EECS 290S: Network Information Flow **Fall 2008**

Lecture 16 — October 21,23

Lecturer: Anant Sahai *Scribe: Se Yong Park*

So far, we have focused on modeling channel interaction and the resulting interference. The simplest model that captures this phenomenon was a multi-level binary deterministic model, which gave us an understanding of information flow in networks. However, it is conspicuously missing the role of noise in communication systems. Even so, this deterministic model is reasonable for the high-SNR case since the gap to a Gaussian model is bounded by a constant. However, when it comes to low-SNR, it is unclear how good/bad the deterministic models are without taking a closer look at the role of noise. We start by reviewing the Point-to-Point Channel, and then taking a close look at the Multiple Access Channel (MAC) and Broadcast Channel, the simplest multiterminal models for communication. These two lectures cover the point-to-point case and the MAC.

16.1 Finishing up on earlier lectures

Before we study the noisy channel, a few variations on the deterministic model should be discussed. As a natural generalization of a deterministic broadcasting network with two destinations, we can think of an arbitrary number of destination. For simplicity, suppose there was one common message whose rate is R_0 destined to every destination exists. In addition, there are private messages at rate R_i that are intended for the i-th destination and nobody else cares about them. A converse for the supportable rates can be easily derived as

$$R_0 + \sum_{i \in S} R_i \leq c(S_i; \{t_i\}_{i \in S}) \quad for \quad \forall S \subset 1, 2, ..., K$$
$$, where \quad K := \quad number \quad of \quad destinations$$

The remaining question is whether this bound is achievable in general. The answer is we don't know. The proof for two destinations relied on the inclusion relationship between two vector spaces that were generated by the global edge transfer matrix from a destination to dummy destinations. For more than 3 destinations, we can not expect such a nice property to hold.

Another generalization is a multiple-source multiple-destination case, where sources cannot cooperate each other and a message to one destination might be interference from the perspective of another. One interesting result for such networks is that nonlinear codes can beat all possible linear codes [Ken Zeger et al., 'Insufficiency of Linear Network Codes', ISIT Sep 2005]. This result hints that the multiple-source multiple-destination network might elude the set of tools that have been used so far.

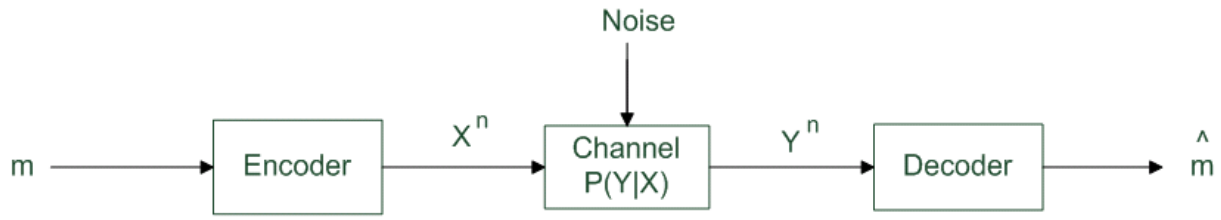

Figure 16.1. Point-to-Point Communication System

16.2 Noisy Channel

16.2.1 Point-to-Point Channel (Discrete-alphabet Channel)

We consider the communication system as shown in Fig 16.1. The message w is uniformly drawn from the set $\{1, 2, ..., 2^{nR}\}$. The encoder maps the message to a particular codeword $X^n(w)$, which is transmitted through the channel. The received signal at the decoder is randomly generated though a discrete memoryless channel specified by transition probability $P(Y|X)$. Finally, the decoder tries to estimate the transmitted message $\hat{w}$ based on the noisy received signal Y^n. Our objective is finding the threshold rate R such that the probability of error can be made asymptotically close to zero as n goes to infinity. This threshold rate can be calculated using Shannon's famous theorem.

Theorem 16.1. *Discrete memoryless channel that is defined by $P(Y|X)$ has a capacity $C = max_{p(x)} I(X;Y) = max_{p(x)} H(Y) - H(Y|X)$.*

This theorem was proved in 229A and you can read Cover and Thomas if you are a bit rusty on this. Here, we just review the key idea. To show that this rate is achievable, we use the trick of increasing the block length n to be large enough. This trick has appeared twice so far in this class. The first was in the proof of the existence of a linear network code in the wireline network. In that proof, the analogy of "large n" was allowing a large enough field size to guarantee the existence of a non-zero root to the appropriate polynomial or a guaranteed solution in Jaggi's algorithm. The second time was in the proof of the existence of a random network code for deterministic linear relay networks. In that proof, rather than trying to construct a particular codebook, we averaged the probability of error over all possible linear codes using a uniform distribution. The instant the probability of error went below 1, we knew there had to exist a zero-error code since the problem was deterministic. In either of these previous cases, the interpretation of large n was that the "packet size" of grains of information was made appropriately large.

In Shannon's theorem, the role of large n is similar with the additional interpretation of imposing typical behavior to the long noise sequence. Given the typical behavior of noise, a random codebook (Table 16.1) is enough to achieve the point-to-point capacity.

The intuitive explanation of the theorem is clear. By using a random codeword drawn iid according to $P(X)$, the number of typical outputs is equal to $2^{nH(Y)}$. Here, $P(Y)$ is

	x_1	x_2	...	x_n
$message0$	$x_1(0)$	$x_2(0)$	...	$x_n(0)$
$message1$	$x_1(1)$	$x_2(1)$	...	$x_n(1)$
.			...	
.				
.			...	
$message2^{nR}-1$	$x_1(2^{nR}-1)$	$x_2(2^{nR}-1)$	...	$x_n(2^{nR}-1)$

Table 16.1. Random Codebook (every component is random variable with pmf P(X))

determined by $P(X)$ and the channel law $P(Y|X)$. Among these typical outputs, any one particular codeword can induce $2^{nH(Y|X)}$ jointly typical received sequences. Therefore, we can conclude that the maximum number of codewords that does not cause any error is simply the ratio of the two numbers, $frac2^{nH(Y)}2^{nH(Y|X)} = 2^{nI(Y;X)}$.

16.2.2 The Gaussian Point-to-Point Channel

First, we have to recognize that an additional constraint is required to make the problem reasonable. If we allow infinite power channel inputs (meaning that the input is entirely unconstrained), an arbitrary number of bits can be sent with an arbitrarily low probability of error. This means that infinite rates are achievable by one channel use, which is not interesting because it is completely aphysical. Therefore, we need to impose either an amplitude or power constraint on the channel inputs X. It turns out that a power constraint is easier to work with when the bandwidth is finite and so under this constraint, the Gaussian channel capacity is given by:

Theorem 16.2. *An additive Gaussian memoryless channel with noise variance N has a capacity*
$C = max_{p(x)s.t.E[X^2]\leq P} I(X;Y) = \max_{p(x)s.t.E[X^2]\leq P} h(Y) - h(Y|X) = \frac{1}{2}\log(1+\frac{P}{N})$ *if the inputs are constrained to have power no greater than P on average.*

Compared to Theorem 16.1, the only main difference is that we've changed from regular entropy to differential entropy. This differential entropy is defined by an integral instead of a summation and involves the probability density functions rather than the probability mass functions. One caveat that should be mentioned is that it would have been mildly improper to write the mutual information be written in the other way because it may not mathematically well-defined. Since random variable X can be a discrete or mixed random variable, $h(X)$ and $h(X|Y)$ may not exist.

The random code construction is also similar to discrete case. Each codeword should be picked following i.i.d. using the marginal distribution $N(0,P)$. In addition, the same intuition as the discrete channel capacity can be also applied to the Gaussian channel with minor revisions. Here, $2^{nh(Y)}$ stands for total effective volume of the Gaussian ball at the output of channel. $2^{h(Y|X)}$ measures the effective volume of the disturbance that can be

caused by noise. The maximum number of codewords that can be transmitted without error is thus given by $2^I(X;Y)$. We can also check this intuition analytically.

$$C = \frac{1}{2}\log(1+\frac{P}{N}) = \frac{1}{2}\log(\frac{P+N}{N})$$

The numerator inside the log, $P+N$, is the total power of the received signal Y. The denominator inside the log, N, is the power of the noise alone.

It is important to be aware that in the literature, two major variations of this codebook are used.

(1) Gaussian codebook : Each codeword component is drawn from $N(0,P)$, and so it can be any value. However the average power of all the codewords is equal to P with probability approaching 1 by the law of large numbers.

(2) Uniform codebook on the shell: The laws of large numbers allow us to instead use a different codebook that is uniformly picked on the shell of the sphere of radius $\sqrt{P}$. Since each codeword is lies on the shell, the maximum codeword power is guaranteed to be P. This allows us to change the average power constraint to a maximum power constraint without difficulty.

In addition, practical codebook constructions are usually made by appropriate mappings of lattice points to appear uniformly distributed upon the sphere.

16.2.3 Achievability for the Multiple Access Channel

The progress in this class proceeds in stages:
(1) No channel interaction : wireline multicast
(2) Interaction, but no noise : wireless deterministic model
(3) Interaction and noise : wireless Gaussian model

In this section, we will start to talk about (3). The MAC communication system (Fig 16.2) is similar to the point-to-point communication system. The difference is that multiple sources want to transmit their independent messages to a common destination. We want to find a feasible pair of rates (R_1, R_2) so that there exists a codebook $(2^{nR_1}, 2^{nR_2}, n)$ achieving asymptotically zero probability of error as n increases. The channel is again memoryless, which means that $P(Y|X_1, X_2)$ only depends on current pair of inputs. The achievable rate region is given in the following theorem.

Theorem 16.3. *The capacity of the multiple-access channel is the closure of convex hull of all (R_1, R_2) satisfying*

$$R_1 + R_2 \leq I(X_1, X_2; Y) \tag{16.1}$$
$$R_1 \leq I(X_1; Y|X_2) \tag{16.2}$$
$$R_2 \leq I(X_2; Y|X_1) \tag{16.3}$$

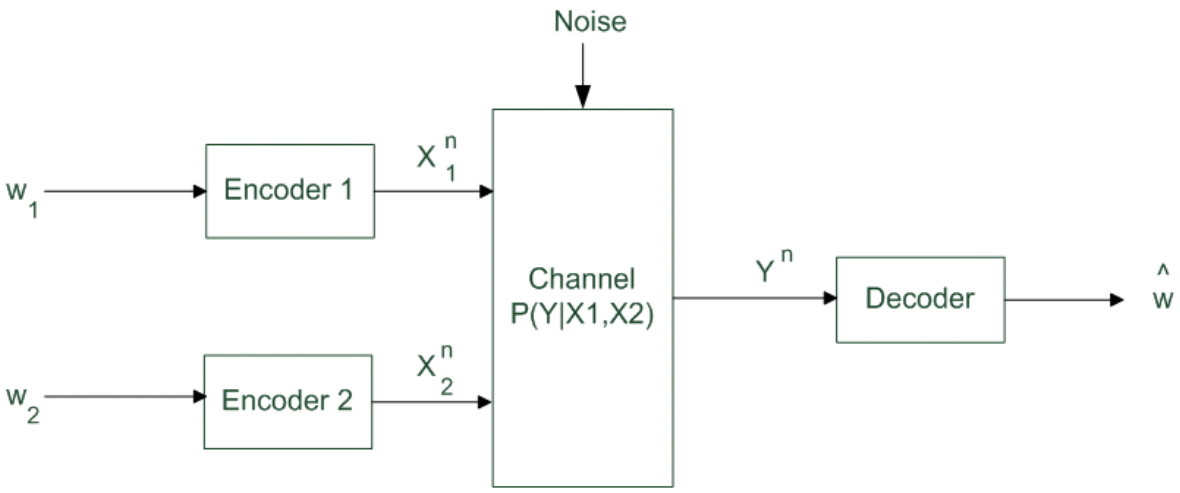

Figure 16.2. Multiple Access Channel Communication System

over independent X_1, X_2 *i.e.* $P(X_1, X_2) = P(X_1)P(X_2)$

(16.1) looks reasonable, since even if encoders are allowed to collaborate $I(X_1, X_2; Y)$ is the maximum rate one can expect. (16.3) can be explained by a genie approach. If we assume x_1 is magically known to the decoder, the reasonable strategy for a decoder is to "subtract" x_1 from the received signal Y and this gives you the bound $I(X_2; Y|X_1)$.

By applying above theorem to the additive Gaussian case, we can find a more interesting result in which the figurative subtraction is interpreted literally.

Theorem 16.4. *The capacity of the additive Gaussian multiple-access channel is all* (R_1, R_2) *satisfying*

$$R_1 + R_2 \leq \frac{1}{2}\log(1 + \frac{P_1 + P_2}{N}) \tag{16.4}$$

$$R_1 \leq \frac{1}{2}\log(1 + \frac{P_1}{N}) \tag{16.5}$$

$$R_2 \leq \frac{1}{2}\log(1 + \frac{P_2}{N}) \tag{16.6}$$

This time, (16.5) and (16.6) look trivial because the conditions are consistent with a point-to-point channel. (16.4) is also not that surprising noticing that the combined power of X_1 and X_2 cannot be bigger than $P_1 + P_2$ (since the independent message assumption prevents a correlation between X_1 and X_2). Expanding (16.4) gives us an even more interesting interpretation.

$$\frac{1}{2}\log(1 + \frac{P_1 + P_2}{N}) = \frac{1}{2}\log(\frac{N + P_1 + P_2}{N + P_2}\frac{N + P_2}{N}) = \frac{1}{2}\log(1 + \frac{P_1}{N + P_2}) + \frac{1}{2}\log(1 + \frac{P_2}{N})$$

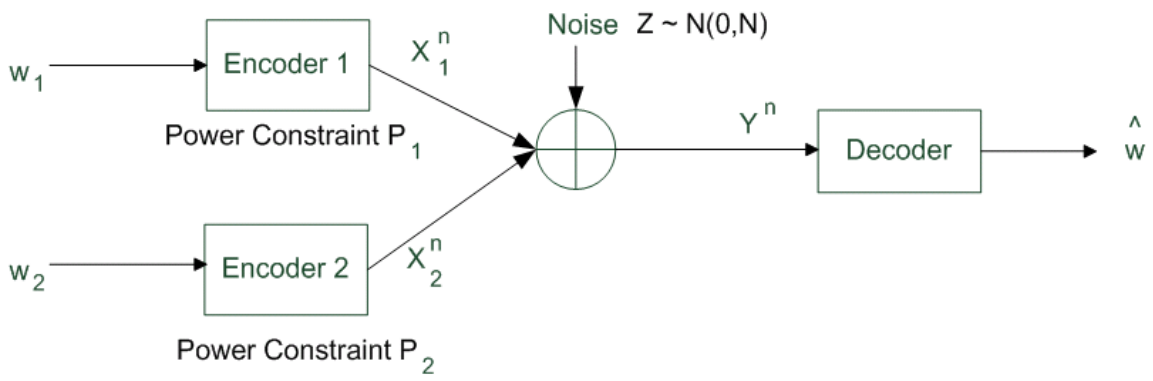

Figure 16.3. Gaussian Multiple Access Channel Communication System

The first term's $\frac{P_1}{P_2+N}$, is the signal-to-interference-plus-noise ratio. The second term is equal to the point-to-point Gaussian capacity. These observations suggest that successive interference cancellation is optimal. That is, first decode the codeword x_1 viewing considering x_2 as Gaussian noise and then subtract the resulting x_1 from y before further decoding x_2 (Fig16.4). Using this decoding method, we can achieve the corner points of the pentagonal rate region depicted in Fig16.5.

All points within the pentagon can be obtained by time-sharing between the four interesting corners and the boring zero-point.

16.2.4 Converse Multiple Access Channel

Now we have an achievability theorem. The remaining question is how good the achievable region is and it turns out to be tight.

Theorem 16.5. *(Converse for rate region of MAC) In A Multiple Access Channel, if (R_1, R_2) is feasible, then there exists X_1, X_2, and Q such that*

$$R_1 \leq I(X_1; Y|X_2, Q) \tag{16.7}$$
$$R_2 \leq I(X_2; Y|X_1, Q) \tag{16.8}$$
$$R_1 + R_2 \leq I(X_1, X_2; Y|Q) \tag{16.9}$$

Q is defined on the finite alphabet $\{1, 2, 3, 4\}$ and Q, X_1, X_2, Y obey a joint distribution that can be factored as $P(q)P(x_1|q)P(x_2|q)P(y|x_1, x_2)$.

Proof: *The rate is achievable so we have a code for some n, which induces a joint distribution:*

$$P(W_1, W_2, X_1^n, X_2^n, Y^n) = \frac{1}{2^{nR_1}} \frac{1}{2^{nR_2}} P(X_1^n|W_1)P(X_2^n|W_2) \prod_{i=1}^{n} P(Y_i|X_{1i}, X_{2i}) \tag{16.10}$$

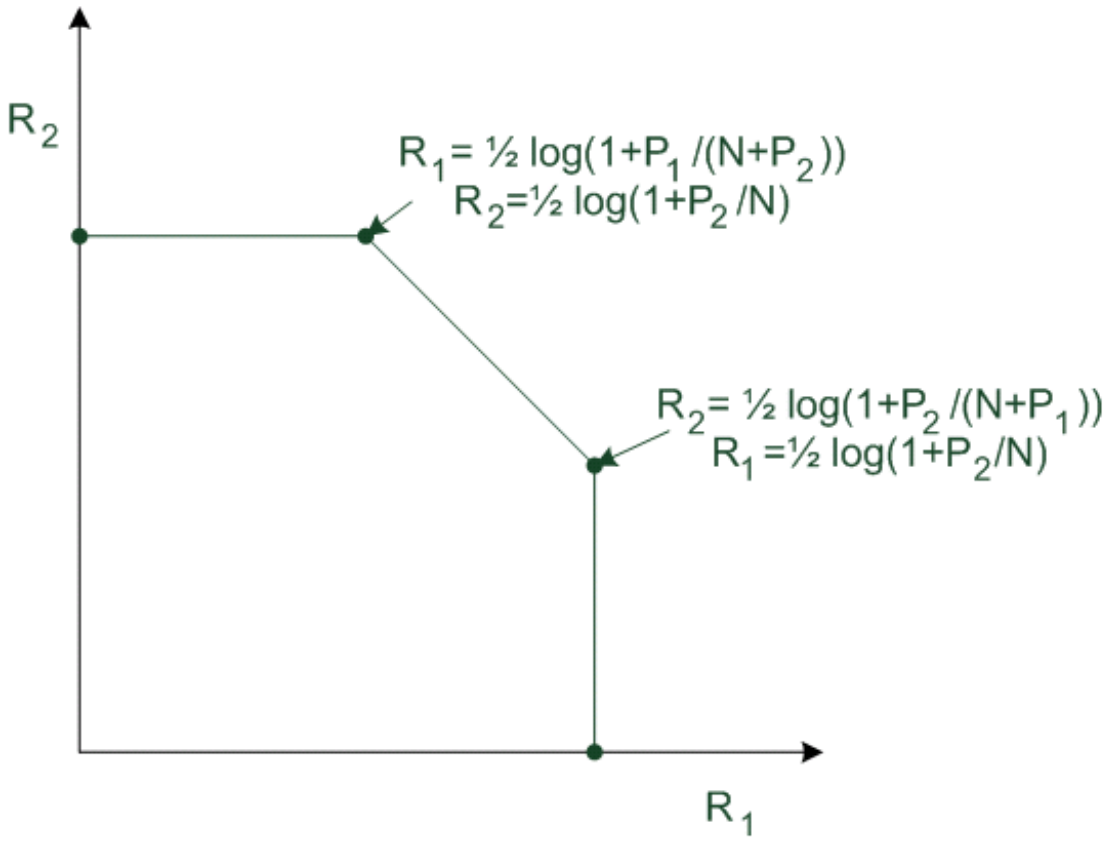

Figure 16.4. Achievable region of MAC Gaussian Channel

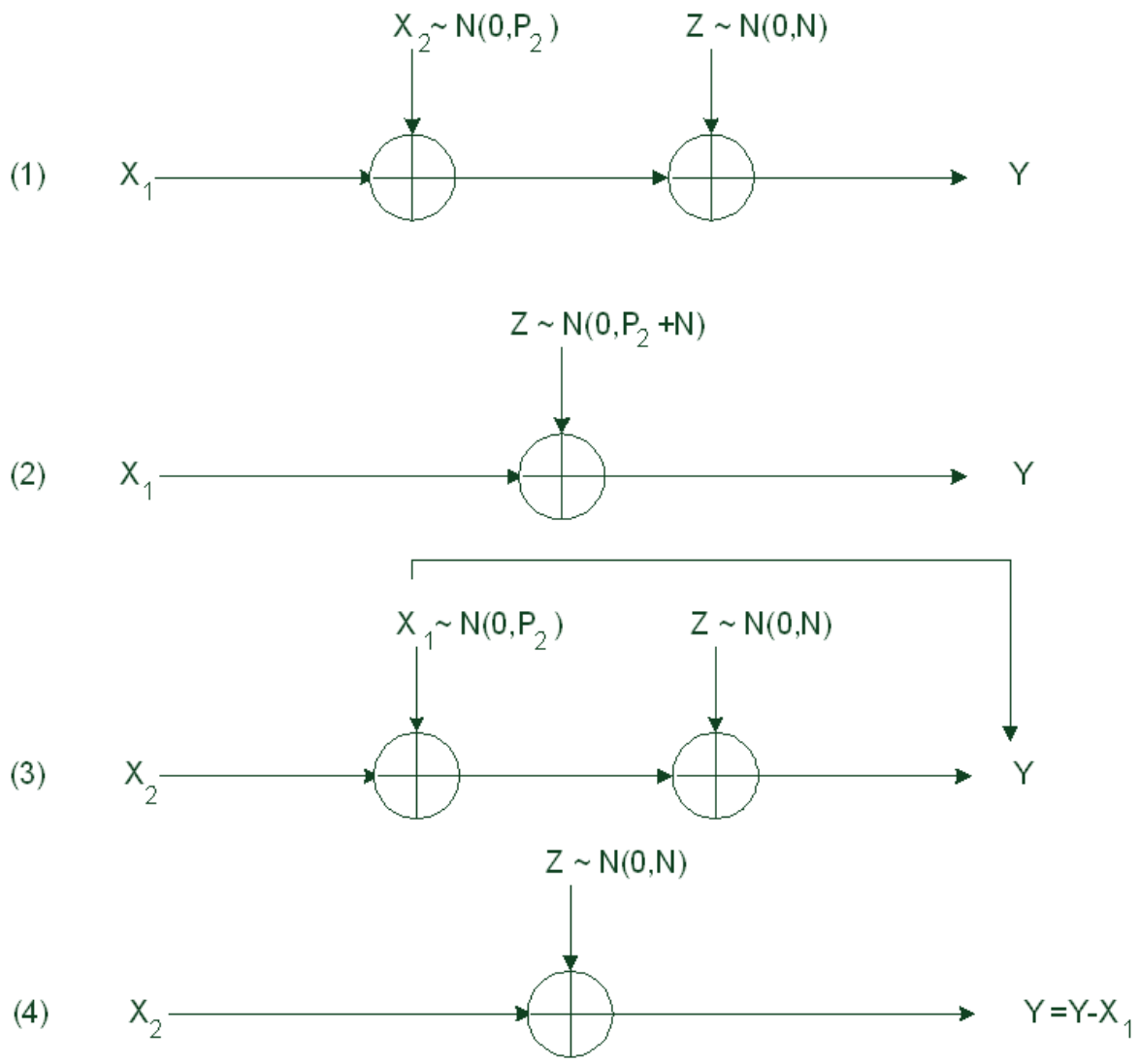

Figure 16.5. Successive Interference Cancellation

Since P_e is arbitrary small, by Fano's inequality we can find a sequence of ϵ_n which converges to zero so that

$$H(W_1, W_2|Y^n) \leq n\epsilon_n. \tag{16.11}$$

Now we can bound R_1.

$$\begin{aligned}
nR_1 = H(W_1) &= I(W_1; Y^n) + H(W_1|Y^n) && (16.12)\\
&\leq I(W_1; Y^n) + n\epsilon_n && (16.13)\\
&\leq I(X_1^n(W_1); Y^n) + n\epsilon_n && (16.14)\\
&= H(X_1^n(W_1)) - H(X_1^n(W_1)|Y^n) + n\epsilon_n && (16.15)\\
&= H(X_1^n(W_1)|X_2^n(W_2)) - H(X_1^n(W_1)|Y^n) + n\epsilon_n && (16.16)\\
&\leq H(X_1^n(W_1)|X_2^n(W_2)) - H(X_1^n(W_1)|Y^n, X_2^n(W_2)) + n\epsilon_n && (16.17)\\
&= I(X_1^n(W_1); Y^n|X_2^n(W_2)) + n\epsilon_n && (16.18)\\
&= H(Y^n|X_2^n(W_2)) - H(Y^n|X_2^n(W_2), X_1^n(W_1)) + n\epsilon_n && (16.19)\\
&= \sum_{i=1}^{n} H(Y_i|X_2^n, Y_1, \ldots, Y_{i-1}) - H(Y_i|X_2^n, X_1^n, Y_1, \ldots, Y_{i-1}) + n\epsilon_n && (16.20)\\
&= \sum_{i=1}^{n} H(Y_i|X_2^n, Y_1, \ldots, Y_{i-1}) - H(Y_i|X_{2i}, X_{1i}) + n\epsilon_n && (16.21)\\
&\leq \sum_{i=1}^{n} H(Y_i|X_{2i}) - H(Y_i|X_{2i}, X_{1i}) + n\epsilon_n && (16.22)\\
&= \sum_{i=1}^{n} I(X_{1i}; Y_i|X_{2i}) + n\epsilon_n && (16.23)
\end{aligned}$$

(16.13) : $H(W_1|Y^n) \leq H(W_1, W_2|Y^n) \leq \epsilon_n$ by (16.11)
(16.14) : Data Processing Inequality
(16.16) : W_1 and W_2 are independent, so $I(X_1^n(W_1)|X_2^n(W_2)) = 0$)
(16.17) : $H(X_1^n(W_1)|Y^n) \geq H(X_1^n(W_1)|Y^n, X_2^n(W_2))$ since conditioning decreases entropy
(16.20) : Entropy Chain Rule
(16.21) : Memoryless Property of Channel
(16.22) : Conditioning Decreases Entropy

By dividing both sides by n, we can get

$$\begin{aligned}
R_1 &\leq \frac{1}{n}[\sum_{i=1}^{n} I(X_{1i}; Y_i|X_{2i})] + \epsilon_n && (16.24)\\
&= \sum_{i=1}^{n} \frac{1}{n} I(X_{1i}; Y_i|X_{2i}) + \epsilon_n && (16.25)
\end{aligned}$$

By symmetry in the problem, we can get the same inequality for R_2.

$$R_1 \leq \sum_{i=1}^{n} \frac{1}{n} I(X_{1i}; Y_i | X_{2i}) + \epsilon_n \tag{16.26}$$

Moreover, we can bound $R_1 + R_2$ in a similar way.

$$n(R_1 + R_2) \leq I(X_1(W_1), X_2(W_2); Y^n) + n\epsilon_n \tag{16.27}$$
$$= H(Y^n) - H(Y^n | X_1(W_1), X_2(W_2)) + n\epsilon_n \tag{16.28}$$
$$= \sum_{i=1}^{n} H(Y_i | Y_1, \ldots, Y_{i-1}) - H(Y_i | Y_1, \ldots, Y_{i-1}, X_1(W_1), X_2 W_2) + n\epsilon_n \tag{16.29}$$
$$= \sum_{i=1}^{n} H(Y_i | Y_1, \ldots, Y_{i-1}) - H(Y_i | X_{1i}, X_{2i}) + n\epsilon_n \tag{16.30}$$
$$\leq \sum_{i=1}^{n} H(Y_i) - H(Y_i | X_{1i}, X_{2i}) + n\epsilon_n \tag{16.31}$$

(16.29):Entropy Chain Rule
(16.30):Memoryless Property of Channel
(16.31):Coditioning decreases entropy
By dividing both side by n,

$$R_1 + R_2 \leq [\sum_{i=1}^{n} \frac{1}{n} I(Y_i; X_{1i}, X_{2i})] + \epsilon_n \tag{16.32}$$

Let Q' be uniform over the large discrete alphabet $\{1, 2, \ldots, n\}$. Define X_1' such that

$$P(X_1' = x | Q = q) = P(X_{1,q} = x)$$

Similarly for X_2' and Y'.

Then, using the definition of Q'

$$R_1 \leq \sum_{i=1}^{n} \frac{1}{n} I(X_{1i}; Y_i | X_{2i}) + \epsilon_n = \sum_{i=1}^{n} I(X_{1i}; Y_i | X_{2i}) P(Q' = i) + \epsilon_n$$
$$= \sum_{q=1}^{n} I(X_{1q}; Y_q | X_{2q}, Q' = q) P(Q' = q) + \epsilon_n = I(X_1'; Y' | X_2', Q') + \epsilon_n \tag{16.33}$$

Since ϵ_n can be chosen to be arbitrary small, we can drop ϵ_n from the right-hand side.

$$R_1 \leq I(X_1'; Y' | X_2', Q') \tag{16.34}$$

Similarly, for R_1 and $R_1 + R_2$ we can get

$$R_2 \leq I(X_1'; Y' | X_2', Q') \tag{16.35}$$
$$R_1 + R_1 \leq I(X_1', X_2'; Y' | Q') \tag{16.36}$$

Here $|Q'|$ is defined on an alphabet of size n which is going to infinity. Even for a finite input alphabet size $|X|$, checking whether a rate pair (R_1, R_2) is feasible induces an infinite dimensional optimization problem. This motivates us to bound the cardinality of $|Q'|$. However, the tuple $(I(X_1';Y'|X_2',Q'), I(X_1';Y'|X_2',Q'), I(X_1',X_2';Y'|,Q'))$ can be written as,

$$(\sum_{q=1}^{|Q'|} I(X_1';Y'|X_2',Q'=q)P(Q'=q), \sum_{q=1}^{|Q'|} I(X_2';Y'|X_1',Q'=q)P(Q'=q), \sum_{q=1}^{|Q'|} I(X_1',X_2';Y'|Q'=q)P(Q'=q))$$

which is just a convex combination of 3-dimensional points

$$(I(X_1';Y'|X_2',Q'=q), I(X_1';Y'|X_2',Q'=q), I(X_1',X_2';Y'|,Q'=q)) \quad for \quad q \in \{1,2,\dots|Q'|\}.$$

Therefore, by Caratheodory's theorem, we could just as well have used another Q (possibly non-uniform) on an alphabet of size ≤ 4. □

Lemma 16.6. *(Caratheodory's Theorem on convex combinations) If x is a convex combination of points $x_1, x_2, ..., x_n$ in d-dimensions, then $\exists i_1, i_2, ...i_{d+1}$ so that $x = \sum_{i=1}^{d+1} \lambda_i x_i$, where $\lambda_i \in [0,1]$ and $\sum_{i=1}^{d+1} \lambda_i = 1$*

Proof: *Because x is a convex combination of $x_1, x_2, ..., x_n$, there exist λ_i's such that*

$$x = \sum_{i=1}^{n'} \lambda_i' x_i \tag{16.37}$$

$$, where \quad \lambda_i' \in (0,1] \quad , \quad \sum_{i=1}^{n'} \lambda_i' = 1, \quad and \quad n' \leq n$$

We will prove this by induction on n'.
If $n' \leq d+1$, then the lemma is self-evident.
For an induction hypothesis, let's assume that the theorem is true until $n'-1$. Then we will show that the theorem is true for n'
We can notice that $n' > d+1$ is only interesting, so for a fixed i there are $n'-1 (> d)$ number of $x_j - x_i$. These $n'-1$ vectors are inevitably linearly dependent because the dimensionality is smaller, and so this linear dependency can be written as

$$\sum_{j=2}^{n'} \mu_j (x_j - x_i) = 0 \quad , where\ mu_j\ are\ not\ all\ zero \tag{16.38}$$

For completeness, denote

$$\mu_1 := -\sum_{j=2}^{n'} \mu_j \tag{16.39}$$

This definition together with the fact that all can't be zero makes us claim that there exists at least one strictly positive μ_k.
Then, (16.38) can be written as

$$\sum_{j=1}^{n'} \mu_j x_j = 0 \tag{16.40}$$

By multiplying (16.40) by α and adding it to (16.37),

$$x = \sum_{i=1}^{n'} (\lambda'_i - \alpha\mu_i) x_i \ \forall \alpha \in \Re \tag{16.41}$$

Here, we know $\sum_{i=1}^{n'}(\lambda'_i - \alpha\mu_i) = 1$. All that remains is to find the proper α that makes at least one of $(\lambda'_k - \alpha\mu_k)$ zero while at the same time keeping the other coefficients non-negative. Now, just consider positive α, then

$$\begin{aligned} \lambda'_i - \alpha\mu_i \geq 0 \ \ \forall i \Leftrightarrow \lambda'_i - \alpha\mu_i \geq 0 \ \ for \ \ \mu_i \geq 0 \\ \Leftrightarrow \lambda'_i \geq \alpha\mu_i \ \ for \ \ \mu_i \geq 0 \Leftrightarrow \alpha \leq \frac{\lambda'_i}{\mu_i} \ \ for \ \ \mu_i \geq 0 \end{aligned} \tag{16.42}$$

Therefore, we choose α as

$$\alpha^* = \min_j \frac{\lambda'_j}{\mu_j} : \mu_j > 0 \tag{16.43}$$

Since we know at least one of μ_k is positive, this α^* is always exists. This α^* leads at least one of $(\lambda'_k - \alpha\mu_k)$ equal to zero, and the others must be non-negative since this is the minimal α. This completes the induction step and so the lemma is proved. □

This theorem proves the achievable region of MAC is tight. Moreover, this converse proof can be easily be extended to the Gaussian case, which almost proves the region in (Fig16.4) is tight except for the pesky problem of not having a cardinality bound. Fortunately, we observe that a Gaussian input maximizes each of the inequalities in the bound and so there is no tension among them regarding the choice of input distribution. This is what is needed to make this proof work for the Gaussian case.