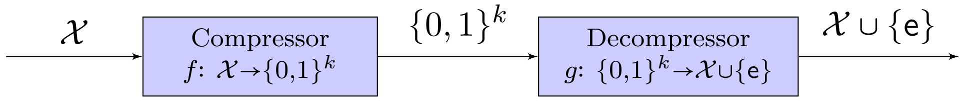
§ 7. Fixed-length (almost lossless) compression. Slepian-Wolf problem.

7.1 Fixed-length code, almost lossless

Coding paradigm:



$\mathcal{X}$
Compressor
f: $\mathcal{X}\to\{0,1\}^k$
$\{0,1\}^k$
Decompressor
g: $\{0,1\}^k\to\mathcal{X}\cup\{\mathsf{e}\}$
$\mathcal{X}\cup\{\mathsf{e}\}$

Note: If we want $g \circ f = \mathbf{1}_{\mathcal{X}}$, then $k \geq \log_2|\mathcal{X}|$. But, the transmission link is erroneous anyway... and it turns out that by tolerating a little error probability ϵ, we gain a lot in terms of code length!

Indeed, the key idea is to **allow errors**: Instead of insisting on $g(f(x)) = x$ for all $x \in \mathcal{X}$, consider only lossless decompression for a subset $\mathcal{S} \subset \mathcal{X}$:

$$g(f(x)) = \begin{cases} x & x \in \mathcal{S} \\ \mathsf{e} & x \notin \mathcal{S} \end{cases}$$

and the probability of error: $\mathbb{P}[g(f(X)) \neq X] = \mathbb{P}[g(f(X)) = \mathsf{e}]$.

Definition 7.1. A compressor-decompressor pair (f,g) is called a (k,ϵ)-code if:

$$\begin{aligned} f &: \mathcal{X} \to \{0,1\}^k \\ g &: \{0,1\}^k \to \mathcal{X} \cup \{\mathsf{e}\} \end{aligned}$$

such that $g(f(x)) \in \{x, \mathsf{e}\}$ and $\mathbb{P}[g(f(X)) = \mathsf{e}] \leq \epsilon$.

Fundamental limit:

$$\epsilon^*(X,k) \triangleq \inf\{\epsilon : \exists (k,\epsilon)\text{-code for } X\}$$

The following result connects the respective fundamental limits of fixed-length almost lossless compression and variable-length lossless compression (Lecture 6):

Theorem 7.1 (Fundamental limit of error probabiliy)**.**

$$\epsilon^*(X,k) = \mathbb{P}[l(f^*(X)) \geq k] = 1 - \text{sum of } 2^k - 1 \text{ largest masses of } X.$$

Proof. The proof is essentially tautological. Note $1 + 2 + \cdots + 2^{k-1} = 2^k - 1$. Let $\mathcal{S} = \{2^k - 1 \text{ most likely realizations of } X\}$. Then

$$\epsilon^*(X,k) = \mathbb{P}[X \notin \mathcal{S}] = \mathbb{P}[l(f^*(X)) \geq k].$$

Optimal codes:

- Variable-length: f^* encodes the 2^k-1 symbols with the highest probabilities to $\{\phi, 0, 1, 00, \dots, 1^{k-1}\}$.
- Fixed-length: The optimal compressor f maps the elements of $\mathcal{S}$ into $(00\dots00), \dots, (11\dots10)$ and the rest to $(11\dots11)$. The decompressor g decodes perfectly except for outputting e upon receipt of $(11\dots11)$. □

Note: In Definition 7.1 we require that the errors are always *detectable*, i.e., $g(f(x)) = x$ or e. Alternatively, we can drop this requirement and allow *undetectable* errors, in which case we can of course do better since we have more freedom in designing codes. It turns out that we do not gain much by this relaxation. Indeed, if we define

$$\tilde{\epsilon}^*(X,k) = \inf\{\mathbb{P}[g(f(X)) \neq X] : f : \mathcal{X} \to \{0,1\}^k, g : \{0,1\}^k \to \mathcal{X} \cup \{\mathsf{e}\}\},$$

then $\tilde{\epsilon}^*(X,k) = 1-$sum of 2^k largest masses of X. This follows immediately from $\mathbb{P}[g(f(X)) = X] = \sum_{x\in C} P_X(x)$ where $C \triangleq \{x : g(f(x)) = x\}$ satisfies $|C| \le 2^k$, because f takes no more than 2^k values. Compared to Theorem 7.1, we see that $\tilde{\epsilon}^*(X,k)$ and $\tilde{\epsilon}^*(X,k)$ do not differ much. In particular, $\epsilon^*(X,k+1) \le \tilde{\epsilon}^*(X,k) \le \epsilon^*(X,k)$.

Corollary 7.1 (Shannon)**.** *Let S^n be i.i.d. Then*

$$\lim_{n\to\infty} \epsilon^*(S^n, nR) = \begin{cases} 0 & R > H(S) \\ 1 & R < H(S) \end{cases}$$

$$\lim_{n\to\infty} \epsilon^*(S^n, nH(S) + \sqrt{nV(S)}\gamma) = 1 - \Phi(\gamma).$$

where $\Phi(\cdot)$ is the CDF of $\mathcal{N}(0,1)$, $H(S) = \mathbb{E}\log\frac{1}{P_S(S)}$ – entropy, $V(S) = \mathrm{Var}\log\frac{1}{P_S(S)}$ – varentropy is assumed to be finite.

Proof. Combine Theorem 7.1 with Theorem 6.1. □

Theorem 7.2 (Converse)**.**

$$\epsilon^*(X,k) \ge \tilde{\epsilon}^*(X,k) \ge \mathbb{P}\left[\log_2\frac{1}{P_X(X)} > k + \tau\right] - 2^{-\tau}, \quad \forall \tau > 0.$$

Proof. Identical to the converse of Theorem 6.4. Let $C = \{x : g(f(x)) = x\}$. Then $|C| \le 2^k$ and $\mathbb{P}[X \in C] \le \mathbb{P}\left[\log_2\frac{1}{P_X(X)} \le k + \tau\right] + \underbrace{\mathbb{P}\left[X \in C, \log_2\frac{1}{P_X(X)} > k + \tau\right]}_{\le 2^{-\tau}}$ □

Two achievability bounds

Theorem 7.3.

$$\epsilon^*(X,k) \le \mathbb{P}\left[\log_2\frac{1}{P_X(X)} \ge k\right] \tag{7.1}$$

and there exists a compressor-decompressor pair that achieves the upper bound.

Proof. Construction: use those $2^k - 1$ symbols with the highest probabilities.

This is essentially the same as the lower bound in Theorem 6.3 from Lecture 6. Note that the m^{th} largest mass $P_X(m) \le \frac{1}{m}$. Therefore

$$\epsilon^*(X,k) = \sum_{m\ge 2^k} P_X(m) = \sum \mathbf{1}_{\{m\ge 2^k\}} P_X(m) \le \sum \mathbf{1}_{\left\{\frac{1}{P_X(m)}\ge 2^k\right\}} P_X(m) = \mathbb{E}\mathbf{1}_{\left\{\log_2\frac{1}{P_X(X)}\ge k\right\}}.$$

□

Theorem 7.4.

$$\epsilon^*(X,k) \le \mathbb{P}\left[\log_2 \frac{1}{P_X(X)} > k - \tau\right] + 2^{-\tau}, \quad \forall \tau > 0 \tag{7.2}$$

and there exists a compressor-decompressor pair that achieves the upper bound.

Note: In fact, Theorem 7.3 is always stronger than Theorem 7.4. Still, we present the proof of Theorem 7.4 and the technology behind it – *random coding* – a powerful technique for proving existence (achievability) which we heavily rely on in this course. To see that Theorem 7.3 gives a better bound, note that even the first term in (7.2) exceeds (7.1). Nevertheless, the method of proof for this weaker bound will be useful for generalizations.

Proof. Construction: **random coding** (Shannon's magic). For a given compressor f, the optimal decompressor which minimizes the error probability is the maximum a posteriori (MAP) decoder, i.e.,

$$g^*(w) = \operatorname*{argmax}_x P_{X|f(X)}(x|w) = \operatorname*{argmax}_{x:f(x)=w} P_X(x),$$

which can be hard to analyze. Instead, let us consider the following (suboptimal) decompressor g:

$$g(w) = \begin{cases} x, & \exists!\ x \in \mathcal{X} \text{ s.t. } f(x) = w \text{ and } \log_2 \frac{1}{P_X(x)} \le k - \tau, \\ & \text{(exists unique high-probability } x \text{ that is mapped to } w\text{)} \\ \mathsf{e}, & \text{o.w.} \end{cases}$$

Denote $f(x) = c_x$ and the codebook $\mathcal{C} = \{c_x : x \in \mathcal{X}\} \subset \{0,1\}^k$. It is instructive to think of $\mathcal{C}$ as a hashing table.

Error probability analysis: There are two ways to make an error ⇒ apply union bound. Before proceeding, define

$$J(x,\mathcal{C}) \triangleq \left\{x' \in \mathcal{X} : c_{x'} = c_x, x' \ne x, \log_2 \frac{1}{P_X(x')} < k - \tau\right\}$$

to be the set of high-probability inputs whose hashes collide with that of x. Then we have the following estimate for probability of error:

$$\begin{aligned}\mathbb{P}[g(f(X)) = \mathsf{e}] &= \mathbb{P}\left[\left\{\log_2 \frac{1}{P_X(X)} \ge k - \tau\right\} \cup \{J(X,\mathcal{C}) \ne \varnothing\}\right] \\ &\le \mathbb{P}\left[\log_2 \frac{1}{P_X(X)} \ge k - \tau\right] + \mathbb{P}[J(X,\mathcal{C}) \ne \phi]\end{aligned}$$

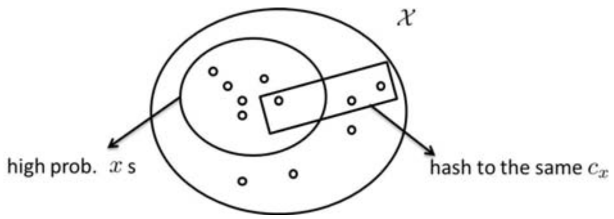
The first term does not depend on the codebook $\mathcal{C}$, while the second term does. The idea now is to randomize over $\mathcal{C}$ and show that when we average over all possible choices of codebook, the second term is smaller than $2^{-\tau}$. Therefore there exists at least one codebook that achieves the desired bound. Specifically, let us consider $\mathcal{C}$ which is uniformly distributed over all codebooks and independently of X. Equivalently, since $\mathcal{C}$ can be represented by a $|\mathcal{X}| \times k$ binary matrix, whose rows correspond to codewords, we choose each entry to be independent fair coin flips.

Averaging the error probability (over $\mathcal{C}$ and over X), we have

$$\begin{aligned}
\mathbb{E}_{\mathcal{C}}[\mathbb{P}[J(X,\mathcal{C}) \neq \phi]] &= \mathbb{E}_{\mathcal{C},X}\left[\mathbf{1}_{\left\{\exists x' \neq X : \log_2 \frac{1}{P_X(x')} < k-\tau, c_{x'} = c_X\right\}}\right] \\
&\leq \mathbb{E}_{\mathcal{C},X}\left[\sum_{x' \neq X} \mathbf{1}_{\left\{\log_2 \frac{1}{P_X(x')} < k-\tau\right\}} \mathbf{1}_{\{c_{x'} = c_X\}}\right] && \text{(union bound)} \\
&= 2^{-k}\mathbb{E}_X\left[\sum_{x' \neq X} \mathbf{1}_{\{P_X(x') > 2^{-k+\tau}\}}\right] \\
&\leq 2^{-k} \sum_{x' \in \mathcal{X}} \mathbf{1}_{\{P_X(x') > 2^{-k+\tau}\}} \\
&\leq 2^{-k} 2^{k-\tau} = 2^{-\tau}.
\end{aligned}$$

□

Note: Why the proof works: Compressor $f(x) = c_x$, hashing $x \in \mathcal{X}$ to a random k-bit string $c_x \in \{0,1\}^k$.



high-probability $x \Leftrightarrow \log_2 \frac{1}{P_X(x)} \leq k - \tau \Leftrightarrow P_X(x) \geq 2^{-k+\tau}$.

Therefore the cardinality of high-probability x's is at most $2^{k-\tau} \ll 2^k$ = number of strings. Hence the chance of collision is small.

Note: The random coding argument is a canonical example of *probabilistic method*: To prove the existence of something with certain property, we construct a probability distribution (randomize) and show that on average the property is satisfied. Hence there exists at least one realization with the desired property. The downside of this argument is that it is not constructive, i.e., does not give us an algorithm to find the object.

Note: This is a subtle point: Notice that in the proof we choose the random codebook to be uniform over all possible codebooks. In other words, $C = \{c_x : x \in \mathcal{X}\}$ consists of iid k-bit strings. In fact, in the proof we only need pairwise independence, i.e., $c_x \perp\!\!\!\perp c_{x'}$ for any $x \neq x'$ (Why?). Now, why should we care about this? In fact, having access to external randomness is also a lot of resources. It is more desirable to use less randomness in the random coding argument. Indeed, if we use zero randomness, then it is a deterministic construction, which is the best situation! Using pairwise independent codebook requires significantly less randomness than complete random coding which needs $|\mathcal{X}|k$ bits. To see this intuitively, note that one can use 2 independent random bits to generate 3 random bits that is pairwise independent but not mutually independent, e.g., $\{b_1, b_2, b_1 \oplus b_2\}$. This observation is related to linear compression studied in the next section, where the codeword we generated are not iid, but related through a linear mapping.

Remark 7.1 (AEP for memoryless sources)**.** Consider iid S^n. By WLLN,

$$\frac{1}{n} \log \frac{1}{P_{S^n}(S^n)} \xrightarrow{\mathbb{P}} H(S). \tag{7.3}$$

For any $\delta > 0$, define the set

$$T_n^\delta = \left\{ s^n : \left| \frac{1}{n} \log \frac{1}{P_{S^n}(s^n)} - H(S) \right| \leq \delta \right\}.$$

As a consequence of (7.3),

1. $\mathbb{P}\left[S^n \in T_n^\delta\right] \to 1$ as $n \to \infty$.
2. $|T_n^\delta| \le 2^{(H(S)+\delta)n} \ll |\mathcal{S}|^n$.

In other words, S^n is concentrated on the set T_n^δ which is exponentially smaller than the whole space. In almost compression we can simply encode this set losslessly. Although this is different than the optimal encoding, Corollary 7.1 indicates that in the large-n limit the optimal compressor is no better.

The property (7.3) is often referred as the *Asymptotic Equipartition Property* (AEP). Note that for any $s^n \in T_n^\delta$, its likelihood is concentrated around $P_{S^n}(s^n) \in 2^{-(H(S)\pm\delta)n}$, called δ-typical sequences.

Next we study fixed-blocklength code, fundamental limit of error probability $\epsilon^*(X,k)$ for the following coding paradigms:

- Linear Compression
- Compression with Side Information
  - side info available at both sides
  - side info available only at decompressor
  - multi-terminal compressor, single decompressor

7.2 Linear Compression

From Shannon's theorem:

$$\epsilon^*(X,nR) \longrightarrow 0 \text{ or } 1 \qquad\qquad R \gtrless H(S)$$

Our goal is to find compressor with structures. The simplest one can think of is probably linear operation, which is also highly desired for its simplicity (low complexity). But of course, we have to be on a vector space where we can define linear operations. In this part, we assume $X = S^n$, where each coordinate takes values in a finite field (Galois Field), i.e., $S_i \in \mathbb{F}_q$, where q is the cardinality of $\mathbb{F}_q$. This is only possible if $q = p^n$ for some prime p and $n \in \mathbb{N}$. So $\mathbb{F}_q = \mathbb{F}_{p^n}$.

Definition 7.2 (Galois Field)**.** F is a finite set with operations $(+,\cdot)$ where

- $a+b$ associative and commutative
- $a \cdot b$ associative and commutative
- $0,1 \in F$ s.t. $0+a = 1\cdot a = a$.
- $\forall a, \exists -a$, s.t. $a+(-a) = 0$
- $\forall a \neq 0, \exists a^{-1}$, s.t. $a^{-1}a = 1$
- distributive: $a\cdot(b+c) = (a\cdot b)+(a\cdot c)$

Example:

- $\mathbb{F}_p = \mathbb{Z}/p\mathbb{Z}$, where p is prime
- $\mathbb{F}_4 = \{0,1,x,x+1\}$ with addition and multiplication as polynomials $\mod (x^2+x+1)$ over $\mathbb{F}_2[x]$.

Linear Compression Problem: $x \in \mathbb{F}_q^n$, $w = Hx$ where $H : \mathbb{F}_q^n \to \mathbb{F}_q^k$ is linear represented by a matrix $H \in \mathbb{F}_q^{k\times n}$.

$$\begin{bmatrix} w_1 \\ \vdots \\ w_k \end{bmatrix} = \begin{bmatrix} h_{11} & \dots & h_{1n} \\ \vdots & & \vdots \\ h_{k1} & \dots & h_{kn} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$$

Compression is achieved if $k \le n$, i.e., H is a fat matrix. Of course, we have to tolerate some error (almost lossless). Otherwise, lossless compression is only possible with $k \ge n$, which not interesting.

Theorem 7.5 (Achievability)**.** *Let $X \in \mathbb{F}_q^n$ be a random vector. $\forall \tau > 0, \exists$ linear compressor $H : \mathbb{F}_q^n \to \mathbb{F}_q^k$ and decompressor $g : \mathbb{F}_q^k \to \mathbb{F}_q^n \cup \{\mathsf{e}\}$, s.t.*

$$\mathbb{P}[g(HX) \neq X] \leq \mathbb{P}\left[\log_q \frac{1}{P_X(X)} > k - \tau\right] + q^{-\tau}$$

Proof. Fix τ. As pointed in the proof of Shannon's random coding theorem (Theorem 7.4), given the compressor H, the optimal decompressor is the MAP decoder, i.e., $g(w) = \operatorname{argmax}_{x:Hx=w} P_X(x)$, which outputs the most likely symbol that is compatible with the codeword received. Instead, let us consider the following (suboptimal) decoder for its ease of analysis:

$$g(w) = \begin{cases} x & \exists! x \in \mathbb{F}_q^n : w = Hx,\ x - h.p. \\ \mathsf{e} & \text{otherwise} \end{cases}$$

where we used the short-hand:

$$x - h.p. \text{ (high probability)} \Leftrightarrow \log_q \frac{1}{P_X(x)} < k - \tau \Leftrightarrow P_X(x) \geq q^{-k+\tau}.$$

Note that this decoder is the same as in the proof of Theorem 7.4. The proof is also mostly the same, except now hash collisions occur under the linear map H. By union bound,

$$\begin{aligned} \mathbb{P}[g(f(X)) = \mathsf{e}] &\leq \mathbb{P}\left[\log_q \frac{1}{P_X(x)} > k - \tau\right] + \mathbb{P}\left[\exists x' - h.p. : x' \neq X, Hx' = HX\right] \\ \text{(union bound)} &\leq \mathbb{P}\left[\log_q \frac{1}{P_X(x)} > k - \tau\right] + \sum_x P_X(x) \sum_{x' - h.p., x' \neq x} \mathbf{1}\{Hx' = Hx\} \end{aligned}$$

Now we use random coding to average the second term over all possible choices of H. Specifically, choose H as a matrix independent of X where each entry is iid and uniform on $\mathbb{F}_q$. For distinct x_0 and x_1, the collision probability is

$$\begin{aligned} \mathbb{P}_H[Hx_1 = Hx_0] &= \mathbb{P}_H[Hx_2 = 0] && (x_2 \triangleq x_1 - x_0 \neq 0) \\ &= \mathbb{P}_H[H_1 \cdot x_2 = 0]^k && \text{(iid rows)} \end{aligned}$$

where H_1 is the first row of the matrix H, and each row of H is independent. This is the probability that H_i is in the orthogonal complement of x_2. On $\mathbb{F}_q^n$, the orthogonal complement of a given non-zero vector has cardinality q^{n-1}. So the probability for the first row to lie in this subspace is $q^{n-1}/q^n = 1/q$, hence the collision probability $1/q^k$. Averaging over H gives

$$\mathbb{E}_H \sum_{x' - h.p., x' \neq x} \mathbf{1}\{Hx' = Hx\} = \sum_{x' - h.p., x' \neq x} \mathbb{P}_H[Hx' = H_x] = |\{x' : x' - h.p., x' \neq x\}| q^{-k} \leq q^{k-\tau} q^{-k} = q^{-\tau}$$

Thus the bound holds. □

Notes:

1. Compared to Theorem 7.4, which is obtained by randomizing over all possible compressors, Theorem 7.5 is obtained by randomizing over only linear compressors, and the bound we obtained is identical. Therefore restricting on linear compression almost does not lose anything.

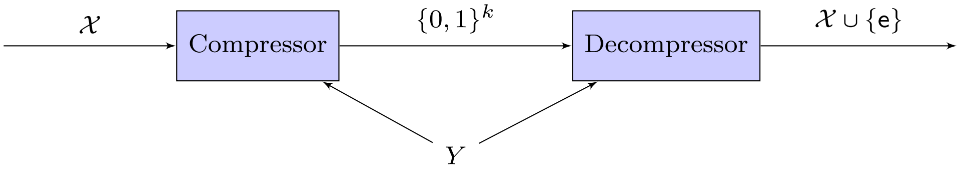
2. Note that in this case it is not possible to make all errors detectable.

3. Can we loosen the requirement on $\mathbb{F}_q$ to instead be a commutative ring? In general, no, since zero divisors in the commutative ring ruin the key proof item of low collision probability in the random hashing. E.g. in $\mathbb{Z}/6\mathbb{Z}$

$$\mathbb{P}\left[H\begin{bmatrix}1\\0\\\vdots\\0\end{bmatrix}=0\right]=6^{-k}\qquad\text{but}\qquad\mathbb{P}\left[H\begin{bmatrix}2\\0\\\vdots\\0\end{bmatrix}=0\right]=3^{-k},$$

since $0\cdot 2=3\cdot 2=0$ in $\mathbb{Z}/6\mathbb{Z}$.

7.3 Compression with Side Information at both compressor and decompressor



Definition 7.3 (Compression wih Side Information)**.** Given P_{XY},

- $f:\mathcal{X}\times\mathcal{Y}\to\{0,1\}^k$
- $g:\{0,1\}^k\times\mathcal{Y}\to\mathcal{X}\cup\{\mathsf{e}\}$
- $\mathbb{P}[g(f(X,Y),Y)\neq X]<\epsilon$
- Fundamental Limit: $\epsilon^*(X|Y,k)=\inf\{\epsilon:\ \exists(k,\epsilon)-S.I.\ code\}$

Note: The side information Y need not be discrete. The source X is, of course, discrete.

Note that conditioned on $Y=y$, the problem reduces to compression without side information where the source X is distributed according to $P_{X|Y=y}$. Since Y is known to both the compressor and decompressor, they can use the best code tailored for this distribution. Recall $\epsilon^*(X,k)$ defined in Definition 7.1, the optimal probability of error for compressing X using k bits, which can also be denoted by $\epsilon^*(P_X,k)$. Then we have the following relationship

$$\epsilon^*(X|Y,k)=\mathbb{E}_{y\sim P_Y}[\epsilon^*(P_{X|Y=y},k)],$$

which allows us to apply various bounds developed before.

Theorem 7.6.

$$\mathbb{P}\left[\log\frac{1}{P_{X|Y}(X|Y)}>k+\tau\right]-2^{-\tau}\le\epsilon^*(X|Y,k)\le\mathbb{P}\left[\log_2\frac{1}{P_{X|Y}(X|Y)}>k-\tau\right]+2^{-\tau},\quad\forall\tau>0$$

Corollary 7.2. *$(X,Y) = (S^n, T^n)$ where $(S_1,T_1),(S_2,T_2),\dots$ are iid pairs $\sim P_{ST}$*

$$\lim_{n\to\infty} \epsilon^*(S^n|T^n, nR) = \begin{cases} 0 & R > H(S|T) \\ 1 & R < H(S|T) \end{cases}$$

Proof. Using the converse Theorem 7.2 and achievability Theorem 7.4 (or Theorem 7.3) for compression without side information, we have

$$\mathbb{P}\left[\log\frac{1}{P_{X|Y}(X|y)} > k + \tau | Y = y\right] - 2^{-\tau} \le \epsilon^*(P_{X|Y=y}, k) \le \mathbb{P}\left[\log\frac{1}{P_{X|Y}(X|y)} > k | Y = y\right]$$

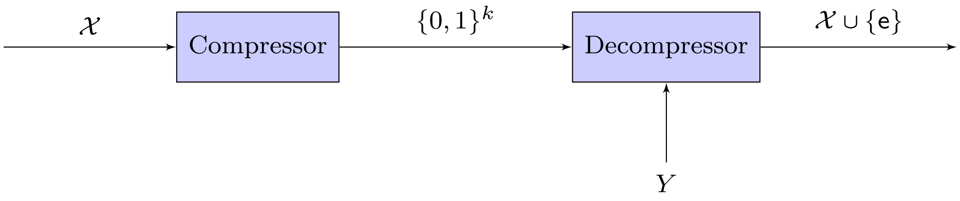
By taking the average over all $y \sim P_Y$, we get the theorem. For the corollary

$$\frac{1}{n}\log\frac{1}{P_{S^n|T^n}(S^n|T^n)} = \frac{1}{n}\sum_{i=1}^{n}\log\frac{1}{P_{S|T}(S_i|T_i)} \longrightarrow H(S|T) \text{ (in probability)}$$

as $n \to \infty$, using the WLLN. □

7.4 Slepian-Wolf (Compression with Side Information at Decompressor only)

Consider the compression with side information problem, except now the compressor has no access to the side information.



Definition 7.4 (S.W. code)**.** Given P_{XY},

- $f : \mathcal{X} \to \{0,1\}^k$
- $g : \{0,1\}^k \times \mathcal{Y} \to \mathcal{X} \cup \{\mathsf{e}\}$
- $\mathbb{P}[g(f(X),Y) \ne X] \le \epsilon$
- Fundamental Limit: $\epsilon^*_{\text{SW}} = \inf\{\epsilon : \exists(k,\epsilon)\text{-S.W. code}\}$

Now the very surprising result: Even without side information at the compressor, we can still compress down to the conditional entropy!

Theorem 7.7 (Slepian-Wolf, ’73)**.**

$$\epsilon^*(X|Y,k) \le \epsilon^*_{\text{SW}}(X|Y,k) \le \mathbb{P}\left[\log\frac{1}{P_{X|Y}(X|Y)} \ge k - \tau\right] + 2^{-\tau}$$

Corollary 7.3.

$$\lim_{n\to\infty} \epsilon^*_{\mathrm{SW}}(S^n|T^n, nR) = \begin{cases} 0 & R > H(S|T) \\ 1 & R < H(S|T) \end{cases}$$

Note: Definition 7.4 does not include the zero-undected-error condition (that is $g(f(x), y) = x$ or e). In other words, we allow for the possibility of undetected errors. Indeed, if we require this condition, the side-information savings will be mostly gone. Indeed, assuming $P_{X,Y}(x,y) > 0$ for all (x,y) it is clear that under zero-undetected-error condition, if $f(x_1) = f(x_2) = c$ then $g(c) = \mathsf{e}$. Thus except for c all other elements in $\{0,1\}^k$ must have unique preimages. Similarly, one can show that Slepian-Wolf theorem does not hold if one uses the setting of variable-length lossless compression (i.e. average length is $H(X)$ not $H(X|Y)$.)

Proof. LHS is obvious, since side information at the compressor and decoder is better than only at the decoder.

For the RHS, first generate a random codebook with iid uniform codewords: $C = \{c_x \in \{0,1\}^k : x \in \mathcal{X}\}$ independently of (X, Y), then define the compressor and decoder as

$$\begin{aligned} f(x) &= C_x \\ g(w,y) &= \begin{cases} x & \exists! x : C_x = w, x - h.p.|y \\ 0 & \text{o.w.} \end{cases} \end{aligned}$$

where we used the shorthand $x - h.p.|y \Leftrightarrow \log_2 \frac{1}{P_{X|Y}(x|y)} < k - \tau$. The error probability of this scheme is

$$\begin{aligned} \mathcal{E}(C) &= \mathbb{P}\left[\log \frac{1}{P_{X|Y}(X|Y)} \geq k - \tau \text{ or } J(X, C|Y) \neq \varnothing\right] \\ &\leq \mathbb{P}\left[\log \frac{1}{P_{X|Y}(X|Y)} \geq k - \tau\right] + \mathbb{P}[J(X, C|Y) \neq \varnothing] \\ &= \mathbb{P}\left[\log \frac{1}{P_{X|Y}(X|Y)} \geq k - \tau\right] + \sum_{x,y} P_{XY}(x,y) \mathbf{1}_{\{J(x,C|y)\neq\varnothing\}}. \end{aligned}$$

where $J(x, C|y) \triangleq \{x' \neq x : x' - h.p.|y, c_x = c_{x'}\}$.

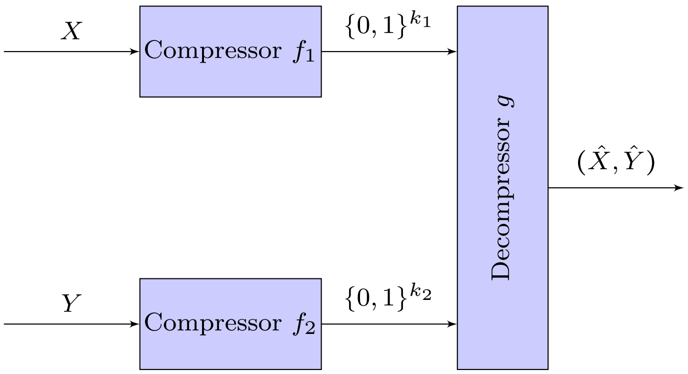
Now averaging over C and applying the union bound: use $|\{x' : x' - h.p.|y\}| \leq 2^{k-\tau}$ and $\mathbb{P}[C_{x'} = C_x] = 2^{-k}$ for any $x \neq x'$,

$$\begin{aligned} \mathbb{P}_C[J(x, C|y) \neq \varnothing] &\leq \mathbb{E}_C\left[\sum_{x' \neq x} \mathbf{1}_{\{x' - h.p.|y\}} \mathbf{1}_{\{C_{x'} = C_x\}}\right] \\ &= 2^{k-\tau} \mathbb{P}[C_{x'} = C_x] \\ &= 2^{-\tau} \end{aligned}$$

Hence the theorem follows as usual from two terms in the union bound. □

7.5 Multi-terminal Slepian Wolf

Distributed compression: Two sources are correlated. Compress individually, decompress jointly. What are those rate pairs that guarantee successful reconstruction?



Definition 7.5. Given P_{XY},

- (f_1,f_2,g) is (k_1,k_2,ϵ)-code if $f_1:\mathcal{X}\to\{0,1\}^{k_1}$, $f_2:\mathcal{Y}\to\{0,1\}^{k_2}$, $g:\{0,1\}^{k_1}\times\{0,1\}^{k_2}\to\mathcal{X}\times\mathcal{Y}$, s.t. $\mathbb{P}[(\hat{X},\hat{Y})\neq(X,Y)]\leq\epsilon$, where $(\hat{X},\hat{Y})=g(f_1(X),f_2(Y))$.
- Fundamental limit: $\epsilon^*_{\text{SW}}(X,Y,k_1,k_2)=\inf\{\epsilon:\exists(k_1,k_2,\epsilon)\text{-code}\}$.

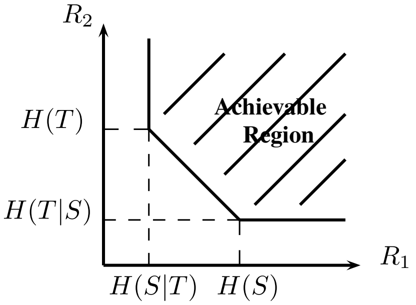
Theorem 7.8. $(X,Y)=(S^n,T^n)$ *- iid pairs*

$$\lim_{n\to\infty}\epsilon^*_{\text{SW}}(S^n,T^n,nR_1,nR_2)=\begin{cases}0 & (R_1,R_2)\in\mathcal{R}_{\text{SW}}\\ 1 & (R_1,R_2)\notin\mathcal{R}_{\text{SW}}\end{cases}$$

where $\mathcal{R}_{\text{SW}}$ *denotes the Slepian-Wolf rate region*

$$\mathcal{R}_{\text{SW}}=\left\{(a,b):\begin{array}{l}a\geq H(S|T)\\ b\geq H(T|S)\\ a+b\geq H(S,T)\end{array}\right.$$

Note: The rate region $\mathcal{R}_{\text{SW}}$ typically looks like:

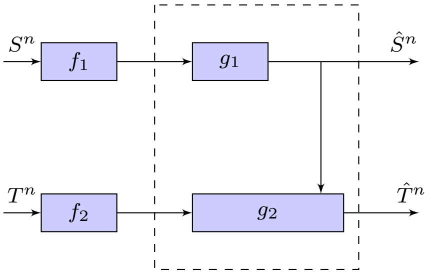


Since $H(T)-H(T|S)=H(S)-H(S|T)=I(S;T)$, the slope is -1.

Proof. Converse: Take $(R_1,R_2)\notin\mathcal{R}_{\text{SW}}$. Then one of three cases must occur:

1. $R_1<H(S|T)$. Then even if encoder and decoder had full T^n, still can't achieve this (from compression with side info result – Corollary 7.2).
2. $R_2<H(T|S)$ (same).
3. $R_1+R_2<H(S,T)$. Can't compress below the joint entropy of the pair (S,T).

Achievability: First note that we can achieve the two corner points. The point $(H(S), H(T|S))$ can be approached by almost lossless compressing S at entropy and compressing T with side information S at the decoder. To make this rigorous, let $k_1 = n(H(S)+\delta)$ and $k_2 = n(H(T|S)+\delta)$. By Corollary 7.1, there exist $f_1 : \mathcal{S}^n \to \{0,1\}^{k_1}$ and $g_1 : \{0,1\}^{k_1} \to \mathcal{S}^n$ s.t. $\mathbb{P}[g_1(f_1(S^n)) \neq S^n] \leq \epsilon_n \to 0$. By Theorem 7.7, there exist $f_2 : \mathcal{T}^n \to \{0,1\}^{k_2}$ and $g_2 : \{0,1\}^{k_1} \times \mathcal{S}^n \to \mathcal{T}^n$ s.t. $\mathbb{P}[g_2(f_2(T^n), S^n) \neq T^n] \leq \epsilon_n \to 0$. Now that S^n is not available, feed the S.W. decompressor with $g(f(S^n))$ and define the joint decompressor by $g(w_1, w_2) = (g_1(w_1), g_2(w_2, g_1(w_1)))$ (see below):



Apply union bound:

$$\begin{aligned}
&\mathbb{P}[g(f_1(S^n), f_2(T^n)) \neq (S^n, T^n)] \\
&= \mathbb{P}[g(f_1(S^n)) \neq S^n] + \mathbb{P}[g_2(f_2(T^n), g(f_1(S^n))) \neq T^n, g(f_1(S^n)) = S^n] \\
&\leq \mathbb{P}[g(f_1(S^n)) \neq S^n] + \mathbb{P}[g_2(f_2(T^n), S^n) \neq T^n] \\
&\leq 2\epsilon_n \to 0.
\end{aligned}$$

Similarly, the point $(H(S), H(T|S))$ can be approached.

To achieve other points in the region, use the idea of **time sharing**: If you can achieve with vanishing error probability any two points (R_1, R_2) and (R_1', R_2'), then you can achieve for $\lambda \in [0,1]$, $(\lambda R_1 + \bar{\lambda} R_1', \lambda R_2 + \bar{\lambda} R_2')$ by dividing the block of length n into two blocks of length λn and $\bar{\lambda} n$ and apply the two codes respectively

$$\begin{aligned}
(S_1^{\lambda n}, T_1^{\lambda n}) &\to \begin{bmatrix} \lambda n R_1 \\ \lambda n R_2 \end{bmatrix} \quad \text{using } (R_1, R_2) \text{ code} \\
(S_{\lambda n+1}^n, T_{\lambda n+1}^n) &\to \begin{bmatrix} \bar{\lambda} n R_1' \\ \bar{\lambda} n R_2' \end{bmatrix} \quad \text{using } (R_1', R_2') \text{ code}
\end{aligned}$$

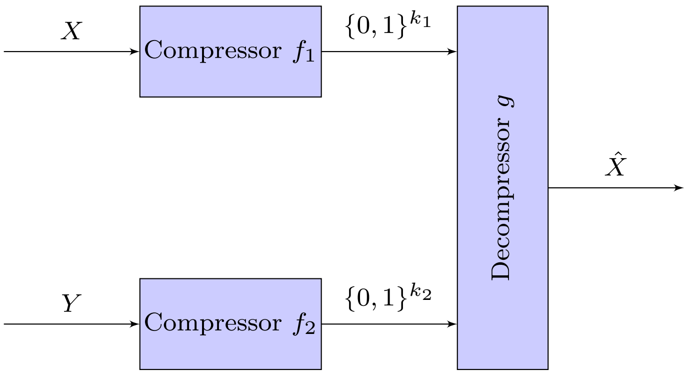
(Exercise: Write down the details rigorously yourself!) Therefore, all convex combinations of points in the achievable regions are also achievable, so the achievable region must be convex. □

7.6* Source-coding with a helper (Ahlswede-Körner-Wyner)

Yet another variation of distributed compression problem is compressing X with a helper, see figure below. Note that the main difference from the previous section is that decompressor is only required to produce the estimate of X, using rate-limited help from an observer who has access to Y. Characterization of rate pairs R_1, R_2 is harder than in the previous section.

Theorem 7.9 (Ahlswede-Körner-Wyner)**.** *Consider i.i.d. source $(X^n, Y^n) \sim P_{X,Y}$ with X discrete. If rate pair (R_1, R_2) is achievable with vanishing probability of error $\mathbb{P}[\hat{X}^n \neq X^n] \to 0$, then there exists an auxiliary random variable U taking values on alphabet of cardinality $|\mathcal{Y}| + 1$ such that $P_{X,Y,U} = P_{X,Y} P_{U|X,Y}$ and*

$$R_1 \geq H(X|U), R_2 \geq I(Y;U). \tag{7.4}$$



Furthermore, for every such random variable U the rate pair $(H(X|U), I(Y;U))$ is achievable with vanishing error.

Proof. We only sketch some crucial details.

First, note that iterating over all possible random variables U (without cardinality constraint) the set of pairs (R_1, R_2) satisfying (7.4) is convex. Next, consider a compressor $W_1 = f_1(X^n)$ and $W_2 = f_2(Y^n)$. Then from Fano's inequality (5.7) assuming $\mathbb{P}[X^n \neq \hat{X}^n] = o(1)$ we have

$$H(X^n|W_1, W_2)) = o(n)\,.$$

Thus, from chain rule and conditioning-decreases-entropy, we get

$$\begin{aligned} nR_1 \geq I(X^n; W_1|W_2) &\geq H(X^n|W_2) - o(n) && (7.5)\\ &= \sum_{k=1}^n H(X_k|W_2, X^{k-1}) - o(n) && (7.6)\\ &\geq \sum_{k=1}^n H(X_k|W_2, X^{k-1}, Y^{k-1}) - o(n) && (7.7) \end{aligned}$$

On the other hand, from (5.2) we have

$$\begin{aligned} nR_2 \geq I(W_2; Y^n) &= \sum_{k=1}^n I(W_2; Y_k|Y^{k-1}) && (7.8)\\ &= \sum_{k=1}^n I(W_2, X^{k-1}; Y_k|Y^{k-1}) && (7.9)\\ &= \sum_{k=1}^n I(W_2, X^{k-1}, Y^{k-1}; Y_k) && (7.10) \end{aligned}$$

where (7.9) follows from $I(W_2, X^{k-1}; Y_k|Y^{k-1}) = I(W_2; Y_k|Y^{k-1}) + I(X^{k-1}; Y_k|W_2, Y^{k-1})$ and the fact that $(W_2, Y_k) \perp\!\!\!\perp X^{k-1}|Y^{k-1}$; and (7.10) from $Y^{k-1} \perp\!\!\!\perp Y_k$. Comparing (7.7) and (7.10) we notice that denoting $U_k = (W_2, X^{k-1}, Y^{k-1})$ we have

$$(R_1, R_2) \geq \frac{1}{n}\sum_{k=1}^n (H(X_k|U_k), I(U_k; Y_k))$$

and thus (from convexity) the rate pair must belong to the region spanned by all pairs $(H(X|U), I(U;Y))$.

To show that without loss of generality the auxiliary random variable U can be taken to be $|\mathcal{Y}| + 1$ valued, one needs to invoke Caratheodory's theorem on convex hulls. We omit the details.

Finally, showing that for each U the mentioned rate-pair is achievable, we first notice that if there were side information at the decompressor in the form of the i.i.d. sequence U^n correlated to X^n, then Slepian-Wolf theorem implies that only rate $R_1 = H(X|U)$ would be sufficient to reconstruct X^n. Thus, the question boils down to creating a correlated sequence U^n at the decompressor by using the minimal rate R_2. This is the content of the so called covering lemma, see Theorem 24.5 below: It is sufficient to use rate $I(U;Y)$ to do so. We omit further details. □