# Lecture #2 Basic Information Theory

(Reading: NIT 2, 3.1, 3.3–3.5, 9.1)

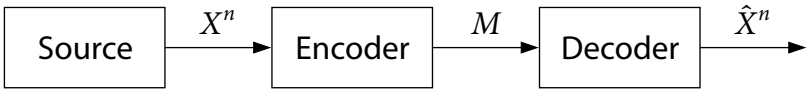
---

- Lossless source coding
- Channel coding
- Channel coding with input cost
- Gaussian channel

---

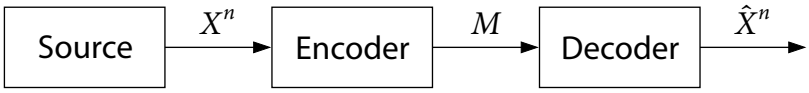


## Point-to-point lossless compression system




- Discrete (stationary) memoryless source (DMS) $(\mathcal{X}, p(x))$ (or $X$ in short)
  - Generates i.i.d. sequence $X_1, X_2, \ldots$ with $X_i \sim p_X(x_i)$
  - Example: $\mathrm{Bern}(p)$ source $X$ generates i.i.d. $\mathrm{Bern}(p)$ sequence
- A $(2^{nR}, n)$ lossless compression code:
  - Encoder: $m(x^n) \in [1 : 2^{nR}) = \{1, 2, \ldots, 2^{\lfloor nR \rfloor}\}$ ($R$ = rate in bits/source symbol)
  - Decoder: $\hat{x}^n(m) \in \mathcal{X}^n$

# Point-to-point lossless compression system




- Probability of error: $P_e^{(n)} = \mathsf{P}\{\hat{X}^n \neq X^n\}$
- $R$ achievable if $\exists$ a sequence of $(2^{nR}, n)$ codes such that $\lim_{n\to\infty} P_e^{(n)} = 0$
- Optimal lossless compression rate $R^*$: Infimum of all achievable $R$

## Lossless source coding theorem (Shannon 1948)

$$R^* = H(X) = -\sum_{x\in\mathcal{X}} p(x)\log p(x) \text{ bits/symbol} \quad \text{(entropy)}$$

- Examples:
  - If $X \sim \text{Bern}(p)$, then $H(X) = -p\log p - (1-p)\log(1-p) = H(p)$ (binary entropy function)
  - If $X \sim \text{Unif}(\mathcal{X})$, then $H(X) = \log|\mathcal{X}|$
  - In general $H(X) \leq \log|\mathcal{X}|$ (by Jensen's inequality)



# Proving the lossless source coding theorem

- To prove this theorem, we need to establish:
  - Achievability: If $R > H(X)$, $\exists$ a sequence of $(2^{nR}, n)$ codes such that $\lim_{n\to\infty} P_e^{(n)} = 0$
  - Converse: For any sequence of $(2^{nR}, n)$ codes with $\lim_{n\to\infty} P_e^{(n)} = 0$, $R \geq H(X)$
- We need to review:
  - Conditional and joint entropy
  - The notion of typicality

## Conditional and joint entropy

- Conditional entropy (equivocation): Let $(X, Y) \sim p(x, y)$, $(x, y) \in \mathcal{X} \times \mathcal{Y}$

$$H(Y|X) = \sum_{x \in \mathcal{X}} H(Y \,|\, X = x) p(x)$$

  - $H(Y|X) \le H(Y)$ (with equality if $X$ and $Y$ are independent)

- Joint entropy: For $(X, Y) \sim p(x, y)$,

$$H(X, Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$$

- Chain rule for entropy:

$$H(X^n) = \sum_{i=1}^{n} H(X_i|X^{i-1})$$

  - $H(X^n) \le \sum_{i=1}^{n} H(X_i)$ (with equality if $X_1, X_2, \ldots, X_n$ are independent)

- Fano's inequality: If $(X, Y) \sim p(x, y)$ and $P_e = \mathsf{P}\{X \neq Y\}$, then

$$H(X|Y) \le H(P_e) + P_e \log |\mathcal{X}| \le 1 + P_e \log |\mathcal{X}|$$



## Typical sequences

- Empirical pmf (type) of $x^n \in \mathcal{X}^n$:

$$\pi(x|x^n) = \frac{\left|\{i: x_i = x\}\right|}{n} \quad \text{for } x \in \mathcal{X}$$

- Example: For $x^n = 0100101$, $\pi(0|x^n) = 4/7$ and $\pi(1|x^n) = 3/7$

- $X_1, X_2, \ldots, X_n$ i.i.d. with $X_1 \sim p(x)$, then by the WLLN

$$\pi(x|X^n) \to p(x) \quad \text{in probability for every } x \in \mathcal{X}$$

- Typical set (Orlitsky–Roche 2001): For $X \sim p(x)$ and $\epsilon > 0$,

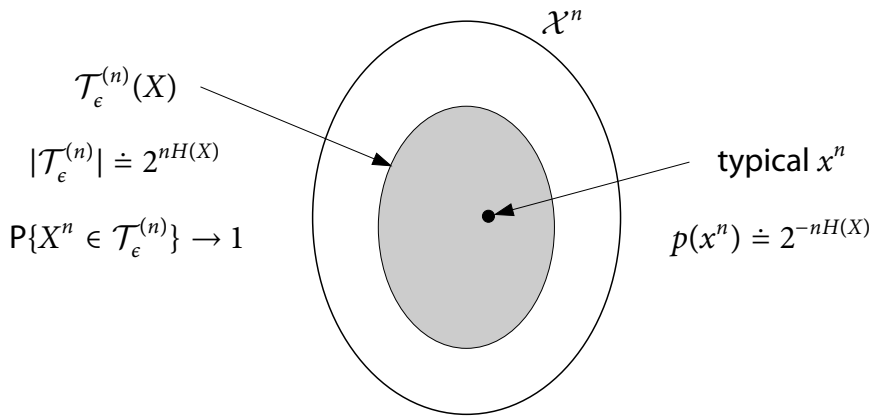
$$\mathcal{T}_\epsilon^{(n)}(X) = \mathcal{T}_\epsilon^{(n)} = \{x^n: \left|\pi(x|x^n) - p(x)\right| \le \epsilon p(x) \text{ for all } x \in \mathcal{X}\}$$

### Typical average lemma

If $x^n \in \mathcal{T}_\epsilon^{(n)}(X)$ and $g(x) \ge 0$, then

$$(1 - \epsilon)\,\mathsf{E}(g(X)) \le \frac{1}{n} \sum_{i=1}^{n} g(x_i) \le (1 + \epsilon)\,\mathsf{E}(g(X))$$

# Properties of typical sequences




- For $x^n \in \mathcal{T}_\epsilon^{(n)}(X)$, $2^{-n(H(X)+\delta(\epsilon))} \le \prod_{i=1}^n p_X(x_i) \le 2^{-n(H(X)-\delta(\epsilon))}$
- $|\mathcal{T}_\epsilon^{(n)}(X)| \le 2^{n(H(X)+\delta(\epsilon))}$
- If $X^n \sim \prod_{i=1}^n p_X(x_i)$, then $\mathrm{P}\{X^n \in \mathcal{T}_\epsilon^{(n)}(X)\} \to 1$ (by the LLN)
- $|\mathcal{T}_\epsilon^{(n)}(X)| \ge (1-\epsilon)2^{n(H(X)-\delta(\epsilon))} = 2^{n(H(X)-\delta'(\epsilon))}$ for $n$ sufficiently large



# Achievability proof of lossless source coding theorem

- If $R > H(X)$, $\exists$ sequence of $(2^{nR}, n)$ codes with $\lim_{n\to\infty} P_e^{(n)} = 0$
- Let $R > H(X) + \delta(\epsilon)$ so that $|\mathcal{T}_\epsilon^{(n)}| \le 2^{n(H(X)+\delta(\epsilon))} < 2^{nR}$
- Codebook:
  - Assign a distinct index $m(x^n)$ to each $x^n \in \mathcal{T}_\epsilon^{(n)}$
  - Assign $m = 1$ to all $x^n \notin \mathcal{T}_\epsilon^{(n)}$
  - Codebook is revealed to both encoder and decoder
- Encoding:
  - Upon observing $x^n$, send $m(x^n)$
- Decoding:
  - Declare $\hat{x}^n = x^n(m)$ for the unique $x^n(m) \in \mathcal{T}_\epsilon^{(n)}$
- Analysis of the probability of error:
  - All typical sequences are correctly recovered
  - Thus, $\lim_{n\to\infty} P_e^{(n)} = \lim_{n\to\infty} \mathrm{P}\{X^n \notin \mathcal{T}_\epsilon^{(n)}\} = 0$, and every $R > H(X)$ is achievable

# Converse proof of lossless source coding theorem

- Given sequence of $(2^{nR}, n)$ codes with $\lim_{n\to\infty} P_e^{(n)} \to 0$, $R \geq H(X)$
- For each code, let $M = m(X^n)$ and $\hat{X}^n = \hat{x}^n(M)$
- Consider

$$\begin{aligned} nR &\geq H(M) \\ &\geq H(\hat{X}^n) \\ &= H(X^n, \hat{X}^n) - H(X^n|\hat{X}^n) \\ &= H(X^n) + H(\hat{X}^n|X^n) - H(X^n|\hat{X}^n) \\ &= H(X^n) - H(X^n|\hat{X}^n) \\ &= nH(X) - H(X^n|\hat{X}^n) \end{aligned}$$

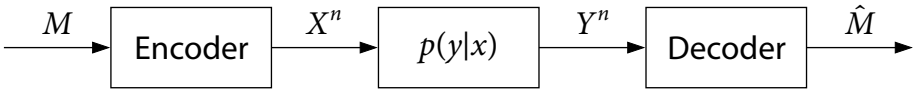
- By Fano's inequality,

$$H(X^n|\hat{X}^n) \leq 1 + nP_e^{(n)} \log|\mathcal{X}| = n(1/n + P_e^{(n)} \log|\mathcal{X}|) = n\epsilon_n,$$

  where $\epsilon_n \to 0$ as $n \to \infty$ by assumption
- Hence as $n \to \infty$, $R \geq H(X)$

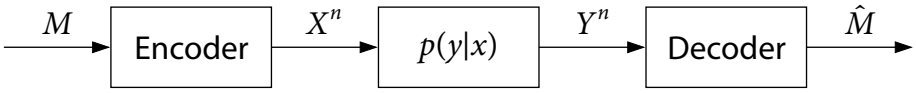


# Point-to-point communication system




- Discrete memoryless channel (DMC) $(\mathcal{X}, p(y|x), \mathcal{Y})$
  - Discrete: $\mathcal{X}$ and $\mathcal{Y}$ are finite
  - Memoryless: $p(y_i|y^{i-1}, x^i, m) = p(y_i|x_i)$, $i \in [1:n]$, i.e., $(M, Y^{i-1}, X^{i-1}) \to X_i \to Y_i$
  - Without feedback: $p(y^n|x^n, m) = \prod_{i=1}^n p_{Y|X}(y_i|x_i)$
- A $(2^{nR}, n)$ code for the DMC:
  - Message set $[1:2^{nR}] = \{1, 2, \ldots, 2^{\lceil nR \rceil}\}$
  - Encoder: a codeword $x^n(m)$ for each $m \in [1:2^{nR}]$
    $\mathcal{C} = \{x^n(1), x^n(2), \ldots, x^n(2^{\lceil nR \rceil})\}$ is the codebook
  - Decoder: an estimate $\hat{m}(y^n) \in [1:2^{nR}] \cup \{e\}$ for each $y^n$

# Point-to-point communication system




- Assume that $M \sim \text{Unif}[1 : 2^{nR}]$
- Average probability of error: $P_e^{(n)} = \mathsf{P}\{\hat{M} \neq M\}$
- $R$ achievable if $\exists$ a sequence of $(2^{nR}, n)$ codes such that $\lim_{n\to\infty} P_e^{(n)} = 0$
- Capacity $C$: Supremum of all achievable rates (operational capacity)
- For $(X, Y) \sim p(x, y)$, define the mutual information as

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) = H(X) + H(Y) - H(X,Y)$$

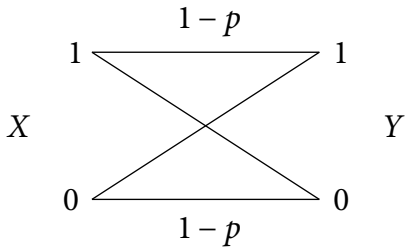
## Channel coding theorem (Shannon 1948)

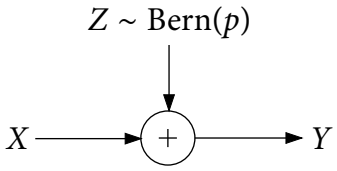
$$C = \max_{p(x)} I(X;Y) \text{ bits/transmission} \quad \text{(information capacity)}$$



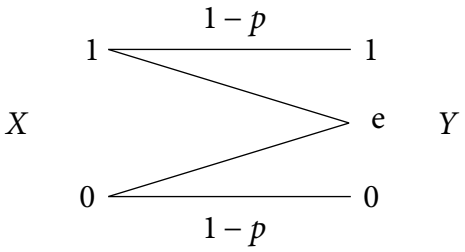
# Examples

- Binary symmetric channel (BSC): $C = 1 - H(p)$




- Binary erasure channel (BEC): $C = 1 - p$



# Proving the channel coding theorem

- Achievability: For every $R < C = \max_{p(x)} I(X;Y)$ $\exists$ a sequence of $(2^{nR}, n)$ codes with $\lim_{n\to\infty} P_e^{(n)} = 0$
  - We will use random coding and joint typicality decoding
- Converse: Given a sequence of $(2^{nR}, n)$ codes with $\lim_{n\to\infty} P_e^{(n)} = 0$, $R \le C = \max_{p(x)} I(X;Y)$
  - Need some properties of mutual information



# Jointly typical sequences

- Joint type of $(x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n$:

$$\pi(x, y|x^n, y^n) = \frac{\left|\{i: (x_i, y_i) = (x, y)\}\right|}{n} \quad \text{for } (x, y) \in \mathcal{X} \times \mathcal{Y}$$

- Jointly typical set: For $(X, Y) \sim p(x, y)$ and $\epsilon > 0$,

$$\begin{aligned} \mathcal{T}_\epsilon^{(n)}(X, Y) &= \mathcal{T}_\epsilon^{(n)}((X, Y)) \\ &= \{(x^n, y^n): |\pi(x, y|x^n, y^n) - p(x, y)| \le \epsilon p(x, y) \text{ for all } (x, y)\} \end{aligned}$$

- If $(x^n, y^n) \in \mathcal{T}_\epsilon^{(n)}(X, Y)$ and $p(x^n, y^n) = \prod_{i=1}^n p_{X,Y}(x_i, y_i)$, then
  - $x^n \in \mathcal{T}_\epsilon^{(n)}(X)$ and $y^n \in \mathcal{T}_\epsilon^{(n)}(Y)$
  - $p(x^n) \doteq 2^{-nH(X)}$, $p(y^n) \doteq 2^{-nH(Y)}$, and $p(x^n, y^n) \doteq 2^{-nH(X,Y)}$
  - $p(x^n|y^n) \doteq 2^{-nH(X|Y)}$ and $p(y^n|x^n) \doteq 2^{-nH(Y|X)}$

# Conditionally typical sequences

- Conditionally typical set: $\mathcal{T}_\epsilon^{(n)}(Y|x^n) = \{y^n : (x^n, y^n) \in \mathcal{T}_\epsilon^{(n)}(X, Y)\}$
- $|\mathcal{T}_\epsilon^{(n)}(Y|x^n)| \le 2^{n(H(Y|X)+\delta(\epsilon))}$

## Conditional typicality lemma

If $x^n \in \mathcal{T}_{\epsilon'}^{(n)}(X)$, $Y^n \sim \prod_{i=1}^n p_{Y|X}(y_i|x_i)$, and $\epsilon > \epsilon'$, then

$$\lim_{n\to\infty} \mathrm{P}\{(x^n, Y^n) \in \mathcal{T}_\epsilon^{(n)}(X, Y)\} = 1$$

- If $x^n \in \mathcal{T}_{\epsilon'}^{(n)}(X)$, $\epsilon > \epsilon'$, then for $n$ sufficiently large,

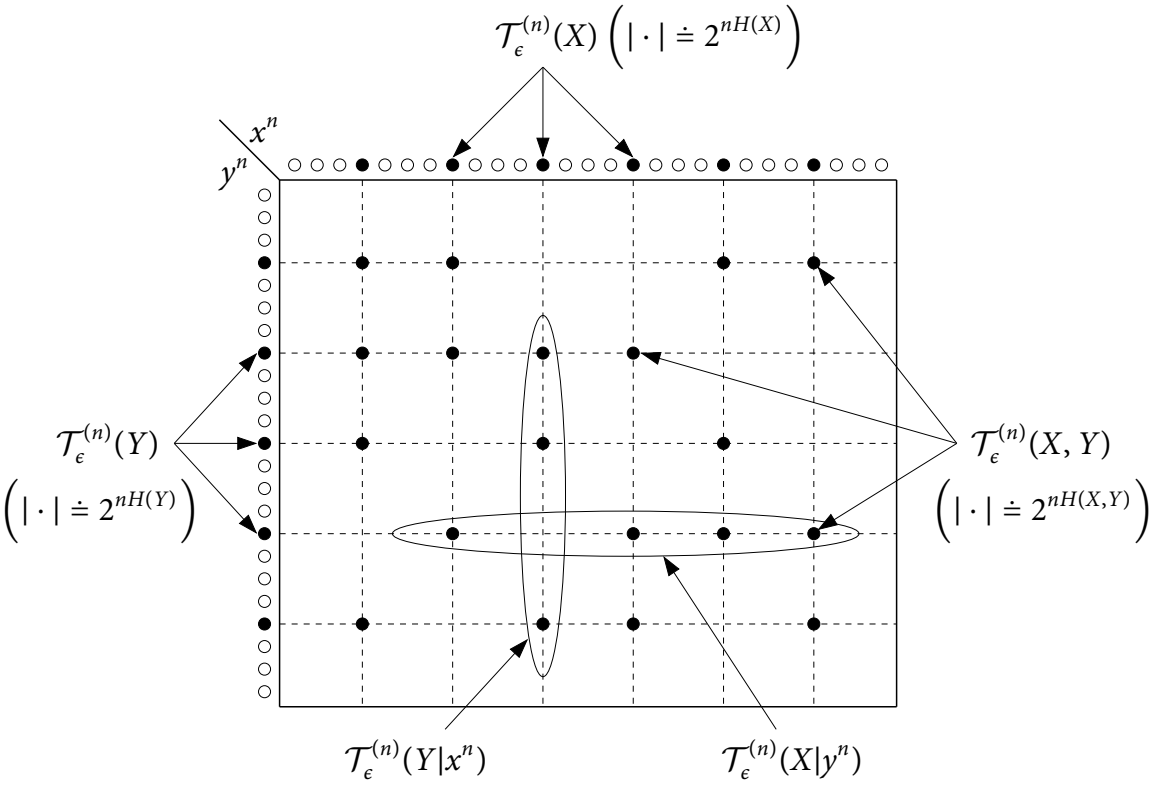
$$|\mathcal{T}_\epsilon^{(n)}(Y|x^n)| \ge 2^{n(H(Y|X)-\delta(\epsilon))}$$

- Let $X \sim p(x)$, $Y = g(X)$, and $x^n \in \mathcal{T}_\epsilon^{(n)}(X)$. Then

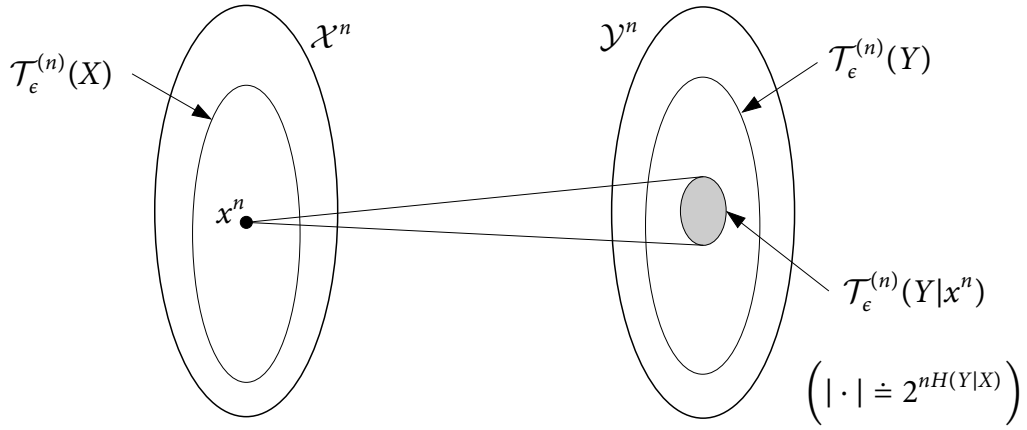
$$y^n \in \mathcal{T}_\epsilon^{(n)}(Y|x^n) \quad \text{iff} \quad y_i = g(x_i),\ i \in [1:n]$$



# Illustration of joint typicality



# Another illustration of joint typicality






# Joint typicality lemma

Let $(X, Y) \sim p(x, y)$ and $\epsilon > \epsilon'$. Then for some $\delta(\epsilon) \to 0$ as $\epsilon \to 0$:

- If $\tilde{x}^n$ is arbitrary and $\tilde{Y}^n \sim \prod_{i=1}^n p_Y(\tilde{y}_i)$, then

$$\mathsf{P}\{(\tilde{x}^n, \tilde{Y}^n) \in \mathcal{T}_\epsilon^{(n)}(X, Y)\} \leq 2^{-n(I(X;Y)-\delta(\epsilon))}$$

- If $x^n \in \mathcal{T}_{\epsilon'}^{(n)}$ and $\tilde{Y}^n \sim \prod_{i=1}^n p_Y(\tilde{y}_i)$, then for $n$ sufficiently large,

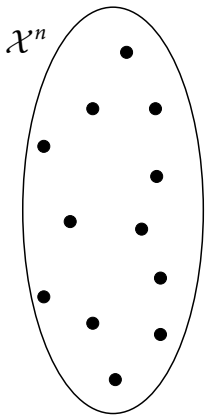
$$\mathsf{P}\{(x^n, \tilde{Y}^n) \in \mathcal{T}_\epsilon^{(n)}(X, Y)\} \geq 2^{-n(I(X;Y)+\delta(\epsilon))}$$

- Corollary: If $(\tilde{X}^n, \tilde{Y}^n) \sim \prod_{i=1}^n p_X(\tilde{x}_i)p_Y(\tilde{y}_i)$, then

$$\mathsf{P}\{(\tilde{X}^n, \tilde{Y}^n) \in \mathcal{T}_\epsilon^{(n)}\} \doteq 2^{-nI(X;Y)}$$

# Achievability proof of channel coding theorem

- For every $R < \max_{p(x)} I(X;Y)$ $\exists$ sequence of $(2^{nR}, n)$ codes with $\lim_{n\to\infty} P_e^{(n)} = 0$
- Key ideas: random coding and joint typicality decoding
- Codebook generation:
  - Fix $p(x)$ that attains $C = \max_{p(x)} I(X;Y)$
  - Independently generate $2^{nR}$ sequences $x^n(m) \sim \prod_{i=1}^n p_X(x_i)$, $m \in [1 : 2^{nR}]$; hence

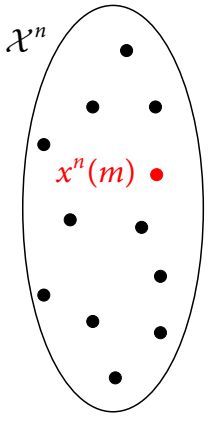



$$p(\mathcal{C}) = \prod_{m=1}^{2^{nR}} \prod_{i=1}^{n} p_X(x_i(m))$$



# Achievability proof of channel coding theorem

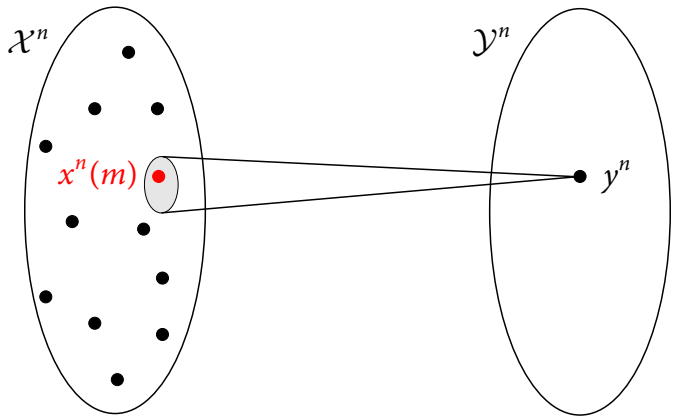
- For every $R < \max_{p(x)} I(X;Y)$ $\exists$ sequence of $(2^{nR}, n)$ codes with $\lim_{n\to\infty} P_e^{(n)} = 0$
- Key ideas: random coding and joint typicality decoding
- Encoding: $\mathcal{C}$ is revealed to both encoder and decoder
  - To send message $m$, transmit $x^n(m)$




19 / 37

# Achievability proof of channel coding theorem

- For every $R < \max_{p(x)} I(X;Y)$ $\exists$ sequence of $(2^{nR}, n)$ codes with $\lim_{n\to\infty} P_e^{(n)} = 0$
- Key ideas: random coding and joint typicality decoding
- Decoding:
  - Declare that $\hat{m}$ is sent if it is unique message such that $(x^n(\hat{m}), y^n) \in \mathcal{T}_\epsilon^{(n)}$
  - Otherwise declare an error e






# Analysis of the probability of error

- Consider $\mathsf{P}(\mathcal{E})$ averaged over codebooks
- Observe that $\mathsf{P}(\mathcal{E}) = \mathsf{P}(\mathcal{E}|M = 1)$; hence assume that $M = 1$ is sent
- Error events:

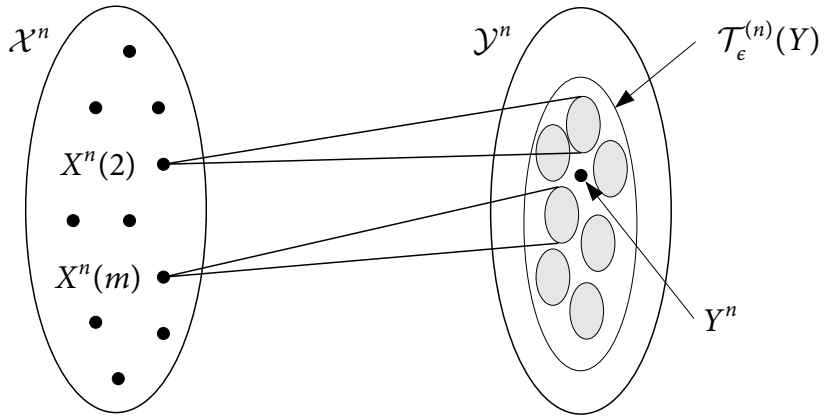
$$\mathcal{E}_1 = \{(X^n(1), Y^n) \notin \mathcal{T}_\epsilon^{(n)}\},$$
$$\mathcal{E}_2 = \{(X^n(m), Y^n) \in \mathcal{T}_\epsilon^{(n)} \text{ for some } m \neq 1\}$$

  By the union of events bound, $\mathsf{P}(\mathcal{E}) = \mathsf{P}(\mathcal{E}_1 \cup \mathcal{E}_2) \leq \mathsf{P}(\mathcal{E}_1) + \mathsf{P}(\mathcal{E}_2)$
- By the LLN, $\mathsf{P}(\mathcal{E}_1) \to 0$ (as $n \to \infty$)
- By the union of events bound and the joint typicality lemma,

$$\mathsf{P}(\mathcal{E}_2) \leq \sum_{m=2}^{2^{nR}} \mathsf{P}\{(X^n(m), Y^n) \in \mathcal{T}_\epsilon^{(n)}\} \leq 2^{-n(C-R-\delta(\epsilon))},$$

  which $\to 0$ as $n \to \infty$ if $R < C - \delta(\epsilon)$
- Hence, $\exists$ sequence of $(2^{nR}, n)$ codes with $\lim_{n\to\infty} P_e^{(n)} = 0$ if $R < C - \delta(\epsilon)$

# Illustration of $\mathcal{E}_2$




- Note that we only needed $X^n(m)$, $m \in [2 : 2^{nR}]$, to be pairwise independent of $Y^n$



# "Little" packing lemma

- Let $(X, Y) \sim p(x, y)$
- Let $\tilde{Y}^n \sim \prod_{i=1}^n p_Y(\tilde{y}_i)$
- Let $X^n(m) \sim \prod_{i=1}^n p_X(x_i)$, $m \in \mathcal{A}$, $|\mathcal{A}| \le 2^{nR}$, be pairwise independent of $\tilde{Y}^n$

## "Little" packing lemma

There exists $\delta(\epsilon) \to 0$ as $\epsilon \to 0$ such that

$$\lim_{n\to\infty} \mathrm{P}\{(X^n(m), \tilde{Y}^n) \in \mathcal{T}_\epsilon^{(n)} \text{ for some } m \in \mathcal{A}\} = 0,$$

if $R < I(X; Y) - \delta(\epsilon)$

- We will generalize this later (see NIT 3.2)

# Application: Achievability using linear codes

- Consider a BSC($p$) and let $m = (u_1, u_2, \ldots, u_k) \in \{0,1\}^k$ (i.e., $k = nR$)
- Random linear codebook: Generator matrix $G$ with i.i.d. $\mathrm{Bern}(1/2)$ entries

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} g_{11} & g_{12} & \cdots & g_{1k} \\ g_{21} & g_{22} & \cdots & g_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ g_{n1} & g_{n2} & \cdots & g_{nk} \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_k \end{bmatrix}$$

  - $X_1(u^k), \ldots, X_n(u^k)$ are i.i.d. $\mathrm{Bern}(1/2)$ for each $u^k \neq 0$
  - $X^n(u^k)$ and $X^n(\tilde{u}^k)$ are independent for each $u^k \neq \tilde{u}^k$
- By the "little" packing lemma, $\mathsf{P}(\mathcal{E}) \to 0$ if $R < 1 - H(p) - \delta(\epsilon)$
- There exists a good sequence of linear codes
- There are now practical randomly generated linear codes (turbo, LDPC)



# Properties of mutual information

- Nonnegativity:

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) = H(X) + H(Y) - H(X,Y) \geq 0$$

- Conditional mutual information:

$$I(X;Y|Z) = \sum_{z \in \mathcal{Z}} I(X;Y|Z=z)p(z) = H(X|Z) - H(X|Y,Z)$$

- Mutual information versus conditional mutual information:
  - Conditional independence: If $Z \to X \to Y$ form a Markov chain, $I(X;Y|Z) \leq I(X;Y)$
  - Independence: If $p(x,y,z) = p(z)p(x)p(y|x,z)$, $I(X;Y|Z) \geq I(X;Y)$
- Chain rule:

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

- Data processing inequality: If $X \to Y \to Z$,

$$I(X;Z) \leq I(X;Y),$$
$$I(X;Z) \leq I(Y;Z)$$

# Converse proof of channel coding theorem

- Need to show: For any sequence of $(2^{nR}, n)$ codes with $P_e^{(n)} \to 0$, $R \le C$
- Each $(2^{nR}, n)$ code induces empirical pmf

$$p(m, x^n, y^n, \hat{m}) = 2^{-nR} p(x^n | m) \prod_{i=1}^{n} p_{Y|X}(y_i | x_i)\, p(\hat{m} | y^n)$$

- Note that

$$\begin{aligned} nR &= H(M) \\ &= I(M; \hat{M}) + H(M|\hat{M}) \end{aligned}$$

- By Fano's inequality,

$$H(M|\hat{M}) \le 1 + P_e^{(n)} nR = n\epsilon_n,$$

  where $\epsilon_n \to 0$ as $n \to \infty$
- By the data processing inequality,

$$\begin{aligned} nR &= I(M; \hat{M}) + n\epsilon_n \\ &\le I(M; Y^n) + n\epsilon_n \end{aligned}$$



# Proof of the converse

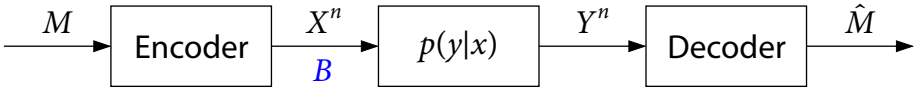
- We have

$$nR \le I(M; Y^n) + n\epsilon_n$$

- Now need to show: $I(M; Y^n) \le n \max_{p(x)} I(X; Y)$

$$\begin{aligned} I(M; Y^n) &= \sum_{i=1}^{n} I(M; Y_i | Y^{i-1}) \\ &\le \sum_{i=1}^{n} I(M, Y^{i-1}; Y_i) \\ &= \sum_{i=1}^{n} I(X_i, M, Y^{i-1}; Y_i) \quad (X_i \text{ is function of } M) \\ &= \sum_{i=1}^{n} I(X_i; Y_i) \quad ((M, Y^{i-1}) \to X_i \to Y_i) \\ &\le n \max_{p(x)} I(X; Y) \end{aligned}$$

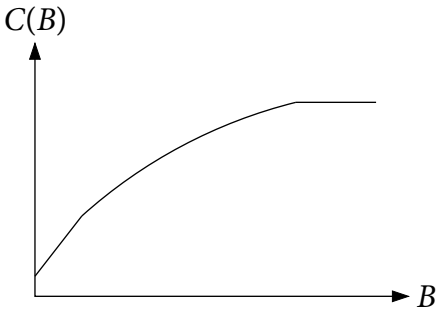
# Channel coding with input cost




- Cost $b(x) \geq 0$ with $b(x_0) = 0$
- Average cost constraint: $\sum_{i=1}^{n} b(x_i(m)) \leq nB, \quad m \in [1 : 2^{nR}]$
- Define capacity–cost function $C(B)$ as $C$

## Capacity–cost function

$$C(B) = \max_{p(x):\mathsf{E}(b(X))\leq B} I(X;Y)$$






# Proof of achievability

- Codebook generation:
  - Fix $p(x)$ that attains $C(B/(1+\epsilon))$
  - Independently generate $2^{nR}$ sequences $x^n(m) \sim \prod_{i=1}^{n} p_X(x_i)$, $m \in [1 : 2^{nR}]$
- Encoding:
  - To send message $m$, transmit $x^n(m)$ if $x^n(m) \in \mathcal{T}_\epsilon^{(n)}$
    (by the typical average lemma, $\sum_{i=1}^{n} b(x_i(m)) \leq nB$)
  - Otherwise transmit $(x_0, \ldots, x_0)$
- Decoding:
  - Declare that $\hat{m}$ is sent if it is unique message such that $(x^n(\hat{m}), y^n) \in \mathcal{T}_\epsilon^{(n)}$
  - Otherwise declare an error
- Analysis of the probability of error: Read NIT 3.3

## Proof of the converse

- Need to show: For any sequence of codes with $P_e^{(n)} \to 0$ and $\sum_{i=1}^n b(x_i(m)) \le nB$,

$$R \le C(B) = \max_{p(x):\mathsf{E}(b(X))\le B} I(X;Y)$$

- By Fano's inequality and the data processing inequality,

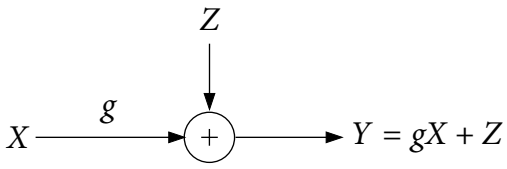
$$\begin{aligned} nR &\le \sum_{i=1}^n I(X_i;Y_i) + n\epsilon_n \\ &\le \sum_{i=1}^n C(\mathsf{E}[b(X_i)]) + n\epsilon_n \quad \text{(by definition)} \\ &\le nC\left(\frac{1}{n}\sum_{i=1}^n \mathsf{E}[b(X_i)]\right) + n\epsilon_n \quad \text{(concavity of } C(B)\text{)} \\ &\le nC(B) + n\epsilon_n \quad \text{(monotonicity of } C(B)\text{)} \end{aligned}$$

- Hence, as $n \to \infty$, $R \le C(B)$



## Gaussian channel

- Discrete-time additive white Gaussian noise channel




  - $g$: channel gain (path loss)
  - $Z \sim \mathrm{N}(0, N_0/2)$ ($\{Z_i\}$: WGN($N_0/2$) process, independent of $M$)
- Average power constraint: $\sum_{i=1}^n x_i^2(m) \le nP$ for every $m \in [1 : 2^{nR}]$
- Assume $N_0/2 = 1$ and label received power $g^2P$ as $S$ (SNR)

**Theorem 3.3. (Shannon 1948)**

$$C = \frac{1}{2}\log(1+S) = \mathsf{C}(S)$$

- To prove this result we need differential entropy

# Differential entropy

- Differential entropy of a continuous random variable $X \sim f(x)$ (pdf):

$$h(X) = -\int f(x) \log f(x)\, dx = -\mathsf{E}_X\big(\log f(X)\big)$$

  - Concave function of $f(x)$ (but not necessarily nonnegative)
  - Examples: $h(\mathrm{Unif}[a,b]) = \log(b-a)$, $h(\mathrm{N}(\mu,\sigma^2)) = (1/2)\log(2\pi e\sigma^2)$
  - Translation: $h(X+a) = h(X)$
  - Scaling: $h(aX) = h(X) + \log|a|$

- Maximum differential entropy under average power constraint:

$$\max_{f(x):\mathsf{E}(X^2)\le P} h(X) = \frac{1}{2}\log(2\pi e P) = h(\mathrm{N}(0,P))$$

  Thus, for any $X \sim f(x)$,

$$h(X) = h(X - \mathsf{E}(X)) \le \frac{1}{2}\log(2\pi e\,\mathrm{Var}(X)) \le \frac{1}{2}\log(2\pi e\,\mathsf{E}(X^2))$$



# Differential entropy

- Conditional differential entropy: If $X \sim F(x)$ and $Y|\{X=x\} \sim f(y|x)$,

$$h(Y|X) = \int h(Y \mid X = x)\, dF(x) = -\mathsf{E}_{X,Y}\big(\log f(Y|X)\big)$$

  - $h(Y|X) \le h(Y)$ (with equality if $X$ and $Y$ are independent)

- For continuous $(X,Y) \sim f(x,y)$,

$$I(X;Y) = h(X) - h(X|Y) = h(Y) - h(Y|X) = h(X) + h(Y) - h(X,Y)$$

- If $X \sim p(x)$ is discrete and $Y|\{X=x\} \sim f(y|x)$ is continuous for each $x$,

$$I(X;Y) = h(Y) - h(Y|X) = H(X) - H(X|Y)$$

## Proof of the converse

- Mutual information extends to arbitrary random variables (Pinsker 1964)
- Hence, by the converse proof for the DMC with cost,

$$C \le \sup_{F(x):\mathsf{E}(X^2)\le P} I(X;Y)$$

- Now consider any $X$ with $\mathsf{E}(X^2) \le P$, thus $\mathsf{E}(Y^2) \le g^2P + 1 = S + 1$

$$\begin{aligned} I(X;Y) &= h(Y) - h(Y|X) \\ &= h(Y) - h(Y - gX|X) \\ &= h(Y) - h(Z|X) \\ &= h(Y) - h(Z) \\ &\le \frac{1}{2}\log(2\pi e(1+S)) - \frac{1}{2}\log(2\pi e) = \mathsf{C}(S) \end{aligned}$$

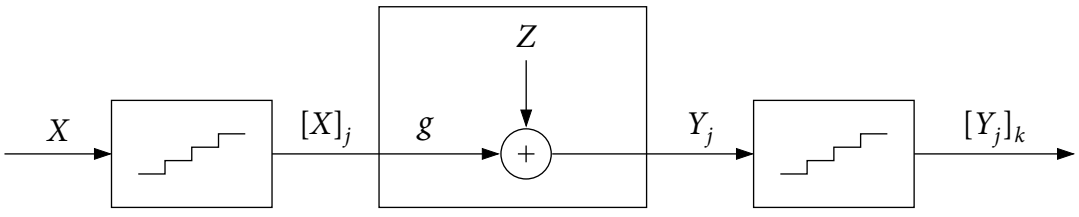
- Finally note that setting $X \sim \mathrm{N}(0,P)$, $I(X;Y) = \mathsf{C}(S)$; hence

$$C \le \max_{F(x):\mathsf{E}(X^2)\le P} I(X;Y) = \mathsf{C}(S)$$



## Proof of achievability

- Extend proof for DMC with cost via discretization procedure (McEliece 1977)
- First note that capacity is attained by $X \sim \mathrm{N}(0,P)$
- Let $[X]_j$ be a finite quantization of $X$ with $[X]_j \to X$ in distribution, $\mathsf{E}([X]_j^2) \le P$



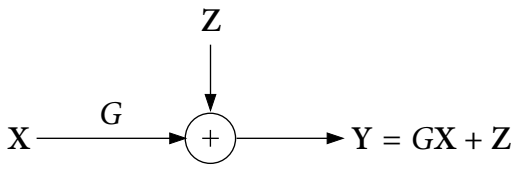

- Let $[Y_j]_k$ be a finite quantization of $Y_j = g[X]_j + Z$ such that $[Y_j]_k \to T_j$ in distribution

- By achievability proof for DMC with cost, $I([X]_j;[Y_j]_k)$ is achievable for every $j,k$
- By weak convergence and the dominated convergence theorem (see NIT 3.4.1),

$$\lim_{j\to\infty}\lim_{k\to\infty} I([X]_j;[Y_j]_k) = \lim_{j\to\infty} I([X]_j;Y_j) = I(X;Y) = \mathsf{C}(S)$$

# Gaussian vector (MIMO) channel

- Discrete-time additive white Gaussian noise multiple-antenna channel




  - $\mathbf{X}$: $t$-vector, $\mathbf{Y}$: $r$-vector
  - $G$: channel gain matrix with gain $G_{jk}$ from transmitter antenna $k$ to receiver antenna $j$
  - $\mathbf{Z} \sim \mathrm{N}(0, I_r)$

- Average power constraint: $\sum_{i=1}^{n} \mathbf{x}^T(m,i)\mathbf{x}(m,i) \le nP$, $m \in [1 : 2^{nR}]$

## Theorem 9.1

$$C = \max_{F(\mathbf{x}):\mathsf{E}(\mathbf{X}^T\mathbf{X})\le P} I(\mathbf{X};\mathbf{Y}) = \max_{K_{\mathbf{X}}\succeq 0:\mathrm{tr}(K_{\mathbf{X}})\le P} \frac{1}{2}\log|GK_{\mathbf{X}}G^T + I_r|$$



# Summary

- Lossless source coding problem
- Discrete memoryless source
- Entropy is the limit on lossless source coding
- Proof of coding theorem: achievability and the converse
- Channel coding problem
- Discrete memoryless channel (DMC), e.g., BSC and BEC
- Information capacity is the limit on channel coding
  - Random codebook generation
  - Joint typicality decoding
  - "Little" packing lemma
  - Capacity with input cost
  - Gaussian channel (discretization procedure)

# References

McEliece, R. J. (1977). *The Theory of Information and Coding*. Addison-Wesley, Reading, MA.

Orlitsky, A. and Roche, J. R. (2001). Coding for computing. *IEEE Trans. Inf. Theory*, 47(3), 903–917.

Pinsker, M. S. (1964). *Information and Information Stability of Random Variables and Processes*. Holden-Day, San Francisco.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell Syst. Tech. J.*, 27(3), 379–423, 27(4), 623–656.