Topics: Strong Converse, Channel Dispersion, Joint Source Channel Coding (JSCC)

20.1 Strong Converse

We begin by stating the main theorem.

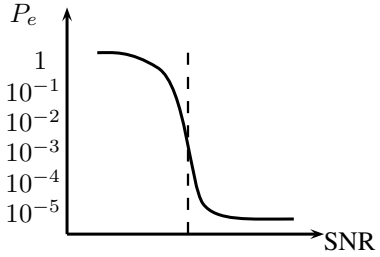
Theorem 20.1. *For any stationary memoryless channel with either $|\mathcal{A}| < \infty$ or $|\mathcal{B}| < \infty$ we have $C_\epsilon = C$ for $0 < \epsilon < 1$.*

Remark: In Theorem 16.4, we showed that $C \le C_\epsilon \le \frac{C}{1-\epsilon}$. Now we are asserting that equality holds for every ϵ. Our previous converse arguments showed that communication with an arbitrarily small error probability is possible only when using rate $R < C$; the strong converse shows that when you try to communicate with any rate above capacity $R > C$, then the probability of error will go to 1 (typically with exponential speed in n). In other words,

$$\epsilon^*(n, \exp(nR)) \to \begin{cases} 0 & R < C \\ 1 & R > C \end{cases}$$

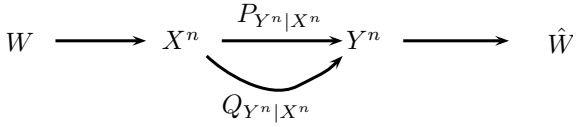
where $\epsilon^*(n, M)$ is the inverse of $M^*(n, \epsilon)$ defined in (16.3).

In practice, engineers observe this effect in the form of *waterfall plots*, which depict the dependence of a given communication system (code+modulation) on the SNR.



Below a certain SNR, the probability of error shoots up to 1, so that the receiver will only see garbage.

Proof. We will give a sketch of the proof. Take an (n, M, ϵ)-code for channel $P_{Y|X}$. The main trick is to consider an auxiliary channel $Q_{Y|X}$ which is easier to analyze.



Sketch 1: Here, we take $Q_{Y^n|X^n} = (P_Y^*)^n$, where P_Y^* is the capacity-achieving output distribution (caod) of the channel $P_{Y|X}$.[1] Note that for communication purposes, $Q_{Y^n|X^n}$ is a useless channel; it ignores the input and randomly picks a member of the output space according to $(P_Y^*)^n$, so that X^n and Y^n are decoupled (independent). Consider the probability of error under each channel:

$$\begin{aligned}\mathbb{Q}[\hat{W} = W] &= \frac{1}{M} \quad \text{(Blindly guessing the sent codeword)}\\ \mathbb{P}[\hat{W} = W] &= 1 - \epsilon\end{aligned}$$

Since the random variable $\mathbf{1}_{\{\hat{W}=W\}}$ has a huge mass under $\mathbb{P}$ and small mass under $\mathbb{Q}$, this looks like a great binary hypothesis test to distinguish the two distributions, $P_{WX^nY^n\hat{W}}$ and $Q_{WX^nY^n\hat{W}}$. Since any hypothesis test can't beat the optimal Neyman-Pearson test, we get the upper bound

$$\beta_{1-\epsilon}(P_{WX^nY^n\hat{W}}, Q_{WX^nY^n\hat{W}}) \le \frac{1}{M} \tag{20.1}$$

(Recall that $\beta_\alpha(P, Q) = \inf_{P[E]\ge\alpha} Q[E]$). Since the likelihood ratio is a sufficient statistic for this hypothesis test, we can test only between

$$\frac{P_{WX^nY^n\hat{W}}}{Q_{WX^nY^n\hat{W}}} = \frac{P_W P_{X^n|W} P_{Y^n|X_n} P_{\hat{W}|Y^n}}{P_W P_{X^n|W} (P_Y^*)^n P_{\hat{W}|Y^n}} = \frac{P_{W|X^n} P_{X^nY^n} P_{\hat{W}|Y^n}}{P_{W|X^n} P_{X^n} (P_Y^*)^n P_{\hat{W}|Y^n}} = \frac{P_{X^nY^n}}{P_{X^n}(P_Y^*)^n}$$

Therefore, inequality above becomes

$$\beta_{1-\epsilon}(P_{X^nY^n}, P_{X^n}(P_Y^*)^n) \le \frac{1}{M} \tag{20.2}$$

Computing the LHS of this bound need not be easy, since generally we know $P_{Y|X}$ and P_Y^*, but can't assume anything about P_{X^n} which depends on the code. (Note that X^n is the output of the encoder and uniformly distributed on the codebook for deterministic encoders). Certain tricks are needed to remove the dependency on codebook. However, in case the channel is "symmetric" the dependence on the codebook disappears: this is shown in the following example for the BSC. To treat the general case one simply decomposes the channel into symmetric subchannels (for example, by considering constant composition subcodes).

Example. For a $\mathrm{BSC}(\delta)^n$, recall that

$$\begin{aligned}P_{Y^n|X^n}(y^n|x^n) &= P_Z^n(y^n - x^n), \quad Z^n \sim \mathrm{Bern}(\delta)^n\\ (P_Y^*)^n(y^n) &= 2^{-n}\end{aligned}$$

From the Neyman Pearson test, the optimal HT takes the form

$$\beta_\alpha(\underbrace{P_{X^nY^n}}_{\mathbb{P}}, \underbrace{P_{X^n}(P_Y^*)^n}_{\mathbb{Q}}) = \mathbb{Q}\left[\log\frac{P_{X^nY^n}}{P_{X^n}(P_Y^*)^n} \ge \gamma\right] \text{ where } \alpha = \mathbb{P}\left[\log\frac{P_{X^nY^n}}{P_{X^n}(P_Y^*)^n} \ge \gamma\right]$$

For the BSC, this becomes

$$\log\frac{P_{X^nY^n}}{P_{X^n}(P_Y^*)^n} = \log\frac{P_{Z^n}(y^n - x^n)}{2^{-n}}$$

[1] Recall from Theorem 4.5 that the caod of a random transformation *always exists and is unique*, whereas a caid may not exist.

So under each hypothesis $\mathbb{P}$ and $\mathbb{Q}$, the difference $Y^n - X^n$ takes the form

$$\mathbb{Q} : Y^n - X^n \sim \text{Bern}(\frac{1}{2})^n$$
$$\mathbb{P} : Y^n - X^n \sim \text{Bern}(\delta)^n$$

Now all the relevant distributions are known, so we can compute β_α

$$\begin{aligned}\beta_\alpha(P_{X^nY^n}, P_{X^n}(P_Y^*)^n) &= \beta_\alpha(\text{Bern}(\delta)^n, \text{Bern}(\frac{1}{2})^n) \\ &= 2^{-nD(\text{Bern}(\delta)\|\text{Bern}(\frac{1}{2}))+o(n)} \qquad \text{(Stein's Lemma Theorem 11.1)} \\ &= 2^{-nd(\delta\|\frac{1}{2})+o(n)}\end{aligned}$$

Putting this all together, we see that any (n, M, ϵ) code for the BSC satisfies

$$2^{-nd(\delta\|\frac{1}{2})+o(n)} \le \frac{1}{M} \implies \log M \le nd(\delta\|\frac{1}{2}) + o(n)$$

Since this is satisfied for all codes, it is also satisfied for the optimal code, so we get the converse bound

$$\liminf_{n\to\infty} \frac{1}{n} \log M^*(n, \epsilon) \le d(\delta\|\frac{1}{2}) = \log 2 - h(\delta)$$

For a general channel, this computation can be much more difficult. The expression for β in this case is

$$\beta_{1-\epsilon}(P_{X^n}P_{Y^n|X^n}, P_{X^n}(P_Y^*)^n) = 2^{-nD(P_{Y|X}\|P_Y^*|\bar{P}_X)+o(n)} \le \frac{1}{M} \tag{20.3}$$

where $\bar{P}_X$ is unknown (depending on the code).

Explanation of (20.3): A statistician observes sequences of (X^n, Y^n):

$$\begin{array}{rl} X^n = & [\,\boxed{0} \;\; 1 \;\; 2 \;\; \boxed{0 \;\; 0} \;\; 1 \;\; 2 \;\; 2\,] \\ Y^n = & [\,\boxed{a} \;\; b \;\; b \;\; \boxed{a \;\; c} \;\; c \;\; a \;\; b\,] \end{array}$$

On the marked three blocks, test between iid samples of $P_{Y|X=0}$ vs P_Y^*, which has exponent $D(P_{Y|X=0}\|P_Y^*)$. Thus, intuitively averaging over the composition of the codeword we get that the exponent of β is given by (20.3).

Recall that from the saddle point characterization of capacity (Theorem 4.4) for any distribution $\bar{P}_X$ we have

$$D(P_{Y|X}\|P_Y^*|\bar{P}_X) \le C\,. \tag{20.4}$$

Thus from (20.3) and (20.1):

$$\log M \le nD(P_{Y|X}\|P_Y^*|\bar{P}_X) + o(n) \le nC + o(n)$$

Sketch 2: (More formal) Again, we will choose a dummy auxiliary channel $Q_{Y^n|X^n} = (Q_Y)^n$. However, choice of Q_Y will depend on one of the two cases:

1. If $|\mathcal{B}| < \infty$ we take $Q_Y = P_Y^*$ (the caod) and note that from (16.16) we have

$$\sum_y P_{Y|X}(y|x_0) \log^2 P_{Y|X}(y|x_0) \le \log^2 |\mathcal{B}| \qquad \forall x_0 \in \mathcal{A}$$

and since $\min_y P_Y^*(y) > 0$ (without loss of generality), we conclude that for any distribution of X on $\mathcal{A}$ we have

$$\operatorname{Var}\left[\log \frac{P_{Y|X}(Y|X)}{Q_Y(Y)}|X\right] \le K < \infty \qquad \forall P_X\,. \tag{20.5}$$

Furthermore, we also have from (20.4) that

$$\mathbb{E}\left[\log \frac{P_{Y|X}(Y|X)}{Q_Y(Y)}|X\right] \le C \qquad \forall P_X\,. \tag{20.6}$$

2. If $|\mathcal{A}| < \infty$, then for each codeword $c \in \mathcal{A}^n$ we define its *composition* as

$$\hat{P}_c(x) \triangleq \frac{1}{n}\sum_{j=1}^n 1\{c_j = x\}\,.$$

By simple counting it is clear that from any (n, M, ϵ) code, it is possible to select an (n, M', ϵ) subcode, such that a) all codeword have the same composition P_0; and b) $M' > \frac{M}{n^{|\mathcal{A}|}}$. Note that, $\log M = \log M' + O(\log n)$ and thus we may replace M with M' and focus on the analysis of the chosen subcode. Then we set $Q_Y = P_{Y|X} \circ P_0$. In this case, from (16.9) we have

$$\operatorname{Var}\left[\log \frac{P_{Y|X}(Y|X)}{Q_Y(Y)}|X\right] \le K < \infty \qquad X \sim P_0\,. \tag{20.7}$$

Furthermore, we also have

$$\mathbb{E}\left[\log \frac{P_{Y|X}(Y|X)}{Q_Y(Y)}|X\right] = D(P_{Y|X}\|Q_Y|P_0) = I(X;Y) \le C \qquad X \sim P_0\,. \tag{20.8}$$

Now, proceed as in (20.2) to get

$$\beta_{1-\epsilon}(P_{X^nY^n}, P_{X^n}(Q_Y)^n) \le \frac{1}{M}\,. \tag{20.9}$$

We next apply the lower bound on β from Theorem 10.5:

$$\gamma\beta_{1-\epsilon}(P_{X^nY^n}, P_{X^n}(Q_Y)^n) \ge \mathbb{P}\Big[\log \frac{\mathrm{d}P_{Y^n|X^n}(Y^n|X^n)}{\mathrm{d}\prod Q_Y(Y_i)} \le \log\gamma\Big] - \epsilon$$

Set $\log\gamma = nC + K'\sqrt{n}$ with K' to be chosen shortly and denote for convenience

$$S_n \triangleq \log \frac{\mathrm{d}P_{Y^n|X^n}(Y^n|X^n)}{\mathrm{d}\prod Q_Y(Y_i)} = \sum_{j=1}^n \log \frac{\mathrm{d}P_{Y|X}(Y_j|X_j)}{\mathrm{d}Q_Y(Y_j)}$$

Conditioning on X^n and using (20.6) and (20.8) we get

$$\mathbb{P}\left[S_n \le nC + K'\sqrt{n}|X^n\right] \ge \mathbb{P}\left[S_n \le n\,\mathbb{E}[S_n|X^n] + K'\sqrt{n}|X^n\right]$$

From here, we apply Chebyshev inequality and (20.5) or (20.7) to get

$$\mathbb{P}\left[S_n \le n\,\mathbb{E}[S_n|X^n] + K'\sqrt{n}|X^n\right] \ge 1 - \frac{K'^2}{K}\,.$$

If we set K' so large that $1 - \frac{K'^2}{K} > 2\epsilon$ then overall we get that

$$\log \beta_{1-\epsilon}(P_{X^nY^n}, P_{X^n}(Q_Y)^n) \ge -nC - K'\sqrt{n} - \log\epsilon\,.$$

Consequently, from (20.9) we conclude that

$$\log M^*(n,\epsilon) \le nC + O(\sqrt{n})\,,$$

implying the strong converse. □

In summary, the take-away points for the strong converse are

1. Strong converse can be proven by using binary hypothesis testing.
2. The capacity saddle point (20.4) is key.

In the homework, we will explore in detail proofs of the strong converse for the BSC and the AWGN channel.

20.2 Stationary memoryless channel without strong converse

It may seem that the strong converse should hold for an arbitrary stationary memoryless channel (it was only showed for the *discrete* ones above). However, it turns out that there exist counterexamples. We construct one next.

Let output alphabet be $\mathcal{B} = [0,1]$. The input $\mathcal{A}$ is going to be countable, it will be convenient to define it as

$$\mathcal{A} = \{(j,m) : j,m \in \mathbb{Z}_+, 0 \le j \le m\}\,.$$

The single-letter channel $P_{Y|X}$ is defined in terms of probability density function as

$$p_{Y|X}(y|(j,m)) = \begin{cases} a_m, & \frac{j}{m} \le y \le \frac{j+1}{m}, , \\ b_m, & \text{otherwise}\,, \end{cases}$$

where a_m, b_m are chosen to satisfy

$$\frac{1}{m}a_m + (1-\frac{1}{m})b_m = 1 \tag{20.10}$$

$$\frac{1}{m}a_m \log a_m + (1-\frac{1}{m})b_m \log b_m = C\,, \tag{20.11}$$

where $C > 0$ is an arbitary fixed constant. Note that for large m we have

$$a_m = \frac{mC}{\log m}(1 + O(\frac{1}{\log m}))\,, \tag{20.12}$$

$$b_m = 1 - \frac{C}{\log m} + O(\frac{1}{\log^2 m}) \tag{20.13}$$

It is easy to see that $P_Y^* = \text{Unif}[0,1]$ is the capacity-achieving output distribution and

$$\sup_{P_X} I(X;Y) = C\,.$$

Thus by Theorem 16.6 the capacity of the corresponding stationary memoryless channel is C. We next show that nevertheless the ϵ-capacity can be strictly greater than C.

Indeed, fix blocklength n and consider a *single letter* distribution P_X assigning equal weights to all atoms (j,m) with $m = \exp\{2nC\}$. It can be shown that in this case, the distribution of a single-letter information density is given by

$$i(X;Y) \approx \begin{cases} 2nC, & w.p. \frac{1}{2n} \\ 0, & w.p. 1 - \frac{1}{2n} \end{cases}$$

Thus, for blocklength-n density we have

$$\frac{1}{n} i(X^n;Y^n) \to 2C\text{Poisson}(1/2)\,.$$

Therefore, from Theorem 15.1 we get that for $\epsilon > 1 - e^{-1/2}$ there exist (n, M, ϵ)-codes with

$$\log M \geq 2nC\,.$$

In particular,

$$C_\epsilon \geq 2C \quad \forall \epsilon > 1 - e^{-1/2}$$

20.3 Channel Dispersion

The strong converse tells us that $\log M^*(n,\epsilon) = nC + o(n)\ \forall \epsilon \in (0,1)$. An engineer sees this, and estimates $\log M^* \approx nC$. However, this doesn't give any information about the dependence of $\log M^*$ on the error probability ϵ, which is hidden in the $o(n)$ term. We unravel this in the following theorem.

Theorem 20.2. *Consider one of the following channels:*

1. *DMC*
2. *DMC with cost constraint*
3. *AWGN or parallel AWGN*

The following expansion holds for a fixed $0 < \epsilon < 1/2$ and $n \to \infty$

$$\log M^*(n,\epsilon) = nC - \sqrt{nV}Q^{-1}(\epsilon) + O(\log n)$$

where Q^{-1} is the inverse of the complementary standard normal CDF, the channel capacity is $C = I(X^*;Y^*) = \mathbb{E}[i(X^*;Y^*)]$, and the channel dispersion[2] is $V = \text{Var}[i(X^*;Y^*)|X^*]$.

[2]There could be multiple capacity-achieving input distributions, in which case P_{X^*} should be chosen as the one that minimizes $\text{Var}[i(X^*;Y^*)|X^*]$. See [PPV10] for more details.

Proof. For achievability, we have shown (Theorem 16.7) that $\log M^*(n,\epsilon) \geq nC - \sqrt{nV}Q^{-1}(\epsilon)$ by refining the proof of the noisy channel coding theorem using the CLT.

The converse statement is $\log M^* \leq -\log \beta_{1-\epsilon}(P_{X^nY^n}, P_{X^n}(P_Y^*)^n)$. For the BSC, we showed that the RHS of the previous expression is

$$-\log \beta_{1-\epsilon}(\text{Bern}(\delta)^n, \text{Bern}(\frac{1}{2})^n) = nd(\delta\|\frac{1}{2}) + \sqrt{nV}Q^{-1}(\epsilon) + o(\sqrt{n})$$

(see homework) where the dispersion is

$$V = \text{Var}_{Z\sim\text{Bern}(\delta)}\left[\log \frac{\text{Bern}(\delta)}{\text{Bern}(\frac{1}{2})}(Z)\right].$$

The general proof is omitted. □

Remark: This expansion only applies for certain channels (as described in the theorem). If, for example, $\text{Var}[i(X;Y)] = \infty$, then the theorem need not hold and there are other stable (non-Gaussian) distributions that we might converge to instead. Also notice that for DMC without cost constraint

$$\text{Var}[i(X^*;Y^*)|X^*] = \text{Var}[i(X^*;Y^*)]$$

since (capacity saddle point!) $\mathbb{E}[i(X^*;Y^*)|X^* = x] = C$ for P_{X^*}-almost all x.

20.3.1 Applications

As stated earlier, direct computation of $M^*(n,\epsilon)$ by exhaustive search doubly exponential in complexity, and thus is infeasible in most cases. However, we can get an easily computable approximation using the channel dispersion via

$$\log M^*(n,\epsilon) \approx nC - \sqrt{nV}Q^{-1}(\epsilon)$$

Consider a BEC ($n = 500, \delta = 1/2$) as an example of using this approximation. For this channel, the capacity and dispersion are

$$C = 1 - \delta$$
$$V = \delta\bar{\delta}$$

Where $\bar{\delta} = 1 - \delta$. Using these values, our approximation for this BEC becomes

$$\log M^*(500, 10^{-3}) \approx nC - \sqrt{nV}Q^{-1}(\epsilon) = n\bar{\delta} - \sqrt{n\delta\bar{\delta}}Q^{-1}(10^{-3}) \approx 215.5 \text{ bits}$$

In the homework, for the BEC$(500, 1/2)$ we obtained bounds $213 \leq \log M^*(500, 10^{-3}) \leq 217$, so this approximation falls in the middle of these bounds.

Examples of Channel Dispersion

For a few common channels, the dispersions are

BEC: $V(\delta) = \delta\bar{\delta}\log^2 2$

BSC: $V(\delta) = \delta\bar{\delta}\log^2 \frac{\bar{\delta}}{\delta}$

AWGN: $V(P) = \frac{P(P+2)}{2(P+1)^2}\log^2 e$ (Real) $\qquad \frac{P(P+2)}{(P+1)^2}\log^2 e$ (Complex)

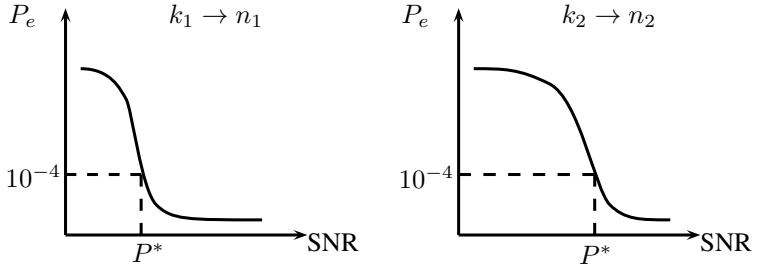
Parallel AWGN: $V(\mathbf{P},\sigma^2) = \sum_{j=1}^{L} V_{AWGN}(\frac{P_j}{\sigma_j^2}) = \frac{\log^2 e}{2}\sum_{j=1}^{L}\left|1-\left(\frac{\sigma_j^2}{T}\right)^2\right|^+$

where $\sum_{j=1}^{L}|T-\sigma_j^2|^+ = P$ is the water-filling solution of the parallel AWGN

Punchline: Although the only machinery needed for this approximation is the CLT, the results produced are incredibly useful. Even though $\log M^*$ is nearly impossible to compute on its own, by only finding C and V we are able to get a good approximation that is easily computable.

20.4 Normalized Rate

Suppose you're given two codes $k_1 \to n_1$ and $k_2 \to n_2$, how do you fairly compare them? Perhaps they have the following waterfall plots



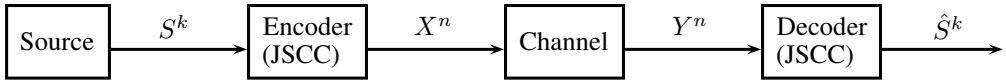
After inspecting these plots, one may believe that the $k_1 \to n_1$ code is better, since it requires a smaller SNR to achieve the same error probability. However, there are many factors, such as blocklength, rate, etc. that don't appear on these plots. To get a fair comparison, we can use the notion of *normalized rate*. To each $(n, 2^k, \epsilon)$-code, define

$$R_{\text{norm}} = \frac{k}{\log_2 M^*_{AWGN}(n,\epsilon,P)} \approx \frac{k}{nC(P) - \sqrt{nV(P)}Q^{-1}(\epsilon)}$$

Take $\epsilon = 10^{-4}$, and P (SNR) according to the water fall plot corresponding to $P_e = 10^{-4}$, and we can compare codes directly (see Fig. 20.1). This normalized rate gives another motivation for the expansion given in Theorem 20.2.

20.5 Joint Source Channel Coding

Now we will examine a slightly different information transmission scenario called *Joint Source Channel Coding*



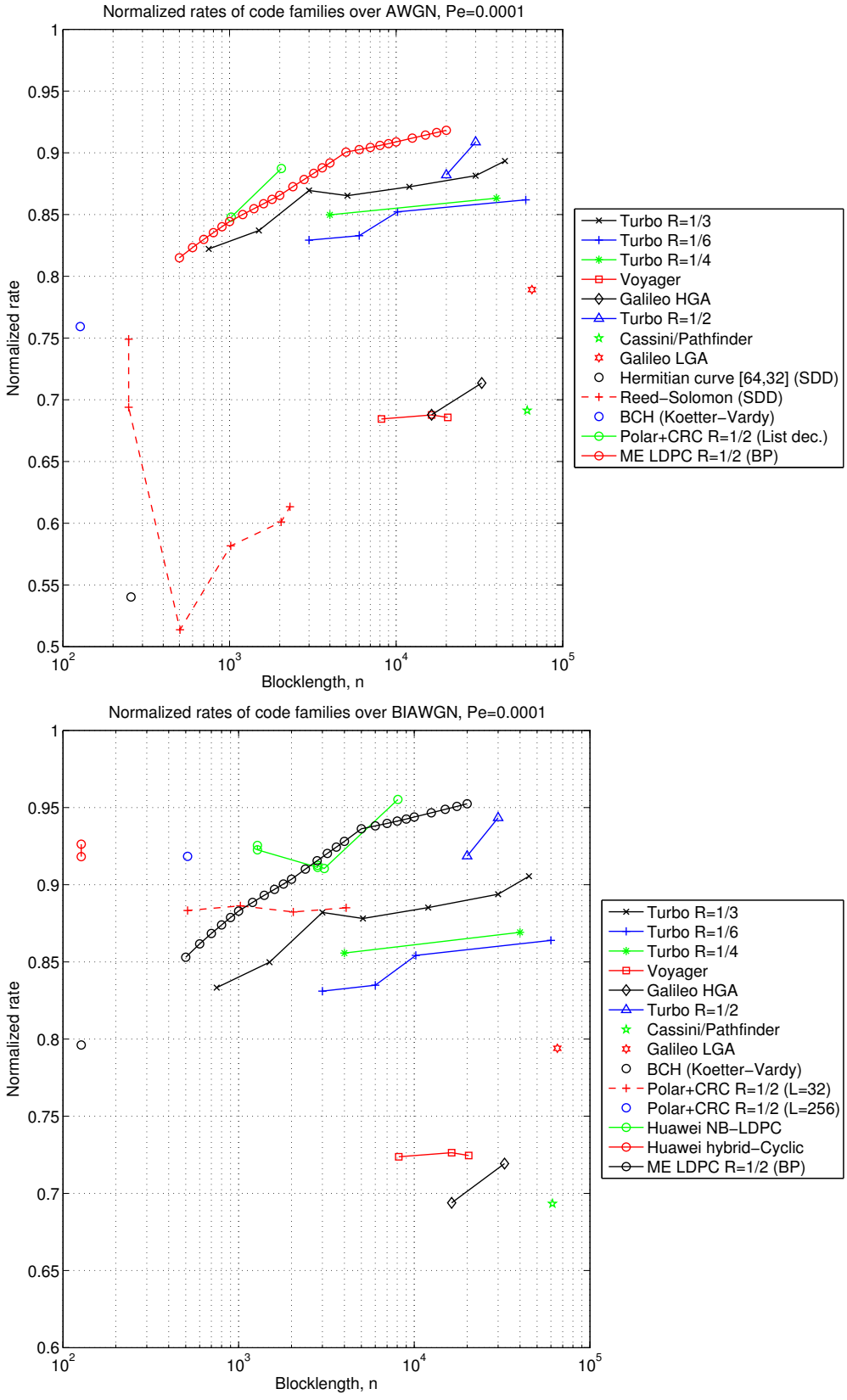


Figure 20.1: Normalized rates for various codes. Plots generated via [Spe15].

Definition 20.1. For a Joint Source Channel Code

- Goal: $\mathbb{P}[S^k \neq \hat{S}^k] \leq \epsilon$
- Encoder: $f : \mathcal{A}^k \to \mathcal{X}^n$
- Decoder: $g : \mathcal{Y}^n \to \mathcal{A}^k$
- Fundamental Limit (Optimal probability of error): $\epsilon^*_{JSCC}(k,n) = \inf_{f,g} \mathbb{P}[S^k \neq \hat{S}^k]$

where the rate is $R = \frac{k}{n}$ (symbol per channel use).

Note: In channel coding we are interested in transmitting M messages and all messages are born equal. Here we want to convey the source realizations which might not be equiprobable (has redundancy). Indeed, if S^k is uniformly distributed on, say, $\{0,1\}^k$, then we are back to the channel coding setup with $M = 2^k$ under average probability of error, and $\epsilon^*_{JSCC}(k,n)$ coincides with $\epsilon^*(n, 2^k)$ defined in Section 20.1.

Note: Here, we look for a clever scheme to directly encode k symbols from $\mathcal{A}$ into a length n channel input such that we achieve a small probability of error over the channel. This feels like a mix of two problems we've seen: compressing a source and coding over a channel. The following theorem shows that compressing and channel coding separately is optimal. This is a relief, since it implies that we do not need to develop any new theory or architectures to solve the Joint Source Channel Coding problem. As far as the leading term in the asymptotics is concerned, the following two-stage scheme is optimal: First use the optimal compressor to eliminate all the redundancy in the source, then use the optimal channel code to add redundancy to combat the noise in the transmission.

Theorem 20.3. *Let the source $\{S_k\}$ be stationary memoryless on a finite alphabet with entropy H. Let the channel be stationary memoryless with finite capacity C. Then*

$$\epsilon^*_{JSCC}(nR,n) \begin{cases} \to 0 & R < C/H \\ \not\to 0 & R > C/H \end{cases} \quad n \to \infty.$$

Note: Interpretation: Each source symbol has information content (entropy) H bits. Each channel use can convey C bits. Therefore to reliably transmit k symbols over n channel uses, we need $kH \leq nC$.

Proof. **Achievability.** The idea is to separately compress our source and code it for transmission. Since this is a feasible way to solve the JSCC problem, it gives an achievability bound. This separated architecture is

$$S^k \xrightarrow{f_1} W \xrightarrow{f_2} X^n \overset{P_{Y^n|X^n}}{\longrightarrow} Y^n \xrightarrow{g_2} \hat{W} \xrightarrow{g_1} \hat{S}^k$$

Where we use the optimal compressor (f_1, g_1) and optimal channel code (<u>maximum</u> probability of error) (f_2, g_2). Let W denote the output of the compressor which takes at most M_k values. Then

$$\text{(From optimal compressor) } \frac{1}{k}\log M_k > H + \delta \implies \mathbb{P}[\hat{S}^k \neq S^k(W)] \leq \epsilon \quad \forall k \geq k_0$$

$$\text{(From optimal channel code) } \frac{1}{n}\log M_k < C - \delta \implies \mathbb{P}[\hat{W} \neq m | W = m] \leq \epsilon \quad \forall m, \forall k \geq k_0$$

Using both of these,

$$\begin{aligned}\mathbb{P}[S^k \neq \hat{S}^k(\hat{W})] &\leq \mathbb{P}[S^k \neq \hat{S}^k, W = \hat{W}] + \mathbb{P}[W \neq \hat{W}] \\ &\leq \mathbb{P}[S^k \neq \hat{S}^k(W)] + \mathbb{P}[W \neq \hat{W}] \leq \epsilon + \epsilon\end{aligned}$$

And therefore if $R(H+\delta) < C - \delta$, then $\epsilon^* \to 0 \overset{\delta \to 0}{\Longrightarrow} R > C/H$.

Converse: channel-substitution proof. Let $Q_{S^k\hat{S}^k} = U_{S^k} P_{\hat{S}^k}$ where U_{S^k} is the uniform distribution. Using data processing

$$D(P_{S^k\hat{S}^k} \| Q_{S^k\hat{S}^k}) = D(P_{S^k} \| U_{S^k}) + D(P_{\hat{S}|S^k} \| P_{\hat{S}} | P_{S^k}) \geq d(1-\epsilon \| \frac{1}{|\mathcal{A}|^k})$$

Rearranging this gives

$$\begin{aligned}I(S^k; \hat{S}^k) &\geq d(1-\epsilon \| \frac{1}{|\mathcal{A}|^k}) - D(P_{S^k} \| U_{S^k}) \\ &\geq -\log 2 + k\bar{\epsilon} \log |\mathcal{A}| + H(S^k) - k \log |\mathcal{A}| \\ &= H(S^k) - \log 2 - k\epsilon \log |\mathcal{A}|\end{aligned}$$

Which follows from expanding out the terms. Now, normalizing and taking the sup of both sides gives

$$\frac{1}{n} \sup_{X^n} I(X^n; Y^n) \geq \frac{1}{n} H(S^k) - \epsilon \frac{k}{n} \log |A| + o(1)$$

letting $R = k/n$, this shows

$$C \geq RH - \epsilon R \log |A| \implies \epsilon \geq \frac{RH - C}{R \log |A|} > 0$$

where the last expression is positive when $R > C/H$.

Converse: usual proof. Any JSCC encoder/decoder induces a Markov chain

$$S^k \to X^n \to Y^n \to \hat{S}^k.$$

Applying data processing for mutual information

$$I(S^k; \hat{S}^k) \leq I(X^n; Y^n) \leq \sup_{P_{X^n}} I(X^n; Y^n) = nC.$$

On the other hand, since $\mathbb{P}[S^k \neq \hat{S}^k] \leq \epsilon_n$, Fano's inequality yields

$$I(S^k; \hat{S}^k) = H(S^k) - H(S^k | \hat{S}^k) \geq kH - \epsilon_n \log |\mathcal{A}|^k - \log 2.$$

Combining the two gives

$$nC \geq kH - \epsilon_n \log |\mathcal{A}|^k - \log 2.$$

Since $R = \frac{k}{n}$, dividing both sides by n and sending $n \to \infty$ yields

$$\liminf_{n \to \infty} \epsilon_n \geq \frac{RH - C}{R \log |\mathcal{A}|}.$$

Therefore ϵ_n does not vanish if $R > C/H$. □

6.441 Information Theory

Spring 2016