**Last time:** For stationary memoryless (iid) sources and separable distortion, under the assumption that $D_{\max} < \infty$.

$$R(D) = R_i(D) = \inf_{P_{\hat{S}|S}: \mathbb{E}d(S,\hat{S}) \le D} I(S;\hat{S}).$$

## 25.1 Evaluation of $R(D)$

So far we've proved some properties about the rate distortion function, now we'll compute its value for a few simple statistical sources. We'll do this in a somewhat unsatisfying way: guess the answer, then verify its correctness. At the end, we'll show that there is a pattern behind this method.

### 25.1.1 Bernoulli Source

Let $S \sim \mathrm{Ber}(p)$, $p \le 1/2$, with Hamming distortion $d(S,\hat{S}) = \mathbf{1}\{S \ne \hat{S}\}$ and alphabets $\mathcal{A} = \hat{\mathcal{A}} = \{0,1\}$. Then $d(s^n,\hat{s}^n) = \frac{1}{n}\|s^n - \hat{s}^n\|_{\text{Hamming}}$ is the bit-error rate.

**Claim:** $R(D) = |h(p) - h(D)|^+$

*Proof.* Since $D_{\max} = p$, in the sequel we can assume $D < p$ for otherwise there is nothing to show.

(Achievability) We're free to choose any $P_{\hat{S}|S}$, so choose $S = \hat{S} + Z$, where $\hat{S} \sim \mathrm{Ber}(p') \perp\!\!\!\perp Z \sim \mathrm{Ber}(D)$, and $p'$ is such that $p'(1-D) + (1-p')D = p$ so that $p' < p$. In other words, the backward channel $P_{S|\hat{S}}$ is a BSC$(D)$. This induces some forward channel $P_{\hat{S}|S}$. Then,

$$I(S;\hat{S}) = H(S) - H(S|\hat{S}) = h(p) - h(D)$$

Since one such $P_{\hat{S}|S}$ exists, we have the upper bound $R(D) \le h(p) - h(D)$.

(Converse) First proof: For any $P_{\hat{S}|S}$ such that $P[S \ne \hat{S}] \le D \le p \le \frac{1}{2}$,

$$\begin{aligned} I(S;\hat{S}) &= H(S) - H(S|\hat{S}) \\ &= H(S) - H(S+\hat{S}|\hat{S}) \\ &\ge H(S) - H(S+\hat{S}) \\ &= h(p) - h(P[S \ne \hat{S}]) \\ &\ge h(p) - h(D) \end{aligned}$$

Second proof: Here is a more general strategy. Denote the random transformation from the achievability proof by $P^*_{\hat{S}|S}$. Now we need to show that there is no better $Q_{\hat{S}|S}$ with $\mathbb{E}_Q[d(S,\hat{S})] \le D$

and a smaller mutual information. Then consider the chain:

$$\begin{aligned} R(D) &\le I(P_S, Q_{\hat{S}|S}) = D(Q_{S|\hat{S}} \| P_S | Q_{\hat{S}}) \\ &= D(Q_{S|\hat{S}} \| P_{S|\hat{S}} | Q_{\hat{S}}) + \mathbb{E}_Q\left[\log \frac{P_{S|\hat{S}}}{P_S}\right] \\ (\text{Marginal } Q_{S\hat{S}} = P_S Q_{\hat{S}|S}) \quad &= D(Q_{S|\hat{S}} \| P_{S|\hat{S}} | Q_{\hat{S}}) + H(S) + \mathbb{E}_Q[\log D\mathbf{1}\{S \neq \hat{S}\} + \log \bar{D}\mathbf{1}\{S = \hat{S}\}] \end{aligned}$$

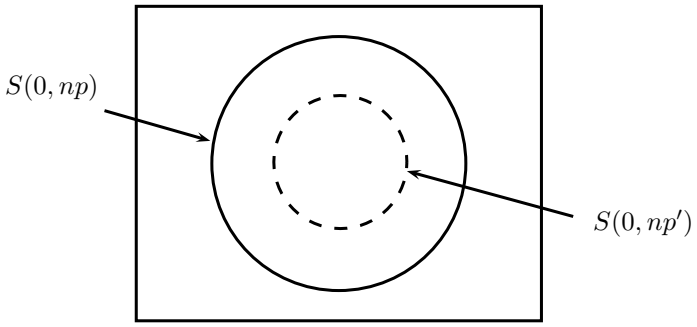
And we can minimize this expression by taking $Q_{S|\hat{S}} = P_{S|\hat{S}}$, giving

$$\ge 0 + H(S) + P[S = \hat{S}] \log(1 - D) + P[S \neq \hat{S}] \log D \ge h(p) - h(D) \quad (D \le 1/2) \tag{25.1}$$

Since the upper and lower bound agree, we have $R(D) = |h(p) - h(D)|^+$. □

For example, when $p = 1/2$, $D = .11$, then $R(D) = 1/2$ bit. In the Hamming game where we compressed 100 bits down to 50, we indeed can do this while achieving 11% average distortion, compared to the naive scheme of storing half the string and guessing on the other half, which achieves 25% average distortion.

**Interpretation:** By WLLN, the distribution $P_S^n = \text{Ber}(p)^n$ concentrates near the Hamming sphere of radius $np$ as $n$ grows large. The above result about Hamming sources tells us that the optimal reconstruction points are from $P_{\hat{S}}^n = \text{Ber}(p')^n$ where $p' < p$, which concentrates on a sphere of radius $np'$ (note the reconstruction points are some exponentially small subset of this sphere).



Hamming Spheres

It is interesting to note that *none* of the reconstruction points are the same as any of the possible source values (with high probability).

### 25.1.2 Gaussian Source

The Gaussian source is defines as $\mathcal{A} = \hat{\mathcal{A}} = \mathbb{R}$, $S \sim \mathcal{N}(0, \sigma^2)$, $d(a, \hat{a}) = |a - \hat{a}|^2$ (MSE distortion).

**Claim:** $R(D) = \frac{1}{2} \log^+ \frac{\sigma^2}{D}$.

*Proof.* Since $D_{\max} = \sigma^2$, in the sequel we can assume $D < \sigma^2$ for otherwise there is nothing to show.

(Achievability) Choose $S = \hat{S} + Z$ , where $\hat{S} \sim \mathcal{N}(0, \sigma^2 - D) \perp\!\!\!\perp Z \sim \mathcal{N}(0, D)$. In other words, the backward channel $P_{S|\hat{S}}$ is AWGN with noise power $D$. Since everything is jointly Gaussian, the forward channel can be easily found to be $P_{\hat{S}|S} = \mathcal{N}(\frac{\sigma^2 - D}{\sigma^2} S, \frac{\sigma^2 - D}{\sigma^2} D)$. Then

$$I(S; \hat{S}) = \frac{1}{2} \log \frac{\sigma^2}{D} \implies R(D) \le \frac{1}{2} \log \frac{\sigma^2}{D}$$

(Converse) Let $P_{\hat{S}|S}$ be any conditional distribution such that $\mathbb{E}_P|S-\hat{S}|^2 \le D$. Denote the forward channel in the achievability by $P^*_{\hat{S}|S}$. We use the same trick as before

$$\begin{aligned}
I(P_S, P_{\hat{S}|S}) &= D(P_{S|\hat{S}}\|P^*_{S|\hat{S}}|P_{\hat{S}}) + \mathbb{E}_P\left[\log\frac{P^*_{S|\hat{S}}}{P_S}\right] \\
&\ge \mathbb{E}_P\left[\log\frac{P^*_{S|\hat{S}}}{P_S}\right] \\
&= \mathbb{E}_P\left[\log\frac{\frac{1}{\sqrt{2\pi D}}e^{-\frac{(S-\hat{S})^2}{2D}}}{\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{S^2}{2\sigma^2}}}\right] \\
&= \frac{1}{2}\log\frac{\sigma^2}{D} + \frac{\log e}{2}\mathbb{E}_P\left[\frac{S^2}{\sigma^2} - \frac{|S-\hat{S}|^2}{D}\right] \\
&\ge \frac{1}{2}\log\frac{\sigma^2}{D}.
\end{aligned}$$

Again, the upper and lower bounds agree. □

The interpretation in the Gaussian case is very similar to the case of the Hamming source. As $n$ grows large, our source distribution concentrates on $S(0,\sqrt{n\sigma^2})$ ($n$-sphere in Euclidean space rather than Hamming), and our reconstruction points on $S(0,\sqrt{n(\sigma^2-D)})$. So again the picture is two nested sphere.

How sensitive is the rate-distortion formula to the Gaussianity assumption of the source?

**Theorem 25.1.** *Assume that* $\mathbb{E}S = 0$ *and* $\operatorname{Var} S = \sigma^2$. *Let the distortion metric be quadratic:* $d(s,\hat{s}) = (s-\hat{s})^2$. *Then*

$$\frac{1}{2}\log^+\frac{\sigma^2}{D} - D(P_S\|\mathcal{N}(0,\sigma^2)) \le R(D) = \inf_{P_{\hat{S}|S}:\mathbb{E}(\hat{S}-S)^2\le D} I(S;\hat{S}) \le \frac{1}{2}\log^+\frac{\sigma^2}{D}.$$

**Note**: This result is in exact parallel to what we proved in Theorem 17.6 for additive-noise channel capacity:

$$\frac{1}{2}\log\left(1+\frac{P}{\sigma^2}\right) \le \sup_{P_X:\mathbb{E}X^2\le P} I(X;X+Z) \le \frac{1}{2}\log\left(1+\frac{P}{\sigma^2}\right) + D(P_Z\|\mathcal{N}(0,\sigma^2)).$$

where $\mathbb{E}Z = 0$ and $\operatorname{Var} Z = \sigma^2$.

**Note**: A simple consequence of Theorem 25.1 is that for source distributions with a density, the rate-distortion function grows according to $\frac{1}{2}\log\frac{1}{D}$ in the low-distortion regime as long as $D(P_S\|\mathcal{N}(0,\sigma^2))$ is finite. In fact, the first inequality, known as the *Shannon lower bound*, is asymptotically tight, i.e., $R(D) = \frac{1}{2}\log\frac{\sigma^2}{D} - D(P_S\|\mathcal{N}(0,\sigma^2)) + o(1)$ as $D\to 0$. Therefore in this regime performing uniform scalar quantization with accuracy $\frac{1}{\sqrt{D}}$ is in fact asymptotically optimal within an $o(1)$ term.

*Proof.* Again, assume $D < D_{\max} = \sigma^2$. Let $S_G \sim \mathcal{N}(0,\sigma^2)$.

(Achievability) Use the same $P^*_{\hat{S}|S} = \mathcal{N}(\frac{\sigma^2-D}{\sigma^2}S, \frac{\sigma^2-D}{\sigma^2}D)$ in the achievability proof of Gaussian rate-distortion function:

$$\begin{aligned}
R(D) &\le I(P_S, P^*_{\hat{S}|S}) \\
&= I(S; \frac{\sigma^2-D}{\sigma^2}S + W) && W \sim \mathcal{N}(0, \frac{\sigma^2-D}{\sigma^2}D) \\
&\le I(S_G; \frac{\sigma^2-D}{\sigma^2}S_G + W) && \text{by Gaussian saddle point (Theorem 4.6)} \\
&= \frac{1}{2}\log\frac{\sigma^2}{D}.
\end{aligned}$$

(Converse) For any $P_{\hat{S}|S}$ such that $\mathbb{E}(\hat{S}-S)^2 \le D$. Let $P^*_{S|\hat{S}} = \mathcal{N}(\hat{S}, D)$ denote AWGN with noise power $D$. Then

$$\begin{aligned}
I(S;\hat{S}) &= D(P_{S|\hat{S}}\|P_S|P_{\hat{S}}) \\
&= D(P_{S|\hat{S}}\|P^*_{S|\hat{S}}|P_{\hat{S}}) + \mathbb{E}_P\left[\log\frac{P^*_{S|\hat{S}}}{P_{S_G}}\right] - D(P_S\|P_{S_G}) \\
&\ge \mathbb{E}_P\left[\log\frac{\frac{1}{\sqrt{2\pi D}}e^{-\frac{(S-\hat{S})^2}{2D}}}{\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{S^2}{2\sigma^2}}}\right] - D(P_S\|P_{S_G}) \\
&\ge \frac{1}{2}\log\frac{\sigma^2}{D} - D(P_S\|P_{S_G}).
\end{aligned}$$

□

**Remark:** The theory of quantization and the rate distortion theory at large have played a significant role in pure mathematics. For instance, Hilbert's thirteenth problem was partially solved by Arnold and Kolmogorov after they realized that they could classify spaces of functions looking at the optimal quantizer for such functions.

## 25.2* Analog of saddle-point property in rate-distortion

In the computation of $R(D)$ for the Hamming and Gaussian source, we guessed the correct form of the rate distortion function. In both of their converse arguments, we used the same trick to establish that any other $P_{\hat{S}|S}$ gave a larger value for $R(D)$. In this section, we formalize this trick, in an analogous manner to the saddle point property of the channel capacity. Note that typically we don't need any tricks to compute $R(D)$, since we can obtain a solution in parametric form to the unconstrained convex optimization

$$\min_{P_{\hat{S}|S}} I(S;\hat{S}) + \lambda\,\mathbb{E}[d(S,\hat{S})]$$

In fact there are also iterative algorithms (Blahut-Arimoto) that computes $R(D)$. However, for peace of mind it is good to know there are some general reasons why tricks like we used in Hamming/Gaussian actually are guaranteed to work.

**Theorem 25.2.** *1. Suppose $P_{Y^*}$ and $P_{X|Y^*} \ll P_X$ are found with the property that $\mathbb{E}[d(X,Y^*)] \le D$ and for any $P_{XY}$ with $\mathbb{E}[d(X,Y)] \le D$ we have*

$$\mathbb{E}\left[\log \frac{dP_{X|Y^*}}{dP_X}(X|Y)\right] \ge I(X;Y^*)\,. \tag{25.2}$$

*Then $R(D) = I(X;Y^*)$.*

*2. Suppose that $I(X;Y^*) = R(D)$. Then for any regular branch of conditional probability $P_{X|Y^*}$ and for any $P_{XY}$ satisfying*

- *$\mathbb{E}[d(X,Y)] \le D$ and*
- *$P_Y \ll P_{Y^*}$ and*
- *$I(X;Y) < \infty$*

*the inequality (25.2) holds.*

**Remarks**:

1. The first part is a sufficient condition for optimality of a given $P_{XY^*}$. The second part gives a necessary condition that is convenient to narrow down the search. Indeed, typically the set of $P_{XY}$ satisfying those conditions is rich enough to infer from (25.2):

$$\log \frac{dP_{X|Y^*}}{dP_X}(x|y) = R(D) - \theta[d(x,y) - D]\,,$$

   for a positive $\theta > 0$.

2. Note that the second part is not valid without $P_Y \ll P_{Y^*}$ condition. The counter-example to this and various other erroneous (but frequently encountered) generalizations is the following: $\mathcal{A} = \{0,1\}$, $P_X = \mathrm{Bern}(1/2)$, $\hat{\mathcal{A}} = \{0,1,0',1'\}$ and

$$d(0,0) = d(0,0') = 1 - d(0,1) = 1 - d(0,1') = 0\,.$$

   The $R(D) = |1 - h(D)|^+$, but there are a bunch of non-equivalent optimal $P_{Y|X}$, $P_{X|Y}$ and $P_Y$'s.

*Proof.* First part is just a repetition of the proofs above, so we focus on part 2. Suppose there exists a counter-example $P_{XY}$ achieving

$$I_1 = \mathbb{E}\left[\log \frac{dP_{X|Y^*}}{dP_X}(X|Y)\right] < I^* = R(D)\,.$$

Notice that whenever $I(X;Y) < \infty$ we have

$$I_1 = I(X;Y) - D(P_{X|Y}\|P_{X|Y^*}|P_Y)\,,$$

and thus

$$D(P_{X|Y}\|P_{X|Y^*}|P_Y) < \infty\,. \tag{25.3}$$

Before going to the actual proof, we describe the principal idea. For every $\lambda$ we can define a joint distribution

$$P_{X,Y_\lambda} = \lambda P_{X,Y} + (1-\lambda)P_{X,Y^*}\,.$$

Then, we can compute

$$I(X;Y_\lambda) = \mathbb{E}\left[\log \frac{P_{X|Y_\lambda}}{P_X}(X|Y_\lambda)\right] = \mathbb{E}\left[\log \frac{P_{X|Y_\lambda}}{P_{X|Y^*}}\frac{P_{X|Y^*}}{P_X}\right] \tag{25.4}$$

$$= D(P_{X|Y_\lambda}\|P_{X|Y^*}|P_{Y_\lambda}) + \mathbb{E}\left[\frac{P_{X|Y^*}(X|Y_\lambda)}{P_X}\right] \tag{25.5}$$

$$= D(P_{X|Y_\lambda}\|P_{X|Y^*}|P_{Y_\lambda}) + \lambda I_1 + (1-\lambda)I_* \,. \tag{25.6}$$

From here we will conclude, similar to Prop. 4.1, that the first term is $o(\lambda)$ and thus for sufficiently small $\lambda$ we should have $I(X;Y_\lambda) < R(D)$, contradicting optimality of coupling $P_{X,Y^*}$.

We proceed to details. For every $\lambda \in [0,1]$ define

$$\rho_1(y) \triangleq \frac{dP_Y}{dP_{Y^*}}(y) \tag{25.7}$$

$$\lambda(y) \triangleq \frac{\lambda\rho_1(y)}{\lambda\rho_1(y)+\bar\lambda} \tag{25.8}$$

$$P^{(\lambda)}_{X|Y=y} = \lambda(y)P_{X|Y=y} + \bar\lambda(y)P_{X|Y^*=y} \tag{25.9}$$

$$dP_{Y_\lambda} = \lambda dP_Y + \bar\lambda dP_{Y^*} = (\lambda\rho_1(y)+\bar\lambda)dP_{Y^*} \tag{25.10}$$

$$D(y) = D(P_{X|Y=y}\|P_{X|Y^*=y}) \tag{25.11}$$

$$D_\lambda(y) = D(P^{(\lambda)}_{X|Y=y}\|P_{X|Y^*=y})\,. \tag{25.12}$$

Notice:

$$\text{On } \{\rho_1 = 0\}: \quad \lambda(y) = D(y) = D_\lambda(y) = 0$$

and otherwise $\lambda(y) > 0$. By convexity of divergence

$$D_\lambda(y) \le \lambda(y)D(y)$$

and therefore

$$\frac{1}{\lambda(y)}D_\lambda(y)1\{\rho_1(y)>0\} \le D(y)1\{\rho_1(y)>0\}\,.$$

Notice that by (25.3) the function $\rho_1(y)D(y)$ is non-negative and $P_{Y^*}$-integrable. Then, applying dominated convergence theorem we get

$$\lim_{\lambda\to 0}\int_{\{\rho_1>0\}} dP_{Y^*}\frac{1}{\lambda(y)}D_\lambda(y)\rho_1(y) = \int_{\{\rho_1>0\}} dP_{Y^*}\rho_1(y)\lim_{\lambda\to 0}\frac{1}{\lambda(y)}D_\lambda(y) = 0 \tag{25.13}$$

where in the last step we applied the result from Lecture 4

$$D(P\|Q) < \infty \qquad \Longrightarrow \qquad D(\lambda P + \bar\lambda Q\|Q) = o(\lambda)$$

since for each $y$ on the set $\{\rho_1 > 0\}$ we have $\lambda(y) \to 0$ as $\lambda \to 0$.

On the other hand, notice that

$$\int_{\{\rho_1>0\}} dP_{Y^*}\frac{1}{\lambda(y)}D_\lambda(y)\rho_1(y)1\{\rho_1(y)>0\} = \frac{1}{\lambda}\int_{\{\rho_1>0\}} dP_{Y^*}(\lambda\rho_1(y)+\bar\lambda)D_\lambda(y) \tag{25.14}$$

$$= \frac{1}{\lambda}\int_{\{\rho_1>0\}} dP_{Y_\lambda}D_\lambda(y) \tag{25.15}$$

$$= \frac{1}{\lambda}\int_{\mathcal{Y}} dP_{Y_\lambda}D_\lambda(y) = \frac{1}{\lambda}D(P^{(\lambda)}_{X|Y}\|P_{X|Y^*}|P_{Y_\lambda})\,, \tag{25.16}$$

where in the penultimate step we used $D_\lambda(y) = 0$ on $\{\rho_1 = 0\}$. Hence, (25.13) shows

$$D(P^{(\lambda)}_{X|Y} \| P_{X|Y^*} | P_{Y_\lambda}) = o(\lambda)\,, \qquad \lambda \to 0\,.$$

Finally, since

$$P^{(\lambda)}_{X|Y} \circ P_{Y_\lambda} = P_X\,,$$

we have

$$I(X;Y_\lambda) = D(P^{(\lambda)}_{X|Y} \| P_{X|Y^*} | P_{Y_\lambda}) + \lambda\,\mathbb{E}\left[\log \frac{dP_{X|Y^*}}{dP_X}(X|Y)\right] + \bar{\lambda}\,\mathbb{E}\left[\log \frac{dP_{X|Y^*}}{dP_X}(X|Y^*)\right] \tag{25.17}$$

$$= I^* + \lambda(I_1 - I^*) + o(\lambda)\,, \tag{25.18}$$

contradicting the assumption

$$I(X;Y_\lambda) \geq I^* = R(D)\,.$$

□

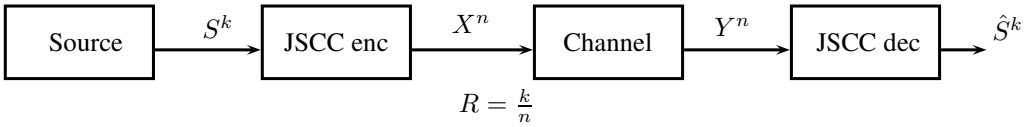
## 25.3 Lossy joint source-channel coding

The *lossy joint source channel coding problem* refers to the fundamental limits of lossy compression followed by transmission over a channel.

**Problem Setup:** For an $\mathcal{A}$-valued $(\{S_1, S_2, \dots\}$ and distortion metric $d : \mathcal{A}^k \times \hat{\mathcal{A}}^k \to \mathbb{R}$, a lossy JSCC is a pair $(f, g)$ such that

$$S^k \xrightarrow{f} X^n \xrightarrow{\text{ch.}} Y^n \xrightarrow{g} \hat{S}^k$$

**Definition 25.1.** $(f, g)$ is a $(k, n, D)$-JSCC if $\mathbb{E}[d(S^k, \hat{S}^k)] \leq D$.




where $\rho$ is the *bandwidth expansion factor*:

$$\rho = \frac{n}{k} \quad \text{channel uses/symbol.}$$

Our goal is to minimize $\rho$ subject to a fidelity guarantee by designing the encoder/decoder pairs smartly. The asymptotic fundamental limit for a lossy JSCC is

$$\rho^*(D) = \limsup_{n\to\infty} \min\{\frac{n}{k} : \exists(k, n, D) - code\}$$

For simplicity in this lecture we will focus on JSCC for stationary memoryless sources with separable distortion + stationary memoryless channels.

### 25.3.1 Converse

The converse for the JSCC is quite simple. Note that since there is no $\epsilon$ under consideration, the strong converse is the same as the weak converse. The proof architecture is identical to the weak converse of lossless JSCC which uses Fano's inequality.

**Theorem 25.3** (Converse)**.** *For any source such that*

$$R_i(D) = \lim_{k\to\infty} \frac{1}{k} \inf_{P_{\hat{S}^k|S^k}:\mathbb{E}[d(S^k,\hat{S}^k)]\le D} I(S^k;\hat{S}^k)$$

*we have*

$$\rho^*(D) \ge \frac{R_i(D)}{C_i}$$

**Remark:** The requirement of this theorem on the source isn't too stringent; the limit expression for $R_i(D)$ typically exists for stationary sources (like for the entropy rate)

*Proof.* Take a $(k, n, D)$-code $S^k \to X^n \to Y^n \to \hat{S}^k$. Then

$$\inf_{P_{\hat{S}^k|S^k}} I(S^k;\hat{S}^k) \le I(S^k;\hat{S}^k) \le I(X^k;Y^k) \le \sup_{P_{X^n}} I(X^n;Y^n)$$

Which follows from data processing and taking inf/sup. Normalizing by $1/k$ and taking the liminf as $n \to \infty$

$$\text{(LHS)} \quad \liminf_{n\to\infty} \frac{1}{n} \sup_{P_{X^n}} I(X^n;Y^n) = C_i$$

$$\text{(RHS)} \quad \liminf_{n\to\infty} \frac{1}{k_n} \inf_{P_{\hat{S}^{k_n}|S^{k_n}}} I(S^{k_n};\hat{S}^{k_n}) = R_i(D)$$

And therefore, any sequence of $(k_n, n, D)$-codes satisfies

$$\limsup_{n\to\infty} \frac{n}{k_n} \ge \frac{R_i(D)}{C_i}$$

□

**Note**: Clearly the assumptions in Theorem 25.3 are satisfied for memoryless sources. If the source $S$ is iid Bern(1/2) with Hamming distortion, then Theorem 25.3 coincides with the weak converse for channel coding under bit error rate in Theorem 14.4:

$$k \le \frac{nC}{1 - h(p_b)}$$

which we proved using ad hoc techniques. In the case of channel with cost constraints, e.g., the AWGN channel with $C(\text{SNR}) = \frac{1}{2}\log(1 + \text{SNR})$, we have

$$p_b \ge h^{-1}\left(1 - \frac{C(\text{SNR})}{R}\right)$$

This is often referred to as the Shannon limit in plots comparing the bit-error rate of practical codes. See, e.g., Fig. 2 from [RSU01] for BIAWGN (binary-input) channel. *This is erroneous*, since the $p_b$ above refers to the bit-error of data bits (or systematic bits), not all of the codeword bits. The latter quantity is what typically called BER in the coding-theoretic literature.

### 25.3.2 Achievability via separation

The proof strategy is similar to the lossless JSCC: We construct a separated lossy compression and channel coding scheme using our tools from those areas, i.e., let the JSCC encoder to be the concatenation of a loss compressor and a channel encoder, and the JSCC decoder to be the concatenation of a channel decoder followed by a loss compressor, then show that this separated construction is optimal.

**Theorem 25.4.** *For any stationary memoryless source $(P_S, \mathcal{A}, \hat{\mathcal{A}}, d)$ satisfying assumption A1 (below), and for any stationary memoryless channel $P_{Y|X}$,*

$$\rho^*(D) = \frac{R(D)}{C}$$

**Note**: The assumption on the source is to control the distortion incurred by the channel decoder making an error. Although we know that this is a low-probability event, without any assumption on the distortion metric, we cannot say much about its contribution to the end-to-end average distortion. This will not be a problem if the distortion metric is bounded (for which Assumption A1 is satisfied of course). Note that we do not have this nuisance in the lossless JSCC because we at most suffer the channel error probability (union bound).

The assumption is rather technical which can be skipped in the first reading. Note that it is trivially satisfied by bounded distortion (e.g., Hamming), and can be shown to hold for Gaussian source and MSE distortion.

*Proof.* The converse direction follows from the previous theorem. For the other direction, we constructed a separated compression / channel coding scheme. Take

$$S^k \to W \to \hat{S}^k \text{ compressor to } W \in [2^{kR(D)+o(k)}] \text{ with } \mathbb{E}[d(S^k, \hat{S}^k)] \le D$$
$$W \to X^n \to Y^n \to \hat{W} \text{ maximal probability of error channel code (assuming } kR(D) \le nC + o(n))$$
$$\text{with } \mathbb{P}[W \neq \hat{W}] \le \epsilon \ \forall P_W$$

So that the overall system is

$$S^k \longrightarrow W \longrightarrow X^n \longrightarrow Y^n \longrightarrow \hat{W} \longrightarrow \hat{S}^k$$

Note that here we need a **maximum** probability of error code since when we concatenate these two schemes, $W$ at the input of the channel is the output of the source compressor, which is not guaranteed to be uniform. Now that we have a scheme, we must analyze the average distortion to show that it meets the end-to-end distortion constraint. We start by splitting the expression into two cases

$$\mathbb{E}[d(S^k, \hat{S}^k)] = \mathbb{E}[d(S^k, \hat{S}^k(W))\mathbf{1}\{W = \hat{W}\}] + \mathbb{E}[d(S^k, \hat{S}^k(\hat{W}))\mathbf{1}\{W \neq \hat{W}\}]$$

By assumption on our lossy code, we know that the first term is $\le D$. In the second term, we know that the probability of the event $\{W \neq \hat{W}\}$ is small by assumption on our channel code, but we cannot say anything about $\mathbb{E}[d(S^k, \hat{S}^k(\hat{W}))]$ unless, for example, $d$ is bounded. But by Lemma 25.1 (below), $\exists$ code $S^k \to W \to \hat{S}^k$ such that

(1) $\mathbb{E}[d(S^k, \hat{S}^k)] \le D$ holds

(2) $d(a_0^k, \hat{S}^k) \le L$ for all quantization outputs $\hat{S}^k$, where $a_0^k = (a_0, \ldots, a_0)$ is some fixed string of length $k$ from the Assumption A1 below.

The second bullet says that all points in the reconstruction space are "close" to some fixed string. Now we can deal with the troublesome term

$$\begin{aligned}\mathbb{E}[d(S^k, \hat{S}^k(\hat{W}))\mathbf{1}\{W \neq \hat{W}\}] &\leq \mathbb{E}[\mathbf{1}\{W \neq \hat{W}\}\lambda(d(S^k, \hat{a}_0^k) + d(a_0^k, \hat{S}^k))] \\ \text{(by point (2) above)} \quad &\leq \lambda\mathbb{E}[\mathbf{1}\{W \neq \hat{W}\}d(S^k, \hat{a}_0^k)] + \lambda\mathbb{E}[\mathbf{1}\{W \neq \hat{W}\}L] \\ &\leq \lambda o(1) + \lambda L\epsilon \to 0 \text{ as } \epsilon \to 0\end{aligned}$$

where in the last step we applied the same uniform integrability argument that showed vanishing of the expectation in (24.20) before.

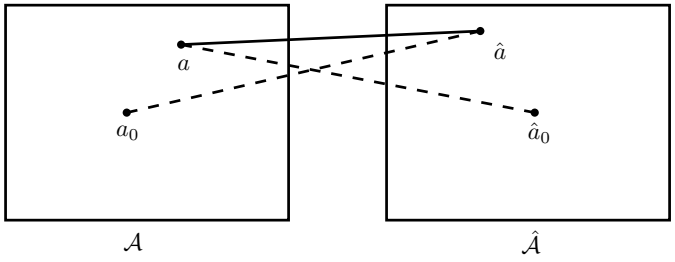
In all, our scheme meets the average distortion constraint. Hence we conclude that for $\forall \rho > \frac{R(D)}{C}$, $\exists$ sequence of $(k, n, D + o(1))$-codes. □

The following assumption is critical to the previous theorem:

**Assumption A1:** For a source $(P_S, \mathcal{A}, \hat{\mathcal{A}}, d)$, $\exists \lambda \geq 0, a_0 \in \mathcal{A}, \hat{a}_0 \in \hat{\mathcal{A}}$ such that

1. $d(a, \hat{a}) \leq \lambda(d(a, \hat{a}_o) + d(a_0, \hat{a}))\ \ \forall a, \hat{a}$ (generalized triangle inequality)
2. $\mathbb{E}[d(S, \hat{a}_0)] < \infty$ (so that $D_{\max} < \infty$ too).
3. $\mathbb{E}[d(a_0, \hat{S})] < \infty$ for any output distribution $P_{\hat{S}}$ achieving the rate-distortion function $R(D)$ at some $D$.
4. $d(a_0, \hat{a}_0) < \infty$.

This assumption says that the spaces $\mathcal{A}$ and $\hat{\mathcal{A}}$ have "nice centers", in the sense that the distance between any two points is upper bounded by a constant times the distance from the centers to each point (see figure below).




But the assumption isn't easy to verify, or clear which sources satisfy the assumptions. Because of this, we now give a few sufficient conditions for Assumption A1 to hold.

**Trivial Condition:** If the distortion function is bounded, then the assumption A1 holds automatically. In other words, if we have a discrete source with finite alphabet $|\mathcal{A}|, |\hat{\mathcal{A}}| < \infty$ and a finite distortion function $d(a, \hat{a}) < \infty$, then A1 holds. More generally, we have the following criterion.

**Theorem 25.5** (Criterion for satisfying A1)**.** *If $\mathcal{A} = \hat{\mathcal{A}}$ and $d(a, \hat{a}) = \rho^q(a, \hat{a})$ for some metric $\rho$ with $q \geq 1$, and $D_{\max} \triangleq \inf_{\hat{a}_0} \mathbb{E}[d(S, \hat{a}_0)] < \infty$, then A1 holds.*

*Proof.* Take $a_0 = \hat{a}_0$ that achieves finite $D_p$ (in fact, any points can serve as centers in a metric space). Then

$$\begin{aligned}(\frac{1}{2}\rho(a, \hat{a}))^q &\leq \left(\frac{1}{2}\rho(a, a_0) + \frac{1}{2}\rho(a_0, \hat{a})\right)^q \\ \text{(Jensen's)} \quad &\leq \frac{1}{2}\rho^q(a, a_0) + \frac{1}{2}\rho^q(a_0, \hat{a})\end{aligned}$$

And thus $d(a,\hat{a}) \le 2^{q-1}(d(a,a_0)+d(a_0,\hat{a}))$. Taking $\lambda = 2^{q-1}$ verifies (1) and (2) in A1. To verify (3), we can use this generalized triangle inequality for our source

$$d(a_0,\hat{S}) \le 2^{q-1}(d(a_0,S)+d(S,\hat{S}))$$

Then taking the expectation of both sides gives

$$\begin{aligned}\mathbb{E}[d(a_0,\hat{S})] &\le 2^{q-1}(\mathbb{E}[d(a_0,S)]+\mathbb{E}[d(S,\hat{S})]) \\ &\le 2^{q-1}(D_{\max}+D) < \infty\end{aligned}$$

So that condition (3) in A1 holds. □

So we see that metrics raised to powers (e.g. squared Euclidean norm) satisfy the condition A1. The lemma used in Theorem 25.4 is now given.

**Lemma 25.1.** *Fix a source satisfying A1 and an arbitrary $P_{\hat{S}|S}$. Let $R > I(S;\hat{S})$, $L > \max\{\mathbb{E}[d(a_0,\hat{S})], d(a_0,\hat{a}_0)\}$ and $D > \mathbb{E}[d(S,\hat{S})]$. Then, there exists a $(k, 2^{kR}, D)$-code such that for every reconstruction point $\hat{x} \in \hat{A}^k$ we have $d(a_0^k,\hat{x}) \le L$.*

*Proof.* Let $\mathcal{X} = \mathcal{A}^k$, $\hat{\mathcal{X}} = \hat{\mathcal{A}}^k$ and $P_X = P_S^k, P_{Y|X} = P_{\hat{S}|S}^k$. Then apply the achievability bound for excess distortion from Theorem 24.4 with

$$d_1(x,\hat{x}) = \begin{cases} d(x,\hat{x}) & d(a_0^k,\hat{x}) \le L \\ +\infty & \text{o/w} \end{cases}$$

Note that this is NOT a separable distortion metric. Also note that without any change in $d_1$-distortion we can remove all (if any) reconstruction points $\hat{x}$ with $d(a_0^k,\hat{x}) > L$. Furthermore, from the WLLN we have for any $D > D' > \mathbb{E}[d(S,\hat{S}')]$

$$\mathbb{P}[d_1(X,Y) > D'] \le \mathbb{P}[d(S^k,\hat{S}^k) > D'] + \mathbb{P}[d(a_0^k,\hat{S}^k) > L] \to 0$$

as $k \to \infty$ (since $\mathbb{E}[d(S,\hat{S})] < D'$ and $\mathbb{E}[a_0,\hat{S}] < L$). Thus, overall we get $M = 2^{kR}$ reconstruction points $(c_1,\dots,c_M)$ such that

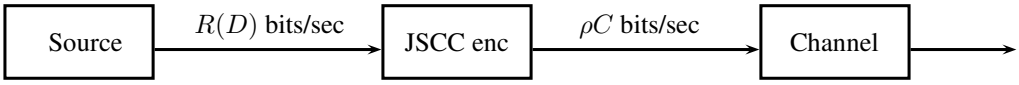
$$\mathbb{P}[\min_{j\in[M]} d(S^k,c_j) > D'] \to 0$$

and $d(a_0^k,c_j) \le L$. By adding $c_{M+1} = (\hat{a}_0,\dots,\hat{a}_0)$ we get

$$\mathbb{E}[\min_{j\in[M+1]} d(S^k,c_j)] \le D' + \mathbb{E}[d(S^k,c_{M+1})1\{\min_{j\in[M]} d(S^k,c_j) > D'\}] = D' + o(1)\,,$$

where the last estimate follows from uniform integrability as in the vanishing of expectation in (24.20). Thus, for sufficiently large $n$ the expected distortion is $\le D$, as required. □

To summarize the results in this section, under stationarity and memorylessness assumptions on the source and the channel, we have shown that the following separately-designed scheme achieves the optimal rate for lossy JSCC: First compress the data, then encode it using your favorite channel code, then decompress at the receiver.



## 25.4 What is lacking in classical lossy compression?

Examples of some issues with the classical compression theory:

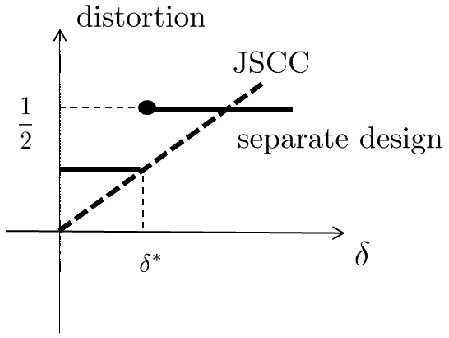
- compression: we can apply the standard results in compression of a text file, but it is extremely difficult for image files due to the strong spatial correlation. For example, the first sentence and the last in Tolstoy's novel are pretty uncorrelated. But the regions in the upper-left and bottom-right corners of one image can be strongly correlated. Thus for practicing the lossy compression of videos and images the key problem is that of coming up with a good "whitening" basis.

- JSCC: Asymptotically the separation principle sounds good, but the separated systems can be very unstable - no graceful degradation. Consider the following example of JSCC.

  **Example**: Source = $Bern(\frac{1}{2})$, channel = $BSC(\delta)$.

  1. separate compressor and channel encoder designed for $\frac{R(D)}{C(\delta)} = 1$
  2. a simple JSCC:

$$\rho = 1, X_i = S_i$$




no graceful degradation of separately designed source channel code

# 6.441 Information Theory

Spring 2016