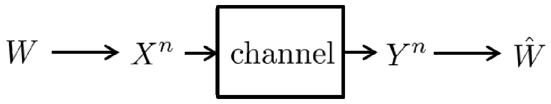
§ 21. Channel coding with feedback

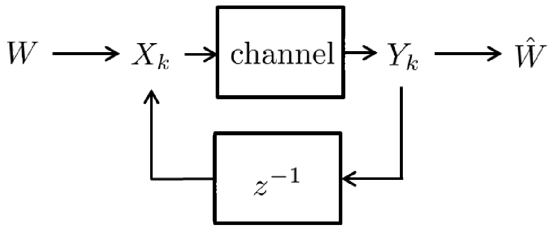
Criticism: Channels without feedback don't exist (except storage).

Motivation: Consider the communication channel of the downlink transmission from a satellite to earth. Downlink transmission is very expensive (power constraint at the satellite), but the uplink from earth to the satellite is cheap which makes virtually noiseless feedback readily available at the transmitter (satellite). In general, channel with noiseless feedback is interesting when such asymmetry exists between uplink and downlink.

In the first half of our discussion, we shall follow Shannon to show that feedback gains "nothing" in the conventional setup, while in the second half, we look at situations where feedback gains a lot.



channel w/o feedback



causality constraint
noiseless / Shannon / ideal feedback channel

channel with feedback

21.1 Feedback does not increase capacity for stationary memoryless channels

Definition 21.1 (Code with feedback)**.** An (n, M, ϵ)-code with feedback is specified by the encoder-decoder pair (f, g) as follows:

- Encoder: (time varying)

$$\begin{aligned} &f_1 : [M] \to \mathcal{A} \\ &f_2 : [M] \times \mathcal{B} \to \mathcal{A} \\ &\vdots \\ &f_n : [M] \times \mathcal{B}^{n-1} \to \mathcal{A} \end{aligned}$$

- Decoder:

$$g : \mathcal{B}^n \to [M]$$

such that $\mathbb{P}[W \neq \hat{W}] \leq \epsilon$.

Note: [Probability space]

$$W \sim \text{ uniform on } [M]$$

$$\left.\begin{array}{l} X_1 = f_1(W) \overset{P_{Y|X}}{\longrightarrow} Y_1 \\ \vdots \\ X_n = f_n(W, Y_1^{n-1}) \overset{P_{Y|X}}{\longrightarrow} Y_n \end{array}\right\} \longrightarrow \hat{W} = g(Y^n)$$

Definition 21.2 (Fundamental limits)**.**

$$\begin{aligned} M^*_{fb}(n,\epsilon) &= \max\{M : \exists (n,M,\epsilon) \text{ code with feedback.}\} \\ C_{fb,\epsilon} &= \liminf_{n\to\infty} \frac{1}{n} \log M^*_{fb}(n,\epsilon) \\ C_{fb} &= \lim_{\epsilon\to 0} C_{fb,\epsilon} && \text{(Shannon capacity with feedback)} \end{aligned}$$

Theorem 21.1 (Shannon 1956)**.** *For a stationary memoryless channel,*

$$C_{fb} = C = C_i = \sup_{P_X} I(X;Y)$$

Proof. Achievability: Although it is obvious that $C_{fb} \ge C$, we wanted to demonstrate that in fact constructing codes achieving capacity with *full feedback* can be done directly, without appealing to a (much harder) problem of non-feedback codes. Let $\pi_t(\cdot) \triangleq P_{W|Y^t}(\cdot|Y^t)$ with the (random) posterior distribution after t steps. It is clear that due to the knowledge of Y^t on both ends, transmitter and receiver have perfectly synchronized knowledge of π_t. Now consider how the transmission progresses:

1. Initialize $\pi_0(\cdot) = \frac{1}{M}$
2. At $(t+1)$-th step, having knowledge of π_t all messages are partitioned into classes $\mathcal{P}_a$, according to the values $f_{t+1}(\cdot, Y^t)$:

   $$\mathcal{P}_a \triangleq \{j \in [M] : f_{t+1}(j, Y^t) = a\} \qquad a \in \mathcal{A}\,.$$

   Then transmitter, possessing the knowledge of the true message W, selects a letter $X_{t+1} = f_{t+1}(W, Y^t)$.

3. Channel perturbs X_{t+1} into Y_{t+1} and both parties compute the updated posterior:

   $$\pi_{t+1}(j) \triangleq \pi_t(j) B_{t+1}(j)\,, \qquad B_{t+1}(j) \triangleq \frac{P_{Y|X}(Y_{t+1}|f_{t+1}(j, Y^t))}{\sum_{a\in\mathcal{A}} \pi_t(\mathcal{P}_a)}\,.$$

   Notice that (this is the crucial part!) the random multiplier satisfies:

   $$\mathbb{E}[\log B_{t+1}(W)|Y^t] = \sum_{a\in\mathcal{A}} \sum_{y\in\mathcal{B}} \pi_t(\mathcal{P}_a) \log \frac{P_{Y|X}(y|a)}{\sum_{a\in\mathcal{A}} \pi_t(\mathcal{P}_a) a} = I(\tilde{\pi}_t, P_{Y|X}) \qquad (21.1)$$

   where $\tilde{\pi}_t(a) \triangleq \pi_t(\mathcal{P}_a)$ is a (random) distribution on $\mathcal{A}$.

The goal of the code designer is to come up with such a partitioning $\{\mathcal{P}_a, a \in \mathcal{A}\}$ that the speed of growth of $\pi_t(W)$ is maximal. Now, analyzing the speed of growth of a random-multiplicative process is best done by taking logs:

$$\log \pi_t(j) = \sum_{s=1}^{t} \log B_s + \log \pi_0(j)\,.$$

Intutively, we expect that the process $\log \pi_t(W)$ resembles a random walk starting from $-\log M$ and having a positive drift. Thus to estimate the time it takes for this process to reach value 0 we need to estimate the upward drift. Appealing to intuition and the law of large numbers we approximate

$$\log \pi_t(W) - \log \pi_0(W) \approx \sum_{s=1}^{t} \mathbb{E}[\log B_s]\,.$$

Finally, from (21.1) we conclude that the best idea is to select partitioning at each step in such a way that $\tilde{\pi}_t \approx P_X^*$ (caid) and this obtains

$$\log \pi_t(W) \approx tC - \log M\,,$$

implying that the transmission terminates in time $\approx \frac{\log M}{C}$. The important lesson here is the following: *The optimal transmission scheme should map messages to channel inputs in such a way that the induced input distribution $P_{X_{t+1}|Y^t}$ is approximately equal to the one maximizing $I(X;Y)$.* This idea is called *posterior matching* and explored in detail in [SF11].[1]

Converse: we are left to show that $C_{fb} \le C_i$.

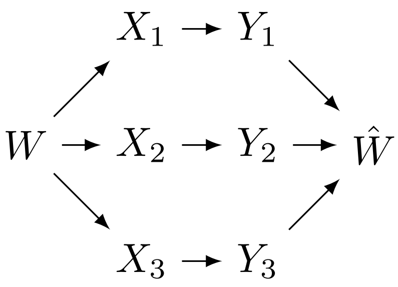
Recall the key in proving weak converse for channel coding without feedback: Fano's inequality plus the graphical model

$$W \to X^n \to Y^n \to \hat{W}. \tag{21.2}$$

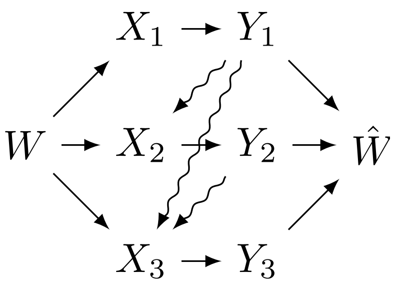
Then

$$h(\epsilon) + \bar{\epsilon} \log M \le I(W; \hat{W}) \le I(X^n; Y^n) \le nC_i.$$

With feedback the probabilistic picture becomes more complicated as the following figure shows for $n = 3$ (dependence introduced by the extra squiggly arrows):



without feedback



with feedback

So, while the Markov chain realtion in (21.2) is still true, we also have

$$P_{Y^n|X^n}(y^n|x^n) \ne \prod_{j=1}^{n} P_{Y|X}(y_j|x_j) \qquad (!)$$

(This is easy to see from the example where $X_2 = Y_1$ and thus $P_{Y_1|X^2}$ has no randomness.) There is still a large degree of independence in the channel, though. Namely, we have

$$(Y^{i-1}, W) \to X_i \to Y_i, \quad i = 1, \ldots, n \tag{21.3}$$

$$W \to Y^n \to \hat{W} \tag{21.4}$$

[1] Note that the magic of Shannon's theorem is that this optimal partitioning can also be done blindly. I.e. it is possible to preselect partitions $\mathcal{P}_a$ in a way independent of π_t (but dependent on t) and so that the $\pi_t(\mathcal{P}_a) \approx P_X^*(a)$ with overwhelming probability and for all $t \in [1, n]$.

Then

$$\begin{aligned}
h(\epsilon) + \bar{\epsilon}\log M &\le I(W;\hat{W}) && \text{(Fano)}\\
&\le I(W;Y^n) && \text{(Data processing applied to (21.4))}\\
&= \sum_{i=1}^n I(W;Y_i|Y^{i-1}) && \text{(Chain rule)}\\
&\le \sum_{i=1}^n I(W,Y^{i-1};Y_i) && (I(W;Y_i|Y^{i-1}) = I(W,Y^{i-1};Y_i) - I(Y^{i-1};Y_i))\\
&\le \sum_{i=1}^n I(X_i;Y_i) && \text{(Data processing applied to (21.3))}\\
&\le nC_i && \square
\end{aligned}$$

The following result (without proof) suggests that feedback does not even improve the speed of approaching capacity either (under fixed-length block coding) and can at most improve smallish $\log n$ terms:

Theorem 21.2 (Dispersion with feedback)**.** *For weakly input-symmetric DMC (e.g. additive noise, BSC, BEC) we have:*

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

(The meaning of this is that for such channels feedback can at most improve smallish $\log n$ terms.)

21.2* Alternative proof of Theorem 21.1 and Massey's directed information

The following alternative proof emphasizes on data processing inequality and the comparison idea (auxiliary channel) as in Theorem 19.1.

Proof. It is obvious that $C_{fb} \ge C$, we are left to show that $C_{fb} \le C_i$.

1. Recap of the steps of showing the strong converse of $C \le C_i$ in the last lecture: take any (n,M,ϵ) code, compare the two distributions:

$$P : W \to X^n \to Y^n \to \hat{W} \tag{21.5}$$

$$Q : W \to X^n \quad Y^n \to \hat{W} \tag{21.6}$$

   two key observations:

   a) Under Q, $W \perp\!\!\!\perp W$, so that $\mathbb{Q}[W = \hat{W}] = \frac{1}{M}$ while $\mathbb{P}[W = \hat{W}] \ge 1 - \epsilon$.

   b) The two graphical models give the factorization:

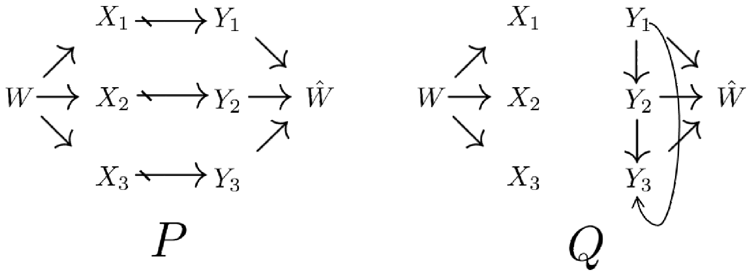
$$P_{W,X^n,Y^n,\hat{W}} = P_{W,X^n}P_{Y^n|X^n}P_{\hat{W}|Y^n}, \qquad Q_{W,X^n,Y^n,\hat{W}} = P_{W,X^n}P_{Y^n}P_{\hat{W}|Y^n}$$

   thus $D(P\|Q) = I(X^n;Y^n)$ measures the information flow through the links $X^n \to Y^n$.

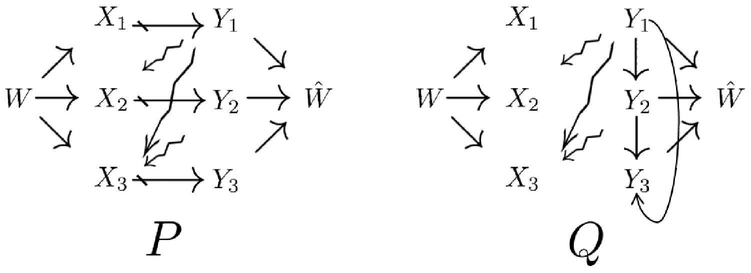
$$h(\epsilon) + \bar{\epsilon}\log M = d(1-\epsilon\|\frac{1}{M}) \overset{\text{d-proc ineq}}{\le} D(P\|Q) = I(X^n;Y^n) \overset{mem-less,stat}{=} \sum_{i=1}^n I(X;Y) \le nC_i \tag{21.7}$$

2. Notice that when feedback is present, $X^n \to Y^n$ is not memoryless due to the transmission protocol, let's unfold the probability space over time to see the dependence. As an example, the graphical model for $n = 3$ is given below:

No feedback



With feedback



If we define Q similarly as in the case without feedback, we will encounter a problem at the second last inequality in (21.7), as with feedback $I(X^n;Y^n)$ can be significantly larger than $\sum_{i=1}^n I(X;Y)$. Consider the example where $X_2 = Y_1$, we have $I(X^n;Y^n) = +\infty$ independent of $I(X;Y)$.

We also make the observe that if Q is defined in (21.6), $D(P\|Q) = I(X^n;Y^n)$ measures the information flow through all the $\nrightarrow$ and $\rightsquigarrow$ links. This motivates us to find a proper Q such that $D(P\|Q)$ only captures the information flow through all the $\nrightarrow$ links $\{X_i \to Y_i : i = 1, \dots, n\}$, so that $D(P\|Q)$ closely relates to nC_i, while still guarantees that $W \perp\!\!\!\perp W$, so that $\mathbb{Q}[W \neq \hat{W}] = \frac{1}{M}$.

3. Formally, we shall restrict $Q_{W,X^n,Y^n,\hat{W}} \in \mathcal{Q}$, where $\mathcal{Q}$ is the set of distributions that can be factorized as follows:

$$Q_{W,X^n,Y^n,\hat{W}} = Q_W Q_{X_1|W} Q_{Y_1} Q_{X_2|W,Y_1} Q_{Y_2|Y_1} \cdots Q_{X_n|W,Y^{n-1}} Q_{Y_n|Y^{n-1}} Q_{\hat{W}|Y^n} \tag{21.8}$$

$$P_{W,X^n,Y^n,\hat{W}} = P_W P_{X_1|W} P_{Y_1|X_1} P_{X_2|W,Y_1} P_{Y_2|X_2} \cdots P_{X_n|W,Y^{n-1}} P_{Y_n|X_n} P_{\hat{W}|Y^n} \tag{21.9}$$

Verify that $W \perp\!\!\!\perp W$ under Q: W and $\hat{W}$ are d-separated by X^n.

Notice that in the graphical models, when removing $\nrightarrow$ we also added the directional links between the Y_is, these links serve to maximally preserve the dependence relationships between variables when $\nrightarrow$ are removed, so that Q is the "closest" to P while $W \perp\!\!\!\perp W$ is satisfied.

Now we have that for $Q \in \mathcal{Q}$, $d(1 - \epsilon \| \frac{1}{M}) \leq D(P\|Q)$, in order to obtain the least upper bound, in Lemma 21.1 we shall show that:

$$\begin{aligned}\inf_{Q\in\mathcal{Q}} D(P_{W,X^n,Y^n,\hat{W}}\|Q_{W,X^n,Y^n,\hat{W}}) &= \sum_{k=1}^{n} I(X_k;Y_k|Y^{k-1})\\ &= \sum_{k=1}^{n} \mathbb{E}_{Y^{k-1}}[I(P_{X_k|Y^{k-1}},P_{Y|X})]\\ &\le \sum_{k=1}^{n} I(\mathbb{E}_{Y^{k-1}}[P_{X_k|Y^{k-1}}],P_{Y|X}) \quad \text{(concavity of } I(P_X,P_{Y|X}) \text{ in } P_X)\\ &= \sum_{k=1}^{n} I(P_{X_k},P_{Y|X})\\ &\le nC_i.\end{aligned}$$

Following the same procedure as in (a) we have

$$h(\epsilon)+\bar{\epsilon}\log M \le nC_i \Rightarrow \log M \le \frac{nC+h(\epsilon)}{1-\epsilon} \Rightarrow C_{fb,\epsilon} \le \frac{C}{1-\epsilon} \Rightarrow C_{fb} \le C.$$

4. Notice that the above proof is also valid even when cost constraint is present.

□

Lemma 21.1.

$$\inf_{Q\in\mathcal{Q}} D(P_{W,X^n,Y^n,\hat{W}}\|Q_{W,X^n,Y^n,\hat{W}}) = \sum_{k=1}^{n} I(X_k;Y_k|Y^{k-1}) \qquad (21.10)$$
$$(\triangleq \vec{I}(X^n;Y^n),\ \textit{\textbf{directed information}})$$

Proof. By chain rule, we can show that the minimizer $Q\in\mathcal{Q}$ must satisfy the following equalities:

$$\begin{aligned}Q_{X,W} &= P_{X,W},\\ Q_{X_k|W,Y^{k-1}} &= P_{X_k|W,Y^{k-1}}, \quad \text{(check!)}\\ Q_{\hat{W}|Y^n} &= P_{W|Y^n}.\end{aligned}$$

and therefore

$$\begin{aligned}&\inf_{Q\in\mathcal{Q}} D(P_{W,X^n,Y^n,\hat{W}}\|Q_{W,X^n,Y^n,\hat{W}})\\ &= D(P_{Y_1|X_1}\|Q_{Y_1}|X_1)+D(P_{Y_2|X_2,Y_1}\|Q_{Y_2|Y_1}|X_2,Y_1)+\cdots+D(P_{Y_n|X_n,Y^{n-1}}\|Q_{Y_n|Y^{n-1}}|X_n,Y^{n-1})\\ &= I(X_1;Y_1)+I(X_2;Y_2|Y_1)+\cdots+I(X_n;Y_n|Y^{n-1})\end{aligned}$$

□

21.3 When is feedback really useful?

Theorems 21.1 and 21.2 state that feedback does not improve communication rate neither asymptotically nor for moderate blocklengths. In this section, we shall examine three cases where feedback turns out to be very useful.

21.3.1 Code with very small (e.g. zero) error probability

Theorem 21.3 (Shannon '56). *For any DMC $P_{Y|X}$,*

$$C_{fb,0} = \max_{P_X} \min_{y\in\mathcal{B}} \log \frac{1}{P_X(S_y)} \tag{21.11}$$

where

$$S_y = \{a \in \mathcal{A} : P_{Y|X}(y|a) > 0\}$$

denotes the set of input symbols that can lead to the output symbol y.

Note: For stationary memoryless channel,

$$C_0 \overset{\text{def.}}{\le} C_{fb,0} \overset{\text{def.}}{\le} C_{fb} = \lim_{\epsilon\to 0} C_{fb,\epsilon} \overset{\text{Thm 21.1}}{=} C = \lim_{\epsilon\to 0} C_\epsilon \overset{\text{Shannon}}{=} C_i = \sup_{P_X} I(X;Y)$$

All capacity quantities above are defined with (fixed-length) block codes.

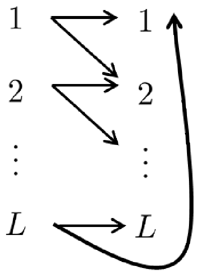
Observations:

1. In DMC for both zero-error capacities (C_0 and $C_{fb,0}$) only the support of the transition matrix $P_{Y|X}$, i.e., whether $P_{Y|X}(b|a) > 0$ or not, matters. The value of $P_{Y|X}(b|a) > 0$ is irrelevant. That is, C_0 and $C_{fb,0}$ are functions of a bipartite graph between input and output alphabets. Furthermore, the C_0 (but not $C_{fb,0}$!) is a function of the *confusability graph* – a simple undirected graph on $\mathcal{A}$ with $a \neq a'$ connected by an edge iff $\exists b \in \mathcal{B}$ s.t. $P_{Y|X}(b|a)P_{Y|X}(b|a') > 0$.

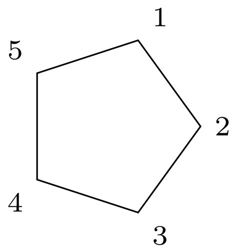
2. That $C_{fb,0}$ is not a function of the confusability graph alone is easily seen from comparing the polygon channel (next remark) with $L = 3$ (for which $C_{fb,0} = \log\frac{3}{2}$) and the useless channel with $\mathcal{A} = \{1,2,3\}$ and $\mathcal{B} = \{1\}$ (for which $C_{fb,0} = 0$). Clearly in both cases confusability graph is the same – a triangle.

3. Usually C_0 is very hard to compute, but $C_{fb,0}$ can be obtained in closed form as in (21.11).

   Example: (Polygon channel)



Bipartite graph



Confusability graph

- Zero-error capacity C_0:
  - $L = 3$: $C_0 = 0$
  - $L = 5$: $C_0 = \frac{1}{2}\log 5$ (Shannon '56-Lovasz '79).
    Achievability:
    a) blocklength one: $\{1,3\}$, rate = 1 bit.
    b) blocklength two: $\{(1,1),(2,3),(3,5),(4,2),(5,4)\}$, rate = $\frac{1}{2}\log 5$ bit – optimal!

  - $L = 7$: $3/5 \log 7 \le C_0 \le \log 3.32$ (Exact value unknown to this day)
  - Even $L = 2k$: $C_0 = \log \frac{L}{2}$ for all k (Why? Homework.).
  - Odd $L = 2k+1$: $C_0 = \log \frac{L}{2} + o(1)$ as $k \to \infty$ (Bohman '03)

- Zero-error capacity with feedback (proof: exercise!)

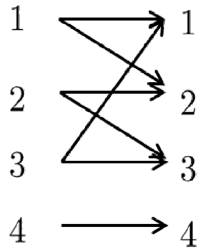
$$C_{fb,0} = \log \frac{L}{2}, \quad \forall L,$$

which can be strictly bigger than C_0.

4. Notice that $C_{fb,0}$ is not necessarily equal to $C_{fb} = \lim_{\epsilon\to 0} C_{fb,\epsilon} = C$. Here is an example when

$$C_0 < C_{fb,0} < C_{fb} = C$$

Example:



Then

$$\begin{aligned} C_0 &= \log 2 \\ C_{fb,0} &= \max_{\delta} -\log \max(\frac{2}{3}\delta, 1-\delta) && (P_X^* = (\delta/3, \delta/3, \delta/3, \bar{\delta})) \\ &= \log \frac{5}{2} > C_0 && (\delta^* = \frac{3}{5}) \end{aligned}$$

On the other hand, Shannon capacity $C = C_{fb}$ can be made arbitrarily close to $\log 4$ by picking the cross-over probability arbitrarily close to zero, while the confusability graph stays the same.

Proof of Theorem 21.3. 1. Fix any $(n, M, 0)$-code. Denote the confusability set of all possible messages that could have produced the received signal $y^t = (y_1, \ldots, y_t)$ for all $t = 0, 1, \ldots, n$ by:

$$E_t(y^t) \triangleq \{m \in [M] : f_1(m) \in S_{y_1}, f_2(m, y_1) \in S_{y_2}, \ldots, f_n(m, y^{t-1}) \in S_{y_t}\}$$

Notice that zero-error means no ambiguity:

$$\epsilon = 0 \Leftrightarrow \forall y^n \in \mathcal{B}^n, |E_n(y^n)| = 1 \text{ or } 0. \tag{21.12}$$

2. The key quantities in the proof are defined as follows:

$$\theta_{fb} = \min_{P_X} \max_{y \in \mathcal{B}} P_X(S_y),$$

$$P_X^* = \operatorname*{argmin}_{P_X} \max_{y \in \mathcal{B}} P_X(S_y)$$

By definition, we have

$$\forall P_X, \exists y \in \mathcal{B}, \text{ such that } P_X(S_y) \ge \theta_{fb} \tag{21.13}$$

Notice the minimizer distribution P_X^* is usually not the caid in the usual sense. This definition also sheds light on how the encoding and decoding should be proceeded and serves to lower bound the uncertainty reduction at each stage of the decoding scheme.

3. "≤" (converse): Let P_{X^n} be he joint distribution of the codewords. Denote $E_0 = [M]$ – original message set.

   $\underline{t=1}$: For P_{X_1}, by (21.13), $\exists y_1^*$ such that:

   $$P_{X_1}(S_{y_1^*}) = \frac{|\{m : f_1(m) \in S_{y_1^*}\}|}{|\{m \in [M]\}|} = \frac{|E_1(y_1^*)|}{|E_0|} \ge \theta_{fb}.$$

   $\underline{t=2}$: For $P_{X_2|X_1 \in S_{y_1^*}}$, by (21.13), $\exists y_2^*$ such that:

   $$P_{X_2}(S_{y_2^*}|X_1 \in S_{y_1^*}) = \frac{|\{m : f_1(m) \in S_{y_1^*}, f_2(m, y_1^*) \in S_{y_2^*}\}|}{|\{m : f_1(m) \in S_{y_1^*}\}|} = \frac{|E_2(y_1^*, y_2^*)|}{|E_1(y_1^*)|} \ge \theta_{fb},$$

   $\underline{t=n}$: Continue the selection process up to y_n^* which satisfies that:

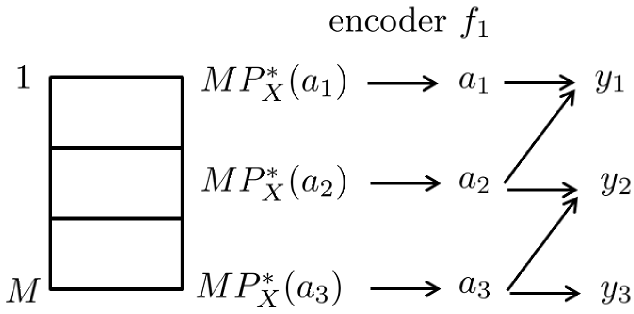
   $$P_{X_n}(S_{y_n^*}|X_k \in S_{y_k^*} \text{ for } k = 1, \dots, n-1) = \frac{|E_n(y_1^*, \dots, y_n^*)|}{|E_{n-1}(y_1^*, \dots, y_{n-1}^*)|} \ge \theta_{fb}.$$

   Finally, by (21.12) and the above selection procedure, we have

   $$\begin{aligned} \frac{1}{M} &\ge \frac{|E_n(y_1^*, \dots, y_n^*)|}{|E_0|} \ge \theta_{fb}^n \\ \Rightarrow M &\le -n \log \theta_{fb} \\ \Rightarrow C_{fb,0} &\le -\log \theta_{fb} \end{aligned}$$

4. "≥" (achievability)

   Let's construct a code that achieves $(M, n, 0)$.



   The above example with $|\mathcal{A}| = 3$ illustrates that the encoder f_1 partitions the space of all messages to 3 groups. The encoder f_1 at the first stage encodes the groups of messages into a_1, a_2, a_3 correspondingly. When channel outputs y_1 and assume that $S_{y_1} = \{a_1, a_2\}$, then the decoder can eliminate a total number of $MP_X^*(a_3)$ candidate messages in this round. The "confusability set" only contains the remaining $MP_X^*(S_{y_1})$ messages. By definition of P_X^* we know that $MP_X^*(S_{y_1}) \le M\theta_{fb}$. In the second round, f_2 partitions the remaining messages into three groups, send the group index and repeat.

By similar arguments, each interaction reduces the uncertainty by a factor of *at least* θ_{fb}. After n iterations, the size of "confusability set" is upper bounded by $M\theta_{fb}^n$, if $M\theta_{fb}^n \le 1$,[2] then zero error probability is achieved. This is guaranteed by choosing $\log M = -n \log \theta_{fb}$. Therefore we have shown that $-n\log\theta_{fb}$ bits can be reliably delivered with $n + O(1)$ channel uses with feedback, thus

$$C_{fb,0} \ge -\log\theta_{fb}$$

□

21.3.2 Code with variable length

Consider the example of BEC(δ) with feedback, send k bits in the following way: repeat sending each bit until it gets through the channel correctly. The expected number of channel uses for sending k bits is given by

$$l = \mathbb{E}[n] = \frac{k}{1-\delta}$$

We state the result for **variable-length feedback** (VLF) code without proof:

$$\log M_{VLF}^*(l,0) \ge lC$$

Notice that compared to the scheme without feedback, there is the improvement of $\sqrt{nV}Q^{-1}(\epsilon)$ in the order of $O(\sqrt{n})$, which is stronger than the result in Theorem 21.2.

This is also true in general:

$$\log M_{VLF}^*(l,\epsilon) = \frac{lC}{1-\epsilon} + O(\log l)$$

Example: For BSC(0.11), without feedback, $n = 3000$ is needed to achieve 90% of capacity C, while with VLF code $l = \mathbb{E}n = 200$ is enough to achieve that.

21.3.3 Code with variable power

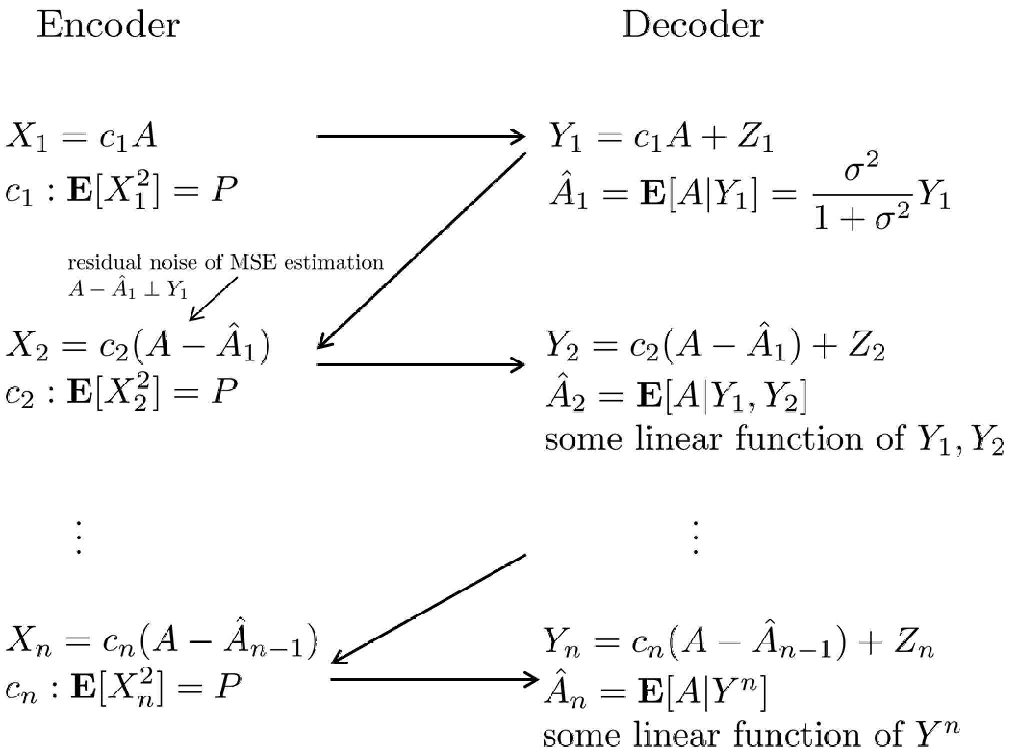
Elias' scheme of sending a number A drawn from a Gaussian distribution $\mathcal{N}(0, \operatorname{Var} A)$ with linear processing.

AWGN setup:

$$Y_k = X_k + Z_k, \quad Z_k \sim \mathcal{N}(0,\sigma^2) \text{ i.i.d.}$$
$$\mathbb{E}[X_k^2] \le P, \quad \text{power constraint in expectation}$$

Note: If we insist the codeword satisfies power constraint almost surely instead on average, i.e., $\sum_{k=1}^n X_k^2 \le nP$ a.s., then the scheme below does not work!

[2] Some rounding-off errors need to be corrected in a few final steps (because P_X^* may not be closely approximable when very few messages are remaining). This does not change the asymptotics though.


Encoder
Decoder
$X_1 = c_1 A$
$c_1 : \mathbf{E}[X_1^2] = P$
$Y_1 = c_1 A + Z_1$
$\hat{A}_1 = \mathbf{E}[A|Y_1] = \frac{\sigma^2}{1+\sigma^2} Y_1$
residual noise of MSE estimation
$A - \hat{A}_1 \perp Y_1$
$X_2 = c_2(A - \hat{A}_1)$
$c_2 : \mathbf{E}[X_2^2] = P$
$Y_2 = c_2(A - \hat{A}_1) + Z_2$
$\hat{A}_2 = \mathbf{E}[A|Y_1, Y_2]$
some linear function of Y_1, Y_2
$X_n = c_n(A - \hat{A}_{n-1})$
$c_n : \mathbf{E}[X_n^2] = P$
$Y_n = c_n(A - \hat{A}_{n-1}) + Z_n$
$\hat{A}_n = \mathbf{E}[A|Y^n]$
some linear function of Y^n

According to the orthogonality principle of the mininum mean-square estimation (MMSE) of A at receiver side in every step:

$$A = \hat{A}_n + N_n, \quad N_n \perp Y^n.$$

Morever, since all operations are lienar and everything is jointly Gaussian, $N_n \perp\!\!\!\perp Y^n$. Since $X_n \propto N_{n-1} \perp\!\!\!\perp Y^{n-1}$, the codeword we are sending at each time slot is independent of the history of the channel output (”innovation”), in order to maximize information transfer.

Note that $Y^n \to \hat{A}_n \to A$, and the optimal estimator $\hat{A}_n$ (a linear combination of Y^n) is a sufficient statistic of Y^n for A under Gaussianity. Then

$$\begin{aligned} I(A;Y^n) =& I(A;\hat{A}_n, Y^n) \\ &= I(A;\hat{A}_n) + I(A;Y^n|\hat{A}_n) \\ &= I(A;\hat{A}_n) \\ &= \frac{1}{2}\log\frac{\mathrm{Var}(A)}{\mathrm{Var}(N_n)}. \end{aligned}$$

where the last equality uses the fact that N follows a normal distribution. $\mathrm{Var}(N_n)$ can be computed directly using standard linear MMSE results. Instead, we determine it information theoretically: Notice that we also have

$$\begin{aligned} I(A;Y^n) &= I(A;Y_1) + I(A;Y_2|Y_1) + \cdots + I(A;Y_n|Y^{n-1}) \\ &= I(X_1;Y_1) + I(X_2;Y_2|Y_1) + \cdots + I(X_n;Y_n|Y^{n-1}) \\ &\overset{\text{key}}{=} I(X_1;Y_1) + I(X_2;Y_2) + \cdots + I(X_n;Y_n) \\ &= n\frac{1}{2}\log(1+P) = nC \end{aligned}$$

Therefore, with Elias' scheme of sending $A \sim \mathcal{N}(0, \operatorname{Var} A)$, after the n-th use of the AWGN(P) channel with feedback,

$$\operatorname{Var} N_n = \operatorname{Var}(\hat{A}_n - A) = 2^{-2nC} \operatorname{Var} A = \left(\frac{P}{P+\sigma^2}\right)^n \operatorname{Var} A,$$

which says that the reduction of uncertainty in the estimation is exponential fast in n.

Schalkwijk-Kailath: Elias' scheme can also be used to send digital data.

Let $W \sim$ uniform on M-PAM constellation in $\in [-1,1]$, i.e., $\{-1, -1+\frac{2}{M}, \cdots, -1+\frac{2k}{M}, \cdots, 1\}$. In the very first step W is sent (after scaling to satisfy the power constraint):

$$X_0 = \sqrt{P} W, \quad Y_0 = X_0 + Z_0$$

Since Y_0 and X_0 are both known at the encoder, it can compute Z_0. Hence, to describe W it is sufficient for the encoder to describe the noise realization Z_0. This is done by employing the Elias' scheme ($n-1$ times). After $n-1$ channel uses, and the MSE estimation, the equivalent channel output:

$$\widetilde{Y}_0 = X_0 + \widetilde{Z}_0, \quad \operatorname{Var}(\widetilde{Z}_0) = 2^{-2(n-1)C}$$

Finally, the decoder quantizes $\widetilde{Y}_0$ to the nearest PAM point. Notice that

$$\begin{aligned}
\epsilon \le \mathbb{P}\left[|\widetilde{Z}_0| > \frac{1}{2M}\right] = \mathbb{P}\left[2^{-(n-1)C}|Z| > \frac{\sqrt{P}}{2M}\right] = 2Q\left(\frac{2^{(n-1)C}\sqrt{P}}{2M}\right) \\
\Rightarrow \log M \ge (n-1)C + \log\frac{\sqrt{P}}{2} - \log Q^{-1}(\frac{\epsilon}{2}) \\
= nC + O(1).
\end{aligned}$$

Hence if the rate is strictly less than capacity, the error probability decays doubly exponentially fast as n increases. More importantly, we gained an $\sqrt{n}$ term in terms of $\log M$, since for the case without feedback we have

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

Example: $P = 1 \Rightarrow$ channel capacity $C = 0.5$ bit per channel use. To achieve error probability 10^{-3}, $2Q\left(\frac{2^{(n-1)C}}{2M}\right) \approx 10^{-3}$, so $\frac{e^{(n-1)C}}{2M} \approx 3$, and $\frac{\log M}{n} \approx \frac{n-1}{n}C - \frac{\log 8}{n}$. Notice that the capacity is achieved to within 99% in as few as $n = 50$ channel uses, whereas the best possible block codes without feedback require $n \approx 2800$ to achieve 90% of capacity.

Take-away message:

Feedback is best harnessed with *adaptive* strategies. Although it does not increase capacity under block coding, feedback greatly boosts reliability as well as reduces coding complexity.