24.1 Recap

Compute $R(D)$.

Recall from the last lecture:

$$R(D) = \limsup_{n\to\infty} \frac{1}{n} \log M^*(n, D), \qquad \text{(rate distortion function)}$$
$$R_i(D) = \limsup_{n\to\infty} \frac{1}{n} \varphi_{S^n}(D), \qquad \text{(information rate distortion function)}$$

and

$$\varphi_S(D) \triangleq \inf_{P_{\hat{S}|S}:\mathbb{E}[d(S,\hat{S})]\le D} I(S;\hat{S})$$
$$\varphi_{S^n}(D) = \inf_{P_{\hat{S}^n|S^n}:\mathbb{E}[d(S^n,\hat{S}^n)]\le D} I(S^n;\hat{S}^n)$$

Also, we showed the general converse: For any (M, D)-code $X \to W \to \hat{X}$ we have

$$\begin{aligned} & \log M \ge \varphi_X(D) \\ \Longrightarrow\ & \log M^*(n, D) \ge \varphi_{S^n}(D) \\ \Longrightarrow\ & R(D) \ge R_i(D) \end{aligned}$$

In this lecture, we will prove the achievability bound and establish the identity $R(D) = R_i(D)$ for stationary memoryless sources.

First we show that $R_i(D)$ can be easily calculated for memoryless source without going through the multi-letter optimization problem.

Theorem 24.1 (Single-letterization)**.** *For stationary memoryless source S^n and separable distortion d,*

$$R_i(D) = \varphi_S(D)$$

Proof. By definition we have that $\varphi_{S^n}(D) \le n\varphi_S(D)$ by choosing a product channel: $P_{\hat{S}^n|S^n} = (P_{\hat{S}|S})^n$. Thus $R_i(D) \le \varphi_S(D)$.

For the converse, take any $P_{\hat{S}^n|S^n}$ such that the constraint $\mathbb{E}[d(S^n, \hat{S}^n)] \le D$ is satisfied, we have

$$\begin{aligned} I(S^n; \hat{S}^n) &\ge \sum_{j=1}^n I(S_j, \hat{S}_j) && (S^n \text{ independent}) \\ &\ge \sum_{j=1}^n \varphi_S(\mathbb{E}[d(S_j, \hat{S}_j)]) \\ &\ge n\varphi_S\left(\frac{1}{n}\sum_{j=1}^n \mathbb{E}[d(S_j, \hat{S}_j)]\right) && (\text{convexity of } \varphi_S) \\ &\ge n\varphi_S(D) && (\varphi_S \text{ non-increasing}) \end{aligned}$$

□

24.2 Shannon's rate-distortion theorem

Theorem 24.2. *Let the source S^n be stationary and memoryless, $S^n \overset{i.i.d.}{\sim} P_S$, and suppose that distortion metric d and the target distortion D satisfy:*

1. *$d(s^n, \hat{s}^n)$ is non-negative and separable*
2. *$D > D_0$*
3. *$D_{\max}$ is finite, i.e.*
$$D_{\max} \triangleq \inf_{\hat{s}} \mathbb{E}[d(S, \hat{s})] < \infty.$$

Then

$$R(D) = R_i(D) = \inf_{P_{\hat{S}|S}:\mathbb{E}[d(S,\hat{S})]\le D} I(S; \hat{S}). \tag{24.1}$$

Remarks:

- Note that $D_{\max} < \infty$ does not imply that $d(\cdot,\cdot)$ only takes values in $\mathbb{R}$, i.e. theorem permits $d(a, \hat{a}) = \infty$.
- It should be remarked that when $D_{\max} = \infty$ typically $R(D) = \infty$. Indeed, suppose that $d(\cdot,\cdot)$ is a metric (i.e. finite valued and satisfies triangle inequality). Then, for any $x_0 \in \mathcal{A}^n$ we have
$$d(X, \hat{X}) \ge d(X, x_0) - d(x_0, \hat{X}).$$
Thus, for any finite codebook $\{c_1, \ldots, c_M\}$ we have $\max_j d(x_0, c_j) < \infty$ and therefore
$$\mathbb{E}[d(X, \hat{X})] \ge \mathbb{E}[d(X, x_0)] - \max_j d(x_0, c_j) = \infty.$$
So that $R(D) = \infty$ for any finite D. This observation, however, should not be interpreted as absolute impossibility of compression for such sources. It is just not possible with fixed-rate codes. As an example, for quadratic distortion and Cauchy-distributed S, $D_{\max} = \infty$ since S has infinite second-order moments. But it is easy to see that $R_i(D) < \infty$ for any $D \in (0, \infty)$. In fact, in this case $R_i(D)$ is a hyperbola-like curve that never touches either axis. A non-trivial compression can be attained with compressors $S^n \to W$ of bounded entropy $H(W)$ (but unbounded alphabet of W). Indeed if we take W to be a Δ-quantized version of S and notice that differential entropy of S is finite, we get from (23.2) that $R_i(\Delta) \le H(W) < \infty$. Interesting question: Is $H(W) = nR_i(D) + o(n)$ attainable?

- Techniques in proving (24.1) for memoryless sources can be applied to prove it for "stationary ergodic" sources with changes similar to those we have discussed in channel coding.

Before giving a formal proof, we illustrate the intuition non-rigorously.

24.2.1 Intuition

Try to throw in M points $\mathcal{C} = \{c_1, \ldots, c_M\} \in \hat{\mathcal{A}}^n$ which are drawn i.i.d. according to a product distribution $Q_{\hat{S}}^n$ where $Q_{\hat{S}}$ is some distribution on $\hat{\mathcal{A}}$. Examine the simple encoder and decoder pair:

$$\text{encoder}: \ f(s^n) = \operatorname*{argmin}_{j \in [M]} d(s^n, c_j) \tag{24.2}$$

$$\text{decoder}: \ g(j) = c_j \tag{24.3}$$

The basic idea is the following: Since the codewords are generated independently of the source, the probability that a given codeword offers good reconstruction is (exponentially) small, say, ϵ. However, since we have many codewords, the chance that there exists a good one can be of high probability. More precisely, the probability that no good codeword exist is $(1-\epsilon)^M$, which can be very close to zero as long as M grows faster than $\frac{1}{\epsilon}$.

To explain the intuition further, let us consider the excess distortion of this code: $\mathbb{P}[d(S^n, \hat{S}^n) > D]$. Define

$$P_{success} \triangleq \mathbb{P}[\exists c \in \mathcal{C}, \text{ s.t. } d(S^n, c) \le D]$$

Then

$$\begin{aligned}
P_{failure} \triangleq & \mathbb{P}[\forall c_i \in \mathcal{C}, d(S^n, c) > D] && (24.4)\\
\approx & \mathbb{P}[\forall c_i \in \mathcal{C}, d(S^n, c) > D | S^n \in T_n] && (24.5)\\
& (\ T_n \text{ is the set of typical strings with empirical distribution } \hat{P}_{S^n} \approx P_S\ ) \\
= & \mathbb{P}[d(S^n, \hat{S}^n) > D | S^n \in T_n]^M \qquad (P_{S^n, \hat{S}^n} = P_S^n Q_{\hat{S}}^n) && (24.6)\\
= & (1 - \underbrace{\mathbb{P}[d(S^n, \hat{S}^n) \le D | S^n \in T_n]}_{\text{since } S^n \perp\!\!\!\perp \hat{S}^n, \text{ this should be small}})^M && (24.7)\\
\approx & (1 - 2^{-nE(Q_{\hat{S}})})^M \qquad \text{(large deviation!)} && (24.8)
\end{aligned}$$

where it can be shown (similar to information projection) that

$$E(Q_{\hat{S}}) = \min_{P_{\hat{S}|S}: \mathbb{E}[d(S,\hat{S})] \le D} D(P_{\hat{S}|S} \| Q_{\hat{S}} | P_S) \tag{24.9}$$

Thus we conclude that $\forall Q_{\hat{S}}, \forall \delta > 0$ we can pick $M = 2^{n(E(Q_{\hat{S}})+\delta)}$ and the above code will have arbitrarily small excess distortion:

$$P_{failure} = \mathbb{P}[\forall c \in \mathcal{C}, d(S^n, c) > D] \to 0 \text{ as } n \to \infty.$$

We optimize $Q_{\hat{S}}$ to get the smallest possible M:

$$\begin{aligned}
\min_{Q_{\hat{S}}} E(Q_{\hat{S}}) &= \min_{P_{\hat{S}|S}: \mathbb{E}[d(S,\hat{S})] \le D} \min_{Q_{\hat{S}}} D(P_{\hat{S}|S} \| Q_{\hat{S}} | P_S) \qquad (24.10)\\
&= \min_{P_{\hat{S}|S}: \mathbb{E}[d(S,\hat{S})] \le D} I(S; \hat{S}) \\
&= \varphi_S(D)
\end{aligned}$$

24.2.2 Proof of Theorem 24.2

Theorem 24.3 (Performance bound of average-distortion codes)**.** *Fix P_X and suppose $d(x,\hat{x}) \geq 0$ for all $x, \hat{x}$. $\forall P_{Y|X}$, $\forall \gamma > 0$, $\forall y_0 \in \hat{\mathcal{A}}$, there exists a code $X \to W \to \hat{X}$, where $W \in [M+1]$ and*

$$\mathbb{E}[d(X,\hat{X})] \leq \mathbb{E}[d(X,Y)] + \mathbb{E}[d(X,y_0)]e^{-M/\gamma} + \mathbb{E}[d(X,y_0)\mathbf{1}_{\{i(X;Y)>\log\gamma\}}]$$
$$d(X,\hat{X}) \leq d(X,y_0) \quad a.s.$$

Notes:

- This theorem says that from an arbitrary $P_{Y|X}$ such that $\mathbb{E}d(X,Y) \leq D$, we can extract a good code with average distortion D plus some extra terms which will vanish in the asymptotic regime.
- The proof uses the random coding argument. The role of the deterministic y_0 is a "fail-safe" codeword (think of y_0 as the default reconstruction with $D_{\max} = \mathbb{E}[d(X,y_0)]$). We add y_0 to the random codebook for damage control, to hedge the (highly unlikely and unlucky) event that we end up with a horrible codebook.

Proof. Similar to the previous intuitive argument, we apply random coding and generate the codewords randomly and independently of the source:

$$\mathcal{C} = \{c_1, \ldots, c_M\} \overset{\text{i.i.d.}}{\sim} P_Y \perp\!\!\!\perp X$$

and add the "fail-safe" codeword $c_{M+1} = y_0$. We adopt the same encoder-decoder pair (24.2) – (24.3) and let $\hat{X} = g(f(X))$. Then by definition,

$$d(X,\hat{X}) = \min_{j\in[M+1]} d(X,c_j) \leq d(X,y_0).$$

To simplify notation, let $\overline{Y}$ be an independent copy of Y (similar to the idea of introducing unsent codeword $\overline{X}$ in channel coding):

$$P_{X,Y,\overline{Y}} = P_{X,Y}P_{\overline{Y}}$$

where $P_{\overline{Y}} = P_Y$. Recall the formula for computing the expectation of a random variable $U \in [0,a]$: $\mathbb{E}[U] = \int_0^a \mathbb{P}[U \geq u]du$. Then the average distortion is

$$\begin{aligned}
\mathbb{E}d(X,\hat{X}) &= \mathbb{E}\min_{j\in[M+1]} d(X,c_j) && (24.11)\\
&= \mathbb{E}_X\mathbb{E}\Big[\min_{j\in[M+1]} d(X,c_j)\Big|X\Big] && (24.12)\\
&= \mathbb{E}_X\int_0^{d(X,y_0)} \mathbb{P}\Big[\min_{j\in[M+1]} d(X,c_j) > u\Big|X\Big]du && (24.13)\\
&\leq \mathbb{E}_X\int_0^{d(X,y_0)} \mathbb{P}\Big[\min_{j\in[M]} d(X,c_j) > u\Big|X\Big]du && (24.14)\\
&= \mathbb{E}_X\int_0^{d(X,y_0)} \mathbb{P}[d(X,\overline{Y}) > u|X]^M du && (24.15)\\
&= \mathbb{E}_X\int_0^{d(X,y_0)} (1-\underbrace{\mathbb{P}[d(X,\overline{Y}) \leq u|X]}_{\triangleq\delta(X,u)})^M du && (24.16)
\end{aligned}$$

Next we upper bound $(1-\delta(X,u))^M$ as follows:

$$\begin{aligned}(1-\delta(X,u))^M &\le e^{-M/\gamma} + |1-\gamma\delta(X,u)|^+ && (24.17)\\ &= e^{-M/\gamma} + \left|1-\gamma\mathbb{E}[\exp\{-i(X;Y)\}\mathbf{1}_{\{d(X,Y)\le u\}}|X]\right|^+ && (24.18)\\ &\le e^{-M/\gamma} + \mathbb{P}[i(X;Y)>\log\gamma|X] + \mathbb{P}[d(X,Y)>u|X] && (24.19)\end{aligned}$$

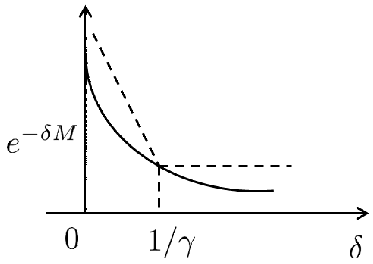
where

- (24.17) uses the following trick in dealing with $(1-\delta)^M$ for $\delta \ll 1$ and $M \gg 1$. First, recall the standard rule of thumb:

$$(1-\epsilon_n)^n \approx \begin{cases}0, & \epsilon_n n \gg 1\\ 1, & \epsilon_n n \ll 1\end{cases}$$

  In order to argue firm bounds of similar flavor, consider

$$\begin{aligned}1-\delta M \overset{\text{union bound}}{\le} (1-\delta)^M &\le e^{-\delta M} && (\log(1-\delta)\le -\delta)\\ &\le e^{-M/\gamma}(\gamma\delta\wedge 1) + |1-\gamma\delta|^+ && (\forall\gamma>0)\\ &\le e^{-M/\gamma} + |1-\gamma\delta|^+\end{aligned}$$



upper bound $e^{-\delta M}$

- (24.18) is simply change of measure using $i(x;y) = \log\frac{P_Y(y)}{P_{Y|X}(y|x)}$ (i.e., conditioning-unconditioning trick for information density, cf. Proposition 15.1.

- (24.19):

$$\begin{aligned}1-\gamma\mathbb{E}[\exp\{-i(X;Y)\}\mathbf{1}_{\{d(X,Y)\le u\}}|X] &\le 1-\gamma\mathbb{E}[\exp\{-i(X;Y)\}\mathbf{1}_{\{d(X,Y)\le u, i(X;Y)\le\log\gamma\}}|X]\\ &\le 1-\mathbb{E}[\mathbf{1}_{\{d(X,Y)\le u, i(X;Y)\le\log\gamma\}}|X]\\ &= \mathbb{P}[d(X,Y)>u \text{ or } i(X;Y)>\log\gamma|X]\\ &\le \mathbb{P}[d(X,Y)>u|X] + \mathbb{P}[i(X;Y)>\log\gamma|X]\end{aligned}$$

Plugging (24.19) into (24.16), we have

$$\begin{aligned}\mathbb{E}[d(X,\hat{X})] &\le \mathbb{E}_X\int_0^{d(X,y_0)}(e^{-M/\gamma} + \mathbb{P}[i(X;Y)>\log\gamma|X] + \mathbb{P}[d(X,Y)>u|X])du\\ &\le \mathbb{E}[d(X,y_0)]e^{-M/\gamma} + \mathbb{E}[d(X,y_0)\mathbb{P}[i(X;Y)>\log\gamma|X]] + \mathbb{E}_X\int_0^\infty \mathbb{P}[d(X,Y)>u|X])du\\ &= \mathbb{E}[d(X,y_0)]e^{-M/\gamma} + \mathbb{E}[d(X,y_0)\mathbf{1}_{\{i(X;Y)>\log\gamma\}}] + \mathbb{E}[d(X,Y)]\end{aligned}$$

□

As a side product, we have the following achievability for excess distortion.

Theorem 24.4 (Performance bound of excess-distortion codes)**.** *$\forall P_{Y|X}$, $\forall \gamma > 0$, there exists a code $X \to W \to \hat{X}$, where $W \in [M]$ and*

$$\mathbb{P}[d(X,\hat{X}) > D] \le e^{-M/\gamma} + \mathbb{P}[\{d(X,Y) > D\} \cup \{i(X;Y) > \log\gamma\}]$$

Proof. Proceed exactly as in the proof of Theorem 24.3, replace (24.11) by $\mathbb{P}[d(X,\hat{X}) > D] = \mathbb{P}[\forall j \in [M], d(X,c_j) > D] = \mathbb{E}_X[(1-\mathbb{P}[d(X,\overline{Y}) \le D|X])^M]$, and continue similarly. □

Finally, we are able to prove Theorem 24.2 rigorously by applying Theorem 24.3 to iid sources $X = S^n$ and $n \to \infty$:

Proof of Theorem 24.2. Our goal is the achievability: $R(D) \le R_i(D) = \varphi_S(D)$.

WLOG we can assume that $D_{\max} = \mathbb{E}[d(S,\hat{s}_0)]$ achieved at some fixed $\hat{s}_0$ – this is our default reconstruction; otherwise just take any other fixed sequence so that the expectation is finite. The default reconstruction for S^n is $\hat{s}_0^n = (\hat{s}_0,\dots,\hat{s}_0)$ and $\mathbb{E}[d(S^n,\hat{s}_0^n)] = D_{\max} < \infty$ since the distortion is separable.

Fix some small $\delta > 0$. Take any $P_{\hat{S}|S}$ such that $\mathbb{E}[d(S,\hat{S})] \le D - \delta$. Apply Theorem 24.3 to $(X,Y) = (S^n,\hat{S}^n)$ with

$$\begin{aligned}
P_X &= P_{S^n}\\
P_{Y|X} &= P_{\hat{S}^n|S^n} = (P_{\hat{S}|S})^n\\
\log M &= n(I(S;\hat{S}) + 2\delta)\\
\log\gamma &= n(I(S;\hat{S}) + \delta)\\
d(X,Y) &= \frac{1}{n}\sum_{j=1}^n d(S_j,\hat{S}_j)\\
y_0 &= \hat{s}_0^n
\end{aligned}$$

we conclude that there exists a compressor $f : \mathcal{A}^n \to [M+1]$ and $g : [M+1] \to \hat{\mathcal{A}}^n$, such that

$$\begin{aligned}
\mathbb{E}[d(S^n,g(f(S^n)))] &\le \mathbb{E}[d(S^n,\hat{S}^n)] + \mathbb{E}[d(S^n,\hat{s}_0^n)]e^{-M/\gamma} + \mathbb{E}[d(S^n,\hat{s}_0^n)\mathbf{1}_{\{i(S^n;\hat{S}^n)>\log\gamma\}}]\\
&\le D - \delta + \underbrace{D_{\max}e^{-\exp(n\delta)}}_{\to 0} + \underbrace{\mathbb{E}[d(S^n,\hat{s}_0^n)\mathbf{1}_{E_n}]}_{\to 0 \text{ (later)}},
\end{aligned} \tag{24.20}$$

where

$$E_n = \{i(S^n;\hat{S}^n) > \log\gamma\} = \left\{\frac{1}{n}\sum_{j=1}^n i(S_j;\hat{S}_j) > I(S;\hat{S}) + \delta\right\} \qquad \overset{\text{WLLN}}{\Longrightarrow} \quad \mathbb{P}[E_n] \to 0$$

If we can show the expectation in (24.20) vanishes, then there exists an $(n,M,\overline{D})$-code with:

$$M = 2^{n(I(S;\hat{S})+2\delta)}, \quad \overline{D} = D - \delta + o(1) \le D.$$

To summarize, $\forall P_{\hat{S}|S}$ such that $\mathbb{E}[d(S,\hat{S})] \le D - \delta$ we have that:

$$\begin{aligned}
&R(D) \le I(S;\hat{S})\\
\overset{\delta\downarrow 0}{\Longrightarrow}\ &R(D) \le \varphi_S(D-) = \varphi_S(D). \quad \text{(continuity, since } D > D_0\text{)}
\end{aligned}$$

It remains to show the expectation in (24.20) vanishes. This is a simple consequence of the uniform integrability of the sequence $\{d(S^n, \hat{s}_0^n)\}$. (Indeed, any sequence $V_n \xrightarrow{L_1} V$ is uniformly integrable.) If you do not know what uniform integrability is, here is a self-contained proof.

Lemma 24.1. *For any positive random variable U, define $g(\delta) = \sup_{H:\mathbb{P}[H]\le\delta} \mathbb{E}[U\mathbf{1}_H]$. Then*[1] $\mathbb{E}U < \infty \Rightarrow g(\delta) \xrightarrow{\delta\to 0} 0$.

Proof. For any $b > 0$, $\mathbb{E}[U\mathbf{1}_H] \le \mathbb{E}[U\mathbf{1}_{\{U>b\}}] + b\delta$, where $\mathbb{E}[U\mathbf{1}_{\{U>b\}}] \xrightarrow{b\to\infty} 0$ by dominated convergence theorem. Then the proof is completed by setting $b = 1/\sqrt{\delta}$. □

Now $d(S^n, \hat{s}_0^n) = \frac{1}{n}\sum U_j$, where U_j are iid copies of U. Since $\mathbb{E}[U] = D_{\max} < \infty$ by assumption, applying Lemma 24.1 yields $\mathbb{E}[d(S^n, \hat{s}_0^n)\mathbf{1}_{E_n}] = \frac{1}{n}\sum \mathbb{E}[U_j\mathbf{1}_{E_n}] \le g(\mathbb{P}[E_n]) \to 0$, since $\mathbb{P}[E_n] \to 0$. We are done proving the theorem. □

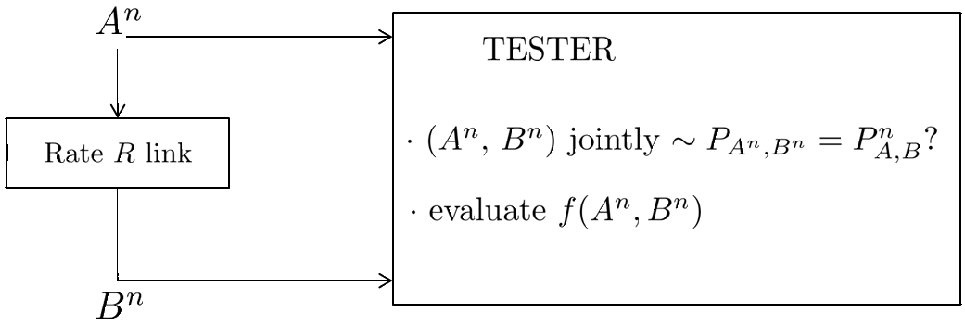
Note: It seems that in Section 24.2.1 and in Theorem 24.2 we applied different relaxations in showing the lower bound, how come they turn out to yield the same *tight* asymptotic result?

This is because the key to both proofs is to estimate the exponent (large deviations) of the underlined probabilities in (24.7) and (24.16), respectively. To get the right exponent, as we know, the key is to apply tilting (change of measure) to the distribution solving the information projection problem (24.9). In the case, when $P_Y = (Q_{\hat{S}})^n = (P_{\hat{S}})^n$ is chosen as the solution to rate-distortion optimization $\inf I(S;\hat{S})$, the resulting tilting is precisely given by $2^{-i(X;Y)}$.

24.3* Covering lemma

Goal:

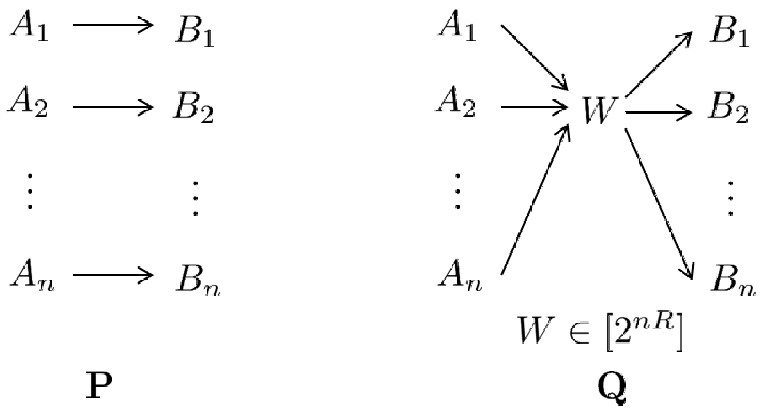
i.i.d. $\sim P_A^n$ generated by nature



What's the minimum rate R needed to fool the tester?

In other words:

[1] In fact, $\Rightarrow$ is $\Leftrightarrow$.



Approximate P with Q such that for any function f, $\forall x$, we have:

$$\mathbb{P}[f(A^n, B^n) \le x] \approx \mathbb{Q}[f(A^n, B^n) \le x], \quad |W| \le 2^{nR}.$$

what is the minimum rate R to achieve this?

Some remarks:

1. The minimal rate will depend (although it is not obvious) on whether the encoder $A^n \to W$ knows about the test that the tester is running (or equivalently whether he knows the function $f(\cdot,\cdot)$).
2. If the function is known to be of the form $f(A^n, B^n) = \sum_{j=1}^n f_1(A_j, B_j)$, then evidently the job of the encoder is the following: For any realization of the sequence A^n, we need to generate a sequence B^n such that joint composition (empirical distribution) is very close to $P_{A,B}$.
3. If $R = H(A)$, we can compress A^n and send it to "B side", who can reconstruct A^n perfectly and use that information to produce B^n through $P_{B^n|A^n}$.
4. If $R = H(B)$, "A side" can generate B^n according to $P_{A,B}^n$ and send that B^n sequence to the "B side".
5. If $A \perp\!\!\!\perp B$, we know that $R = 0$, as "B side" can generate B^n independently.

Our previous argument turns out to give a sharp answer for the case when encoder is aware of the tester's algorithm. Here is a precise result:

Theorem 24.5 (Covering Lemma)**.** *$\forall P_{A,B}$ and $R > I(A;B)$, let $\mathcal{C} = \{c_1, \ldots, c_M\}$ where each codeword c_j is i.i.d. drawn from distribution P_B^n. $\forall \epsilon > 0$, for $M \ge 2^{n(I(A;B)+\epsilon)}$ we have that:*

$$\mathbb{P}[\exists c \in \mathcal{C} \text{ such that } \hat{P}_{A^n,c} \approx P_{A,B}] \to 1$$

Stronger form: $\forall F$

$$\mathbb{P}[\exists c : (A^n, c) \in F] \ge \mathbb{P}[(A^n, B^n) \in F] + \underbrace{o(1)}_{\text{uniform in } F}$$

Proof. Following similar arguments of the proof for Theorem 24.3, we have

$$\begin{aligned}\mathbb{P}[\forall c \in \mathcal{C} : (A^n, c) \notin F] &\le e^{-\gamma} + \mathbb{P}[\{(A^n, B^n) \notin F\} \cup \{i(A^n; B^n) > \log \gamma\}] \\ &= \mathbb{P}[(A^n, B^n) \notin F] + o(1) \\ \Rightarrow \mathbb{P}[\forall c \in \mathcal{C} : (A^n, c) \in F] &\ge \mathbb{P}[(A^n, B^n) \in F] + o(1)\end{aligned}$$

□

Note: [Intuition] To generate B^n, there are around $2^{nH(B)}$ high probability sequences; for each A^n sequence, there are around $2^{nH(B|A)}$ B^n sequences that have the same joint distribution, therefore, it is sufficient to describe the class of B^n for each A^n sequence, and there are around $\frac{2^{nH(B)}}{2^{nH(B|A)}} = 2^{nI(A;B)}$ classes.

Although Covering Lemma is a powerful tool, it does not imply that the constructed joint distribution $Q_{A^nB^n}$ can fool any permutation invariant tester. In other words, it is not guaranteed that

$$\sup_{F \subset \mathcal{A}^n \times \mathcal{B}^n, \text{permut.invar.}} |Q_{A^n,B^n}(F) - P_{A,B}^n(F)| \to 0\,.$$

Indeed, a sufficient statistic for a permutation invariant tester is a joint type $\hat{P}_{A^n,c}$. Our code satisfies $\hat{P}_{A^n,c} \approx P_{A,B}$, but it might happen that $\hat{P}_{A^n,c}$ although close to $P_{A,B}$ still takes highly unlikely values (for example, if we restrict all c to have the same composition P_0, the tester can easily detect the problem since P_B^n-measure of all strings of composition P_0 cannot exceed $O(1/\sqrt{n})$). Formally, to fool permutation invariant tester we need to have small total variation between the distribution on the joint types under P and Q. (It is natural to conjecture that rate $R = I(A;B)$ should be sufficient to achieve this requirement, though).

A related question is about the minimal possible rate (i.e. cardinality of $W \in [2^{nR}]$) required to have small total variation:

$$\mathrm{TV}(Q_{A^n,B^n}, P_{AB}^n) \le \epsilon \tag{24.21}$$

Note that condition (24.21) guarantees that any tester (permutation invariant or not) is fooled to believe he sees the truly iid (A^n, B^n). The minimal required rate turns out to be (Cuff'2012):

$$R = \min_{A \to U \to B} I(A,B;U)$$

a quantity known as Wyner's common information $C(A;B)$. Showing that Wyner's common information is a lower-bound is not hard. Indeed, since $Q_{A^n,B^n} \approx P_{AB}^n$ (in TV) we have

$$I(Q_{A^{t-1},B^{t-1}}, Q_{A_tB_t|A^{t-1},B^{t-1}}) \approx I(P_{A^{t-1},B^{t-1}}, P_{A_tB_t|A^{t-1},B^{t-1}}) = 0$$

(Here one needs to use finiteness of the alphabet of A and B and the bounds relating $H(P) - H(Q)$ with $\mathrm{TV}(P,Q)$). We have (under Q!)

$$\begin{aligned} nR = H(W) &\ge I(A^n, B^n; W) && (24.22)\\ &\ge \sum_{t=1}^T I(A_t, B_t; W) - I(A_t, B_t; A^{t-1}B^{t-1}) && (24.23)\\ &\approx \sum_{t=1}^T I(A_t, B_t; W) && (24.24)\\ &\gtrsim nC(A;B) && (24.25) \end{aligned}$$

where in the last step we used the crucial observation that under Q there is a Markov chain

$$A_t \to W \to B_t$$

and that Wyner's common information $P_{A,B} \mapsto C(A;B)$ should be continuous in the total variation distance on $P_{A,B}$. Showing achievability is a little more involved.

6.441 Information Theory

Spring 2016