6.441: Information Theory	Spring 2012
Lecturer: Yury Polyanskiy	**Lecture 1: Information measures**

Feb. 4, 2013 [1]

Two methods to describe a random variable X.

- r.v. X – a distribution P_X on some measurable space $\mathcal{X}$.
  Note: $\mathcal{X}$ can be a space of apples or oranges — unlike in Probability Theory.
- r.v. X – a function $X : \Omega \to \mathcal{X}$ from the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ to a target sapce $\mathcal{X}$.

Definition 1.1. *Random variable X — discrete if there exists a countable set $\{x_j, j = 1, \ldots\}$ such that $\sum_{j=1}^{\infty} P_X(x_j) = 1$.*

Definition 1.2. *For a d.R.V. X with distribution P_X:*

$$\begin{aligned} H(X) &= E\Big[\log \frac{1}{P_X(X)}\Big] \leftarrow \textit{[Measure of randomness of } X \textit{ (intuitively)]} \\ &= \sum_{x \in \mathcal{X}} P_X(x) \log \frac{1}{P_X(X)} \ . \end{aligned}$$

Why such definition, why log, why entropy? — later.

Note:

- $0 \log \frac{1}{0} = 0$ (by continuity)
- Will also write $H(P_X)$ instead of $H(X)$ (abuse of notation).
- Basis of log — units

$$\begin{aligned} \log_2 &\leftrightarrow \text{bits} \\ \log_e &\leftrightarrow \text{nats} \\ \log &\leftrightarrow \text{arbitrary units, base always matches exp} \end{aligned}$$

Note: Elementary properties of H:

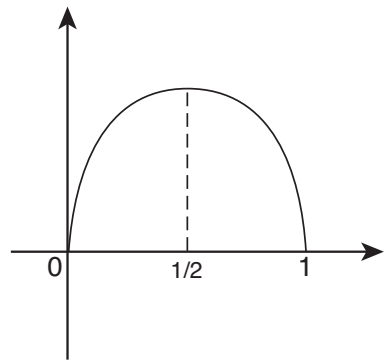
1. $0 \le H(x) \le \log |\mathcal{X}|$ (*proof*)

[1] Typed by Rachel Cohen rcohen@mit.edu, Qingqing H. qqh@mit.edu

2. Can $H(x) = +\infty$?
   A: Yes, example: $\mathbb{P}[X = k] = \frac{c}{k \ln^2 k}$

Example: $X \in \{0, 1\}$ $\quad \mathbb{P}[X = 1] = P_X(1) \triangleq p$

$$\begin{aligned} H(X) &= h(p)\ , \\ & h(p) = p \log \frac{1}{p} + \overline{p} \log \frac{1}{\overline{p}} \end{aligned}$$



where $h(\cdot)$ is the binary entropy function and $\overline{p} = 1 - p$.

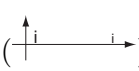
Proposition 1.1. *$h(p)$ -continuous , concave and*

$$h'(p) = \log \frac{\overline{p}}{p}$$

Example: $X \in \{0, 1, 2, \ldots\}$ $\quad \mathbb{P}[X = i] = P_x(i) = p \cdot (\overline{p})^i$

$$\begin{aligned} H(X) &= \sum_{i=0}^{\infty} p \cdot \overline{p}^i \log \frac{1}{p \cdot \overline{p}^i} = \sum_{i=0}^{\infty} p \overline{p}^i \Big(i \log \frac{1}{\overline{p}} + \log \frac{1}{p} \Big) \\ &= \log \frac{1}{p} + p \cdot \log \frac{1}{\overline{p}} \cdot \frac{1-p}{p^2} = \frac{h(p)}{p} \end{aligned}$$

Theorem 1.1. *H — only depends on the values of P_X, not locations*

$H($ $) = H($  $)$

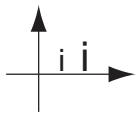
i.e.

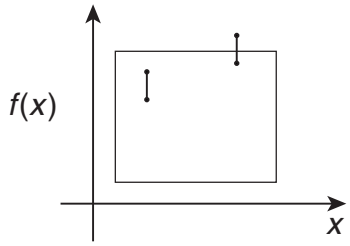
$$H(X) = H(f(X)) \ \forall \textit{1-1 function } f$$

Definition 1.3. *Joint entropy $H(X_1, X_2, \ldots, X_n) \triangleq H(X^n)$. X^n – a vector with n components.*

$$H(X_1, \ldots, X_n) = \mathbb{E}\Big[\log \frac{1}{P_{X_1, \ldots, X_n}(X_1, \ldots, X_n)} \Big]$$

Theorem 1.2. $H(X, f(X)) = H(X)$

Proof. 



□

Definition 1.4. *Conditional entropy:*

$$H(X|Y) = \mathbb{E}_{y\sim P_Y}[H(P_{X|Y=y})] = \mathbb{E}\Big[\log \frac{1}{P_{X|Y}(X|Y)}\Big]$$

Theorem 1.3. $H(X|Y) \leq H(X)$*: Pf: later.*

Theorem 1.4. $H(X,Y) = H(X) + H(Y|X)$*:*

Proof.

$$\mathbb{E}\log \frac{1}{P_{XY}[X,Y]} = \mathbb{E}\Big[\log \frac{1}{P_X(X)\cdot P_{Y|X}(Y|X)}\Big]$$

□

Theorem 1.5. $H(f(X)) \leq H(X)$*.* $H(f(X))$ *would be equal to* $H(X)$ *iff* f *is* $1-1$ *on* supp X.

Proof.

$$H(X) = H(X, f(X)) = H(f(X)) = H(X|f(X)) \geq H(f(X))$$

They are equal iff $H(X|f(X)) = 0 \Leftrightarrow$ given $f(X)$ — X is *constant* □

Theorem 1.6.

$$\begin{aligned} H(X_1,\ldots,X_n) = \sum_{i=1}^{n} H(X_i|X_1^{i-1}) \quad &\leq \quad \sum_{i=1}^{n} H(X_i) \\ &\uparrow \quad \textit{equality iff } X_1 \perp\!\!\!\perp X_n \end{aligned}$$

Proof. (telescoping)

$$P_{X_1X_2\cdots X_n} = P_{X_1}P_{X_2|X_1}\cdots P_{X_n|X^{n-1}}$$

Recall notation $X^n = (X_1,\ldots,X_n)$. □

Definition 1.5. *Divergence: Let* P, Q *— distribution on it*

- *Alphabet $\mathcal{A}$ — discrete (finite or countably infinite)*

$$D(P||Q) \triangleq \sum_{a\in\mathcal{A}} P(a)\log\frac{P(a)}{Q(a)} \ ,$$

where we agree:

(1) $0\cdot\log\frac{0}{0}=0$

(2) $\exists a : Q(a)=0, P(a)>0 \Rightarrow D(P||Q)=\infty$

- *$\mathcal{A}=\mathbb{R}^k$, P and Q have densities f_P and f_Q*

$$D(P||Q) = \begin{cases} \int_{\mathbb{R}^k} \log\frac{f_P(x^k)}{f_Q(x^k)} f_P(x^k)dx^k & , \ \mathrm{Leb}\{f_P > f_Q = 0\} = 0 \\ +\infty & , \ \textit{otherwise} \end{cases}$$

- *$\mathcal{A}$ — measurable space:*

$$D(P||Q) = \begin{cases} \mathbb{E}_Q \frac{dP}{dQ}\log\frac{dP}{dQ} & , \ P \ll Q \\ +\infty & , \ \textit{otherwise} \end{cases}$$

- *Also known as Information divergence, Kullback–Leibler divergence, relative entropy.*

Note: $D(P||Q)$ can be ∞ also when $P \ll Q$, but the two cases of $D(P||Q)=+\infty$ are consistent since $D(P||Q)=\sup_\Pi D(P_\Pi||Q_\Pi)$, where Π is a finite partition of the underlying space $\mathcal{A}$ (proof: later)

Note: Radon-Nikodym theorem: if two measure on $\mathcal{X}$ satisfy that $\nu << \mu$, i.e. ν is absolutely continuous w.r.t. μ, then there exists a function $f:\mathcal{X}\to\mathbb{R}_+$ such that

$$\nu[E] = \int_E f(x)d\mu = \int_E \frac{d\nu}{d\mu}d\mu$$

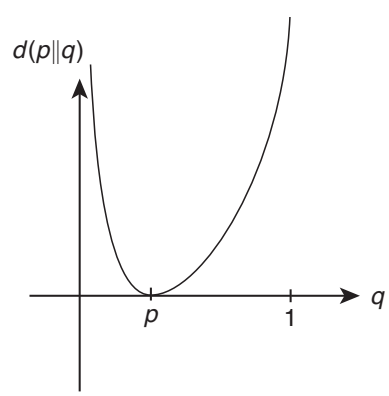
Theorem 1.7. $H(P) = \log|\mathcal{A}| - D(P||U_{\mathcal{A}})$

$\uparrow$ *uniform on $\mathcal{A}$*

Intuition: dissimilarity of P vs. Q.

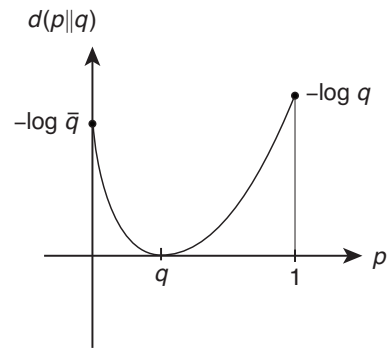
Example: $\mathcal{A}=\{0,1\}$; $P=[p,\overline{p}]$; $Q=[q,\overline{q}]$

$$D(P||Q) \triangleq d(p||q) = p\log\frac{p}{q} + \overline{p}\log\frac{\overline{p}}{\overline{q}}$$



Quadratic lower bound:

$$d(p||q) \geq 2(p-q)^2 \log e$$



Example: $\mathcal{A} = \mathbb{R}$

$$D(\mathcal{N}(m_1, \sigma_1^2)||\mathcal{N}(m_0, \sigma_0^2)) = \frac{1}{2}\log\frac{\sigma_0^2}{\sigma_1^2} + \frac{1}{2}\Big[\frac{(m_1-m_0)^2}{\sigma_0^2} + \frac{\sigma_1^2}{\sigma_0^2} - 1\Big]\log e$$

Example: $\mathcal{A} = \mathbb{C}$, $\mathcal{N}_c(m, \sigma^2) \sim \frac{1}{\pi\sigma^2} e^{-|x-m|^2/\sigma^2}$ where $\mathcal{N}_c(m, \sigma^2) = \mathcal{N}\left(\begin{bmatrix} m & m \end{bmatrix}, \begin{bmatrix} \sigma^2/2 & 0 \\ 0 & \sigma^2/2 \end{bmatrix}\right)$

$$D(\mathcal{N}_c(m_1, \sigma_1^2)||\mathcal{N}_c(m_0, \sigma_0^2)) = \log\frac{\sigma_0^2}{\sigma_1^2} + \Big[\frac{|m_1-m_0|^2}{\sigma_0^2} + \frac{\sigma_1^2}{\sigma_0^2} - 1\Big]\log e$$

Example: $\mathcal{A} = \mathbb{C}^k$

$$\begin{aligned} D(\mathcal{N}_c(m_1, \Sigma_1)||\mathcal{N}_c(m_0, \Sigma_0)) &= \log\det\Sigma_0 - \log\det\Sigma_1 + (m_1-m_0)^H\Sigma_0^{-1}(m_1-m_0)\log e \\ &\quad + \operatorname{tr}(\Sigma_0^{-1}\Sigma_1 - I)\log e \end{aligned}$$

(assume $\det\Sigma_0 \neq 0$).

6.441: Information Theory	Spring 2013
Lecturer: Yury Polyanskiy	**Lecture 2: Information Measures**

Topics: Properties of D; Mutual Information

Theorem 2.1 (Information Inequality)**.**

$$D(P||Q) \geq 0 \; ; \quad D(P||Q) = 0 \quad \textit{iff } P = Q$$

Proof. Let $\varphi(x) \triangleq x \log x$, and use Jensen's Inequality:

$$D(P||Q) = \sum_{\mathcal{X}} P(x) \log \frac{P(x)}{Q(x)} = \sum_{\mathcal{X}} Q(x) \varphi\left(\frac{P(x)}{Q(x)}\right) \geq \varphi\left(\sum_{\mathcal{X}} Q(x) \frac{P(x)}{Q(x)}\right) = \varphi(1) = 0$$

□

2.1 Conditional divergence

Definition 2.1. *Conditional probability kernels (also called random transformations) $P_{Y|X}$:*

1. $P_{Y|X}(\,\cdot\,|x_0)$ *— prob. measure on* Y
2. $x \to P_{Y|X}(A|x)$ *— measurable on* $\mathcal{X}$

Notation: $X \xrightarrow{P_{Y|X}} Y$

Multiplication:

$$\begin{tikzcd} & X \arrow[r, "P_{Y|X}"] & Y \end{tikzcd}$$

to get $P_{XY} = P_X P_{Y|X}$:

$$P_{XY}(x, y) = P_{Y|X}(y|x) P_X(x)\,.$$

Composition: $P_Y = P_{Y|X} \circ P_X$, that is $P_{Y|X}$ acts on P_X to produce P_Y:

$$P_Y(y) = \sum_{x \in \mathcal{X}} P_{Y|X}(y|x) P_X(x)\,.$$

Will also write $P_X \overset{P_{Y|X}}{\to} P_Y$.

Typed by Rachel Cohen rcohen@mit.edu, Austin Collins austinc@mit.edu, Feb 7, 2013

Definition 2.2. *Conditional divergence*

$$\begin{aligned} D(P_{Y|X}||Q_{Y|X}|P_X) &= \mathbb{E}_{x\sim P_X}[D(P_{Y|X=x}||Q_{Y|X=x})] && (1)\\ &= \sum_{x\in\mathcal{X}} P_X(x) D(P_{Y|X=x}||Q_{Y|X=x})\,. && (2)\end{aligned}$$

Note: $H(X|Y) = \log|\mathcal{A}| - D(P_{Y|X}||U_X|P_Y)$, U_X *— uniform on* $\mathcal{X}$.

Theorem 2.2. *(Properties of Divergences)*

1. $D(P_{Y|X}||Q_{Y|X}|P_X) = D(P_X P_{Y|X}||P_X Q_{Y|X})$

   Proof. $\mathbb{E}_{x\sim P_X}[D(P_{Y|X=x}||Q_{Y|X=x})] = \mathbb{E}_{(X,Y)\sim P_X P_{Y|X}}\left[\log \frac{P_{Y|X}}{Q_{Y|X}}\frac{P_X}{P_X}\right]$ □

2. *(Chain Rule)* $D(P_{XY}||Q_{XY}) = D(P_{Y|X}||Q_{Y|X}|P_X) + D(P_X||Q_X)$

   Proof. Disintegration: $\mathbb{E}_{(X,Y)}\left[\log \frac{P_{XY}}{Q_{XY}}\right] = \mathbb{E}_{(X,Y)}\left[\log \frac{P_{Y|X}}{Q_{Y|X}} + \log \frac{P_X}{Q_X}\right]$ □

   Corollary 2.1.
   $$D(P_{XY}||Q_{XY}) \geq D(P_X||Q_X) \Rightarrow H(X|Y) \leq H(X)$$

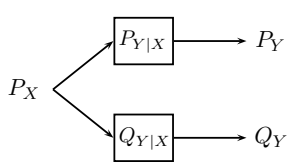
3. $D(P_{X_1\cdots X_n}||Q_{X_1}\cdots Q_{X_n}) = D(P_{X_1\cdots X_n}||P_{X_1}\cdots P_{X_n}) + \sum D(P_{X_i}||Q_{X_i})$

   Corollary 2.2.
   $$\begin{aligned} D(P_{X_1\cdots X_n}||Q_{X_1}\cdots Q_{X_n}) &\geq \sum D(P_{X_i}||Q_{X_i}) \quad \textit{or}\\ &= \textit{iff } X_1 \perp\!\!\!\perp X_2 \perp\!\!\!\perp \ldots \perp\!\!\!\perp X_n\end{aligned}$$

   Question*: Is* $D(P_{XY}||Q_{XY}) \geq D(P_X||Q_X) + D(P_Y||Q_Y)$*?*
   No! See, for example, $X = Y$*. (Hint: apply Theorem 2.)*

4. *(**Conditioning increases divergence**) Let* $P_{Y|X}$ *and* $Q_{Y|X}$ *be two kernels, let* $P_Y = P_{Y|X} \circ P_X$ *and* $Q_Y = Q_{Y|X} \circ P_X$*. Then*
   $$\begin{aligned} D(P_Y||Q_Y) &\leq D(P_{Y|X}||Q_{Y|X}|P_X)\\ &\quad \textit{equality iff } D(P_{X|Y}||Q_{X|Y}|P_Y) = 0\end{aligned}$$

   Picture:



   Proof. the "chain rule" for D:
   $$\begin{aligned} D(P_{XY}||Q_{XY}) &= D(P_{Y|X}||Q_{Y|X}|P_X) + D(P_X||P_X)\\ &= D(P_{X|Y}||Q_{X|Y}|P_Y) + D(P_Y||Q_Y)\end{aligned}$$

   □

5. *(**Data-processing for divergences**) Let* $P_Y = P_{Y|X} \circ P_X$

$$\left.\begin{array}{rcl} P_Y & = & \int P_{Y|X}(\cdot|x)dP_X \\ Q_Y & = & \int P_{Y|X}(\cdot|x)dQ_X \end{array}\right\} \Longrightarrow D(P_Y||Q_Y) \le D(P_X||Q_X)$$

Pictorially:

$$\begin{tikzcd} P_X \arrow[dr] & & & P_Y \\ & X \arrow[r, "P_{Y|X}"] & Y \arrow[ur] \arrow[dr] & \\ Q_X \arrow[ur] & & & Q_Y \end{tikzcd} \Longrightarrow D(P_X||Q_X) \ge D(P_Y||Q_Y)$$

Corollary 2.3. $Y = f(x) \quad \& \exists f^{-1} : D(P_Y||Q_Y) = D(P_X||Q_X)$

Corollary 2.4 (Large deviations estimate)**.** *For any subset* $E \subset \mathcal{X}$ *we have*

$$d(P_X[E]||Q_X[E]) \le D(P_X||Q_X)$$

2.2 Mutual information

Definition 2.3. *Mutual information:*

$$I(X;Y) = D(P_{XY}||P_X P_Y)$$

- *Defined by Shannon (in a different form), in this form by Fano.*
- *Note: not restricted to discrete.*

Theorem 2.3. *(Properties of I)*

1. $I(X;Y) = I(Y;X)$
2. $I(X;Y) \ge 0$ *or* $I(X;Y) = 0$ *iff* $X \perp\!\!\!\perp Y$
3. $I(X;X) = \begin{cases} H(X) & , \ \lambda\text{-discrete} \\ +\infty & , \ \sigma/\omega \end{cases}$

   Proof. $I(q(X);q(X)) = \log|\mathrm{Im}\, q|$ where $q(X)$ is a uniform quantizer. □

4. f *– one-to-one functions:* $I(f(X);Y) = I(X;Y)$

   Proof. From data processing: $I(X;Y) \ge I(f(X);Y) \ge I(f^{-1}(f(X));Y) = I(X;Y)$ □

5. *More data* $\Rightarrow$ *more info:* $I(X_1, X_2; Z) \ge I(X_1; Z)$
6. *If* X, Y *discrete:* $I(X;Y) = H(X) + H(Y) - H(X,Y)$

   Proof. $\mathbb{E}\left[\log\frac{1}{P_X} + \log\frac{1}{P_Y} - \log\frac{1}{P_{XY}}\right] = \mathbb{E}\log\frac{P_{XY}}{P_X P_Y}$ □

7. *If Y discrete:* $I(X;Y) = H(Y) - H(Y|X)$

8. $I(X;Y) \leq \min(H(X), H(Y))$, *equality iff* $H(X|Y)H(Y|X) = 0$

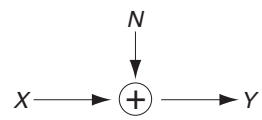
Example: X, Y — jointly Gaussian

$$I(X;Y) = \frac{1}{2}\log\frac{1}{1-\rho^2}$$

where ρ is the correlation coefficient.

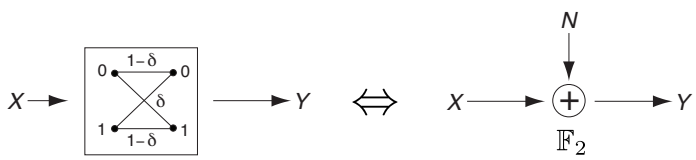
Example: $X \perp\!\!\!\perp N$ — jointly Gaussian

$$I(X;X+N) = \frac{1}{2}\log\left(1+\frac{\sigma_X^2}{\sigma_N^2}\right)$$



Example:

$$\begin{aligned} X &\sim \mathrm{Bern}\left(\frac{1}{2}\right),\ Z \sim \mathrm{Bern}(\delta) \\ Y &= X + Z \\ I(X;Y) &= \log 2 - h(\delta) \end{aligned}$$



Example: Addition over finite abelian group:

$$I(X;X+N) = \log|\mathcal{A}| - H(N)$$

Example: Gaussian vectors:

$$I(\vec{X};\vec{Y}) = \frac{1}{2}\log\frac{|\Sigma_X||\Sigma_Y|}{|\Sigma_{XY}|}$$

2.3 Conditional M.I. and conditional independence

Definition 2.4. *Conditional M.I.:*

$$I(X;Y|Z) = D(P_{XY|Z}||P_{X|Z}P_{Y|Z}|P_Z)$$

Where the product of two random transformations is $(P_{X|Z=z}P_{Y|Z=z})(x,y) \triangleq P_{X|Z}(x,z)P_{Y|Z}(y,z)$.

Remark: A family of distributions can be represented by a directed acyclic graph. A simple example is a Markov chain, which represents distributions that factor as $\{P_{XYZ} : P_{XYZ} = P_X P_{Y|X} P_{Z|Y}\}$.

$$\text{Cond. indep. notation} \begin{cases} X \to Y \to Z & \Leftrightarrow \quad P_{XZ|Y} = P_{X|Y} \cdot P_{Z|Y} \\ & \Leftrightarrow \quad P_{XYZ} = P_X \cdot P_{Z|X} \cdot P_{Z|Y} \\ & \Leftrightarrow \quad X, Y, Z \text{ form a Markov channel} \\ & \Leftrightarrow \quad Z \to Y \to X \Leftrightarrow X \to Y \to Z \end{cases}$$

Theorem 2.4. *Further properties of Mutual Information*

1. $I(X;Z|Y) \geq 0$ *or* $I(X;Z|Y) = 0$ *if* $X \to Y \to Z$
2. *(Kolmogorov identity)*

$$I(X,Y;Z) = I(X;Z) + I(Y;Z|X) = I(Y;Z) + I(X;Z|Y)$$

Proof.

$$\frac{P_{XYZ}}{P_{XY}P_Z} = \frac{P_{XZ}}{P_X P_Z} \cdot \frac{P_{Y|XZ}}{P_{Y|X}}$$

□

3. *(**Data Processing for M.I.**) If* $X \to Y \to Z$*:*
   (a) $I(X;Z) \leq I(X;Y)$
   (b) $I(X;Y|Z) \leq I(X;Y)$

Proof. Apply Kolmogorov identity to $I(Y,Z;X)$. □

Note: In general, conditioning does not always decrease M.I.

$$\begin{aligned} X \perp\!\!\!\perp Z \ , \ \sim U_{[0,1]} \qquad & Y = X + Z \\ I(X;Y) = 0 \qquad & X \to Y \leftarrow Z \\ I(X;Y|Z) = H(X) = \log Z & \end{aligned}$$

Example*: 1-to-1 function* $\Rightarrow I(X;Y) = I(X;f(Y))$

4. *(Chain Rule for M.I) Using the notation* $X^k \triangleq (X_1, \cdots, X_k)$

$$I(X^n;Y) = \sum_{k=1}^{n} I(X_k;Y|X^{k-1})$$

Proof. Recursive application of Kolmogorov identity. □

Question: $X - Y - Z \implies I(X;Y|Z) \leq I(X;Y)$. Does this hold in general?

Hint: Look for counterexamples where X, Y, Z *do not form a Markov chain, there is only one directed acyclic graph non-isomorphic to* $X - Y - Z$*.*

6.441: Information Theory	Spring 2012
Lecturer: Yury Polyanskiy	**Lecture 3: Representation of M.I.**

Feb. 12, 2013 [1]

3.1 Sufficient Statistics and data-processing

Definition 3.1. *Let P_X^θ be a collection of distributions of X parameterized by θ and $P_{Y|X}$ be some probability kernel. For every θ let $P_{X|Y}^\theta$ be a conditional distribution of X given Y. We say that Y is a sufficient statistic (s.s.) of X for θ if $P_{X|Y}^\theta$ can be chosen to not depend on θ.*

Interpretation: Know Y, can forget X (dimension reduction).

Theorem 3.1 (Fisher)**.** *Let $X \to Y \to T$. Then, T—s.s. of Y for X iff $X \to T \to Y$, i.e. can be factorized:*

$$P_{YT}^x(y,t) = a(t,x)b(y,t)$$

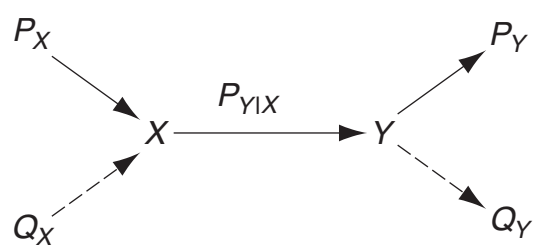
$\Leftrightarrow$

$$\forall P_X, I(X;T) = I(X;Y)$$

Example: $\theta \in \mathbb{R}$ and observations $Y^n \sim \mathcal{N}(\theta, \mathbf{I}_n)$. Then $T = \sum_j Y_j$ is a s.s.
Verify: P_{TY}^θ can be factorized.

Example: $\theta = \{0,1\}$ (binary hypothesis testing). Then T - s.s. iff $D(P_{Y|T}^0 || P_{Y|T}^1 | P_T^0) = 0$.
The concept of s.s. gives a natural condition for the equality in the data-processing:

"Processing decreases divergence":



$$\begin{aligned} D(P_X||Q_X) &\leq D(P_Y||Q_Y) \\ &\quad \text{equality iff } Y\text{– s.s. for testing } P_X \text{ vs. } Q_X \end{aligned}$$

[1] Typed by Rachel Cohen rcohen@mit.edu, Qingqing H. qqh@mit.edu

Proof: Let $Q_{XY} = Q_X P_{Y|X}$, $P_{XY} = P_X P_{Y|X}$, then

$$\begin{aligned} D(P_{XY}||Q_{XY}) &= \underbrace{D(P_{Y|X}||Q_{Y|X}|P_X)}_{=0} + D(P_X||Q_X) \\ &= D(P_{X|Y}||Q_{X|Y}|P_Y) + D(P_Y||Q_Y) \\ &\geq D(P_Y||Q_Y) \end{aligned}$$

Theorem 3.2. *If $X \to Y \to Z$ then*

$$\begin{aligned} I(X;Z) &\leq I(X;Y) \\ & \quad \textit{equality iff } Z\text{—s.s. for } X \end{aligned}$$

3.2 Geometric interpretation of M.I.

$$I(X;Y) = D(P_{Y|X}\|P_Y|P_X) = \sum_\theta D(P_{Y|X=\theta}\|P_Y)P_X(\theta)$$

M.I. as weighted distance.

Theorem 3.3 (Golden Formula / Duality method)**.** *$\forall Q_Y$ such that $D(P_{Y|X}\|Q_Y|P_X) < \infty$*

$$I(X;Y) = D(P_{Y|X}\|Q_Y|P_X) - D(P_Y\|Q_Y)$$

Proof. $I(X;Y) = \mathbf{E}[log\frac{P_{Y|X}Q_Y}{P_Y Q_Y}]$, group $P_{Y|X}$ and Q_Y □

Theorem 3.4 (M.I. as center of gravity)**.**

$$I(X;Y) = \min_Q D(P_{Y|X}||Q|P_X)$$

Proof. use the Golden formula, pick $Q_Y = P_Y$. □

Note: it is useful in proving M.I. upperbound

Theorem 3.5 (M.I. as the min distance to space $X \perp Y$)**.**

$$I(X;Y) = \min_{Q_X,Q_Y} D(P_{XY}||Q_X Q_Y)$$

Proof. $I(X;Y) = \mathbf{E}[log\frac{P_{Y|X}Q_X Q_Y}{P_X P_Y Q_X Q_Y}]$, group P_{XY} and $Q_X Q_Y$ □

Note: More R.V.s? think about the graphical model, take distance to subgraphs.
example: X, Y, Z

$$I(X;Z|Y) = \min_{Q_{XYZ}:X\to Y\to Z} D(P_{XYZ}\|Q_{XYZ})$$

Proof.

$$\begin{aligned}D(P_{XYZ}\|Q_XQ_{Y|X}Q_{Z|X}) =& D(P_{XYZ}\|P_XP_{Y|X}P_{Z|Y}) + D(P_X\|Q_X) + D(P_{Y|X}\|Q_{Y|X}) + D(P_{Z|Y}\|Q_{Z|Y})\\ =& D(P_{XYZ}\|P_YP_{X|Y}P_{Z|Y}) + \dots\\ =& D(P_{XZ|Y}\|P_{X|Y}P_{Z|Y}|P_Y) + \dots\end{aligned}$$

$I(X;Z|Y)$ is the information flow on edge $X - Z$ in the graphical model for P_{XYZ}. □

3.3 Variational characterizations of divergence

Theorem 3.6 (Donsker-Varadhan)**.** *Let P, Q be probability measures on $\mathcal{X}$ and let $\mathcal{B}$ be a class of all functions $f : \mathcal{B} \to \mathbb{R}$ such that $\mathbb{E}_Q[\exp\{f(X)\}] < \infty$. If $D(P||Q) < \infty$ then for every $f \in \mathcal{B}$ expectation $\mathbb{E}_P[f(X)]$ exists and furthermore*

$$D(P||Q) = \sup_{f\in\mathcal{C}} \mathbb{E}_P[f(X)] - \log \mathbb{E}_Q[\exp\{f(X)\}]\,. \tag{1}$$

Proof. The $\leq$ is easy: $f = \log \frac{dP}{dQ}$. For $\geq$, fix $f \in \mathcal{C}$ and define a tilted measure Q^f via

$$Q^f(dx) \triangleq \exp\{f(x) - Z_f\}Q(dx)\,, \qquad Z_f = \log \int_{\mathcal{X}} \exp\{f(x)\}Q(dx) \triangleq \log \mathbb{E}_Q[\exp\{f(X)\}]\,.$$

Then, we have

$$\mathbb{E}_P[f(X)] - Z_f = \mathbb{E}_P\left[\log \frac{dQ^f}{dQ}\right] = D(P||Q) - D(P||Q^f) \leq D(P||Q)\,.$$

□

Interpretation: By setting $f(x) = \epsilon \cdot g(x)$ with $\epsilon \ll 1$ and linearizing exp and log we can see that when $D(P||Q) \ll 1$, expectations under P can be approximated by expectations over Q (change of measure): $\mathbb{E}_P[g(X)] \approx \mathbb{E}_Q[g(X)]$.

Note: In the next lecture we show that $P \mapsto D(P||Q)$ is convex. A general method of obtaining variational formulas like (1) is by Young-Fenchel inequality. Indeed, (1) is exactly this inequality since the Fenchel-Legendre conjugate of $D(\cdot||Q)$ is given by a convex map $f \mapsto Z_f$.

Note: the inequality $\mathbf{E}_P[f(X)] - \log E_Q[\exp f(X)] \leq D(P\|Q)$ is very useful in approximating $E_P[f(X)]$ for complicated distribution P

Theorem 3.7 (weak lower-semicontinuity of divergence)**.** *Let $\mathcal{X}$ be a metric space with Borel σ-algebra $\mathcal{H}$. If P_n and Q_n converge weakly (in distribution) to P, Q, then*

$$D(P||Q) \leq \liminf_{n\to\infty} D(P_n||Q_n)\,. \tag{2}$$

Proof. <u>First method:</u> On metric $\mathcal{X}$ bounded continuous functions are dense in the set of all integrable functions. Then from Donsker-Varadhan (1) we have

$$D(P_n||Q_n) = \sup_{f\text{-bdd. cts.}} \mathbb{E}_{P_n}[f(X)] - \log \mathbb{E}_{Q_n}[\exp\{f(X)\}]\,.$$

Taking the limit concludes the proof.

<u>Second method</u> (less mysterious): Let $\mathcal{A}$ be the algebra of Borel sets E whose boundary has zero $(P+Q)$ measure, i.e.

$$\mathcal{A} = \{E \in \mathcal{H} : (P+Q)(\partial E) = 0\}\,.$$

By the property of weak convergence P_n and Q_n converge pointwise on $\mathcal{A}$. Thus by (6) we have

$$D(P_{\mathcal{A}}||Q_{\mathcal{A}}) \le \lim_{n\to\infty} D(P_{n,\mathcal{A}}||Q_{n,\mathcal{A}})$$

If we show $\mathcal{A}$ is $(P+Q)$-dense in $\mathcal{H}$, we are done by (5). To get an idea, consider $\mathcal{X} = \mathbb{R}$. Then open sets are $(P+Q)$-dense in $\mathcal{H}$ (since finite measures are regular), while the algebra $\mathcal{F}$ generated by open intervals is $(P+Q)$-dense in the open sets. Since there are at most countably many points $a \in \mathcal{X}$ with $P(a)+Q(a) > 0$, we may further approximate each interval (a,b) whose boundary has non-zero $(P+Q)$ measure by a slightly larger interval from $\mathcal{A}$. $\square$

3.4 Variational characterizations of M.I.

Finally, note that since $I(X;Y) = D(P_{XY}||P_XP_Y)$, each of these properties has a natural counterpart for mutual information, e.g.

1. $$I(X;Y) = \sup_f \mathbb{E}[f(X,Y)] - \log \mathbb{E}[\exp\{f(X,\bar{Y})\}]\,,$$

   where $\bar{Y}$ is a copy of Y, independent of X and supremum is over bounded, or even bounded continuous functions.

2. If $(X_n, Y_n) \xrightarrow{d} (X,Y)$ converge in distribution, then

   $$I(X;Y) \le \liminf_{n\to\infty} I(X_n;Y_n)\,.$$

3. $$I(X;Y) = \sup_{\{E_i\}\times\{F_j\}} \sum_{i,j} P_{XY}[E_i \times F_j] \log \frac{P_{XY}[E_i \times F_j]}{P_X[E_i]P_Y[F_j]}\,,$$

   where supremum is over finite partitions of spaces $\mathcal{X}$ and $\mathcal{Y}$.[2]

4. (Monotone convergence):

   $$I(X^\infty;Y) = \lim_{n\to\infty} I(X^n;Y)\,.$$

[2] To prove this from (3) one needs to notice that algebra of measurable rectangles is dense in the σ-algebra on $\mathcal{X}\times\mathcal{Y}$.

3.5 Characterizations of divergence: Gelfand-Yaglom-Peres

Theorem 3.8. *Let P, Q be two probability measures on $\mathcal{X}$ with σ-algebra $\mathcal{F}$. Then*

$$D(P||Q) = \sup_{\{E_1,\ldots,E_n\}} \sum_{i=1}^{n} P[E_i] \log \frac{P[E_i]}{Q[E_i]}\,, \tag{3}$$

where supremum is over all finite $\mathcal{F}$-measurable partitions: $\bigcup_{j=1}^{n} E_j = \mathcal{X}, E_j \cap E_i = \emptyset$.

Proof. To show $D(P||Q)$ is indeed achievable, let $E_j = \{x : \log \frac{dP}{dQ} \in \epsilon \cdot [j - n/2, j + 1 - n/2)\}$, $j = 1, \ldots, n$ and let $\epsilon \to 0$, $n\epsilon \to \infty$. Conversely, fix a finite partition $E_1, \ldots E_n$ and let X be $\mathcal{X}$-valued random variable (an identity map $\mathcal{X} \to \mathcal{X}$). Denote by $Y = Y(X)$ the (random) index $j \in \{1, \ldots, n\}$ of the set E_j to which X belongs. Then

$$\begin{aligned} D(P||Q) &= D(P_X||Q_X) \\ &= D(P_{XY}||Q_{XY}) \\ &\geq D(P_Y||Q_Y) \\ &= \sum_i P[E_i] \log \frac{P[E_i]}{Q[E_i]}\,, \end{aligned}$$

where the inequality is by data-processing. □

Note that divergence $D(P||Q)$ implicitly depends on the σ-algebra $\mathcal{F}$. To emphasize the dependence on $\mathcal{F}$ we will write

$$D(P_{\mathcal{F}}||Q_{\mathcal{F}})$$

Comment: (3) can be taken as the definition of $D(P_{\mathcal{F}}||Q_{\mathcal{F}})$ for an arbitrary algebra of sets $\mathcal{F}$ and two positive additive set-functions P, Q on $\mathcal{F}$.

Corollary 3.1 (Measure-theoretic properties of divergence)**.** *Let P, Q be probability measures on the measurable space $(\mathcal{X}, \mathcal{H})$. Assume all algebras below are sub-algebras of $\mathcal{H}$. Then:*

- *(Monotonicity) If $\mathcal{F} \subseteq \mathcal{G}$ then*

$$D(P_{\mathcal{F}}||Q_{\mathcal{F}}) \leq D(P_{\mathcal{G}}||Q_{\mathcal{G}})\,. \tag{4}$$

- *Let $\mathcal{F}_1 \subseteq \mathcal{F}_2 \ldots$ be an increasing sequence of algebras and let $\mathcal{F} = \bigcup_n \mathcal{F}_n$ be their limit, then*

$$D(P_{\mathcal{F}_n}||Q_{\mathcal{F}_n}) \nearrow D(P_{\mathcal{F}}||Q_{\mathcal{F}})\,.$$

- *If $\mathcal{F}$ is $(P + Q)$-dense in $\mathcal{G}$ then*[3]

$$D(P_{\mathcal{F}}||Q_{\mathcal{F}}) = D(P_{\mathcal{G}}||Q_{\mathcal{G}})\,. \tag{5}$$

[3]Note: $\mathcal{F}$ is μ-dense in $\mathcal{G}$ if $\forall E \in \mathcal{G}, \epsilon > 0 \exists E' \in \mathcal{F}$ s.t. $\mu[E \Delta E'] \leq \epsilon$.

- *(Monotone convergence theorem) Let $\mathcal{F}_1 \subseteq \mathcal{F}_2 \ldots$ be an increasing sequence of algebras and let $\mathcal{F} = \bigvee_n \mathcal{F}_n$ be the σ-algebra generated by them, then*

$$D(P_{\mathcal{F}_n}||Q_{\mathcal{F}_n}) \nearrow D(P_{\mathcal{F}}||Q_{\mathcal{F}})\,.$$

  In particular,

$$D(P_{X^\infty}||Q_{X^\infty}) = \lim_{n\to\infty} D(P_{X^n}||Q_{X^n})\,.$$

- *(Lower-semicontinuity of divergence) If $P_n \to P$ and $Q_n \to Q$ pointwise on the algebra $\mathcal{F}$, then*[4]

$$D(P_{\mathcal{F}}||Q_{\mathcal{F}}) \leq \liminf_{n\to\infty} D(P_{n,\mathcal{F}}||Q_{n,\mathcal{F}})\,. \tag{6}$$

Proof. Straighforward applications of (3) and the observation that any algebra $\mathcal{F}$ is μ-dense in the σ-algebra $\sigma\{\mathcal{F}\}$ it generates, for any μ on $(\mathcal{X}, \mathcal{H})$.[5] □

Note: pointwise convergence on $\mathcal{H}$ is weaker than convergence in total variation and stronger than convergence in distribution (aka "weak convergence"). However, (6) can be extended to this mode of convergence (see Theorem 3.7).

[4] $P_n \to P$ pointwise on some algebra $\mathcal{F}$ if $\forall E \in \mathcal{F} : P_n[E] \to P[E]$.

[5] This may be shown by transfinite induction: to each ordinal ω associate an algebra $\mathcal{F}_\omega$ generated by monotone limits of sets from $\mathcal{F}_{\omega'}$ with $\omega' < \omega$. Then $\sigma\{\mathcal{F}\} = \mathcal{F}_{\omega_0}$, where ω_0 is the first ordinal for which $\mathcal{F}_\omega$ is a monotone class. But $\mathcal{F}$ is μ-dense in each $\mathcal{F}_\omega$ by transfinite induction.

4.1 Convexity of information measures

Theorem 4.1 (Divergence)**.** $(P, Q) \mapsto D(P||Q)$ *is convex.*

Proof. Let $X \in \{0, 1\}$, $P_X = [\lambda, 1-\lambda]$. Select two conditional kernels:

$$P_{Y|X=0} = P_0, P_{Y|X=1} = P_1 \tag{1}$$

$$Q_{Y|X=0} = Q_0, Q_{Y|X=1} = Q_1 \tag{2}$$

Conditioning increases divergence, hence

$$D(P_{Y|X}||Q_{Y|X}|P_X) \geq D(P_Y||Q_Y)$$

□

Lemma 4.1. *When* $D(P||Q) < \infty$

$$\frac{d}{d\lambda}\Big|_{\lambda=0} D(\lambda P + \bar{\lambda} Q||Q) = 0$$

Proof.

$$\frac{1}{\lambda} D(\lambda P + \bar{\lambda} Q||Q) = \mathbb{E}_Q \left[\frac{1}{\lambda}(\lambda f + \bar{\lambda}) \log(\lambda f + \bar{\lambda}) \right]$$

where $f = \frac{dP}{dQ}$. As $\lambda \to 0$ the function under expectation converges to $(f-1)\log e$. Moreover,

$$\frac{1}{\lambda}(\lambda f + \overline{\lambda})(\log(\lambda f + \overline{\lambda})) \leq \frac{1}{\lambda}(\lambda f \log f + \overline{\lambda} 1 \log 1) = f \log f$$

and thus the Dominated Convergence Theorem applies. □

Note: More generally:

$$\frac{d}{d\lambda}\Big|_{\lambda=0} D(\lambda P + \overline{\lambda} Q||R) = \mathbb{E}_P \left[\log \frac{dQ}{dR} \right] - D(Q||R)\,.$$

Typed by Rachel Cohen rcohen@mit.edu, Austin Collins austinc@mit.edu, Feb 14, 2013

Theorem 4.2 (Entropy)**.** *$P_X \mapsto H(P_X)$ — concave*

Proof. If P_X is on a finite alphabet, then proof is complete by $H(X) = \log|\mathcal{X}| - D(P_X||U_X)$. Otherwise, set

$$P_{X|Y} = \begin{cases} P_0 & Y = 0 \\ P_1 & Y = 1 \end{cases}, \quad P_Y(Y=0) = \lambda$$

Then apply $H(X|Y) \le H(X)$. □

Theorem 4.3 (Mutual Information)**.** *$(P_X, P_{Y|X}) \mapsto I(P_X, P_{Y|X})$ is concave in P_X and convex in $P_{Y|X}$.*

Proof.

$$\begin{aligned} I(X;Y) &= D(P_{Y|X}||P_Y|P_X) \\ &= D(P_{Y|X}||Q|P_X) - D(P_Y||Q) \end{aligned}$$

□

4.2 Extremization of mutual information

Two problems of interest

- Fix $P_{Y|X} \to \max_{P_X} I(X;Y)$.
  Note: This maximum is called "capacity" of a set of distributions $\{P_{Y|X=x}, x \in \mathcal{X}\}$.
- Fix $P_X \to \min_{P_{Y|X}} I(X;Y)$

Theorem 4.4 (Saddle point)**.** *Let $\mathcal{P}$ be a convex set of distributions on $\mathcal{X}$. Suppose there exists P_X^* such that*

$$\sup_{P_X \in \mathcal{P}} I(P_X, P_{Y|X}) = I(P_X^*, P_{Y|X}) \triangleq C$$

and let $P_X^* \overset{P_{Y|X}}{\longmapsto} P_Y^*$. Then for all $P_X \in \mathcal{P}$ and for all Q_Y we have

$$D(P_{Y|X}||P_Y^*|P_X) \le D(P_{Y|X}||P_Y^*|P_X^*) \le D(P_{Y|X}||Q_Y|P_X^*)$$

Note: P_X^* (resp., P_Y^*) is called a capacity-achieving input (resp., output) distribution, or a caid (resp., the caod).

Proof. RHS is obvious from $C = D(P_{Y|X}||Q_Y|P_X^*) - D(P_Y^*||Q_Y)$.

For the LHS, let $P_{X_\lambda} = \lambda P_X + \overline{\lambda} P_X^*$, and introduce $\theta \sim \mathrm{Bern}(\lambda)$, so that $\theta \to X \to Y$, then

$$\begin{aligned} C \ge I(X_\lambda; Y_\lambda) &= I(\theta, X_\lambda; Y_\lambda) = I(\theta; Y_\lambda) + I(X_\lambda; Y_\lambda|\theta) \\ &= D(P_{Y|\theta}||P_{Y_\lambda}|P_\theta) + \lambda I(P_X, P_{Y|X}) + \overline{\lambda} C \\ &= \lambda D(P_Y||P_{Y_\lambda}) + \overline{\lambda} D(P_Y^*||P_{Y_\lambda}) + \lambda I(P_X, P_{Y|X}) + \overline{\lambda} C \\ &\ge \lambda D(P_Y||P_{Y_\lambda}) + \lambda I(P_X, P_{Y|X}) + \overline{\lambda} C \\ \lambda(C - I(P_X, P_{Y|X})) &\ge \lambda D(P_Y||P_{Y_\lambda}) \end{aligned}$$

Dividing through by λ, taking the lim inf and using lower semicontinuity of D

$$\begin{aligned}
\liminf_{\lambda\to 0} D(P_Y||P_{Y_\lambda}) &\le C - I(P_X, P_{Y|X}) \\
\implies D(P_Y||P_Y^*) &\le C - D(P_{Y|X}||P_Y|P_X) \\
&= C - D(P_{Y|X}||P_Y^*|P_X) + D(P_Y||P_Y^*) \\
D(P_{Y|X}||P_Y^*|P_X) &\le D(P_{Y|X}||P_Y^*|P_X^*)
\end{aligned}$$

□

Corollary 4.1. *Caod P_Y^* is unique.*

Proof. The statement is: $I(P_X, P_{Y|X}) = C \Rightarrow P_Y = P_Y^*$. Indeed:

$$\begin{aligned}
C = D(P_{Y|X}||P_Y|P_X) &= D(P_{Y|X}||P_Y^*|P_X) - D(P_Y||P_Y^*) \\
&\le D(P_{Y|X}||P_Y^*|P_X^*) - D(P_Y||P_Y^*) \\
&= C - D(P_Y||P_Y^*) \Rightarrow P_{Y_1} = P_Y^*
\end{aligned}$$

□

Corollary 4.2 (Minimax)**.**

$$\begin{aligned}
\max_{P_X\in\mathcal{P}} I(X;Y) &= \max_{P_X\in\mathcal{P}} \min_{Q_Y} D(P_{Y|X}||Q_Y|P_X) \\
&= \min_{Q_Y} \max_{P_X\in\mathcal{P}} D(P_{Y|X}||Q_Y|P_X)
\end{aligned}$$

Proof. Follows from saddle-point trivially. □

4.3 Capacity = information radius

Corollary 4.3. *Let $\mathcal{P} = \{$all dist. on $\mathcal{X}\}$ and $\mathcal{X}$-finite, then*

$$\begin{aligned}
\max_X I(X;Y) &= \max_x D(P_{Y|X=x}||P_Y^*) \\
&= D(P_{Y|X=x}||P_Y^*) \qquad \forall x : P_X^*(x) > 0 \ .
\end{aligned}$$

The last corollary gives geometric interpretation to capacity: it equals the radius of the smallest divergence-ball that encompassess all distributions $\{P_{Y|X=x}, x \in \mathcal{X}\}$. It turns out this characterization is true even without the assumption of existence of the maximizer P_X^* of $I(P_X, P_{Y|X})$ and in the presence of (convex) input-distribution constraints.

Theorem 4.5. *For any $P_{Y|X}$ and a convex set of distributions $\mathcal{P}$ such that*

$$C = \sup_{P_X\in\mathcal{P}} I(P_X, P_{Y|X}) < \infty \tag{3}$$

there exists a unique P_Y^* with the property that

$$C = \sup_{P_X\in\mathcal{P}} D(P_{Y|X}||P_Y^*|P_X) \, . \tag{4}$$

In particular,

$$C = \min_{Q_Y} \sup_{P_X \in \mathcal{P}} D(P_{Y|X}||Q_Y|P_X) \tag{5}$$

$$= \min_{Q_Y} \sup_{x \in \mathcal{X}} D(P_{Y|X=x}||Q_Y)\,, \qquad \textit{(when } \mathcal{P} = \textit{“all dist.”)} \tag{6}$$

Remark: Condition (3) is automatically satisfied if there is any Q_Y such that

$$\sup_{P_X \in \mathcal{P}} D(P_{Y|X}||Q_Y|P_X) < \infty\,. \tag{7}$$

Example: Let $Z_j = \pm 1$ be $\mathrm{Bern}(2^{-j})$, $\mathcal{X} = \mathbb{Z}\backslash\{0\}$ and $Y = \mathrm{sign}(X)\cdot Z_{|X|} \in \{-1, 1\}$. Then, $C = \log 2$ but supremum in (3) is not attained. Nevertheless, (4) holds with P_Y^* uniform on $\{-1, 1\}$.

Proof. Let P'_{X_n} be a sequence of input dist. achieving C and let $\mathcal{P}_n$ be a convex hull of $\{P'_{X_1}, \ldots, P'_{X_n}\}$. Since $\mathcal{P}_n$ is a finite-dimensional simplex, concave function $I(P_X)$ attains its maximum at some point $P_{X_n} \in \mathcal{P}_n$. I.e.

$$I_n \triangleq I(P_{X_n}, P_{Y|X}) = \max_{P_X \in \mathcal{P}_n} I(P_X, P_{Y|X})\,.$$

Denote by P_{Y_n} be the sequence of output distributions corresponding to P_{X_n}. We have then:

$$D(P_{Y_n}||P_{Y_{n+k}}) = D(P_{Y|X}||P_{Y_{n+k}}|P_{X_n}) - D(P_{Y|X}||P_{Y_n}|P_{X_n}) \tag{8}$$

$$\leq I(P_{X_{n+k}}, P_{Y|X}) - I(P_{X_n}, P_{Y|X}) \tag{9}$$

$$\leq C - I_n\,, \tag{10}$$

where in (9) we applied Theorem 4.4 to $(\mathcal{P}_{n+k}, P_{Y_{n+k}})$. Pinsker-Csiszár inequality bounds total variation distance in terms of divergence:

$$||P - Q||_{TV} \leq \sqrt{2 \log e D(P||Q)}\,.$$

Thus, since $I_n \nearrow C$ we conclude that the sequence P_{Y_n} is Cauchy in total variation:

$$\sup_{k \geq 1} ||P_{Y_n} - P_{Y_{n+k}}||_{TV} \to 0\,, \qquad n \to \infty\,.$$

Since the space of distributions is complete in total variation, the sequence must have a limit point $P_{Y_n} \to P_Y^*$. By taking a limit as $k \to \infty$ in (10) and applying lower-semicontinuity of divergence we get

$$D(P_{Y_n}||P_Y^*) \leq \lim_{k \to \infty} D(P_{Y_n}||P_{Y_{n+k}}) \leq C - I_n\,,$$

and therefore, $P_{Y_n} \to P_Y^*$ in the (stronger) sense of $D(P_{Y_n}||P_Y^*) \to 0$. Therefore,

$$D(P_{Y|X}||P_Y^*|P_{X_n}) = I_n + D(P_{Y_n}||P_Y^*) \to C\,. \tag{11}$$

Take any $P_X \in \bigcup_{k \geq 1} \mathcal{P}_k$. Then $P_X \in \mathcal{P}_n$ for all $n \geq n_0$ and thus by Theorem 4.4

$$D(P_{Y|X}||P_{Y_n}|P_X) \leq I_n \leq C\,,$$

which by lower-semicontinuity of divergence implies

$$D(P_{Y|X}||P_Y^*|P_X) \leq C\,. \tag{12}$$

Finally, to prove that (12) holds for arbitrary $P_X \in \mathcal{P}$, we may repeat the argument above with $\mathcal{P}_n$ replaced by $\tilde{\mathcal{P}}_n = \mathrm{conv}(P_X \cup \mathcal{P}_n)$, denoting the resulting sequences by $\tilde{P}_{X_n}, \tilde{P}_{Y_n}$ and the limit

point by $\tilde{P}_Y^*$ we have:

$$\begin{aligned} D(P_{Y_n}||\tilde{P}_{Y_n}) &= D(P_{Y|X}||\tilde{P}_{Y_n}|P_{X_n}) - D(P_{Y|X}||P_{Y_n}|P_{X_n}) && (13)\\ &\leq C - I_n\,, && (14)\end{aligned}$$

where (14) follows from (12) since $P_{X_n} \in \tilde{\mathcal{P}}_n$. Hence taking limit as $n \to \infty$ we have $\tilde{P}_Y^* = P_Y^*$ and therefore (12) holds. □

4.4 Gaussian Saddle Point

For additive noise, there is also a different kind of saddle point between P_X and the distribution of noise:

Theorem 4.6. *Let $X_g \sim \mathcal{N}(0, \sigma_x^2)$, $N_g \sim \mathcal{N}(0, \sigma_N^2)$, $X_g \perp\!\!\!\perp N_g$. Then:*

1. *"Gaussian capacity":*
$$C = I(X_g; X_g + N_g) = \frac{1}{2}\log\left(1 + \frac{\sigma_x^2}{\sigma_N^2}\right)$$
2. *"Gaussian input is the best":*
$$I(X; X + N_g) \leq I(X_g; X_g + N_g) \qquad \textit{(eq. iff } X \stackrel{d}{=} X_g\textit{)}$$
$$\forall X \perp\!\!\!\perp N_g\,, \qquad \mathrm{Var}\ X \leq \sigma_x^2$$
3. *"Gaussian noise is the worst":*
$$I(X_g; X_g + N) \geq I(X_g; X_g + N_g) \qquad \textit{(eq. iff } N \stackrel{d}{=} N_g\textit{)}$$
$$\forall N : \mathbb{E}[X_g N] = 0\ , \mathbb{E}N^2 \leq \sigma_N^2$$

Proof. Let $Y_g = X_g + N_g$.

$$g(x) = D(P_{Y_g|X_g=x}||P_{Y_g}) = \frac{1}{2}\log\left(1 + \frac{\sigma_x^2}{\sigma_N^2}\right) + \frac{\log e}{2}\frac{x^2 - \sigma_x^2}{\sigma_x^2 - \sigma_N^2}$$

1. Compute $\mathbb{E}[g(X_g)] \Rightarrow$ 1. proved
2. $I(X; X + N_g) = \mathbb{E}[g(X)] - D(P_Y||P_{Y_g}) \Rightarrow$ 2. proved.
   Alternatively, by the information-radius characterization (5) we always have
$$I(X; X + N_g) \leq D(P_{Y_g|X_g}||P_{Y_g}|P_X) = \mathbb{E}[g(X)]\,.$$
3. Let $Y = X_g + N$ and let $P_{Y|X_g}$ be the respective kernel. Then
$$\begin{aligned} I(X_g; X_g + N) &= D(P_{Y|X_g}||P_{Y_g|X_g}|P_{X_g}) - D(P_Y||P_{Y_g}) + \mathbb{E}\left[\log\frac{p_{Y_g|X_g}(Y|X_g)}{p_{Y_g}(Y)}\right] && (15)\\ &\geq 0 + \mathbb{E}\left[\log\frac{p_{Y_g|X_g}(Y|X_g)}{p_{Y_g}(Y)}\right] && (16)\\ &= C + \frac{\log e}{2}\,\mathbb{E}\left[\frac{Y^2}{\sigma_x^2 + \sigma_N^2} - \frac{N^2}{\sigma_N^2}\right] && (17)\\ &\geq C\,, && (18)\end{aligned}$$

where in (15) $p_{Y_g|X_g}$ and p_{Y_g} are pdfs of respective measures, (16) follows by "conditioning increases divergence", (17) is elementary and (18) is by assumptions on N.

Finally, the conditions for equality in (16) say

$$D(P_{X_g|Y}||P_{X_g|Y_g}|P_Y) = 0$$

Thus, $P_{X_g|Y} = P_{X_g|Y_g}$ – the X_g is conditionally Gaussian: $P_{X_g|Y=y} = \mathcal{N}(by, b)$. In other words, under P_{X_gY}, we have

$$X_g = bY + \sqrt{b}Z \quad , \quad Z \sim \text{Gaussian} \perp\!\!\!\perp Y$$

But then Y must be Gaussian itself by Cramer's Theorem or simply by considering characteristic functions:

$$\Psi_Y(t) \cdot e^{ct^2} = e^{c't^2} \Rightarrow \Psi_Y(t) = e^{c''t^2} \implies \text{Y– Gaussian}$$

Thus, it is only possible to attain $I(X_g; X_g + N) = C$ if N is Gaussian of variance σ_N^2.

□

4.4.1 Proof via symmetry method

Consider a multi-dimensional analog of maximizing mutual information over AWGN channel:

$$\max_{\mathbb{E}[\sum X_k^2] \le nP} I(X^n; X^n + Z^n)$$

Since the channel is memoryless (see Lect. 5):

$$\max_{\mathbb{E}[\sum X_k^2] \le nP} I(X^n; X^n + Z^n) \le \max_{\mathbb{E}[\sum X_k^2] \le nP} \sum_{k=1}^n I(X_k; X_k + Z_k)$$

And thus the optimal distribution must be $X_1 \perp\!\!\!\perp \ldots \perp\!\!\!\perp X_n$.

Given a distribution $P_{X_1} \cdots P_{X_n}$ satisfying the constraint, form the "average of marginals" distribution $\bar{P}_X = \frac{1}{n}\sum_{k=1}^n P_{X_k}$, which also satisfies the single letter constraint $\mathbb{E}[X^2] = \frac{1}{n}\sum_{k=1}^n \mathbb{E}[X_k^2] \le P$. Then from concavity in P_X of $I(P_X, P_{Y|X})$

$$I(\bar{P}_X; P_{Y|X}) \ge \frac{1}{n}\sum_{k=1}^n I(P_{X_k}, P_{Y|X})$$

So $\bar{P}_X$ gives the same or better MI, which shows that the extremization above ought to have the form $nC(P)$ where $C(P)$ is the single letter capacity:

$$\max_{\mathbb{E}[\sum X_k^2] \le nP} I(X^n; X^n + Z^n) = nC(P)$$

Now suppose $Y^n = X^n + Z_G^n$ where $Z_G^n \sim \mathcal{N}(0, \mathbf{I}_n)$. Since an isotropic Gaussian is rotationally symmetric, for any orthogonal transformation $U \in O(n)$, the additive noise has the same distribution $Z_G^n \sim UZ_G^n$, so that $P_{UY^n|UX^n} = P_{Y^n|X^n}$, and

$$I(P_{X^n}, P_{Y^n|X^n}) = I(P_{UX^n}, P_{UY^n|UX^n}) = I(P_{UX^n}, P_{Y^n|X^n})$$

Averaging over all rotations U and applying concavity in P_{X^n} we figure out that the optimal distribution $P_{X^n}^*$ must be spherically symmetric (invariant under orthogonal transformations). The only

distribution in $\mathbb{R}^n$ that is iid and spherically symmetric is Gaussian (just think of the characteristic function). Thus, we conclude that $P_X^* \sim \mathcal{N}(0, P\mathbf{I}_n)$.

For the other direction in the Gaussian saddle point problem:

$$\min_{P_N : \mathbb{E}[N^2]=1} I(X_G; X_G + N)$$

Again, we may first go to n-dimensional case, and average over all rotations of the noise P_{N^n}, except this time we apply concavity of mutual information in $P_{Y|X}$.

5.1 Extremization of Mutual Information for Memoryless iid Sources

Theorem 5.1. *(Joint mutual information vs. marginal)*

(1) If $P_{X^n|Y^n} = \prod P_{X_i|Y_i}$ *then*

$$I(X^n;Y^n) \leq \sum I(X_i;Y_i) \quad \textit{or}$$
$$\textit{equality iff } P_{Y^n} = \prod P_{Y_i}$$

(2) If $X_1 \perp\!\!\!\perp \ldots \perp\!\!\!\perp X_n$ *then*

$$I(X^n;Y^n) \geq \sum I(X_i;Y_i) \quad \textit{or}$$
$$\textit{equality iff } P_{X^n|Y^n} = \prod P_{X_i|Y_i}$$

Proof.

(1) $I(X^n;Y^n) - \sum I(X_j;Y_j) = D(P_{X^n|Y^n} || \prod P_{X_i|Y_i})$

(2) same.

□

Example:

1. X_1, X_2 – Uniform on $\{0,1\} = \mathbb{F}_2$:

$$\begin{aligned} Y_1 &= X_1 + X_2 \\ Y_2 &= X_1 \\ I(X_1;Y_1) &= I(X_2;Y_2) = 0 \text{ but } I(X^2;Y^2) = 2 \text{ bits} \end{aligned}$$

2. $X_0 \perp\!\!\!\perp \ldots \perp\!\!\!\perp X_{n-1}$

$$Y_k = X_{k-1} \mod n \quad \Rightarrow \quad I(X_k;Y_k) = 0$$
$$I(X^n;Y^n) = \sum H(X_i) \geq \sum I(X_k;Y_k)$$

3.

$$\forall k \; Y_k = X_k = U \sim \text{Bern}(1/2) \quad \Rightarrow \quad I(X_k;Y_k) = 1$$
$$I(X^n;Y^n) = 1 \leq \sum I(X_k;Y_k)$$

Typed by Rachel Cohen rcohen@mit.edu, Austin Collins austinc@mit.edu, Feb 21, 2013

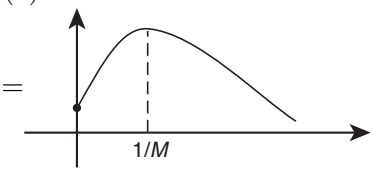
5.2 Information measures and probability of error

Theorem 5.2. *Let $M = |\mathcal{X}| < \infty$ and $P_{\max} \triangleq \max_{x\in\mathcal{X}} P_X(x)$. Then*

$$H(X) \le (1 - P_{\max}) \log(M-1) + h(P_{\max}) \triangleq F_M(P_{\max}) \tag{1}$$

Proof. Consider p.m.f. + averaging + concavity. □

The function of $P_{\max}$ in the RHS of (1) is

$F_M =$ 

—unimodal with max at $1/M$, but not monotone. However, $P_{\max} \ge \frac{1}{M}$.

Interpretation: Suppose one is trying to estimate the value of X, the optimal performance among all estimators

$$\max_{\hat{X} \perp\!\!\!\perp X} \mathbb{P}[X = \hat{X}] = P_{\max}$$

Thus (1) means: It is hard to predict something of large entropy.

Conceptual question: Is it true (for every predictor $\hat{X} \perp\!\!\!\perp X$) that

$$H(X) \le F_M(\mathbb{P}[X = \hat{X}])\,?$$

Not obvious from (1) since $p \mapsto F_M(p)$ is not monotone. To show (5.2) consider:

$$\begin{array}{rclcrcl} P_{X\hat{X}} & = & P_X P_{\hat{X}} & & \mathbb{P}[X=\hat{X}] & \triangleq & P_S \\ Q_{X\hat{X}} & = & U_X P_{\hat{X}} & \Rightarrow & \mathbb{Q}[X=\hat{x}] & = & \frac{1}{M} \end{array} \quad \Rightarrow \quad \begin{array}{rcl} d\Big(P_S || \frac{1}{M}\Big) & \le & D(P_{X\hat{X}} || Q_{X\hat{X}}) \\ & = & \log M - H(X) \end{array}$$ □

where inequality follows by the data-processing for divergence. The benefit of this proof is that it trivially generalizes to estimators $\hat{X}(Y)$, which depend on some observation Y correlated with X:

Theorem 5.3 (Fano inequality: uniform X)**.** *Let $|\mathcal{X}| < \infty$ then for any $\hat{X}(Y)$ – a (possibly randomized) function of Y we have*

$$H(X|Y) \le F_M(\mathbb{P}[X = \hat{X}(Y)])$$

Proof. P_{XY} vs. $U_X P_Y$ + apply data processing. □

Remark: We can also derive Fano's Inquality by starting with a uniform X, then analyzing the $\min_{P_{Y|X}} I(P_X, P_{Y|X})$ with the constraint that X and Y agree with $\mathbb{P}[A = B] \ge 1 - \epsilon$ (Hint: look for invariants that the matrix $P_{Y|X}$ must satisfy under permutations $(X, Y) \mapsto (\pi(X), \pi(Y))$.

Theorem 5.4 (Fano inequality: general X)**.** *Let $X, Y \in \mathcal{X}$, $|\mathcal{X}| = M$ and let $Q_{XY} = P_X P_Y$, then*

$$\begin{aligned} I(X;Y) &\ge d(\mathbb{P}[X=Y] || \mathbb{Q}[X=Y]) \\ &\ge \mathbb{P}[X=Y] \log \frac{1}{\mathbb{Q}[X=Y]} - h(\mathbb{P}[X=Y]) \\ &(= F_{M+1}(\mathbb{P}[X=Y]) \quad \textit{if } P_X \textit{ or } P_Y = \textit{uniform}) \end{aligned}$$

Corollary 5.1. *Let $W \to X \to Y \to \hat{W}$ and W – uniform on $[M] \triangleq \{1, \ldots, M\}$. Then*

$$\begin{aligned} P_e &\geq 1 - \frac{I(X;Y) + h(P_e)}{\log M} && (2) \\ &\geq 1 - \frac{I(X;Y) + \log 2}{\log M}\,. && (3) \end{aligned}$$

where $P_e = \mathbb{P}[W \neq \hat{W}]$.

5.3 Entropy Rates

Theorem 5.5. *For every statistical source:*

1. $H(X_n|X^{n-1}) \leq H(X_{n-1}|X^{n-2})$
2. $\frac{1}{n}H(X^n) \geq H(X_n|X^{n-1})$
3. $\frac{1}{n}H(X^n) \leq \frac{1}{n-1}H(X^{n-1})$
4. $H(\mathbb{X}) \triangleq \lim_{n\to\infty} \frac{1}{n}H(X^n) = \lim_{n\to\infty} H(X_n|X^{n-1})$
5. $H(\mathbb{X}) = H(X_1|X_{-\infty}^0)$

Proof.

1. Further conditioning + stationarity
2. $\frac{1}{n}H(X^n) = \frac{1}{n}\sum H(X_i|X^{i-1}) \geq H(X_n|X^{n-1})$
3. $H(X^n) = H(X^{n-1}) + H(X_n|X^{n-1}) \leq H(X^{n-1}) - \frac{1}{2}H(X^n)$
4. Suppose

$$\begin{aligned} \lim_{n\to\infty} \frac{1}{n}H(X^n) &= Q > \lim_{n\to\infty} H(X_n|X^{n-1}) \\ &\Rightarrow \text{all except first few } M(X_i|X^{i-1}) \leq Q - \mathcal{E} \\ &\Rightarrow \frac{1}{n}\sum H(X_i|X^{i-1}) \leq \frac{1}{n}[kH(X_0) + (n-k)(Q - \mathcal{E})] \\ &- \text{contradiction!} \end{aligned}$$

5. $\lim_{n\to\infty} H(X_1) - H(X_1|X_{-n}^0) = \lim_{n\to\infty} I(X_1; X_{-n}^0) = I(X_1; X_{-\infty}^0) = H(X_1) - H(X_1|X_{-\infty}^0)$

□

Example:

1. $\mathbb{X}$ – iid source $\Rightarrow H(\mathbb{X}) = H(X_1)$

2. $\mathbb{X}$ – Markov (stat.) : $\quad X_1 \to X_2 \to X_3 \to \cdots$

$$H(X_n|X^{n-1}) = H(X_n|X_{n-1}) \Rightarrow H(\mathbb{X}) = H(X_2|X_1) = \sum_{a,b} \mu(a) P_{b|a} \log \frac{1}{P_{b|a}}$$

where μ is an invariant measure (possibly non-unique).

Definition 5.1. *Mutual information rate:*

$$I(\mathbb{X};\mathbb{Y}) = \lim_{n\to\infty} \frac{1}{n} I(X^n;Y^n)$$

Example: Consider $\mathbb{X}, \mathbb{N}$ – discrete stationary processes independent of each other, taking values on finite Abelian group G. Then $\mathbb{Y} = \mathbb{X} + \mathbb{N}$ is also stationary. In this case

$$I(\mathbb{X};\mathbb{X}+\mathbb{N}) = H(\mathbb{X}+\mathbb{N}) - H(\mathbb{N})\,.$$

Example: Consider $\mathbb{X}, \mathbb{N}$ two stationary Gaussian processes, independent of each other. Assume that their auto-covariance functions are summable and thus there exist continuous power spectral density functions f_X and f_N. Finally, assume $f_N \geq \delta > 0$. Then:

$$\begin{aligned}
I(X^n; X^n + N^n) &= D(P_{Y^n|X^n} || P_{Y^n} | P_{X^n}) \\
&= \mathbb{E}_{X^n}[\log\det\Sigma_{Y^n} - \log\det\Sigma_{N^n} + \mathrm{tr}(\Sigma_{Y^n}^{-1}\Sigma_{N^n} - I)\log e + (X^n, \Sigma_{Y^n}^{-1} X^n)\log e] \\
&= \log\det\Sigma_{Y^n} - \log\det\Sigma_{N^n} + \mathrm{tr}(\Sigma_{Y^n}^{-1}\Sigma_{N^n} - I) + (\Sigma_{Y^n}^{-1}\Sigma_{X^n}) \\
&= \log\det\Sigma_{Y^n} - \log\det\Sigma_{N^n} + \underbrace{\mathrm{tr}(\Sigma_{Y^n}^{-1}\Sigma_{Y^n} - I)\log e}_{=0} \\
&= \sum_{i=1}^n \log\sigma_i - \sum_{i=1}^n \log\lambda_i\,,
\end{aligned}$$

where σ_j, λ_j are the eigenvalues of the covariance matrices Σ_{Y^n} and Σ_{N^n}. By Szegö's theorem (see below):

$$\frac{1}{n}\sum_{i=1}^n \log\sigma_i \to \frac{1}{2\pi}\int_0^{2\pi} \log f(x)dx$$

Thus, we have

$$\frac{1}{n} I(X^n; X^n + N^n) \to I(\mathbb{X};\mathbb{X}+\mathbb{N}) = \frac{1}{2\pi}\int_0^{2\pi} \log\frac{f_X(w) + f_N(\omega)}{f_N(\omega)} d\omega$$

(Note: maximizing this over $f_X(\omega)$ leads to the famous waterfilling solution $f_X^*(\omega) = |T - f_N(\omega)|^+$.)

Theorem 5.6 (Szegö)**.** *Let $f : [0, 2\pi) \to \mathbb{R}$ be the Fourier transform of a summable sequence a_k, that is*

$$f(\omega) = \sum_{k=-\infty}^{\infty} e^{ik\omega} a_k\,, \qquad \sum |a_k| < \infty$$

Then for any $\phi : \mathbb{R} \to \mathbb{R}$ continuous on the closure of the range of f, we have

$$\lim_{n\to\infty} \frac{1}{n}\sum_{j=1}^n \phi(\sigma_{n,j}) = \frac{1}{2\pi}\int_0^{2\pi} \phi(f(\omega))d\omega\,,$$

where $\{\sigma_{n,j}, j = 1, \ldots, n\}$ are the eigenvalues of the Toeplitz matrix $T_n = \{a_{\ell-m}\}_{\ell,m=1}^n$.

Proof. The idea is to approximate ϕ by polynomials, while for polynomials the statement is checked directly. E.g. for $\phi(x) = x^2$ we have

$$\begin{aligned}
\frac{1}{n}\sum_{j=1}^{n}\sigma_{n,j}^2 &= \frac{1}{n}\operatorname{tr} T_n^2 \\
&= \frac{1}{n}\sum_{\ell,m=1}^{n}(T_n)_{\ell,m}(T_n)_{m,\ell} \\
&= \frac{1}{n}\sum_{\ell,m} a_{\ell-m}a_{m-\ell} \\
&= \frac{1}{n}\sum_{\ell=-n-1}^{n-1}(n-|\ell|)a_\ell a_{-\ell} \\
&= \sum_{x\in(-1,1)\cap\frac{1}{n}\mathbb{Z}}(1-|x|)a_{nx}a_{-nx}\,,
\end{aligned}$$

Substituting $a_\ell = \frac{1}{2\pi}\int_0^{2\pi} f(\omega)e^{i\omega\ell}$ we get

$$\frac{1}{n}\sum_{j=1}^{n}\sigma_{n,j}^2 = \frac{1}{(2\pi)^2}\int\int f(\omega)f(\omega')\theta_n(\omega-\omega')\,, \tag{4}$$

where

$$\theta_n(u) = \sum_{x\in(-1,1)\cap\frac{1}{n}\mathbb{Z}}(1-|x|)e^{-inux}$$

converges to a δ-function: $\theta_n(u) \to 2\pi\delta(u)$. Thus from (4) we get

$$\frac{1}{n}\sum_{j=1}^{n}\sigma_{n,j}^2 \to \frac{1}{(2\pi)^2}\int\int f(\omega)f(\omega')2\pi\delta(\omega-\omega')d\omega d\omega' = \frac{1}{2\pi}\int_0^{2\pi} f^2(\omega)d\omega$$

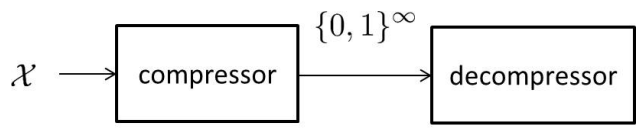
as claimed.

6.441: Information Theory	Spring 2012
Lecturer: Yury Polyanskiy	**Lecture 6: compression**

Idea of compression:
for $X \in \mathcal{A}^n \to \{0,1\}^K$, if R.V. $X \sim P_X$, there exists efficient representation (min # bits)

- Lossy
  $X \to W \to \hat{X}$ s.t. $\mathbb{E}[(X-\hat{X})^2] \leq$ distortion.
- Lossless
  $P(X \neq \hat{X}) = 0$. variable length code (prefix codes (Huffman), one shot codes)
- Almost lossless (fixed blocklength)
  $P(X \neq \hat{X}) \leq \epsilon$.

3.1 variable length, lossless, optimal compressor



$f : \mathcal{X} \to \{0,1\}^\infty \qquad g : \{0,1\}^\infty \to \mathcal{X}$

Note:

- Since $\{0,1\}^\infty = \{\phi, 0, 1, 00, 01, \dots\}$ is countable, lossless compression is only possible for dicrete R.V.;
- if we want $g \circ f = 1_{\mathcal{X}}$ (lossless), then f must be injective;
- relabel $\mathcal{X}$ such that $\mathcal{X} = \mathbb{Z}_+ = \{1, 2, \dots\}$ and $P_X(i) \geq P_X(i+1)$.

Length function:

$$l : \{0,1\}^\infty \to \mathbb{Z}_+$$

Measures to minimize

$$\mathbb{E}[l(f(X))]$$
$$\sup l(f(X))$$
$$\text{median} l(f(X))$$

there is a compressor f^* that minimizes all together. Idea: assign shorter length codes to high probability symbols.

Typed by qqh@mit.edu, Feb 22, 2013

Theorem 3.1 (optimal f^*)**.** *Consider the compressor* f^*

i	1	2 3	4 5 6
$f^*(i)$	ϕ	$\{0,1\}$	$\{(0,0),(0,1),(1,0),(1,1)\}$

$$l(f^*(i)) = \lfloor \log_2 i \rfloor$$

$$\forall f : P[l(f(X)) \leq k] \leq P[l(f^*(X)) \leq k]$$

Proof. show by contradiction □

Theorem 3.2 (expected code length $\mathbb{E}\left[l(f^*(X))\right]$)**.**

$$H(X) - \log_2[e(H(X)+1)] \leq \mathbb{E}[l(f^*(X))] \leq H(X)$$

Proof. Define $L(X) = l(f^*(X)))$,
RHS: observe that $P_X(m) \leq 1/m$, so $L(m) \leq log_2 m \leq \log_2(1/P_X(m))$, take exp., $\mathbb{E}[L(X)] \leq H(X)$.
LHS:

$$\begin{aligned}
H(X) = H(X,L) &= H(X|L) + H(L) \\
&\leq \mathbb{E}[L] + h(\frac{1}{1+\mathbb{E}[L]})(1+\mathbb{E}[L]) \qquad (\max H(L) \text{ s.t. } \mathbb{E}[L] \text{is fixed}) \\
&= \mathbb{E}[L] + \log_2(1+\mathbb{E}[L]) + \mathbb{E}[L]\log(1+\frac{1}{\mathbb{E}[L]}) \\
&\leq \mathbb{E}[L] + \log_2(1+\mathbb{E}[L]) + \log_2 e \\
&\leq \mathbb{E}[L] + \log(e(1+H(X)) \text{ by RHS}
\end{aligned}$$

□

Note: if S^n is i.i.d. sequence, then

$$\mathbb{E}[l(f^*(S^n))] \geq nH(S) - \log n + O(1)$$

However, this lower bound is not tight, can improve to $nH(S) - \frac{1}{2}\log n$.

Theorem 3.3 (code length distribution of f^*)**.**

$$\forall \tau, k \quad P[\log_2(\frac{1}{P_X(X)}) \leq k] \leq P[l(f^*(X)) \leq k] \leq P[\log_2(\frac{1}{P_X(X)}) \leq k+\tau] + 2^{-\tau+1}$$

Proof. LHS: easy, use $P_X(m) \leq 1/m$
RHS: (truncation)

$$\begin{aligned}
P[l \leq k] &= P[l \leq k, \log_2(\frac{1}{P_X(X)}) \leq k+\tau] + P[l \leq k, \log_2(\frac{1}{P_X(X)}) > k+\tau] \\
&\leq P[\log_2(\frac{1}{P_X(X)}) \leq k+\tau] + \sum_{x\in\mathcal{X}} P_X(x)\mathbf{1}[l(f^*(x)) \leq k]\mathbf{1}[P_X(x) \leq 2^{-k-\tau}] \\
&\leq P[\log_2(\frac{1}{P_X(X)}) \leq k+\tau] + 2^{k+1}\cdot 2^{-k-\tau}
\end{aligned}$$

□

Corollary 3.1. *$S^n = (S_1, \ldots, S_n)$ some process,*

$$\frac{1}{n}\frac{1}{\log P_{S^n}(S^n)} \to U \quad \textit{in dist.} \quad \Leftrightarrow \quad \frac{1}{n} l(f^*(S^n)) \to U \quad \textit{in dist.}$$

Proof. pick $\tau = \sqrt{n}$ □

Note: if further S^n i.i.d., by LLN we know that limit dist. U is degenerate $H(S)$.

Corollary 3.2. *$S^n = (S_1, \ldots, S_n)$ some process,*

$$\frac{1}{\sqrt{n}}(\frac{1}{\log P_{S^n}(S^n)} - nH(S)) \to U \quad \textit{in dist.} \quad \Leftrightarrow \quad \frac{1}{\sqrt{n}}(l(f^*(S^n)) - nH(S)) \to U \quad \textit{in dist.}$$

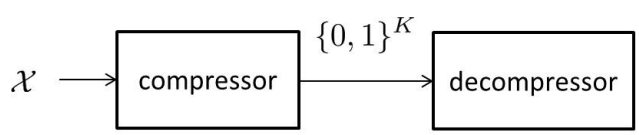
Proof. pick $\tau = n^{1/4}$, by CLT, we know that U is Normal dist. □

Note: if further that S^n i.i.d. sequence,

$$l(f^*(S^n)) \sim nH + \sqrt{nV}\mathcal{N}(0,1) \text{ in dist.}$$

where $V = Var[\log(1/P_{S^n}(S^n)]$.

4 fixed blocklength code, almost lossless



$$f : \mathcal{X} \to \{0,1\}^K \qquad g : \{0,1\}^K \to \mathcal{X}$$

Note: If we want $g \circ f = \mathbf{1}_{\mathcal{X}}$, then $K \geq log_2|\mathcal{X}|$. But, the transmission link is erroneous anyway...and it turns out that by tolerating with a little error probability ϵ, we gain a lot in terms of code length!

Definition 4.1 ((K,ϵ) compressor). *define compressor-decompressor pair (f,g) to be a (K,ϵ) compressor if:*

$$\begin{aligned} &f : \mathcal{X} \to \{0,1\}^K \\ &g : \{0,1\}^K \to \mathcal{X} \cup \{e\} \\ &P(g \circ f(X) = e) \leq \epsilon \end{aligned}$$

fundamental limit:

$$\epsilon^*(X,K) = \inf\{\epsilon : \exists (K,\epsilon)\text{compressor for } X\}$$

Theorem 4.1 (fundamental limit of error probabiliy)**.**

$$\epsilon^*(X,K) = P[l(f^*(X)) \geq K]$$

Proof. (construction) encode the $2^K - 1$ symbols with highest prob. to $\{\phi, 0, 1, 00, \ldots, 1^{K-1}\}$. □

Corollary 4.1 (Shannon's)**.** *S^n i.i.d. source*

$$\lim_{n\to\infty} \epsilon^*(S^n, nR) = \begin{cases} 0 & R > H(S) \\ 1 & R < H(S) \end{cases}$$

$$\lim_{n\to\infty} \epsilon^*(S^n, nH(S) + \sqrt{nV}\gamma) = 1 - P(\gamma)$$

where $P(\cdot)$ is the CDF of $\mathcal{N}(0,1)$

Proof. use Corrolary 3.2. □

Two achievability bound

Theorem 4.2.

$$\epsilon^*(X,K) \leq P[\log_2 \frac{1}{P_X(X)} \geq K]$$

Proof. (construction) use the $2^K - 1$ symbols with highest prob. □

Theorem 4.3.

$$\epsilon^*(X,K) \leq P[\log_2 \frac{1}{P_X(X)} > K - \tau] + 2^{-\tau}, \quad \forall \tau > 0$$

Proof. (construction) **random coding** (Shannon's magic)
compressor $f(x) = c_x$, hashing $x \in \mathcal{X}$ to a random K bit string c_x in the 2^K universe W
decompressor g is given by:

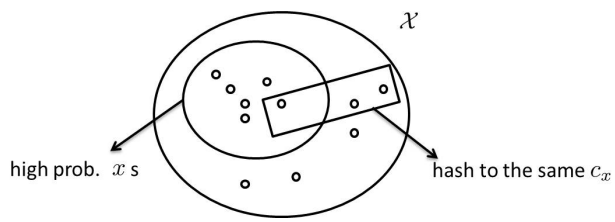
$$g(w) = \begin{cases} x, & \exists!\ x \in \mathcal{X}\ s.t. f(x) = w \text{ and } \log_2(\frac{1}{P_X(x)}) \leq K - \tau \\ e, & o.w. \end{cases}$$

Show that the error probability (average over Hashing choices and over X) achieves the bound:

$$
\begin{aligned}
P(g \circ f(X) = e) &= P\left(\{\log_2(\frac{1}{P_X(x)}) \geq K - \tau\} \cup \{\exists x' \neq x, c_{x'} = c_x \text{ and } \frac{1}{P_X(x)}) < K - \tau\}\right) \\
&\leq P[\log_2 \frac{1}{P_X(X)} \geq K - \tau] + P[\exists x' \neq x, c_{x'} = c_x \text{ and } \frac{1}{P_X(x)}) < K - \tau] \\
&\leq P[\log_2 \frac{1}{P_X(X)} > K - \tau] + 2^{K-\tau} \cdot 2^{-K}
\end{aligned}
$$

$\square$

Note: idea:



Distribution of $\ell(f^*(X))$

4.1 Compressing iid ternary source

Consider the source outputing n ternary letters each independent and distributed as

$$P_X = [.445 \quad .445 \quad .11]\,.$$

For iid source it can be shown

$$\mathbb{E}[\ell(f^*(X))] = nH(X) - \frac{1}{2}\log 2\pi e V n + O(1)\,,$$

where we denoted the *varentropy* of X by

$$V \triangleq \mathrm{Var}\left[\log \frac{1}{P_X(X)}\right]\,.$$

The Gaussian approximation to $\ell(f^*(X))$ is defined as

$$nH(X) - \frac{1}{2}\log 2\pi e V n + \sqrt{nV}Z\,,$$

where $Z \sim \mathcal{N}(0,1)$.

On Fig. 1, 2, 3 we plot the distribution of the length of the optimal compressor for different values of n and compare with the Gaussian approximation.

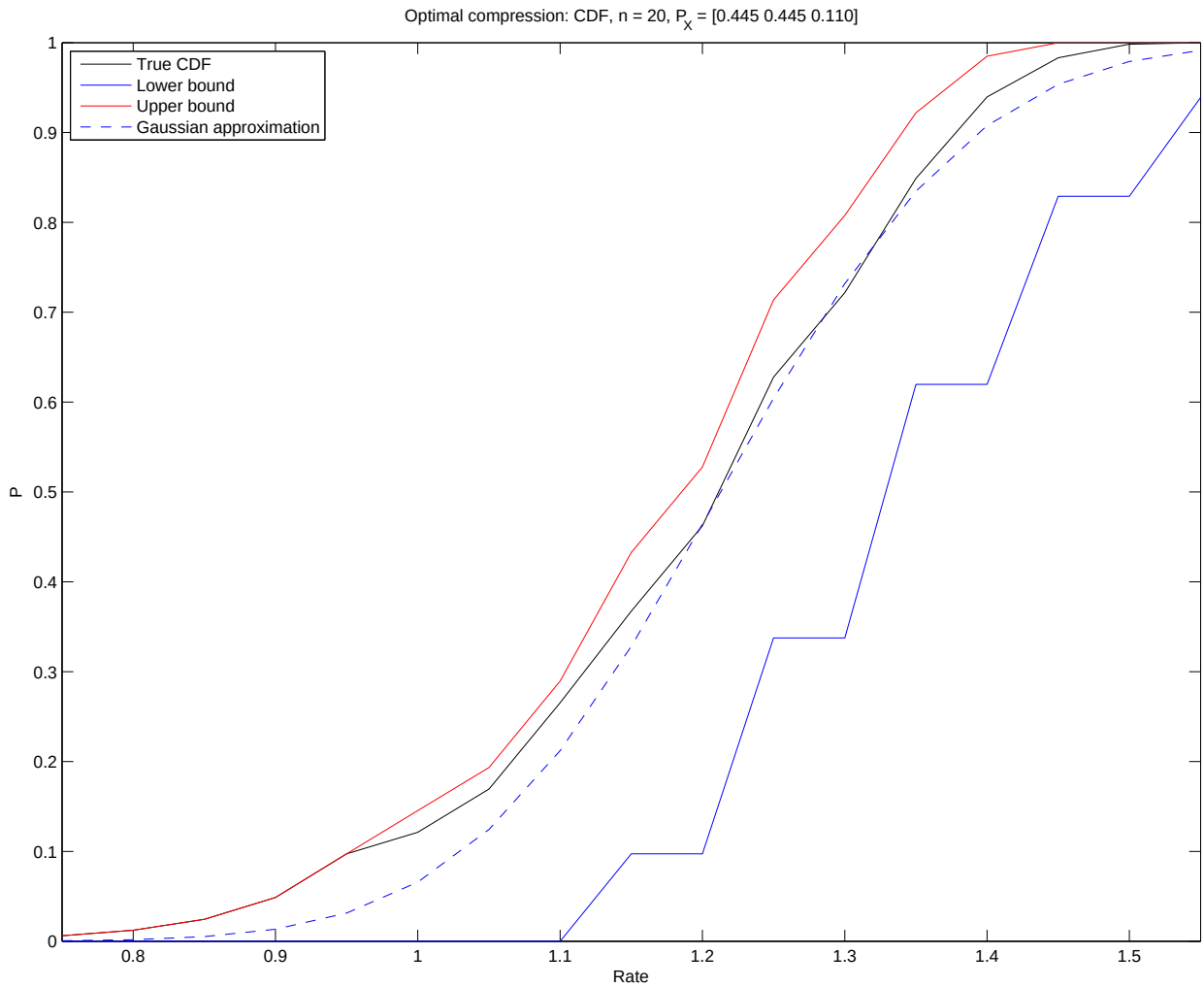
Upper/lower bounds on the expectation:

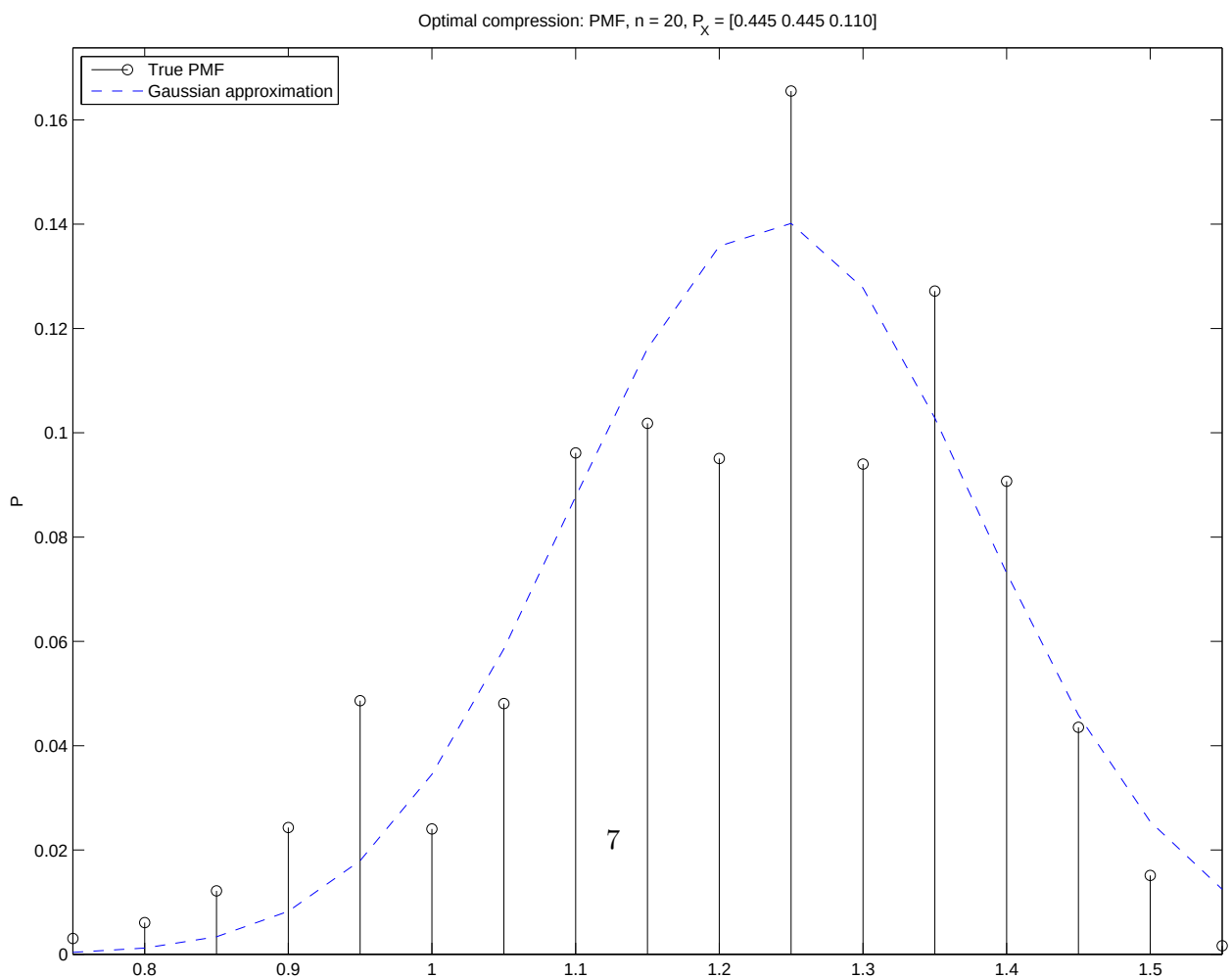
$$H(X^n) - \log(H(X^n)+1) - \log e \le \mathbb{E}[\ell(f^*(X))] \le H(X^n)$$

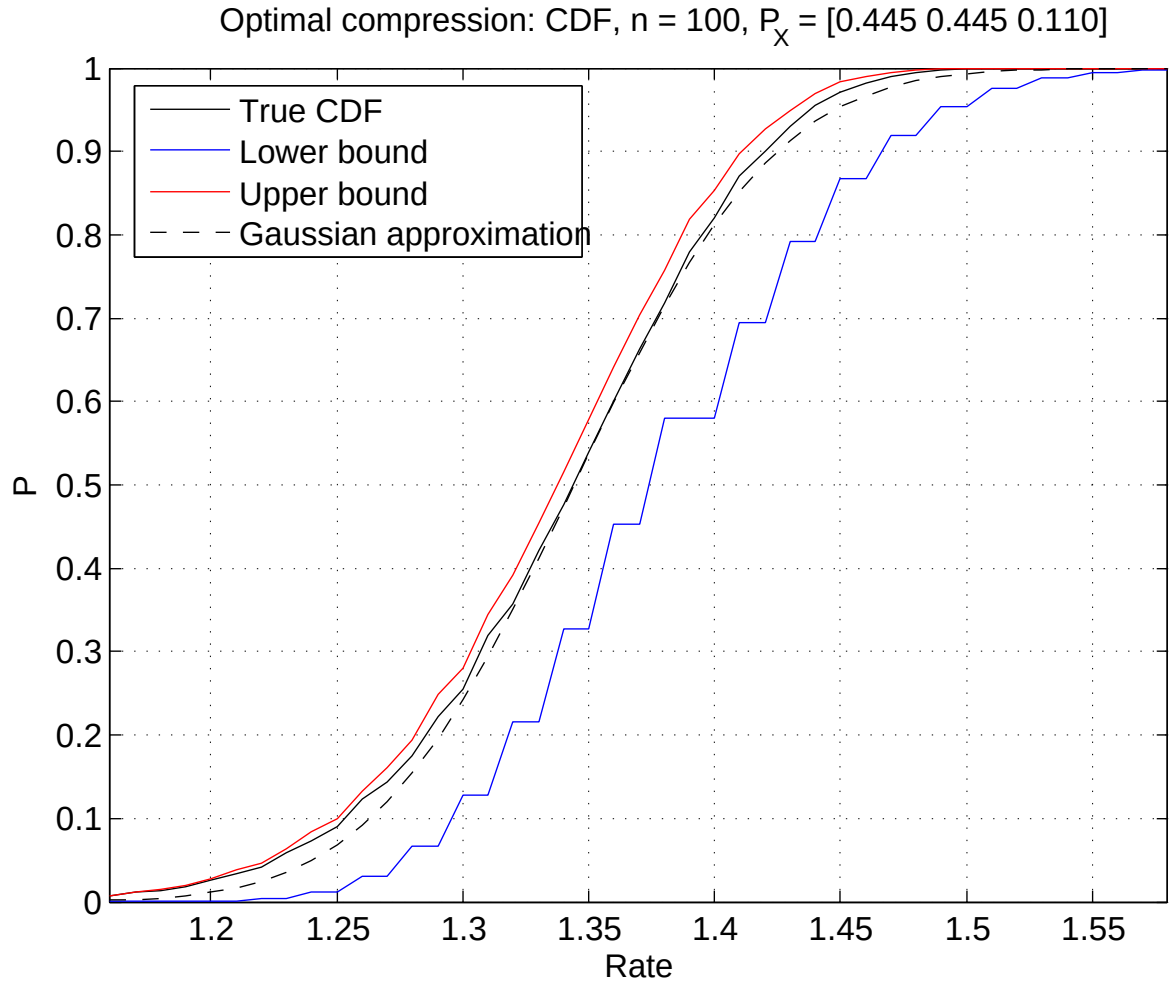
Here are the numbers for different n

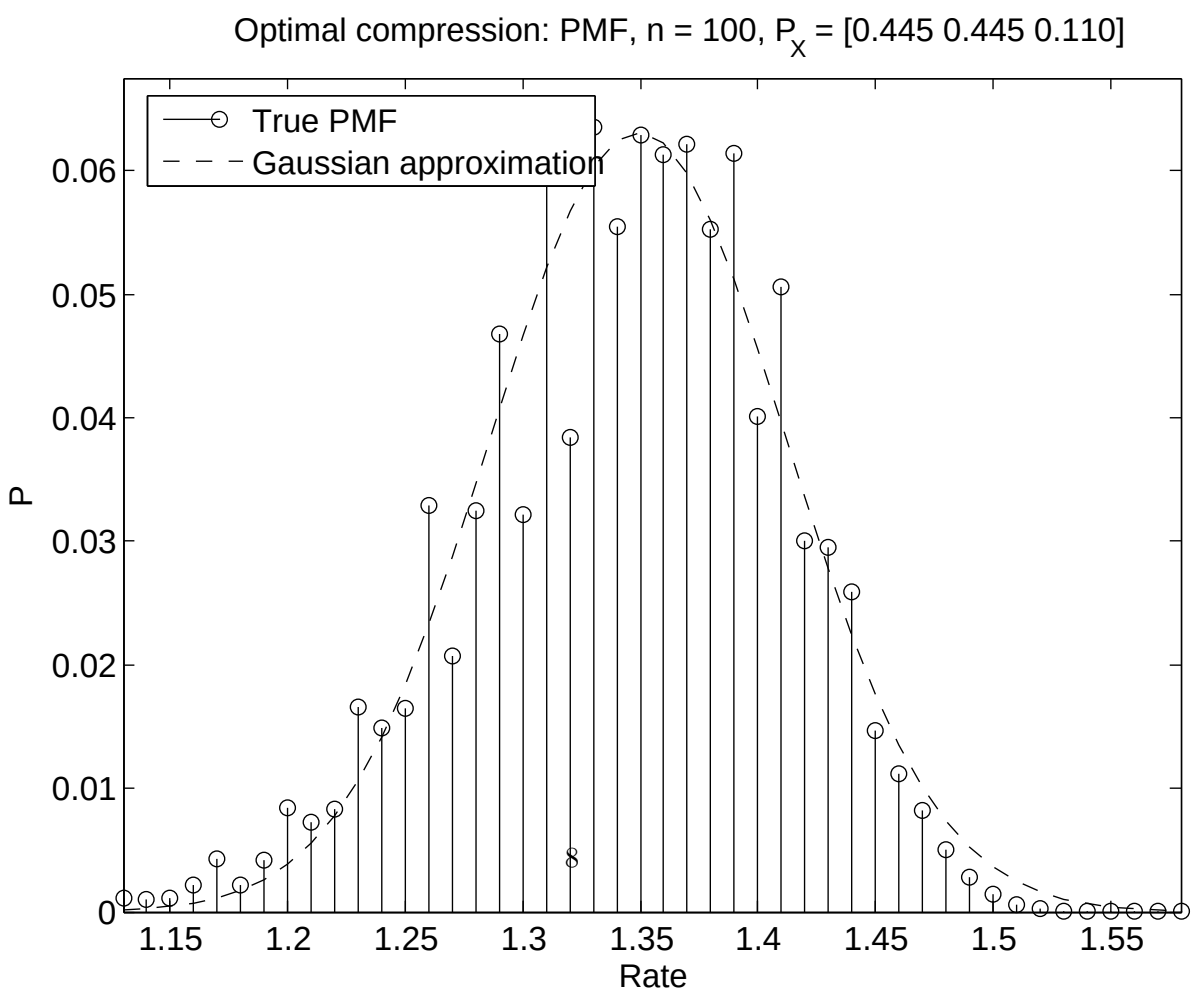
$n = 20$	21.5	$\le$	24.3	$\le$	27.8
$n = 100$	130.4	$\le$	134.4	$\le$	139.0
$n = 500$	684.1	$\le$	689.2	$\le$	695.0

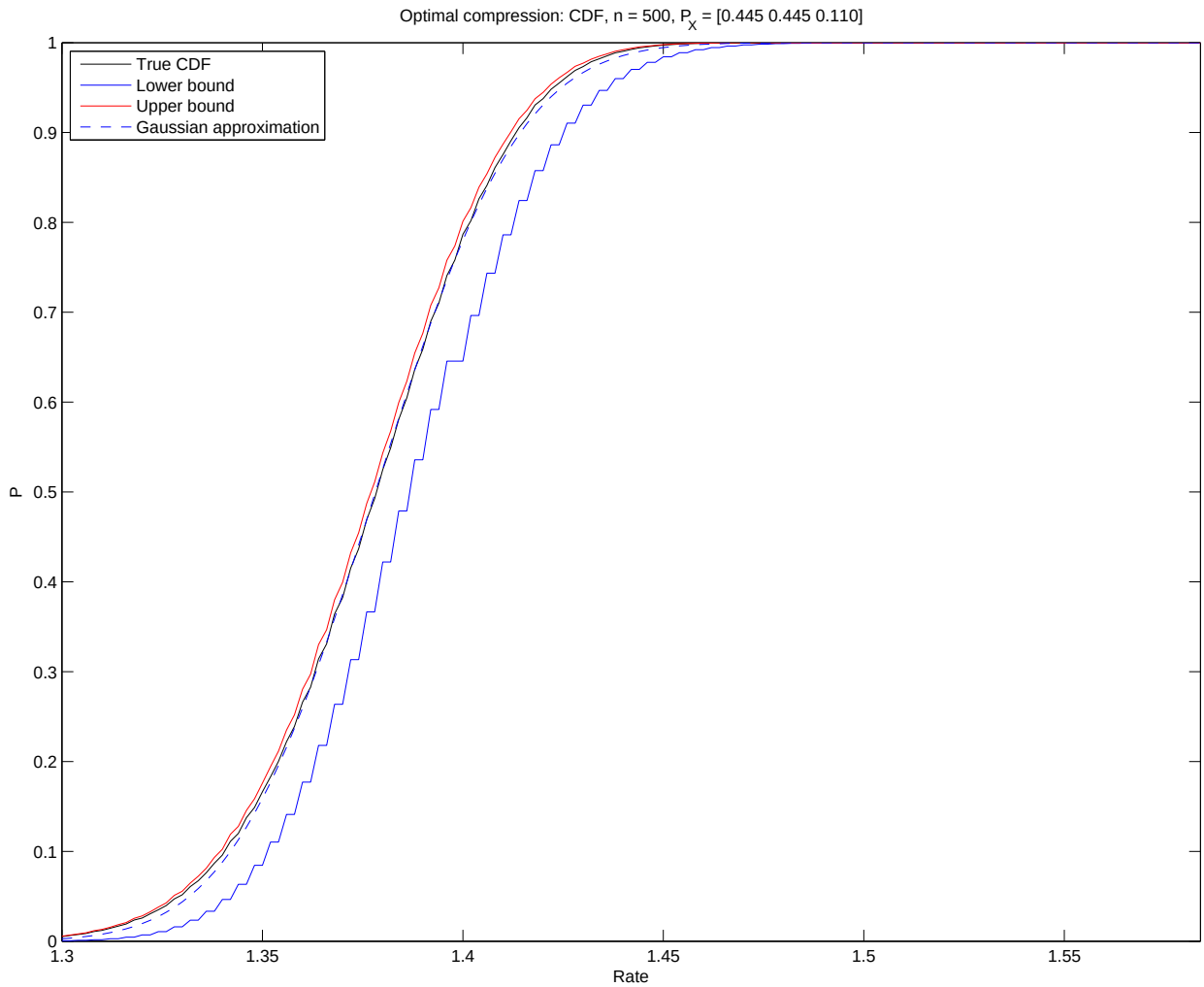
In all cases above $\mathbb{E}\,\tau$ is close to a midpoint between the two.

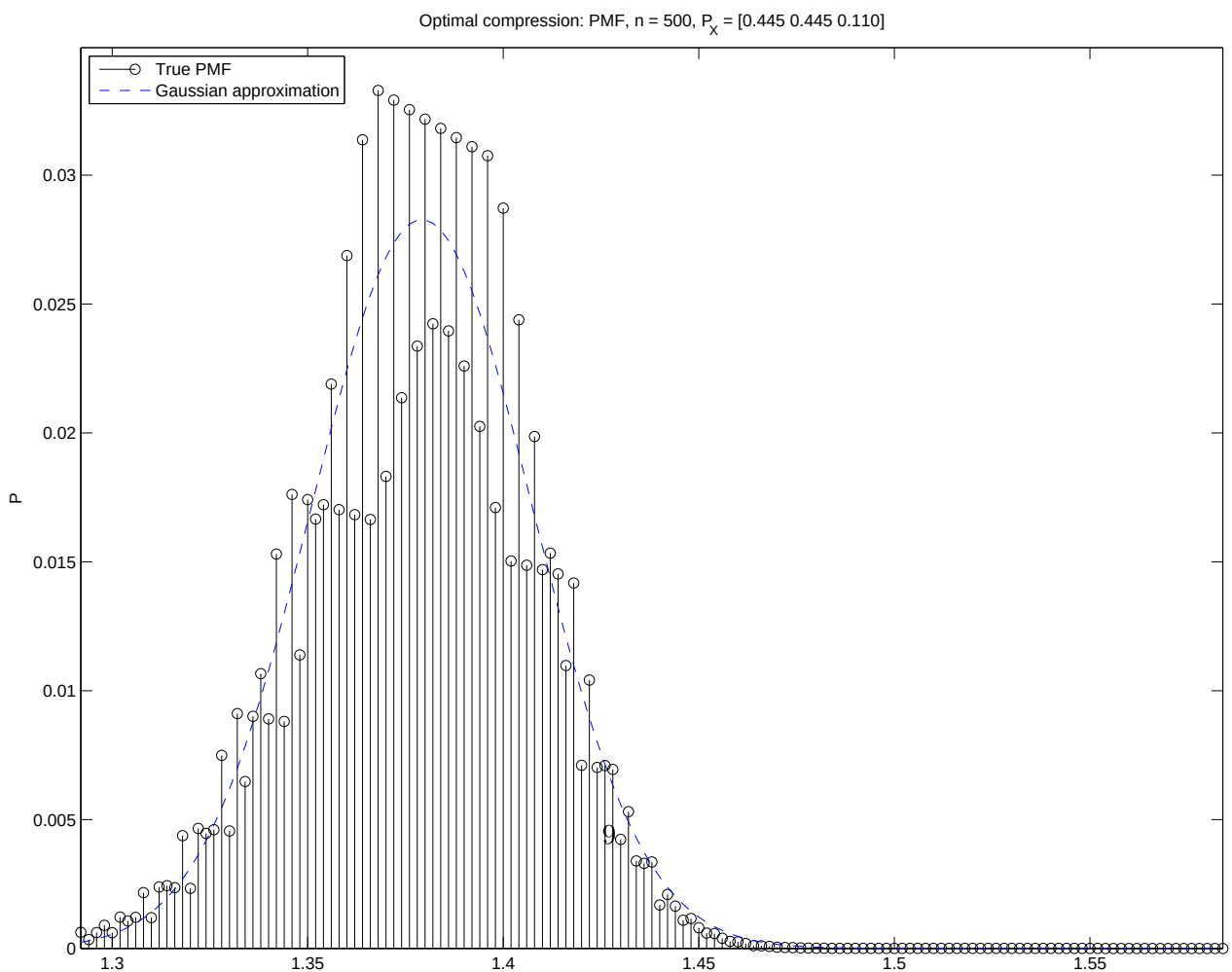

Optimal compression: CDF, n = 20, P_X = [0.445 0.445 0.110]
True CDF
Lower bound
Upper bound
Gaussian approximation
P
Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0.8
0.9
1
1.1
1.2
1.3
1.4
1.5


Optimal compression: PMF, n = 20, P_X = [0.445 0.445 0.110]
True PMF
Gaussian approximation
P
0
0.02
0.04
0.06
0.08
0.1
0.12
0.14
0.16
0.8
0.9
1
1.1
1.2
1.3
1.4
1.5


Optimal compression: CDF, n = 100, P_X = [0.445 0.445 0.110]
True CDF
Lower bound
Upper bound
Gaussian approximation
P
Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
1.2
1.25
1.3
1.35
1.4
1.45
1.5
1.55


Optimal compression: PMF, n = 100, P_X = [0.445 0.445 0.110]
True PMF
Gaussian approximation
P
Rate
0
0.01
0.02
0.03
0.04
0.05
0.06
1.15
1.2
1.25
1.3
1.35
1.4
1.45
1.5
1.55


Optimal compression: CDF, n = 500, P_X = [0.445 0.445 0.110]
True CDF
Lower bound
Upper bound
Gaussian approximation
P
Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
1.3
1.35
1.4
1.45
1.5
1.55


Optimal compression: PMF, n = 500, P_X = [0.445 0.445 0.110]
True PMF
Gaussian approximation
P
0
0.005
0.01
0.015
0.02
0.025
0.03
1.3
1.35
1.4
1.45
1.5
1.55

6.441: Information Theory	Spring 2013
Lecturer: Yury Polyanskiy	**Lecture 7: Lossless Compression**

Lossless Compression

- Linear Compression
- Compression with Side Information
- Slepian Wolf

3.1 Linear Compression

$$\epsilon^*(X, nR) \longrightarrow 0 \text{ or } 1 \qquad\qquad R \lessgtr H(S)$$

Assume $S_i \in \mathbb{F}_{p^n} = \mathbb{F}_q$ (Galois Field) where p is prime, $n \in \mathbb{N}$, and q is number of elements in $\mathbb{F}_q$.

Definition 3.1 (Galois Field)**.** *F is a set with operations $(+, \cdot)$ where $\forall a, b \in F$*

- *$a + b$ associative and commutative*
- *$a \cdot b$ associative and commutative*
- $0, 1 \in F$
- $a + (-a) = 0,\ a^{-1}a = 1$
- $a \cdot (b + c) = (a \cdot b) + (a \cdot c)$

Example:

- $\mathbb{F}_p = \mathbb{Z}/p\mathbb{Z}$
- $\mathbb{F}_2^k = \{0, 1\}^k$ with usual mod 2 sum and product

Linear Compression Problem: $x \in \mathbb{F}_q^n$, $w = Hx$ where $H : \mathbb{F}_q^n \to \mathbb{F}_q^k$ is linear.

$$\begin{bmatrix} w_1 \\ \vdots \\ w_k \end{bmatrix} = \begin{bmatrix} h_{11} & \dots & h_{1n} \\ \vdots & & \vdots \\ h_{k1} & \dots & h_{kn} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$$

Typed by Austin Collins, austinc@mit.edu, Feb 28, 2013

Theorem 3.1. *$\forall \tau > 0, \exists$ linear compressor $H : \mathbb{F}_q^n \to \mathbb{F}_q^k$ s.t.*

$$\epsilon \leq \mathbb{P}[\log_q \frac{1}{P_X(x)} > k - \tau] + q^{-\tau}$$

Proof.

$$g(w) = \begin{cases} x_0 & w = Hx, \exists! x_0 : w = Hx_0,\ x_0 - h.p. \\ e & o/w \end{cases}$$

Where $x_0 - h.p.$ (high probability) $\Leftrightarrow \log \frac{1}{P_X(x_0)} < k - \tau$

Note that this decoder is the same as in the proof of Shannon's random coding theorem in Lec. 5 This proof is mostly the same, except now hash collisions occur under the linear map H. Collision have probability

$$\begin{aligned} \mathbb{P}_H[Hx_1 = Hx_0] &= \mathbb{P}_H[Hx_2 = 0] \\ &= \mathbb{P}_H[H_1 \cdot x_2 = 0]^k \end{aligned}$$

Where H_1 is the first row of the matrix H, and each row of H is independent. This is the probability that H_i is in the orthogonal complement of x_2. On $\mathbb{F}_q^n$, the orthogonal complement of a vector has size q^{n-1}, so its probability is $q^{n-1}/q^n = 1/q$, hence the collision probability is $1/q^k$. Therefore,

$$\begin{aligned} \mathbb{P}[g(f(X) = e] &\leq \mathbb{P}[\log_q \frac{1}{P_X(x)} > k - \tau] + \mathbb{P}[\{x_1 \neq x, Hx_1 = Hx, x_1 - h.p.\}] \\ \text{(union bound)} &\leq \mathbb{P}[\log_q \frac{1}{P_X(x)} > k - \tau] + \sum_{x' - h.p., x' \neq x} \mathbf{1}\{Hx' = Hx\} \end{aligned}$$

And averaging over all linear compressors H gives

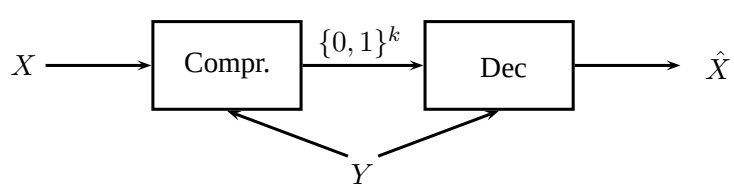
$$\sum_{x' - h.p., x' \neq x} \mathbb{P}[Hx' = H_x] = q^{k-\tau} q^{-k} = q^{-\tau}$$

Thus the bound holds. □

Remark: Can we loosen the requirement on $\mathbb{F}_q$ to instead be a commutative ring? In general, no, since zero divisors in the commutative ring ruin the key proof item of low collision probability in the random hashing. E.g. in $\mathbb{Z}/6\mathbb{Z}$

$$\mathbb{P}\left[H \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = 0\right] = 6^{-k} \qquad \text{but} \qquad \mathbb{P}\left[H \begin{bmatrix} 2 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = 0\right] = 3^{-k}$$

3.2 Compression with Side Information



Definition 3.2 (Compression wih Side Information Code)**.**

- $f : \mathcal{X} \times \mathcal{Y} \to \{0,1\}^k$
- $g : \{0,1\}^k \times \mathcal{Y} \to \mathcal{X} \cup \{e\}$
- $\forall y,\ g(f(x,y),y) = x \ \ or\ e$
- $\mathbb{P}[g(f(X,Y),Y) = e] < \epsilon$
- *Fundamental Limit:* $\epsilon^*(X|Y,k) = \inf\{\epsilon :\ \exists (k,\epsilon) - S.I.\ code\}$

Theorem 3.2.

$$\mathbb{P}[\log \frac{1}{P_{X|Y}(X|Y)} > k + \tau] - 2^{-\tau} \leq \epsilon^*(X|Y,k) \leq \mathbb{P}[\log \frac{1}{P_{X|Y}(X|Y)} > k - \tau] + 2^{-\tau}$$

Corollary 3.1. $(X,Y) = (S^n, T^n)$ *where* $(S_1,T_1),(S_2,T_2),\ldots$ *are iid pairs* $\sim P_{ST}$

$$\lim_{n\to\infty} \epsilon^*(S^n|T^n, nR) = \begin{cases} 0 & R > H(S|T) \\ 1 & R < H(S|T) \end{cases}$$

Proof. Fix $Y = y_0 \implies P_{X|Y=y0}$, then using the bound for compression without side information

$$\mathbb{P}[\log \frac{1}{P_{X|Y}(X|y_0)} > k + \tau | Y = y_0] - 2^{-\tau} \leq \epsilon^*(X|Y = y_0, k) \leq \mathbb{P}[\log \frac{1}{P_{X|Y}(X|y_0)} > k - \tau | Y = y_0] + 2^{-\tau}$$

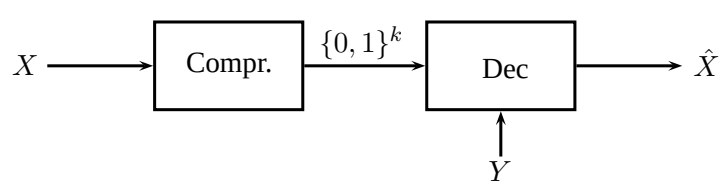
By taking the average over all $y_o \in Y$, we get the theorem. For the corollary

$$\frac{1}{n} \log \frac{1}{P_{S^n|T^n}(S^n|T^n)} = \frac{1}{n} \sum_{i=1}^{n} \log \frac{1}{P_{S|T}(S_i|T_i)} \longrightarrow H(S|T) \text{ (in probability)}$$

As $n \to \infty$, using the WLLN. □

3.3 Slepian-Wolf

Consider the compression with side information problem, except now the compressor has no access to the side information.



Definition 3.3 (S.W. code)**.**

- $f : \mathcal{X} \to \{0,1\}^k$
- $g : \{0,1\}^k \times \mathcal{Y} \to \mathcal{X} \cup \{e\}$

- *For* $\hat{X} = g(f(X), Y)$*, require* $\mathbb{P}[\hat{X} \neq X] \leq \epsilon$
- *Fundamental Limit:* $\epsilon^*_{SW} = \inf\{\epsilon : \exists (k, \epsilon) - S.W.\ code\}$

Note: This definition does not include the usual condition that $g(f(x), y) = x$ or e, hence we allow for the possibility of undetected errors. Indeed, if we require this condition, the theorem no longer holds.

Now the very surprising result: even without side information at the compressor, we can still compress down to the conditional entropy!

Theorem 3.3.

$$\epsilon^*(X|Y, k) \leq \epsilon^*_{SW}(X|Y, k) \leq \mathbb{P}[\log \frac{1}{P_{X|Y}} \geq k - \tau] + 2^{-\tau}$$

Corollary 3.2.

$$\lim_{n\to\infty} \epsilon^*_{SW}(S^n|T^n, nR) = \begin{cases} 0 & R > H(S|T) \\ 1 & R < H(S|T) \end{cases}$$

Proof. LHS is obvious, since side information at the compressor and decoder is better than only at the decoder.

For the RHS, first generate a random codebook uniformly $\forall x \in \mathcal{X}, C_x \in \{0, 1\}^k$, then define the compressor and decoder as

$$f(x_0) = C_{x_0}$$
$$g(w, y) = \begin{cases} x_0 & \exists! x_0 : C_{x_0} = w, x_0 - h.p. \\ 0 & o.w. \end{cases}$$

Where $x_0 - h.p. \Leftrightarrow \log \frac{1}{P_{X|Y}(x_0|Y)} < k - \tau$. The error probability of this scheme is

$$\mathcal{E}(C) = \sum_{\mathcal{X}\mathcal{Y}} P_{XY}(x, y) 1\{\log \frac{1}{P_{X|Y}(x|y)} \geq k - \tau \ \cup \ J(X, C) \neq \emptyset\}$$

Apply the union bound, and upper bound by the cardinality

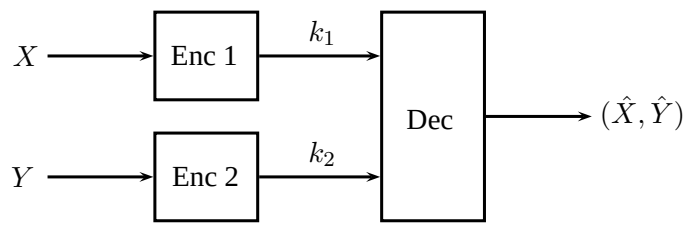
$$\mathbb{P}_X[J(X, C|Y = y) \neq \emptyset] \leq |J(x, C|Y = y)|$$

Now averaging over all codebooks, using $\#\{x' - h.p.|y\} \leq 2^{k-\tau}$ and $\mathbb{P}[C_{x'} = C_x] \leq 2^{-k}$

$$\begin{aligned} \mathbb{E}_C[|J(x, C|Y = y)|] &= \mathbb{E}\left[\sum_{x'} 1\{x' - h.p.|y\} 1\{C_{x'} = C_x\}\right] \\ &= 2^{k-\tau} \mathbb{P}[C_{x'} = C_x] \\ &= 2^{-\tau} \end{aligned}$$

Hence the theorem follows as usual from two terms in the union bound. □

3.4 Multiterminal Slepian Wolf



Definition 3.4.

- (f_1, f_2, g) *is* (k_1, k_2, ϵ) *code ...* $\mathbb{P}[\hat{X} \neq X \text{ or } \hat{Y} \neq Y]\epsilon$
- *Fundamental limit:* $\epsilon^*_{SW}(X, Y, k_1, k_2)$

Theorem 3.4. $(X, Y) = (S^n, T^n)$ *- iid pairs*

$$\lim_{n\to\infty} \epsilon^*_{SW}(X, Y, nR_1, nR_2) = \begin{cases} 0 & (R_1, R_2) \in \mathcal{R}_{SW} \\ 1 & (R_1, R_2) \notin \mathcal{R}_{SW} \end{cases}$$

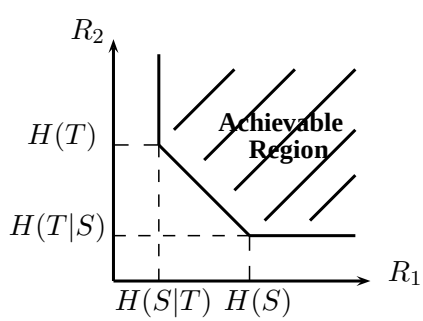
Where

$$\mathcal{R}_{SW} = \left\{ (a, b) : \begin{array}{l} a \geq H(S|T) \\ b \geq H(T|S) \\ a + b \geq H(S, T) \end{array} \right.$$

Proof. Take $(R_1, R_2) \notin \mathcal{R}_{SW}$. Then one of three cases must occur:

1. $R_1 < H(S|T)$. Then even if encoder and decoder had full T^n, still can't achieve this (from compression with side channel info result).
2. $R_2 < H(T|S)$ (same).
3. $R_1 + R_2 < H(S, T)$. Can't compress below the joint entropy of the pair (S, T).

□



Note: Since $H(T) - H(T|S) = H(S) - H(S|T) = I(S; T)$, the slope is -1.

3.4.1 Time Sharing

If you can achieve with vanishing error probability any two points (R_1, R_2) and (R_1', R_2'), then you can achieve for $\lambda \in [0,1]$, $(\lambda R_1 + \bar{\lambda} R_1', \lambda R_2 + \bar{\lambda} R_2')$ via

$$(S_1^{\lambda n}, T_1^{\lambda n}) \to \begin{bmatrix} \lambda n R_1 \\ \lambda n R_2 \end{bmatrix} \quad \text{using } (R_1, R_2) \text{ code}$$

$$(S_{\lambda n+1}^{n}, T_{\lambda n+1}^{n}) \to \begin{bmatrix} \bar{\lambda} n R_1' \\ \bar{\lambda} n R_2' \end{bmatrix} \quad \text{using } (R_1', R_2') \text{ code}$$

Therefore, all convex combinations of points in the achievable regions are also achievable, so the achievable region must be convex.

6.441: Information Theory	Spring 2012
Lecturer: Yury Polyanskiy	**Lecture 8: Compression, Ergodic process**

We have examined the compression of i.i.d. sequence, for which

$$\frac{1}{n} l(f^*(X^n)) \to H(X) \quad \text{in prob.}$$
$$\lim_{n\to\infty} \epsilon^*(X^n, nR) = \begin{cases} 0 & R > H(X) \\ 1 & R < H(X) \end{cases}$$

In this lecture, we shall examine similar results for Ergodic processes and we first state the main theory as follows:

Theorem 8.1 (Shannon-McMillan)**.** *Let $\{S_0, S_1, \dots\}$ be a stationary and Ergodic dicrete process, then*

$$\frac{1}{n} \log \frac{1}{P_{S^n}(S^n)} \to \mathcal{H} \quad \textit{in prob.} \tag{1}$$

where $\mathcal{H} = \lim_{n\to\infty} \frac{1}{n} H(S^n)$ is the entropy rate.

Note: In fact convergence holds in the almost sure sense too (Breiman).

8.1 Review of Ergodic theory

Let's start with a dynamic system view and introduce a few definitions:

Definition 8.1 (stationary process)**.** *A collection of maps $\{S^n\}$ ($S_i : \Omega \to \mathbf{R}$) is stationary if there exists a measure preserving transformation $\tau : \Omega \to \Omega$ such that:*

$$S_j = S_{j-1} \circ \tau = S_0 \circ \tau^j$$

Note: Therefore a stationary process can be described by the tuple $(\Omega, \mathcal{F}, \mathbb{P}, \tau, X_0)$ and $X_k = X_0 \circ \tau^k$. We can think of τ as the shift operator:

$$(\theta_0, \theta_1, \dots) \xrightarrow{\tau} (\theta_1, \theta_2, \dots)$$

Note: Statistics view of stationary process: a process $(X_0, X_1, \dots)$ on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is stationary if $X_n^m \sim X_{n+t}^{m+t}$ in dist. $\forall n, m, t$.

Note: Show the two definitions are equivalent. (DIY)

Typed by Qingqing Huang (qqh@mit.edu), Mar. 5, 2013

Definition 8.2 (Measure preserving transformation)**.** *$\tau : \Omega \to \Omega$ is measure preserving if*

$$\forall E \in \mathcal{F}, P(E) = P(\tau^{-1}E).$$

Definition 8.3 (Shift invariant event)**.** *An event $E \in \mathcal{F}$ is shift invariant if*

$$(\theta_0, \theta_1, \dots) \in E \to (\theta_1, \theta_2, \dots) \in E$$

or equivalently $E = \tau^{-1}E$.

Note: Given a measure preserving trsf. τ, $B_\tau = \{E : E = \tau^{-1}E\}$ forms a σ-algebra on Ω. (DIY)

Definition 8.4 (Ergodic process)**.** *A process is Ergodic if for any shift invariant event E, $P(E) = 0$ or 1 (B_τ consists of determinstic events).*

Example:

- $\{S_k = k^2\}$, ergo. but not stat.
- $\{S_k\}$ i.i.d. is ergo. (by komogorov's 0-1 law, tail event)
- (Sliding the window construction of Ergodic process)
  If $\{S_i\}$ is ergo., then $\{X_i = f(S_i, S_{i+1}, \dots)\}$ is also ergo. It is called **B-process** if S_i i.i.d.
  Example, $S_i \sim Ber(\frac{1}{2})$ i.i.d. $X_k = \sum_{n=0}^{\infty} 2^{-n-1} S_{k+n} = 2X_{k-1}$ *mod* 1.
- $\{S_i\}$ follows finite Markov chain, with recurrent states.
  Toy example: kernel $P(0|1) = P(1|0) = 1$ with initial dist. $P(S_0 = 0) = 0.5$.
- let $\{S_i\}$ be stationary 0-mean Gaussian process,

$$R(n) = \mathbb{E}[S_0 S_n'] \to 0 \Leftrightarrow \{S_i\} \text{ ergo.} \tag{2}$$

$$F(\omega) = \text{F.T. of } R(n) \text{ is abs. cont's} \Leftrightarrow \{S_i\}\text{B-process} \tag{3}$$

8.2 Proof of Shannon-McMillan

We shall only show the convergence in probability.

In order to prove Shannon-McMillan, let's first introduce the Khintchine's convergence theorem for Ergodic processes, the proof of which is presented in the next subsection.

Theorem 8.2 (Birkoff-Khintchine's convergence theorem)**.** *if $\{S_i\}$ stationary and Ergodic, $\forall$ function $f \in \mathcal{L}_1$,*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} f(S_k, \dots) \to \mathbb{E}[f(S_0, \dots)] \quad \textit{a.s. and in } \mathcal{L}_1$$

Note: layman: time average converges to ensemble average.

Note: Showing (1) for an m-th order time homogeneous Markov chain $\{S_i\}$ is a direct application of Birkoff-Khintchine.

$$\frac{1}{n}P_{S^n}(S^n) = \underbrace{\frac{1}{n}\log\frac{1}{P_{S_1}(S_1^m)}}_{\to 0} + \underbrace{\frac{1}{n}\sum_{l=m+1}^{n}\log\frac{1}{P_{S_l|S_{l-m}^{l-1}}(S_l|S_{l-m}^{l-1})}}_{\to H(S_{l+1}|S_l^m) \text{ by Khintchine}} \tag{4}$$

Note: Now let's prove (1) for a general Stationary Ergodic process $\{S_i\}$. The idea is to approximate the distribution of that Ergodic process by an m-th order MC and make use of (4); then let $m \to \infty$ to make the the approximation accurate.

Proof. Define the distribution

$$Q^{(m)}(S_1^\infty) = P_{S_1^m}(S_1^m)\prod_{l=m+1}^{\infty} P_{S_l|S_{l-m}^{l-1}}(S_l|S_{l-m}^{l-1})$$

To show (1) holds in probability, we want to show that

$$\begin{aligned}
P[\Big|\frac{1}{n}\log\frac{1}{P_{S^n}(S^n)} - \mathcal{H}\Big| > 3\epsilon] \leq & P[\underbrace{\Big|\frac{1}{n}\log\frac{1}{P_{S^n}(S^n)} - \frac{1}{n}\log\frac{1}{Q_{S^n}^{(m)}(S^n)}\Big|}_{A} > \epsilon] \\
& + P[\underbrace{\Big|\frac{1}{n}\log\frac{1}{Q_{S^n}^{(m)}(S^n)} - H_m\Big|}_{B} > \epsilon] \\
& + P[\underbrace{\Big|H_m - \mathcal{H}\Big|}_{C} > \epsilon] \to 0 \text{ as } n \to \infty, m \to \infty
\end{aligned}$$

where $H_m = H(S_{m+1}|S_1^m)$.

$C = |H_m - \mathcal{H}| \to 0$ as $m \to \infty$ by definition of entropy rate.
$B \to 0$ a.s. and in $\mathcal{L}_1$ as $n \to \infty$ by Khintchine for any fixed m.

$$\begin{aligned}
E[A] = \frac{1}{n}E_P\Big|\log\frac{dP_{S^n}}{dQ_{S^n}^{(m)}}\Big| = \frac{1}{n}E_Q\Big|\frac{dP_{S^n}}{dQ_{S^n}^{(m)}}\log\frac{dP_{S^n}}{dQ_{S^n}^{(m)}}\Big| \leq \frac{1}{n}E_Q[\frac{dP_{S^n}}{dQ_{S^n}^{(m)}}\log\frac{dP_{S^n}}{dQ_{S^n}^{(m)}}] + 2\frac{\log e}{en} \\
= \frac{1}{n}D(P_{S^n}\|Q_{S^n}^{(m)}) + 2\frac{\log e}{en} \to 0
\end{aligned}$$

where we used that

$$\begin{aligned}
\frac{1}{n}D(P_{S^n}\|Q_{S^n}^{(m)}) &= \frac{1}{n}\mathbb{E}[\log\frac{P_{S^n}}{P_{S^m}\prod_{l=m+1}^{n}P_{S_k|S_{k-m}^{k-1}}}] \\
&= \frac{1}{n}(-H(S^n) + H(S^m) + (n-m)H_m) \\
&\to -\mathcal{H} + H_m \text{ as } n \to \infty \\
&\to 0 \text{ as } m \to \infty
\end{aligned}$$

and the next Lemma. □

Lemma 8.1.

$$\mathbb{E}_P\left[\left|\log\frac{dP}{dQ}\right|\right] \leq D(P||Q) + \frac{2\log e}{e}\,.$$

Proof. $|xlogx| - xlogx \leq \frac{2\log e}{e}$ □

8.3 Proof of Birkhoff-Khintchine

Proof. $\forall$ function $\tilde{f} \in \mathcal{L}_1$, $\forall\epsilon$, there exists a decomposition $\tilde{f} = f + h$ such that f is bounded, and $h \in \mathcal{L}_2$, $\|h\|_1 \leq \epsilon$.
Let's first focus on the bounded function f. Note that in the bounded domain $\mathcal{L}_1 \subset \mathcal{L}_2$, thus $f \in \mathcal{L}_2$ Hilbert space with inner product $< f, g >= \mathbb{E}[f(S_1^\infty)\overline{g(S_1^\infty)}]$.

For the measure preserving transformation τ that generates the stationary process $\{S_i\}$, define the operator $T(f) = f \circ \tau$. Since τ is measure preserving, we know that $\|Tf\|_2^2 = \|f\|_2^2$, thus T is a unitary and bounded operator.

Define the operator

$$A_n(f) = \frac{1}{n}\sum_{k=1}^{n} f \circ \tau^k$$

Intuitively:

$$A_n = \frac{1}{n}\sum_{k=1}^{n} T^k = \frac{1}{n}(I - T^n)(I - T)^{-1}$$

Then, if $f \perp \ker(I - T)$ we should have $A_n f \to 0$, since only components in the kernel can blow up. This intuition is formalized in the proof below.

Let's further decompose f into two parts $f = f_1 + f_2$, where $f_1 \in Ker(I-T)$ and $f_2 \in Ker(I-T)^\perp$.

Observations:

- if $g \in Ker(I - T)$, g must be a constant function. This is due to the Ergodicity. Consider indicator function $\mathbf{1}_A$, if $\mathbf{1}_A = \mathbf{1}_A \circ \tau = \mathbf{1}_{\tau^{-1}A}$, then $\mathbb{P}[A] = 0$ or 1. For a general case, suppose $g = Tg$ and g is not constant, then at least some set $\{g \in (a, b)\}$ will be shift-invariant and have non-trivial measure, violating ergodicity.
- $Ker(I - T) = Ker(I - T^*)$. This is due to the fact that T is unitary:

  $g = Tg \Rightarrow < g - Tg, g >= 0 \Rightarrow \|g\|_2^2 =< Tg, g >\Rightarrow Tg = cg($ for some complex const $c, |c| = 1)$
- $Ker(I - T)^\perp = Ker(I - T^*)^\perp = [Im(I - T)]$, where $[Im(I - T)]$ is an $\mathcal{L}_2$ closure.
- $g \in Ker(I - T)^\perp \quad \Longleftrightarrow \quad \mathbb{E}[g] = 0$. Indeed, only zero-mean functions are orthogonal to constants.

With these observations, we know that $f_1 = m$ is a const. Also, $f_2 \in [Im(I - T)]$ so we further approximate it by $f_2 = f_0 + h_1$, where $f_0 \in Im(I - T)$, namely $f_0 = g - g \circ \tau$ for some function

$g \in \mathcal{L}_2$, and $\|h_1\|_1 \leq \|h_1\|_2 < \epsilon$. Therefore we have

$$\begin{aligned}
&A_n f_1 = f_1 = \mathbb{E}[f] \\
&A_n f_0 = \frac{1}{n}(g - g \circ \tau^n) \to 0 \text{ a.s. and } \mathcal{L}_1 \\
&(\text{since } \mathbb{E}[\sum_{k=1}^{n}(\frac{g \circ \tau^k}{n})^2] = \mathbb{E}[g^2]\frac{n}{n^2} < \infty,\ \frac{1}{n} g \circ \tau^n \to 0 \text{ a.s.})
\end{aligned}$$

The proof completes by showing

$$\mathbb{P}[\limsup_n A_n(h + h_1) \geq \delta] \leq \frac{2\epsilon}{\delta}\,. \tag{5}$$

Indeed, then by taking $\epsilon \to 0$ we will have shown

$$\mathbb{P}[\limsup_n A_n(f) \geq \mathbb{E}[f] + \delta] = 0$$

as required.

□

Proof of (5) makes use of the Maximal Ergodic Theorem stated as follows:

Theorem 8.3 (Maximal Ergodic Lemma)**.** *Let $(\mathbb{P}, \tau)$ be a probability measure and a measure-preserving transformation. Then for any $f \in \mathcal{L}_1(\mathbb{P})$ we have*

$$\mathbb{P}\left[\sup_{n\geq 1} A_n f > a\right] \leq \frac{\mathbb{E}[f \mathbf{1}_{\sup_{n\geq 1} A_n f > a}]}{a} \leq \frac{\|f\|_1}{a}$$

where $A_n f = \frac{1}{n}\sum_{k=0}^{n-1} f \circ \tau^k$.

Note: weak $\mathcal{L}_1$ estimate.

Proof. Denote the partial sum $S_n = \sum_{k=0}^{n-1} f \circ \tau^k$, we have $S_{K+1} = f + S_K \circ \tau$.
Denote $M_K = \max(0, S_1, \ldots, S_K)$, we have

$$\begin{aligned}
&f + M_K\tau \geq f + S_i \circ \tau \geq S_{i+1} \\
\Rightarrow &f \geq S_{i+1} - M_K\tau,\ \text{for } i = 1, \ldots, K \\
\Rightarrow &f \geq M_K - M_K\tau \\
\Rightarrow &\mathbb{E}[f\mathbf{1}_{\{M_K>0\}}] \geq \mathbb{E}[(M_K - M_K\tau)\mathbf{1}_{\{M_K>0\}}] \geq \mathbb{E}[(M_K - M_K\tau)] = 0
\end{aligned}$$

(This $(\mathbb{E}[f\mathbf{1}_{\{M_K>0\}}] \geq 0)$ is already a very nice and useful result.)

$\lim_{K\to\infty}\{M_K > 0\} = \{sup_{n\geq 1} A_n f > 0\}$. Let $f = h - \delta$, then $sup_{n\geq 1} A_n f = sup_{n\geq 1} A_n h - \delta$. Finally, we have

$$\mathbb{E}[(h - \delta)\mathbf{1}_{\{\sup_n A_n h > \delta\}}] \geq 0 \quad \Rightarrow \quad \mathbb{E}[|h|] \geq \delta P[\{\sup_n A_n h > \delta]$$

□

6.441: Information Theory	Spring 2013
Lecturer: Yury Polyanskiy	**Lecture 9: Hypothesis Testing**

3.1 Binary Hypothesis Testing

Two possible distributions on a space $\mathcal{X}$

$$\begin{aligned} H_0 &: X \sim P \\ H_1 &: X \sim Q \end{aligned}$$

Where under hypothesis H_0 (the null hypothesis) X is distributed according to P, and under H_1 (the alternative hypothesis) X is distributed according to Q. A *test* between two distributions chooses either H_0 or H_1 based on an observation of X

- Deterministic test: $f : \mathcal{X} \to \{0, 1\}$
- Randomized test: $P_{Z|X} : \mathcal{X} \to [0, 1]$

Let $Z = 1$ denote that the test chooses P, and $Z = 0$ when the test chooses Q.

3.1.1 Performance Metrics

In order to determine the "effectiveness" of a test, we look at two metrics. Let $\pi_{i|j}$ denote the probability of the test choosing i when the correct hypothesis is j. With this

$$\begin{aligned} \alpha &= \pi_{0|0} = P[Z = 0] \quad \text{(Probability of success given } H_0 \text{ true)} \\ \beta &= \pi_{0|1} = Q[Z = 0] \quad \text{(Probability of error given } H_1 \text{ true)} \end{aligned}$$

Remark: $P[Z = 0]$ is a slight abuse of notation, more accurately $P[Z = 0] = \sum_{x \in \mathcal{X}} P(x) P_{Z|X}(0|x) = \mathbb{E}_{X \sim P_X}[1 - f(x)]$. Also, the choice of these two metrics to judge the test is not unique, we can use many other pairs from $\{\pi_{0|0}, \pi_{0|1}, \pi_{1|0}, \pi_{1|1}\}$.

So for any test $P_{Z|X}$ there is an associated (α, β). There are a few ways to determine the "best test"

- Bayesian - Assume prior distributions $\mathbb{P}[H_0] = \pi_0$ and $\mathbb{P}[H_1] = \pi_1$, minimize the expected error

$$P_b^* = \min_{\text{tests}} \pi_0 \pi_{1|0} + \pi_1 \pi_{0|1}$$

Typed by Austin Collins, austinc@mit.edu, March 7, 2013

- Minimax - Assume there is a prior distribution but it is unknown, so choose the test that preforms the best for the worst case priors

$$P_m^* = \min_{\text{tests}} \max_{\pi_0} \pi_0 \pi_{1|0} + \pi_1 \pi_{0|1}$$

- Neyman Pearson - Minimize error β subject to success probability at least α

In this course, the Neyman Pearson test will play the most central role.

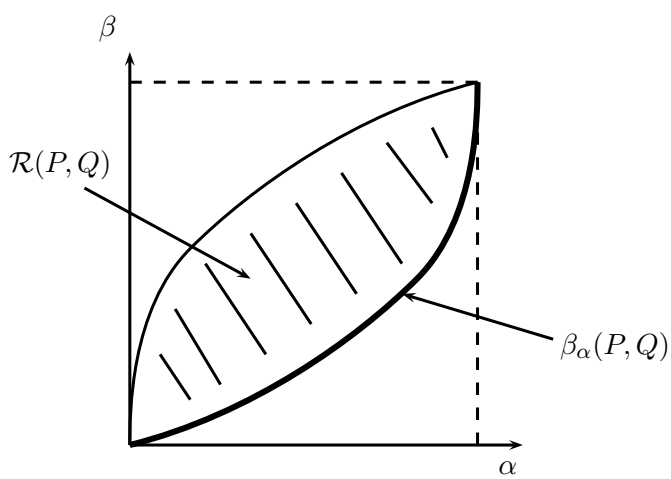
3.2 Neyman Pearson Test

Definition 3.1. *Given that we require $P[Z=0] \geq \alpha$,*

$$\beta_\alpha(P,Q) \triangleq \inf_{P[Z=0]\geq\alpha} Q[Z=0]$$

Definition 3.2. *For given (P,Q), the region of achievable tests is*

$$\mathcal{R}(P,Q) = \{(\alpha,\beta) \ : \ \exists P_{Z|X} \ s.t. \ \alpha = P[Z=0], \beta = Q[Z=0]\} \subset [0,1]^2$$



Remark: This region encodes a lot of useful information about the relationship between P and Q.

Theorem 3.1 (Properties of $\mathcal{R}(P,Q)$)**.**

1. *$\mathcal{R}(P,Q)$ - closed, convex subset of $[0,1]^2$*
2. *Symmetric:* $\alpha \mapsto 1-\alpha$
   $\beta \mapsto 1-\beta$

Proof. **(1)** For convexity, suppose $(\alpha_0,\beta_0),(\alpha_1,\beta_1) \in \mathcal{R}(P,Q)$, then each specifies a test $P_{Z_0|X}, P_{Z_1|X}$ respectively. Randomize between these two test to get the test $\lambda P_{Z_0|X} + \bar{\lambda} P_{Z_1|X}$ for $\lambda \in [0,1]$, which achieves the point $(\lambda\alpha_0 + \bar{\lambda}\alpha_1, \lambda\beta_0 + \bar{\lambda}\beta_1) \in \mathcal{R}(P,Q)$. Closed - later.

(2) If $(\alpha,\beta) \in \mathcal{R}(P,Q)$, then form the test that chooses P whenever $P_{Z|X}$ choses Q, and chooses Q whenever $P_{Z|X}$ choses P, which gives $(1-\alpha, 1-\beta) \in \mathcal{R}(P,Q)$. □

3.3 Bounds on $\mathcal{R}(P, Q)$

Theorem 3.2 (Weak Converse)**.** *If $(\alpha, \beta) \in \mathcal{R}$ then*

$$d(\alpha||\beta) \leq D(P||Q)$$
$$d(\beta||\alpha) \leq D(Q||P)$$

Where $d(\cdot||\cdot)$ is the binary divergence.

Proof. Use data processing with $P_{Z|X}$. □

Definition 3.3. *The log likelihood ratio is $F = \log \frac{dP}{dQ} \ : \mathcal{X} \to \mathbb{R} \cup \{\pm\infty\}$*

Theorem 3.3.

1. *F is a sufficient statistic for testing H_0 vs H_1.*
2. *$Q_F(f) = \exp(-f)P_F(f)$ (or $dQ_F = e^f dP_F$)*

Proof. **(2)**

$$Q_F(f) = \sum_{\mathcal{X}} Q(x)\mathbf{1}\{\log \frac{P(x)}{Q(x)} = f\} = \sum_{\mathcal{X}} Q(x)\mathbf{1}\{e^f Q(x) = P(x)\}$$
$$= e^{-f} \sum_{\mathcal{X}} P(x)\mathbf{1}\{\log \frac{P(x)}{Q(x)} = f\} = e^{-f} P_F(f)$$

(1) To show F is a s.s, we need to show $P_{X|F} = Q_{X|F}$

$$P_{X|F} = \frac{P_X(x)\mathbf{1}\{e^f = \frac{P(x)}{Q(x)}\}}{P_F(f)} = \frac{e^f Q(x)\mathbf{1}\{e^f = \frac{P(x)}{Q(x)}\}}{P_F(f)}$$
$$= \frac{Q_{XF}}{e^{-f}P_F(f)} = \frac{Q_{XF}}{Q_F} = Q_{X|F}$$

□

Lemma 3.1. $\forall E,\ \forall \gamma :\ P[E] - \gamma Q[E] \leq P[\log \frac{dP}{dQ} > \log \gamma]$

Proof. Suppose $P, Q \ll \mu$ for some measure μ (this is WLOG, since we can always take $\mu = P + Q$), then $P(dx) = p(x)d\mu,\ Q(dx) = q(x)d\mu$. Then

$$P[E] - \gamma Q[E] = \int_E d\mu(p(x) - \gamma q(x)) \leq \int_{\mathcal{X}} d\mu |p(x) - \gamma q(x)|^+$$

Now $p(x) - \gamma q(x) > 0 \Leftrightarrow \log \frac{p(x)}{q(x)} > \log \gamma \implies |p(x) - \gamma q(x)|^+ \leq p(x)\mathbf{1}\{\log \frac{p(x)}{q(x)} > \log \gamma\}$ for all x, so continuing

$$\leq \int d\mu\ p(x)\mathbf{1}\{\log \frac{p(x)}{q(x)}\} = P[\log \frac{dP}{dQ} > \log \gamma]$$

□

Note: In this case, we do not need $P \ll Q$, since $\pm\infty$ is a reasonable and meaningful value for the log likelihood ratio.

Lemma 3.2. $P[Z=0] - \gamma Q[Z=0] \leq P[F > \log\gamma]$

Proof. Follows directly from the previous theorem. □

Theorem 3.4 (Strong Converse)**.** $\forall(\alpha,\beta) \in \mathcal{R}(P,Q)$,

$$\alpha - \gamma\beta \leq P[\log\frac{dP}{dQ} > \log\gamma]$$
$$\beta - \frac{1}{\gamma}\alpha \leq Q[\log\frac{dP}{dQ} > \log\gamma]$$

Proof. Apply lemma. □

Remark: To apply the strong converse, we need to know the CDF of F, whereas to apply the weak converse we need only to know the expectation of F.

Theorem 3.5 (Neyman Pearson)**.** *Tests* $X \to \mathbf{1}\{\log\frac{dP}{dQ} \leq \tau\}$ *are optimal, namely*

$$\alpha = P[\log\frac{dP}{dQ} > \tau]$$
$$\beta = Q[\log\frac{dP}{dQ} > \tau]$$

Parametrize all extremal points on the boundary of $\mathcal{R}(P,Q)$.

Proof. For any extremal point on the boundary (α^*,β^*), $\exists$ supporting line (hyperplane) touching the boundary at the point

$$\alpha^* - \gamma\beta^* = \max_{tests}(\alpha - \gamma\beta) = \max_{P_{Z|X}}\sum_{\mathcal{X}}(P(x) - \gamma Q(x))P_{Z|X}(0|x) = \sum_{\mathcal{X}}|P(x) - \gamma Q(x)|^+$$

In the last equality, we are free to choose $P_{Z|X}(0|x)$, select $P_{Z|X}(0|x) = 1$ whenever $P(x) - \gamma Q(x) > 0$, and 0 otherwise. Hence the test is

$$Z = \mathbf{1}\{\log\frac{dP}{dQ} > \log\gamma\}$$

Therefore, for any τ, there is an optimal log likelihood threshold test achieving a point on the boundary of $\mathcal{R}(P,Q)$. □

Corollary 3.1. $\forall\tau$, *and when*

$$\alpha = P[\log\frac{dP}{dQ} > \tau]$$
$$\beta = \exp(-\tau)P[\log\frac{dP}{dQ} > \tau] \leq \exp(-\tau)$$

Then $(\alpha,\beta) \in \mathcal{R}(P,Q)$

Proof.

$$
\begin{aligned}
Q[\log \frac{dP}{dQ} > \tau] &= \sum Q(x)\mathbf{1}\{\frac{P(x)}{Q(x)} > e^{\tau}\} \\
&\leq \sum P(x)e^{-\tau}\mathbf{1}\{\frac{P(x)}{Q(x)} > e^{\tau}\} = e^{-\tau}P[F > \tau]
\end{aligned}
$$

□

3.4 Asymptotics

Now we have many samples from the underlying distribution

$$
\begin{aligned}
H_0 &: X^n \sim P_{X^n} \\
H_1 &: X^n \sim Q_{X^n}
\end{aligned}
$$

We're interested in the asymptotics of the performance metrics. There are two main types of tests, both which the convergence rate to zero error is exponential.

1. Stein Regime: What is the best exponential rate of convergence for $\pi_{0|1}$ when $\pi_{1|0}$ has to be $\leq \epsilon$?

$$
\begin{cases}
\pi_{1|0} \leq \epsilon \\
\pi_{0|1} \longrightarrow 0
\end{cases}
$$

2. Chernoff Regime: What is the trade off between exponents of the convergence rates of $\pi_{1|0}$ and $\pi_{0|1}$ when we want both errors to go to 0?

$$
\begin{cases}
\pi_{1|0} \longrightarrow 0 \\
\pi_{0|1} \longrightarrow 0
\end{cases}
$$

6.441: Information Theory	Spring 2012
Lecturer: Yury Polyanskiy **Lecture 10: Binary Hypothesis Testing, asymptotics**	

setup:

$$\begin{aligned}
&H_0 : X^n \sim P_{X^n} \qquad H_1 : X^n \sim Q_{X^n} \\
&\text{test } P_{Z|X^n} : \mathcal{X}^n \to \{0,1\} \\
&\text{specification } 1-\alpha = \pi_{1|0} \qquad \beta = \pi_{0|1}
\end{aligned}$$

10.1 Stein's regime

$$\begin{aligned}
&1-\alpha = \pi_{1|0} \le \epsilon \\
&\beta = \pi_{0|1} \to 0 \quad \text{at the rate } 2^{-nV_\epsilon}
\end{aligned}$$

Note: interpretation of this specification, usually a "miss"$(0|1)$ is much worse than a "false alarm" $(1|0)$.

Definition 10.1 (ϵ-optimal exponent)**.** *V_ϵ is called an ϵ-optimal exponent in Stein's regime if*

$$\begin{aligned}
V_\epsilon &= \sup\{E : \forall n, \exists P_{Z|X^n} \ s.t. \ \alpha > 1-\epsilon, \beta < 2^{-nE}, \} \\
\Leftrightarrow V_\epsilon &= \lim \inf_{n\to\infty} -\frac{1}{n} \log \beta_{1-\epsilon}(P_{X^n}, Q_{X^n})
\end{aligned}$$

where $\beta_\alpha(P,Q) = \min_{P_{Z|X}, P(Z=0)\ge\alpha} Q(Z=0)$.

Note: show the equivalence (DIY)

Definition 10.2 (Stein's exponent)**.**

$$V = \lim_{\epsilon\to 0} V_\epsilon$$

Theorem 10.1 (Stein's lemma)**.** *let $P_{X^n} = P_X^n$ i.i.d. and $Q_{X^n} = Q_X^n$ i.i.d. then $\forall \epsilon \in (0,1)$,*

$$V_\epsilon = D(P\|Q)$$

Example: If it is required that $\alpha \ge 1-10^{-3}$, and $\beta \le 10^{-40}$, what's the number of samples needed? Stein's lemma provides a rule of thumb: $n \gtrsim -\frac{\log 10^{-40}}{D(P\|Q)}$.

Typed by Qingqing Huang (qqh@mit.edu), Mar. 12, 2013

Proof. Denote $F = \log dP/dQ$, and $F_n = \log dP_{X^n}/dQ_{X^n}$.

Recall Neyman Pearson's lemma on optimal tests: $\forall \tau$,

$$\alpha = P(F > \tau), \quad \beta = Q(F > \tau) \leq e^{-\tau}$$

Also notice that by LLN,

$$\lim_{n\to\infty} \frac{1}{n}\mathbb{E}_P[F_n] = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} E_P[\frac{dP(X_i)}{dQ(X_i)}] = D(P\|Q) \tag{1}$$

1. show $V_\epsilon \geq D(P\|Q) = D$.
   pick $\tau = n(D-\delta)$, for some $0 < \delta \ll 1$. Then the optimal test achieves:

   $$\alpha = P(F_n > n(D-\delta)) \to 1, \text{ by } (\textbf{??})$$
   $$\beta \leq e^{-n(D-\delta)}$$

   then pick n large enough (depends on ϵ, δ) such that $\alpha \geq 1-\epsilon$, we have the exponent $E = D-\delta$ achievable, $V_\epsilon \geq E$. Further let $\delta \to 0$, we have that $V_\epsilon \geq D$.

2. show $V_\epsilon \leq D(P\|Q) = D$.

   (a) (weak converse) $\forall(\alpha,\beta) \in \mathcal{R}(P_{X^n}, Q_{X^n})$, we have

   $$-h(\alpha) + \alpha \log\frac{1}{\beta} \leq d(\alpha\|\beta) \leq D(P_{X^n}\|Q_{X^n}) \tag{2}$$

   where the first inequality is due to

   $$d(\alpha\|\beta) = \alpha\log\frac{\alpha}{\beta} + \bar{\alpha}\log\frac{\bar{\alpha}}{\bar{\beta}} = -h(\alpha) + \alpha\log\frac{1}{\beta} + \underbrace{\bar{\alpha}\log\frac{1}{\bar{\beta}}}_{\text{close to 0 for small } \beta}$$

   and the second is due to data processing inequality.
   $\forall$ achivable exponent E, construct a sequence of tests $P_{Z|X^n}$ such that $\alpha_n \geq 1-\epsilon$ and $\beta_n \leq 2^{-nE}$, plug it in the above inequality,

   $$-h(1-\epsilon) + (1-\epsilon)nE \leq nD(P\|Q) \Rightarrow E \leq \frac{D(P\|Q)}{1-\epsilon} + \underbrace{\frac{1}{n}\frac{h(1-\epsilon)}{1-\epsilon}}_{\to 0, \text{ as } n\to 0}$$

   therefore

   $$V_\epsilon \leq \frac{D(P\|Q)}{1-\epsilon}$$

   Notice that the weak converse result is tight for $\epsilon \to 0$ (Stein's exponent).

   (b) (strong converse) In proving the weak converse, we only made use of the expectation of F_n in (**??**), we need to make use of the entire distribution in order to obtain stronger results.
   Recall the strong converse result which we showed in the last lecture:

   $$\forall(\alpha,\beta) \in \mathcal{R}(P,Q), \forall\gamma, \quad \alpha - \gamma\beta \leq P(F > \log\gamma)$$

Here, let's take a sequence of tests $P_{Z|X_n}$ we have $\alpha_n \geq 1-\epsilon$ and $\beta_n \leq 2^{-nE}$, and thus

$$1-\epsilon-\gamma 2^{-nE} \leq \alpha_n - \gamma\beta_n \leq P_{X^n}[F_n > \log\gamma]$$

pick $\log\gamma = n(D+\delta)$, by (**??**) the RHS goes to 0, and we have

$$\begin{aligned}
&1-\epsilon-2^{n(D+\delta)}2^{-nE} \leq o(1)\\
\Rightarrow &D+\delta-E \geq \frac{1}{n}\log(1-\epsilon+o(1)) \to 0\\
\Rightarrow &E \leq D \text{ as } \delta \to 0\\
\Rightarrow &V_\epsilon \leq D
\end{aligned}$$

□

Note: [Ergodic] Just like in last section of data compression. Ergodic assumptions on P_{X^n} and Q_{X^n} allow one to show that

$$V_\epsilon = \lim_{n\to\infty}\frac{1}{n}D(P_{X^n}\|Q_{X^n})$$

the counterpart of (**??**), which is the key for picking the appropriate τ, for Ergodic sequence X^n is the Birkhoff-Khintchine convergence theorem.

Note: The theoretical importance of knowing the Stein's exponents is that:

$$\forall E \subset \mathcal{X}^n, \quad P_{X^n}[E] \geq 1-\epsilon \quad \Rightarrow Q_{X^n}[E] \geq 2^{-nV_\epsilon+o(n)}$$

Thus knowledge of Stein's exponent V_ϵ allows one to prove exponential bounds on probabilities of arbitrary sets, the technique is known as "change of measure".

10.2 (E_0, E_1) Tradeoff regime

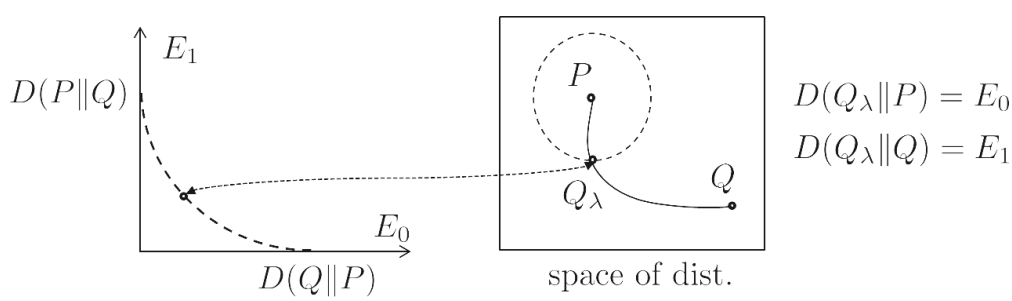
We are still considering i.i.d. sequence X^n, and binary hypothesis

$$H_0 : X^n \sim P_X^n \qquad H_1 : X^n \sim Q_X^n$$

In this section, we shall look at the following specification:

$$\begin{aligned}
&1-\alpha = \pi_{1|0} \to 0 \quad \text{at the rate } 2^{-nE_0}\\
&\beta = \pi_{0|1} \to 0 \quad \text{at the rate } 2^{-nE_1}
\end{aligned}$$

what's the achievable region of (E_0, E_1)?



characterize the boundary of the achievable region of (E_0, E_1)

The optimal tests give the explict error probability:

$$\alpha_n = P[F_n > \tau], \quad \beta_n = Q[F_n > \tau]$$

We showed that when fix $\alpha_n \sim 1 - \epsilon$, namely $E_0 = 0$, picking $\tau = nD(P\|Q)$ gives the exponential convergence rate of β_n as $E_1 = D(P\|Q)$. More generally, to show the tradeoff between E_0 and E_1, we need to introduce a powerful tool – Large deviation theory.

10.3 Review of Large deviation theory

Let X^n be an i.i.d. sequence and $X_i \sim P$. Large deviation focuses on the following inequality:

$$P[\sum_{i=1}^{n} X_i \geq n\gamma] = 2^{-nE(\gamma)+o(n)}$$

what is the rate function $E(\gamma) = -\lim_{n\to\infty} \frac{1}{n} \log P[\frac{\sum_{i=1}^{n} X_i}{n} \geq \gamma]$? (Chernoff's ineq.)

Let's first introduce the two key quantities: log MGF $\psi_X(\lambda)$ and tilted distribution Q_λ.

10.3.1 log MGF

Definition 10.3 (Cumulant generating function)**.** *(log MGF)*

$$\psi_X(\lambda) = \log(\mathbb{E}[e^{\lambda X}]), \quad \lambda \in \mathbf{R}.$$

Note: In this section, we shall restrict to the distributions $P(X)$ such that $\forall \lambda \in \mathbf{R}, \psi_X(\lambda) < \infty$, and $X \neq$const.
An example of R.V. such that $\psi_X(\lambda)$ does not exist: Z^3, $Z \sim$ Gaussian.

Theorem 10.2 (Properties of $\psi_X(\lambda)$)**.**

1. *$\psi_X(\lambda)$ is convex;*
2. *$\psi_X(\lambda)$ is continuous;*
3. *$\psi_X(\lambda)$ is differentiable and*
$$\psi'_X(\lambda) = e^{-\psi_X(\lambda)}\mathbb{E}[Xe^{\lambda X}];$$
4. *if the support of $P(X) \subset [a, b]$, then $a \leq \psi'_X \leq b$*
5. *let $A = \inf_{\lambda\in\mathbf{R}} \psi'_X(\lambda)$, $B = \sup_{\lambda\in\mathbf{R}} \psi'_X(\lambda)$, then the support of $P_X \subset [A, B]$;*
   Note*: 4). and 5). together show that supp P_X = range ψ'_X.*
6. *$\psi_X(\lambda)$ is strictly convex, and $\psi'_X(\lambda)$ is strictly increasing.*

Proof.

1. Recall Holder's inequality:

$$\mathbb{E}[UV] \leq \|U\|_p \|V\|_q, \quad \text{for } p, q > 0, \frac{1}{p} + \frac{1}{q} = 1$$

Apply it to $\mathbb{E}[e^{(\theta\lambda_1 + \bar{\theta}\lambda_2)X}]$, pick $p = 1/\theta, q = 1/\bar{\theta}$, and show that $e^{\psi_X(\theta\lambda_1 + \bar{\theta}\lambda_2)} \leq e^{\psi_X(\lambda_1)\theta} e^{\psi_X(\lambda_2)\bar{\theta}}$.

2. By assumption on X, domain of ψ_X is $\mathbf{R}$, and by the fact that convex function must be continuous on the interior of its domain, we have that ψ_X is continuous on $\mathbf{R}$.

3. Be careful when exchanging the order of differentiation and expectation.

   Assume $\lambda > 0$ (similar for $\lambda \leq 0$).
   First, we show that $\log \mathbb{E}[|Xe^{\lambda X}|]$ exists. Since

$$\begin{aligned} e^{|X|} &\leq e^X + e^{-X} \\ |Xe^{\lambda X}| &\leq e^{|(\lambda+1)X|} \leq e^{(\lambda+1)X} + e^{-(\lambda+1)X} \end{aligned}$$

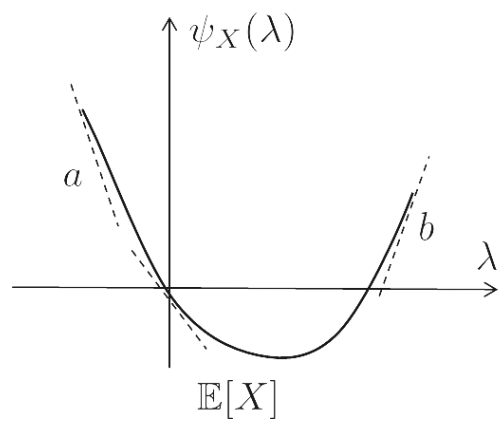
   by assumption on X, both of the summands are absolutely integrable in X. Therefore by DCT, $\mathbb{E}[|Xe^{\lambda X}|]$ exists and is continuous in λ.

   Second, by the existence and continuity of $E[|Xe^{\lambda X}|]$, we can switch order of integration and differentiation and have

$$\begin{aligned} e^{\psi_X(\lambda)} &= \mathbb{E}[e^{\lambda X}] = \mathbb{E}[1 + \int_{u=0}^{\lambda} Xe^{uX} du] \\ \psi_X'(\lambda) e^{\psi_X(\lambda)} &= \mathbb{E}[Xe^{\lambda X}] \end{aligned}$$

   thus $\psi_X'(\lambda) = e^{-\psi_X(\lambda)} \mathbb{E}[Xe^{\lambda X}]$ exists and is continuous in λ on $\mathbf{R}$.

4. $$a \leq X \leq b \Rightarrow a \leq e^{-\psi_X(\lambda)} \mathbb{E}[ae^{\lambda X}] \leq \psi_X' \leq e^{-\psi_X(\lambda)} \mathbb{E}[be^{\lambda X}] = b$$



5. Suppose $P_X[X > B] > 0$, then $P_X[X > B + 2\epsilon] > 0$ for some $\epsilon > 0$. But then $Q_\lambda[X \leq B + \epsilon] \to 0$ for $\lambda \to \infty$. On the other hand, we know $E_{Q_\lambda}[X] \leq B$. A contradiction.

6. Suppose $\psi_X(\lambda)$ is not strictly convex. Then there exists a small neighborhood of λ_0 such that $\psi_X(\lambda_0 + u) = \psi_X(\lambda_0) + ur$ Then $\psi_{Q_\lambda}(u) = ur$ for all u in small neighborhood of zero, or equivalently $\mathbb{E}_{Q_\lambda}[e^{u(X-r)}] = 1$ for u small.

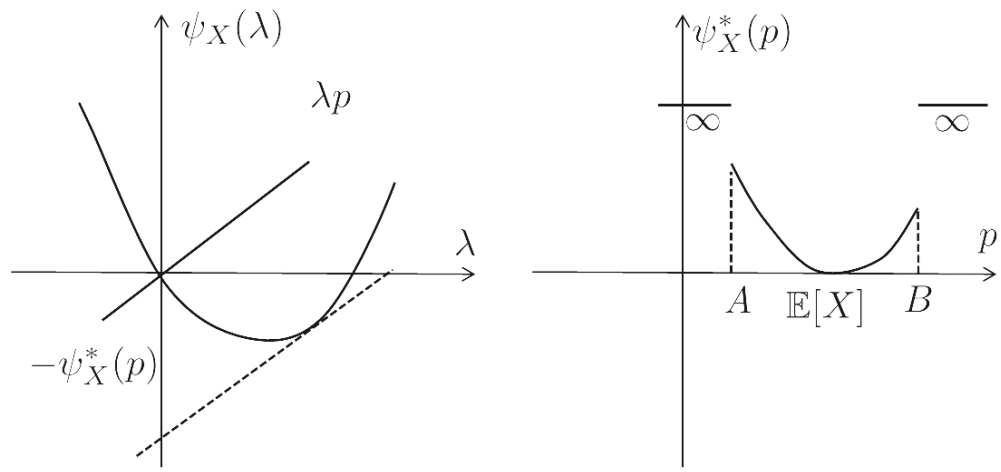
□

Lemma 10.1. $\mathbb{E}[e^{uS}] = 1$ *for all* $u \in (-\epsilon, \epsilon)$ *then* $S = 0$.

Proof. expand in Taylor series around $u = 0$ to obtain $E[S] = 0$, $E[S^2] = 0$. □

Definition 10.4 (rate function)**.** *The rate function* $\psi_X^*(p)$ *is given by the Legendre-Fenchel dual of the log MGF:*

$$\psi_X^*(p) = \sup_{\lambda \in \mathbf{R}} \lambda p - \psi_X(\lambda)$$



Theorem 10.3 (Properties of $\psi_X^*(p)$)**.**

1. $$\psi_X^*(p) = \begin{cases} \lambda_0 p - \psi_X(\lambda_0) & p = \psi_X'(\lambda_0) \text{ for some } \lambda_0 \in (A, B) \\ ? & p = A \text{ or } B \\ +\infty & p < A \text{ or } p > B \end{cases}$$

2. $\psi_X^*(p)$ *is strictly convex and strictly positive.*

3. $\psi_X^*(p)$ *is decreasing in the region* $(A, \mathbb{E}[X])$*, and is increasing in the region* $[\mathbb{E}[X], B)$

10.3.2 Tilted distribution

The tilted measure Q_λ of X is defined as follows:

$$dQ_\lambda = e^{\lambda X - \psi_X(\lambda)} dP \tag{3}$$

Theorem 10.4 (Properties of Q_λ)**.**

1. $$\psi_{Q_\lambda}(u) = \psi_X(\lambda + u) - \psi_X(\lambda)$$

2.

$$\mathbb{E}_{Q_\lambda}[X] = \psi'_X(\lambda)$$
$$D(Q_\lambda \| P) = \psi^*_X(\mathbb{E}_{Q_\lambda}[X])$$

Note: *tilting trades mean for divergence*

3.

$$P(X > b) > 0 \Rightarrow \forall \epsilon > 0, Q_\lambda(X \leq b - \epsilon) \to 0 \ \textit{as} \ \lambda \to \infty$$
$$P(X < a) > 0 \Rightarrow \forall \epsilon > 0, Q_\lambda(X \geq a - \epsilon) \to 0 \ \textit{as} \ \lambda \to \infty$$

Proof. 1. and 2. are simple. (DIY)
3.

$$\begin{aligned} Q_\lambda(X \leq b - \epsilon) &= \mathbb{E}_P[e^{\lambda X - \psi_X(\lambda)} \mathbf{1}[X \leq b - \epsilon]] \\ &\leq \mathbb{E}_P[e^{\lambda(b-\epsilon) - \psi_X(\lambda)} \mathbf{1}[X \leq b - \epsilon]] \\ &\leq e^{-\lambda\epsilon} e^{\lambda b - \psi_X(\lambda)} P[X \leq b - \epsilon] \\ &\leq e^{-\lambda\epsilon} \frac{P[X \leq b - \epsilon]}{P[X \leq b]} \to 0 \text{ as } \lambda \to \infty \end{aligned}$$

where the last inequality is due to:

$$\begin{aligned} Q_\lambda(X > b) &= \mathbb{E}_P[e^{\lambda X - \psi_X(\lambda)} \mathbf{1}[X > b] \geq P[X > b] e^{\lambda b - \psi_X(\lambda)} \\ \Rightarrow e^{\lambda b - \psi_X(\lambda)} &\leq \frac{Q_\lambda(X > b)}{P[X > b]} \leq \frac{1}{P[X \leq b]} \end{aligned}$$

□

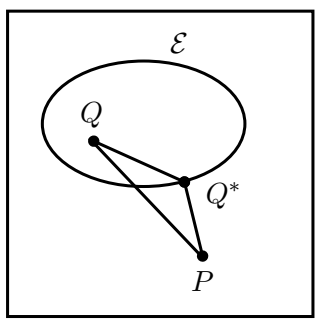
6.441: Information Theory	Spring 2013
Lecturer: Yury Polyanskiy	**Lecture 11: Information Projection**

11.1 Information Projection

Information Projection Problem: Let $\mathcal{E}$ be a convex set of distributions on $\mathcal{X}$, then for the distribution P on $\mathcal{X}$, we want

$$\inf_{Q \in \mathcal{E}} D(Q||P)$$

Denote the minimizing distribution Q by Q^*, and let $X \sim P$. The main idea is that the "line" between P and optimal Q^* is orthogonal to $\mathcal{E}$.



Distributions on $\mathcal{X}$

Theorem 11.1. *Suppose* $\exists Q^* \in \mathcal{E}$ *such that* $Q^* = \min_{Q \in \mathcal{E}} D(Q^*||P)$*, then* $\forall Q \in \mathcal{E}$

$$D(Q||P) \geq D(Q||Q^*) + D(Q^*||P)$$

Proof. If $D(Q||P) = \infty$, then we're done, so assume that $D(Q^*||P)D(Q||P) < \infty$. For $\theta \in [0, 1]$, form the convex combination $Q' = \bar{\theta}Q^* + \theta Q \in \mathcal{E}$. Since Q^* is the minimizer of $D(Q||P)$

$$0 \leq \frac{\partial}{\partial \theta} D(Q'||P) = D(Q||P) - D(Q||Q^*) - D(Q^*||P)$$

(found by taking the derivative and matching terms, DIY) and we're done. □

Remark: If we view the picture above in the Euclidean setting, the triangle formed by P, Q^* and Q (for Q^*, Q in a convex set, P outside the set) is always obtuse, and is a right triangle only when the convex set has a "flat face". In this sense, the divergence is similar to the squared Euclidean distance, and the above theorem is sometimes known as a "Pythagorean" theorem.

Typed by Austin Collins, austinc@mit.edu, March 14, 2013

The interesting set of Q's that we'll particularize to is the half-space of distributions $\mathcal{E} = \{Q : \mathbb{E}_Q[X] \geq \gamma\}$. The following theorem characterizes the minimizers over this set of distributions.

Theorem 11.2.

1. *The information projection problem over $\mathcal{E} = \{Q : \mathbb{E}_Q[X] \geq \gamma\}$ has solution*

$$\min_{Q\,:\,\mathbb{E}_Q[X]\geq\gamma} D(P||Q) = \begin{cases} 0 & \gamma < \mathbb{E}_P[X] \\ \psi^*(\gamma) & \mathbb{E}_P[X] \leq \gamma < B \\ ? & \gamma = B \\ +\infty & \gamma > B \end{cases}$$

   Where $B = \sup_{\lambda\in\mathbb{R}} \psi'_X(\lambda)$ (so that along with $A = \inf_{\lambda\in\mathbb{R}} \psi'_X(\lambda)$, we have supp$(P) \subset [A, B]$), and "?" denotes that we can't say anything about the value when $\gamma = B$.

2. *The minimizing distribution is unique and equal to a tilted measure Q_λ.*

3. $\min_{\mathbb{E}_Q[X]\geq\gamma} D(P||Q) = \inf_{\mathbb{E}_Q[X]>\gamma} D(P||Q) = \min_{\mathbb{E}_Q[X]=\gamma} D(P||Q)$, *where the last equality holds in the region $\mathbb{E}_P[X] \leq \gamma \leq B$.*

Proof. The first, third, and fourth are obvious (to see the fourth, when $\mathbb{E}_Q[X] > B$, then $Q \not\ll P \implies D(Q||P) = \infty$).

Second case: Fix $\mathbb{E}_P[X] \leq \gamma < B$, and find the unique λ such that $\psi'_X(\lambda) = \gamma = \mathbb{E}_{Q_\lambda}[X]$ where $Q_\lambda = \exp(\lambda X - \psi_X(\lambda))dP$. This corresponds to tilting P far enough to the right to move its mean to γ, in particular $\lambda \geq 0$. Take any Q such that $\mathbb{E}_Q[X] \geq \gamma$, then

$$\begin{aligned} D(P||Q) &= \mathbb{E}_Q[\log \frac{dQdQ_\lambda}{dPdQ_\lambda}] \\ &= D(Q||Q_\lambda) + \mathbb{E}_Q[\log \frac{dQ_\lambda}{dP}] \\ &= D(Q||Q_\lambda) + \mathbb{E}_Q[\lambda X - \psi_X(\lambda)] \\ &\geq D(Q||Q_\lambda) + \lambda\gamma - \psi_X(\lambda) = D(Q||Q_\lambda) + \psi^*_X(\gamma) \end{aligned}$$

The last expression is minimized when $Q = Q_\lambda$. In addition, this shows the minimizer is unique, proving 2. 3 follows from the next corollary. □

Corollary 11.1. *$\forall Q$ with $\mathbb{E}_Q[X] \in (A, B)$ and $D(Q||P) < \infty$, $\exists! Q_\lambda$ such that*

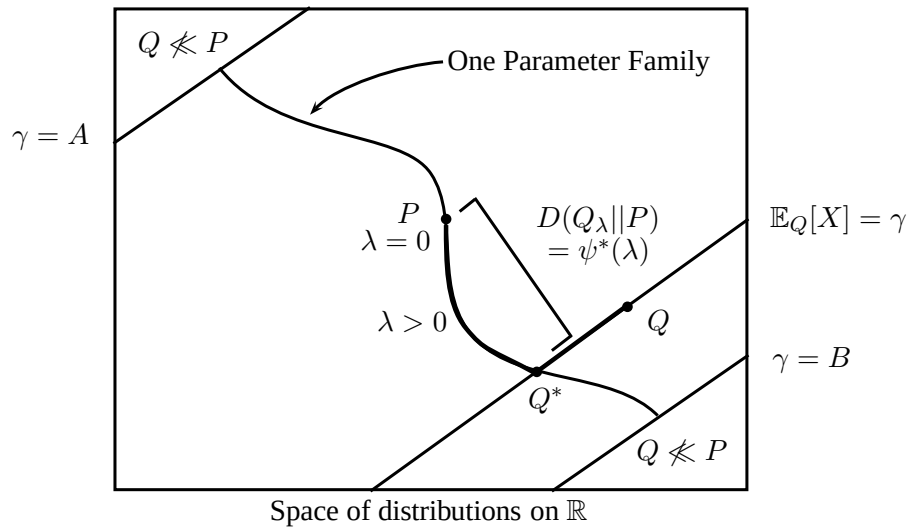
$$\begin{aligned} &\mathbb{E}_{Q_\lambda}[X] = \mathbb{E}_Q[X] \\ &D(Q_\lambda||P) \leq D(Q||P) \end{aligned}$$

Proof. $D(Q||P) = D(Q||Q_\lambda) + D(Q_\lambda||P)$ along with the previous theorem. □

Remark: For any Q, this allows us to find a tilted measure Q_λ that has the same mean yet smaller divergence.

11.2 Interpretation of Information Projection

The following picture describes many properties of information projections.



- Each set $\{Q \ : \ \mathbb{E}_Q[X] = \gamma\}$ corresponds to a slice. As γ varies from A to B, the curves fill the entire space minus the corner regions.
- When $\gamma < A$ or $\gamma > B$, $Q \not\ll P$.
- As λ varies, the Q_λ's trace out a curve via $\psi^*(\lambda) = D(Q_\lambda || P)$. This set of distributions is called a one parameter family, or exponential family.
- **Key Point:** The one parameter family curve intersects each γ-slice orthogonally at the minimizing Q^* in $\{Q \ : \ \mathbb{E}_Q[X] = \gamma\}$, and the distance from P to Q^* is given by $\psi^*(\lambda)$.

11.3 Application to Large Deviations

Large Deviations Problem: Large deviations problems make statements about the tail probabilities of a sequence of distributions. We're interested in the rate $E(\gamma)$ at which for iid X_k

$$\mathbb{P}\left[\frac{1}{n}\sum_{k=1}^{n} X_k \geq \gamma\right] = e^{-nE(\gamma)+O(n)}$$

Theorem 11.3. *The exponent above is given by*

$$E[\gamma] = \begin{cases} 0 & \gamma \leq \mathbb{E}_P[X] \\ \psi^*(\gamma) & \mathbb{E}_P[X] \leq \gamma < B \\ ? & \gamma = B \\ +\infty & \gamma > B \end{cases}$$

Proof. Again, the first, third, and fourth cases are obvious. For the second case, we give a bound from below and above, then show that these bounds agree.

Lower Bound on $\mathbb{P}[\cdot]$: Define Q such that $X^n \sim$ iid Q, then $\mathbb{E}_Q[X] > \gamma$. Let $E_n = \{\sum_{k=1}^n X_k \geq n\gamma\}$. We want to know $P[E_n]$. First

$$Q[E_n] = Q[\sum_{k=1}^n X_k \geq n\gamma] = 1 - o(1)$$

Now data processing gives

$$d(Q[E_n]||P[E_n]) \leq D(Q_{X^n}||P_{X^n}) = nD(P||Q)$$

And a lower bound for the binary divergence is

$$d(Q[E_n]||P[E_n]) \geq -h(Q[E_n]) + Q[E_n] \log \frac{1}{P[E_n]}$$

Combining the two bounds on on $d(Q[E_n]||P[E_n])$ gives

$$P[\sum_{k=1}^n X_k \geq n\gamma] \geq \exp\left(\frac{-nD(Q||P) + \log 2}{Q[E_n]}\right)$$

We've proved $E[\gamma] \leq D(Q||P)$ for any $Q \,:\, \mathbb{E}_Q[X] > \gamma$. The best bound occurs at the minimizer

$$E[\gamma] \leq \inf_{\{Q:\mathbb{E}_Q[X]>\gamma\}} D(Q||P) = \min_{\{Q\,:\,\mathbb{E}_Q[X]\geq\gamma\}} D(Q||P) = \psi^*(\gamma)$$

Upper Bound: For all $\lambda \geq 0$, the Chernoff bound gives

$$P[\sum_{k=1}^n X_k \geq n\gamma] \leq e^{-n\gamma\lambda}(\mathbb{E}_P[e^{\lambda X}])^n = e^{-n(\lambda\gamma - \psi_X(\lambda))}$$

Where we used iid X_k's in the expectation. To get the lowest upper bound, we take

$$\sup_{\lambda\geq 0} \lambda\gamma - \psi_X(\lambda) = \sup_{\lambda\in\mathbb{R}} \lambda\gamma - \psi_X(\lambda) = \psi_X^*(\gamma)$$

Where the first equally follows since $\gamma \geq \mathbb{E}_P[X]$.

Finally, by the properties of information projections, the tilting measure minimizes divergence, so the Chernoff bound is tight, hence the upper and lower bounds match. □

Remark: The Chernoff bound is tight precisely because, from information projecting, the lower bound showed that the best change of measure is to change to the tilted measure Q_λ. This shows that the inequality in Chernoff's bound at equality.

11.4 Generalizations

Theorem 11.4 (Sanov's Theorem)**.** *Let $\hat{P}$ be the empirical distribution after observing n samples $X_1, \ldots, X_n \sim$ iid P (so that $\hat{P} = \frac{1}{n}\sum_{j=1}^{n} \delta_{X_j}$). Then*

$$\mathbb{P}[\hat{P} \in \mathcal{E}] = e^{-n \min D(Q||P)}$$

Proof. Write the convex set $\mathcal{E}$ as an intersection of half spaces, then we've already solved the problem for half-spaces $\{Q \,:\, \mathbb{E}_Q[X] \geq \gamma\}$ $\square$

Remark: Sanov's theorem tells us the probability that, after observing n iid samples of a distribution, our empirical distribution is still far away from the true distribution.

6.441: Information Theory	Spring 2012
Lecturer: Yury Polyanskiy	**Lecture 12: Hypothesis Testing Asymptotics**

setup:

$$\begin{aligned}
&H_0 : X^n \sim P_{X^n} \qquad H_1 : X^n \sim Q_{X^n} \quad \text{(i.i.d.)}\\
&\text{test } P_{Z|X^n} : \mathcal{X}^n \to \{0,1\}\\
&\text{specification: } 1-\alpha = \pi_{1|0}^{(n)} \le 2^{-nE_0} \qquad \beta = \pi_{0|1}^{(n)} \le 2^{-nE_1}
\end{aligned}$$

Bounds:

- achievability (Neyman Pearson)

$$\alpha = 1 - \pi_{1|0} = P(F > \tau), \qquad \beta = \pi_{0|1} = Q(F > \tau)$$

- converse (strong)

$$\forall(\alpha,\beta)\text{achievable, } \alpha - \gamma\beta \le P(F > \log\gamma)$$

  where

$$F = \log\frac{dP(X)}{dQ(X)}, \quad F_n = \log\frac{dP(X^n)}{dQ(X^n)}$$

12.1 (E_0, E_1) Tradeoff regime

Define

$$T = \log\frac{dQ}{dP}, \quad T_k = \log\frac{dQ(X_k)}{dP(X_k)}, \quad \text{thus } \log\frac{dQ(X^n)}{dP(X^n)} = \sum_{k=1}^n T_k$$

$$\begin{aligned}
\psi_P(\lambda) &= \log \mathbb{E}_P[e^{\lambda T}]\\
&= \log\sum_x P(x)^{1-\lambda}Q(x)^\lambda = \log\int (dP)^{1-\lambda}(dQ)^\lambda\\
\psi_P^*(\theta) &= \sup_\lambda \theta\lambda - \psi_P(\lambda)
\end{aligned}$$

The large deviation theory will be applied to the sum $\sum_{k=1}^n T_k$. Specifically, let's rewrite the bounds in terms of T:

- achievability (Neyman Pearson)

$$\text{let } \tau = -n\theta, \quad \pi_{1|0}^{(n)} = P(\sum_{k=1}^n T_k \ge n\theta) \quad \pi_{0|1}^{(n)} = Q(\sum_{k=1}^n T_k < n\theta)$$

Typed by Qingqing Huang (qqh@mit.edu), Mar. 21, 2013

- converse (strong)

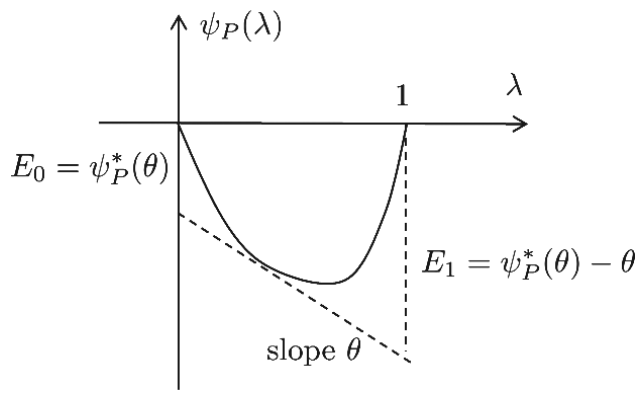
$$\text{let } \gamma = 2^{-n\theta}, \quad \pi_{1|0} + 2^{-n\theta}\pi_{0|1} \geq P(\sum_{k=1}^{n} T_k \geq n\theta)$$

Theorem 12.1. *let $P \ll Q$, $Q \ll P$ then $\forall\theta$ such that $-D(P\|Q) \leq \theta \leq D(Q\|P)$,*

$$E_0(\theta) = \psi_P^*(\theta), \qquad E_1(\theta) = \psi_P^*(\theta) - \theta \tag{1}$$

which characterize the best exponents on the boundary of achievable (E_0, E_1).

Note: the geometric interpretation of the above theorem is shown in the following picture, which rely on the properties of $\psi_P(\lambda)$ and $\psi_P^*(\theta)$:



Corollary 12.1 (Bayesian)**.** *Given the prior (π_0, π_1), the best exponent for the Bayesian error criteria*

$$P_e = \pi_0\pi_{1|0} + \pi_1\pi_{0|1} \leq 2^{-nE}$$

is

$$E = \max_\theta \min(E_0, E_1) = \psi_P^*(0) = C(P, Q)$$

and $\psi_P^*(0)$ is called Chernoff exponent.

Proof.

achievability:
Using Neyman Pearson test, for fixed $\tau = -n\theta$, apply the large deviation theorem:

$$\pi_{1|0}^{(n)} = P(\sum_{k=1}^{n} T_k \geq n\theta) = e^{-n\psi_P^*(\theta)+o(n)}, \qquad \text{for } \theta \geq \mathbb{E}_P T = -D(P\|Q)$$
$$\pi_{0|1}^{(n)} = Q(\sum_{k=1}^{n} T_k < n\theta) = e^{-n\psi_Q^*(\theta)+o(n)}, \qquad \text{for } \theta \leq \mathbb{E}_Q T = D(Q\|P)$$

Notice that by the definition of T we have

$$\psi_Q(\lambda) = \log \mathbb{E}_Q[e^{\lambda \log(Q/P)}] = \log \mathbb{E}_P[e^{(\lambda+1)\log(Q/P)}] = \psi_P(\lambda+1)$$
$$\Rightarrow \psi_Q^*(\theta) = \psi_P^*(\theta) - \theta$$

thus (E_0, E_1) in (1) achievable.

converse (strong):
We want to show that any achievable (E_0, E_1) pair must be below the curve $(E_0(\theta), E_1(\theta))$ in the above Neyman-Pearson test with parameter θ. Apply the strong converse bound we have:

$$\begin{aligned}
&2^{-nE_0} + 2^{-n\theta}2^{-nE_1} \geq 2^{-n\psi_P^*(\theta)+o(n)} \\
\Rightarrow &\min(E_0, E_1 + \theta) \leq \psi_P^*(\theta),\ \forall n, \theta, -D(P\|Q) \leq \theta \leq D(Q\|P) \\
\Rightarrow &\text{ either } E_0 \leq \psi_P^*(\theta) \text{ or } E_1 \leq \psi_P^*(\theta) - \theta
\end{aligned}$$

□

12.2 Alternative forms of Theorem 12.1

Theorem 12.2.

1. $$\textit{for } 0 \leq \lambda \leq 1, \quad E_0 = D(Q_\lambda\|P), \qquad E_1 = D(Q_\lambda\|Q) \tag{2}$$

   where the distribution $Q_\lambda(X)$ which moves distribution P to Q smoothly as λ ranges from 0 to 1 is defined as:

   $$\begin{aligned}
   Q_\lambda(X) &= P^{1-\lambda}(X)Q^\lambda(X) \\
   dQ_\lambda &= (dP)^{1-\lambda}(dQ)^\lambda
   \end{aligned}$$

2. $$E_1 = \min_{Q':D(Q'\|P)\leq E_0} D(Q'\|Q) \tag{3}$$

(2) and (3) characterize the same boundary as in Theorem 12.1.

Proof.

1. Pick $\tau = -n\theta$ in the Neyman-Pearson test,

   $$\begin{aligned}
   E_0(\theta) = \psi_P^*(\theta) &= \min_{Q':\mathbb{E}_{Q'}[T]\geq\theta} D(Q'\|P) = \min_{Q':\mathbb{E}_{Q'}[T]=\theta} D(Q'\|P) \\
   &= \min_{Q':\mathbb{E}_{Q'}[T]=\theta} D(Q'\|Q) + \theta
   \end{aligned}$$

   since

   $$\theta = \mathbb{E}_{Q'}[T] = \mathbb{E}_{Q'}[\log \frac{dQ}{dQ'}\frac{dQ'}{dP}] = D(Q'\|P) - D(Q'\|Q)$$

   also

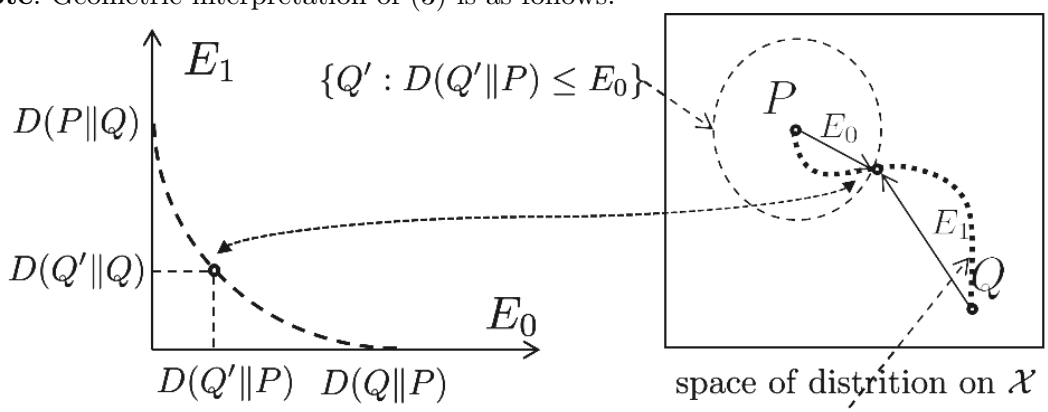
   $$E_1(\theta) = \min_{Q':\mathbb{E}_{Q'}[T]\geq\theta} D(Q'\|Q)$$

   we know that for both of the information projection problems achieve their optimality with some tilted distributions, and in this case, they correspond to the same tilted distribution Q_λ.

2. Suppose that for distribution Q', $D(Q'\|P) = E_0(\theta)$ for some θ, we want to show that $D(Q'\|Q) \geq E_1(\theta)$ in (3).
   Suppose that for distribution Q_λ, $D(Q_\lambda\|P) = E_0(\theta), D(Q_\lambda\|Q) = E_1(\theta)$, and assume (to arrive at contradiction) $D(Q'\|Q) < E_1(\theta)$.
   But then recall that $E_1(\theta) = \min_{Q^*:D(Q^*\|P)-D(Q^*\|P)\geq\theta} D(Q^*\|Q)$. By picking $Q^* = Q'$ and since $D(Q'\|P) - D(Q'\|Q) > E_0(\theta) - E_1(\theta) = \theta$ we have $E_1(\theta) \leq D(Q'\|Q) < E_1(\theta)$ contradiction.

□

Note: Geometric interpretation of (3) is as follows:

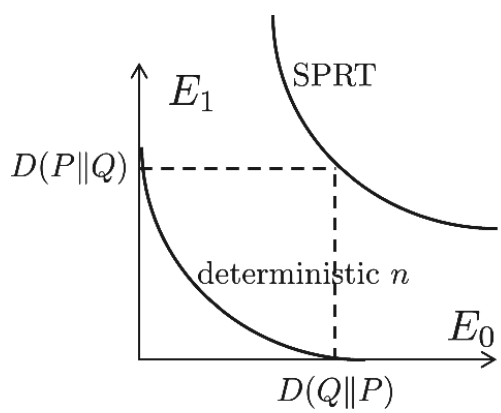


Non-linearity of the boundary corresponds to the scenario when the triangle inequality is not "="

$\forall$ distribution Q' in the tilted family, it minimizes E_0, E_1 simultaneously.
$\exists$ a unique optimal path from P to Q

12.3 Sequential Hypothesis Testing

Different realizations of X_k are informative to different levels, the total "information" we receive follows a random process. Therefore, instead of fixing the sample size n, we can make n a stopping time τ, which gives a "better" (E_0, E_1) tradeoff.



Sample size are now in expectation, define:

$$l_0 = \mathbb{E}_{\mathbb{P}}[\tau], \qquad l_1 = \mathbb{E}_{\mathbb{Q}}[\tau]$$

Sequential test Z: at each step declair either "H_0" or "H_1" or "give me one more sample".
Note: Put it in σ-algebra language, let $\mathcal{F}_n = \sigma(X_k : k = 1, \ldots, n)$ be a filtration on $(\Omega, \mathcal{F}_\infty)$, and for a stopping time τ, i.e. $\{\tau \leq n\} \in \mathcal{F}_n$, we denote the stopping time σ-algebra $\mathcal{F}_\tau = \sigma(\{E : E \cap \{\tau \leq n\} \in \mathcal{F}_n\})$. The sequential test Z is "sequential" in the sense that it should be adaptive to $\mathcal{F}_\tau$, namely $Z \in \mathcal{F}_\tau$.

Theorem 12.3. *Assume that* [1]

$$\left| log \frac{P(x)}{Q(x)} \right| \leq c_0, \forall x$$

where c_0 is some positive constant. If the error probabilities satisfy:

$$\pi_{1|0} \leq 2^{-l_0 E_0}, \qquad \pi_{0|1} \leq 2^{-l_1 E_1}$$

for large l_0, l_1, then the following inequality for the exponents holds

$$E_0 E_1 \leq D(P\|Q) D(Q\|P).$$

with optimal boundary achieved by the sequential probability ratio test $SPRT(A, B)$ (A, B are large positive numbers) defined as follows:

$$\tau = \inf\{n : S_n \geq B \ \ or \ S_n \leq -A\}$$
$$Z = \begin{cases} 0, & if \ S_\tau \geq B \\ 1, & if \ S_\tau < -A \end{cases}$$

where

$$S_n = \sum_{k=1}^{n} \log \frac{P(X_k)}{Q(X_k)}$$

is the log likelihood function of the first k observations.

Proof. **achievability**

$$\begin{aligned}
\pi_{1|0} &= \mathbb{P}(S_\tau \leq -A) \quad (\mathbb{P}\big|_{\mathcal{F}_\tau} \text{ is the probability measure w.r.t } \mathcal{F}_\tau) \\
&= \int_{X^\infty} \mathbf{1}[S_\tau \leq -A] d\mathbb{P}\big|_{\mathcal{F}_\tau} \\
&= \int_{X^\infty} d\mathbb{Q} \mathbf{1}[S_\tau \leq -A] \frac{d\mathbb{P}}{d\mathbb{Q}}\Big|_{\mathcal{F}_\tau} \\
&= \int_{X^\infty} d\mathbb{Q} \mathbf{1}[S_\tau \leq -A] e^{S_\tau} \quad \text{(notice that though by definition } \frac{d\mathbb{P}}{d\mathbb{Q}}\Big|_{\mathcal{F}_n} = e^{S_n}\text{, replacing } n \text{ by } \tau \text{ is not trivial)} \\
&\leq e^{-A}
\end{aligned}$$

Denote $\tau_0 = \inf\{n : S_n \geq B\}$, we have

$$S_{\tau_0} \leq B + c_0$$

[1] This assumption is satisfied for discrete distributions on finite spaces.

where c_0 is due to possible jumps in distribution.

Notice that $\{S_n - nD(P\|Q) : n \geq 1\}$ is a martingale with respect to filtration $\{\mathcal{F}_n : n \geq 1\}$ and measure $\mathbb{P}$. By Optional stopping time theorem,

$$\mathbb{E}_{\mathbb{P}}[S_{\tau_0} - \tau_0 D(P\|Q)] = 0$$

Since $\mathbb{E}_{\mathbb{P}}[S_{\tau_0}] \leq B + c_0$, we have

$$l_0 = \mathbb{E}_{\mathbb{P}}[\tau] \leq \mathbb{E}_{\mathbb{P}}[\tau_0] \leq \frac{B + c_0}{D(P\|Q)} \approx \frac{B}{D(P\|Q)} \text{ for large } B$$

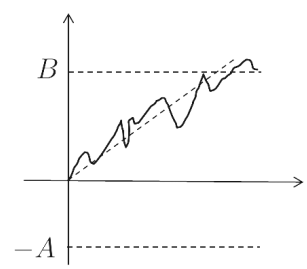
Similarly we can show $\pi_{0|1} \leq e^{-B}$ and $l_1 \leq \frac{A}{D(Q\|P)}$ for large A. Take $B = l_0 D(P\|Q), A = l_1 D(Q\|P)$, this shows the achievability.

converse: Assume (E_0, E_1) achievable for large l_0, l_1,
apply data processing inequality of divergence:

$$\begin{aligned} d(\mathbb{P}(Z=1)\|\mathbb{Q}(Z=1)) &\leq D(\mathbb{P}\|\mathbb{Q})\big|_{\mathcal{F}_\tau} \\ &= \mathbb{E}_{\mathbb{P}}[\tau] D(P\|Q) \quad \text{(optional stopping time)} \\ &= l_0 D(P\|Q) \end{aligned}$$

notice that for $l_0 E_0$ and $l_1 E_1$ large enough, we have the approximation $d(\mathbb{P}(Z=1)\|\mathbb{Q}(Z=1)) \approx l_1 E_1$, therefore $l_1 E_1 \leq l_0 D(P\|Q)$. Similarly we can show that $l_0 E_0 \leq l_1 D(Q\|P)$, finally we have

$$E_0 E_1 \leq D(P\|Q) D(Q\|P), \text{ as } l_0, l_1 \to \infty$$



under P, $S_n - nD(P\|Q)$ is a martingale

$\square$

6.441: Information Theory	Spring 2013
Lecturer: Yury Polyanskiy	**Lecture 13: Channel Coding**

Objects of study so far:

1. P_X - Single distribution, Compression
2. P_X vs Q_X - Comparing two distributions, Hypothesis testing
3. Now: $P_{Y|X} : \mathcal{X} \to \mathcal{Y}$ (called a *random transformation*) - Infinitely many distributions

13.1 Channel Coding

Definition 13.1. *An M-code for $P_{Y|X}$ is an encoder/decoder pair (f, g) of (randomized) functions*

- *encoder* $f : \{1, \ldots, m\} \to \mathcal{X}$
- *decoder* $g : \mathcal{Y} \to \{1, \ldots, m\} \cup \{e\}$

Notation: We'll often use $[m]$ to denote the set $\{1, \ldots, m\}$.

In most cases f and g are deterministic functions, in which case we think of them in terms of codewords, codebooks, and decoding regions

- $\forall i \in [m] : c_i = f(i)$ are *codewords*, the collection $\mathcal{C} = \{c_1, \ldots, c_m\}$ is called a *codebook*.
- $\forall i \in [m], D_i = g^{-1}(\{i\})$ is the *decoding region* for i.

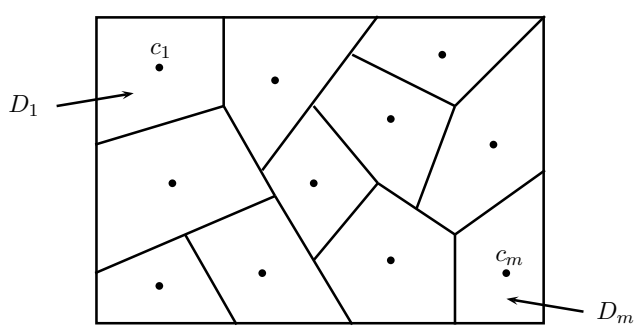


Figure 1: When $\mathcal{X} = \mathcal{Y}$, the decoding regions can be pictured as a partition of the space, each containing one codeword.

Typed by Austin Collins, austinc@mit.edu, April 2, 2013

Remark: The underlying probability space for channel coding problems will always be

$$W \xrightarrow{f} X \xrightarrow{P_{Y|X}} Y \xrightarrow{g} \hat{W}$$

Rigorously, when the source alphabet is $[M]$:

$$\text{(general) } P_{WXY\hat{W}}(m, a, b, \hat{m}) = \frac{1}{M} f(a|m) P_{Y|X}(b|a) g(\hat{m}|b)$$
$$\text{(deterministic) } P_{WXY\hat{W}}(c_m, b, \hat{m}) = \frac{1}{M} P_{Y|X}(b|c_m) \mathbf{1}\{b \in D_{\hat{m}}\}$$

Throughout the notes, these quantities will be called:

- W - Original message
- X - (Induced) Channel input
- Y - Channel output
- $\hat{W}$ - Message estimate

13.1.1 Performance Metrics

To judge the quality of a code, we have three forms for the probability of an error:

1. $P_{e,avg} \triangleq \mathbb{P}[W \neq \hat{W}]$ - Average probability of (block) error.
2. $P_{e,max} \triangleq \max_{1 \leq m \leq M} \mathbb{P}[\hat{W} \neq m | W = m]$ - Maximum probability of error.
3. In the special case when $M = 2^k$, think of $W = S^k \in \mathbb{F}_2^k$ as a length k bit string. Then the bit error is $P_b \triangleq \frac{1}{k} \sum_{j=1}^{k} \mathbb{P}[S_j \neq \hat{S}_j]$

The most typical metric is average probability of error, but the others will be used occasionally in the course as well.

13.1.2 Fundamental Limit of $P_{Y|X}$

Definition 13.2. *A code (f, g) is an $(M, \epsilon)_{avg} - code$ for $P_{Y|X}$ if the source alphabet size is M, and $P_{e,avg} \leq \epsilon$. Similarly, an $(M, \epsilon)_{max}$ code must satisfy $P_{e,max} \leq \epsilon$.*

Then the fundamental limits of channel codes are defined as

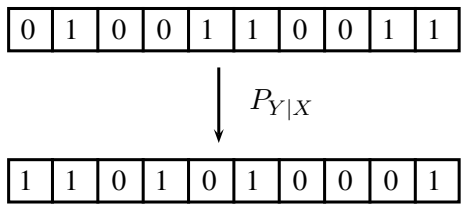
$$M^*_{avg}(\epsilon) = \max\{M : \exists (M, \epsilon)_{avg} - code\}$$
$$M^*_{max}(\epsilon) = \max\{M : \exists (M, \epsilon)_{max} - code\}$$

Remark: $\log_2 M^*$ gives the maximum number of bits that we can pump through a channel $P_{Y|X}$ while still having the error probability (in the appropriate sense) be $\leq \epsilon$.

Example: The random transformation BSC(n, δ) (binary symmetric channel) is defined as

$$\begin{aligned}
&\mathcal{X} = \{0,1\}^n \\
&\mathcal{Y} = \{0,1\}^n \\
&P_{Y|X}(y|x) = \delta^{|y-x|_h}\bar{\delta}^{n-|y-x|_h}
\end{aligned}$$

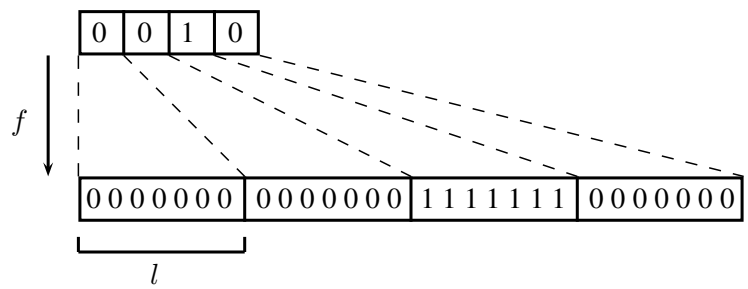
Where $|\cdot|_h$ denotes the Hamming distance, and $\bar{\delta} = 1 - \delta$. Graphically, BSC$(n, \delta)$ takes a binary sequence length n and flips bits iid with probability δ



Question: When $\delta = .11$, $n = 1000$, what is the max number of bits you can send with $P_e \leq 10^{-3}$?

Ideas:

1. Take each bit and repeat it l times (l-repetition code).



   The probability of error of this scheme is $P_e \approx k\mathbb{P}[\text{Binom}(l, \delta) > l/2]$ and $kl \leq n = 1000$, which for $P_e \leq 10^{-3}$ gives $l = 21$, $k = 47$ bits.

2. Reed-Muller Codes $(1, r)$. Interpret a message $a_0, \ldots, a_{r-1} \in \mathbb{F}_2^r$ as the polynomial $\sum_{i=1}^{r-1} a_i x_i + a_0$, then codewords are formed by evaluating the polynomial at $x^{r-1} \in \mathbb{F}_2^{r-1}$. This code has minimum distance 2^{r-2}. For $r = 7$, there is a $[7, 64, 32]$ Reed Muller code. If we use this for the BSC$(1000, \delta)$, we have $P_{e,block} \approx 10^{-4}$ and can send $7 \cdot 16 = 112$ bits in 1000 channel uses, more than double that of the repetition code.

3. Shannon told us that the a BSC(n, δ) has capacity

$$I(X;Y) = n \max_A I(A; A+Z) = n(1 - h(\delta)) \approx \frac{n}{2}$$

   For our running example, this says we should be able to send 500 bits reliably. Turns out, for this blocklength this is not quite possible.

4. Even though calculating $\log M^*$ is not computationally feasible (doubly exponential), we can find bounds on $\log M^*$ that are easy to compute. We will show later in the course that in fact, for BSC$(1000, .11)$

$$414 \leq \log M^* \leq 416$$

5. The first codes to approach the bounds on $\log M^*$ are called *turbo codes* (after the turbocharger engine – where exhaust is fed back in to power the engine). This class of codes is known as *sparse graph codes*, of which LDPC codes are particularly well studied. As a rule of thumb, these codes typically approach $80 \ldots 90\%$ of $\log M^*$ when $n \approx 10^3 \ldots 10^4$.

13.2 Basic Results

Recall that the object of our study is $M^*_{avg}(\epsilon) = \max\{M : \exists (M, \epsilon)_{avg} - code\}$.

13.2.1 Determinism

1. Suppose an encoder $f : [m] \to \mathcal{X}$ is given. Then the decoder that minimizes $P_{e,avg}$ is the MAP decoder

$$g(y) = \operatorname*{argmax}_m \mathbb{P}[W = m | Y = y]$$

Furthermore, for a fixed f, the MAP decoder g is deterministic

2. For a given M, $P_{Y|X}$, the $P_{e,avg}$-minimizing encoder is deterministic

Proof. Let $f : [m] \to \mathcal{X}$ be a random transformation. Instead of $f(a|m)$, consider the deterministic encoder $\tilde{f}(m, u)$, that receives auxiliary randomness u. Then looking at all possible values of the randomness,

$$P_e = P_{U, P_{Y|X}}[W \neq \hat{W}] = \mathbb{E}_u[\mathbb{P}[W \neq \hat{W}|U] = \mathbb{E}_U[P_e(U)]$$

Each u in the expectation gives a deterministic encoder, hence there is a deterministic encoder that is at least as good as the average of the collection, ie $\exists u_0$ s.t. $P_e(u_0) \leq \mathbb{P}[W \neq \hat{W}]$ $\square$

Remark: If instead we use maximal probability of error as our performance criterion, then these results don't hold; randomized encoders and decoders may improve performance.

13.2.2 Bit Error Rate vs Block Error Rate

Now we give a bound on the average probability of error in terms of the bit error probability.

Theorem 13.1. *For all* (f, g), $M = 2^k \implies P_b \leq P_{e,avg} \leq kP_b$

Proof. Recall that $M = 2^k$ gives us the interpretation of $W = S^k$ sequence of bits.

$$\frac{1}{k}\sum_{i=1}^{k} \mathbf{1}\{S_i \neq \hat{S}_i\} \leq \mathbf{1}\{S^k \neq \hat{S}^k\} \leq \sum_{i=1}^{k} \mathbf{1}\{S_i \neq \hat{S}_i\}$$

Where the first inequality is obvious and the second follow from the union bound. Taking expectation of the above expression gives the theorem. □

Remark: We often use the direction $P_b \geq \frac{1}{k}P_e$, although this is a very loose bound for large k.

Theorem 13.2 (Assouad)**.** *Let* $P_{c'|c} = \mathbb{P}[\hat{W} = c'|W = c]$. *Then* $P_b \geq \min_{|c-c'|_h=1} P_{c'|c}$

Proof. Let e_i be a length k vector that is 1 in the i-th position, and zero everywhere else. Then

$$\sum_{i=1}^{k} \mathbf{1}\{S_i \neq \hat{S}_i\} \geq \sum_{i=1}^{k} \mathbf{1}\{S^k = \hat{S}^k + e_i\}$$

Dividing by k and taking expectation gives

$$P_b \geq \frac{1}{k}\sum_{i=1}^{k} \mathbb{P}[S^k = \hat{S}^k + e_i] \geq \min_{|c'-c|_h=1} P_{c'|c}$$

□

Remark: In statistics, Assouad's Lemma gives a lower bound on the minimax risk of an estimator. Say you have any estimator $\hat{\theta}$ for a parameter $\theta \in \mathbb{R}^k$ and want to minimize the maximum of the risk function $\|\theta - \hat{\theta}\|_1$. A lower bound to this maximum error is the average error $\mathbb{E}[\|\theta - \hat{\theta}\|_1]$. We look at a hypercube $\epsilon \cdot \mathbb{F}_2^k$ with side lengths ϵ embedded in the space of parameters. Then our error can be lower bounded by (using Assouad's Lemma)

$$\mathbb{E}[\|\theta - \hat{\theta}\|_1] \geq \mathbb{E}[\|\theta - \hat{\theta}_{dis}\|_1] = \epsilon\mathbb{E}[|\theta - \hat{\theta}_{dis}|_h] = \epsilon k P_b \geq \epsilon k \min_{c':|c'-c|_h=1} P_{c'|c} \geq \frac{\epsilon k}{2}\min_{\theta,\theta'}(1 - TV(P_\theta, P_{\theta'}))\,,$$

where θ, θ' belong to endpoints of a hypercube, $\hat{\theta}_{dis}$ is a discretized version of the estimate, i.e. estimator that chooses the closest point on the hypercube to $\hat{\theta}$. The first inequality above can be seen by lower bounding $|a - b|$ by a step function that is 0 when $|a| \leq b/2$ and 1 elsewhere.

13.3 General (Weak) Converse Bounds

Theorem 13.3.

1. *Any M-code for $P_{Y|X}$ satisfies*

$$\log M \leq \frac{\sup_X I(X;Y) + h(P_{e,avg})}{1 - P_{e,avg}}$$

2. *When* $M = 2^k$

$$\log M \leq \frac{\sup_X I(X;Y)}{\log 2 - h(P_b)}$$

Proof. **(1)** Since $W \to X \to Y \to \hat{W}$, we can form a sequence of inequalities

$$\begin{aligned}\sup_X I(X;Y) \geq I(X;Y) &\geq I(W;\hat{W}) \\ &\geq d(\mathbb{P}[W = \hat{W}] || \frac{1}{M}) \\ &\geq -h(\mathbb{P}[W \neq \hat{W}]) + \mathbb{P}[W = \hat{W}] \log M\end{aligned}$$

Plugging in $P_{e,avg} = \mathbb{P}[W \neq \hat{W}]$ finishes the first proof.

(2) Now $S^k \to X \to Y \to \hat{S}^k$. Recall from Lecture 5 that for an iid, $\sum I(S_i; \hat{S}_i) \leq I(S^k; \hat{S}^k)$. This gives us

$$\begin{aligned}\sup_X I(X;Y) \geq I(X;Y) &\geq \sum_{i=1}^k I(S_i, \hat{S}_i) \\ &\geq k\frac{1}{k}\sum d(\mathbb{P}[S_i = \hat{S}_i] || \frac{1}{2}) \\ &\geq kd(\frac{1}{k}\sum_{i=1}^k \mathbb{P}[S_i = \hat{S}_i] || \frac{1}{2}) \\ &= kd(1 - P_b || \frac{1}{2}) = k(\log 2 - h(P_b))\end{aligned}$$

Where the third line used the convexity of divergence. □

13.4 General achievability bounds

Remark: Regarding differences between information theory and statistics: in statistics, there is a parametrized set of distributions on a space (determined by the model) from which we try to estimate the underlying distribution from samples. In information theory, the challenge is to *choose* the structure on the parameter space (channel coding) such that, upon observing a sample, we can estimate the correct parameter with high probability. With this in mind, it is natural to view

$$\log \frac{P_{Y|X=x}}{P_Y}$$

As an LLR of a binary hypothesis test, where we compare the hypothesis $X = x$ to the distribution induced by our codebook: $P_Y = P_{Y|X} \circ P_X$ (so compare c_i to "everything else"). To decode, we ask M different questions of this form. This motivates the following definition.

Definition 13.3. *For all* P_X*, the information density is defined as*

$$i(a,b) = \log \frac{P_{Y|X}(b|a)}{P_Y(b)}$$

Where $P_Y = P_{Y|X} \circ P_X$.

Remark: $\mathbb{E}[i(X,Y)] = I(X;Y)$.

The next theorem, called Shannon's Random Coding Theorem, is of utmost importance.

Theorem 13.4. $\forall P_X$, $\forall \tau$, $\exists (M,\epsilon)_{avg} - code$ *with*

$$\epsilon \leq \mathbb{P}[i(X,Y) \leq \log M + \tau] + e^{-\tau}$$

Proof in the next lecture.

Notation: in the following proofs, we shall make use of the *independent pairs* $(X, Y) \perp (\overline{X}, \overline{Y})$

$$\begin{aligned} X &\to Y \quad (X: \text{ sent codeword}) \\ \overline{X} &\to \overline{Y} \quad (\overline{X}: \text{ unsent codeword}) \end{aligned}$$

The joint distribution is given by:

$$P_{XY\overline{XY}}(a, b, \overline{a}, \overline{b}) = P_X(a)P_{Y|X}(b|a)P_X(\overline{a})P_{Y|X}(\overline{b}|\overline{a}).$$

14.1 Information density

Definition 14.1 (Information density)**.**

$$i(a, b) = \log \frac{P_{Y|X}(b|a)}{P_Y(b)} = \log \frac{dP_{Y|X=a}(b)}{dP_Y(b)} \tag{1}$$

and we define $i(x, y) = +\infty$ for all y in the singular set where $P_{Y|X=x}$ is not absolutely continuous w.r.t. P_Y. We also define $i(x, y) = -\infty$ for all y such that $dP_{Y|X=x}/dP_Y$ equals zero.
Notice that $i(a, b)$ depends on the underlying P_X and $P_{Y|X}$, which should be understood from the context.

Note: [Intuition] Information density is a useful concept in understanding decoding. In discriminating between two codewords, one concerns with (as we learned in binary hypothesis testing) the LLR, $\log \frac{P_{Y|X=c_1}}{P_{Y|X=c_2}}$. In M-ary hypothesis testing, a similar role is played by information density $i(c_1, y)$, which loosely speaking evaluates the likelihood of c_1 against the average noise, or "everything else", which we model by P_Y.

Note: [Joint measurability] There is a measure-theoretic subtlety in (1): the so defined function $i(\cdot, \cdot)$ may not be a measurable function on the product space $\mathcal{X} \times \mathcal{Y}$. This is resolved by the following technical assumption: suppose there exists a σ-finite measure μ on $\mathcal{Y}$ and a measurable function $\rho : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ such that

$$P_{Y|X=a}[E] = \int_E \rho(a, b) d\mu(b)\,,$$

then

$$i(a, b) = \log \frac{\rho(a, b)}{\int_{\mathcal{X}} dP_X(x)\rho(x, b)}$$

is jointly measurable in (a, b).

Typed by Qingqing Huang (qqh@mit.edu), Apr. 4, 2013

Note: [Alternative definition] Observe that for discrete $\mathcal{X}, \mathcal{Y}$, (1) is equivalently written as

$$i(a,b) = \log \frac{P_{X,Y}(a,b)}{P_X(a)P_Y(b)} = \log \frac{P_{X|Y}(a|b)}{P_X(a)}$$

For the continuous case, people often use the alternative definition, which is symmetric in X and Y and is measurable w.r.t. $\mathcal{X} \times \mathcal{Y}$:

$$i(a,b) = \log \frac{dP_{X,Y}}{dP_X \times P_Y}(a,b) \tag{2}$$

Notice a subtle difference between (1) and (2) for the continuous case: In (2) the Radon-Nikodym derivative is only defined upto sets of measure zero, therefore whenever $P_X(a) = 0$ the value of $P_Y(i(a,Y) > t)$ is undefined. This problem does not occur with definition (1). In any case, for discrete $\mathcal{X}$, $\mathcal{Y}$, or under other regularity conditions, all the definitions are equivalent.

Proposition 14.1 (Properties of information density)**.**

1. *(conditioning and unconditioning trick)*
   Let $f(x,y) = 0$ whenever $i(x,y) = -\infty$, then

$$\mathbb{E}[f(Y)] = \mathbb{E}[e^{-i(a,Y)} f(Y) \Big| X = a], \forall a \tag{3}$$

$$\mathbb{E}[f(X)] = \mathbb{E}[e^{-i(X,b)} f(X) \Big| Y = b], \forall b \tag{4}$$

2.

$$\mathbb{E}[f(\overline{X}, Y)] = \mathbb{E}[e^{-i(X,Y)} f(X,Y)] \tag{5}$$

$$\mathbb{E}[f(X, \overline{Y})] = \mathbb{E}[e^{-i(X,Y)} f(X,Y)] \tag{6}$$

Proof. 2.3.:

$$\mathbb{E}f(Y) = \sum_{y \in \mathcal{Y}} P_Y(y) f(y) = \sum_{y \in \mathcal{Y}} P_{Y|X}(y|a) \frac{P_Y(y)}{P_{Y|X}(y|a)} f(y)$$

notice that by the assumption on $f(\cdot)$, the summation is valid even if for some a we have that $P_{Y|X}(y|a) = 0$. The rest are by symmetry. □

Corollary 14.1.

$$P(i(a,Y) > t) \leq e^{-t} \tag{7}$$

$$P(i(\overline{X}, Y) > t) \leq e^{-t} \tag{8}$$

Proof. pick $f(Y) = \mathbf{1}[i(a,Y) > t]$ in (3). □

(3) holds when $i(x,y)$ is defined as $i = \log \frac{dP_{Y|X}}{P_Y}$, and (4) holds when $i(x,y)$ is defined as $i = \log \frac{dP_{X|Y}}{P_X}$. (5) and (6) hold under either of the definitions. Since in the following we shall only make use of (3) and (5), we adopt the definition in (1).

14.2 Shannon's achievability bound

Theorem 14.1 (Shannon's achievability bound)**.**
For a given $P_{Y|X}$, $\forall P_X$, $\forall \tau > 0$, $\exists (M, \epsilon)_{avg}$ code with

$$\epsilon \leq P(i(X,Y)) \leq log(M+\tau) + e^{-\tau} \tag{9}$$

Proof. Recall that for a given codebook $\{c_1, \ldots, c_M\}$, the optimal decoder is MAP / ML as follows:

$$g^*(y) = \arg\max_{m \in [M]} P_{Y|X}(y|c_m) = \arg\max_m i(c_m, y).$$

The step of selecting the maximum likelihood can make evaluating the error probability difficult. The magic in showing the following two ahievability bounds is to consider a suboptimal decoder. In Shannon's bound, we consider a threshold-based suboptimal decoder $g(y)$ as follows:

$$g(y) = \begin{cases} m, & \exists! c_m \text{ s.t. } i(c_m, y) \geq M + \tau \\ e, & o/w \end{cases}$$

For a given codebook $(c_1, \ldots, c_M)$, the error probability is:

$$P_e(c_1, \ldots, c_M) = P(\{i(c_W, Y) \leq \log M + \tau\} \cup \{\exists \overline{m} \neq W, i(c_{\overline{m}}, Y) > \log M + \tau\})$$

where W is a uniform distribution on $[M]$.
We generate the codebook $(c_1, \ldots, c_M)$ randomly with $c_m \sim P_X$ i.i.d. for $m \in [M]$. By symmetry, the error probability averaging over all possible codebooks is given by:

$$\begin{aligned}
\mathbb{E}[P_e(c_1, \ldots, c_M)] &= \mathbb{E}[P_e(c_1, \ldots, c_M)|W = 1] \\
&= P_{W=1}(\{i(c_1, Y) \leq \log M + \tau\} \cup \{\exists \overline{m} \neq 1, i(c_{\overline{m}}, Y) > \log M + \tau\}) \\
&\leq P_{W=1}(\{i(c_1, Y) \leq \log M + \tau\}) + \sum_{m=2}^{M} P(i(c_m, Y) > \log M + \tau\}) \quad \text{(union bound)} \\
&\leq P_{W=1}(\{i(c_1, Y) \leq \log M + \tau\}) + (M-1)P(i(\overline{X}, Y) > \log M + \tau\}) \quad \text{(random codebook)} \\
&\leq P_{W=1}(\{i(c_1, Y) \leq \log M + \tau\}) + (M-1)e^{-(\log M + \tau)} \quad \text{(by Corollary 14.1)} \\
&\leq P_{W=1}(\{i(c_1, Y) \leq \log M + \tau\}) + e^{-\tau}
\end{aligned}$$

Finally, since the average error probability satisfies the upper bound, there must exists some code allocation that leads to error probability that is smaller or equal to the bound.

□

Note: [Typicality constraints]

- The property of a pair (x, y) encoded by the condition $\{i(x, y) \geq \gamma\}$ is called "joint typicality". Such version of joint typicality is useful when random coding is done in product spaces with $c_j \sim P_X^n$ (i.e. coordinates of the codeword are iid).
- A popular alternative to the definition of typicality is to require that $\hat{P}_{x^n, y^n} \approx P_{XY}$, where

  $$\hat{P}_{x^n, y^n}(a, b) = \frac{1}{n} \cdot \#\{j : x_j = a, y_j = b\}.$$

  This definition is natural for cases when random coding is done with $c_j \sim$ uniform on the set $\{x^n : \hat{P}_{x^n} \approx P_X\}$ (type class).

14.3 Dependence-testing bound

Theorem 14.2 (DT bound)**.** *$\forall P_X$, $\exists (M,\epsilon)_{avg}$ code with*

$$\epsilon \leq \mathbb{E}[exp(-\Big|i(X,Y) - \log\frac{M-1}{2}\Big|^+)] \tag{10}$$

where $|x|^+ = \max(x,0)$.

Proof. For a fixed γ, consider the following suboptimal decoder:

$$g(y) = \begin{cases} m, & \text{for the smallest } m \text{ s.t. } i(c_m,y) \geq \gamma \\ e, & o/w \end{cases}$$

notice that this decoder may give undetected decoding error, and the codewords are not symmetric. For the random generated codebooks, the expected error probability is given by:

$$\begin{aligned}
\mathbb{E}[P_e(c_1,\ldots,c_M)] &= \frac{1}{M}\sum_{j=1}^{M}\mathbb{E}[P_e|W=j] \\
&\leq \frac{1}{M}\sum_{j=1}^{M}\Big(P(i(X,Y)\leq\gamma) + \sum_{k=1}^{j-1}P(i(\overline{X},Y)>\gamma)\Big) \\
&= P(i(X,Y)\leq\gamma) + \frac{M-1}{2}\mathbb{E}[e^{-i(X,Y)}\mathbf{1}[i(X,Y)>\gamma]] \quad \text{(by (3))} \\
&= \mathbb{E}\Big[\mathbf{1}[i(X,Y)\leq\gamma] + \frac{M-1}{2}e^{-i(X,Y)}\mathbf{1}[i(X,Y)>\gamma]\Big] \\
&= \mathbb{E}\Big[\min(1,\frac{M-1}{2}e^{-i(X,Y)})\Big] \quad \text{(Take } \gamma = \log\frac{M-1}{2} \text{ for the minimum upper bound)} \\
&= \mathbb{E}[exp(-\Big|i(X,Y) - \log\frac{M-1}{2}\Big|^+)]
\end{aligned}$$

□

Note: Dependence-testing: the R.H.S. of (10) is equivalent to the minimum error probability in Bayesian hypothesis testing with the setting:

$$\begin{aligned}
&H_0 : X,Y \sim P_{X,Y}; \quad H_1 : X,Y \sim P_X P_Y \\
&\text{prior prob.: } \pi_0 = \frac{2}{M+1}, \pi_1 = \frac{M-1}{M+1}.
\end{aligned}$$

Note: Here we avoid minimizing over τ to get the minimum upperbound in (9) . Moreover, DT bound is stronger than the best Shannon's bound.

14.4 Feinstein's Lemma

Theorem 14.3 (Feinstein's lemma)**.** *$\forall P_X$, $\forall\gamma$, $\forall\epsilon$, $\exists (M,\epsilon)_{max}$ code such that*

$$M \geq \gamma(\epsilon - P(i(X,Y) < \log\gamma)) \tag{11}$$

Proof. The idea is to construct the codebook of size M in a greedy way.
For every $x \in \mathcal{X}$, associate it with a preliminary decode region defined as follows:

$$E_x \equiv \{y : i(x, y) \geq \log \gamma\}$$

Notice that the preliminary decode regions $\{E_x\}$ may be overlapping, and we denote the final decode region partition regions by $\{D_x\}$.
We first claim that there exists some c such that $P[E_c] \geq 1 - \epsilon$. Show by contradiction. Assume that $\forall c \in \mathcal{X}$, $P[i(c, Y) \geq \log \gamma] < 1 - \epsilon$, then pick $c \sim P_X$, we have $P[i(X, Y) \geq \log \gamma] < 1 - \epsilon$, thus $\epsilon < P[i(X, Y) < \log \gamma]$, then the R.H.S. of (11) is negative, and the bound is trivial.
Then we construct the codebook in the following greedy way:

1. pick c_1 to be any codeword such that $P[E_{c_1}] \geq 1 - \epsilon$, and set $D_1 = E_{c_1}$;
2. pick c_2 to be any codeword such that $P[E_x \backslash D_1 | x = c_2] \geq 1 - \epsilon$, and set $D_2 = E_{c_2} \backslash D_1$;
   ...
3. pick c_M to be any codeword such that $P[E_x \backslash \cup_{j=1}^{M-1} D_j | x = c_M] \geq 1 - \epsilon$, and set $D_M = E_{c_M} \backslash \cup_{j=1}^{M-1} D_j$. M is determined by the stopping condition:

$$\forall x_0 \in \mathcal{X}, P[E_{x_0} \backslash \cup_{j=1}^{M} D_j | x = x_0] < 1 - \epsilon$$

averaging over $x_0 \sim P_X$, the stopping condition gives that

$$P(\{i(X, Y) \geq \log \gamma\} \backslash \cup_{j=1}^{M} D_j) < 1 - \epsilon$$

by union bound $P(A \backslash B) \geq P(A) - P(B)$, we have

$$P(i(X, Y) \geq \log \gamma) - \sum_{j=1}^{M} P_Y(D_j) < 1 - \epsilon$$
$$\Rightarrow P(i(X, Y) \geq \log \gamma) - \frac{M}{\gamma} < 1 - \epsilon$$

where the last step makes use of the following key observation:

$$P_Y(D_j) \leq P_Y(E_j) = P_Y(i(c_j, Y) \geq \log \gamma) < \frac{1}{\gamma}, \quad \text{(by Corollary 14.1).}$$

Note: (comparison with Shannon's bound) We can also interpret (11) as for fixed M, there exists a $(M, \epsilon)_{max}$ code that achieves the error probability bounded as follows:

$$\epsilon \leq P(i(X, Y) < \log \gamma) + \frac{M}{\gamma}$$

Take $\gamma = \log M + \tau$, this gives the bound of exactly the same form in (9). However, the two are proved in seemingly quite different ways: Shannon's bound is by random coding, while Feinstein's bound is by greedily selecting the codewords.
Nevertheless, Feinstein's bound is stronger in the sense that it concerns about the max error probability instead of the average.

Recall that last time we showed some achievability bounds:

$$\begin{aligned}
\text{Shannon's:}\quad & P_{e,avg} \le P[i(X,Y) \le \log M + \tau] + e^{-\tau} \\
& \Uparrow \\
\text{DT:}\quad & P_{e,avg} \le \mathbb{E}\left[\exp\left\{-\left|i(X,Y) - \log\frac{M-1}{2}\right|^{+}\right\}\right] \\
& \Uparrow \\
\text{RCU :}\quad & P_{e,avg} \le \mathbb{E}\left[\tfrac{M-1}{2} P[(\overline{X},Y) \ge i(X,Y)|X,Y] \wedge 1\right] \qquad a \wedge b = \min(a,b)) \\
\text{Feinstein's:}\quad & P_{e,max} \le P[i(X,Y) \le \log M + \tau] + e^{-\tau}
\end{aligned}$$

This time we shall use a shortcut to prove the above bounds and in which case $P_{e,avg} = P_{e,max}$.

15.1 Linear coding

Definition 15.1 (Linear code)**.** *Let $\mathcal{X} = \mathcal{Y} = \mathbb{F}_q^n$, $M = q^k$. We index the code words by $(c_0, \ldots, c_u, \ldots, c_{[-1,\cdots,-1]})$ where $u \in \mathbb{F}_q^k$. A code $f : \mathbb{F}_q^k \to \mathbb{F}_q^n$ is a linear code if $\forall u \in \mathbb{F}_q^k$, $c_u = uG$, where G is the generating matrix over $\mathbb{F}_q^{k\times n}$.*

Proposition 15.1.

$$\begin{aligned}
& x \in c \triangleq \{c_u : u \in \mathbb{F}_q^k\} \\
\Leftrightarrow & x \in \text{ row span of } G \\
\Leftrightarrow & x \in KerH, \text{ where } H \text{ is some } n \times (n-k) \text{ matrix over } \mathbb{F}_q.
\end{aligned}$$

Example: [Hamming code] The $[7,4,3]_2$ Hamming code over $\mathbb{F}_2$ is a linear code with $G = [I; P]$ and $H = \begin{bmatrix} -P \\ I \end{bmatrix}$. H is called the parity check matrix.

$$G = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{bmatrix}$$

Note: Linear codes are almost always examined with channels of additive noise.

Definition 15.2 (Additive noise)**.** *$P_{Y|X}$ is additive noise over $\mathbb{F}_q^n$ if*

$$P_{Y|X}(y|x) = P_{Z^n}(y-x) \Leftrightarrow Y = X + Z^n \text{ where } P_{Z^n} \in \mathcal{P}(\mathbb{F}_q^n)$$

Typed by Qingqing Huang (qqh@mit.edu), Apr. 9, 2013

Theorem 15.1. *Any $[k,n]_{\mathbb{F}_q}$ linear code over an additive $P_{Y|X}$ has a maximum likelihood decoder $g(\cdot)$ such that:*

1. $g(y) = y - g_{synd}(yH)$
2. $P_{e,max} = P_{e,avg}$

where $g_{synd}()$ is a (randomized) look up table, called "syndrome decoder".

Theorem 15.2. *Let $P_{Y|X}$ be additive noise over $\mathbb{F}_q^n$. $\forall k, \exists$ a linear code $f : \mathbb{F}_q^k \to \mathbb{F}_q^n$ with the error probability:*

$$P_{e,max} = P_{e,avg} \leq \mathbb{E}\Big[q^{-\left|n-k-\log_q \frac{1}{P_{Z^n}(z^n)}\right|^+}\Big] \tag{1}$$

Proof. Recall that in proving the Shannon's achievability bounds, we select the code words $c_1, \ldots, c_m$ i.i.d $\sim P_X$ and showed that

$$\mathbb{E}[P_e(c_1, \ldots, c_m)] \leq P[i(X,Y) \leq \gamma] + \frac{M-1}{2} P(i(\overline{X}, Y) \geq \gamma)$$

Here we will adopt a similar approach.

Let's first do a quick check of the capacity achieving input distribution for $P_{Y|X}$ with additive noise over $\mathbb{F}_q^n$:

$$\max_{P_X} I(X;Y) = \max_{P_X} H(Y) - H(Y|X) = \max_{P_X} H(Y) - H(Z^n) \Rightarrow P_X^* \text{ uniform on } \mathbb{F}_q^n$$

We shall use the uniform distribution P_X in the "random coding" trick.

Moreover, the maxium likelihood decoder for linear code with uniform input has all decoding regions translational invariant, namely $D_u = c_u + D_0, \forall u$, and therefore:

$$P_{e,max} = P_{e,avg} = P[\hat{W} \neq j | W = j], \forall j.$$

Step 1: Random linear coding with dithering:

$$\forall u \in \mathbb{F}_q^k, c_u = uG + h$$

G and h are drawn from the new emsemble, where the $k \times n$ entries of G and the $1 \times n$ entries of h are i.i.d. uniform over $\mathbb{F}_q$.

Step 2: Construct the independent and uniform distributed pair: $(c_u, c_{u'}) \sim (X, \overline{X})$, such that $P_{X,\overline{X}}(x, \overline{x}) = 1/q^{2n}$.

$$c_u \sim \text{uniform on } \mathbb{F}_q^n$$
$$c_u = uG + h = u'G + h + (u - u')G = c_{u'} + (u - u')G$$

We claim that $c_u \perp G$ because conditional on the generating matrix $G = G_0$, $c_u \sim$ uniform on $\mathbb{F}_q^n$ due to the dithering h.
We also claim that $c_u \perp c_{u'}$ because conditional on $c_{u'}$, $(u - u')G \sim$ uniform on $\mathbb{F}_q^n$.
Thus Random linear coding with dithering indeed gives codewords $c_u, c_{u'}$ pairwise independent and are uniformly distributed.

Step 3: Repeat the same argument in proving DT bound for the symmetric and pairwise independent codewords, we have

$$\frac{1}{M}\sum_j \mathbb{E}[P_e|W=j] \le P[i(X,Y)\le\gamma] + \frac{M-1}{2}P(i(\overline{X},Y)\ge\gamma)$$

$$\Rightarrow P_{e,avg} \le \mathbb{E}[e^{-\left|i(X,Y)-\log_q\frac{q^k-1}{2}\right|^+}] \le \mathbb{E}[e^{-\left|i(X,Y)-k\right|^+}]$$

Step 4: compute $i(X,Y)$:

$$i(a,b) = \log_q \frac{P_{Z^n}(b-a)}{q^{-n}} = n - \log_q \frac{1}{P_{Z^n}(b-a)}$$

therefore

$$P_{e,avg} \le \mathbb{E}[e^{-\left|-\log_q\frac{1}{P_{Z^n}(z^n)}+n-k\right|^+}] \tag{2}$$

Step 5: Kill h. We claim that there exists a linear code without dithering such that (2) is satisfied. We modify the coding scheme with G, h which achieves the bound in the following way: modify the decoder input $Y' = Y - h$, then when c_u is sent, the additive noise $P_{Y'|X}$ becomes then $Y' = uG + h + Z^n - h = uG + Z^n$, which is equivalent to that the linear code generated by G is used.

□

Note: An interesting research problem: The ensemble $c_u = uG + h$ has the pairwise independence property. The joint entropy $H(c_1,\ldots,c_M) = (nk+n)\log q$ is significantly smaller than the "fully random" ensemble we used in the previous lecture. Recall that in that ensemble each c_j was selected independently uniform over $\mathbb{F}_q^n$, implying $H(c_1,\ldots,c_M) = q^k n \log q$. Question:

$$\min H(c_1,\ldots,c_M) = ??$$

where minimum is over all distributions with $P[c_i = a, c_j = b] = q^{-2n}$ when $i \ne j$ (pairwise independent, uniform codewords).

Note: With some non vanishing probability G may fail to be full rank. In such a case, there are two identical codewords and hence $P_{e,max} \ge 1/2$. There are two alternative ensembles of codes which do not contain such degenerate codebooks:

1. $G \sim$ uniform on all full rank matrices
2. search codeword $c_u \in \{c : cH = 0\}$ where $H \sim$ uniform on all $n \times (n-k)$ full row rank matrices. (random parity check construction)

Analysis of random coding over such ensemble is similar, except that this time $(X, \bar{X})$ have distribution

$$P_{X,\bar{X}} = \frac{1}{q^{2n}} 1\{X \ne X'\}$$

uniform on all pairs of *distinct* codewords.

15.2 Channel capacity

We shall formalize the concept of a channel and channel capacity.

Definition 15.3 (Channel)**.** *A channel is specified by:*

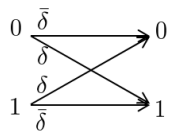
- *input alphabet* $\mathcal{A}$
- *output alphabet* $\mathcal{B}$
- *a sequence of random transformation kernels* $P_{Y^n|X^n} : \mathcal{A}^n \to \mathcal{B}^n, n = 1, 2, \ldots$.

Definition 15.4 (Channel characteristics)**.**

- *A channel is discrete if* $\mathcal{A}$ *and* $\mathcal{B}$ *are finite.*
- *A channel is an additive noise if* $\mathcal{A} = \mathcal{B}$ *are abelian group, and*
$$P_{Y^n|X^n} = P_{Z^n}(Y^n - X^n) \Leftrightarrow Y^n = X^n + Z^n.$$
- *A channel is memoryless if* $P_{Y^n|X^n} = \prod_{k=1}^n P_{Y_k|X_k}$.
- *A channel is stationary memoryless if* $P_{Y^n|X^n} = \prod_{k=1}^n P_{Y_1|X_1}$.
- **DMC** *(discrete memoryless stationary channel)*
A DMC can be specified in two ways: a $|\mathcal{A}| \times |\mathcal{B}|$*-dimensional matrix* $P_{Y|X}$ *where elements specify the transition probabilities, or a biparttite graph with link weight specifying the transition probabilities.*

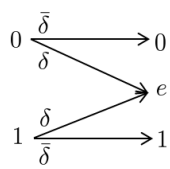
  Example*:*

BSC



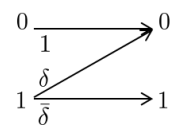
$$\begin{bmatrix} \bar{\delta} & \delta \\ \delta & \bar{\delta} \end{bmatrix}$$

BEC



$$\begin{bmatrix} \bar{\delta} & \delta & 0 \\ 0 & \delta & \bar{\delta} \end{bmatrix}$$

Z channel



$$\begin{bmatrix} 1 & 0 \\ \delta & \bar{\delta} \end{bmatrix}$$

Definition 15.5 (Channel capacity and fundamental limit)**.**
For any channel, define the fundamental limits:

$$M^*_{avg}(n, \epsilon) = \max\{M : \exists (M, \epsilon)_{avg} \text{ code for } P_{Y^n|X^n}\}$$
$$M^*_{max}(n, \epsilon) = \max\{M : \exists (M, \epsilon)_{max} \text{ code for } P_{Y^n|X^n}\}$$

Define the ϵ*-capacity* C_ϵ *and the Shannon capacity* C *of the channel as:*

$$C_\epsilon = \lim \inf_{n\to\infty} \frac{1}{n} \log M^*(n, \epsilon)$$
$$C = \lim_{\epsilon \to 0^+} C_\epsilon$$

16.1 Channel Capacity

Recall the basic definitions that will be used in this lecture:

Definition 16.1. *A channel is a sequence of random transformations*

$$\{P_{Y^n|X^n} : \mathcal{A}^n \to \mathcal{B}^n, n = 1, 2, \ldots\}$$

Definition 16.2. *An $(M, \epsilon)_{avg/max}$ code for the n-th random transformation $P_{Y^n|X^n}$ is called an $(n, M, \epsilon)_{avg/max}$ code. The parameter n is called the* blocklength.

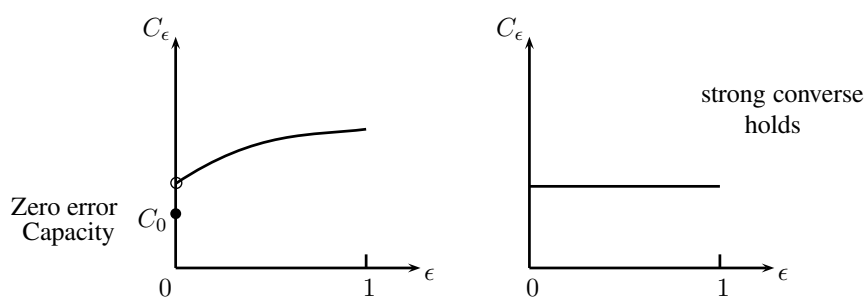
Definition 16.3 (Fundamental Limits)**.** *For any channel*

$$M^*(n, \epsilon) = \max\{M : \exists\ (n, M, \epsilon)_{avg}\ code\}$$
$$M^*_{max}(n, \epsilon) = \max\{M : \exists\ (n, M, \epsilon)_{max}\ code\}$$

Definition 16.4. *The ϵ-Capacity C_ϵ and Shannon Capacity C are*

$$C_\epsilon \triangleq \liminf_{n\to\infty} \frac{1}{n} \log M^*(n, \epsilon)$$
$$C = \lim_{\epsilon\to 0^+} C_\epsilon$$

This operational definition of the capacity represents maximum rate at which one can communicate through a channel with probability of error less than ϵ. Generally, the ϵ-capacity looks like the plot on the LHS, where C_0 is called the *zero error capacity*, which represents the maximal achievable rate with no error (this is often zero). If C_ϵ is constant for all ϵ (RHS plot), then we say that the strong converse holds (more on this later).



Typed by Austin Collins, austinc@mit.edu, April 11, 2013

Theorem 16.1. *Equivalent formulations of C_ϵ and C are*

$$C_\epsilon = \sup\{R : \forall \delta > 0, \exists n_0(\delta), \forall n \geq n_0(\delta), \exists (n, 2^{n(R-\delta)}, \epsilon)_{avg} \text{ code}\}$$
$$C = \sup\{R : \forall \epsilon > 0, \forall \delta > 0, \exists n_0(\delta, \epsilon), \forall n \geq n_0(\delta, \epsilon), \exists (n, 2^{n(R-\delta)}, \epsilon)_{avg} \text{ code}\}$$

Proof. This trivially follows from applying the definitions of $M^*(n, \epsilon)$ (DIY). □

Question: Why do we define C_ϵ and C with respect to average probability of error? Why not maximal probability of error? It turns out that these two definitions are equivalent, as the next theorem shows.

Theorem 16.2. $\forall \tau \in (0, 1)$,

$$\tau M^*(n, \epsilon(1-\tau)) \leq M^*_{max}(n, \epsilon) \leq M^*(n, \epsilon)$$

Proof. The second inequality is obvious, since any code that achieves a maximum error probability ϵ also achieves an average error probability of ϵ. For the first inequality, take an $(n, M, \epsilon(1-\tau))_{avg}$ code, and define the error probability for the j-th codeword as

$$\lambda_j \triangleq \mathbb{P}[\hat{W} \neq j | W = j]$$

WLOG assume the codewords are sorted in order of increasing probability of error $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_{M\tau} \leq \cdots \leq \lambda_M$. We lower bound the average probability of error by taking $\lambda_j = 0$ when $j < \tau M$ and $\lambda_j = \lambda_{\tau M}$ otherwise. Since this code has average error $\leq \epsilon(1-\tau)$, we have

$$\epsilon(1-\tau) \geq \frac{1}{M}\sum_{j=1}^{M} \lambda_j \geq \frac{1}{M}\sum_{j=\tau M}^{M} \lambda_j \geq \frac{1}{M}(1-\tau)M\lambda_{\tau M} = (1-\tau)\lambda_{\tau M}$$

Hence $\lambda_1 \leq \cdots \leq \lambda_{\tau M - 1} \leq \lambda_{\tau M} \leq \epsilon$, so we can extract an $(n, \tau M, \epsilon)_{max}$ code. Finally, take $M = M^*(n, \epsilon(1-\tau))$ to finish the proof. □

Corollary 16.1. $C_\epsilon^{(max)} = C_\epsilon$ *and* $C^{(max)} = C$, $\forall \epsilon$ *such that* $C_\epsilon = C_{\epsilon^-}$.

Notation: The notation $f(x-) \triangleq \lim_{y \nearrow x} f(y)$.

Proof. Using the definition of M^* and the previous theorem, for any fixed $\tau > 0$

$$C_\epsilon^{(max)} \geq \liminf_{n\to\infty} \frac{1}{n} \log \tau M^*(n, \epsilon(1-\tau)) \geq C_\epsilon(1-\tau)$$

And the second piece follows since $C_\epsilon \geq C_\epsilon^{(max)} \geq C_{\epsilon^-}$. □

16.1.1 General Bounds on C_ϵ

Now that we have the basic definitions for C_ϵ, we define another type of capacity, and show that for a certain class of channels, the two notions ("operational" and "information" capacity) agree.

Definition 16.5. *The information capacity of a channel is*

$$C_i = \liminf_{n\to\infty} \frac{1}{n} \sup_{P_{X^n}} I(X^n; Y^n)$$

Remark: This is not the same as the shannon capacity, and has no direct interpretation as a quantity related to coding. Rather, it is best to think of this only as taking the n-th random transformation in the channel, maximizing over input distributions, then normalizing, then looking at the limit of this sequence.

Theorem 16.3 (Upper Bound for C_ϵ)**.** $\forall \epsilon \in [0,1)$, $C_\epsilon \leq \frac{C_i}{1-\epsilon}$ *and* $C \leq C_i$

Proof. Recall the general converse bound introduced in Lecture 13:

$$\log M^*(n,\epsilon) \leq \frac{\sup_{P_{X^n}} I(X^n;Y^n) + h(\epsilon)}{1-\epsilon}$$

Normalizing this by n the taking the lim inf gives

$$C_\epsilon = \liminf_{n\to\infty} \frac{1}{n} \log M^*(n,\epsilon) \leq \liminf_{n\to\infty} \frac{1}{n} \frac{\sup_{P_{X^n}} I(X^n;Y^n) + h(\epsilon)}{1-\epsilon} = \frac{C_i}{1-\epsilon}$$

□

Now a lower bound on C_ϵ.

Theorem 16.4 (Lower Bound for C_ϵ)**.** *For a stationary, memoryless channel,* $C_\epsilon \geq C_i$

Corollary 16.2. *For a stationary memoryless channel,* $C = C_i = \sup_{P_X} I(X;Y)$.

The corollary follows from pairing the upper and lower bounds on C_ϵ.

Remark: The above corollary, known as *Shannon's Noisy Channel Theorem*, is perhaps the most significant result in information theory. For communications engineers, the major surprise was that $C > 0$, i.e. communication over a channel is possible with strictly positive rate for any arbitrarily small probability of error. This result influenced the evolution of communication systems to block architectures that used bits as a universal currency for data, along with encoding and decoding procedures.

Before giving the proof, we show the second equality in the corollary. Notice that C_i for stationary memoryless channels is easy to compute:

Proposition 16.1. *For a stat. mem-less channel,* $C_i = \sup_{P_X} I(X;Y)$

Proof. Recall that for stat. mem-less channels, $I(X^n;Y^n) \leq \sum_{k=1}^n I(X;Y)$, with equality when X_i's are independent. Then

$$C_i = \liminf_{n\to\infty} \frac{1}{n} \sup_{P_{X^n}} I(X^n;Y^n) = \liminf_{n\to\infty} \sup_{P_X} I(X;Y) = \sup_{P_X} I(X;Y)$$

□

Now we prove the theorem

Proof. $\forall P_X$, and let $P_{X^n} = P_X^n$ (iid product). Recall Shannon's bound:

$$\epsilon_1 \leq \mathbb{P}[i(X^n, Y^n) \leq \log M + \gamma] + \exp(-\gamma) \quad \forall M, \forall \gamma$$

Here the information density is defined as $i(X^n, Y^n) = \log \frac{dP_{Y^n|X^n}}{dP_{Y^n}}(Y^n|X^n) = \sum_{k=1}^n i(X^k, Y^k)$, which is a sum of iid r.v.'s with mean $I(X;Y)$. Set $\log M = n(I(X;Y) - 2\delta)$ for $\delta > 0$, and taking $\gamma = \delta n$ in Shannon's bound

$$\epsilon_1 \leq \mathbb{P}[\sum_{k=1}^n i(X^k, Y^k) \leq nI(X;Y) - \delta n] + \exp(-\delta n) \overset{n\to\infty}{\longrightarrow} 0$$

The second terms goes to zero since $\delta > 0$, and the first terms goes to zero by WLLN.

Therefore, $\forall P_X$, $\forall \delta > 0$, there exists a sequence of $(n, M_n, \epsilon_1)_{max}$ codes with $\epsilon_1 \to 0$ (where $\log M_n = n(I(X;Y) - 2\delta)$). Hence, for all n such that $\epsilon_1 \leq \epsilon$

$$\log M^*(n, \epsilon) \geq n(I(X;Y) - 2\delta)$$

And so

$$C_\epsilon = \liminf_{n\to\infty} \frac{1}{n} M^*(n, \epsilon) \geq I(X;Y) - 2\delta \quad \forall P_X, \forall \delta$$

Since this holds for all P_X and all δ, we conclude $C_\epsilon \geq \sup_{P_X} I(X;Y) = C_i$. □

16.1.2 Bound Using CLT

The main tools in the previous proof was the WLLN. The lower bound $C_\epsilon \geq C_i$ in Theorem 16.4 shows that $\log M^*(n, \epsilon) \geq nC + o(n)$ (since normalizing by n and taking the liminf must result in something $\geq C$). If instead we do a more refined analysis using the CLT, we find

Theorem 16.5. *For any stationary memoryless channel with $C = \max_{P_X} I(X;Y)$ (i.e. $\exists P_X^* = \operatorname{argmax}_{P_X} I(X;Y)$) such that $V = Var[i(X^*, Y^*)] < \infty$, then*

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

Where $Q(\cdot)$ is the complementary Gaussian CDF and $Q^{-1}(\cdot)$ is its functional inverse.

Proof. Recall Feinstein's bound

$$\exists (n, M, \epsilon)_{max}: \quad M \geq \beta\left(\epsilon - \mathbb{P}[i(X^n, Y^n) \leq \log \beta]\right)$$

Take $\log \beta = nC + \sqrt{nV}t$, then applying the CLT gives

$$\begin{aligned}
&\log M \geq nC + \sqrt{nV}t + \log\left(\epsilon - \mathbb{P}\left[\sum i(X_k, Y_k) \leq nC + \sqrt{nV}t\right]\right) \\
&\implies \log M \geq nC + \sqrt{nV}t + \log\left(\epsilon - \Phi(t)\right) \quad \forall \Phi(t) < \epsilon \\
&\implies \frac{\log M - nC}{\sqrt{nV}} \leq t + \frac{\log(\epsilon - \Phi(t))}{\sqrt{nV}}
\end{aligned}$$

Where $\Phi(t)$ is the standard normal CDF. Taking the liminf of both sides

$$\liminf_{n\to\infty} \frac{\log M^*(n,\epsilon) - nC}{\sqrt{nV}} \geq t \quad \forall t \ s.t. \ \Phi(t) < \epsilon$$

Taking $t \nearrow \Phi^{-1}(\epsilon)$, and writing the liminf in little o form completes the proof

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

□

16.1.3 Information Stability

We saw that $C = C_i$ for stationary memoryless channels, but what other channels does this hold for? To answer this question, we introduce the notion of *information stability.*

Definition 16.6. *A channel is called information stable if there exists a sequence of input distribution* $\{P_{X^n}, n = 1, 2, \ldots\}$ *such that*

$$\frac{1}{n} i(X^n, Y^n) \longrightarrow C_i \ \ \textit{in probability}$$

For example $P_{X^n} = (P_X^*)^n$ for stationary memoryless channels. The next theorem gives conditions to check for information stability in memoryless channels.

Theorem 16.6. *A memoryless, non-stationary channel is information stable if either of the following two conditions hold*

$$\exists X_k^* \ \ s.t \ \ \frac{1}{n}\sum_{k=1}^{n} I(X_k^*; Y_k^*) \to C_i \ \ \textit{and} \ \ \sum_{n=1}^{\infty} \frac{1}{n^2} Var[i(X_n^*, Y_n^*)] < \infty$$

Or

$$|\mathcal{A}| < \infty \ \ \textit{or} \ \ |\mathcal{B}| < \infty$$

Proof. The second condition follows from the first since $Var[i(X;Y)] \leq \log^2(\min(|\mathcal{A}|, |\mathcal{B}|))$ and X_k^* such that $\frac{1}{n}\sum I(X_k^*; Y_k^*) \to C_i$ always exists. To show the first condition, it is sufficient to prove

$$\mathbb{P}\left[\frac{1}{n}\left|\sum_{k=1}^{n} i(X_k^*, Y_k^*) - I(X_k^*, Y_k^*)\right| > \delta\right] \to 0$$

So that $\frac{1}{n} i(X^n, Y^n) \to C_i$ in probability. We bound this by Chebyshev's inequality

$$\mathbb{P}\left[\frac{1}{n}\left|\sum_{k=1}^{n} i(X_k^*, Y_k^*) - I(X_k^*, Y_k^*)\right| > \delta\right] \leq \frac{\frac{1}{n^2}\sum_{k=1}^{n} Var[i(X_k^*, Y_k^*)]}{\delta^2} \xrightarrow{?} 0$$

The question mark indicates that we need the following lemma to make this conclusion. □

Lemma 16.1 (Kronecker Lemma)**.** *Let a sequence* $0 < b_n \nearrow \infty$ *and a non-negative sequence* $\{x_n\}$ *such that* $\sum_{n=1}^{\infty} x_n < \infty$*, then*

$$\frac{1}{b_n}\sum_{j=1}^{n} b_j x_j \longrightarrow 0$$

Proof. Since b_n's are strictly increasing, we can split up the summation and bound them from above

$$\sum_{k=1}^{n} b_k x_k \leq b_m \sum_{k=1}^{m} x_k + \sum_{k=m+1}^{n} b_k x_k$$

Now throw in the rest of the x_k's in the summation

$$\begin{aligned}
&\implies \frac{1}{b_n}\sum_{k=1}^{n} b_k x_k \leq \frac{b_m}{b_n}\sum_{k=1}^{\infty} x_k + \sum_{k=m+1}^{n} \frac{b_k}{b_n} x_k \leq \frac{b_m}{b_n}\sum_{k=1}^{\infty} x_k + \sum_{k=m+1}^{\infty} x_k \\
&\implies \lim_{n\to\infty} \frac{1}{b_n}\sum_{k=1}^{n} b_k x_k \leq \sum_{k=m+1}^{\infty} x_k \to 0
\end{aligned}$$

Since this holds for any m, we can make the last term arbitrarily small. □

Example: $X \sim \mathcal{N}(0, \sigma^2)$, $Y = X + Z$ where $Z \sim \mathcal{N}(0, \sigma_Z^2)$, and $Var[i(X,Y)] = \frac{\sigma_X^2}{\sigma_X^2 + \sigma_Y^2}\log^2 e < \log^2 e$, so the AWGN channel is information stable.

The purpose for defining information stability is the following theorem.

Theorem 16.7. *For an information stable channel,* $C = C_i$*.*

Proof. Like the stationary, memoryless case, the upper bound comes from the general converse in Lecture 13, and the lower bound uses a similar strategy as Theorem 16.4, except utilizing the definition of information stability. □

6.441: Information Theory	Spring 2013
Lecturer: Yury Polyanskiy	**Lecture 17: Capacity of Gaussian channels**

17.1 Channels coding with input constraint

Definition 17.1. *An (n, M, ϵ) code satisfies the input constraint $F_n \subset \mathcal{A}^n$ if the encoder $f : [m] \to \mathcal{A}^n$ satisfies $Im\ f \subset F_n$.*

Definition 17.2 (Separable cost constraint)**.** *A channel with separable cost constraint is specified as follows:*

1. *$\mathcal{A}, \mathcal{B}$ - input/output spaces*
2. *$P_{Y^n|X^n} : \mathcal{A}^n \to \mathcal{B}^n$, $n = 1, 2, \ldots$*
3. *Cost $\mathsf{c} : \mathcal{A} \to \bar{\mathbb{R}}$*

Note: We will abuse notation and write

$$\mathsf{c}(x^n) = \frac{1}{n}\sum_{k=1}^{n} \mathsf{c}(x_k)$$

to denote an average per-letter cost of a codeword x^n.

Definition 17.3 (Admissible constraint)**.** *P is an admissible constraint if $\exists x_0 \in \mathcal{A}$ s.t. $\mathsf{c}(x_0) \le P \Leftrightarrow \exists P_X : \mathbb{E}[\mathsf{c}(X)] \le P$. The set of admissible P's is denoted by $\mathcal{D}_{\mathsf{c}}$, and can be either in the form (P_0, ∞) or $[P_0, \infty]$, where $P_0 \triangleq \inf_{x\in\mathcal{A}} \mathsf{c}(x)$.*

Definition 17.4. *Some basic definitions in parallel with the channel capacity without input constraint.*

- *A code is an (n, M, ϵ, P) code ($P \in \mathbb{R}$) if it is an (n, M, ϵ) code satisfying input constraint $F_n \triangleq \{x^n : \frac{1}{n}\sum \mathsf{c}(x_k) \le P\}$*
-

$$\begin{aligned}
M^*(n, \epsilon, P) &= \max\{M : \exists (n, M, \epsilon, P) - code\} \\
M^*_{max}(n, \epsilon, P) &= \max\{M : \exists (n, M, \epsilon, P)_{max} - code\} \\
C_\epsilon(P) &= \liminf_{n\to\infty} \frac{1}{n} \log M^*(n, \epsilon, P) \\
C_i(P) &= \liminf_{n\to\infty} \frac{1}{n} \sup_{P_{X^n}:\mathbb{E}[\sum_{k=1}^n \mathsf{c}(X_k)]\le nP} I(X^n; Y^n)
\end{aligned}$$

Typed by Qingqing Huang (qqh@mit.edu), Apr. 9 , 2013

Note: These are the usual definitions, except that in $C_i(P)$, we are permitted to maximize $I(X^n;Y^n)$ using input distributions from the constraint set $\{P_{X^n} : \mathbb{E}[\sum_{k=1}^n \mathsf{c}(X_k)] \le nP\}$ instead of the distributions with support of F_n.

Proposition 17.1. *Function $f(P) = \sup_{P_X:\mathbb{E}[\mathsf{c}(X)]\le P} I(X;Y)$ is*

1. *Concave, increasing, dom $f = \mathcal{D}_{\mathsf{c}}$*
2. *One of the following is true: $f(P)$ is continuous and finite on (P_0,∞), or $f=\infty$ on (P_0,∞).*

Furthermore, both properties hold for the function $P \mapsto C_i(P)$.

Proof. In (1) all statements are obvious, except for concavity, which follows from the concavity of $P_X \mapsto I(X;Y)$. (2) follows from concavity of $f(\cdot)$.

To extend these results to $C_i(P)$ observe that for every n

$$P \mapsto \frac{1}{n} \sup_{P_{X^n}:\mathbb{E}[\mathsf{c}(X^n)]\le P} I(X^n;Y^n)$$

is concave. Then taking $\liminf_{n\to\infty}$ the same holds for $C_i(P)$. □

Corollary 17.1. *For stationary memoryless channel $C_i(P) = f(P) = \sup_{\mathbb{E}[\mathsf{c}(X)]\le P} I(X;Y)$.*

Proof. $C_i(P) \ge nf(P)$ is obvious and for the reverse, use the concavity of $f(\cdot)$, we have that for any P_{X^n},

$$I(X^n;Y^n) \le \sum_{j=1}^n I(X_j;Y_j) \le \sum_{j=1}^n f(\mathbb{E}[\mathsf{c}(X_j)]) \le nf(\frac{1}{n}\sum_{j=1}^n \mathbb{E}[\mathsf{c}(X_j)]) \le nf(P).$$

□

17.2 Capacity under input constraint $C(P) \stackrel{?}{=} C_i(P)$

Theorem 17.1 (General inverse)**.**

$$C_\epsilon(P) \le \frac{C_i(P)}{1-\epsilon}$$

Proof. Take any (n,M,ϵ,P) code, $W \to X^n \to Y^n \to \hat{W}$. Apply Fano's inequality, we have

$$-h(\epsilon) + (1-\epsilon)\log M \le I(W;\hat{W}) \le I(X^n;Y^n) \le \sup_{P_{X^n}:\mathbb{E}[\mathsf{c}(X^n)]\le P} I(X^n;Y^n) \le nf(P)$$

□

Theorem 17.2 (Extended Feinstein's Lemma)**.** *$\forall P_X, \forall P_{Y|X}, \forall F \in \mathcal{X}, \forall\gamma, \forall M$, there exists a $(M,\epsilon)_{max}$ code with:*

- *encoder satisfies the input constraint:* $f : [M] \mapsto F \in \mathcal{X}$;
- $$\epsilon P_X(F) \le P(i(X,Y) < \log\gamma) + \frac{M}{\gamma}$$

Notice that when $F = \mathcal{X}$, it is consistent with the original Feinstein's Lemma.

Proof. Similar to the proof of the original Feinstein's Lemma, define the preliminary decode regions $E_c = \{y : i(c,y) \ge \log\gamma\}$. Sequentially pick codewords $\{c_1, \dots, c_M\}$ from the set F and the real decode region $\{D_1, \dots, D_M\}$ where $D_j \triangleq E_{c_j} \setminus \cup_{k=1}^{j-1} D_k$. The stopping criterion for M is:

$$\begin{aligned}
&\forall x_0 \in F, P[E_x \setminus \cup_j^M D_j \big| x = x_0] < 1-\epsilon \\
\Leftrightarrow &\forall x_0 \in \mathcal{X}, P[E_x \setminus \cup_j^M D_j \big| x = x_0] < (1-\epsilon)\mathbf{1}[x_0 \in F] + \mathbf{1}[x_0 \in F^c] \\
\Rightarrow &\text{average over } x_0 \sim P_X,\ P(\{i(X,Y) \ge \log\gamma\} \setminus \cup_j^M D_j) \le (1-\epsilon)P_X(F) + P_X(F^c) = 1 - \epsilon P_X(F)
\end{aligned}$$

repeat the rest of proof as in Feinstein's Lemma and we are done.

□

Theorem 17.3 (Achievability)**.** *For any information stable channel with input constraints and $P > P_0$ we have*

$$C_\epsilon(P) \ge C_i(P-) \tag{1}$$

Proof. Let's first focus on stationary memoryless channel, and proof for general information stable channel follows similarly. Thus $P_{Y^n|X^n} = (P_{Y|X})^n$.

Fix $n \ge 1$, choose a P_X such that $\mathbb{E}[\mathsf{c}(X)] < P$, Pick $\log M = n(I(X;Y) - 2\delta)$ and $\log\gamma = n(I(X;Y) - \delta)$.

With the input constraint set $F = \{x^n : \frac{1}{n}\sum \mathsf{c}(x_k) \le P\}$, and distribution $P_{X^n} = P_X^n$, we apply the extended Feinstein's Lemma, there exists a $(n, M, \epsilon_1, P)_{max}$ code with the encoder satisfying input constraint F and the error probability

$$\epsilon_1 P_X(F) \le \underbrace{P(i(X^n, Y^n) \le n(I(X;Y) - \delta))}_{\to 0 \text{ as } n\to\infty \text{ by WLLN and stationary memoryless assumption}} + \underbrace{e^{-n\delta}}_{\to 0}$$

Also, since $\forall k$, $\mathbb{E}[\mathsf{c}(X_k)] < P$, we have $P_X(F) = P(\frac{1}{n}\sum \mathsf{c}(x_k) \le P) \to 1$.

$$\begin{aligned}
&\epsilon_1(1 + o(1)) \le o(1) \\
\Rightarrow &\epsilon_1 \to 0 \text{ as } n \to \infty \\
\Rightarrow &\forall \epsilon, \exists n_0, \text{ s.t. } \forall n \ge n_0, \exists (n, M, \epsilon_1, P)_{max} \text{ code, with } \epsilon_1 \le \epsilon
\end{aligned}$$

Therefore

$$\begin{aligned}
&C_\epsilon(P) \ge \frac{1}{n}\log M = I(X;Y) - 2\delta, \quad \forall \delta > 0, \forall P_X \text{ s.t. } \mathbb{E}[\mathsf{c}(X)] < P \\
\Rightarrow &C_\epsilon(P) \ge \sup_{\{P_X : \mathbb{E}[\mathsf{c}(X)] < P\}} \lim_{\delta\to 0}(I(X;Y) - 2\delta) \\
\Rightarrow &C_\epsilon(P) \ge C_i(P-)
\end{aligned}$$

Notice that for general information stable channel, we just need to use the definition to show that $P(i(X^n, Y^n) \leq n(C_i - \delta)) \to 0$, and all the rest follows.

□

Theorem 17.4 (Shannon capacity)**.** *For an information stable channel with cost constraint, $\forall P > P_0$*

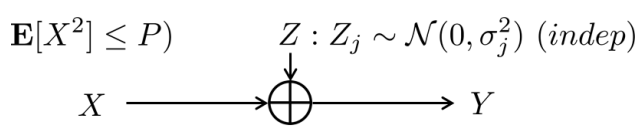
$$C(P) = C_i(P)$$

Proof. Theorem 17.1 shows $C_\epsilon(P) \leq \frac{C_i(P)}{1-\epsilon}$, thus $C(P) \leq C_i(P)$.. Theorem 1 shows $C_\epsilon(P) \geq C_i(P-)$. Also, $C_i(P)$ is continous at $P, \forall P > P_0$, so $C_i(P-) = C_i(P)$. □

Note: in homework, you will show that $C(P_0) = C_i(P_0)$ also holds, even though $C_i(P)$ may be discontinuous at P_0.

17.3 Applications

17.3.1 Stationary AWGN channel



AWGN

Definition 17.5 (AWGN)**.** *A stationary additive Gaussian noise (AWGN) channel is defined as follows: $\mathcal{A} = \mathcal{B} = \mathbb{R}$, $\mathsf{c}(x) = x^2$, $P_{Y^n|X^n} : Y^n = X^n + Z^n$, where $Z^n \sim \mathcal{N}(0, I_n)$.*

Note: complex AWGN channel is similarly defined: $\mathcal{A} = \mathcal{B} = \mathbb{C}$, $\mathsf{c}(x) = |x|^2$, and $Z^n \sim \mathbb{C}\mathcal{N}(0, I_n)$

Theorem 17.5. *for stationary ($\mathbb{C}$)-AWGN channel, the channel capacity is equal to information capacity, and is given by:*

$$C(P) = C_i(P) = \frac{1}{2}\log(1+P) \quad \textit{for AWGN}$$
$$C(P) = C_i(P) = \log(1+P) \quad \textit{for } \mathbb{C}\textit{-AWGN}$$

Proof.

$$C_i = \liminf_n \frac{1}{n} \sup_{\{P_{X^n}:\mathbb{E}[\sum_k \mathsf{c}(X_k)] \leq nP\}} I(X^n; Y^n) = \sup_{\{P_X:\mathbb{E}[\mathsf{c}(X) \leq P]\}} I(X; X+Z)$$

□

17.3.2 Parallel AWGN channel

Definition 17.6 (Parallel AWGN)**.** *A parallel AWGN channel with L branches is defined as follows: $\mathcal{A} = \mathcal{B} = \mathbb{R}^L$; $\mathsf{c}(x) = \sum_{k=1}^L |x_k|^2$; $P_{Y|X} : Y_k = X_k + Z_k$, for $k = 1, \ldots, L$, and $Z_k \sim \mathcal{N}(0, \sigma_k^2)$ are independent across k.*

Theorem 17.6 (Waterfilling)**.** *The capacity of L-parallel AWGN channel is given in the following parametric form: for $T \in [0, +\infty]$, we have:*

$$C = \frac{1}{2}\sum_{j=1}^{L} \log^+ \frac{T}{\sigma_j^2}$$

$$P = \sum_{j=1}^{L} |T - \sigma_j^2|^+$$

where $\log^+(x) = \max(\log x, 0)$.

Proof.

$$\begin{aligned} C_i(P) &= \sup_{\{P_{X_1,\ldots,X_L}:\mathbb{E}[X_1^2]+\cdots+\mathbb{E}[X_L^2]\le P\}} I(Y_1,\ldots,Y_L;Y_1,\ldots,Y_L) \\ &\le \sup_{\sum P_k \le P, P_k \ge 0} \sum_{k=1}^{L} \sup_{\mathbb{E}[X_k^2]\le P_k} I(X_k;Y_k) \\ &\le \sup_{\sum P_k \le P, P_k \ge 0} \sum_{k=1}^{L} \frac{1}{2}\log(1+\frac{P_k}{\sigma_k^2}) \end{aligned}$$

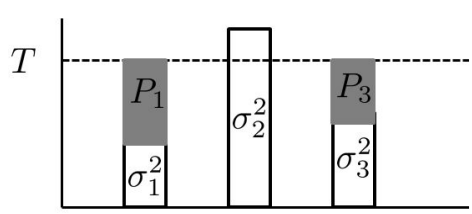
For the last optimization problem above, denote the Lagragian multipliers for the constraint $\sum P_k \le P$ by λ and for the constraint $P_k \ge 0$ by μ_k. First order condition on P_k gives that

$$\frac{1}{\sigma_k^2 + P_k} = \lambda - \mu_k, \quad \mu_k P_k = 0$$

therefore the optimal solution is given by

$$P_k = |T - \sigma_k^2|^+, \quad T \text{ is chosen such that } P = \sum_k |T - \sigma_k^2|^+$$

□



waterfilling across 3 parallel channels

Note: [Significance of the waterfilling theorem] In the high SNR regime, the capacity for 1 AWGN channel is approximately $\frac{1}{2}\log P$, while the capacity for L parallel AWGN channel is approximately $\frac{L}{2}\log(\frac{P}{L}) \approx \frac{L}{2}\log P$ for large P. This L-fold increase in capacity at high SNR regime leads to the powerful technique of spatial multiplexing in MIMO.

Also notice that this gain does not come from multipath diversity. Consider the scheme that a single stream of data is sent through every parallel channel simultaneously, with multipath diversity, the effective noise level is reduced to $\frac{1}{L}$, and the capacity is approximately $\log(LP)$, which is much smaller than $\frac{L}{2}\log(\frac{P}{L})$ for P large.

17.3.3 Non-stationary AWGN

Definition 17.7 (Non-stationary AWGN)**.** *A non-stationary AWGN channel is defined as follows:* $\mathcal{A} = \mathcal{B} = \mathbb{R}$, $\mathsf{c}(x) = x^2$, $P_{Y_j|X_j} : Y_j = X_j + Z_j$, *where* $Z_j \sim \mathcal{N}(0, \sigma_j^2)$.

Theorem 17.7. *Assume that for every* T *the following limits exist:*

$$C_i(T) = \lim_{n\to\infty} \frac{1}{n} \sum_{j}^{n} \frac{1}{2} \log^+ \frac{T}{\sigma_j^2}$$

$$P(T) = \lim_{n\to\infty} \frac{1}{n} \sum_{j}^{n} |T - \sigma_j^2|^+$$

then the capacity of the non-stationary AWGN channel is given by the parameterized form: $C(T) = C_i(T)$ *with input power constraint* $P(T)$.

Proof. Compute capacity by remembering waterfilling:

$$\frac{1}{n} \sup_{P_{X^n}} I(X^n; Y^n) \overset{\text{waterfilling}}{=} \frac{1}{n} \sup_{P_{X^n}} \sum_{j=1}^{n} \log^+ \frac{T}{\sigma_j^2} \to C_i(T)$$

Verify that the channel is information stable:

$$\mathrm{Var}(i(X_j, Y_j)) = \frac{\log^2 e}{2} \frac{P_j}{P_j + \sigma_j^2} \leq \frac{\log^2 e}{2}$$

and thus

$$\sum_{j=1}^{n} \frac{1}{n^2} \mathrm{Var}(i(X_j, Y_j)) < \infty$$

□

Note: Non-stationary AWGN is primarily interesting due to its relationship to the stationary Additive Colored Gaussian noise channel in the following discussion.

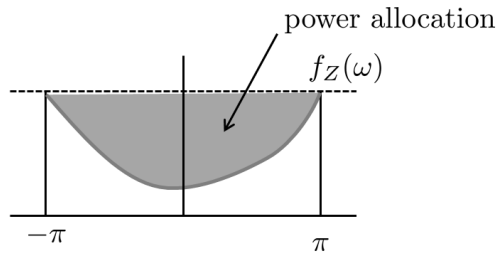
17.3.4 Stationary Additive Colored Gaussian noise channel

Definition 17.8 (Additive colored Gaussian noise channel)**.** *An Additive Colored Gaussian noise channel is defined as follows:* $\mathcal{A} = \mathcal{B} = \mathbb{R}$, $\mathsf{c}(x) = x^2$, $P_{Y_j|X_j} : Y_j = X_j + Z_j$, *where* Z_j *is a stationary Gaussian process with spectral density* $f_Z(\omega) > 0, \omega \in [-\pi, \pi]$.

Theorem 17.8. *The capacity of stationary ACGN channel is given by the parameterized form:*

$$C(T) = \frac{1}{2\pi} \int_0^{2\pi} \frac{1}{2} \log^+ \frac{T}{f_Z(\omega)} d\omega$$

$$P(T) = \frac{1}{2\pi} \int_0^{2\pi} \left| T - f_Z(\omega) \right|^+ d\omega$$



waterfilling across spectrum for stationary ACGN channel

Proof. Take $n \geq 1$, consider the diagonalization of the covariance matrix of Z^n:

$$Cov(Z^n) = \Sigma = U^* \widetilde{\Sigma} U, \text{ such that } \widetilde{\Sigma} = diag(\sigma_1, \ldots, \sigma_n)$$

Since $Cov(Z^n)$ is positive semi-definite, U is a unitary matrix. Define $\widetilde{X}^n = UX^n$ and $\widetilde{Y}^n = UY^n$, the channel between $\widetilde{X}^n$ and $\widetilde{Y}^n$ is thus

$$\begin{aligned} &\widetilde{Y}^n = \widetilde{X}^n + UZ^n, \\ &Cov(UZ^n) = UCov(Z^n)U^* = \widetilde{\Sigma} \end{aligned}$$

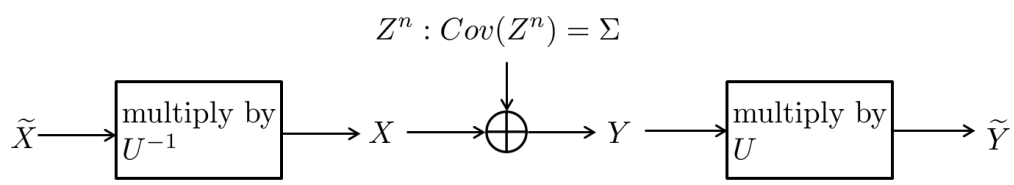
Therefore we have the equivalent channel as follows:

$$\widetilde{Y}^n = \widetilde{X}^n + \widetilde{Z}^n, \ \widetilde{Z}_j^n \sim \mathcal{N}(0, \sigma_j^2) \text{ indep across } j$$

By Theorem 17.7, we have that

$$\begin{aligned} \widetilde{C} &= \lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} \log^+ \frac{T}{\sigma_j^2} = \frac{1}{2\pi} \int_0^{2\pi} \frac{1}{2} \log^+ \frac{T}{f_Z(\omega)} d\omega. \quad (Szego) \\ &\lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} |T - \sigma_j^2|^+ = P(T) \end{aligned}$$

Finally since U is unitary, $C = \widetilde{C}$.



stationary additive Gausian noise channel

□

Note: Noise is born white, the colored noise is essentially due to some filtering.

17.4 Additive White Gaussian Noise channel with Intersymbol Interference

Definition 17.9 (AWGN with ISI)**.** *An AWGN channel with ISI is defined as follows:* $\mathcal{A} = \mathcal{B} = \mathbb{R}$, $\mathsf{c}(x) = x^2$, *and the channel law* $P_{Y^n|X^n}$ *is given by*

$$Y_k = \sum_{j=1}^{n} h_{k-j} X_j + Z_k \,, \qquad k = 1, \ldots, n$$

where $Z_k \sim \mathcal{N}(0,1)$ is white Gaussian noise, $\{h_k, k = -\infty, \ldots, \infty\}$ are coefficients of a discrete-time channel filter.

Theorem 17.9. *Suppose that the sequence $\{h_k\}$ is an inverse Fourier transform of a frequency response $H(\omega)$:*

$$h_k = \frac{1}{2\pi} \int_0^{2\pi} e^{i\omega k} H(\omega) d\omega \,.$$

Assume also that $H(\omega)$ is a continuous function on $[0, 2\pi]$. Then the capacity of the AWGN channel with ISI is given by

$$C(T) = \frac{1}{2\pi} \int_0^{2\pi} \frac{1}{2} \log^+(T|H(\omega)|^2) d\omega$$
$$P(T) = \frac{1}{2\pi} \int_0^{2\pi} \left| T - \frac{1}{|H(\omega)|^2} \right|^+ d\omega$$

Proof. (Sketch) At the decoder apply the inverse filter with frequency response $\omega \mapsto \frac{1}{H(\omega)}$. The equivalent channel then becomes a stationary colored-noise Gaussian channel:

$$\tilde{Y}_j = X_j + \tilde{Z}_j \,,$$

where $\tilde{Z}_j$ is a stationary Gaussian process with spectral density

$$f_{\tilde{Z}}(\omega) = \frac{1}{|H(\omega)|^2} \,.$$

Then apply Theorem 17.8 to the resulting channel.

Remark: to make the above argument rigorous one must simply carefully analyze the non-zero error introduced by truncating the deconvolution filter to finite n. □

17.5 More on Gaussian noise channels

We have examined some classical results of additive Gaussian noise channels, in the following, we will list some more recent results without proof.

Theorem 17.10 (Amplitude-constrained capacity of AWGN channel)**.** *For an AWGN channel $Y_i = X_i + Z_i$ with amplitude constraint $|X_i| \le A$ and energy constraint $\sum_{i=1}^n X_i^2 \le nP$, we denote the capacity by:*

$$C(A, P) = \max_{P_X:} I(X; X + Z).$$

The optimal input distribution P_X^* has finitely many stems on $[-A, A]$, namely X^* is a discrete R.V. Moreover the convergence speed of $\lim_{A\to\infty} C(A, P) = \frac{1}{2}\log(1 + P)$ is of the order $O(e^{-A^2})$.

Fading channel Fading channels are often used to model the urban signal propagation with multipath or shadowing. The received signal Y_i is modeled to be affected by multiplicative fading coefficient H_i and additive noise Z_i:

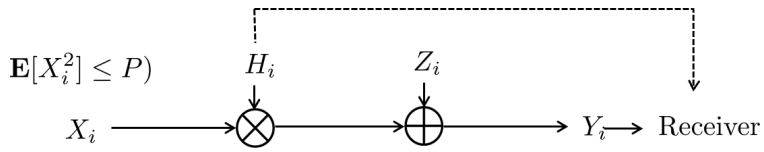
$$Y_i = X_i + Z_i, \quad Z_i \sim \mathcal{N}(0,1)$$

In the coherent case, the receiver can estimate the channel state information of H_i, namely the channel output is (Y_i, H_i). Wheneer H_j is a stationary ergodic process, we have the channel capacity given by:

$$C(P) = \mathbb{E}[\frac{1}{2}\log(1 + P|H|^2)]$$

and the capacity achieving input distribution is Normal. Note that the capacity $C(P)$ is in the order of $\log(P)$ and we call the channel "energy efficient".

In the non-coherent case where the receiver does not have the information of H_i, the channel capacity is unknown, but we know that the capacity achieving distribution is disrete, and the capacity $C(P) \sim O(\log\log P)$ for $P \gg 1$, and the channel is energy inefficient.



Fading channel

6.441: Information Theory	Spring 2012
Lecturer: Yury Polyanskiy **Lecture 18: Channel coding: energy-per-bit, cts time**	

18.1 Energy per bit

Consider the additive Gaussian noise channel:

$$Y_i = X_i + Z_i, \quad Z_i \sim \mathcal{N}(0, \frac{N_0}{2}). \tag{1}$$

In the last lecture, we analyzed the maximum number of information bits ($M^*(n, \epsilon, P)$) that can be pumped through for given n time use of the channel under the energy constraint P. Today we shall study the counterpart of it: without any time constraint, in order to send k information bits, what is the minimum energy needed? ($E^*(k, \epsilon)$)

Definition 18.1 ($(E, 2^k, \epsilon)$ code)**.** *For a channel $W \to X^\infty \to Y^\infty \to \hat{W}$, where $Y^\infty = X^\infty + Z^\infty$, a $(E, 2^k, \epsilon)$ code is a pair of encoder-decoder:*

$$\begin{aligned} & f : [2^k] \to \mathbb{R}^\infty, \quad g : \mathbb{R}^\infty \to [2^k], \\ \textit{such that } 1). \quad & \forall m, \|f(m)\|_2^2 \le E, \\ 2). \quad & P[g(f(W) + Z^\infty) \neq W] \le \epsilon. \end{aligned}$$

Definition 18.2 (Fundamental limit)**.**

$$E^*(k, \epsilon) = \min\{E : \exists (E, 2^k, \epsilon) \ \textit{code}\}$$

Note: Operational meaning of $\lim_{\epsilon \to 0} E^*(k, \epsilon)$: it suggests the smallest battery one needs in order to send k bits without any time constraints, below that level reliable communication is impossible.

Theorem 18.1 ($(E_b/N_0)_{min} = -1.6dB$)**.**

$$\lim_{\epsilon \to 0} \limsup_{k \to \infty} \frac{E^*(k, \epsilon)}{k} = \frac{N_0}{\log_2 e}, \quad \frac{1}{\log_2 e} = -1.6dB \tag{2}$$

Proof.

Typed by Qingqing Huang (qqh@mit.edu), Apr. 23 , 2013

1. ("$\geq$" converse part)

$$
\begin{aligned}
-h(\epsilon) + \bar{\epsilon}k &\leq d((1-\epsilon)\|\frac{1}{M}) \qquad \text{(Fano)} \\
&\leq I(W;\hat{W}) \qquad \text{(data processing for divergence)} \\
&\leq I(X^\infty;Y^\infty) \qquad \text{(data processing for M.I.)} \\
&\leq \sum_{i=1}^{\infty} I(X_i;Y_i) \qquad (\lim_{n\to\infty} I(X^n;U) = I(X^\infty;U)) \\
&\leq \sum_{i=1}^{\infty} \frac{1}{2}\log(1+\frac{\mathbb{E}X_i^2}{N_0/2}) \qquad \text{(Gaussian)} \\
&\leq \frac{\log e}{2}\sum_{i=1}^{\infty}\frac{\mathbb{E}X_i^2}{N_0/2} \qquad \text{(linearization)} \\
&\leq \frac{E}{N_0}\log e \\
\Rightarrow \frac{E^*(k,\epsilon)}{k} &\geq \frac{N_0}{\log e}(\bar{\epsilon} - \frac{h(\epsilon)}{k}).
\end{aligned}
$$

2. ("$\leq$" achievability part)
Notice that a $(n, 2^k, \epsilon, P)$ code for AWGN channel is also a $(nP, 2^k, \epsilon)$ code for the energy problem without time constraint. Therefore,

$$\log_2 M^*_{max}(n,\epsilon,P) \geq k \Rightarrow E^*(k,\epsilon) \leq nP.$$

$\forall P$, take $k_n = \lfloor \log M^*_{max}(n,\epsilon,P)\rfloor$, we have $\frac{E^*(k_n,\epsilon)}{k_n} \leq \frac{nP}{k_n}$, $\forall n$, and take the limit:

$$
\begin{aligned}
\limsup_{n\to\infty}\frac{E^*(k_n,\epsilon)}{k_n} &\leq \limsup_{n\to\infty}\frac{nP}{\log M^*_{max}(n,\epsilon,P)} \\
&= \frac{P}{\liminf_{n\to\infty}\frac{1}{n}\log M^*_{max}(n,\epsilon,P)} \\
&= \frac{P}{\frac{1}{2}\log(1+\frac{P}{N_0/2})}
\end{aligned}
$$

Choose P for the lowest upper bound:

$$
\begin{aligned}
\limsup_{n\to\infty}\frac{E^*(k_n,\epsilon)}{k_n} &\leq \inf_{P\geq 0}\frac{P}{\frac{1}{2}\log(1+\frac{P}{N_0/2})} \\
&= \lim_{P\to 0}\frac{P}{\frac{1}{2}\log(1+\frac{P}{N_0/2})} \\
&= \frac{N_0}{\log_2 e}
\end{aligned}
$$

□

Note: [Remark] In order to send information reliably at $E_b/N_0 = -1.6dB$, infinitely many time slots are needed, and the information rate (spectral efficiency) is thus 0. In order to have non-zero spectral efficiency, one necessarily has to step back from $-1.6\ dB$.

Note: [PPM code] The following code, pulse-position modulation (PPM), is very efficient in terms of E_b/N_0.

$$\text{PPM encoder: } \forall m, f(m) = (0, 0, \ldots, \underbrace{\sqrt{E}}_{m\text{-th location}}, \ldots)$$

In fact, the code can be further slightly optimized by subtracting the common center of gravity $(2^{-k}\sqrt{E}, \ldots, 2^{-k}\sqrt{E} \ldots)$ and rescaling each codeword to satisfy the power constraint. The resulting constellation (simplex code) is conjectured to be non-asymptotic optimum in terms of E_b/N_0 for small ϵ.

18.2 What is N_0?

In the above discussion, we have assumed $Z_i \sim \mathcal{N}(0, N_0/2)$, but how do we determine N_0?

In reality the signals are continuous time (CT) process, the continuous time AWGN channel for the RF signals is modeled as:

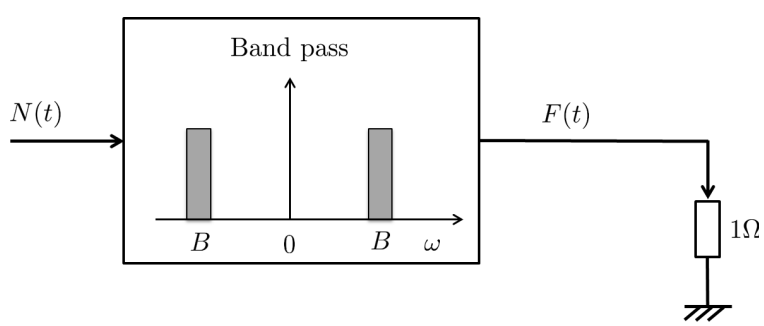
$$Y(t) = X(t) + N(t) \tag{3}$$

where noise $N(t)$ (added at the receiver antenna) is a real stationary ergodic process and is assumed to be "white Gaussian noise" with single-sided PSD N_0. Figure 1 at the end illustrates the communication architecture. In the following discussion, we shall find the equivalent discrete time (DT) AWGN model for the continuous time (CT) AWGN model in (3), and identify the relationship between N_0 in the DT model and $N(t)$ in the CT model.

- Goal: communication in $f_c \pm B/2$ band.
  (the (possibly complex) baseband signal lies in $[-W, +W]$, where $W = B/2$)
- observations:
  1. Any signal band limited to $f_c \pm B/2$ can be produced by this architecture
  2. At the step of C/D conversion, the LPF followed by sampling at B samples/sec is sufficient statistics for estimating $X(t), X_B(t)$, as well as $\{X_i\}$.

First of all, what is $N(t)$ in (3)?

Engineers' definition of $N(t)$



Testing whether a process $N(t)$ is "white noise"

Estimate the average power dissipation at the resistor:

$$\lim_{T\to\infty}\frac{1}{T}\int_{t=0}^{T}F_t^2dt \overset{\text{ergodic}}{=\!=\!=} \mathbb{E}[F^2] \overset{(*)}{=\!=} N_0B$$

If for some constant N_0, (*) holds for any narrow band with center frequency f_c and bandwidth B, then $N(t)$ is called a "white noise" with one-sided PSD N_0.

Typically, white noise comes from thermol noise at the receiver antenna. Thus:

$$N_0 \approx k\mathbf{T} \tag{4}$$

where $k = 1.38 \times 10^{-23}$ is the Boltzmann constant, and $\mathbf{T}$ is the absolute temprature. The unit of N_0 is $(Watt/Hz = J)$.

An intuitive explanation to (4) is as follows: the thermal energy carried by each microscopic degree of freedom (dof) is approximately $\frac{k\mathbf{T}}{2}$; for bandwidth B and duration T, there are in total $2BT$ dof; by "white noise" definition we have the total energy of the noise to be:

$$N_0BT = \frac{k\mathbf{T}}{2}2BT \;\Rightarrow\; N_0 = k\mathbf{T}.$$

Mathematicians' definition of $N(t)$

Denote the set of all real finite energy signals $f(t)$ by $\mathcal{L}_2(\mathbb{R})$, it is a vector space with with the inner product of two signals $f(t), g(t)$ defined by

$$< f, g >= \int_{t=-\infty}^{\infty} f(t)g(t)dt.$$

Definition 18.3 (White noise)**.** *$N(t)$ is a white noise with two-sided PSD being constant $N_0/2$ if $\forall f, g \in \mathcal{L}_2(\mathbb{R})$ such that $\int_{-\infty}^{\infty} f^2(t)dt = \int_{-\infty}^{\infty} g^2(t)dt = 1$, we have that*

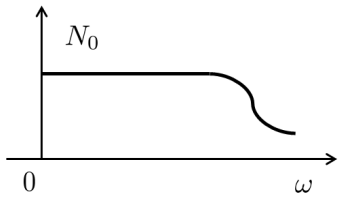
1. $$< f, N > \overset{\triangle}{=} \int_{-\infty}^{\infty} f(t)N(t)dt \sim \mathcal{N}(0, \frac{N_0}{2}). \tag{5}$$

2. *The joint distribution of $(< f, N >, < g, N >)$ is jointly Gaussian with covariance equal to inner product $< f, g >$.*

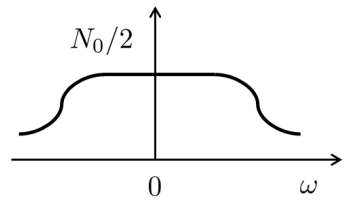
Note: By this definition, $N(t)$ is not a stochastic process, rather it is a collection of linear mappings that map any $f \in \mathcal{L}_2(\mathbb{R})$ to a Gaussian random variable.

Note: Informally, we write:

$$N(t) \text{ is white noise with one-sided PSD } N_0 (or \text{ two-sided PSD } N_0/2) \Longleftrightarrow \mathbb{E}[N(t)N(s)] = \frac{N_0}{2}\delta(t-s) \tag{6}$$



Engineers' white noise



Mathematicians' white noise

Note: The concept of one-sided PSD arises when $N(t)$ is necessarily real, since in that case power spectrum density is symmetric around 0, and thus to get the noise power in band $[a, b]$ one can get

$$\text{noise power} = \int_a^b F_{\text{one-sided}}(f)df = \int_a^b + \int_{-b}^{-a} F_{\text{two-sided}}(f)df\,,$$

where $F_{\text{one-sided}}(f) = 2F_{\text{two-sided}}(f)$. In theory of stochastic processes it is uncommon to talk about one-sided PSD, but in engineering it is.

Verify the equivalence between CT /DT models

First, consider the relation between RF signals and baseband signals.

$$X(t) = Re(X_B(t)\sqrt{2}e^{j\omega_c t}),$$
$$Y_B(t) = \sqrt{2}LPF_2(Y(t)e^{j\omega_c t}),$$

where $\omega_c = 2\pi f_c$. The LPF_2 with high cutoff frequency $\sim \frac{3}{4}f_c$ serves to kill the high frequency component after demodulation, and the amplifier of magnitude $\sqrt{2}$ serves to preserve the total energy of the signal, so that in the absence of noise we have that $Y_B(t) = X_B(t)$. Therefore,

$$Y_B(t) = X_B(t) + \widetilde{N}(t) \sim \mathbb{C}$$

where $\widetilde{N}(t)$ is a complex Gaussian white noise and

$$\mathbb{E}[\widetilde{N}(t)\widetilde{N}(s)^*] = N_0\delta(t-s).$$

Notice that after demodulation, the PSD of the noise is $N_0/2$ with $N_0/4$ in the real part and $N_0/4$ in the imaginary part, and after the $\sqrt{2}$ amplifier the PSD of the noise is restored to $N_0/2$ in both real and imaginary part.

Next, consider the equivalent discrete time signals.

$$X_B(t) = \sum_{i=-\infty}^{\infty} X_i sinc_B(t - \frac{i}{B})$$
$$Y_i = \int_{t=-\infty}^{\infty} Y_B(t) sinc_B(t - \frac{i}{B})dt$$
$$Y_i = X_i + Z_i$$

where the additive noise Z_i is given by:

$$Z_i = \int_{t=-\infty}^{\infty} \widetilde{N}(t) sinc_B(t - \frac{i}{B})dt \sim i.i.d\ \mathbb{C}\mathcal{N}(0, N_0). \qquad \text{(by (5))}$$

if we focus on the real part of all signals, it is consistent with the real AWGN channel model in (1). Finally, the energy of the signal is preserved:

$$\sum_{i=-\infty}^{\infty} |X_i|^2 = \|X_B(t)\|_2^2 = \|X(t)\|_2^2.$$

Note: [Punchline]

$$\begin{aligned} \text{CT AWGN (band limited)} &\Longleftrightarrow \text{DT } \mathbb{C}\text{-AWGN} \\ \text{two-sided PSD } \frac{N_0}{2} &\Longleftrightarrow Z_i \sim \mathbb{C}\mathcal{N}(0, N_0) \\ \text{energy}=\int X(t)^2 dt &\Longleftrightarrow \text{energy}=\sum |X_i|^2 \end{aligned}$$

18.3 Capacity of the continuous-time bandlimited AWGN channel

Theorem 18.2. *Let $M^*_{CT}(T, \epsilon, P)$ the maximum number of waveforms that can be sent through the channel*

$$Y(t) = X(t) + N(t)\,, \qquad \mathbb{E}\, N(t)N(s) = \frac{N_0}{2}\delta(t-s)$$

such that:

1. *in the duration $[0, T]$;*
2. *band limited to $[f_c - \frac{B}{2}, f_c + \frac{B}{2}]$ for some large carrier frequency*
3. *input energy constrained to $\int_{t=0}^{T} x^2(t) \le TP$;*
4. *error probability $P[\hat{W} \neq W] \le \epsilon$.*

Then

$$\lim_{\epsilon\to 0} \liminf_{n\to\infty} \frac{1}{T} \log M^*_{CT}(T, \epsilon, P) = B \log(1 + \frac{P}{N_0 B})\,, \tag{7}$$

Proof. Consider the DT equivalent $\mathbb{C}$-AWGN channel of this CT model, we have that

$$\frac{1}{T} \log M^*_{CT}(T, \epsilon, P) = \frac{1}{T} \log M^*_{\mathbb{C}-\text{AWGN}}(BT, \epsilon, P/B)$$

This is because:

- in time T we get to choose BT complex samples
- The power constraint in the DT model changed because for blocklength BT we have

$$\sum_{i=1}^{BT} |X_i|^2 = \|X(t)\|_2^2 \le PT\,,$$

  thus per-letter power constraint is $\frac{P}{B}$.

Calculate the rate of the equivalent DT AWGN channel and we are done. $\square$

Note the above "theorem" is not rigorous, since conditions 1 and 2 are mutually exclusive: any time limited non-trivial signal cannot be band limited. Rigorously, one should relax 2 by constraining the signal to have a vanishing out-of-band energy as $T \to \infty$. Rigorous approach to this question lead to the theory of prolate spheroidal functions.

18.4 Capacity of the continuous-time band-unlimited AWGN channel

In the limit of large bandwidth B the capacity formula (8) yields

$$C_{B=\infty}(P) = \lim_{B\to\infty} B \log(1 + \frac{P}{N_0 B}) = \frac{P}{N_0} \log e \,.$$

It turns out that this result is easy to prove rigorously.

Theorem 18.3. *Let $M^*(T, \epsilon, P)$ the maximum number of waveforms that can be sent through the channel*

$$Y(t) = X(t) + N(t)\,, \qquad \mathbb{E}\, N(t)N(s) = \frac{N_0}{2}\delta(t-s)$$

such that each waveform $x(t)$

1. *is non-zero only on $[0, T]$;*
2. *input energy constrained to $\int_{t=0}^{T} x^2(t) \le TP$;*
3. *error probability $P[\hat{W} \neq W] \le \epsilon$.*

Then

$$\lim_{\epsilon\to 0} \liminf_{T\to\infty} \frac{1}{T} \log M^*(T, \epsilon, P) = \frac{P}{N_0} \log e \tag{8}$$

Proof. Note that the space of all square-integrable functions on $[0, T]$, denoted $L_2[0, T]$ has countable basis (e.g. sinusoids). Thus, by changing to that basis we may assume that equivalent channel model

$$\tilde{Y}_j = \tilde{X}_j + \tilde{Z}_j\,, \qquad \tilde{Z}_j \sim \mathcal{N}(0, \frac{N_0}{2})\,,$$

and energy constraint (dependent upon duration T):

$$\sum_{j=1}^{\infty} \tilde{X}_j^2 \le PT\,.$$

But then the problem is equivalent to energy-per-bit one and hence

$$\log_2 M^*(T, P, \epsilon) = k \iff E^*(k, \epsilon) = PT\,.$$

Thus,

$$\lim_{\epsilon\to 0} \liminf_{n\to\infty} \frac{1}{T} \log_2 M^*(T, P, \epsilon) = \frac{P}{\lim_{\epsilon\to 0} \limsup_{k\to\infty} \frac{E^*(k,\epsilon)}{k}} = \frac{P}{N_0} \log_2 e\,,$$

where the last step is by Theorem 18.1. $\square$

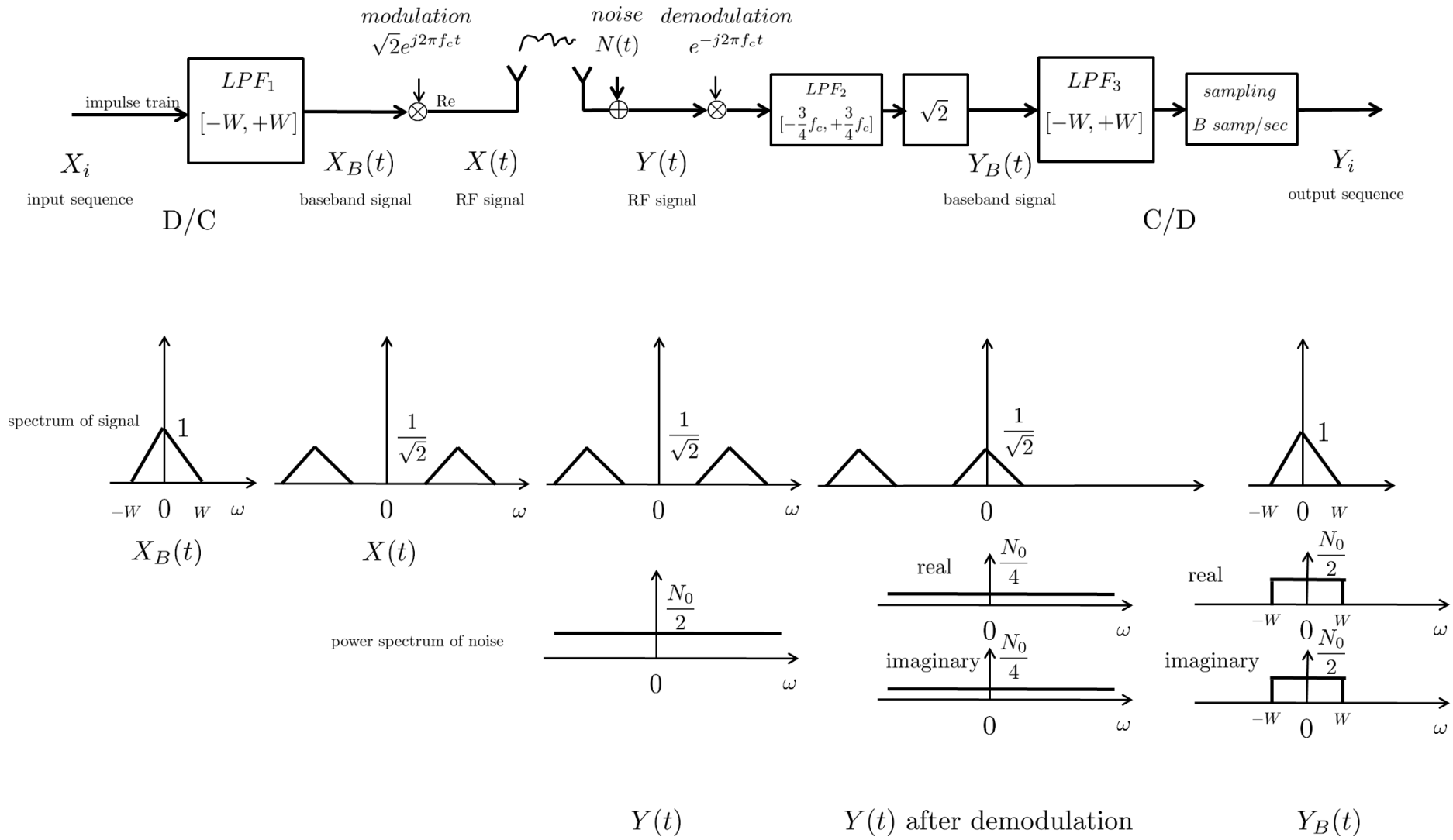


Figure 1: DT / CT AWGN model

6.441: Information Theory	Spring 2013
Lecturer: Yury Polyanskiy	**Lecture 19: Advanced Channel Coding**

Topics: Strong Converse, Channel Dispersion, Joint Source Channel Coding

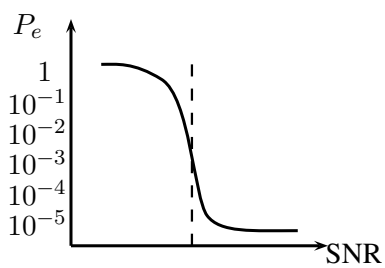
19.1 Strong Converse

We begin by stating the main theorem.

Theorem 19.1. *For any memoryless channel, $C_\epsilon = C$ for $0 < \epsilon < 1$*

Remark: Before, we showed that $C \leq C_\epsilon \leq \frac{C}{1-\epsilon}$. 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 exponentially fast to 1.

In practice, engineers look at *waterfall plots* to see this phenomenon, as shown below.



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

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

$$\begin{tikzcd} W \arrow[r] & X^n \arrow[r, "P_{Y^n|X^n}"] \arrow[r, bend right, "Q_{Y^n|X^n}"'] & Y^n \arrow[r] & \hat{W} \end{tikzcd}$$

Here, we take $Q_{Y^n|X^n} = \prod_{j=1}^n P_Y^*(y_j)$, where P_Y^* is the caod of the channel $P_{Y|X}$. Recall that the caod of a random transformation *always exists and is unique*, whereas a caid may not exist. Note that $Q_{Y^n|X^n}$ is a useless channel; it ignores the input and randomly picks a member of the output

Typed by Austin Collins, austinc@mit.edu, April 25, 2013

space according to $(P_Y^*)^n$, so that X^n and Y^n are decoupled. Consider the probability of error under each channel:

$$\mathbb{Q}[\hat{W} = W] = \frac{1}{M} \quad \text{(Guessing the sent codeword)}$$
$$\mathbb{P}[\hat{W} = W] = 1 - \epsilon$$

Since the set $\{\hat{W} = W\}$ has a huge mass under $\mathbb{P}$ and small mass under $\mathbb{Q}$, this looks like a great binary hypothesis tests between 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

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

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

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

Therefore, inequality above becomes

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

Computing this bound isn't often easy, since generally we know $P_{Y|X}$ and P_Y^*, but can't assume anything about P_X. Tricks are needed to get ride of this uncertainty, as shown in the following example for a BSC.

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

$$P_{Y^n|X^n}(y^n|x^n) = P_Z^n(y^n - x^n), \quad Z^n \sim \text{Ber}(\delta)^n$$
$$P_Y^*(y) = 2^{-n}$$

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

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

For the BSC, this becomes

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

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

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

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

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

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

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

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

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

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

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

Where $\bar{P}_X$ is unknown (depends on the code).

Explanation of (2): A statistician observes sequences of (X^n, Y^n):

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

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 (2).

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

$$D(P_{Y|X=0}||P_Y^*|\bar{P}_X) \leq C\,. \tag{3}$$

Thus from (2) and (1):

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

□

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

1. Strong converse is proven by using binary hypothesis testing.
2. The capacity saddle point (**??**) is key.

19.2 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 19.2. *For the following 3 classes of channels:*

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

The following expansion holds

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

Where Q^{-1} is the inverse of the complementary standard normal CDF, $C = I(X^*, Y^*)$, and $V = Var[i(X^*, Y^*)|X^*]$.

Proof. For achievability, earlier we showed 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

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

Where the dispersion is

$$V = Var\left[\log \frac{P_Z \sim \mathrm{Ber}(\delta)^n}{Q_Z \sim \mathrm{Ber}(\frac{1}{2})}\right]$$

The general proof is omitted. □

Remark: This expansion only applies for certain channels (as described in the theorem). If, for example, $Var[i(X,Y)] = \infty$, then there are other stable (non-Gaussian) distributions that we might converge to instead.

19.2.1 Applications

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

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

For certain channels, such as BSC, AWGN, or anything with zeros in $P_{Y|X}$, the next term in this approximation is known to be $1/2 \log n$, and for the BEC the $\log n$ term has coefficient 0. 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)

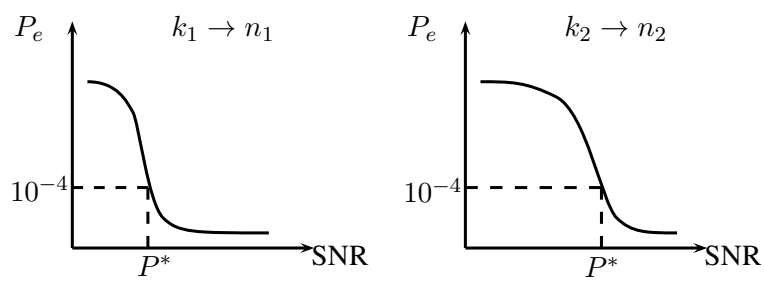
Parallel AWGN: $V(\mathbf{P}, \sigma^2) = \sum_{j=1}^{L} V_{AWGN}(\frac{P_j}{\sigma_j^2}) = \frac{\log^2 e}{2}\sum_{j=1}^{L}\left|1 - \left(\frac{\delta_j^2}{T}\right)^2\right|^+$

where $\sum_{j=1}^{L}|T - \sigma_j^2|^+ = P$ is the waterfilling 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.

19.3 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



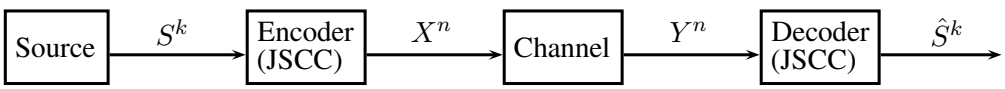
After inspecting these plots, one may believe that the $k_1 \to n_1$ code is better, since it requires a smaller SNR to achieve the same error probability. However, there are many factors, such as blocklength, rate, etc. that don't appear on these plots. To get a fair comparision, we can use the notion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

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

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 the plot on the last page). This normalized rate gives another motivation for the expansion given in Theorem 19.2.

19.4 Joint Source Channel Coding

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



Definition 19.1. *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:* $\epsilon^*_{JSCC}(k, n) = \inf_{all\ JSCC\ enc/dec} \mathbb{P}[S^k \neq \hat{S}^k]$

Where $R = n/k$ is called the bandwidth expansion factor.

Remark: 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.

Theorem 19.3. *Let source $\{S_k\}$ be ergodic on a finite alphabet, and have entropy rate $\mathcal{H}$. Let the channel have capacity C and be information stable. Then*

$$\lim_{k\to\infty} \epsilon^*_{JSCC}(k, kR) = \begin{cases} 0 & R > \mathcal{H}/C \\ > 0 & R < \mathcal{H}/C \end{cases}$$

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

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

Where we use the optimal compressor (f_1, g_1) and optimal channel code (maximum probability of error) (f_2, g_2). Let $\log M_k = |W|$. Then from optimality of our compressor/encoder

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

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

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 $(C-\delta)R > \mathcal{H}+\delta$, then $\epsilon^* \to 0 \implies R > \mathcal{H}/C$.

(**Converse**) 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

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

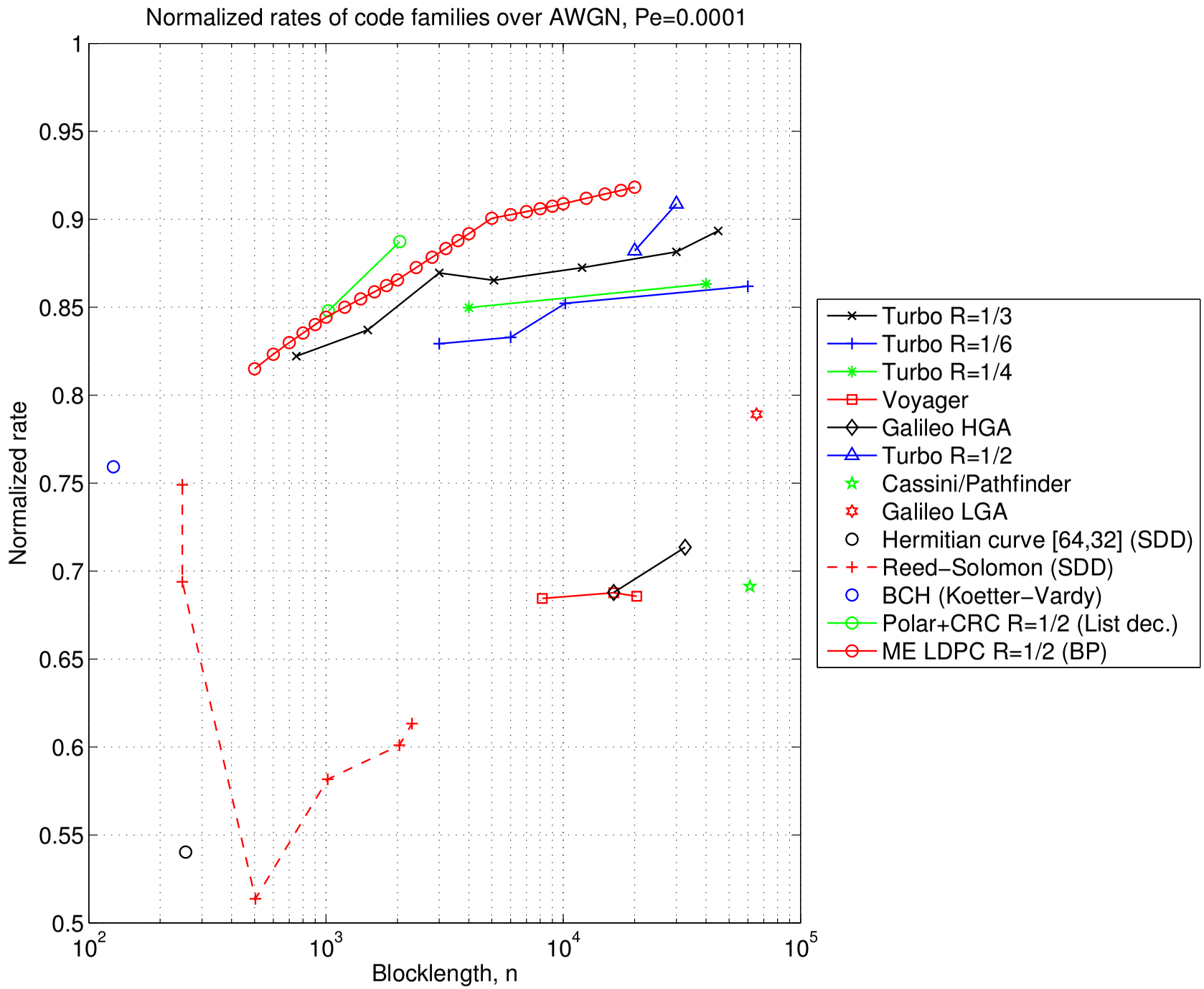
Which follows from expanding out the terms. Now, normalizing and taking the sup of both sides gives

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

letting $R = n/k$, this shows

$$C \geq R^{-1}\mathcal{H} - \epsilon R^{-1}\log|A| \implies \epsilon \geq \frac{R^{-1}\mathcal{H} - C}{R^{-1}\log|A|} > 0$$

Where the last expression is positive when $R^{-1} > C/\mathcal{H}$. $\square$

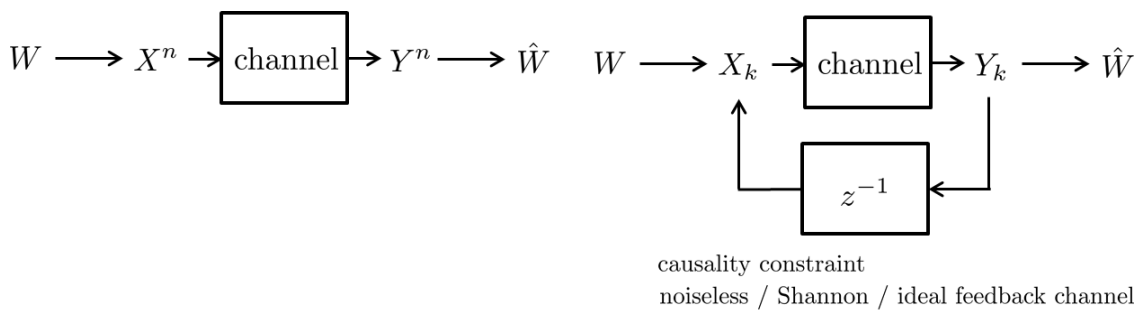

Normalized rates of code families over AWGN, Pe=0.0001
Normalized rate
Blocklength, n
1
0.95
0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
10^2
10^3
10^4
10^5
Turbo R=1/3
Turbo R=1/6
Turbo R=1/4
Voyager
Galileo HGA
Turbo R=1/2
Cassini/Pathfinder
Galileo LGA
Hermitian curve [64,32] (SDD)
Reed−Solomon (SDD)
BCH (Koetter−Vardy)
Polar+CRC R=1/2 (List dec.)
ME LDPC R=1/2 (BP)

6.441: Information Theory	Spring 2013
Lecturer: Yury Polyanskiy	**Lecture 20: Channel coding with feedback**

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

Note: [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 (satelite). 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 traditional setup, while in the second half, we look at situations where feedback gains a lot.



channel w/o feedback channel with feedback

20.1 $C_{fb} = C$ for memoryless channels

Definition 20.1 (Code with feedback)**.** *An $(n, M, \epsilon)_{avg}$ 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:*

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

Typed by Qingqing Huang (qqh@mit.edu), Apr. 30, 2013

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

Note: [Probability space]

$$\begin{aligned} &W \sim \text{ uniform on } [M] \\ &\left.\begin{aligned} &X_1 = f_1(W) \xrightarrow{P_{Y|X}} Y_1 \\ &\vdots \\ &X_1 = f_1(W, Y_1^{n-1}) \xrightarrow{P_{Y|X}} Y_n \end{aligned}\right\} \longrightarrow \hat{W} = g(Y^n) \end{aligned}$$

Definition 20.2 (Fundamental limits)**.**

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

Theorem 20.1 (Shannon)**.** *For a stationary memoryless channel,*

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

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

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

$$P : W \to X^n \to Y^n \to \hat{W} \tag{1}$$

$$Q : W \to X^n \quad Y^n \to \hat{W} \tag{2}$$

   two key observations:

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

   (b) The two graphical models give the factorization:

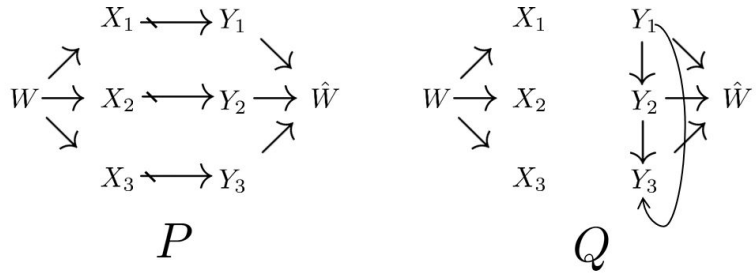
$$P_{W,X^n,Y^n,\hat{W}} = P_{W,X^n} P_{Y^n|X^n} P_{\hat{W}|Y^n}, \qquad Q_{W,X^n,Y^n,\hat{W}} = P_{W,X^n} P_{Y^n} P_{\hat{W}|Y^n}$$

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

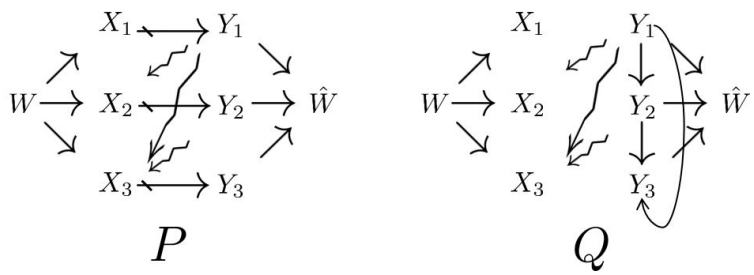
$$h(\epsilon) + \bar{\epsilon} \log M = d(1 - \epsilon \| \frac{1}{M}) \overset{\text{d-proc ineq}}{\leq} D(P\|Q) = I(X^n; Y^n) \overset{mem-less,stat}{=} \sum_{i=1}^{n} I(X; Y) \leq nC_i \tag{3}$$

2. Notice that when feedback is present, $X^n \to Y^n$ is not memoryless due to the transmission protocal, 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 (3), as with feedback $I(X^n;Y^n)$ can be significant 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 as in (2), $D(P\|Q) = I(X^n, lY^n)$ measures the information flow through all the $\not\to$ and $\rightsquigarrow$ links. This motivates us to find a proper Q such that $D(P\|Q)$ only captures the information flow through all the $\not\to$ links $\{X_i \to Y_i : i = 1, \ldots, n\}$, so that $D(P\|Q)$ closely relates to nC_i, while still guarantees that $W \perp \hat{W}$.

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:

$$Q_{W,X^n,Y^n,\hat{W}} = Q_W Q_{X_1|W} Q_{Y_1} Q_{X_2|W,Y_1} Q_{Y_2|Y_1} \cdots Q_{X_n|W,Y^{n-1}} Q_{Y_n|Y^{n-1}} Q_{\hat{W}|Y^n} \tag{4}$$

$$P_{W,X^n,Y^n,\hat{W}} = P_W P_{X_1|W} P_{Y_1|X_1} P_{X_2|W,Y_1} P_{Y_2|X_2} \cdots P_{X_n|W,Y^{n-1}} P_{Y_n|X_n} P_{\hat{W}|Y^n} \tag{5}$$

Verify that $W \perp \hat{W}$: W and $\hat{W}$ are d-seperated 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 $W \perp \hat{W}$ 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 20.1 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 20.1.

$$\inf_{Q\in\mathcal{Q}} D(P_{W,X^n,Y^n,\hat{W}}\|Q_{W,X^n,Y^n,\hat{W}}) = \sum_{k=1}^{n} I(X_k;Y_k|Y^{k-1}) \tag{6}$$
$$(\triangleq \vec{I}(X^n;Y^n), \ \textit{directed information})$$

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

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

□

Theorem 20.2 (Dispersion with feedback)**.** *For weakly input-symmetric DMC (e.g. additive noise, BSC, BEC) we have:*

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

This suggests that feedback can at most improve smallish $\log n$ *terms.*

20.2 When is feedback really useful?

Theorems 20.1 and 20.2 say 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.

20.2.1 code with very small (e.g. zero) error probability

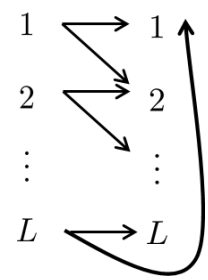
Theorem 20.3 (Shannon)**.** *For any DMC*

$$C_{fb,0} = \max_{P_X}\{-\log \max_{y\in\mathcal{B}} P_X(S_y)\} \tag{7}$$

Note: Observations

1. Usually C_0 is very hard to compute, but $C_{fb,0}$ can be obtained in closed form as above.

   Example:

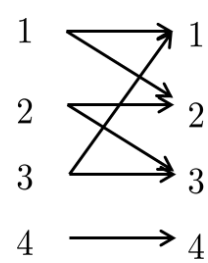


   for $L = 5, C_0 = \log\frac{5}{2}$, (Shannon-Lovasz)

   for $L = 7, \log 7 \le C_0 \le \log 3.32$, (exact value unknown to this day)

2. Notice that $C_{fb,0}$ is not necessarily equal to $C_{fb} = lim_{\epsilon\to 0} C_{fb,\epsilon} = C$, in fact $C_{fb,0}$ can be larger than C.

   Example:



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

Proof.

1. First notice that in DMC for code that achieves zeros error probability, in terms of coding and decoding, only whether $P_{Y|X}(b|a) > 0$ or $= 0$ matters, the magnitude of $P_{Y|X}(b|a)$ does not matter. Denote the set of all input symbols that could have produced the output symbol y by S_y:

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

2. Take 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

$$\epsilon = 0 \Leftrightarrow \forall y^n \in \mathcal{B}^n, |E_n(y^n)| = 1 \text{ or } 0. \tag{8}$$

3. The key quantity in the proof is define as follows:

$$
\begin{aligned}
\theta_{fb} &= \min_{P_X} \max_{y \in \mathcal{B}} P_X(S_y), \\
P_X^* &= \arg\min_{P_X} \max_{y \in \mathcal{B}} P_X(S_y)
\end{aligned}
$$

By definition, we have

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

Notice the relationship between the minimizer distribution P_X^* and 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.

4. "$\leq$":

$\forall P_{X_1}$, denote $E_0 = [M]$, $\exists y_1^*$ such that:

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

Fix y_1^*, $\forall P_{X_2}$, $\exists y_2^*$ such that:

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

Continue the selection process up to y_n^* which satisfies that:

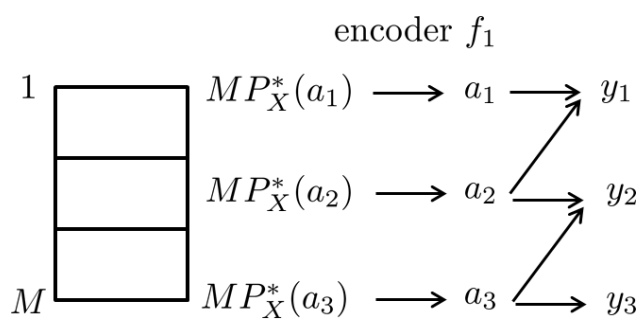
$$P_{X_n}(S_{y_n^*}|X_k \in S_{y_k^*} \text{ for } k = 1 \text{ to } n-1) = \frac{|E_n(y_1^*, \ldots, y_n^*)|}{|E_{n-1}(y_1^*, \ldots, y_{n-1}^*)|} \geq \theta_{fb}.$$

Finally, by (8) and the above selection procedure, we have

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

5. "$\geq$" (achiev)

   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 rest $MP_X^*(S_{y_1})$ messages. By definition of P_X^* we know that $MP_X^*(S_{y_1}) \leq M\theta_{fb}$.

   By similar arguments, after n iterations, the size of "confusability set" is bounded by $M\theta_{fb}^n$, if $M\theta_{fb}^n \leq 1$, 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 delievered with $n + O(1)$ channel uses with feedback, thus

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

□

20.2.2 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

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.

sending k bits is given by

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

We state the dispersion result without proof:

$$\log M^*_{VLF}(l, 0) \geq lC$$

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 20.2.

This is also true in general:

$$\log M^*_{VLP}(l, \epsilon) = \frac{lC}{1-\epsilon} + O(\log e)$$

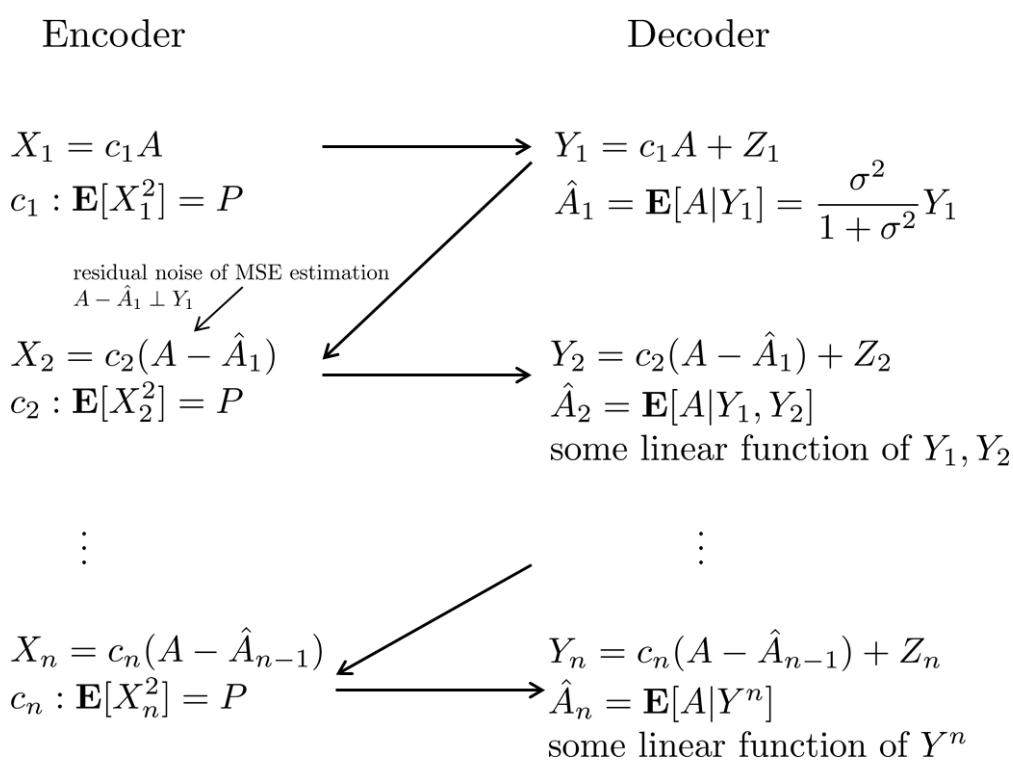
Example: For BSC(0.11), without feedback, $n = 3000$ is needed to achieve 90% of capacity C, while with variable length feedback (VLF) code $l = 200$ is enough to achieve that.

20.2.3 code with variable power

Elias's scheme of sending a number A drawn from a Gaussian distribution $\mathcal{N}(0, \sigma^2)$.

AWGN setup:

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



According to the orthogonality principle of the mininum MSE estimation of A at receiver side in every step:

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

Also notice that $\hat{A}_n$ will be linear combinations of Y^n, therefore we have

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

where the last equality uses the fact that N follows a normal distribution. 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's scheme of sending $A \sim \mathcal{N}(0,\sigma^2)$, after the n-th use of the AWGN(P) channel with feedback,

$$Var(\hat{A}_n - A) = \sigma^2 2^{-2nC},$$

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

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

Let $W \sim$ uniform on M-PAM $\in [-1,1]$ ($[-1, -1+\frac{2}{M}, \cdots, -1+\frac{2k}{M}, \cdots, 1]$). 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's scheme ($n-1$ times). After $n-1$ channel uses, and the MSE estimation, the equivalent channel output:

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

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

$$\epsilon \le \mathbb{P}[|\widetilde{Z}_0| > \frac{1}{2M}] = \mathbb{P}[2^{-(n-1)C}|Z| > \frac{\sqrt{P}}{2M}] = 2Q\left(\frac{2^{(n-1)C}}{2M}\right)$$

$$\begin{aligned} \Rightarrow \log M &\ge (n-1)C + \log(\frac{\sqrt{P}}{2M}) - \log Q^{-1}(\frac{\epsilon}{2}) \\ &= nC + O(1) \end{aligned}$$

the error probability drops almost doubly exponentially fast as n increases. More importantly, we gained a $\sqrt{n}$ 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 for the best possible block codes without feedback $n \approx 2800$ achieves about 90%.

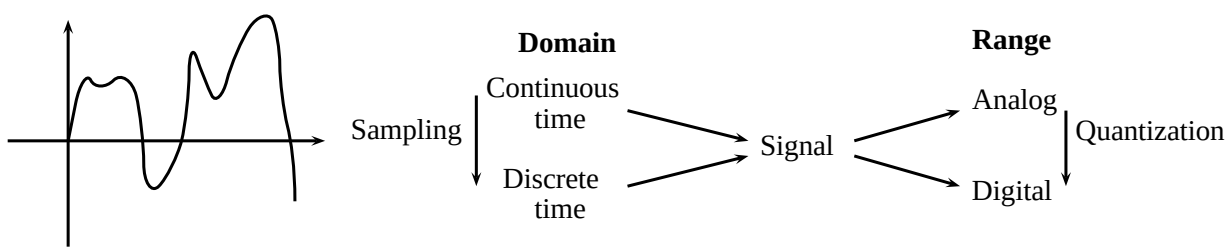
6.441: Information Theory	Spring 2013
Lecturer: Yury Polyanskiy	**Lecture 21: Rate Distortion Theory**

Big picture so far:

1. Given a discrete ergodic source S^k, we know how to encode to pure bits $W \in [2^k]$.
2. Know how to send bits over a channel $[2^k] \ni W \to X \to Y$.
3. JSCC showed how to send discrete data optimally over a noisy channel.

21.1 Quantization

Problem: Data isn't discrete! Often, a signal (function) comes from voltage levels or other continuous quantities. The question of how to map (naturally occurring) continuous time/analog signals into (electronics friendly) discrete/digital signals is known as *quantization*, or in information theory, as *rate distortion theory*.

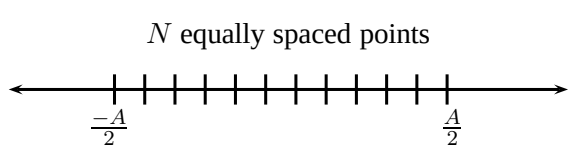


We will look at a few ways to do quantization in the next few sections.

21.1.1 Scalar Uniform Quantization

Quantization was analyzed before Shannon published his influential information theory paper in 1948. Still though, this work on scalar uniform quantizers was done by Oliver, Pierce, and Shannon (same Shannon) in 1947.

For a random variable $X \in [-A/2, A/2] \subset \mathbb{R}$, the scalar uniform quantizer $q_U(X)$ with N quantization points partitions the interval $[-A/2, A/2]$ uniformly



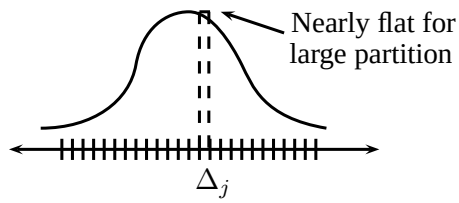
Typed by Austin Collins, austinc@mit.edu, May 2nd, 2013

Where the points are in $\{\frac{-A}{2} + \frac{kA}{N}, k = 0, \ldots, N-1\}$.

What is the *quality* (or fidelity) of this quantization? Most of the time, mean squared error is used as the quality criterion:

$$D(N) = \mathbb{E}|X - q_U(X)|^2$$

Where D is the *distortion*. Often $R = \log_2 N$ is used instead of N, so that we think about the number of bits we can use for quantization instead of the number of points. To analyze this scalar uniform quantizer, we'll look at the high rate regime ($R \gg 1$). The key idea in the high rate regime is that (assuming a smooth density P_X), each quantization interval Δ_j looks nearly flat, so conditioned on Δ_j, the distribution is accurately approximately by a uniform distribution.



Let c_j be the j-th quantization point, and Δ_j be the j-th quantization interval. Here we have

$$\begin{aligned}
\mathbb{E}|X - q_U(X)|^2 &= \sum_{j=1}^{N} \mathbb{E}[|X - c_j|^2 | X \in \Delta_j]\mathbb{P}[X \in \Delta_j] \\
\text{(high rate assumption)} \quad &= \sum_{j=1}^{N} \frac{|\Delta_j|^2}{12}\mathbb{P}[X \in \Delta_j] \\
&= \frac{(\frac{A}{N})^2}{12} = \frac{A^2}{12}2^{-2R}
\end{aligned}$$

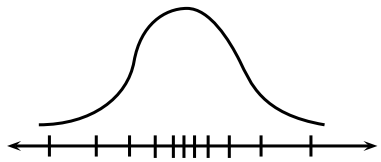
How much do we gain per bit?

$$\begin{aligned}
10\log_{10} SNR &= 10\log_{10} \frac{Var(X)}{\mathbb{E}|X - q_U(X)|^2} \\
&= 10\log_{10} \frac{12Var(X)}{A^2} + (20\log_{10} 2)R \\
&= \text{constant} + (6.02dB)R
\end{aligned}$$

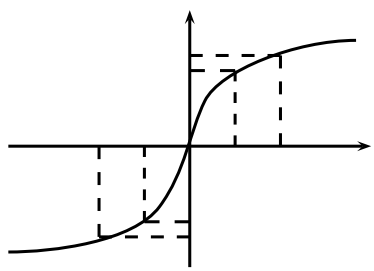
For example, when X is uniform, the constant is 0. Every engineer knows the rule of thumb "6dB per bit"; adding one more quantization bit gets you 6 dB improvement in SNR. However, here we can see that this rule of thumb is valid only in the high rate regime.

21.1.2 Scalar Non-uniform Quantization

Since our source has density P_X, a good idea might be to use more quantization points where P_X is larger, and less where P_X is smaller.



Often the way such quantizers are implemented is to take a monotone function of the source $f(X)$, preform uniform quantization, then take the inverse function: $q(X) = f^{-1}(q_U(f(X)))$. For example, the human ear can detect sounds with volume as small as $0dB$, and a painful, ear-damaging sound occurs around $140dB$. Furthermore, human hearing is very non uniform; it it decently approximated by the log-normal distribution, so to apply a uniform quantizer, we want to take the inverse of this. A popular choice for f is the μ-companding (compression-expanding) function



$$f(X) = \text{sign}(X)\frac{\ln(1+\mu|X|)}{\ln(1+\mu)}$$

This uses more bits for smaller $|X|$'s, e.g. $|X|$'s in the range of human hearing, and less quantization bits outside this region.

21.1.3 Optimal Scalar Quantizers

Now we look for the optimal scalar quantizer given R bits for reconstruction. Formally, this is

$$D_{scalar}(R) = \min_{q:|Im\ q|\leq 2^R} \mathbb{E}|X - q(X)|^2$$

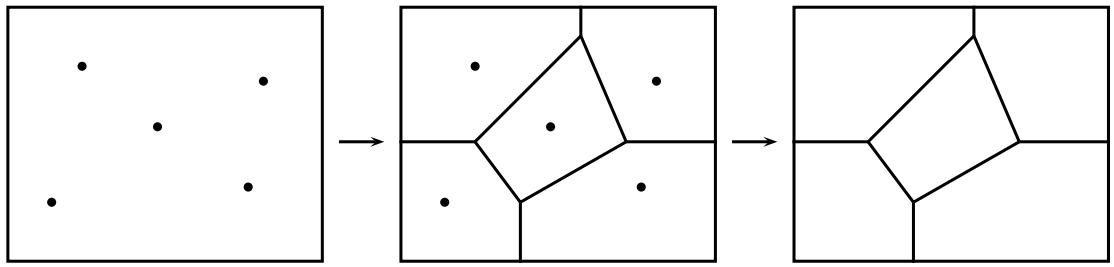
Intuitively, we would think that the optimal quantization regions should be contiguous

$$q(X) = j\mathbf{1}\{T_j \leq X \leq T_{j+1}\}$$

Otherwise, given a point c_j, our reconstruction error will be larger. With ideas like this, in 1982 Stuart Lloyd developed an algorithm (called Lloyd's algorithm) for iteratively finding optimal quantization regions and points. This works for both the scalar and vector cases, and goes as follows:

1. Pick any $N = 2^k$ points
2. Draw the Voronoi regions around the chosen quantization points (aka minimum distance tessellation, or set of points closest to c_j), which forms a partition of the space.
3. Erase the quantization points, choose a new set of quantization points as the centroids ($\mathbb{E}[X|X \in D]$) of each Voronoi region.

4. Repeat.



Steps of Lloyd's algorithm

Lloyd's clever observation is that the centroid of each Voronoi region is (in general) different than the region generated by the quantization point. Therefore, iterating through this procedure gives the *Centroidal Voronoi Tessellation* (CVT - which are very beautiful objects in their own right), which can be viewed as the fixed point of this iterative mapping. The following theorem gives the results about Lloyd's algorithm

Theorem 21.1 (Lloyd)**.**

1. *Lloyd's algorithm always converges to a Centroidal Voronoi Tessellation.*
2. *The optimal quantization strategy is always a CVT.*
3. *CVT's are non-unique, and the algorithm may converge to non-global optima.*

Remark: The third point tells us that Lloyd's algorithm isn't guaranteed to give the optimal quantization strategy. Inf fact, even for $P_X = \mathcal{N}(0, \sigma^2)$, $N > 3$, optimal thresholds are not known in closed form! So it's hard to say too much about optimal quantizers. Because of this, we look for an approximation.

21.1.4 High SNR Approximation

(Panter-Dite '51) Now we look at the high SNR approximation. For this, introduce the distribution $\lambda(x)$, which represents the density of our quantization points. With this, the quantizer associated with q puts $\approx N \int_a^b \lambda(x)dx$ points in any $[a, b]$. Around any point x, the closest point to x will be at distance $\Delta(x)$, so that $N\lambda(x)\Delta(x) \approx 1 \implies \Delta(x) \approx \frac{1}{N\lambda(x)}$. With this approximation, the quality of reconstruction is

$$\begin{aligned}
\mathbb{E}|X - q(X)|^2 &= \sum_{j=1}^{N} \mathbb{E}[|X - c_j|^2 | X \in \Delta_j]\mathbb{P}[X \in \Delta_j] \\
&\approx \sum_{j=1}^{N} \mathbb{P}[X \in \Delta_j]\frac{|\Delta_j|^2}{12} \approx \int f(x)\frac{\Delta^2(x)}{12}dx \\
&= \frac{1}{12N^2}\int f(x)\lambda^{-2}(x)dx
\end{aligned}$$

To find the distribution $\lambda(x)$ that gives the best reconstruction (minimum MSE) when X has distribution $f(x)$, we solve the variational problem

$$\begin{aligned}
&\int f\lambda^{-2}dx - \theta \int \lambda dx \\
&\implies \delta\lambda : \int \left(\frac{f}{\lambda^3} - \text{constant}\right)\delta(x)dx = 0 \\
&\implies \lambda^*(x) = f^{1/3}(x) \cdot \text{constant}
\end{aligned}$$

And therefore

$$D_{scalar}(R) \approx \frac{1}{12}2^{-2R}\left(\int f^{1/3}dx\right)^3$$

So our optimal quantizer density in the high rate regime is proportional to the cube root of the density of our source. This approximation is called the Panter-Dite approximation. For example, when $X \sim \mathcal{N}(0, \sigma^2)$, this is

$$D_{scalar}(R) = \sigma^2 2^{-2R}\frac{\pi\sqrt{3}}{2}$$

21.1.5 Information Theoretic Quantization

By doing vector quantization (namely, compressing $(X_1, \ldots, X_n) \to 2^{nR}$ points), we can achieve

$$D_{vec}(R) = \sigma^2 2^{-2R}$$

Which saves 4.35 dB! Remarkably, even when $X_1, \ldots, X_n$ are iid, we can get better performance than compressing each X_i individually. One instance of this is called the Hamming Game,

Hamming Game: Given 100 bits uniformly chosen, we want to store 50 bits, what is the minimum bit error rate achievable? Naively, one might think that the optimal strategy would be to store half of the bits then guess on the rest, which gives %25 BER. However, as we will show in the next few lectures, the optimal strategy amazingly achieves a BER of %11.

Math Formalism: A lossy compressor is an encoder/decoder pair (f, g) where

$$X \xrightarrow{f} W \xrightarrow{g} \hat{X}$$

- X - continuous source
- W - discrete data
- $\hat{X}$ - reproduction

A *distortion metric* is a function $d : \mathcal{X} \times \hat{\mathcal{X}} \to \mathbb{R}$. There are a few varieties of the lossy compression problem:

1. Fixed length (fixed rate), avg distortion: $W \in [M]$, minimize $\mathcal{E}[d(X, \hat{X})]$.

2. Fixed length, excess distortion: $W \in [M]$, minimize $\mathbb{P}[d(X, \hat{X}) > D]$.

3. Variable length, max distortion: $W \in \{0,1\}^*$, $d(X, \hat{X}) \leq D$ a.s., minimize $\mathbb{E}[length(W)]$ or $H(\hat{X}) = H(W)$.

Definition 21.1. *A source is a sequence of $\mathcal{A}$-valued r.v.'s $(S_1, S_2, \ldots) \in \mathcal{A}$. A separable distortion metric is a distortion metric such that*

$$d(a^n, \hat{a}^n) = \frac{1}{n} \sum d(a_i, \hat{a}_i)$$

i.e. we can look at the per-letter distortion. An (n, M, D)-code is

- *Encoder $f : \mathcal{A}^n \to [M]$*
- *Decoder $g : [M] \to \hat{\mathcal{A}}^n$*
- *With $\mathbb{E}[d(S^n, g(f(S^n)))] \leq D$*

Fundamental limit:

$$M^*(n, D) = \min\{M : \exists (n, M, D) - code\}$$
$$R(D) = \limsup_{n \to \infty} \frac{1}{n} \log M^*(n, D)$$

We will also make the following technical assumption on (S, d).

Assumption 12.1. Assume that (S, d) is such that $D_p < \infty$, where

$$D_p \triangleq \sup_n \min_a E^{1/p}[|d(S^n, a)|^p] < +\infty$$

i.e. that our separable distortion metric d doesn't grow too fast. By Minkowski's inequality, this implies that for memoryless sources we have (single-letter)

$$D_p = \min_a E^{1/p}[|d(S, a)|^p]$$

Now that we have the definition, we give the (surprisingly simple) general converse

Theorem 21.2 (General Converse)**.** *For all lossy codes $X \to W \to \hat{W}$ such that $\mathbb{E}[d(X, \hat{X})] \leq D$, we have*

$$\log M \geq \varphi_X(D) \triangleq \inf_{P_{Y|X} : \mathbb{E}[d(X,Y)] \leq D} I(X; Y)$$

Where $W \in [M]$ is the cardinality of the alphabet.

Proof.

$$\log M \geq H(W) \geq I(X; W) \geq I(X; \hat{X}) \geq \varphi_X(D)$$

Where the last inequality follows from the fact that $P_{\hat{X}|X}$ is a valid choice (by assumption). $\square$

Theorem 21.3 (Properties of φ_X)**.**

1. *φ_X is convex, non-increasing.*
2. *φ_X continuous on (D_0, ∞), where $D_0 = \inf\{D : \varphi_X(D) < \infty\}$.*
3. *If*

$$d(x,y) = \begin{cases} D_0 & x = y \\ > D_0 & x \neq y \end{cases}$$

Then $\varphi_X(D_0) = I(X;X)$.

Proof.

1. Convexity follows from the convexity of $I(P_X, \cdot)$.
2. Continuity on interior of the domain follows from convexity.
3. The only way to satisfy the constraint is to take $X = Y$.

□

In channel coding, we looked at the capacity and the information capacity. We define the *Information Rate-Distortion function* in an analogous way here.

Definition 21.2. *The Information Rate-Distortion function for a source is*

$$R_i(D) = \limsup_{n\to\infty} \frac{1}{n}\varphi_{S^n}(D) \;\; \textit{where} \;\; \varphi_{S^n}(D) = \inf_{P_{\hat{S}^n|S^n}:\mathbb{E}[d(S^n,\hat{S}^n)\leq D} I(S^n;\hat{S}^n)$$

And $D_0 = \inf\{D : R_i(D) < \infty\}$.

Theorem 21.4 (Properties of R_i)**.**

1. *$R_i(D)$ is convex, non-increasing*
2. *$R_i(D)$ is continuous on (D_0, ∞).*
3. *If*

$$d(x,y) = \begin{cases} D_0 & x = y \\ > D_0 & x \neq y \end{cases}$$

Then $R_i(D) = \mathcal{H}$ (entropy rate) or $+\infty$ if S_k is not discrete.

Corollary 21.1. *$\forall D$, $R(D) \geq R_i(D)$.*

The following theorem demonstrates the need of the technical assumption in our definition.

Theorem 21.5 (Excess-to-Average)**.** *Suppose there exists $X \to W \to \hat{X}$ such that $W \in [M]$. If for some $p > 1$ and $x_0 \in \mathcal{X}$,*

$$\mathbb{P}[d(X, \hat{X}) > D] \leq \epsilon \quad \textit{and} \quad (\mathbb{E}[d(X, x_0)]^p)^{1/p} = D_p < \infty$$

Then there exists $X \to W' \to \hat{X}'$ code such that $W' \in [M+1]$, $\mathbb{E}[d(X, \hat{X})] \leq D(1-\epsilon) + D_p \epsilon^{1-1/p}$

Proof. We transform the first code into the second as follows

$$f'(x) = \begin{cases} f(x) & d(x, g(f(x))) \leq D \\ M+1 & o/w \end{cases}$$

$$g'(j) = \begin{cases} g(j) & j \leq M \\ x_0 & j = M+1 \end{cases}$$

Then

$$\begin{aligned} \mathbb{E}[d(X, g' \circ f'(X)) &\leq \mathbb{E}[d(X, \hat{X}) | \hat{W} = M+1](1-\epsilon) + \mathbb{E}[d(X, x_0)\mathbf{1}\{\hat{W} = M+1\}] \\ \text{(Holders Inequality)} \quad &\leq D(1-\epsilon) + D_p \epsilon^{1-1/p} \end{aligned}$$

□

6.441: Information Theory	Spring 2012
Lecturer: Yury Polyanskiy	**Lecture 22: Rate distortion: achievability bounds**

20.1 Goal

Compute $R(D)$.

Recall from the last lecture:

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

and

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

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

$$\log M \ge \varphi_X(D)\,.$$

In this lecture, we will prove the achievability bound and establish the identity $R(D) = R_i(D)$ for memoryless sources (can be extended to "information stable" sources).

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

Theorem 20.1. *For memoryless source S^n,*

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

Proof. By definition we have that $\varphi_{S^n}(D) \le n\varphi_S(D)$ as it allows a larger optimization domain. Thus $R_i(D) \le \varphi_S(D)$.

Typed by Qingqing Huang (qqh@mit.edu), May 7, 2013

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

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

□

20.2 $R(D) \leq R_i(D)$

Theorem 20.2. *For a memoryless source S^n, with the technical assumption that $D_p < \infty$, $\forall D > D_0$ we have:*

$$R(D) = R_i(D) \tag{1}$$

Same techniques in proving (1) for memoryless sources can be applied to prove it for "information stable" sources.

Before diving into the proof, we shall adopt a less rigorous approach to show the intuition behind the proof.

20.2.1 Intuition

The idea is to apply Excess-to-Average Lemma to show the existence of a good code from some simple excess distortion code with ϵ-small excess probability $\mathbb{P}[d(S^n, \hat{S}^n) > D]$.

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

$$\text{encoder}: \ f(s^n) = \operatorname*{argmin}_{c \in \mathcal{C}} d(s^n, c) \tag{2}$$

$$\text{encoder}: \ g(c_j) = c_j \tag{3}$$

We are interested in the excess distortion of this code, define

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

$$
\begin{aligned}
P_{failure} \triangleq & \mathbb{P}[\forall c_i \in \mathcal{C}, d(S^n, c) > D] && (4)\\
\approx & \mathbb{P}[\forall c_i \in \mathcal{C}, d(S^n, c) > D | S^n \in \widetilde{S}^n] && (5)\\
& (\ \widetilde{S}^n \text{ is the set of typical strings with empirical distribution } \hat{P}_{S^n} \approx P_S\ ) \\
= & \mathbb{P}[d(S^n, B^n) > D | S^n \in \widetilde{S}^n]^M \qquad (P_{S^n, B^n} = P_S^n Q_B^n) && (6)\\
= & (1 - \underbrace{\mathbb{P}[d(S^n, B^n) \leq D | S^n \in \widetilde{S}^n]}_{\text{since } S^n \perp B^n, \text{ this should be small}}\ )^M && (7)\\
\approx & 1 - M\mathbb{P}[d(S^n, B^n) \leq D | S^n \in \widetilde{S}^n] && (8)\\
\approx & 1 - M 2^{-nE(Q_B)} \qquad \text{(large deviation!)} && (9)
\end{aligned}
$$

where (we have shown in homework)

$$
E(Q_B) = \min_{P_{B|S}: \mathbb{E}[d(S,B)] \leq D} D(P_{B|S} \| Q_B | P_S) \tag{10}
$$

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

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

We optimize Q_B to get the smallest possible M:

$$
\begin{aligned}
\min_{Q_B} E(Q_B) &= \min_{P_{B|S}: \mathbb{E}[d(S,B)] \leq D} \min_{Q_B} D(P_{B|S} \| Q_B | P_S) \\
&= \min_{P_{B|S}: \mathbb{E}[d(S,B)] \leq D} I(S, B) \\
&= \varphi_S(D)
\end{aligned} \tag{11}
$$

Finally, by the Excess-to-Average Lemma, we know that there exists a code $(S^n \to W' \to \hat{S}^n)$ with $W' \in [M+1]$, and

$$
\begin{aligned}
\mathbb{E}[d(S^n, \hat{S}^n)] \leq & D(1 - P_{failure}) + D_p P_{failure}^{1-1/p} \\
\to & D \text{ as } n \to \infty.
\end{aligned}
$$

therefore $R(D) \leq \liminf_{n\to\infty} \frac{\log(M+1)}{n} = \varphi_S(D) + \delta, \forall \delta > 0$.

20.2.2 Rigorous proof to Theorem 20.2

Theorem 20.3 (Performance bound of excess distortion codes)**.** *$\forall P_{Y|X}$ for $X \to Y$, $\forall \gamma, \exists$ M-code $X \to W \to \hat{X}$, where $W \in [M]$ and*

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

Proof. This theorem says that we can extract a good code from an arbitrary $P_{Y|X}$.

Construct an independent copy $\overline{Y}$:

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

(similar to the idea of introducing unsent codeword $\overline{X}$)

Similar to the steps in the previous intuition, we generate the codewords:

$$\mathcal{C} = \{c_1, \ldots, c_M\} \sim P_{\overline{Y}}^M \perp\!\!\!\perp X$$

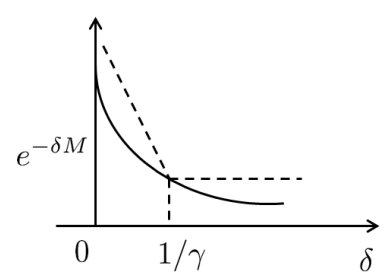
and we adopt the same encoder-decoder pair as in (2, 3). The excess distortion is given by:

$$\begin{aligned}
\mathbb{P}[d(X;\hat{X}) > D] &= \mathbb{P}[\forall j \in [M], d(X;c_j) > D] && (12)\\
&= \mathbb{E}_X[(1 - \underbrace{\mathbb{P}[d(X,\overline{Y}) \le D \Big| X]}_{\text{intuition: exponentially small!}})^M] && (13)\\
&\overset{(1)}{\le} e^{-M/\gamma} + \mathbb{E}_X \left|1 - \gamma \mathbb{P}[d(X,\overline{Y}) \le D \Big| X]\right|^+ && (14)\\
&\overset{(2)}{=} e^{-M/\gamma} + \mathbb{E}_X \left|1 - \gamma \mathbb{E}[2^{-i(X,Y)} \mathbf{1}[d(X,Y) \le D] \Big| X]\right|^+ && (15)\\
&\le e^{-M/\gamma} + \left|1 - \gamma \mathbb{E}[2^{-i(X,Y)} \mathbf{1}[d(X,Y) \le D]]\right|^+ \quad \text{(convexity of } |\cdot|^+) && (16)\\
&\overset{(3)}{\le} e^{-M/\gamma} + \mathbb{P}[i(X,Y) > \log\gamma] + \mathbb{P}[i(X,Y) \le \log\gamma, d(X,Y) > D] && (17)\\
&= e^{-M/\gamma} + \mathbb{P}[\{i(X,Y) > \log\gamma\} \cup \{d(X,Y) > D\}] && (18)
\end{aligned}$$

justifications:

- (1) makes use of the following trick in dealing with $(1-\delta)^M$ for $\delta \ll 1$ and $M \gg 1$:

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



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

- (2) uses the conditionin-unconditioning trick for information density (Lec14)
- (3):

$$\begin{aligned}
&\text{on } i(X,Y) \le \log\gamma, \quad \left|1 - \gamma 2^{-i(X,Y)} \mathbf{1}[d(X,Y) \le D]\right|^+ \le \mathbf{1}[d(X,Y) > D];\\
&\text{on } i(X,Y) > \log\gamma, \quad \left|1 - \gamma 2^{-i(X,Y)} \mathbf{1}[d(X,Y) \le D]\right|^+ \le 1.
\end{aligned}$$

□

Finally, we are able to provide the rigorous proof for Theorem 20.2.

Proof. Take any $P_{\hat{S}|S}$ such that $\mathbb{E}[d(S,\hat{S})] < D - \delta$ (strictly less than). For some $\delta > 0$, apply Theorem 20.3, with

$$\begin{aligned}
&P_X = P_{S^n} \\
&P_{Y|X} = P^n_{\hat{S}|S} \\
&\log M = n(I(S;\hat{S}) + 2\delta) \\
&\log \gamma = n(I(S;\hat{S}) + \delta)
\end{aligned}$$

we have that there exists a (n, M) code $X \to W \to \hat{X}$ such that:

$$\mathbb{P}[d(X,\hat{X}) > D - \frac{\delta}{2}] \leq \underbrace{e^{-e^{n\delta}}}_{\to 0} + \underbrace{\mathbb{P}(\sum_{j=1}^{n} d(S_j,\hat{S}_j) > n(D - \frac{\delta}{2}))}_{\to 0 \text{ WLLN and assumption}} + \underbrace{\mathbb{P}(\sum_{j=1}^{n} i(S_j,\hat{S}_j) > n(I(S;\hat{S}) + \delta)}_{\to 0 \text{ WLLN}}.$$

By Excess-to-Average Lemma, there exists a $(n, M, \overline{D})$ code with:

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

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

$$\begin{aligned}
&R(D) \leq I(S;\hat{S}) \\
\Rightarrow &R(D) \leq \varphi_S(D-) = \varphi_S(D). \quad (\text{since } D > D_0)
\end{aligned}$$

□

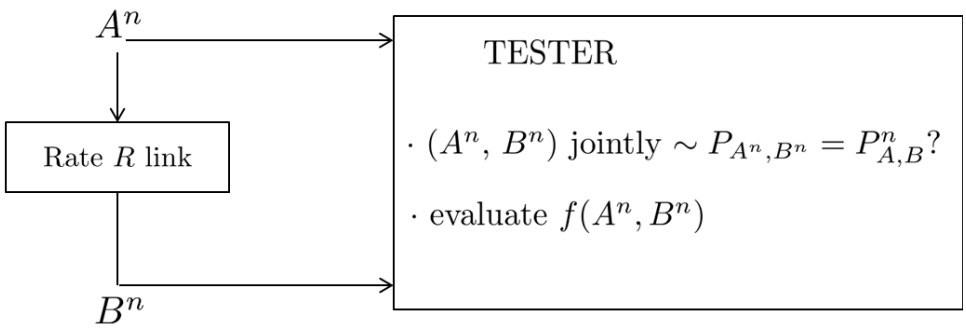
Note: It seems that in 20.2.1 and in Theorem 20.2 we applied different relaxations in showing the lower bound, how come they turn out to yield the same *tight* asymptotic result?

This is because the key to both of the proofs is to estimate the exponent (large deviations) of the underlined probabilities in (7) and (13), respectively. To get the right exponent, as we know, the key is to apply tilting (change of measure) to the distribution solving the information projection problem (10). In the case, when $P_{\bar{Y}} = (Q_B)^n = (P_{\hat{S}})^n$ is chosen as the solution to rate-distortion optimization $\inf I(S;\hat{S})$, the resulting tilting is precisely given by $2^{-i(X;Y)}$.

20.3 Covering lemma

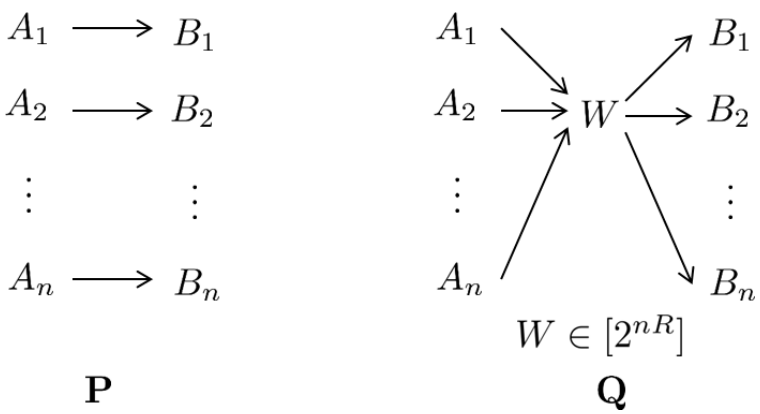
Goal:

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



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

In other words:



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

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

what's the minimum rate R to achieve this?

In other words:

for any realization of sequence A^n, what's the minimum amount of information we need in order to generate a sequence B^n such that joint composition can well approximate the joint distribution $P^n_{A,B}$?

Ideas:

1. If $R = H(A)$, we can compress A^n and send it to "B side", and that information to produce B^n.
2. If $R = H(B)$, "A side" can generate B^n according to $P^n_{A,B}$ and send that B^n sequence to "B side".
3. If $A \perp B$, we know that $R = 0$, as "B side" can generate B^n independently.

In general, $R = I(A, B)$ is sufficient to fool any permutation invariant tester. This is formalized in the following Covering Lemma:

Theorem 20.4 (Covering Lemma)**.** *$\forall P_{A,B}$ and $R > I(A;B)$, let $\mathcal{C} = \{c_1, \ldots, c_M\}$ where each codeword c_j is i.i.d. drawn from distribution P_B^n. $\forall \epsilon > 0$, for $M \geq 2^{n(I(A,B)+\epsilon)}$ we have that:*

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

Stronger form: $\forall F$

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

Proof. Following similar arguments of the proof for Theorem 20.3, we have

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

□

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

6.441: Information Theory	Spring 2013
Lecturer: Yury Polyanskiy	**Lecture 23: Lossy Joint Source Channel Coding**

Last time: For memoryless sources, $R(D) = R_i(D)$.

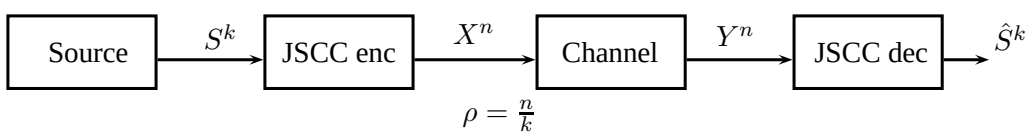
23.1 Lossy Joint Source Channel Coding

The *lossy joint source channel coding problem* refers to the fundamental limits of lossy compression followed by transmission over a channel.

Problem Setup: For a source described by $(\{S_1, S_2, \dots\}, \mathcal{A}, \hat{\mathcal{A}}, d : \mathcal{A} \times \hat{\mathcal{A}} \to \mathbb{R})$, a lossy JSCC is a "hook up" (f, g) such that

$$S^k \xrightarrow{f} X^n \xrightarrow{\text{ch.}} Y^n \xrightarrow{g} \hat{S}^k$$

Definition 23.1. *(f, g) is a (k, n, D)-JSCC if $\mathbb{E}[d(S^k, \hat{S}^k)] \le D$.*



Where $\rho = n/k$ is called the bandwidth expansion factor. The asymptotic fundamental limit for a lossy JSCC is

$$\rho^*(D) = \limsup_{n\to\infty} \min\{\frac{n}{k} : \exists (k, n, D) - code\}$$

The converse for the JSCC is quite simple. Note that since there is no ϵ under consideration, the strong converse is the same as the weak converse.

Theorem 23.1 (Converse)**.** *For any source such that*

$$R_i(D) = \lim_{k\to\infty} \frac{1}{k} \inf_{P_{\hat{S}^k|S^k} : \mathbb{E}[d(S^k, \hat{S}^k)] \le D} I(S^k; \hat{S}^k)$$

We have

$$\rho^*(D) \ge \frac{R_i(D)}{C_i}$$

Remark: The requirement of this theorem on the source isn't too stringent; the limit expression for $R_i(D)$ typically exists for stationary sources (like for the entropy rate)

Typed by Austin Collins, austinc@mit.edu, May 9nd, 2013

Proof. Take a (k, n, D)-code $S^k \to X^n \to Y^n \to \hat{S}^k$. Then

$$\inf_{P_{\hat{S}^k|S^k}} I(S^k; \hat{S}^k) \le I(S^k; \hat{S}^k) \le I(X^k; Y^k) \le \sup_{P_{X^n}} I(X^n; Y^n)$$

Which follows from data processing and taking inf/sup. Normalizing by $1/k$ and taking the liminf as $n \to \infty$

$$\text{(LHS)} \quad \liminf_{n\to\infty} \frac{1}{n} \sup_{P_{X^n}} I(X^n; Y^n) = C_i$$

$$\text{(RHS)} \quad \liminf_{n\to\infty} \frac{1}{k_n} \inf_{P_{\hat{S}^{k_n}|S^{k_n}}} I(S^{k_n}; \hat{S}^{k_n}) = R_i(D)$$

And therefore, any sequence of (k_n, n, D)-codes satisfies

$$\limsup_{n\to\infty} \frac{n}{k_n} \ge \frac{R_i(D)}{C_i}$$

□

We now give the achievability side for memoryless sources, and upgrade from information rate distortion and capacity to $R(D)/C$. The proof strategy is similar to the non-lossy JSCC: we construct a separated lossy compression and channel coding scheme using our tools from those areas, then show that this separated construction is optimal.

Theorem 23.2. *For any memoryless source $(P_S, \mathcal{A}, \hat{\mathcal{A}}, d)$ satisfying assumption A1 (below), and for any memoryless channel $P_{Y|X}$,*

$$\rho^*(D) = \frac{R(D)}{C}$$

Proof. One direction, $\rho^*(D) \ge R(D)/C$ follows from the previous theorem. For the other direction, we constructed a separated compression / channel coding scheme. Take

$S^k \to W \to \hat{S}^k$ compressor to $W \in [2^{kR(D)+o(k)}]$ with $\mathbb{E}[d(S^k, \hat{S}^k)] \le D$

$W \to X^n \to Y^n \to \hat{W}$ max probability of error channel code (assuming $kR(D) \le nC + o(n)$) with $\mathbb{P}[W \ne \hat{W}] \le \epsilon \ \forall P_W$

So that the overall system is

$$S^k \longrightarrow W \longrightarrow X^n \longrightarrow Y^n \longrightarrow \hat{W} \longrightarrow \hat{S}^k$$

Note that here we need a maximum probability of error code since when we "hook up" these two schemes, W at the input of the channel is not guaranteed to be uniform. Now that we have a scheme, we must analyze the average distortion to show that it meets the distortion constraint. We start by splitting the expression into two cases

$$\mathbb{E}[d(S^k, \hat{S}^k)] = \mathbb{E}[d(S^k, \hat{S}^k(W))\mathbf{1}\{W = \hat{W}\}] + \mathbb{E}[d(S^k, \hat{S}^k(\hat{W}))\mathbf{1}\{W \ne \hat{W}\}]$$

By assumption on our lossy code, we know that the first term is $\le D$. In the second term, we know that the probability of the event $\{W \ne \hat{W}\}$ is small by assumption on our channel code, but we can't say anything about $\mathbb{E}[d(S^k, \hat{S}^k(\hat{W}))]$. But under assumption A1, by the lemma (below), $\exists$ code $S^k \to W \to \hat{S}^k$ such that

1. $\mathbb{E}[d(S^k, \hat{S}^k)] \leq D$ holds
2. $d(a_0^k, \hat{S}^k) \leq L$ for all quantization outputs $\hat{S}^k$, where $a_0^k = (a_0, \ldots, a_0)$ is some fixed string of length k from assumption A1.

The second bullet says that all points in the reconstruction space are "close" to some fixed string. Now we can deal with the troublesome term

$$\begin{aligned}
\mathbb{E}[d(S^k, \hat{S}^k(\hat{W}))\mathbf{1}\{W \neq \hat{W}\}] &\leq \mathbb{E}[\mathbf{1}\{W \neq \hat{W}\}\lambda(d(S^k, \hat{a}_0^k) + d(a_0^k, \hat{S}^k))] \\
\text{(by point (2) above)} \quad &\leq \lambda\mathbb{E}[\mathbf{1}\{W \neq \hat{W}\}d(S^k, \hat{a}_0^k)] + \lambda\mathbb{E}[\mathbf{1}\{W \neq \hat{W}\}L] \\
\text{(Holder's Inequality)} \quad &\leq \lambda D_p \epsilon^{1-1/p} + \lambda L\epsilon \to 0 \text{ as } \epsilon \to 0
\end{aligned}$$

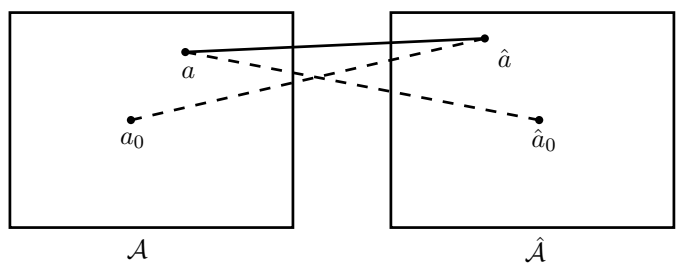
Therefore our scheme meets the average distortion constraint. Hence we conclude that for $\forall \rho > \frac{R(D)}{C}$, $\exists$ sequence of $(k, n, D + o(1))$-codes. $\square$

The following assumption was critical to the previous theorem:

Assumption A1: For a source $(P_S, \mathcal{A}, \hat{\mathcal{A}}, d)$, $\exists \lambda \geq 0, a_0 \in \mathcal{A}, \hat{a}_0 \in \hat{\mathcal{A}}$ such that

1. $d(a, \hat{a}) \leq \lambda(d(a, \hat{a}_o) + d(a_0, \hat{a}))$ $\forall a, \hat{a}$ (generalized triangle inequality)
2. $\mathbb{E}[d(S, \hat{a}_0)^p]^{1/p} \triangleq D_p < \infty$
3. $\mathbb{E}[d(a_0, \hat{S})] < \infty$ for any rate-distortion output achieving $P_{\hat{S}}$.

This assumption says that the spaces $\mathcal{A}$ and $\hat{\mathcal{A}}$ have "nice centers", in the sense that the distance between any two points is upper bounded by a constant times the distance from the centers to each point (see figure below).



But the assumption isn't easy to verify, or clear which sources satisfy the assumptions. Because of this, we now give a few sufficient conditions for Assumption A1 to hold.

Trivial Condition: If $|\mathcal{A}| < \infty$ and $d(a, \hat{a}) \in \mathcal{R}$, then A1 holds. I.e. when we have a discrete source with a real valued distortion function, A1 holds.

Theorem 23.3. *If $d(a, \hat{a}) = \rho^q(a, \hat{a})$ for some metric ρ with $q \geq 1$, and $D_p < \infty$, then A1 holds.*

Proof. Recall D_p was defined as $\mathbb{E}[d(S,\hat{a}_0)^p]^{1/p} < \infty$. Take $a_0 = \hat{a}_0$ that achieves finite D_p (in fact, any points can serve as centers in a metric space). Then

$$\begin{aligned}
(\frac{1}{2}\rho(a,\hat{a}))^q &\leq \left(\frac{1}{2}\rho(a,a_0) + \frac{1}{2}\rho(a_0,\hat{a})\right)^q \\
\text{(Jensen's)} \quad &\leq \frac{1}{2}\rho^q(a,a_0) + \frac{1}{2}\rho^q(a_0,\hat{a})
\end{aligned}$$

And thus $d(a,\hat{a}) \leq 2^{q-1}(d(a,a_0) + d(a_0,\hat{a}))$. Taking $\lambda = 2^{q-1}$ verifies (1) and (2) in A1. To verify (3), we can use this generalized triangle inequality for our source

$$d(a_0,\hat{S}) \leq 2^{q-1}(d(a_0,S) + d(S,\hat{S}))$$

Then taking the expectation of both sides gives

$$\begin{aligned}
\mathbb{E}[d(a_0,\hat{S})] &\leq 2^{q-1}(\mathbb{E}[d(a_0,S)] + \mathbb{E}[d(S,\hat{S})]) \\
&\leq 2^{q-1}(D_p + D) < \infty
\end{aligned}$$

So that point (3) in A1 holds. □

So we see that metrics raised to powers (e.g. squared Euclidean norm) satisfy the condition A1. The lemma used in Theorem 23.2 is now given.

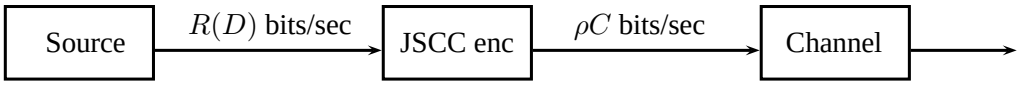
Lemma 23.1. *Take $\mathcal{X} = A^k$, $\hat{\mathcal{X}} = \hat{\mathcal{A}}^k$, $P_{Y|X} : \mathcal{X} \to \hat{\mathcal{X}}$, and $P_{Y|X} = (P_{\hat{S}|S})^k$. Under assumption A1, $\forall R > R(D), L > \mathbb{E}[d(a_0,\hat{S})]$, $\exists(n, 2^{nR+o(a)}, D)$ code such that $d(a_o^k, \hat{X}) \leq L$ for any reconstruction level $\hat{X}$.*

Proof. Apply the achievability bound from Theorem 20.3 with

$$d_1(x,\hat{x}) = \begin{cases} d(x,\hat{x}) & if\ d(a_0^k,\hat{x}) \leq L \\ +\infty & o/w \end{cases}$$

□

To summarize the results in this section; we have shown that the optimal lossy compression scheme for transmission over a channel is the separated one: first compress the data, then encode it using your favorite channel code, then decompress at the receiver.



23.2 Evaluation of $R(D)$

So far we've proved some properties about the rate distortion function, now we'll compute its value for a few simple statistical sources. We'll do this in a somewhat unsatisfying way: guess the answer, then verify its correctness. At the end, we'll show that there is a pattern behind this method.

23.2.1 Hamming Source

Let $S \sim \text{Ber}(p)$, $p \leq 1/2$, with distortion $d(S, \hat{S}) = \mathbf{1}\{S \neq \hat{S}\}$ and alphabets $\mathcal{A} = \hat{\mathcal{A}} = \{0, 1\}$.

Claim: $R(D) = |h(p) - h(D)|^+$

Proof. (Achievability) We're free to choose any $P_{\hat{S}|S}$, so choose $S = \hat{S} + Z$ (cf. BSC(D)), where $\hat{S} \sim \text{Ber}(p')$ and $Z \sim \text{Ber}(D)$, and p' is such that $p'(1-D) + (1-p')D = p$ so that $p' < p$. Then,

$$I(S; \hat{S}) = H(S) - H(S|\hat{S}) = h(p) - h(D)$$

Since one such $P_{\hat{S}|S}$ exists, we have the upper bound $R(D) \leq h(p) - h(D)$.

(Converse) Now we need to show that there is no better choice $Q_{\hat{S}|S}$ such that $\mathbb{E}_Q[d(S, \hat{S})] \leq D$. Since any $I(P_S, Q_{\hat{S}|S})$ is an upper bound for $R(D)$, we have

$$\begin{aligned} R(D) &\leq I(P_S, Q_{\hat{S}|S}) = D(Q_{S|\hat{S}} || P_S | Q_{\hat{S}}) \\ &= D(Q_{S|\hat{S}} || P_{S|\hat{S}} | Q_{\hat{S}}) + \mathbb{E}_Q \left[\log \frac{P_{S|\hat{S}}}{P_S} \right] \\ \text{(Marginal } Q_{S\hat{S}} = P_S Q_{\hat{S}|S}) \quad &= D(Q_{S|\hat{S}} || P_{S|\hat{S}} | Q_{\hat{S}}) + H(S) + \mathbb{E}_Q[\log D \mathbf{1}\{S \neq \hat{S}\} + \log \bar{D} \mathbf{1}\{S = \hat{S}\}] \end{aligned}$$

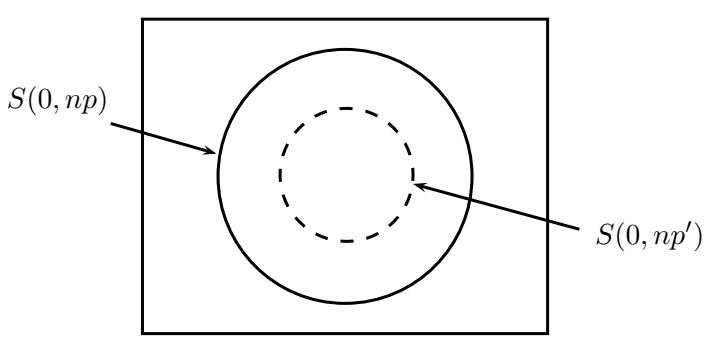
And we can minimize this expression by taking $Q_{S|\hat{S}} = P_{S|\hat{S}}$, giving

$$\geq 0 + H(S) + Q[S = \hat{S}] \log \bar{D} + Q[S \neq \hat{S}] \log D \geq h(p) - h(D) \tag{1}$$

Since the upper and lower bound agree, we have $R(D) = |h(p) - h(D)|^+$. □

For example, when $p = 1/2$, $D = .11$, then $R(D) = 1/2$ bit. In the Hamming game where we compressed 100 bits down to 50, we indeed can do this while achieving %11 average distortion, compared to the naive scheme of storing half the string and guessing on the other half, which achieves %25 average distortion.

Interpretation: The distribution $P_S^n = \text{Ber}(p)^n$ concentrates on the Hamming sphere of radius np as n grows large. The above result about Hamming sources tells us that the optimal reconstruction points are from $P_{\hat{S}}^n = \text{Ber}(p')^n$ where $p' < p$, which concentrates on a sphere of radius np' (note the reconstruction points are some exponentially small subset of this sphere).



Hamming Spheres

It is interesting to note that *none* of the reconstruction points are the same as any of the possible source values (with high probability).

23.2.2 Gaussian Source

The Gaussian source is defines as $\mathcal{A} = \hat{\mathcal{A}} = \mathbb{R}$, $S \sim \mathcal{N}(0, \sigma^2)$, $d(a, \hat{a}) = |a - \hat{a}|^2$.

Claim: $R(D) = \frac{1}{2} \log^+ \frac{\sigma^2}{D}$

Proof. (Achievability) Choose $S = \hat{S} + Z$ (cf. AWGN with noise power D), where $\hat{S} \sim \mathcal{N}(0, \sigma^2 - D)$ and $Z \sim \mathcal{N}(0, D)$. Then

$$I(S; \hat{S}) = \frac{1}{2} \log \frac{\sigma^2}{D} \implies R(D) \leq \frac{1}{2} \log \frac{\sigma^2}{D}$$

(Converse) Let $Q_{\hat{S}|S}$ be any conditional distribution such that $\mathbb{E}_Q |S - \hat{S}|^2 \leq D$. We use the same trick as before

$$R(D) \leq I(P_S, Q_{\hat{S}|S}) = D(Q_{S|\hat{S}} || P_{S|\hat{S}} | Q_{\hat{S}}) + \mathbb{E}_Q \left[\log \frac{P_{S|\hat{S}}}{P_S} \right]$$

Minimizing this by taking $Q_{\hat{S}|S} = P_{S|\hat{S}}$ gives

$$\geq \mathbb{E}_Q \log \frac{\frac{1}{\sqrt{2\pi D}} e^{-\frac{|S-\hat{S}|^2}{2D}}}{\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{S^2}{2\sigma^2}}} = \frac{1}{2} \log \frac{\sigma^2}{D} + \frac{\log e}{2} \mathbb{E}_Q \left[\frac{S^2}{\sigma^2} - \frac{|S - \hat{S}|^2}{D} \right] \geq \frac{1}{2} \log \frac{\sigma^2}{D}$$

Again, the upper and lower bounds agree. □

The interpretation in the Gaussian case is very similar to the case of the Hamming source. As n grows large, our source distribution concentrates on $S(0, \sqrt{n\sigma^2})$ (n-sphere in Euclidean space rather than Hamming), and our reconstruction points on $S(0, \sqrt{n(\sigma^2 - D)})$. So again the picture is two nested sphere.

Remark: The computation of rate distortion functions, and rate distortion theory in general, has played a significant role in pure mathematics. For instance, Hilbert's thirteenth problem was partially solved by Arnold and Kolomogorov after they realized that they could classify spaces of functions looking at the optimal quantizer for such functions.

23.2.3 General Case

In the computation of $R(D)$ for the Hamming and Gaussian source, we guessed the correct form of the rate distortion function. In both of their converse arguments, we used the same trick to establish that any other $Q_{\hat{S}|S}$ gave a larger value for $R(D)$. In this section, we formalize this trick, in an analogous manner to the saddle point property of the channel capacity. Note that typically we don't need any tricks to compute $R(D)$, since we can obtain a solution in parametric form to the unconstrained convex optimization

$$\min_{P_{\hat{S}|S}} I(S; \hat{S}) + \lambda \, \mathbb{E}[d(S, \hat{S})]$$

However, for peace of mind it is good to know there are some general reasons why tricks like we used in Hamming/Gaussian actually are guaranteed to work.

Theorem 23.4. *1. Suppose P_{Y^*} and $P_{X|Y^*} \ll P_X$ are found with the property that $\mathbb{E}[d(X,Y^*)] \le D$ and for any P_{XY} with $\mathbb{E}[d(X,Y)] \le D$ we have*

$$\mathbb{E}\left[\log \frac{dP_{X|Y^*}}{dP_X}(X|Y)\right] \ge I(X;Y^*)\,. \tag{2}$$

Then $R(D) = I(X;Y^*)$.

2. Suppose that $I(X;Y^*) = R(D)$. Then for any regular branch of conditional probability $P_{X|Y^*}$ and for any P_{XY} satisfying

- *$\mathbb{E}[d(X,Y)] \le D$ and*
- *$P_Y \ll P_{Y^*}$ and*
- *$I(X;Y) < \infty$*

the inequality (2) holds.

Remarks:

1. The first part is a sufficient condition for optimality of a given P_{XY^*}. The second part gives a necessary condition that is convenient to narrow down the search. Indeed, typically the set of P_{XY} satisfying those conditions is rich enough to infer from (2):

$$\log \frac{dP_{X|Y^*}}{dP_X}(x|y) = R(D) - \theta[d(x,y) - D]\,,$$

   for a positive $\theta > 0$.

2. Note that the second part is not valid without $P_Y \ll P_{Y^*}$ condition. The counter-example to this and various other erroneous (but frequently encountered) generalizations is the following: $\mathcal{A} = \{0,1\}$, $P_X = \mathrm{Bern}(1/2)$, $\hat{\mathcal{A}} = \{0,1,0',1'\}$ and

$$d(0,0) = d(0,0') = 1 - d(0,1) = 1 - d(0,1') = 0\,.$$

   The $R(D) = |1 - h(D)|^+$, but there are a bunch of non-equivalent optimal $P_{Y|X}$, $P_{X|Y}$ and P_Y's.

Proof. First part is just a repetition of the proofs above, so we focus on part 2. Suppose there exists a counter-example P_{XY} achieving

$$I_1 = \mathbb{E}\left[\log \frac{dP_{X|Y^*}}{dP_X}(X|Y)\right] < I^* = R(D)\,.$$

Notice that whenever $I(X;Y) < \infty$ we have

$$I_1 = I(X;Y) - D(P_{X|Y}||P_{X|Y^*}|P_Y)\,,$$

and thus

$$D(P_{X|Y}||P_{X|Y^*}|P_Y) < \infty\,. \tag{3}$$

For every $\lambda \in [0,1]$ define

$$\rho_1(y) \triangleq \frac{dP_Y}{dP_{Y^*}}(y) \tag{4}$$

$$\lambda(y) \triangleq \frac{\lambda\rho_1(y)}{\lambda\rho_1(y)+\bar{\lambda}} \tag{5}$$

$$P^{(\lambda)}_{X|Y=y} = \lambda(y)P_{X|Y=y} + \bar{\lambda}(y)P_{X|Y^*=y} \tag{6}$$

$$dP_{Y_\lambda} = \lambda dP_Y + \bar{\lambda}dP_{Y^*} = (\lambda\rho_1(y)+\bar{\lambda})dP_{Y^*} \tag{7}$$

$$D(y) = D(P_{X|Y=y}||P_{X|Y^*=y}) \tag{8}$$

$$D_\lambda(y) = D(P^{(\lambda)}_{X|Y=y}||P_{X|Y^*=y})\,. \tag{9}$$

Notice:

$$\text{On } \{\rho_1 = 0\}: \quad \lambda(y) = D(y) = D_\lambda(y) = 0$$

and otherwise $\lambda(y) > 0$. By convexity of divergence

$$D_\lambda(y) \le \lambda(y)D(y)$$

and therefore

$$\frac{1}{\lambda(y)}D_\lambda(y)1\{\rho_1(y)>0\} \le D(y)1\{\rho_1(y)>0\}\,.$$

Notice that by (3) the function $\rho_1(y)D(y)$ is non-negative and P_{Y^*}-integrable. Then, applying dominated convergence theorem we get

$$\lim_{\lambda\to 0}\int_{\{\rho_1>0\}} dP_{Y^*}\frac{1}{\lambda(y)}D_\lambda(y)\rho_1(y) = \int_{\{\rho_1>0\}} dP_{Y^*}\rho_1(y)\lim_{\lambda\to 0}\frac{1}{\lambda(y)}D_\lambda(y) = 0 \tag{10}$$

where in the last step we applied the result from Lecture 4

$$D(P||Q) < \infty \qquad \Longrightarrow \qquad D(\lambda P + \bar{\lambda}Q||Q) = o(\lambda)$$

since for each y on the set $\{\rho_1 > 0\}$ we have $\lambda(y) \to 0$ as $\lambda \to 0$.

On the other hand, notice that

$$\int_{\{\rho_1>0\}} dP_{Y^*}\frac{1}{\lambda(y)}D_\lambda(y)\rho_1(y)1\{\rho_1(y)>0\} = \frac{1}{\lambda}\int_{\{\rho_1>0\}} dP_{Y^*}(\lambda\rho_1(y)+\bar{\lambda})D_\lambda(y) \tag{11}$$

$$= \frac{1}{\lambda}\int_{\{\rho_1>0\}} dP_{Y_\lambda}D_\lambda(y) \tag{12}$$

$$= \frac{1}{\lambda}\int_{\mathcal{Y}} dP_{Y_\lambda}D_\lambda(y) = \frac{1}{\lambda}D(P^{(\lambda)}_{X|Y}||P_{X|Y^*}|P_{Y_\lambda})\,, \tag{13}$$

where in the penultimate step we used $D_\lambda(y) = 0$ on $\{\rho_1 = 0\}$. Hence, (10) shows

$$D(P^{(\lambda)}_{X|Y}||P_{X|Y^*}|P_{Y_\lambda}) = o(\lambda)\,, \qquad \lambda \to 0\,.$$

Finally, since

$$P^{(\lambda)}_{X|Y} \circ P_{Y_\lambda} = P_X\,,$$

we have

$$I(X;Y_\lambda) = D(P^{(\lambda)}_{X|Y}||P_{X|Y^*}|P_{Y_\lambda}) + \lambda\,\mathbb{E}\left[\log\frac{dP_{X|Y^*}}{dP_X}(X|Y)\right] + \bar{\lambda}\,\mathbb{E}\left[\log\frac{dP_{X|Y^*}}{dP_X}(X|Y^*)\right] \tag{14}$$

$$= I^* + \lambda(I_1 - I^*) + o(\lambda)\,, \tag{15}$$

contradicting the assumption

$$I(X;Y_\lambda) \geq I^* = R(D)\,.$$

□

6.441: Information Theory	Spring 2013
Lecturer: Yury Polyanskiy	**Lecture 24: Multi-access channel**

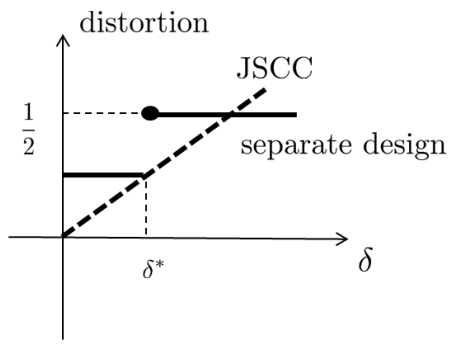
Note: Open questions in the topics of classical information theory:

- lossless compression: we can apply the standard results in compression of a text file, but it is extremely difficult for image files due to the strong spatial correlation.
- JSCC: Asymptotically the separation principle sounds good, but the separated systems can be very unstable - no graceful degradation. Consider the following example of JSCC.

  Example: Source $= Bern(\frac{1}{2})$, channel $= BSC(\delta)$.

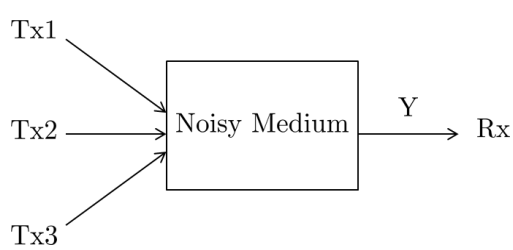
  1. separate compressor and channel encoder designed for $\frac{R(D)}{C(\delta)} = 1$
  2. a simple JSCC:

$$\rho = 1, X_i = S_i$$



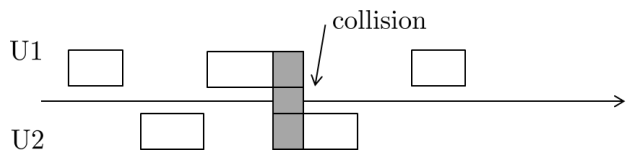
no graceful degradation of separately designed source channel code

24.1 MAC



Note: In network community, people are mostly interested in channel access control mechanisms that help to detect or avoid data packet collisions so that the channel is shared among multiple users.

Typed by Qingqing Huang (qqh@mit.edu), May 14, 2013



The famous ALOHA protocal achieves

$$\sum_i R_i \approx 0.37C$$

where C is the (single-user) capacity of the channel.

In information theory community, the goal is to achieve

$$\sum_i R_i > C$$

The key to achieve this is to use coding so that collisions are resolvable.

In the following discussion we shall focus on the case with two users. This is without loss of much generality, as all the results can easily be extended to N users.

Definition 24.1.

- *Multi-access channel:* $\{P_{Y^n|A^n,B^n} : \mathcal{A}^n \times \mathcal{B}^n \to \mathcal{Y}^n, n = 1, 2, \dots\}$.
- *a* (n, M_1, M_2, ϵ) *code is specified by*

$$\begin{aligned} &f_1 : [M_1] \to \mathcal{A}^n, \quad f_2 : [M_2] \to \mathcal{B}^n \\ &g : \mathcal{Y}^n \to [M_1] \times [M_2] \end{aligned}$$

$$\begin{tikzcd}
W_1 \arrow[r, "f_1"] & A^n \arrow[dr] & & \hat{W}_1 \\
& & Y^n \arrow[ur, "g"] \arrow[dr] & \\
W_2 \arrow[r, "f_2"] & B^n \arrow[ur] & & \hat{W}_2
\end{tikzcd}$$

$$P$$

  $W_1, W_2 \sim$ *uniform, and the codes achieves*

$$\mathbb{P}[\{W_1 \neq \hat{W}_1\} \bigcup \{W_2 \neq \hat{W}_2\}] \leq \epsilon$$

- *Fundamental limit of capacity region*

$$\mathcal{R}^*(n, \epsilon) = \{(R_1, R_2) : \exists \text{ a } (n, 2^{nR_1}, 2^{nR_2}, \epsilon) \text{ code}\}$$

- *Asymptotoics:*

$$\mathcal{C}_\epsilon = \left[\liminf_{n\to\infty} \mathcal{R}^*(n, \epsilon)\right]$$

  where $[\cdot]$ *denotes the closure of a set.*

Note: $\liminf$ *and* $\limsup$ *of a sequence of sets* $\{A_n\}$*:*

$$\liminf_n A_n = \{\omega : \omega \in A_n, \forall n \geq n_0\}$$
$$\limsup_n A_n = \{\omega : \omega \textit{ infinitely occur}\}$$

- $$\mathcal{C} = \lim \mathcal{C}_\epsilon = \bigcap_{\epsilon > 0} \mathcal{C}_\epsilon$$

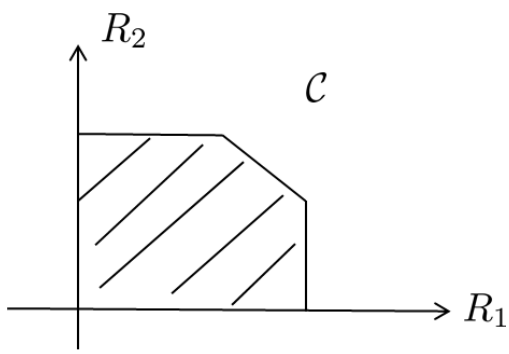
Theorem 24.1 (Capacity region)**.**

$$\mathcal{C}_\epsilon = \overline{co} \bigcup_{P_A, P_B} Penta(P_A, P_B) \tag{1}$$
$$= \Big[\bigcup_{P_{U,A,B} = P_U P_{A|U} P_{B|U}} Penta(P_{A|U}, P_{B|U} | P_U) \Big] \tag{2}$$

where $\overline{co}$ *is the set operator of constructing the convex hull followed by taking the closure, and* $Penta(\cdot, \cdot)$ *is defined as follows:*

$$Penta(P_A, P_B) = \left\{ (R_1, R_2) : \begin{array}{c} 0 \leq R_1 \leq I(A; Y|B) \\ 0 \leq R_2 \leq I(B; Y|A) \\ R_1 + R_2 \leq I(A, B; Y) \end{array} \right\}$$
$$Penta(P_{A|U}, P_{B|U} | P_u) = \left\{ (R_1, R_2) : \begin{array}{c} 0 \leq R_1 \leq I(A; Y|B, U) \\ 0 \leq R_2 \leq I(B; Y|A, U) \\ R_1 + R_2 \leq I(A, B; Y|U) \end{array} \right\}$$

Note: the two forms in (1) and (2) are equivalent without cost constraints. In the case when constraints such as $\mathbb{E}\mathsf{c}_1(A) \leq P_1$ and $\mathbb{E}\mathsf{c}_2(B) \leq P_2$ are present, only the second expression yields the true capacity region.



Penta

24.2 MAC achievability bound

First, we introduce a lemma which will be used in the proof of Theorem 24.1.

Lemma 24.1. *$\forall P_A, P_B, P_{Y|A,B}$ such that $P_{A,B,Y} = P_A P_B P_{Y|A,B}$, and $\forall \gamma_1, \gamma_2, \gamma_{12} > 0$, $\forall M_1, M_2$, there exists a (M_1, M_2, ϵ) MAC code such that:*

$$\begin{aligned}\epsilon \leq & \mathbb{P}\Big[\{i_{12}(A,B;Y) \leq \log\gamma_{12}\} \bigcup \{i_1(A;Y|B) \leq \log\gamma_1\} \bigcup \{i_2(B;Y|A) \leq \log\gamma_2\}\Big] \\ & + (M_1-1)(M_2-1)e^{-\gamma_{12}} + (M_1-1)e^{-\gamma_1} + (M_2-1)e^{-\gamma_2} \end{aligned} \tag{3}$$

Proof. We again use the idea of random coding.

Generate the codebooks

$$c_1, \ldots, c_{M_1} \in \mathcal{A}, \quad d_1, \ldots, d_{M_2} \in \mathcal{B}$$

where the codes are drawn i.i.d from distribuions: $c_1, \ldots, c_{M_1} \sim$ *i.i.d.* P_A, $d_1, \ldots, d_{M_2} \sim$ *i.i.d.* P_B.

The decoder operates in the following way: report (m, m') if it is the unique pair that satisfies:

$$\begin{aligned} (P_{12}) \quad & i_{12}(c_m, d_{m'}; y) > \log\gamma_{12} \\ (P_1) \quad & i_1(c_m; y|d_{m'}) > \log\gamma_1 \\ (P_2) \quad & i_2(d_{m'}; y|c_m) > \log\gamma_2 \end{aligned}$$

Evaluate the expected error probability:

$$\begin{aligned}\mathbb{E}P_e(c_1^{M_1}, d_1^{M_2}) = \mathbb{P}\Big[& \{(W_1, W_2) \text{ violate } (P_{12}) \text{ or } (P_1) \text{ or } (P_2)\} \\ & \bigcup \{\exists \text{ impostor } (W_1', W_2') \text{ that satisfy } (P_{12}) \text{ and } (P_1) \text{ and } (P_2)\}\Big]\end{aligned}$$

by symmetry of random codes, we have

$$\begin{aligned}P_e = \mathbb{E}[P_e|W_1 = m, W_2 = m'] = \mathbb{P}\Big[& \{(m, m') \text{ violate } (P_{12}) \text{ or } (P_1) \text{ or } (P_2)\} \\ & \bigcup \{\exists \text{ impostor } (i \neq m, j \neq m') \text{ that satisfy } (P_{12}) \text{ and } (P_1) \text{ and } (P_2)\}\Big]\end{aligned}$$

$$\Rightarrow P_e \leq \mathbb{P}\Big[\{i_{12}(A,B;Y) \leq \log\gamma_{12}\} \bigcup \{i_1(A;Y|B) \leq \log\gamma_1\} \bigcup \{i_2(B;Y|A) \leq \log\gamma_2\}\Big] + \mathbb{P}[E_{12}] + \mathbb{P}[E_1] + \mathbb{P}[E_2]$$

where

$$\begin{aligned}\mathbb{P}[E_{12}] &= \mathbb{P}[\{\exists (i \neq m, j \neq m') \text{ s.t. } i_{12}(c_m, d_{m'}; y) > \log\gamma_{12}\}] \\ &\leq (M_1-1)(M_2-1)\mathbb{P}[i_{12}(\overline{A}, \overline{B}; Y) > \log\gamma_{12}] \\ &= \mathbb{E}[e^{-i_{12}(A,B;Y)}\mathbf{1}\{i_{12}(A,B;Y) > \log\gamma_{12}\}] \\ &\leq e^{-\gamma_{12}} \\ \mathbb{P}[E_2] &= \mathbb{P}[\{\exists (j \neq m') \text{ s.t. } i_2(d_j; y|c_i) > \log\gamma_2\}] \\ &\leq (M_2-1)\mathbb{P}[i_2(\overline{B}; Y|A) > \log\gamma_2] \\ &= \mathbb{E}_A[e^{-i_2(B;Y|A)}\mathbf{1}\{i_2(B;Y|A) > \log\gamma_2\}|A] \\ &\leq \mathbb{E}_A[e^{-\gamma_2}|A] = e^{-\gamma_2} \\ \text{similarly } \ \mathbb{P}[E_1] &\leq e^{-\gamma_1}\end{aligned}$$

□

Note: [Intuition] Consider the decoding step when a random codebook is used. We observe Y and need to solve an M-ary hypothesis testing problem: Which of $\{P_{Y|A=c_m,B=d_{m'}}\}_{m,m'\in[M_1]\times[M_2]}$ produced the sample Y?

Recall that in P2P channel coding, we had a similar problem and the M-ary hypothesis testing problem was converted to M binary testing problems:

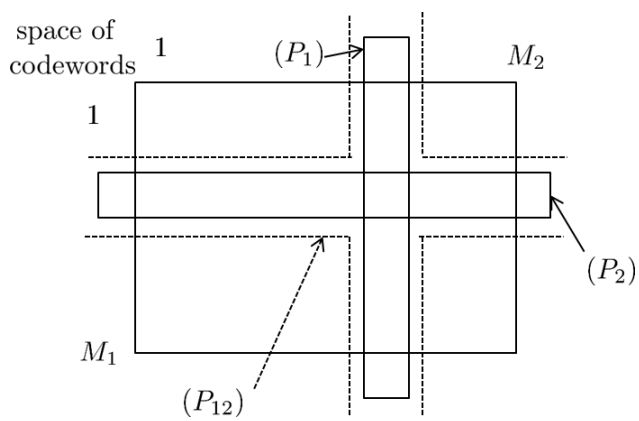
$$P_{Y|X=c_j} \quad \text{vs} \quad P_{Y_{-j}} \triangleq \sum_{i\neq j} \frac{1}{M-1} P_{Y|X=c_i} \approx P_Y$$

I.e. distinguish c_j (hypothesis H_0) against the average distribution induced by all other codewords (hypothesis H_1), which for a random coding ensemble $c_j \sim P_X$ is very well approximated by $P_Y = P_{Y|X} \circ P_X$. The optimal test for this problem is roughly

$$\frac{P_{Y|X=c}}{P_Y} \gtrsim \log(M-1) \quad \Longrightarrow \quad \text{decide } P_{Y|X=c_j} \tag{4}$$

since the prior for H_0 is $\frac{1}{M}$, while the prior for H_1 is $\frac{M-1}{M}$.

The proof above followed the same idea except that this time because of the two-dimensional grid structure:



there are in fact binary HT of three kinds

$$(P12) \quad \sim \quad \text{test } P_{Y|A=c_m,B=d_{m'}} \text{ vs } \quad \frac{1}{(M_1-1)(M_2-1)} \sum_{i\neq m} \sum_{j\neq m'} P_{Y|A=c_i,B=d_j} \approx P_Y$$

$$(P1) \quad \sim \quad \text{test } P_{Y|A=c_m,B=d_{m'}} \text{ vs } \quad \frac{1}{M_1-1} \sum_{i\neq m} P_{Y|A=c_i,B=d_{m'}} \approx P_{Y|B=d_{m'}}$$

$$(P2) \quad \sim \quad \text{test } P_{Y|A=c_m,B=d_{m'}} \text{ vs } \quad \frac{1}{M_2-1} \sum_{j\neq m'} P_{Y|A=c_m,B=d_j} \approx P_{Y|A=c_m}$$

And analogously to (4) the optimal tests are given by comparing the respective information densities with $\log M_1M_2$, $\log M_1$ and $\log M_2$.

Another observation following from the proof is that the following decoder would also achieve exactly the same performance:

- Step 1: rule out all cells (i,j) with $i_{12}(c_i,d_j;Y) \lesssim \log M_1M_2$.

- Step 2: If the messages remaining are NOT all in one row or one column, then FAIL.
- Step 3a: If the messages remaining are all in one column m' then declare $\hat{W}_2 = m'$. Rule out all entries in that column with $i_1(c_i; Y|d_{m'}) \lesssim \log M_1$. If more than one entry remains, FAIL. Otherwise declare the unique remaining entry m as $\hat{W}_1 = m$.
- Step 3b: Similarly with column replaced by row, i_1 with i_2 and $\log M_1$ with $\log M_2$.

The importance of this observation is that in the regime when RHS of (3) is small, the decoder always finds it possible to basically decode one message, "subtract" its influence and then decode the other message. Which of the possibilities 3a/3b appears more often depends on the operating point (R_1, R_2) inside $\mathcal{C}$.

24.3 MAC capacity region proof

Proof. 1. Show $\mathcal{C}$ is convex.

Take $(R_1, R_2) \in \mathcal{C}_{\epsilon/2}$, and take $(R_1', R_2') \in \mathcal{C}_{\epsilon/2}$.
We merge the $(n, 2^{nR_1}, 2^{nR_2}, \epsilon/2)$ code and the $(n, 2^{nR_1}, 2^{nR_2}, \epsilon/2)$ code in the following time-sharing way: in the first n channels, use the first set of codes, and in the last n channels, use the second set of codes.
Thus we formed a new $(2n, 2^{R_1+R_1'}, 2^{R_2+R_2'}, \epsilon)$ code, we know that

$$\frac{1}{2}\mathcal{C}_{\epsilon/2} + \frac{1}{2}\mathcal{C}_{\epsilon/2} \subset \mathcal{C}_\epsilon$$

take limit at both sides

$$\frac{1}{2}\mathcal{C} + \frac{1}{2}\mathcal{C} \subset \mathcal{C}$$

also we know that $\mathcal{C} \subset \frac{1}{2}\mathcal{C} + \frac{1}{2}\mathcal{C}$, therefore $\mathcal{C} = \frac{1}{2}\mathcal{C} + \frac{1}{2}\mathcal{C}$ is convex.

Note: the set addition is defined in the following way:

$$\mathcal{A} + \mathcal{B} \triangleq \{(a+b) : a \in \mathcal{A}, b \in \mathcal{B}\}$$

2. Achievability

STP: $\forall P_A, P_B, \forall (R_1, R_2) \in Penta(P_A, P_B)$, $\exists (n, 2^{nR_1}, 2^{nR_2}, \epsilon) code$.

Apply Lemma 24.1 with:

$$P_A \to P_A^n, \quad P_B \to P_B^n, \quad P_{Y|A,B} \to P_{Y|A,B}^n$$
$$M_1 = 2^{nR_1}, \quad M_2 = 2^{nR_2},$$
$$\log \gamma_{12} = n(I(A,B;Y) - \delta), \quad \log \gamma_1 = n(I(A;Y|B) - \delta), \quad \log \gamma_2 = n(I(B;Y|A) - \delta).$$

we have that there exists a (M_1, M_2, ϵ) code with

$$\begin{aligned}\epsilon \leq & \mathbb{P}\Big[\{\frac{1}{n}\sum_{k=1}^{n} i_{12}(A_k, B_k; Y_k) \leq \log\gamma_{12} - \delta\} \bigcup \{\frac{1}{n}\sum_{k=1}^{n} i_1(A_k; Y_k|B_k) \leq \log\gamma_1 - \delta\} \\ & \underbrace{\bigcup \{\frac{1}{n}\sum_{k=1}^{n} i_2(B_k; Y_k|A_k) \leq \log\gamma_2 - \delta\}\Big]}_{①} \\ & + \underbrace{(2^{nR_1}-1)(2^{nR_2}-1)e^{-\gamma_{12}} + (2^{nR_1}-1)e^{-\gamma_1} + (2^{nR_2}-1)e^{-\gamma_2}}_{②}\end{aligned}$$

by WLLN, the first part goes to zero, and for any (R_1, R_2) such that $R_1 < I(A; Y|B) - \delta$ and $R_2 < I(B; Y|A) - \delta$ and $R_1 + R_2 < I(A, B; Y) - \delta$, the second part goes to zero as well. Therefore, if $(R_1, R_2) \in$ interia of the Penta, there exists a $(M_1, M_2, \epsilon = o(1))$ code.

3. Weak converse

$$\begin{tikzcd} W_1 \arrow[r, "f_1"] & A^n \arrow[dr] & & \hat{W}_1 \\ & & Y^n \arrow[ur, "g"] \arrow[dr] & \\ W_2 \arrow[r, "f_2"] & B^n \arrow[ur] & & \hat{W}_2 \end{tikzcd}$$

P

$$\begin{tikzcd} W_1 \arrow[r, "f_1"] & A^n & & \hat{W}_1 \\ & & Y^n \arrow[ur, "g"] \arrow[dr] & \\ W_2 \arrow[r, "f_2"] & B^n & & \hat{W}_2 \end{tikzcd}$$

$Q \in (*)$

$$\begin{tikzcd} W_1 \arrow[r, "f_1"] & A^n \arrow[dr] & & \hat{W}_1 \\ & & Y^n \arrow[ur, "g"] \arrow[dr] & \\ W_2 \arrow[r, "f_2"] & B^n & & \hat{W}_2 \end{tikzcd}$$

$Q_1 \in (*1)$

$$\begin{tikzcd} W_1 \arrow[r, "f_1"] & A^n & & \hat{W}_1 \\ & & Y^n \arrow[ur, "g"] \arrow[dr] & \\ W_2 \arrow[r, "f_2"] & B^n \arrow[ur] & & \hat{W}_2 \end{tikzcd}$$

$Q_2 \in (*2)$

$$\mathbb{Q}[W_1 = \hat{W}_1, W_2 = \hat{W}_2] = \frac{1}{M_1 M_2}, \quad \mathbb{P}[W_1 = \hat{W}_1, W_2 = \hat{W}_2] \geq 1 - \epsilon$$

d-proc:

$$\begin{aligned} & d(1-\epsilon \| \frac{1}{M_1 M_2}) \leq \inf_{\mathbb{Q}\in(*)} D(P\|Q) = I(A^n, B^n; Y^n) \\ \Rightarrow & R_1 + R_2 \leq \frac{1}{n} I(A^n, B^n; Y^n) + o(1) \end{aligned}$$

To get separate bounds, we apply the same trick to evaluate the information flow from the link between $A \to Y$ and $B \to Y$ separately:

$$\mathbb{Q}_1[W_2 = \hat{W}_2] = \frac{1}{M_2}, \quad \mathbb{P}[W_2 = \hat{W}_2] \geq 1 - \epsilon$$

d-proc:

$$\begin{aligned} & d(1-\epsilon \| \frac{1}{M_2}) \leq \inf_{\mathbb{Q}_1\in(*1)} D(P\|Q_1) = I(B^n; Y^n|A^n) \\ \Rightarrow & R_2 \leq \frac{1}{n} I(B^n; Y^n|A^n) + o(1) \end{aligned}$$

similarly we can show that

$$R_2 \leq \frac{1}{n} I(A^n; Y^n | B^n) + o(1)$$

For memoryless channels, we know that $\frac{1}{n} I(A^n, B^n; Y^n) \leq \frac{1}{n} \sum_k I(A_k, B_k; Y_k)$. Notice that each (A_i, B_i) pair corresponds to (P_{A_k}, P_{B_k}), $\forall k$ define

$$Penta_k(P_{A_k}, P_{B_k}) = \left\{ (R_{1,k}, R_{2,k}) : \begin{array}{c} 0 \leq R_{1,k} \leq I(A_k; Y_k | B_k) \\ 0 \leq R_{2,k} \leq I(B_k; Y_k | A_k) \\ R_{1,k} + R_{2,k} \leq I(A_k, B_k; Y_k) \end{array} \right\}$$

therefore

$$\begin{aligned} &(R_1, R_2) \in \left[\frac{1}{n} \sum_k Penta_k \right] \\ \Rightarrow &C \in \overline{co} \bigcup_{P_A, P_B} Penta \end{aligned}$$

$\square$

25.1 Examples of Multiple Access Channels

Last time we defined the multiple access channel as the sequence of random transformations

$$\{P_{Y^n|A^nB^n} : \mathcal{A}^n \times \mathcal{B}^n \to \mathcal{Y}^n, n = 1, 2, \ldots\}$$

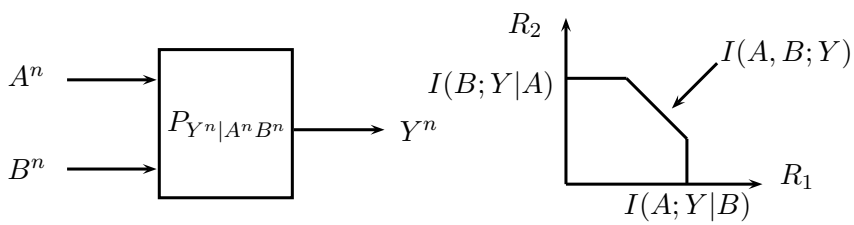
Furthermore, we showed that its capacity region is

$$C = \{(R_1, R_2) : \exists (n, 2^{nR_1}, 2^{nR_2}, \epsilon) - MAC\ code\} = \overline{co} \bigcup_{P_A P_B} \text{Penta}(P_A, P_B)$$

were $\overline{co}$ denotes the convex hull of the sets Penta, and Penta is

$$\text{Penta}(P_A, P_B) = \begin{cases} R_1 \le I(A; Y|B) \\ R_2 \le I(B; Y|A) \\ R_1 + R_2 \le I(A, B; Y) \end{cases}$$

So a general MAC and one Penta region looks like



Note that the union of Pentas need not look like a Penta region itself, as we will see in a later example.

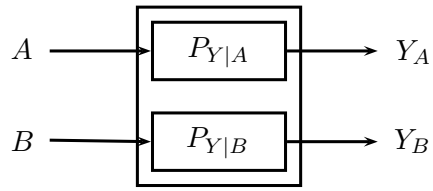
25.1.1 Orthogonal MAC

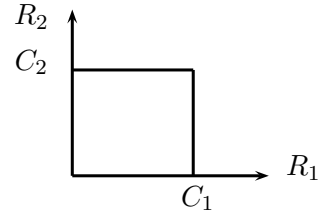
The trivial MAC is when each input sees its own independent channel: $P_{Y|AB} = P_{Y|A}P_{Y|B}$ where the receiver sees (Y_A, Y_B). In this situation, we expect that each transmitter can achieve it's own capacity, and no more than that. Indeed, our theorem above shows exactly this:

$$\text{Penta}(P_A, P_B) = \begin{cases} R_1 \le I(A; Y|B) = I(A; Y) \\ R_2 \le I(B; Y|A) = I(B; Y) \\ R_1 + R_2 \le I(A, B; Y) \end{cases}$$

Where in this case the last constraint is not applicable; it does not restrict the capacity region.

Typed by Austin Collins, austinc@mit.edu, May 16th, 2013





Hence our capacity region is a rectangle bounded by the individual capacities of each channel.

25.1.2 BSC MAC

Before introducing this channel, we need a definition and a theorem:

Definition 25.1 (Sum Capacity)**.** $C_{sum} \triangleq \max\{R_1 + R_2 : (R_1, R_2) \in C\}$

Theorem 25.1. $C_{sum} = \max_{A \perp\!\!\!\perp B} I(A, B; Y)$

Proof. Since the max above is achieved by an extreme point on one of the Penta regions, we can drop the convex closure operation to get

$$\max\{R_1 + R_2 : (R_1, R_2) \in \overline{co} \bigcup \text{Penta}(P_A, P_B)\} = \max\{R_1 + R_2 : (R_1, R_2) \in \bigcup \text{Penta}(P_A, P_B)\}$$
$$\max_{P_A, P_B} \{R_1 + R_2 : (R_1, R_2) \in \text{Penta}(P_A, P_B)\} \leq \max_{P_A, P_B} I(A, B; Y)$$

Where the last step follows from the definition of Penta. Now we need to show that the constraint on $R_1 + R_2$ in Penta is active at at least one point, so we need to show that $I(A, B; Y) \leq I(A; Y|B) + I(B; Y|A)$ when $A \perp\!\!\!\perp B$, which follows from applying Kolmogorov identities

$$\begin{aligned} I(A; Y, B) = 0 + I(A; Y|B) = I(A; Y) + I(A; B|Y) &\implies I(A; Y) \leq I(A; Y|B) \\ &\implies I(A, B; Y) = I(A; Y) + I(B; Y|A) \leq I(A; Y|B) + I(B; Y|A) \end{aligned}$$

Hence $\max_{P_A, P_B}\{R_1 + R_2 : (R_1, R_2) \in \text{Penta}(P_A, P_B)\} = \max_{P_A P_B} I(A, B; Y)$ □

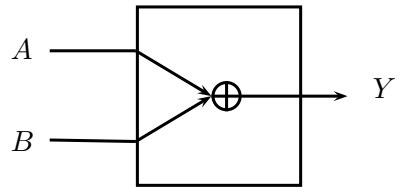
We now look at the BSC MAC, defined by

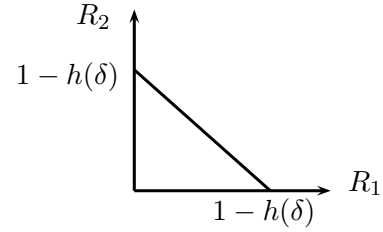
$$\begin{aligned} &Y = A + B + Z \mod 2 \\ &Z \sim \text{Ber}(\delta) \\ &A, B \in \{0, 1\} \end{aligned}$$

Since the output Y can only be 0 or 1, the capacity of this channel can be no larger than 1 bit. If B doesn't transmit at all, then A can achieve capacity $1 - h(\delta)$ (and B can achieve capacity when A doesn't transmit), so that $R_1, R_2 \leq 1 - h(\delta)$. By time sharing we can obtain any point between these two. This gives an inner bound on the capacity region. For an outer bound, we use Theorem 25.1, which gives

$$\begin{aligned} C_{sum} = \max_{P_A P_B} I(A, B; Y) &= \max_{P_A P_B} I(A, B; A + B + Z) \\ &= \max_{P_A P_B} H(A + B + Z) - H(Z) = 1 - h(\delta) \end{aligned}$$

Hence $R_1 + R_2 \leq 1 - h(\delta)$, so by this outer bound, we can do no better than time sharing between the two individual channel capacity points.





Remark: Even though this channel seems so simple, there are still hidden things about it, which we'll see later.

25.1.3 Adder MAC

Now we analyze the Adder MAC, which is a noiseless channel defined by:

$$\begin{aligned} &Y = A + B \quad (\text{over } \mathbb{Z}) \\ &A, B \in \{0, 1\} \end{aligned}$$

Intuitively, the game here is that when both A and B send either 0 or 1, we receiver 0 or 2 and can decode perfectly. However, when A sends 0 and B send 1, the situation is ambiguous. To analyze this channel, we start with an interesting fact

Interesting Fact 1: Any deterministic MAC ($Y = f(A, B)$) has $C_{sum} = \max H(Y)$. To see this, just expand $I(A, B; Y)$.

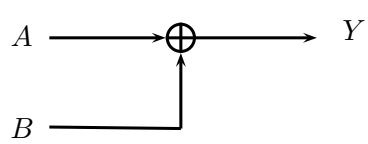
Therefore, the sum capacity of this MAC is

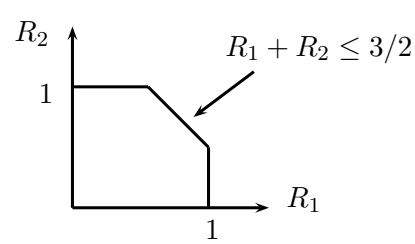
$$C_{sum} = \max_{A \perp\!\!\!\perp B} H(A + B) = H\left(\frac{1}{4}, \frac{1}{2}, \frac{1}{4}\right) = \frac{3}{2} \text{ bits}$$

Which is achieved when both A and B are Ber(1/2). With this, our capacity region is

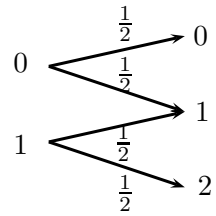
$$\text{Penta}(\text{Ber}(1/2), \text{Ber}(1/2)) = \begin{cases} R_1 \leq I(A; Y|B) = H(A) = 1 \\ R_2 \leq I(B; Y|A) = H(B) = 1 \\ R_1 + R_2 \leq I(A, B; Y) = 3/2 \end{cases}$$

So the channel can be described by

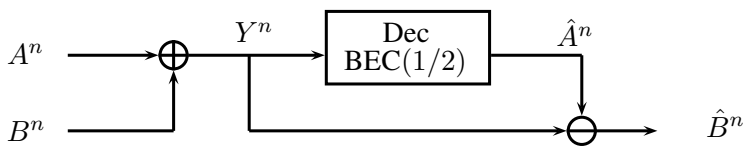




Now we can ask: how do we achieve the corner points of the region, e.g. $R_1 = 1/2$ and $R_2 = 1$? The answer gives insights into how to code for this channel. Take the greedy codebook $B = \{0, 1\}^n$ (the entire space), then the channel $A \to Y$ is a DMC:



Which we recognize as a BEC(1/2) (no preference to either -1 or 1), which has capacity 1/2. How do we decode? The idea is *successive cancellation*, where first we decode A, then remove $\hat{A}$ from Y, then decode B.



Using this strategy, we can use a single user code for the BEC (an object we understand well) to attain capacity.

25.1.4 Multiplier MAC

The Multiplier MAC is defined as

$$\begin{aligned} &Y = AB \\ &A \in \{0,1\},\ B \in \{-1,1\} \end{aligned}$$

Note that $A = |Y|$ can always be resolved, and B can be resolved whenever $A = 1$. To find the capacity region of this channel, we'll use another interesting fact:

Interesting Fact 2: If $A = g(Y)$, then each $\text{Penta}(P_A, P_B)$ is a rectangle with

$$\text{Penta}(P_A, P_B) = \begin{cases} R_1 \leq H(A) \\ R_2 \leq I(A,B;Y) - H(A) \end{cases}$$

Proof. Using the assumption that $A = g(Y)$ and expanding the mutual information

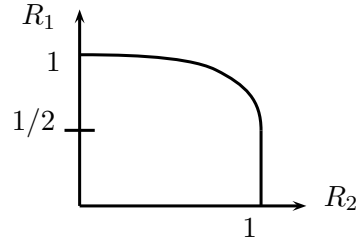
$$\begin{aligned} I(A;Y|B) + I(B;Y|A) &= H(A) - H(Y|A) - H(Y|A,B) = H(A,Y) - H(Y|A,B) \\ &= H(Y) - H(Y|A,B) = I(A,B;Y) \end{aligned}$$

Therefore the $R_1 + R_2$ constraint is not active, so our region is a rectangle. □

By symmetry, we take $P_B = \text{Ber}(1/2)$. When $P_A = \text{Ber}(p)$, the output has $H(Y) = p + h(p)$. Using the above fact, the capacity region for the Multiplier MAC is

$$C = \overline{co} \bigcup \begin{cases} R_1 \leq H(A) = h(p) \\ R_2 \leq H(Y) - H(A) = p \end{cases}$$

We can view this as the graph of the binary entropy function on its side, parametrized by p:



To achieve the extreme point $(1, 1/2)$ of this region, we can use the same scheme as for the Adder MAC: take the codebook of A to be $\{0,1\}^n$, then B sees a BEC(1/2). Again, successive cancellation decoding can be used.

For future reference we note:

Lemma 25.1. *The full capacity region of multiplier MAC is achieved with zero error.*

Proof. For a given codebook D of user B the number of messages that user A can send equals the total number of erasure patters that codebook D can tolerate with vanishing probability of error. Fix rate $R_2 < 1$ and let D bew a row-span of a random linear $nR_2 \times n$ binary matrix. Then randomly erase each column with probability $1 - R_2 - \epsilon$. Since on average there will be $n(R_2 + \epsilon)$ columns left, the resulting matrix is still full-rank and the decoding is possible. In other words,

$$\mathbb{P}[D \text{ is decodable}, \# \text{ of erasures} \approx n(1 - R_2 - \epsilon)] \to 1\,.$$

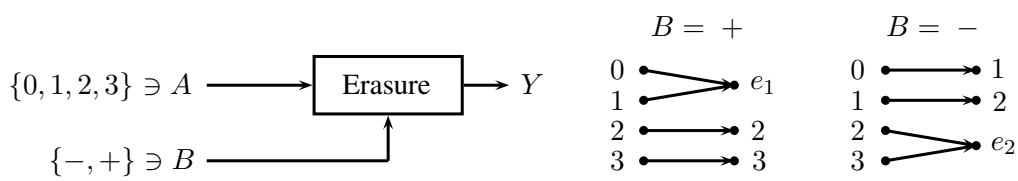
Hence, by counting the total number of erasures, for a random linear code we have

$$\mathbb{E}[\# \text{ of decodable erasure patterns for } D] \approx 2^{nh(1-R_2-\epsilon)+o(n)}\,.$$

And result follows by selecting a random element of the D-ensemble and then taking the codebook of user A to be the set of decodable erasure patterns for a selected D. □

25.1.5 Contraction MAC

The Contraction MAC is defines as



Here, B is received perfectly, We can use the fact above to see that the capacity region is

$$C = \begin{cases} R_1 \le \frac{3}{2} \\ R_2 \le 1 \end{cases}$$

For future reference we note the following:

Lemma 25.2. *The zero-error capacity of the contraction MAC satisfies*

$$R_1 \le h(1/3) + (2/3 - p)\log 2\,, \tag{1}$$

$$R_2 \le h(p) \tag{2}$$

for some $p \in [0, 1/2]$. In particular, the point $R_1 = \frac{3}{2}$, $R_2 = 1$ is not achievable with zero error.

Proof. Let C and D denote the zero-error codebooks of two users. Then for each string $b^n \in \{+,-\}^n$ denote

$$U_{b^n} = \{a^n : a_j \in \{0,1\} \text{ if } b_j = +, a_j \in \{2,3\} \text{ if } b_j = -\}\,.$$

Then clearly for each b^n we have

$$|U_{b^n}| \le 2^{d(b^n, D)}\,,$$

where $d(b^n, D)$ denotes the minimum Hamming distance from string b^n to the set D. Then,

$$|C| \le \sum_{b^n} 2^{d(b^n,D)} \tag{3}$$

$$= \sum_{j=0}^{n} 2^j |\{b^n : d(b^n, D) = j\}| \tag{4}$$

For a given cardinality $|D|$ the set that maximizes the above sum is the Hamming ball. Hence, $R_2 = h(p) + O(1)$ implies

$$R_2 \le \max_{q \in [p,1]} h(q) + (q-p)\log 2 = h(1/3) + (2/3 - p)\log 2\,.$$

□

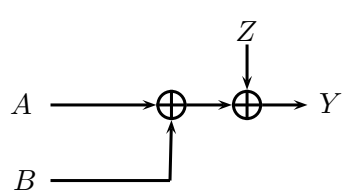
25.1.6 Gaussian MAC

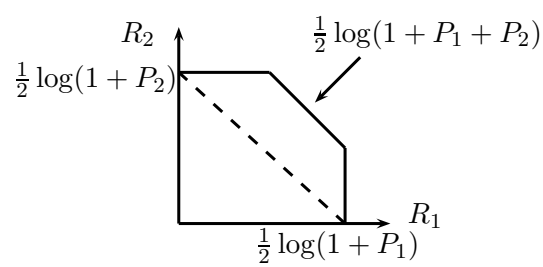
Perhaps the most important MAC is the Gaussian MAC. This is defined as

$$\begin{aligned} &Y = A + B + Z \\ &Z \sim \mathcal{N}(0,1) \\ &\mathbb{E}[A^2] \le P_1,\ \mathbb{E}[B^2] \le P_2 \end{aligned}$$

Evaluating the mutual information, we see that the capacity region is

$$\begin{aligned} I(A;Y|B) &= I(A; A+Z) \le \frac{1}{2}\log(1+P_1) \\ I(B;Y|A) &= I(B; B+Z) \le \frac{1}{2} \le (1+P_2) \\ I(A,B;Y) &= h(Y) - h(Z) \le \frac{1}{2}\log(1+P_1+P_2) \end{aligned}$$





Where the region is Penta($\mathcal{N}(0, P_1), \mathcal{N}(0, P_2)$). How do we achieve the rates in this region? We'll look at a few schemes.

1. TDMA: A and B switch off between transmitting at full rate and not transmitting at all. This achieves any rate pair in the form

$$R_1 = \lambda \frac{1}{2} \log(1 + P_1), \quad R_2 = \bar{\lambda} \frac{1}{2} \log(1 + P_2)$$

   Which is the dotted line on the plot above. Clearly, there are much better rates to be gained by smarter schemes.

2. FDMA (OFDM): Dividing users into different frequency bands rather than time windows gives an enormous advantage. Using frequency division, we can attain rates

$$R_1 = \lambda \frac{1}{2} \log\left(1 + \frac{P_1}{\lambda}\right), \quad R_2 = \bar{\lambda} \frac{1}{2} \log\left(1 + \frac{P_2}{\bar{\lambda}}\right)$$

   In fact, these rates touch the boundary of the capacity region at its intersection with the $R_1 = R_2$ line. The optimal rate occurs when the power at each transmitter makes the noise look white:

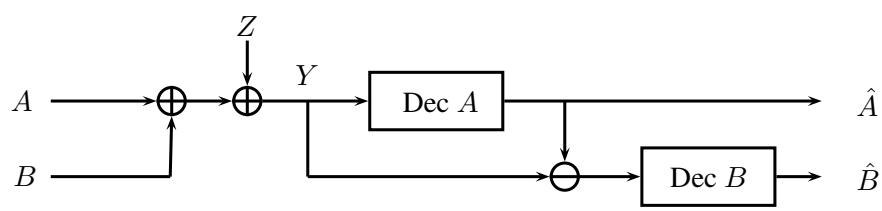
$$\frac{P_1}{\lambda} = \frac{P_2}{\bar{\lambda}} \implies \lambda^* = \frac{P_1}{P_1 + P_2}$$

   While this touches the capacity region at one point, it doesn't quite reach the corner points.

3. Rate Splitting/Successive Cancellation: To reach the corner points, we can use successive cancellation, similar to the decoding schemes in the Adder and Multiplier MACs. We can use rates:

$$\begin{aligned} R_2 &= \frac{1}{2} \log(1 + P_2) \\ R_1 &= \frac{1}{2}(\log(1 + P_1 + P_2) - \log(1 + P_2)) = \frac{1}{2} \log\left(1 + \frac{P_1}{1 + P_2}\right) \end{aligned}$$

   The second expression suggests that A transmits at a rate for an AWGN channel that has power constraint P_1 and noise $1 + P_2$, i.e. the power used by B looks like noise to A.



Theorem 25.2. *There exists a (E_1, E_2, D_1, D_2) code that achieves the corner points.*

Proof. Random coding: $B^n \sim \mathcal{N}(0, P_2)^n$. Since A^n now sees noise $1 + P_2$, there exists a code for A with rate $R_1 = \frac{1}{2} \log(1 + P_1/(1 + P_2))$. □

This scheme (unlike the above two) can tolerate frame un-synchronization between the two transmitters. This is because any chunk of length n has distribution $\mathcal{N}(0, P_2)^n$.

25.1.7 MAC Peculiarities

Now that we've seen some nice properties and examples of MAC's, we'll look at cases where MAC's differ from the point to point channels we've seen so far.

1. Max probability of error $\neq$ average probability of error.

   Theorem 25.3. $C^{(max)} \neq C$

   Proof. The key observation for deterministic MAC is that $C^{(max)} = C_0$ (zero error capacity) when $\epsilon \leq 1/2$. This is because when any two strings can be confused, the maximum probability of error

   $$\max_{m,m'} \mathbb{P}[\hat{W}_1 \neq m \ \cup \ \hat{W}_2 \neq m' | W_1 = m, W_2 = m']$$

   Must be larger than 1/2. □

   For some of the channels we've seen

   - Contraction MAC: $C_0 \neq C$
   - Multiplier MAC: $C_0 = C$
   - Adder MAC: $C_0 \neq C$. For this channel, no one yet can show that $C_{0,sum} < 3/2$. The idea is combinatorial in nature: produce two sets (Sidon sets) such that all pairwise sums between the two do not overlap.

2. Separation does not hold: In the point to point channel, through joint source channel coding we saw that an optimal architecture is to do source coding then channel coding separately. This doesn't hold for the MAC. Take as a simple example the Adder MAC with a correlated source and bandwidth expansion factor $\rho = 1$. Let the source (S, T) have joint distribution

   $$P_{ST} = \begin{bmatrix} 1/3 & 1/3 \\ 0 & 1/3 \end{bmatrix}$$

   We encode S^n to channel input A^n and T^n to channel input B^n. The simplest possible scheme is to not encoder at all; simply take $S_j = A_j$ and $T_j = B_j$. Take the decoder

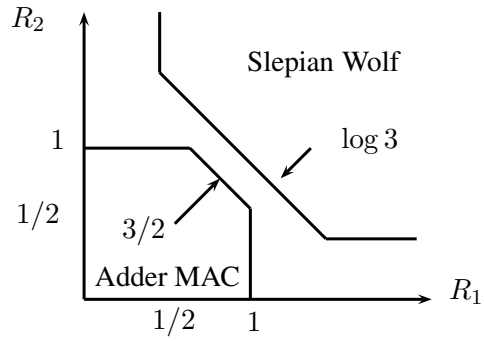
   $$\begin{array}{rcl} & & \hat{S} \quad \hat{T} \\ Y_j = 0 & \Longrightarrow & 0 \quad 0 \\ Y_j = 1 & \Longrightarrow & 0 \quad 1 \\ Y_j = 2 & \Longrightarrow & 1 \quad 1 \end{array}$$

   Which gives $\mathbb{P}[\hat{S}^n = S^n, \hat{T}^n = T^n] = 1$, since we are able to take advantage of the zero entry in joint distribution of our correlated source.

   Can we achieve this with a separated source? Amazingly, even though the above scheme is so simple, we can't! The compressors in the separated architecture operate in the Slepian Wolf region

   $$\begin{cases} R_1 \geq H(S|T) \\ R_2 \geq H(T|S) \\ R_1 + R_2 \geq H(S,T) = \log 3 \end{cases}$$

Hence the sum rate for compression must be $\geq \log 3$, while the sum rate for the Adder MAC must be $\leq 3/2$, so these two regions do not overlap, hence we can not operate at a bandwidth expansion factor of 1 for this source and channel.



3. Finally, even for the BSC-MAC there is some pecularity. Indeed, suppose that two users A and B each have a k-bit message W_1, W_2. Suppose the receiver is only interested in estimating $W_1 + W_2$. What is the largest ratio k/n? Clearly, separation can achieve

$$k/n \approx \frac{1}{2}(\log 2 - h(\delta))$$

by simply creating a scheme in which both W_1 and W_2 are estimated and then their sum is computed.

A more clever solution is however to encode

$$\begin{aligned} A^n &= G \cdot W_1 \,, \\ B^n &= G \cdot W_2 \,, \\ Y^n &= A^n + B^n + Z^n = G(W_1 + W_2) + Z^n \,. \end{aligned}$$

where G is a generating matrix of a good k-to-n linear code. Then, provided that

$$k < n(\log 2 - h(\delta)) + o(n)$$

the sum $W_1 + W_2$ is decodable. Hence even for a simple BSC-MAC there exist clever ways to exceed MAC capacity for certain scenarios. Note that this "distributed computation" can also be viewed as lossy source coding with a distortion metric that is only sensitive to discrepancy between $W_1 + W_2$ and $\hat{W}_1 + \hat{W}_2$.