating evidence: the KL chain rule

To use Fano in an inference problem, we often need to upper-bound information in the observations. Divergence gives a convenient way to account for evidence in stages. The next identities prepare for that use.

Proposition 3.4 (Chain rule for relative entropy). *Let P_{XY}, Q_{XY} be distributions on the same finite product alphabet. Assume $P_{XY}(x, y) > 0$ implies $Q_{XY}(x, y) > 0$. Then*

$$D(P_{XY} \| Q_{XY}) = D(P_X \| Q_X) + \sum_{x: P_X(x) > 0} P_X(x) D(P_{Y|x} \| Q_{Y|x}).$$

Proof. Evidence about a pair consists of evidence about the first coordinate and then evidence about the second conditioned on the first. On the support of P_{XY} , all needed denominators are positive and

$$\log \frac{P_{XY}(x, y)}{Q_{XY}(x, y)} = \log \frac{P_X(x)}{Q_X(x)} + \log \frac{P_{Y|x}(y)}{Q_{Y|x}(y)}.$$

Taking expectation under P_{XY} gives

$$\begin{aligned} D(P_{XY} \| Q_{XY}) &= \sum_x \left(\sum_y P_{XY}(x, y) \right) \log \frac{P_X(x)}{Q_X(x)} \\ &\quad + \sum_x P_X(x) \sum_y P_{Y|x}(y) \log \frac{P_{Y|x}(y)}{Q_{Y|x}(y)}. \end{aligned}$$

The inner sum in the first line is $P_X(x)$, so that line is $D(P_X \| Q_X)$. The inner sum in the second line is the conditional divergence at x . In particular, the conditional divergences are weighted by P_X , not by Q_X . $\square$

For independent observations this yields

$$D(P^{\otimes n} \| Q^{\otimes n}) = nD(P \| Q), \quad n \geq 1. \quad (3.5)$$

For finite divergence, the log likelihood ratio is the sum of the n individual log ratios. If the one-sample divergence is infinite, an outcome possible under P but impossible under Q also makes the product divergence infinite. The same expansion works for independent observations with different distributions:

$$D\left(\bigotimes_{i=1}^n P_i \left\| \bigotimes_{i=1}^n Q_i\right.\right) = \sum_{i=1}^n D(P_i \| Q_i).$$

For finite divergences, write the log of the product likelihood ratio as the sum of its coordinate logs. Taking expectation, coordinate i has distribution P_i , so its term is exactly $D(P_i \| Q_i)$. If a coordinate divergence is infinite, a positive-probability outcome in that coordinate forces infinite divergence of the product as well. Independence is the reason for this exact addition; correlated observations instead require conditional chain-rule terms.

3.4 Processing cannot create evidence

Theorem 3.5 (Data processing for divergence). *Let $K(y | x)$ be a common channel, meaning that it is a probability distribution over y for each input x . Write $(PK)(y) = \sum_x P(x)K(y | x)$, and similarly for QK . Then*

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

Proof. Keep both the input and its processed output, and expand their divergence in two orders. Generating the output by the same rule under both hypotheses adds no divergence; discarding the input can lose some.

If $D(P \| Q) = \infty$, the inequality is automatic. Otherwise set $\tilde{P}(x, y) = P(x)K(y | x)$ and $\tilde{Q}(x, y) = Q(x)K(y | x)$. Expanding with X first gives $D(\tilde{P} \| \tilde{Q}) = D(P \| Q)$ because both conditional channels are K . Expanding with Y first gives the exact identity

$$D(P \| Q) = D(PK \| QK) + \sum_{y: (PK)(y) > 0} (PK)(y) D(\tilde{P}_{X|y} \| \tilde{Q}_{X|y}).$$

Gibbs' inequality makes the last term nonnegative. $\square$

The lost divergence is precisely the remaining average disagreement about the input after the output is known. Equality holds when these conditional input distributions agree for every output of positive PK probability, assuming finite input divergence. A deterministic statistic is a special case, as is adding fresh noise independent of the hypothesis conditional on the input. The word *common* is essential: using different processing rules under the two hypotheses can create a difference even if $P = Q$.

This parallels $I(X; Z) \leq I(X; Y)$ for $X \rightarrow Y \rightarrow Z$ from Lecture 1. It does not assert $H(Z) \leq H(Y)$ for a randomized computation; adding noise can increase output entropy while reducing information about its source.

3.5 Total variation and the best hypothesis test

KL is well suited to adding evidence, but a numerical KL value is not itself a success probability. Total variation provides the operational quantity we need. For distributions P, Q on the same finite alphabet $\mathcal{Z}$, define

$$\text{TV}(P, Q) = \frac{1}{2} \sum_z |P(z) - Q(z)|.$$

It lies between zero and one. It is zero exactly when the distributions agree, and is one exactly when their positive-probability supports are disjoint.

Proposition 3.6 (Event characterization). *For $A = \{z : P(z) \geq Q(z)\}$,*

$$\text{TV}(P, Q) = P(A) - Q(A) = \max_{B \subseteq \mathcal{Z}} |P(B) - Q(B)|.$$

Proof. Put $\Delta(z) = P(z) - Q(z)$. Since both distributions have total mass one, $\sum_z \Delta(z) = 0$, so the total positive mass of Δ equals the absolute value of its total negative mass. Each is half of $\sum_z |\Delta(z)|$. Therefore $\sum_{z \in A} \Delta(z) = \text{TV}(P, Q)$. For any event B , its sum of Δ cannot exceed all the positive terms combined, and cannot be smaller than all the negative terms combined. Thus $|P(B) - Q(B)| \leq \text{TV}(P, Q)$, with equality for $B = A$. $\square$

A binary testing experiment. An unknown hypothesis is chosen uniformly from two possibilities. Under the first, we observe $Z \sim P$; under the second, $Z \sim Q$. The test receives Z and must name the hypothesis. If a deterministic test answers P on an event B , its success is

$$\begin{aligned} \frac{1}{2}P(B) + \frac{1}{2}Q(B^c) &= \frac{1}{2}(P(B) + 1 - Q(B)) \\ &= \frac{1}{2}(1 + P(B) - Q(B)). \end{aligned}$$

The event characterization now gives the exact formulas

$$\boxed{p_{\text{success}}^* = \frac{1 + \text{TV}(P, Q)}{2}, \quad p_{\text{error}}^* = \frac{1 - \text{TV}(P, Q)}{2}.} \quad (3.6)$$

These expressions use equal prior probabilities. Advantage over random *guessing* in this experiment is $\text{TV}(P, Q)/2$; the maximum difference between a test's acceptance probabilities under P and Q is $\text{TV}(P, Q)$. These are two different conventions for the word advantage.

A randomized test answers P with a probability $a(z) \in [0, 1]$. Its success is $\frac{1}{2} + \frac{1}{2} \sum_z \Delta(z) a(z)$. To maximize this expression, take $a(z) = 1$ where $\Delta(z) > 0$ and $a(z) = 0$ where $\Delta(z) < 0$; ties do not matter. Randomization therefore cannot improve the optimum. The same argument also proves the useful general statement

$$|\mathbb{E}_P a - \mathbb{E}_Q a| \leq \text{TV}(P, Q) \quad (0 \leq a \leq 1). \quad (3.7)$$

Example 3.7 (A test on one bit). For $P = \text{Bernoulli}(3/4)$ and $Q = \text{Bernoulli}(1/2)$,

$$\text{TV}(P, Q) = \frac{1}{2}(|3/4 - 1/2| + |1/4 - 1/2|) = 1/4.$$

The best test answers P on seeing a 1 and Q on seeing a 0. Its success is $(3/4 + 1/2)/2 = 5/8$, so its error is $3/8$. A difference of $1/4$ between the distributions gives advantage $1/8$ over guessing.

3.6 Pinsker: turning divergence into distinguishability

The event description of total variation tells us which feature to retain: one bit indicating membership in the most distinguishing event. KL data processing says that even this one bit cannot have more divergence than the original observation. A scalar inequality will complete the comparison.

Theorem 3.8 (Pinsker's inequality, in bits). *For finite distributions P, Q ,*

$$\boxed{\text{TV}(P, Q)^2 \leq \frac{\ln 2}{2} D(P \| Q).} \quad (3.8)$$

Proof. If $D(P \| Q) = \infty$, the inequality is automatic. Otherwise let $A = \{z : P(z) \geq Q(z)\}$, $p = P(A)$, and $q = Q(A)$. By Proposition 3.6, $p - q = \text{TV}(P, Q)$. Apply the same deterministic map $z \mapsto \mathbf{1}\{z \in A\}$ under both distributions. KL processing gives

$$D(P \| Q) \geq d_2(p \| q), \quad d_2(p \| q) = p \log \frac{p}{q} + (1 - p) \log \frac{1 - p}{1 - q}.$$

It remains to prove $d_2(p \| q) \geq 2(p - q)^2 / \ln 2$.

Fix $0 < q < 1$ and differentiate with respect to $p \in (0, 1)$:

$$\begin{aligned} \frac{\partial}{\partial p} d_2(p \| q) &= \log \frac{p(1 - q)}{q(1 - p)}, \\ \frac{\partial^2}{\partial p^2} d_2(p \| q) &= \frac{1}{\ln 2} \left(\frac{1}{p} + \frac{1}{1 - p} \right) = \frac{1}{(\ln 2)p(1 - p)} \geq \frac{4}{\ln 2}. \end{aligned}$$

The last inequality follows from $p(1-p) \leq 1/4$. Define $g(p) = d_2(p\|q) - 2(p-q)^2/\ln 2$. Then $g''(p) \geq 0$, while $g(q) = 0$ and $g'(q) = 0$. A differentiable convex function is minimized at a point with derivative zero, so $g(p) \geq 0$. This proves the scalar inequality in the interior; continuity gives it at $p = 0, 1$. If $q = 0$ or $q = 1$, either $p = q$ and the scalar inequality is trivial, or $d_2(p\|q) = \infty$. Combining the scalar bound with KL processing proves (3.8). $\square$

The coefficient $\ln 2$ converts the base-two divergence to natural-log units. It is not a choice of normalization for total variation: our definition always includes the factor $1/2$. When using the square-root bound, we can also take the minimum with one, since TV never exceeds one.

For independent samples, product divergence gives the useful consequence

$$\text{TV}(P^{\otimes n}, Q^{\otimes n}) \leq \min \left\{ 1, \sqrt{\frac{n \ln 2}{2} D(P\|Q)} \right\}. \quad (3.9)$$

The method is now systematic: bound the one-sample KL, add it over samples, convert to TV, and use the exact testing formula (3.6).

Checks and further exercises

These reinforce the models and proofs above; they are not additional required homework problems.

1. In the twenty-bit erasure example, replace the erasure probability 0.4 by $a \in [0, 1]$. Compute residual entropy, information, and optimal recovery error. At which values does the coarse Fano bound become uninformative?
2. Suppose the encoder in the coordinate-query problem learns the requested index before creating its message. Give a one-bit scheme and identify the hypothesis of the lower bound that has changed.
3. Let $P = (3/4, 1/4)$ and $Q = (1/4, 3/4)$. Compute their divergence and total variation. Then process both distributions through a channel that outputs a fair bit independently of its input. Identify the lost divergence using the posterior formula in the processing proof.