Logistics

Class Road Map - IT-I

- L19 (1/6): Overview, Communications, Gaussian Channel
- L20 (1/8): Gaussian Channel, band limitation, parallel channels, optimization and duality
- L21 (1/13): parallel channels, colored noise, feedback, matrix inequalities
- L22 (1/15): matrix inequalities, rate distortion.
  - (1/20): Monday holiday
- L23 (1/22): rate distortion for Bernoulli, Gaussian, and Multiple Gaussians with unequal noise
- L24 (1/27): main rate distortion theorem, geometry
- L25 (1/29): computing $R(D)$
- L26 (2/3): computing $R(D)$, alternating minimization
- L27 (2/5): Kolmogorov complexity
- L28 (2/10): algorithmic randomness, universal prob.,
  L29 (2/12): universal compression, LZ compression
  - (2/17): Monday, Holiday
- L30 (2/19): LZ compression, info measures
  L31 (2/24): Info measures
  L32 (2/26): Info measures
  L33 (3/3):
  L34 (3/5):
  L35 (3/10):
  L36 (3/12):

Cumulative Outstanding Reading

- Read Ch. 15 in our book (Cover & Thomas, “Information Theory”).
- Read Ch. 13 in our book (Cover & Thomas, “Information Theory”).
- Read Ch. 14 in our book (Cover & Thomas, “Information Theory”).
- Read Ch. 10 in our book (Cover & Thomas, “Information Theory”).
- Read Ch. 17 in our book (Cover & Thomas, “Information Theory”) on matrix inequalities.
- Read Ch. 9 in our book (Cover & Thomas, “Information Theory”)
- Read Ch. 5 in Boyd and Vandenberghe’s Convex Optimization book
- Read all readings assigned in EE514a, Fall 2013. (see later lectures on our previous web page (http://j.ee.washington.edu/~bilmes/classes/ee514a_fall_2013/)).

Announcements

- Office hours on Mondays, 3:30-4:30.
- As always, email me if you want to skype/google hangout rather than come to office hours, also at different times.
On Final Presentations

- Your task is to give a 10-15 minute presentation that summarizes 2-3 related and significant papers that come from IEEE Transactions on Information Theory (or a very related area).
- The papers must not be ones that we covered in class, although they can be related.
- You need to do the research to find the papers yourself (i.e., that is part of the assignment).
- The majority of the papers must have been published in the last 10 years (so no old or classic papers).
- Your grade will be based on how clear, understandable, and accurate your presentation is (and also milestones).
- This is a real challenge and will require significant work! Many of the papers are complex. To get a good grade, you will need to work very hard to present very complex ideas in an extremely simple yet still precise way.
- Again, don’t expect this to be easy, you might need to try a few topics until you find one that is suitable.

Final Presentation Milestones

All submissions done in PDF file format via our assignment dropbox (https://canvas.uw.edu/courses/880971/assignments)

- Monday, March 3rd 11:45pm: progress report (at most 1 page). Any background papers you needed to read to better understand your core set. Thoughts on coherent and simple unifying presentation.
- Monday, March 10th, 11:45pm: updated short (≤ 1 page) writeup on more details of how you will present the ideas in a simple fashion.
- Final presentations: Monday, March 17, 2014, 2:30–4:20pm, LOW 102. What to turn in: your slides and a short at most 4 page summary of the papers.
Homework

- Next slide summarizes remaining assignments.

Upcoming assignments summary

- Final project status update 2, due Today at 11:45pm
- Three one-page summary of assigned papers to read, due Friday at 5pm,
- Final project status update 3, due Mar 10 at 11:45pm,
- Graded 1 page writeups, due Mar 11 at 11:45pm
- Three one-page summaries of assigned papers (second round), due Mar 14 at 5pm,
- Final slides and 4 page paper, no lates accepted! due Mar 17 at 1pm,
- Graded 1 page writeups (round 2), due Mar 18 at 11:45pm
More on $I(X_1; X_2; X_3)$

- We’ve seen $I(X_1; X_2; X_3)$ before in Venn diagrams

- We can thus view $I(X_1; X_2; X_3)$ as a form of information “among” three random variables, or the information “common to” the three r.v.s.

An alternate to $I(X_1; X_2; X_3)$

- As an alternative (note the prime), consider:

\[ I'(X_1; X_2; X_3) \triangleq H(X_1) + H(X_2) + H(X_3) - H(X_1, X_2, X_3) \]  

(33.6)

sometimes called the “total correlation” (Watanabe, 1960).

- Visualized, we have:
For $n > 3$ random variables, and multiple groups \( \{G_i\}_i \), we can define:

\[
\mu((\tilde{X}_{G_1} \cap \tilde{X}_{G_2} \cap \cdots \cap \tilde{X}_{G_m}) \setminus \tilde{X}_F) \triangleq I(X_{G_1}; X_{G_2}; \ldots; X_{G_m} | X_F)
\]

(33.9)

as the mutual information “between” or “among” the groups \( \{X_{G_i}\}_i \) conditioned on \( X_F \)

---

**Example: Two views of two views of a source**

- Can correlated noisy signals ever reveal more about a source than independent (noisy) looks at a source?
- We wish to learn about \( X \) and we have two ways we might do it:

```
 X
 / \
 Y1  Y2
 / \  vs.  \
 Z1  Z2
```

- I.e., you have either \((Z_1, Z_2)\) or \((W_1, W_2)\) but not both, where \(I(Z_1; X) = I(Z_2; X) = I(W_1; X) = I(W_2; X)\), so each individual variable is equally useful to learn about \( X \).
- Ex: communications: might get “stuck” with noise given in \( Y_1 \).
- Example: population theory. When testing about the instance, say, of a disease in a population, we can either test one individual twice (each might have errors) or two individuals once. Which is better?
Markov Chain Restrictions: Empty measures

- In fact, 5 atoms have empty measures.

\[ A = H(X_1|X_2, X_3, X_4), B = I(X_1; X_2|X_3, X_4), \]
\[ C = I(X_1; X_3|X_4), \ldots, \text{and } G = I(X_2; X_4|X_1). \]

- For example: \( I(X_1; X_3|X_2) = 0 \). What else? \( I(X_1; X_4|X_2) = 0, \)
\( I(X_1; X_4|X_3) = 0, I(X_2; X_4|X_3) = 0, \) and \( I(X_1, X_2; X_4|X_3) = 0. \)

- The 10 non-empty atoms are then: \( H(X_1|X_2, X_3, X_4), \)
\( I(X_1; X_2|X_3, X_4), I(X_1; X_3|X_4), I(X_1; X_4), H(X_2|X_1, X_2, X_4), \)
\( I(X_2; X_3|X_1, X_4), I(X_2; X_4|X_1), H(X_3|X_1, X_2, X_4), \)
\( I(X_3; X_4|X_1, X_2), \) and \( H(X_4|X_1, X_2, X_3). \)

A defining property of entropy?

- This begs the question: Consider all functions \( f : 2^{[n]} \to \mathbb{R}_+ \) that
satisfy the following three properties:

\[
\begin{align*}
f(\emptyset) &= 0 & (33.23) \\
f(A) &\leq f(B) \text{ whenever } A \subseteq B \subseteq [n] & (33.24) \\
f(A) + f(B) &\geq f(A \cup B) + f(A \cap B) \text{ for all } A, B \subseteq [n] & (33.25)
\end{align*}
\]

and call this set of functions \( \Gamma_n \) (note that this set is called the cone of
polymatroidal functions, the reason being that they are closed under conic combinations).

- Now, consider the set of all discrete entropy functions on \( n \) random
variables, call it \( \Gamma_n^H \) (also close this set as well).

- Is there a relationship between \( \Gamma_n^H \) and \( \Gamma_n \)?

- Clearly, \( \Gamma_n^H \subseteq \Gamma_n \) since entropy satisfies the three properties.
### Entropy as a class of functions

- Consider the function $f : 2^V \rightarrow \mathbb{R}_+$ defined as $f(A) = H(X_A)$, with $V = [n]$, so $|V| = n$.
- This function is specified by $2^n - 1$ values, a vector in $\mathbb{R}_{2^n-1}^+$. 
- Now consider all possible distributions defined on these $n$ random variables $X_1, X_2, \ldots, X_n$.
- Each such distribution defines a vector $\in \mathbb{R}_{2^n-1}^+$ and thus defines a function $f : 2^V \rightarrow \mathbb{R}_+$ defined by $f(A) = H(X_A)$.
- Call this entire class of functions $\Gamma_n^H$.
- We ask is $\Gamma_n^H \neq \Gamma_n$?

### Non-Shannon Entropic Inequality

Via information measures ideas presented earlier, we can get:

**Theorem 33.2.4**

*For any four discrete random variables $X, Y, Z, U$, define*

$$\Delta(Z, U | X, Y) = I(Z; U) - I(Z; U, X) - I(Z; U, Y)$$  \hspace{1cm} (33.23)

*then the following inequality holds:*

$$\Delta(Z, U | X, Y) \leq \frac{1}{2} [I(X; Y) + I(X; Z, U) + I(Z; U, X) - I(Z; U | Y)]$$  \hspace{1cm} (33.24)

*and while the l.h.s. is symmetric in $X$ and $Y$ but the r.h.s. is not, we also have:*

$$\Delta(Z, U | X, Y) \leq \frac{1}{2} [I(X; Y) + I(Y; Z, U) - I(Z; U, X) + I(Z; U | Y)]$$  \hspace{1cm} (33.25)
Entropic Functions $\subset$ Polymatroid Functions

Theorem 33.2.4

$$\Gamma^H_n \neq \Gamma_n$$ (33.26)

Proof.

- Define a function $F$ as follows:

  $$F(\emptyset) = 0$$
  $$F(X) = F(Y) = F(Z) = F(U) = 2a > 0$$
  $$F(X, Y) = 4a$$
  $$F(X, U) = F(X, Z) = F(Y, U) = F(Y, Z) = F(Z, U) = 3a$$
  $$F(X, Y, Z) = F(X, Y, U) = F(X, Z, U) = F(Y, Z, U) = 4a$$

Notation

- Notation: let $V = [n] = \{1, 2, \ldots, n\}$ be a finite set, and let $S \subseteq V$ be any subset, and $S^c = V \setminus S$ be its complement.
- We have a set of random variables $X_V$ and $X_S$ is a subset of random variables.
- Let $C = \{C_1, C_2, \ldots\}$, with $C_i \subseteq V$, be an arbitrary collection of subsets of $V$. 
Information Inequalities, and Shannon Type

- Most generally, an “information inequality” is given by defining a set of \( k \) sets \( C = \{C_1, \ldots, C_k\} \) and a set of values \( \{\alpha_1, \ldots, \alpha_k\} \) with \( \alpha_i \in \mathbb{R} \). An information inequality takes the form:

\[
\sum_i \alpha_i H(X_{C_i}) \geq 0 \tag{33.1}
\]

- E.g., non-negativity of mutual information \( I(X_{C_1}; X_{C_2}) \) takes the form \( H(X_{C_1}) + H(X_{C_2}) - H(X_{C_1 \cup C_2}) \geq 0 \) when \( C_1 \cap C_2 = \emptyset \).

- As we’ve seen, simple inequalities of the form \( H(X) \geq 0 \), \( H(X|Y) \geq 0 \), \( I(X; Y) \geq 0 \) are all derivable by \( I(A; B|C) \geq 0 \).

- General “shannon type” inequality takes the form:

\[
\sum_i \alpha_i I(A_i; B_i|C_i) \geq 0 \tag{33.2}
\]

where \( \alpha_i \geq 0 \) for all \( i \).

Information Non-Shannon Entropy Inequalities

- There are others that do not have this form. We in fact saw one already (the one not necessarily satisfied by all polymatroid functions), for four variables \( A, B, C, D \) i.e.:

\[
2I(C; D) \leq I(A; B) + I(A; C, D) + 3I(C; D|A) + I(C; D|B) \tag{33.3}
\]

- There are others as well (Dougherty et al., 2006), i.e.,:

\[
\begin{align*}
2I(A; B) &\leq 3I(A; B|C) + 3I(A; c|B) + 3I(B; c|A) + 2I(A; D) + 2I(B; C|D) \\
2I(A; B) &\leq 4I(A; B|C) + I(A; C|B) + 2I(B; C|A) + 3I(A; B|D) + I(B; C|A) + 2I(C; D) \\
2I(A; B) &\leq 3I(A; B|C) + 2I(A; C|B) + 4I(B; C|A) + 2I(A; C|D) + I(A; D|C) + 2I(B; D) + I(C; D|A) \\
2I(A; B) &\leq 5I(A; B|C) + 3I(A; C|B) + I(B; C|A) + 2I(A; D) + 2I(B; C|D) \\
2I(A; B) &\leq 4I(A; B|C) + 4I(A; B|C) + I(B; C|A) + 3I(A; D) + 3I(B; C|D) + I(C; D|B) \\
2I(A; B) &\leq 3I(A; B|C) + 2I(A; C|B) + 2I(B; C|A) + 2I(A; B|D) + I(A; D|B) + I(B; D|A) + 2I(C; D)
\end{align*}
\]

- These were apparently discovered using computer search.
Still other entropic inequalities

- **Degree**: For any $v \in V$, define the degree $r(v)$ of $v$ within $C$ as the number of times $v$ occurs in sets of $C$. That is
  \[ r(v) = |\{C \in C : v \in C\}| \] (33.10)

- **Min degree**: For any $S \subseteq V$, define $r_-(S) = \min_{v \in S} r(v)$ as the minimal degree, and $r_+(S) = \max_{v \in S} r(v)$ as the maximal degree.

- Consider bipartite graph $G = (V, C, E)$ formulation. l.h.s. is $V$ (one vertex for each $v \in V$) and r.h.s. corresponds to the sets $C$ (one vertex for each $C \in C$), edge $e = (v, C) \in E$ if $v \in C$.

---

**Theorem 33.3.1 (Madiman & Tetali 2010)**

Let $X_1, X_2, \ldots, X_n$ be any arbitrary set of discrete random variables jointly distributed according to some distribution. Let $C$ be any set of subsets of indices of the random variables, and assume that for all $v \in [n]$, $r(v) > 0$ (non-zero degree), then we have:

\[
\sum_{C \in C} \frac{H(X_C | X_{C^c})}{r_+(C)} \leq H(X_V) \leq \sum_{C \in C} \frac{H(X_C)}{r_-(C)}
\] (33.11)

- This is a very general and potentially very powerful set of inequalities, and generalizes a number of results, some of which we already know.
- These are still Shannon! (hold for polymatroid functions).
- This is generalizable to the notion of fractional covers and fractional packings (and continuous entropy for fractional partitions).
Degree based inequalities

- For example, assume $C = \{\{1\}, \{2\}, \ldots, \{n\}\}$ is the set of singletons.
- Then we immediately get the following:

$$\sum_{i=1}^{n} H(X_i | X_{V \setminus \{i\}}) \leq H(X_V) \leq \sum_{i=1}^{n} H(X_i) \quad (33.12)$$

- The upper bound is the subadditivity of entropy.

If set set $C = \{V \setminus \{1\}, V \setminus \{2\}, \ldots, V \setminus \{n\}\}$, all sets of $n - 1$ elements.
- Then we get:

$$\frac{1}{n-1} \sum_{i=1}^{n} H(X_V \setminus \{i\} | X_i) \leq H(X_V) \leq \frac{1}{n-1} \sum_{i=1}^{n} H(X_V \setminus \{i\}) \quad (33.13)$$

- This has been called “Han’s inequality”
**Degree based inequalities**

- Note that for any \( C \in \mathcal{C} \), we have \( r_+([n]) \geq r_+(C) \) (which is the overall maximum degree) and \( r_-([n]) \leq r_-(C) \) (which is the minimum degree).

- Then we have:

\[
\frac{1}{r_+([n])} \sum_{C \in \mathcal{C}} H(X_C | X_{C^c}) \leq H(X_V) \leq \frac{1}{r_-([n])} \sum_{C \in \mathcal{C}} H(X_C)
\]

(33.14)

- The r.h.s. is known as Shearer’s lemma.

**Fractional Covers, Packings, and Partitions**

- A function \( \alpha : \mathcal{C} \rightarrow \mathbb{R}_+ \) is called a **fractional covering** if for each \( v \in V \), we have that

\[
\sum_{C \in \mathcal{C} : v \in C} \alpha(C) \geq 1
\]

(33.15)

- A function \( \beta : \mathcal{C} \rightarrow \mathbb{R}_+ \) is called a **fractional packing** if for each \( v \in V \), we have that:

\[
\sum_{C \in \mathcal{C} : v \in C} \beta(C) \leq 1
\]

(33.16)

- A function \( \gamma : \mathcal{C} \rightarrow \mathbb{R}_+ \) is called a **fractional partition** if it is both a fractional cover and a fractional packing. I.e., if for each \( v \in V \), we have that:

\[
\sum_{C \in \mathcal{C} : v \in C} \gamma(C) = 1
\]

(33.17)
**Theorem 33.3.2 (Madiman & Tetali 2010)**

then, for any collection $C$ of subsets of $V$, we have that:

$$
\sum_{C \in C} \beta(C) H(X_C | H_{C^c \setminus C^>}) \leq H(X_V) \leq \sum_{C \in C} \alpha(C) H(X_C | X_{C^<}) \quad (33.18)
$$

and

$$
\sum_{C \in C} \gamma(C) h(X_C | H_{C^c \setminus C^>}) \leq h(X_V) \leq \sum_{C \in C} \gamma(C) h(X_C | X_{C^<}) \quad (33.19)
$$

where $h$ is the continuous entropy function, $C^>$ are all elements with index greater than the max index in $C$, and $C^<$ are all elements with index less than the min index in $C$.

Note, these are also still Shannon (hold for polymatroid functions).
Rényi Entropy

- If instead we use $\varphi(x) = 2^{(1-\alpha)x}$ for $\alpha \neq 1$ we still have activity.
- Doing so, gets us the Rényi Entropy, i.e.,

$H_\alpha(X) = \frac{1}{1-\alpha} \log \sum_{i=1}^{n} p_i^\alpha$  \hspace{1cm} (33.21)

for $\alpha > 0, \alpha \neq 1$.
- Note $H_\alpha(X)$ is monotonic decreasing with increasing $\alpha$, and

$\log n \geq H_\alpha(X) \geq -\log p_{\max}$  \hspace{1cm} (33.22)

where the bounds are achieved at $\alpha = 0$ and $\alpha \to \infty$.
- Also, $\lim_{\alpha \to 1} H_\alpha(X) = H(X)$, so this generalizes Shannon’s entropy.
- There are other entropies as well (Daroczy’s, Quadratic, R-norm, Havrda-Charvát, etc.), all of which are studied both for their mathematical as well as their practical properties.

Point-to-Point Communications

- So far, our model has been that there is one sender $X$ and one receiver $Y$.
- We have usefully quantified information using entropy $H$ and have studied limits of communication under this measure of information.
- We have found that one can communicate at a rate of $R$ bits per channel use (on average for large $n$) as long as $R < C$ where $C$ is the capacity of the channel. This occurs with vanishingly small error as $n \to \infty$.
- And if $R > C$, there is no hope to communicate with vanishingly small error.
- We’ve moreover realized that source and channel coding can be separated. I.e., we can entropy compress the source (whatever it is) to what looks like uniformly at random bits, and then communicate this bit stream using channel coding. The channel coder can be shared between all types of sources.
Point-to-Point Communications

- We can visualize this as follows:

General Network Information Theory

- Very important part of modern IT (still currently being actively researched).
- A very general case (first). We have an arbitrary network:

Each sender \( X_i \) is trying to communicate simultaneously with each receiver \( Y_i \) (i.e., for all \( i \), \( X_i \) is sending to \( \{Y_i\}_i \)).
- The \( X_i \) are not necessarily independent.
General Network Information Theory

• The goal is to compute the achievable region of capacities. I.e., a collective vector-valued function $\vec{C}(\Pr(x_1, x_2, \ldots, x_m))$.

• This is the capacity over which the sources can communicate without error (as $n \to \infty$).

• More generally, let $V = \{1, 2, \ldots, m\} = [m]$, and let $S \subseteq V$.

• We want a function $C : 2^V \to \mathbb{R}_+$ that gives constraints on the rate limits for communicating sources in $S$. I.e., constraints would be of the form:

$$\sum_{s \in S} R_s \leq C(S) \quad \forall S \subseteq V$$

(33.23)

• General communication network is then:

$$\Pr(y_1, y_2, \ldots, y_m | x_1, x_2, \ldots, x_m)$$

(33.24)

so a single overall rate is not specific enough.

There are many special cases of this general case, and often these special cases are very realistic and practical.

Some of these cases are solved (in that we know the achievable region for a vector of rates).

Many other important problems are only partially known (e.g., outer bounds on the achievable region, or no bounds) and these are currently of great interest.

Sometimes nodes themselves are “active” in that they may perform computation.

Active nodes can be important: with point-to-point, we only need to send the source bits and do channel coding.

In network coding theory, nodes might need to “re-code” (multiple) inputs before doing channel coding in order to achieve a given rate.
Typicality: Notation

- We'll first use typicality theory in a more general setting which will be quite useful for addressing some of these cases.
- Let $X^{1:m} \sim p(x^1, x^2, \ldots, x^m)$ be $m$ separate random variables, with $V = \{1, 2, \ldots, m\}$ and $S \subseteq V$
- Notation: $X^{1:m}$ means that we have $n$ independent instances of this length-$m$ vector (so this really is a matrix).
- $X^S_{1:n}$ for $S \subseteq [m] = V$ means $n$ independent instances of the variables in $S \subseteq V$.
- Thus,

$$\Pr(X^S_{1:n} = x^S_{1:n}) = \prod_{i=1}^{n} \Pr(X_{i}^S = x_{i}^S) \quad (33.25)$$

- Ex: for $S = (j, \ell)$

$$\Pr(X_{1:n}^{\{j, \ell\}} = x_{1:n}^{\{j, \ell\}}) = \prod_{i=1}^{n} \Pr(X_{i}^{j} = x_{i}^{j}, X_{i}^{\ell} = x_{i}^{\ell}) \quad (33.26)$$

WLLN and typicality

- the weak law of large numbers, again, says that $\forall S \subseteq V$:

$$-\frac{1}{n} \log \Pr(X^S_{1:n}) = -\frac{1}{n} \sum_{i=1}^{n} \log \Pr(X_{i}^S) \to H(X^S) \quad (33.27)$$

when $x^S_{i} \sim \Pr(x^S)$, and this is true for all $S \subseteq V$ (note again, there are $2^{|V|}$ such subsets here.)
- Define: $\forall S \subseteq V$

$$A^{(n)}(\epsilon) = \left\{ x^{S}_{1:n} : \left| -\frac{1}{n} \log \Pr(x^{S'}_{1:n}) - H(X^{S'}) \right| < \epsilon, \forall S' \subseteq S \right\} \quad (33.28)$$

- Note that this notion of typicality on $S$ requires typicality to hold for all subsets $S'$ of $S$.
- Note, however, that $S = \emptyset$ or $S' = \emptyset$ is vacuous.
Typicality, a few more definitions

- Ex: If \( S = \{1, 2\} \), then we get joint typicality we saw earlier
  \[
  A_{\epsilon}^{(n)}(S) = \left\{ (x_{1:n}^1, x_{1:n}^2) : \left| -\frac{1}{n} \log p(x_{1:n}^1, x_{1:n}^2) - H(X^1, X^2) \right| < \epsilon, \right. \\
  \left. \left| -\frac{1}{n} \log p(x_{1:n}^1) - H(X^1) \right| < \epsilon, \right. \\
  \left. \left| -\frac{1}{n} \log p(x_{1:n}^2) - H(X^2) \right| < \epsilon \right\}
  \]

- Also, \( A_{\epsilon}^{(n)} \triangleq A_{\epsilon}^{(n)}(V) \)
- \( A_{\epsilon}^{(n)}(S_1, S_2) \triangleq A_{\epsilon}^{(n)}(S_1 \cup S_2) \)

Typicality

- Notation: \( a_n \triangleq 2^{n(b+\epsilon)} \iff |\frac{1}{n} \log a_n - b| < \epsilon \). Stated another way, \( a_n = \text{poly}(n)2^{n(b+\epsilon)} \)

**Theorem 33.5.1 (Typicality)**

\( \forall \epsilon > 0, \exists n_0 \text{ s.t. for } n > n_0, we have: \)

1. \( \Pr(A_{\epsilon}^{(n)}(S)) \geq 1 - \epsilon \) for all \( S \subseteq V \)
2. If \( x_S^S \in A_{\epsilon}^{(n)}(S) \) then \( \Pr(x_S^S) \triangleq 2^{-n(H(X^S)\pm\epsilon)} \)
3. \( |A_{\epsilon}^{(n)}(S)| \triangleq 2^{n(H(X^S)\pm\epsilon)} \)
4. For \( S_1, S_2 \subseteq V \), if \( x_1^S \in A_{\epsilon}^{(n)}(S_1 \cup S_2) \) then \( \Pr(x_1^S | x_1^S) \triangleq 2^{-n(H(X^S_1|X^S_2)\pm2\epsilon)} \).

**Proof.**

Obvious from previous proofs of typicality.
Typicality: we also have

**Theorem 33.5.2**

For all $S_1, S_2 \subseteq V$ and for all $\epsilon > 0$, we have

$$A_{\epsilon}^{(n)}(X_{1:n}^{S_1} | x_{1:n}^{S_2}) = \left\{ (x_{1:n}^{S_1} : x_{1:n}^{S_1 \cup S_2} \in A_{\epsilon}^{(n)}(S_1 \cup S_2) \right\} \quad (33.30)$$

(i.e., the set of $S_1$ sequences jointly-typical with a given $S_2$ sequence $x_{1:2}^{S_2}$). Then, if $x_{1:n}^{S_2} \in A_{\epsilon}^{(n)}(S_2)$, then for large enough $n$, we have:

$$\left| A_{\epsilon}^{(n)}(X_{1:n}^{S_1} | x_{1:n}^{S_2}) \right| \leq 2^{n(H(X_{S_1}^{S_2} | X_{S_2}) + 2\epsilon)} \quad (33.31)$$

And also,

$$(1 - \epsilon)2^{n(H(X_{S_1}^{S_2} | X_{S_2}) - 2\epsilon)} \leq \sum_{x_{1:n}^{S_2}} \Pr(x_{1:n}^{S_2}) \left| A_{\epsilon}^{(n)}(X_{1:n}^{S_1} | x_{1:n}^{S_2}) \right| \quad (33.32)$$

Proof is again obvious given what we’ve done previously.

Conditional Independence and Typicality

- Before we wanted the probability that independent $X, Y$ were jointly typical (i.e., if $(X, Y) \sim p(x)p(y)$ generated from marginals $p(x)p(y)$ of $p(x, y)$, we found that $p((x, y) \in A_{\epsilon}^{(n)}) \approx 2^{-nI(X;Y)}$)
- Here, we do a similar thing but use conditional independence.
- I.e., we have $S_1, S_2, S_3 \subseteq V$. If $X^{S_1} \perp \perp X^{S_2} | X^{S_3}$, then $X^{S_1} \rightarrow X^{S_3} \rightarrow X^{S_2}$ forms a Markov chain, and

$$\Pr(x_{1:n}^{S_1 \cup S_2 \cup S_3}) = \prod_{i=1}^{n} p(x_{S_1}^{S_2} | x_{S_2}^{S_3}) p(x_{S_2}^{S_3}) p(x_{S_3}) \quad (33.33)$$

**Theorem 33.5.3**

The probability that such conditionally independent variables are typical is:

$$\Pr(x_{1:n}^{S_1 \cup S_2 \cup S_3} \in A_{\epsilon}^{(n)}(S_1 \cup S_2 \cup S_3)) \approx 2^{-n(I(S_1;S_2|S_3) + 6\epsilon)} \quad (33.34)$$
Conditional Independence and Typicality

**proof of theorem 33.5.3.**

\[
\Pr(x_{1:n}^{S_1 \cup S_2 \cup S_3} \in A^{(n)}_\epsilon(S_1 \cup S_2 \cup S_3)) = \sum_{x_{1:n}^{S_1 \cup S_2 \cup S_3} \in A^{(n)}_\epsilon(S_1 \cup S_2 \cup S_3)} p(x_{1:n}^{S_1} | x_{1:n}^{S_2}) p(x_{1:n}^{S_2} | x_{1:n}^{S_3}) p(x_{1:n}) \\
\geq |A^{(n)}_\epsilon(S_1 \cup S_2 \cup S_3)| 2^{-n(H(X^{S_3}) \pm \epsilon)} 2^{-n(H(S_1 | S_2) \pm 2\epsilon)} 2^{-n(H(S_2 | S_3) \pm 2\epsilon)} \\
= 2^{-n(I(S_1; S_2 | S_3) \pm 6\epsilon)}
\]

Depending on how we define \(S_1\), \(S_2\), and \(S_3\), we will use this theorem in various ways.

**Multiple Access Channel**

- Multiple senders to one receiver, goal is to have the rate of information between the multiple senders and single receiver be as large as possible.
- Senders can’t cooperate/communicate!
- More importantly, goal is to understand the achievable region: what set of rate vectors is achievable (such that as block length gets large, error probability goes to zero).
- Visualized:
Multiple Access Channel

- Clearly, $I(X_1, X_2; Y)$ is the rate of transmission but we can’t maximize over $p(x_1, x_2)$ since that would just be point-to-point and would require communication between $X_1$ and $X_2$. We want $X_1 \perp X_2$.
- Senders must deal with noise between each sender and the receiver, but the senders are like noise to each other and must therefore communicate in the presence of this (additional) noise.
- In general, senders can’t communicate (so no chance for cooperate between senders, but we’ll visit this again in TDMA case later).

We want to know relationship between $I(X_1, X_2; Y)$, and $R_1, R_2$, and also a coding/decoding algorithm, so that the two senders need not communicate with each other while sending in a way that we can still achieve capacity.

Discrete Memoryless Multi-Access Channel (MAC), is $\mathcal{X}_1, \mathcal{X}_2, \mathcal{Y}$, and $p(y|x_1, x_2)$.

**Definition:** A $((2^nR_1, 2^nR_2), n)$ code for a MAC is the pair of message indices $W_1 = \{1, \ldots, 2^nR_1\}$, $W_2 = \{1, \ldots, 2^nR_2\}$; encoders $X_1 : W_1 \rightarrow \mathcal{X}_1^n$ and $X_2 : W_2 \rightarrow \mathcal{X}_2^n$ (so to length-$n$ strings); a decoding function $g : \mathcal{Y}^n \rightarrow W_1 \times W_2$ (we decode both simultaneously).
- Assume $p(w_1, w_2) = \frac{1}{|W_1||W_2|}$ uniform.
- Probability of error:

$$P_e^{(n)} = \frac{1}{2^{n(R_1+R_2)}} \sum_{w_1, w_2} \Pr(g(Y_1^n) \neq (w_1, w_2) | (w_1, w_2) \text{ sent})$$
**Multiple Access Channel**

- **Definition:** a pair \((R_1, R_2)\) is achievable for a MAC if there exists a sequence \(((2^nR_1, 2^nR_2), n)\) of codes with \(P_e^n \to 0\) as \(n \to \infty\).
- **Definition:** The Capacity region is the set of achievable \((R_1, R_2)\) pairs.

**Theorem 33.6.1**

The MAC capacity of a channel is the closure of the convex hull of all \((R_1, R_2)\) satisfying:

\[
R_1 \leq I(X_1; Y | X_2) \tag{33.36}
\]
\[
R_2 \leq I(X_2; Y | X_1) \tag{33.37}
\]
\[
R_1 + R_2 \leq I(X_1, X_2; Y) \tag{33.38}
\]

under a given product distribution \(p(x_1)p(x_2)\).

- We can view this as a polytope (or more simply a pentagon) in \(\mathbb{R}^2\) since \(\max \{I(X_1; Y | X_2), I(X_2; Y | X_1)\} \leq I(X_1X_2; Y) \leq I(X_1; Y | X_2) + I(X_2; Y | X_1)\)
- For a particular \(p_1(x_1), p_2(x_2)\) pair, we have:

\[
\begin{align*}
R_1 & = I(X_1; Y | X_2) \\
R_2 & = I(X_2; Y | X_1) \\
R_1 + R_2 & = I(X_1, X_2; Y)
\end{align*}
\]

Any pair of rates \((R_1, R_2)\) within polytope is achievable.
Multiple Access Channel

- We have $I(X_1; Y | X_2)$ and $I(X_2; Y | X_1)$, $X_1 = \{X_1, X_2\} \setminus X_2$.
- Since, $I(A; B|C) = H(A, C) + H(B, C) - H(C) - H(A, B, C)$ we have that $I(S; Y | S^c) = I(S; Y | V \setminus S) = H(V) - H(V \setminus S) - H(Y, V) + H(Y, V \setminus S)$

so no immediately apparent nice polyhedral structure here.
- However, the function $f(S) = I(S; Y | V \setminus S) = \text{const.} + H(Y | V \setminus S)$ in fact is polymatroidal (non-negative, monotone non-decreasing, submodular) under the MAC model ($X_i$'s are independent). In other words, the function:

$$f(A) = I(X_A; Y | X_{V \setminus A})$$

is polymatroidal under the Mac assumption (Han 1979).
- In fact, we want to find $p(x_1), p(x_2)$ to make the region as large as possible, so that we have capacity constraints $C_1, C_2,$ and $C_{12}$.

Some simple examples: Suppose we have two independent BSCs with no interference between channels.
- Each channel is a BSC and so has rate $R_i = 1 - H(p_i)$ for $i \in \{1, 2\}$.
- The achievable region is a square.
Multiple Access Channel

- Suppose we have a binary multiplier channel, i.e., $Y = X_1 X_2$.
- If $X_1 = 1$, then $X_2$ sent to $Y$ at rate $1 - H(0) = 1$.
- If $X_2 = 1$, then $X_1$ sent to $Y$ at a rate $1 - H(0) = 1$.
- Max $H(Y) = 1$ and gives limit on rate, so $R_1 + R_2 = 1$.
- Thus, we have triangle shaped polytope:

![Triangle Polytope Diagram]

Binary Erasure Channel

- $e$ is an erasure symbol, if that happens we don’t have access to the transmitted bit.
- The probability of dropping a bit is then $\alpha$.
- We want to compute capacity. Obviously, $C = 1$ if $\alpha = 0$.

$$C = \max_{p(x)} I(X;Y) = \max_{p(x)} (H(Y) - H(Y|X))$$

$$= \max_{p(x)} H(Y) - H(\alpha)$$

- So while $H(Y) \leq \log 3$, we want actual value of the capacity.
Binary Erasure Channel

- Let $E = \{Y = e\}$. Then
  
  $$H(Y) = H(Y, E) = H(E) + H(Y|E)$$

- Let $\pi = \Pr(X = 1)$. Then
  
  $$H(Y) = H\left((1 - \pi)(1 - \alpha), \alpha \pi (1 - \alpha)\right)$$

  (33.7)

  $$= H(\alpha) + (1 - \alpha)H(\pi)$$

  (33.8)

  This last equality follows since $H(E) = H(\alpha)$, and
  
  $$H(Y|E) = \alpha H(Y|Y = e) + (1 - \alpha)H(Y|Y \neq e) = \alpha \cdot 0 + (1 - \alpha)H(\pi)$$

  Then we get

  $$C = \max_{p(x)} H(Y) - H(\alpha)$$

  (33.7)

  $$= \max_{\pi} \left( (1 - \alpha)H(\pi) + H(\alpha) \right) - H(\alpha)$$

  (33.8)

  $$= \max_{\pi} (1 - \alpha)H(\pi) = 1 - \alpha$$

  (33.9)

  Best capacity when $\pi = 1/2 = \Pr(X = 1) = \Pr(X = 0)$.

  This makes sense, loose $\alpha\%$ of the bits of original capacity.