INFORMATION THEORY AND APPLICATIONS IN COMPUTER SCIENCE

Lecture 9 | 16 November 2026

Student notes

9 Information equals amortized communication

Conventions. All alphabets are finite unless stated otherwise. Logarithms are base two, except ln; information is in bits. Terms with probability zero contribute zero to entropy sums, and conditional distributions are used only on events of positive probability. We write $h_2(p) = -p \log p - (1-p)\log(1-p)$ with its continuous endpoint values.

The question. Lecture 8 showed that randomness can collapse communication: equality of n-bit strings costs $\Theta(n)$ bits deterministically and a constant number with public coins. Now Alice and Bob hold subsets of $[n]$ and ask whether they intersect. This is the OR of n two-bit ANDs. What is the price per coordinate, and can solving many coordinates together save communication?

Task	Leading cost in bits	Outputs
One AND, error $< 1/3$	2 bits	public
All n ANDs (intersection)	$C_\wedge n$, $C_\wedge \approx 1.4923$	both players locally
Their OR (disjointness)	$C_{\mathrm{DISJ}} n$, $C_{\mathrm{DISJ}} \approx 0.4827$	public

The first row is Exercise 4. The other coefficients are quoted sharp results of Braverman, Garg, Pankratov and Weinstein. Their order of limits is important: choose a small fixed error allowance on the *whole* answer, let n grow, then let the error allowance tend to zero. For any $\eta > 0$, sufficiently small fixed error and then sufficiently large n give cost between $(C-\eta)n$ and $(C+\eta)n$ in the indicated output model. The constants come from information about a single AND; determining them requires work beyond the proofs in this lecture.

The principle. The organizing principle, due to Braverman and Rao, is:

> *When many copies of a problem are solved together, the communication needed per copy equals the information that a protocol for one copy must reveal.*

The per-copy cost in this limit is the *amortized* cost. We prove the principle for independent inputs, where the information measure and the compression construction are especially simple. Disjointness connects its lower and upper halves.

1. *Lower half: direct sum.* Information adds up over copies, so k copies need k times the information of one. For disjointness this gives the lower bound $\Omega(n)$, which we prove completely.
2. *Upper half: compression.* Public proposals replace private noise, and concentration brings the per-copy communication down to the information cost. We end by explaining the scope of this principle.

Guide to the notes. Sections 9.1–9.6 develop the lower bounds and the full one-way compression argument. Section 9.7 sketches conversation compression; its detailed proof, the correlated-input extension, and Section 9.8 are further reading.

Prerequisites. We use transcript factorization from Lecture 8, entropy chain rules, KL divergence, total variation, the law of large numbers, Jensen and Cauchy–Schwarz. Factorization is restated; Hellinger distance is developed from its definition.

9.1 Disjointness, and the model

Alice holds $x \in \{0,1\}^n$ and Bob holds $y \in \{0,1\}^n$, read as indicator vectors of two subsets of $[n]$. Define

$$\mathrm{DISJ}_n(x,y) = 1 \iff x_i y_i = 0 \text{ for every } i, \qquad \mathrm{INT}_n(x,y) = 1 - \mathrm{DISJ}_n(x,y) = \bigvee_{i=1}^{n} (x_i \wedge y_i).$$

Flipping a publicly known output bit is free, so the two functions have the same complexity. We work with INT_n, whose one-coordinate case is $\mathrm{INT}_1 = \mathrm{AND}$. Asking for all n coordinate answers $(x_1 \wedge y_1, \ldots, x_n \wedge y_n)$ instead of their OR is the *set intersection* problem: n copies of AND.

The model for the proved core. A protocol π has a public random string R, which both parties see, and private random strings R_A and R_B, seen only by Alice and only by Bob. The three are independent of each other and of the inputs. The transcript T is the sequence of bits sent, and $W = (R, T)$ is the *full public record*: everything that an observer of the conversation sees. The *cost* of π is the largest number of bits sent, over all inputs and all values of the coins. The output is a function of W; if the task has several answers, all of them are functions of W. Random strings take finitely many values. Quoted results with *local outputs* will be marked explicitly: there each player may also use its own input and private coins to recover the answer. Publishing a single Boolean local answer costs one bit; publishing a vector can cost linearly many bits.

Why keep both kinds of coins? For communication, making private coins public preserves cost and error. For information, it can reveal more: $x \oplus \sigma$ hides x from an observer when σ is a private fair bit, but reveals x when σ is public. The private filler inputs in the direct-sum proof must therefore stay private. Public proposals in the compression proof have a different role: they replace noise that need not be sent.

9.2 Information cost and the amortized theorem

Two results of Lecture 8 are used throughout. Write $P_{xy}(r,t) = \Pr[W = (r,t) \mid x, y]$ for the distribution of the public record on the fixed input pair (x, y), and $p(r)$ for the distribution of R.

Transcript factorization (Theorem 8.10). There are nonnegative functions $a_x(r,t)$ and $b_y(r,t)$ such that

$$P_{xy}(r,t) = p(r)\, a_x(r,t)\, b_y(r,t). \tag{9.1}$$

Here $a_x(r,t)$ is the probability that Alice's private string is consistent with the path t when her input is x; it does not involve y. Symmetrically for $b_y(r,t)$.

Information budget (Theorem 8.12). Let (X,Y) be random inputs and let D be any random variable jointly distributed with them, with the coins of the protocol independent of the triple (X,Y,D). If the protocol has cost c, then for each value of R there are at most 2^c transcripts, hence

$$I(X,Y;T \mid D,R) \le H(T \mid D,R) \le c. \tag{9.2}$$

Definition 9.1 (Information cost)**.** Let λ be a distribution on triples (X,Y,D): a pair of inputs together with an auxiliary variable D. Run π on (X,Y) with coins independent of (X,Y,D). The protocol is not given D. The *information cost of π under λ* is

$$\mathrm{IC}_\lambda(\pi) = I(X,Y;T \mid D,R) = I(X,Y;W \mid D).$$

The two expressions agree by the chain rule, because R is independent of (X,Y,D). Usually there is no auxiliary variable: λ is a distribution μ on input pairs, and $\mathrm{IC}_\mu(\pi) = I(X,Y;T \mid R)$

is what an observer of the public record learns about the inputs. The variable D is needed only for disjointness (Section 9.4); with it, the cost is what the observer learns if she is also told D. The public string is in the conditioning because the observer sees it. The private strings appear nowhere, because she does not.

Two ways to ask for correctness. A protocol computes f *with error* ε *on every input* if, for each fixed input pair, its output differs from f with probability at most ε over the coins. This is the requirement of Lecture 8, and $\mathrm{CC}_\varepsilon(f)$ denotes the least cost of such a protocol. A protocol computes f *with error* ε *on average under* μ if its output differs from $f(X,Y)$ with probability at most ε when $(X,Y)\sim\mu$, the probability now being over inputs and coins. The default task error is strictly below $1/3$. Parameterized bounds use error at most ε, including $\varepsilon=0$ for zero error. Disjointness uses correctness on every input; amortization uses average correctness.

Definition 9.2 (Information complexity)**.** For a function f and $\varepsilon\geq 0$,

$$\mathrm{IC}_\lambda(f,\varepsilon)=\inf\{\mathrm{IC}_\lambda(\pi):\ \pi \text{ computes } f \text{ with error } \varepsilon \text{ on every input}\},$$
$$\mathrm{IC}^{\mathrm{av}}_\mu(f,\varepsilon)=\inf\{\mathrm{IC}_\mu(\pi):\ \pi \text{ computes } f \text{ with error } \varepsilon \text{ on average under } \mu\}.$$

In the first line the distribution is used only to measure information; the error condition also covers inputs outside its support. Clearly $\mathrm{IC}^{\mathrm{av}}_\mu(f,\varepsilon)\leq \mathrm{IC}_\mu(f,\varepsilon)$.

Many copies. For $f:\mathcal{X}\times\mathcal{Y}\to\mathcal{Z}$ on finite sets, write f^k for the task of computing f on k input pairs at once: $f^k(x_1,\dots,x_k;y_1,\dots,y_k)=(f(x_1,y_1),\dots,f(x_k,y_k))$. Alice holds all the x_i and Bob all the y_i. The pairs are drawn independently from μ, so the joint distribution is μ^k. A protocol for f^k has *error* ε *in each copy* if, for each i, its i-th answer differs from $f(X_i,Y_i)$ with probability at most ε. Let

$$C_k(\varepsilon)=\text{the least cost of a protocol for } f^k \text{ with error } \varepsilon \text{ in each copy},$$

and define $\mathrm{IC}^{\mathrm{av}}_{\mu^k}(f^k,\varepsilon)$ as the infimum of IC_{μ^k} over the same protocols.

Proposition 9.3 (Communication is at least information)**.** *A protocol of cost c has information cost at most c under every λ. Consequently*

$$\mathrm{CC}_\varepsilon(f)\geq \mathrm{IC}_\lambda(f,\varepsilon)\qquad \textit{and}\qquad C_k(\varepsilon)\geq \mathrm{IC}^{\mathrm{av}}_{\mu^k}(f^k,\varepsilon).$$

Proof. The first sentence is (9.2). Apply it to a protocol of least cost. □

For a particular protocol the inequality can be far from tight, and the upper half of the theorem is about closing this gap.

Example 9.4 (A message that is half noise)**.** Alice holds a fair bit X and sends $M=X\oplus N$, where N is a private coin with $\Pr[N=1]=0.11$. The cost is one bit. The information is

$$I(X;M)=1-h_2(0.11)\approx 0.50 \text{ bits},$$

the capacity of a binary symmetric channel (Lecture 5). Half of the bit that travels is Alice's private noise. An extreme case: if Alice sends a thousand private coin flips, the cost is 1000 and the information is zero.

Theorem 9.5 (Information equals amortized communication)**.** *Let $f : \mathcal{X} \times \mathcal{Y} \to \mathcal{Z}$ be a function on finite sets, let $\mu = \mu_X \times \mu_Y$ be a product distribution on its inputs, and let $\varepsilon > 0$. Then*

$$\lim_{k\to\infty} \frac{C_k(\varepsilon)}{k} = \mathrm{IC}^{\mathrm{av}}_{\mu}(f, \varepsilon).$$

For independent inputs, external and internal information coincide. The general Braverman–Rao theorem uses internal information and local outputs; see Section 9.7. The lower bound is information additivity. The upper bound compresses many independent conversations: Section 9.6 proves the one-message case, and the complete argument appears in the further-reading part of Section 9.7.

Disjointness asks for one OR rather than the whole vector f^k. Sections 9.4–9.5 adapt direct sum to prove its $\Omega(n)$ lower bound. After compression we explain the further results behind its sharp constant.

9.3 Lower half: information adds up

A protocol for k copies is one conversation about all of them, and a single message may concern many copies at once. The lower half says that the conversation must nevertheless reveal as much as k separate conversations would.

Setting. Let λ be a distribution on triples (X, Y, D) such that X and Y are independent given D. Take k independent triples $(X_i, Y_i, D_i) \sim \lambda$, write $Z_i = (X_i, Y_i)$, and let X, Y, D denote the vectors. *On a first reading, ignore D*: then λ is a product distribution $\mu_X \times \mu_Y$, which is the case of Theorem 9.5. The auxiliary variable costs nothing in the proofs, and disjointness will need it.

Lemma 9.6 (Information adds up)**.** *For every protocol on k pairs of inputs, with public string R and transcript T,*

$$\mathrm{IC}_{\lambda^k}(\pi) = I(Z_1, \dots, Z_k; T \mid D, R) \;\geq\; \sum_{i=1}^{k} I(Z_i; T \mid D, R). \tag{9.3}$$

Proof. The Z_i are independent given D, and R is independent of all of them, so

$$H(Z_1, \dots, Z_k \mid D, R) = \sum_i H(Z_i \mid D, R).$$

After the transcript is revealed the coordinates may be dependent, but entropy is always subadditive (chain rule, then drop conditioning):

$$H(Z_1, \dots, Z_k \mid T, D, R) \leq \sum_i H(Z_i \mid T, D, R).$$

Subtract the second line from the first. □

Only independence before the conversation was used. The inequality can be strict: the parity of two independent fair bits carries one bit about the pair and none about either bit alone. Independence is essential: if $Z_1 = Z_2 = T$ is a fair bit, the left side is 1 and the sum is 2.

Each term is the cost of a protocol for one copy. Fix a protocol Π on k pairs of inputs, and a coordinate i. From Π we build a protocol π_i on *one* pair of inputs: a for Alice and b for Bob. The parties invent the other $k - 1$ coordinates.

1. *Public coins.* The public string R of Π, and values D_j for every $j \neq i$, drawn independently from the distribution of D.

2. *Private coins.* For each $j \neq i$, Alice privately draws X_j from its distribution given D_j, and Bob privately draws Y_j from its distribution given D_j.

3. *Run* Π. Alice sets $X_i = a$ and Bob sets $Y_i = b$. They run Π on (X, Y) with public string R and fresh private strings.

This is a legal protocol: each party acts on its own input, its own private coins, the public coins, and the messages received. Its public coins are $R' = (R, D_{-i})$, where $D_{-i} = (D_j)_{j \neq i}$, and its transcript is the transcript T of Π. Because X_j and Y_j are independent given D_j, drawing them separately gives the pair its correct joint distribution. What π_i outputs is decided later, according to the task.

Lemma 9.7 (Embedding)**.** *Run π_i on inputs (a, b) drawn together with D_i from λ, independently of all coins. Then (X, Y, D, R, T) has the same joint distribution as when Π is run on $(X, Y, D) \sim \lambda^k$. In particular*

$$\mathrm{IC}_\lambda(\pi_i) = I(Z_i; T \mid D, R),$$

where the right side refers to Π run on λ^k.

Proof. In the experiment of the lemma, $(X_i, Y_i, D_i) = (a, b, D_i)$ has distribution λ, and so does each (X_j, Y_j, D_j) with $j \neq i$, by construction. The k triples are independent of each other, of R, and of the private strings of Π. Hence (X, Y, D, R) has the same joint distribution as in the experiment that defines $\mathrm{IC}_{\lambda^k}(\Pi)$, and T is generated from (X, Y, R) by the same protocol. For the information cost, recall that the public string of π_i is $R' = (R, D_{-i})$:

$$\mathrm{IC}_\lambda(\pi_i) = I(a, b; T \mid D_i, R') = I(X_i, Y_i; T \mid D_i, D_{-i}, R) = I(Z_i; T \mid D, R). \qquad \square$$

Two features of the construction carry the proof. They are the reason the model has both kinds of coins.

The invented inputs are private. If they were read from the public string, the public coins of π_i would include Z_{-i}, and the lemma would read $I(Z_i; T \mid D, R, Z_{-i})$. These terms are not controlled by Lemma 9.6. For illustration, let Alice hold $k \geq 2$ independent fair bits and let Π send a single bit, their parity. Once the other coordinates are known this bit reveals X_i, so $I(Z_i; T \mid R, Z_{-i}) = 1$ for every i, and the sum k exceeds the cost 1. With private inputs the other coordinates mask the parity, and $I(Z_i; T \mid R) = 0$.

The shared choices are public. Each party must know the values D_j in order to draw its inputs. They are independent of (a, b), as public coins must be, and they end up in the conditioning, where (9.3) already has them.

Theorem 9.8 (Information is additive)**.** *For a product distribution μ and every $\varepsilon \geq 0$,*

$$\mathrm{IC}^{\mathrm{av}}_{\mu^k}(f^k, \varepsilon) = k \cdot \mathrm{IC}^{\mathrm{av}}_\mu(f, \varepsilon). \qquad \textit{Consequently} \qquad C_k(\varepsilon) \;\geq\; k \cdot \mathrm{IC}^{\mathrm{av}}_\mu(f, \varepsilon).$$

Proof. At least. Let Π be a protocol for f^k with error ε in each copy. Build $\pi_1, \ldots, \pi_k$ as above, with no auxiliary variable, and let π_i output the i-th answer of Π. By Lemma 9.7, on inputs $(a, b) \sim \mu$ the protocol π_i errs exactly as often as the i-th answer of Π does under μ^k, so π_i computes f with error ε on average. By Lemmas 9.6 and 9.7,

$$\mathrm{IC}_{\mu^k}(\Pi) \;\geq\; \sum_{i=1}^{k} I(Z_i; T \mid R) = \sum_{i=1}^{k} \mathrm{IC}_\mu(\pi_i) \;\geq\; k \cdot \mathrm{IC}^{\mathrm{av}}_\mu(f, \varepsilon).$$

At most. Let π compute f with error ε on average. Run k independent copies of π, one on each pair, with independent coins. The k tuples (X_i, Y_i, R_i, T_i) are independent, so the information costs add up exactly, to $k\ \mathrm{IC}_\mu(\pi)$, and each copy keeps its error. Take the infimum over π.

The consequence is Proposition 9.3. $\square$

9.4 Disjointness: one answer, and a hidden selector

Disjointness has n coordinates but one answer, so π_i cannot read off "the i-th answer". The embedding still works if the invented coordinates cannot decide the answer. Fill them with *zeros of* AND: if $X_jY_j = 0$ for every $j \neq i$, then

$$\mathrm{INT}_n(X, Y) = a \wedge b,$$

and the output of a protocol for INT_n is an answer to the one-coordinate question. Two things are needed, and they conflict. The invented coordinates must never intersect. And Alice and Bob must invent them without talking, which requires X_j and Y_j to be independent: but if they are independent and each equals 1 with positive probability, then $(1, 1)$ has positive probability. Fixing $Y_j = 0$ would resolve the conflict. But the same distribution measures information in the coordinate that is not invented, and if Bob's bit is always 0, a protocol in which Bob speaks first reveals nothing. The way out is a mixture of two product distributions, with the auxiliary variable saying which.

Let D be uniform on $\{A, B\}$ and let U be an independent fair bit. Set

$$D = A: \ (X, Y) = (U, 0), \qquad D = B: \ (X, Y) = (0, U).$$

Let ν denote the distribution of the triple (X, Y, D). The *selector* D names the player whose bit is random.

(x, y)	$(0, 0)$	$(1, 0)$	$(0, 1)$	$(1, 1)$
given $D = A$	$1/2$	$1/2$	0	0
given $D = B$	$1/2$	0	$1/2$	0
overall	$1/2$	$1/4$	$1/4$	0

We use three properties.

(i) *Only zero-inputs.* $X \wedge Y = 0$ always.

(ii) *Product given the selector.* X and Y are independent given D. They are not independent without it.

(iii) *Something to hide.* Given $D = A$, Alice's bit is a fair coin; given $D = B$, Bob's is.

Example 9.9 (The cost of the obvious protocol)**.** Alice sends her bit a. If $a = 0$ the protocol stops with output 0. If $a = 1$ Bob sends b, which is the output. The transcripts are 0, 10, 11, and the protocol never errs. It is deterministic, so $\mathrm{IC}_\nu(\pi) = H(T \mid D)$. Given $D = A$ the transcript is 0 or 10 according to the fair bit X, so it has one bit of entropy. Given $D = B$ we have $X = 0$ and the transcript is always 0. Therefore

$$\mathrm{IC}_\nu(\pi) = \tfrac{1}{2} \cdot 1 + \tfrac{1}{2} \cdot 0 = \tfrac{1}{2}, \qquad \text{and so} \qquad \mathrm{IC}_\nu(\mathrm{AND}, 0) \leq \tfrac{1}{2}.$$

If Bob speaks first the cost is again $1/2$, by symmetry. Each half of ν is free for one of the two orders of speaking, but a protocol does not know which half it is facing.

With $\lambda = \nu$ the embedding of Section 9.3 looks as follows, for $n = 5$ and $i = 3$ and one outcome of the public selectors.

coordinate j	1	2	3	4	5
public selector D_j	A	B	none	B	A
Alice's bit X_j	private coin	0	a	0	private coin
Bob's bit Y_j	0	private coin	b	private coin	0

Corollary 9.10 (Direct sum for disjointness)**.** *For every $n \geq 1$ and $0 \leq \varepsilon < 1/2$,*

$$\mathrm{CC}_\varepsilon(\mathrm{DISJ}_n) \;\geq\; \mathrm{IC}_{\nu^n}(\mathrm{INT}_n, \varepsilon) \;\geq\; n \cdot \mathrm{IC}_\nu(\mathrm{AND}, \varepsilon).$$

Proof. The first inequality is Proposition 9.3, with $\mathrm{CC}_\varepsilon(\mathrm{DISJ}_n) = \mathrm{CC}_\varepsilon(\mathrm{INT}_n)$. For the second, let Π compute INT_n with error ε on every input. Build π_i as in Section 9.3 with $\lambda = \nu$, and let it output what Π outputs. Every invented coordinate has $X_j Y_j = 0$, whatever the coins, so $\mathrm{INT}_n(X, Y) = a \wedge b$. For each fixed pair of vectors, Π errs with probability at most ε; averaging over the invented coordinates keeps the error at most ε for each of the four inputs (a, b). So π_i computes AND with error ε on every input, and by Lemmas 9.6 and 9.7,

$$\mathrm{IC}_{\nu^n}(\Pi) \;\geq\; \sum_{i=1}^{n} I(Z_i; T \mid D, R) = \sum_{i=1}^{n} \mathrm{IC}_\nu(\pi_i) \;\geq\; n \cdot \mathrm{IC}_\nu(\mathrm{AND}, \varepsilon). \qquad \square$$

Two points are specific to disjointness.

The target selector is not part of the protocol. D_i is correlated with (a, b), since it says which of the two bits is random, so it cannot be a coin. It appears only as the auxiliary variable in $\mathrm{IC}_\nu(\pi_i)$. Moreover π_i must also work on $(1, 1)$, an input that no selector value describes.

Correctness on every input. On $(a, b) = (1, 1)$ the vectors given to Π intersect, so they lie outside the support of ν^n. A protocol that is correct only on average under ν^n, such as the one that always answers "disjoint", would give a useless π_i. This is why disjointness uses the first notion of correctness. The distribution measures information; the hardness comes from the inputs outside its support.

9.5 The cost of one AND

It remains to show that $\mathrm{IC}_\nu(\mathrm{AND}, \varepsilon)$ is at least a positive constant. Given $D = A$ the input is $(X, 0)$ with X a fair bit, so the information cost is about how different the public record looks on the inputs 00 and 10. Given $D = B$ it is about 00 versus 01. We need a distance between distributions of the record with three properties: small information forces small distance, correctness forces a large distance between the records on 00 and 11, and the distance works well with the product form (9.1). Total variation has the first two properties. Hellinger distance has all three.

Definition 9.11 (Overlap and Hellinger distance)**.** For distributions P, Q on a finite set, the *overlap* is

$$\rho(P, Q) = \sum_w \sqrt{P(w)\, Q(w)},$$

and the *Hellinger distance* is $h(P, Q) = \sqrt{1 - \rho(P, Q)}$.

Why square roots? The overlap is 1 for identical distributions and 0 for disjoint supports. Total variation uses a related overlap:

$$1 - \mathrm{TV}(P, Q) = \sum_w \min\{P(w), Q(w)\}, \qquad 1 - h^2(P, Q) = \sum_w \sqrt{P(w) Q(w)}. \tag{9.4}$$

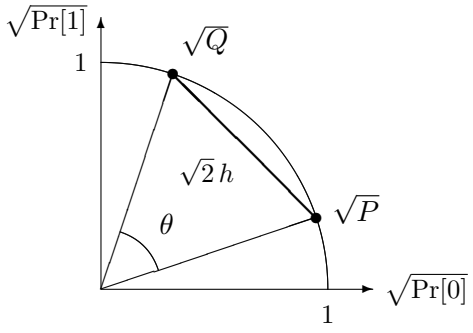
Since $\min\{p, q\} \leq \sqrt{pq}$,

$$h^2(P, Q) \leq \mathrm{TV}(P, Q). \tag{9.5}$$

The square-root vectors have unit length. If their angle is θ, then $\rho(P, Q) = \cos\theta$, and

$$h^2(P, Q) = \frac{1}{2} \sum_w (\sqrt{P(w)} - \sqrt{Q(w)})^2. \tag{9.6}$$

Thus h is Euclidean distance scaled by $1/\sqrt{2}$; in particular it satisfies the triangle inequality.



Two outcomes: $P = (0.9,\, 0.1)$, $Q = (0.1,\, 0.9)$

$\cos\theta = \rho(P,Q) = 2\sqrt{0.09} = 0.6$

$h^2(P,Q) = 1 - \cos\theta = 0.4$

the chord has length $\sqrt{2}\,h(P,Q) \approx 0.89$

$P = Q$: one point. Disjoint supports: $\theta = 90°$.

Under products. Overlaps multiply: $\rho(P \times P', Q \times Q') = \rho(P,Q)\rho(P',Q')$. Consequently k samples have overlap ρ^k, and distinguishing P from Q takes order $1/h^2(P,Q)$ samples (Exercise 3). The same product algebra will give the protocol's cut-and-paste identity.

Lemma 9.12 (Hellinger against TV and KL)**.** *For all distributions P, Q on a finite set,*

$$h^2(P,Q) \;\le\; \mathrm{TV}(P,Q) \le \sqrt{2}\,h(P,Q), \tag{9.7}$$

$$\mathrm{D}(P\|Q) \ge \frac{2}{\ln 2}\,h^2(P,Q). \tag{9.8}$$

Proof. The first inequality is (9.5). Factoring the difference and applying Cauchy–Schwarz gives

$$\begin{aligned}\mathrm{TV}(P,Q) &\le \frac{1}{2}\sqrt{\sum_w (\sqrt{P(w)} - \sqrt{Q(w)})^2}\sqrt{\sum_w (\sqrt{P(w)} + \sqrt{Q(w)})^2} \\ &= \sqrt{(1-\rho)(1+\rho)} \le \sqrt{2}\,h(P,Q).\end{aligned}$$

For KL, if $P(w) > 0 = Q(w)$ the bound is automatic. Otherwise Jensen gives

$$\begin{aligned}\mathrm{D}(P\|Q) &= -2\mathbb{E}_{w\sim P} \log \sqrt{Q(w)/P(w)} \\ &\ge -2\log \rho(P,Q) \ge \frac{2}{\ln 2}(1 - \rho(P,Q)) = \frac{2}{\ln 2}h^2(P,Q),\end{aligned}$$

using $-\ln u \ge 1 - u$. □

Lemma 9.13 (One fair bit)**.** *Let V be a fair bit. Let W have distribution P when $V = 0$ and Q when $V = 1$. Then*

$$I(V;W) \;\ge\; \frac{h^2(P,Q)}{2\ln 2}. \tag{9.9}$$

Proof. The marginal distribution of W is $M = (P+Q)/2$, and $I(V;W) = \frac{1}{2}\,\mathrm{D}(P\|M) + \frac{1}{2}\,\mathrm{D}(Q\|M)$. By (9.8),

$$I(V;W) \;\ge\; \frac{h^2(P,M) + h^2(Q,M)}{\ln 2}.$$

By the triangle inequality and $(u+v)^2 \le 2u^2 + 2v^2$,

$$h^2(P,Q) \le \big(h(P,M) + h(M,Q)\big)^2 \le 2h^2(P,M) + 2h^2(Q,M).$$

Combine the two displays. □

Thus little information about the fair selector forces the two conditional record distributions to be close.

Lemma 9.14 (Cut-and-paste)**.** *For any protocol, any inputs x, x' of Alice and any inputs y, y' of Bob,*

$$h(P_{xy}, P_{x'y'}) = h(P_{xy'}, P_{x'y}). \tag{9.10}$$

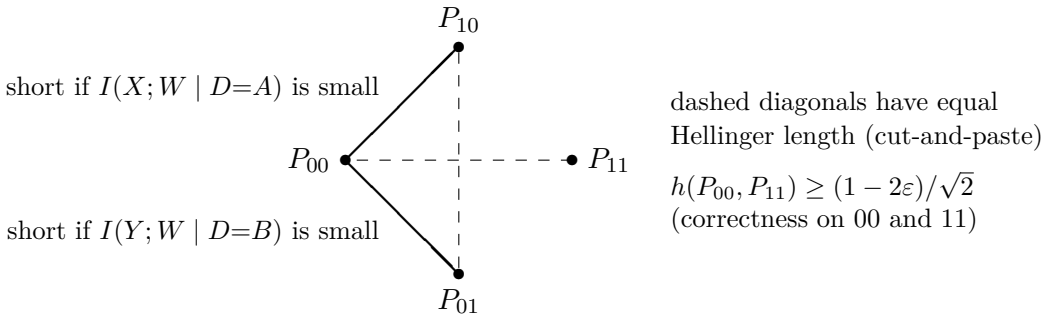
Proof. By (9.1), for every (r, t),

$$\sqrt{P_{xy}(r,t)\,P_{x'y'}(r,t)} = p(r)\sqrt{a_x(r,t)\,a_{x'}(r,t)\,b_y(r,t)\,b_{y'}(r,t)} = \sqrt{P_{xy'}(r,t)\,P_{x'y}(r,t)}.$$

Sum over (r, t): the two pairs of distributions have the same overlap ρ, hence the same Hellinger distance. □

This is the only place where the proof uses the fact that the conversation is held by two separate parties. For a deterministic protocol every P_{xy} is a point mass and h takes only the values 0 and 1. The lemma then says: if (x, y') and (x', y) produce the same transcript, so do (x, y) and (x', y'). That is the rectangle property (Lemma 8.1); cut-and-paste is its version for randomized protocols. The record must include the public string for this to hold (Example 8.11). For one-bit inputs,

$$h(P_{00}, P_{11}) = h(P_{01}, P_{10}). \tag{9.11}$$



The figure is the whole argument. If the protocol reveals little, the two solid sides are short. Then the diagonal from P_{10} to P_{01} is short by the triangle inequality, and so is the other diagonal, by (9.11). But a correct protocol must tell 00 from 11.

Theorem 9.15 (Information cost of AND)**.** *For every $0 \leq \varepsilon < 1/2$,*

$$\mathrm{IC}_\nu(\mathrm{AND}, \varepsilon) \;\geq\; \kappa_\varepsilon, \qquad \textit{where} \qquad \kappa_\varepsilon = \frac{(1-2\varepsilon)^2}{16 \ln 2}.$$

Proof. Let π compute AND with error ε on all four inputs, and let P_{ab} be the distribution of its record $W = (R, T)$ on input (a, b).

Correctness separates P_{00} from P_{11}. Let E be the set of records on which the output is 1. Then $P_{00}(E) \leq \varepsilon$ and $P_{11}(E) \geq 1 - \varepsilon$, so $\mathrm{TV}(P_{00}, P_{11}) \geq 1 - 2\varepsilon$ by the event characterization of total variation. With (9.7),

$$h^2(P_{00}, P_{11}) \;\geq\; \tfrac{1}{2}\,\mathrm{TV}(P_{00}, P_{11})^2 \;\geq\; \tfrac{1}{2}(1-2\varepsilon)^2. \tag{9.12}$$

Information controls two sides. The protocol does not see D, and its coins are independent of D. So given $D = A$ and $X = x$ the record has distribution P_{x0}, where X is a fair bit; given $D = B$ and $Y = y$ it has distribution P_{0y}, where Y is a fair bit. By Lemma 9.13,

$$\begin{aligned} \mathrm{IC}_\nu(\pi) &= \tfrac{1}{2} I(X; W \mid D = A) + \tfrac{1}{2} I(Y; W \mid D = B) \\ &\geq \frac{h^2(P_{00}, P_{10}) + h^2(P_{00}, P_{01})}{4 \ln 2}. \end{aligned} \tag{9.13}$$

Cut-and-paste closes the triangle. By (9.11) and the triangle inequality,

$$h(P_{00}, P_{11}) = h(P_{01}, P_{10}) \le h(P_{01}, P_{00}) + h(P_{00}, P_{10}).$$

Squaring and using $(u+v)^2 \le 2u^2 + 2v^2$, the numerator in (9.13) is at least $\frac{1}{2}h^2(P_{00}, P_{11})$. With (9.12),

$$\mathrm{IC}_\nu(\pi) \;\ge\; \frac{h^2(P_{00}, P_{11})}{8 \ln 2} \;\ge\; \frac{(1-2\varepsilon)^2}{16 \ln 2}. \qquad \square$$

Theorem 9.16 (Disjointness needs linear communication)**.** *For every $n \ge 1$ and $0 \le \varepsilon < 1/2$,*

$$\mathrm{CC}_\varepsilon(\mathrm{DISJ}_n) \;\ge\; n \cdot \mathrm{IC}_\nu(\mathrm{AND}, \varepsilon) \;\ge\; \frac{(1-2\varepsilon)^2}{16 \ln 2}\, n. \tag{9.14}$$

For $\varepsilon = 1/3$ the right side is $n/(144 \ln 2) > n/100$.

Proof. Corollary 9.10 and Theorem 9.15. $\square$

Where the bound came from. Not from the entropy of the answer, which is zero under ν. Correctness was used only on the inputs 00 and 11, and 11 is not even in the support of ν. The supported inputs 10 and 01 entered through the information terms. The four distributions were tied together by cut-and-paste. In the zero-error limit the coefficient is $\kappa_0 \approx 0.09$ bits per coordinate, while Example 9.9 shows that this ν can give at most $1/2$. The sharp small-error coefficient 0.4827 requires the further results discussed after compression.

9.6 Upper half: compressing one message

Start with a single-instance protocol whose information cost is near $\mathrm{IC}^{\mathrm{av}}_\mu(f, \varepsilon)$. Running it k times independently may communicate far more than that. Instead, we simulate its messages jointly, preserving their joint distribution with the inputs while reducing the bits sent.

The task. Alice knows a distribution p on a finite set $\mathcal{M}$: the distribution of the message she would send. Both parties know a second distribution q on $\mathcal{M}$; think of it as the observer's prediction of the message. Bob should end up with a sample from p. Alice could draw the sample privately and send it. The idea is to draw it *together* instead. The public string offers candidate messages drawn from q, and Alice only says which candidate to take. Her private noise is replaced by shared noise, and shared noise does not have to be sent.

Lemma 9.17 (Choosing among public proposals)**.** *Let p and q be distributions on a finite set $\mathcal{M}$, let $s \ge 0$ be real and $0 < \delta \le 1/e$. Put*

$$G = \{m : \; p(m) \le 2^s\, q(m)\}, \qquad N = \lceil 2^{s+1} \ln(1/\delta) \rceil.$$

There is a one-way public-coin protocol in which Alice, who knows p, sends $\lceil \log(N+1) \rceil \le s + \log\ln(1/\delta) + 3$ bits to Bob, who knows only q, s and δ, and Bob outputs a message whose distribution is within total variation $\delta + 2\,p(G^{\mathrm{c}})$ of p.

Proof. The public string holds independent *proposals* $(M_1, U_1), \ldots, (M_N, U_N)$, with $M_n \sim q$ and U_n uniform on $[0, 1]$. (A fine finite grid changes the output distribution arbitrarily little; the estimates below have positive slack in the claimed bound, so the grid can be chosen to preserve it.) Alice *accepts* proposal n if

$$M_n \in G \qquad \text{and} \qquad U_n \le \frac{p(M_n)}{2^s\, q(M_n)},$$

a ratio that is at most 1 on G. She sends the index J of the first accepted proposal, or $J = 0$ if there is none. Bob outputs M_J, or a fixed message if $J = 0$.

For each $m \in G$,

$$\Pr[M_n = m \text{ and proposal } n \text{ is accepted}] = q(m) \cdot \frac{p(m)}{2^s q(m)} = 2^{-s} p(m).$$

So each proposal is accepted with probability $\alpha = 2^{-s} p(G)$, and an accepted proposal has the distribution of p conditioned on G, which is within total variation $p(G^{\mathrm{c}})$ of p. The proposals are independent, so $\Pr[J = 0] = (1 - \alpha)^N \le e^{-\alpha N}$.

If $p(G) \ge 1/2$ then $\alpha N \ge \ln(1/\delta)$, so $\Pr[J = 0] \le \delta$ and the output is within $\delta + p(G^{\mathrm{c}})$ of p. If $p(G) < 1/2$ the claimed bound exceeds 1. For the number of bits, $N + 1 \le 2^{s+1} \ln(1/\delta) + 2 \le 2^{s+2} \ln(1/\delta)$. □

Reading the lemma. The cost is essentially s, the logarithm of the largest likelihood ratio $p(m)/q(m)$ that Alice is prepared to produce. Messages with a larger ratio are given up, at a price $2p(G^{\mathrm{c}})$ in accuracy. The more a message would surprise someone who expects q, the more it costs. Nothing is assumed about how p and q are related, and Bob never learns p.

For the noisy bit of Example 9.4, p is $(0.89, 0.11)$ or $(0.11, 0.89)$ and $q = (\frac{1}{2}, \frac{1}{2})$, so the largest ratio is $1.78 = 2^{0.83}$. For one copy the lemma is useless: it sends more than the single bit we started with. Its strength shows with many copies. The likelihood ratio of k independent messages is a product of k ratios, its logarithm is a sum, and the law of large numbers says that the sum is typically near its mean. The mean is the information.

Theorem 9.18 (One message, many copies)**.** *Let $X \sim \mu_X$ take values in a finite set, and let the message M have distribution p_x on a finite set $\mathcal{M}$ when $X = x$. Put $I = I(X; M)$, and let $(X_1, M_1), \dots, (X_k, M_k)$ be independent copies of (X, M). For all $\gamma > 0$ and $0 < \delta \le 1/2$ there is k_0 such that for every $k \ge k_0$ there is a one-way public-coin protocol in which Alice, given $x_1, \dots, x_k$, sends at most $k(I + \gamma)$ bits and Bob outputs $\widehat{M}_1, \dots, \widehat{M}_k$ with*

$$\mathrm{TV}\big((X^k, \widehat{M}^k),\, (X^k, M^k)\big) \le \delta.$$

Proof. Let q be the distribution of M, the observer's prediction, and define the *information density*

$$\ell(x, m) = \log \frac{p_x(m)}{q(m)}$$

for pairs of positive probability. By the identity $I(X; M) = \mathbb{E}_X \, \mathrm{D}(p_X \| q)$, its mean is $\mathbb{E}\, \ell(X, M) = I$. Its variance V is finite, because ℓ takes finitely many values.

Given $x^k = (x_1, \dots, x_k)$, the k messages have the product distribution $p_{x^k} = p_{x_1} \times \cdots \times p_{x_k}$, the prediction is q^k, and

$$\log \frac{p_{x^k}(m^k)}{q^k(m^k)} = \sum_{i=1}^{k} \ell(x_i, m_i).$$

This is a sum of k independent terms with mean I. Exactly as in the proof of Theorem 4.2, Chebyshev's inequality gives

$$\eta_k := \Pr\Big[\sum_{i=1}^{k} \ell(X_i, M_i) > k\big(I + \tfrac{\gamma}{2}\big)\Big] \le \frac{4V}{k\gamma^2}. \tag{9.15}$$

Alice applies Lemma 9.17 to $p = p_{x^k}$ and q^k, with $s = k(I + \gamma/2)$ and with $\delta/2$ in place of δ. The set G^{c} of the lemma consists of the m^k whose sum exceeds s, so on average over x^k its probability is η_k, and

$$\mathrm{TV}\big((X^k, \widehat{M}^k),\, (X^k, M^k)\big) \le \frac{\delta}{2} + 2\eta_k.$$

This is at most δ once $k \ge 16V/(\gamma^2 \delta)$. The number of bits is at most $k(I + \gamma/2) + \log \ln(2/\delta) + 3$, which is at most $k(I + \gamma)$ once $k\gamma/2 \ge \log \ln(2/\delta) + 3$. □

The noisy bit, compressed. Take $k = 1000$ copies of Example 9.4, with $s = 600$ and $\delta = 0.005$ in Lemma 9.17. Alice sends 604 bits instead of 1000, and Bob's thousand bits are within total variation 0.006 of the real thing: the sum of the densities exceeds 600 only if fewer than 77 of the thousand noise coins come up 1, where 110 are expected. As k grows the rate approaches $I \approx 0.50$ bits per copy.

Two old theorems in a new light. If $M = X$, the density is the surprise $\log(1/\mu_X(x))$, the set G is a typical set, the proposals are a random codebook, and the theorem is Shannon's source coding theorem at rate $H(X)$ (Theorem 4.5). And compare the channel coding theorem (Theorem 5.1): a noisy channel can carry I bits per use. The theorem here is its reverse: I noiseless bits per use, together with shared randomness, can imitate the noisy channel.

9.7 From one message to a conversation

Proof sketch. Write the transcript as $T = (M_1, \dots, M_r)$ for a fixed number $r \geq 1$ of alternating rounds, allowing empty messages. Fix a round j, its speaker's input z, and a history $h = (R, M_{<j})$. Define

$$p^j_{z,h}(m) = \Pr[M_j = m \mid Z = z, h], \qquad q^j_h(m) = \Pr[M_j = m \mid h].$$

Factorization implies that the first distribution is unchanged by also conditioning on the other player's input. The speaker therefore knows $p^j_{z,h}$; both parties know the prediction q^j_h. The information density $\ell_j(z, h, m) = \log(p^j_{z,h}(m)/q^j_h(m))$ has mean

$$I_j = I(X, Y; M_j \mid R, M_{<j}), \qquad \sum_{j=1}^{r} I_j = \mathrm{IC}_\mu(\pi) =: I.$$

Across k independent runs, apply the public-proposal sampler to the product of the k next-message distributions. Concentration gives cost $k(I_j + \gamma/(2r)) + O(\log(r/\delta))$ and total variation error at most δ/r, for sufficiently large k. Since r is fixed, summing the costs gives at most $k(I + \gamma)$.

To handle approximate histories, compare each simulated round with a real round started from the ideal history distribution. Applying the next simulation step cannot increase an existing total variation discrepancy, so the new errors add to at most δ. Start with a single-instance protocol whose error is slightly below the target to leave room for this error. For product inputs, combining compression with direct sum proves the amortized characterization.

Further reading: detailed proof and extensions. The remainder of this section supplies the full proof, the small error adjustment, and the extension to correlated inputs. Section 9.8 discusses one-copy compression.

Theorem 9.19 (A conversation, many copies)**.** *Let μ be any distribution on input pairs, let π be a protocol, and let $I = \mathrm{IC}_\mu(\pi) = I(X, Y; T \mid R)$. For all $\gamma > 0$ and $\delta > 0$ there is k_0 such that for every $k \geq k_0$ there is a protocol σ on k pairs of inputs with the following properties.*

(a) The cost of σ is at most $k(I + \gamma)$.

(b) From the public record of σ one can compute $\widehat{W}_1, \dots, \widehat{W}_k$ such that, for $(X^k, Y^k) \sim \mu^k$,

$$\mathrm{TV}((X^k, Y^k, \widehat{W}^k), (X^k, Y^k, W^k)) \leq \delta,$$

where W_i is the record of an independent run of π on (X_i, Y_i).

In particular, if π computes f with error ε' on average under μ, then $C_k(\varepsilon'+\delta) \le k(I+\gamma)$ for all large k.

Proof. We may assume $\delta \le 1/2$. The public string of σ holds public strings $R_1, \dots, R_k$ for k runs of π, and randomness for proposals. The protocol has r rounds. Before round j both parties know the histories $h_1, \dots, h_k$ of the k simulated runs. In round j the speaker, say Alice, uses Lemma 9.17 for the product distributions

$$P = p^j_{x_1,h_1} \times \cdots \times p^j_{x_k,h_k}, \qquad Q = q^j_{h_1} \times \cdots \times q^j_{h_k},$$

with $s_j = k(I_j + \gamma')$ and error parameter δ', where $\gamma' = \gamma/(2r)$ and $\delta' = \delta/(2r)$. The proposals are samples from Q, computed by both parties from the public randomness and the histories. Bob's output is the vector of k simulated messages, which both parties append to the histories. (For an impossible history use fixed default distributions; assign a fixed answer to an invalid final record.) Three points complete the proof.

1. Cost. By the lemma, the rounds cost at most $\sum_j (s_j + \log\ln(1/\delta') + 3) = k(I + \gamma/2) + r(\log\ln(2r/\delta) + 3)$ bits. The second term does not depend on k, so the total is at most $k(I+\gamma)$ for large k.

2. One round, started correctly. Call *ideal* the experiment in which π is really run k times. Suppose that the inputs and the histories before round j are distributed as in the ideal experiment. Given their values, the ideal messages of round j have distribution P, and by the lemma the simulated ones are within $\delta' + 2P(G^{\mathrm{c}})$ of P. Here G^{c} is the event that $\sum_i \ell_j$ exceeds s_j. In the ideal experiment the k runs are independent and each ℓ_j has mean I_j and a finite variance V_j, so as in (9.15) the average of $P(G^{\mathrm{c}})$ is at most $V_j/(k\gamma'^2)$. Hence the inputs and histories after round j are within $\delta' + 2V_j/(k\gamma'^2)$ of ideal.

3. Errors add. Round j of σ applies one and the same random map to the inputs and histories, whatever their distribution, and applying a map cannot increase total variation. So if the distribution before round j is within e of ideal, the distribution after it is within $e + \delta' + 2V_j/(k\gamma'^2)$. After r rounds the distance is at most $\delta/2 + 2\sum_j V_j/(k\gamma'^2)$, which is at most δ for large k.

For the last statement, let σ output for copy i what π outputs on the record $\widehat{W}_i$. Whether this answer is wrong is an event about $(X_i, Y_i, \widehat{W}_i)$, and its probability differs from the ideal one by at most δ. □

The theorem has no assumption on μ. Its weak point is the quantifier: k must be large compared with the number of rounds, because every round pays a few bits of overhead that do not depend on k. We return to this in Section 9.8.

Lemma 9.20 (A little less error costs a little more information)**.** *For $0 < \delta < \varepsilon$,*

$$\mathrm{IC}^{\mathrm{av}}_\mu(f, \varepsilon - \delta) \;\le\; \mathrm{IC}^{\mathrm{av}}_\mu(f, \varepsilon) + \frac{\delta}{\varepsilon}\,\log(|\mathcal{X}|\,|\mathcal{Y}|).$$

Proof. Let π compute f with error ε on average. A public coin decides: with probability δ/ε Alice sends x, Bob sends y, and the output is $f(x,y)$; otherwise the parties run π. The error is at most $(1 - \delta/\varepsilon)\,\varepsilon = \varepsilon - \delta$. Given the coin, the information is $H(X,Y)$ in the first case and $\mathrm{IC}_\mu(\pi)$ in the second. Take the infimum over π. □

Proof of Theorem 9.5. Lower half. By Theorem 9.8, $C_k(\varepsilon)/k \ge \mathrm{IC}^{\mathrm{av}}_\mu(f, \varepsilon)$ for every k.

Upper half. Fix $0 < \delta < \varepsilon$ and $\gamma > 0$, and let π compute f with error $\varepsilon - \delta$ on average, with $\mathrm{IC}_\mu(\pi) \le \mathrm{IC}^{\mathrm{av}}_\mu(f, \varepsilon - \delta) + \gamma$. By Theorem 9.19, $C_k(\varepsilon) \le k(\mathrm{IC}_\mu(\pi) + \gamma)$ for all large k. With Lemma 9.20,

$$\limsup_{k\to\infty} \frac{C_k(\varepsilon)}{k} \le \mathrm{IC}^{\mathrm{av}}_\mu(f, \varepsilon) + \frac{\delta}{\varepsilon}\log(|\mathcal{X}|\,|\mathcal{Y}|) + 2\gamma.$$

Let δ and γ tend to zero. □

Only the lower half used the independence of X and Y: Alice and Bob had to invent inputs privately.

Quoted extension: correlated inputs and local outputs. For correlated inputs, a player's own input may predict a message better than the observer can. The relevant measure becomes

$$\mathrm{IC}_\mu^{\mathrm{int}}(\pi) = I(X;T \mid Y,R) + I(Y;T \mid X,R).$$

Here we change the output convention: both players recover their answers *locally*, using their own inputs and private coins as well as the public record. Let $C_k^{\mathrm{loc}}(\varepsilon)$ be the corresponding cost with average error at most ε in each copy (an error means either player's answer is wrong), and let $\mathrm{IC}_\mu^{\mathrm{int,av}}(f,\varepsilon)$ be the infimum of internal information cost over such single-instance protocols. The general Braverman–Rao theorem is

$$\lim_{k\to\infty} \frac{C_k^{\mathrm{loc}}(\varepsilon)}{k} = \mathrm{IC}_\mu^{\mathrm{int,av}}(f,\varepsilon), \qquad \varepsilon > 0.$$

Its upper bound needs a sampler for a reference distribution known to the receiver but not the sender. We quote that result here. Local output is essential: when $X = Y$ is a shared fair bit and $f(x,y) = x$, local answers require no communication, whereas publicly reconstructing k answers with coordinate error $\varepsilon < 1/2$ costs at least $k(1 - h_2(\varepsilon))$ bits. Thus the displayed theorem does not assert the same rate with public outputs.

Back to disjointness: interpreting the sharp constants. Braverman, Garg, Pankratov and Weinstein optimize the information cost of one AND and use amortization to obtain the coefficients quoted at the start. The intersection coefficient $C_\wedge \approx 1.4923$ uses the convention that *both players learn the intersection locally*; it is not a public-output upper bound proved by our theorem. For $C_{\mathrm{DISJ}} \approx 0.4827$, the measuring distribution is restricted to zero-inputs of AND, but the protocol must still be correct on *all four inputs*. This optimization differs from using our fixed ν. The DISJ answer can be made public for one extra bit, preserving its leading coefficient. In the symmetric case, the optimal AND protocol has a useful limiting picture. A player holding 0 chooses a private uniform time in $[0,1)$; a player holding 1 waits until time 1. The first interruption announces output 0; if neither interrupts before 1, the output is 1. The time reveals less than a direct request for an input bit. View this continuous clock as the limit of increasingly fine finite protocols. Proving its optimality and deriving the two coefficients require the further analysis in the cited paper.

9.8 Further reading: one copy and the limits of compression

The overhead in Lemma 9.17 is paid once per round. Across k copies it is amortized over k messages, while for one conversation it may dominate the information. Even one noisy bit illustrates the gap: Example 9.4 reveals about 0.50 bits, but an exact single-copy simulation needs 0.78 communicated bits on average (Exercise 5).

The general one-copy theory uses internal information and local reconstruction. At fixed positive distributional simulation error, Barak–Braverman–Chen–Rao obtain roughly $\sqrt{CI}$ bits up to logarithmic factors in C; for product inputs, Sherstov obtains $O((I+1)\log^2(I+2))$ bits independent of C.

Ganor–Kol–Raz give a partial Boolean function under a specified correlated distribution with internal information complexity $O(a)$ but distributional communication complexity at least 2^a. Its amortized local-output cost is $O(a)$ per copy. The resulting failure of tight direct sum is a *distributional* statement, not an every-input claim for total functions.

9.9 Summary

Step	What made it work
Information is at most communication	Given the public coins, a depth-c tree has at most 2^c leaves.
Information adds up	The copies are independent before the conversation.
One term is the cost of one copy	The other inputs can be invented privately: X and Y independent, given D if necessary.
Disjointness reduces to AND	Zero-inputs do not change an OR; correctness on every input.
AND costs a constant	Factorization, hence cut-and-paste.
Little information, little communication	Public proposals; the law of large numbers for the likelihood ratio.

Sources. The amortized characterization is due to Braverman and Rao, *Information equals amortized communication* (FOCS 2011). The public-proposal method is a basic reverse Shannon construction; see Harsha, Jain, McAllester and Radhakrishnan, *The communication complexity of correlation*. The hidden-selector/Hellinger method for DISJ is due to Bar-Yossef, Jayram, Kumar and Sivakumar, JCSS 68 (2004); the original linear lower bounds are due to Kalyanasundaram–Schnitger and Razborov. The sharp constants and clock protocol are from Braverman, Garg, Pankratov and Weinstein, *From information to exact communication*, Theorems 2.5–2.6 and Remark 5.5. For the one-copy comparison, see Barak–Braverman–Chen–Rao, Sherstov, and Ganor–Kol–Raz. The general direct-sum proof is further reading in Braverman's survey, Section 3.1.

Checks and exercises

1. Under ν, compute $H(X)$, $I(X;Y)$ and $I(X;Y \mid D)$. Compute IC_ν when both players always send their bits, and when Bob speaks first in Example 9.9. Why can no product distribution give positive probability to $10, 01$ but zero probability to 11?

2. Write out the DISJ embedding for $n = 2$, target $i = 1$, and each value of D_2. Then let Π send only $x_1 \oplus x_2$ and compute $I(Z_1; T \mid D)$ and $I(Z_1; T \mid D, Z_2)$ under ν^2. Which is the cost of the private-filler embedding?

3. Compare TV and h^2 for $P = (1,0), Q = (1-\gamma, \gamma)$ and for $P = (1/2, 1/2), Q = (1/2+\gamma, 1/2-\gamma)$. Show that
$$1 - e^{-kh^2(P,Q)} \le \mathrm{TV}(P^k, Q^k) \le \sqrt{2k}\, h(P,Q),$$
and estimate the sample count needed to distinguish each pair.

4. Show that a cost-1 public-output protocol for AND has error at least $1/3$ on some input, and construct one attaining exactly $1/3$. Thus strict error $< 1/3$ requires 2 bits. For fixed public coins, at most one player speaks and the record ignores the other input.

5. Alice holds $x \in \{0,1\}$ and Bob must output a bit equal to x with probability exactly 0.89, for each x. With public coins, prove that expected communication is at least 0.78 bits and attain it. For a fixed public string, the root either terminates or requires a bit.

6. A p-pass streaming algorithm uses s memory bits and decides whether any item from $[n]$ occurs twice, with error at most $\varepsilon < 1/2$ on every stream. Reduce DISJ to it and prove $(2p-1)s + 1 \ge \kappa_\varepsilon n$. The extra bit announces Bob's final answer.

Hints. (4) Choose equally among "send x, output x", "send y, output y", and "output 0". (5) With probability 0.22 output a public fair bit; otherwise send x.