2. Fundamentals of Information Theory

Outline

2. Fundamentals of Information Theory

1. **Information measures and basic inequalities**
   - Information quantities (entropy, divergence, mutual information)
   - Important properties (chain rule, conditioning reduces entropy, convexity/concavity)
   - Information inequalities (non-negativity, data processing inequality, Fano's inequality)

2. **Typicality**
   - Typical sequences and typical set
   - Joint and conditional typicality
   - Important properties and bounds
   - Packing lemma

3. **Point-to-point channel**
   - Formulation of point-to-point communication problem
   - Capacity (achievability, converse)

Entropy

Definition (Entropy)

The *entropy* $H(X)$ of a discrete random variable $X \sim P_X$ over $\mathcal{X}$ is defined by:

$$H(X) = -\sum_{x\in\mathcal{X}} P_X(x)\log(P_X(x)) = -\mathbb{E}[\log(P_X(X))].$$

- $H(X)$ measures the average amount of information contained in X or, equivalently, the amount of uncertainty
- Sometimes interchangeably denoted by $H(X) = H(P_X)$ to emphasize dependency on P_X
- Properties:
  - $H(X)$ is non-negative
  - $H(X)$ is a concave function of P_X
  - $H(X) \leq \log|\mathcal{X}|$ with equality if P_X is the uniform distribution

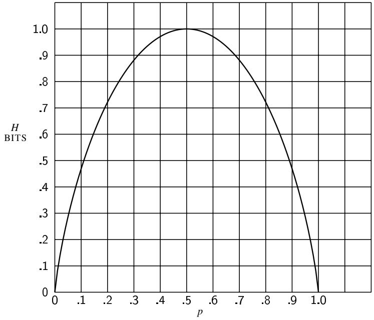
Binary Entropy Function

Example

Consider Bernoulli distribution with $\mathcal{X} = \{0, 1\}$ and $P_X(0) = p$, $P_X(1) = 1 - p$, i.e., $X \sim$ Bernoulli-p. The entropy of X is

$$H(X) = H_2(p) = -p \log p - (1 - p) \log(1 - p)$$

and $H_2(p)$ is called *binary entropy function*.



- $H_2(0) = H_2(1) = 0$
- $H_2(0.11) = H_2(0.89) \approx 0.5$
- $H_2(0.5) = 1$

Conditional Entropy

Definition (Conditional Entropy)

For two jointly distributed random variables X and Y over $\mathcal{X}$ and $\mathcal{Y}$, respectively, with joint pmf P_{XY}, the *conditional entropy of X given Y* is:

$$H(X|Y) = -\sum_{x\in\mathcal{X},y\in\mathcal{Y}} P_{XY}(x,y)\log(P_{X|Y}(x|y))$$
$$= -\mathbb{E}[\log(P_{X|Y}(X|Y))].$$

- Alternatively, conditional entropy $H(X|Y)$ can be expressed as the average of values $H(X|Y=y)$, i.e.,

$$H(X|Y) = \sum_{y\in\mathcal{Y}} P_Y(y)H(X|Y=y)$$

- $0 \leq H(X|Y) \leq H(X) \leq \log|\mathcal{X}|$ (*"conditioning reduces entropy"*)
- $H(Y|X) \neq H(X|Y)$ (non-symmetric)

Joint Entropy

Definition (Joint Entropy)

The *joint entropy of X and Y* over $\mathcal{X}$ and $\mathcal{Y}$, respectively, with joint pmf P_{XY} is:

$$\begin{aligned} H(X,Y) &= -\sum_{x\in\mathcal{X},y\in\mathcal{Y}} P_{XY}(x,y)\log(P_{XY}(x,y)) \\ &= -\mathbb{E}[\log(P_{XY}(X,Y))]. \end{aligned}$$

- Using Bayes' rule, we have

$$\begin{aligned} H(X,Y) &= H(X) + H(Y|X) \\ &= H(Y) + H(X|Y) \end{aligned}$$

- $\max\{H(X),H(Y)\} \le H(X,Y) \le \log(|\mathcal{X}||\mathcal{Y}|)$

Chain Rule for Entropy

Lemma (Chain Rule for Entropy)

For a random vector $X^n = (X_1, X_2, ..., X_n)$ *we have*

$$\begin{aligned} H(X^n) &= H(X_1) + H(X_2|X_1) + H(X_3|X_1, X_2) + ... + H(X_n|X_1, X_2, ..., X_{n-1}) \\ &= \sum_{i=1}^{n} H(X_i|X_1, ..., X_{i-1}) \\ &= \sum_{i=1}^{n} H(X_i|X^{i-1}) \end{aligned}$$

which is known as chain rule for entropy.

- "Developing" entropy in different ways:

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

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

Divergence

Definition (Divergence)

Let P_X and Q_X denote two pmfs over $\mathcal{X}$. The *divergence of P_X and Q_X* is given by

$$D(P_X \| Q_X) = \sum_{x \in \mathcal{X}} P_X(x) \log \frac{P_X(x)}{Q_X(x)} = \mathbb{E}\left[\log \frac{P_X(X)}{Q_X(X)}\right].$$

- Also known as *Kullback-Leibler Distance*, *Information Divergence*, or *Relative Entropy*
- $D(P_X \| Q_X) \geq 0$ with equality if $P_X(x) = Q_X(x)$ for all $x \in \mathcal{X}$
- Non-symmetric: $D(P_X \| Q_X) \neq D(Q_X \| P_X)$ in general
- If for some $x \in \mathcal{X}$ we have $Q_X(x) = 0$ and $P_X(x) > 0$, then $D(P_X \| Q_X) = \infty$
- It is a "*sort of distance*" between two pmfs

Conditional Divergence

Definition (Conditional Divergence)

Let $P_{Y|X}$ and $Q_{Y|X}$ denote two conditional pmfs for Y given X and let P_X denote a pmf for X. The *conditional divergence of $P_{Y|X}$ and $Q_{Y|X}$ with respect to P_X* is given by

$$\begin{aligned} D(P_{Y|X} \| Q_{Y|X} | P_X) &= \sum_{x \in \mathcal{X}} P_X(x) \sum_{y \in \mathcal{Y}} P_{Y|X}(y|x) \log \frac{P_{Y|X}(y|x)}{Q_{Y|X}(y|x)} \\ &= \mathbb{E}\left[\log \frac{P_{Y|X}(Y|X)}{Q_{Y|X}(Y|X)} \right]. \end{aligned}$$

Lemma (Chain Rule for Divergence)

For two joint pmfs $P_{XY} = P_X P_{Y|X}$ and $Q_{XY} = Q_X Q_{Y|X}$ we have

$$D(P_{XY} \| Q_{XY}) = D(P_X \| Q_X) + D(P_{Y|X} \| Q_{Y|X} | P_X).$$

Mutual Information

Definition (Mutual Information)

Let $X, Y \sim P_{XY}$. The *mutual information between X and Y* is given by

$$\begin{aligned} I(X;Y) &= \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} P_{XY}(x,y) \log \frac{P_{XY}(x,y)}{P_X(x) P_Y(y)} \\ &= D(P_{XY} \| P_X P_Y). \end{aligned}$$

- Mutual information in terms of conditional divergence:

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

Properties of Mutual Information

- *Symmetry* of mutual information:

$$I(X;Y) = I(Y;X)$$

- *Non-negativity* of mutual information:

$$I(X;Y) \geq 0$$

  with equality if X and Y are independent

- Mutual information in terms of entropy

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X) \\ &= H(X) + H(Y) - H(X,Y) \\ &= H(X,Y) - H(X|Y) - H(Y|X) \end{aligned}$$

- $I(X;X) = H(X)$

Chain Rule for Mutual Information

Lemma (Chain Rule for Mutual Information)

Let X^n and Y be jointly distributed as P_{X^nY}, then we have

$$\begin{aligned} I(X^n;Y) &= I(X_1;Y) + I(X_2;Y|X_1) + \ldots + I(X_n;Y|X_1,\ldots,X_{n-1}) \\ &= \sum_{i=1}^{n} I(X_i;Y|X_1,\ldots,X_{i-1}) \\ &= \sum_{i=1}^{n} I(X_i;Y|X^{i-1}) \end{aligned}$$

- In contrast to entropy, no general inequality relationship between $I(X;Y|Z)$ and $I(X;Y)$ exists (only special cases)

Information Inequalities (1)

Theorem (Information Inequality)

Let P_X and Q_X be two pmfs defined on $\mathcal{X}$, then

$$D(P_X \| Q_X) \geq 0$$

with equality iff $P_X(x) = Q_X(x)$ for all $x \in \mathcal{X}$ where they are both non-zero.

Corollary

$$I(X;Y) \geq 0$$

with equality iff X and Y are independent.

Information Inequalities (2)

Corollary

$$I(X;Y|Z) \geq 0$$

with equality iff X *and* Y *are conditionally independent given* Z*.*

Corollary (Conditioning reduces entropy)

$$H(Y) \geq H(Y|X)$$

with equality iff X *and* Y *are independent.*

Information Inequalities (3)

Corollary (Uniform pmf maximizes entropy)

For $X \sim P_X$ on $\mathcal{X}$ of size $|\mathcal{X}|$ we have

$$H(X) \leq \log |\mathcal{X}|$$

with equality iff X is uniform over $|\mathcal{X}|$.

Theorem (Independence bound on joint entropy)

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

with equality iff X^n has independent components.

Data Processing and Fano's Inequalities

Theorem (Data Processing Inequality)

If $X - Y - Z$ forms a Markov chain, i.e., $P_{XYZ} = P_X P_{Y|X} P_{Z|Y}$, then

$$I(X;Z) \leq I(Y;Z) \qquad \text{and} \qquad I(X;Z) \leq I(X;Y).$$

Theorem (Fano's Inequality)

Let $(X, \hat{X}) \sim P_{X\hat{X}}$ bet two jointly distributed random variables taking values in the same alphabet $\mathcal{X}$, and define $P_e = \mathbb{P}(X \neq \hat{X})$. Then

$$H(X|\hat{X}) \leq H_2(P_e) + P_e \log(|\mathcal{X}| - 1) \leq 1 + P_e \log |\mathcal{X}|.$$

Convexity Properties

Theorem (Convexity of divergence)

The divergence $D(P_X \| Q_X)$ is convex in the pair (P_X, Q_X), i.e., for distributions $P_X^{(1)}$, $P_X^{(2)}$, $Q_X^{(1)}$, $Q_X^{(2)}$ on the same alphabet $\mathcal{X}$ we have

$$\begin{aligned}\lambda D(P_X^{(1)} \| Q_X^{(1)}) + (1-\lambda) D(P_X^{(2)} \| Q_X^{(2)}) \\ \geq D(\lambda P_X^{(1)} + (1-\lambda) P_X^{(2)} \| \lambda Q_X^{(1)} + (1-\lambda) Q_X^{(2)})\end{aligned}$$

for any λ satisfying $0 \leq \lambda \leq 1$.

Corollary (Concavity of entropy)

The entropy $H(X) = H(P_X)$ is concave in P_X, i.e., for two distributions $P_X^{(1)}$ and $P_X^{(2)}$ on the same alphabet $\mathcal{X}$ we have

$$\lambda H(P_X^{(1)}) + (1-\lambda) H(P_X^{(2)}) \leq H(\lambda P_X^{(1)} + (1-\lambda) P_X^{(2)})$$

for any λ satisfying $0 \leq \lambda \leq 1$.

Convexity Properties (2)

Corollary (Concavity/convexity of mutual information)

Mutual information $I(X;Y) = I(P_X, P_{Y|X})$ *is concave in* P_X *if* $P_{Y|X}$ *is fixed, and* $I(P_X, P_{Y|X})$ *is convex in* $P_{Y|X}$ *if* P_X *is fixed.*

Outline

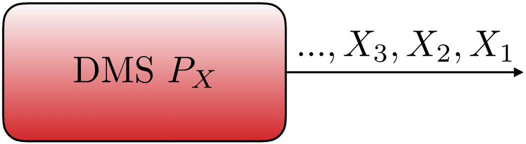
2. Fundamentals of Information Theory

1. **Information measures and basic inequalities**
   - Information quantities (entropy, divergence, mutual information)
   - Important properties (chain rule, conditioning reduces entropy, convexity/concavity)
   - Information inequalities (non-negativity, data processing inequality, Fano's inequality)

2. **Typicality**
   - Typical sequences and typical set
   - Joint and conditional typicality
   - Important properties and bounds
   - Packing lemma

3. **Point-to-point channel**
   - Formulation of point-to-point communication problem
   - Capacity (achievability, converse)

Discrete Memoryless Source



Example

Consider discrete memoryless source (DMS) that emits i.i.d. symbols $X_1, X_2, X_3, \ldots$ from a discrete and finite alphabet $\mathcal{X} = \{0, 1\}$. The source output distribution P_X is

$$P_X(0) = 2/3 \qquad \text{and} \qquad P_X(1) = 1/3.$$

Discrete Memoryless Source (2)

- Consider sequences of length 18. One generated by the DMS and three artificially generated sequences.
  a) 0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
  b) 1,0,1,1,0,1,0,1,1,1,0,0,0,0,1,0,1,0
  c) 0,0,0,1,1,0,0,1,0,0,1,1,0,0,0,1,1,0
  d) 1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1

- Which sequence has been generated by the DMS?

Discrete Memoryless Source (3)

- Consider sequences of length 18. One generated by the DMS and three artificially generated sequences.
  a) 0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
  b) 1,0,1,1,0,1,0,1,1,1,0,0,0,0,1,0,1,0
  c) 0,0,0,1,1,0,0,1,0,0,1,1,0,0,0,1,1,0
  d) 1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1

- Compute the probabilities that these sequences were emitted by the DMS
  a) $(2/3)^{18} \cdot (1/3)^{0} \approx 6.77 \cdot 10^{-4}$
  b) $(2/3)^{9} \cdot (1/3)^{9} \approx 1.32 \cdot 10^{-6}$
  c) $(2/3)^{11} \cdot (1/3)^{7} \approx 5.29 \cdot 10^{-6}$
  d) $(2/3)^{0} \cdot (1/3)^{18} \approx 2.58 \cdot 10^{-9}$

- Which sequence has been generated by the DMS?

Discrete Memoryless Source (4)

- Consider sequences of length 18. One generated by the DMS and three artificially generated sequences.
  a) 0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
  b) 1,0,1,1,0,1,0,1,1,1,0,0,0,0,1,0,1,0
  c) **0,0,0,1,1,0,0,1,0,0,1,1,0,0,0,1,1,0**
  d) 1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1

- Compute the probabilities that these sequences were emitted by the DMS
  a) $(2/3)^{18} \cdot (1/3)^{0} \approx 6.77 \cdot 10^{-4}$
  b) $(2/3)^{9} \cdot (1/3)^{9} \approx 1.32 \cdot 10^{-6}$
  c) $\mathbf{(2/3)^{11} \cdot (1/3)^{7} \approx 5.29 \cdot 10^{-6}}$
  d) $(2/3)^{0} \cdot (1/3)^{18} \approx 2.58 \cdot 10^{-9}$

- Answer: **Sequence c)** was emitted by the DMS. Why is this intuition correct? We need a concept of *"typical"* sequences.

Typical Sequences

- Let $x^n \in \mathcal{X}^n$. The *empirical pmf of x^n* is defined as

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

  This is also referred to as the "*type*" of x^n.

- Let X^n denote an i.i.d. random vector with $X_i \sim P_X$. By the (weak) law of large number

$$\lim_{n\to\infty} \pi(x|X^n) \overset{p}{=} P_X(x), \quad \text{for } x \in \mathcal{X}$$

Definition (Typical set)

For a given pmf P_X on $\mathcal{X}$ and $\epsilon > 0$, the *ϵ-typical set of sequences $x^n \in \mathcal{X}^n$* is defined as

$$\mathcal{T}_\epsilon^{(n)}(X) = \{x^n \in \mathcal{X}^n : |\pi(x|x^n) - P_X(x)| \le \epsilon P_X(x), \quad \forall x \in \mathcal{X}\}$$

Asymptotic Equipartition Property (AEP)

Lemma (Asymptotic Equipartition Property (AEP))

All typical sequences have roughly the same probability. For each $x^n \in \mathcal{T}_\epsilon^{(n)}(X)$ we have:

$$2^{-n(H(X)+\delta(\epsilon))} \leq P_{X^n}(x^n) \leq 2^{-n(H(X)-\delta(\epsilon))}$$

where $\delta(\epsilon) \downarrow 0$ as $\epsilon \to 0$. In short, we write $P_{X^n}(x^n) \doteq 2^{-nH(X)}$.

Properties of the Typical Set

- Typical set *cardinality upper bound*:

$$\left|\mathcal{T}_{\epsilon}^{(n)}(X)\right| \leq 2^{n(H(X)+\delta(\epsilon))}$$

- Law of Large Numbers (LLN): if X^n is an i.i.d. sequence with $X_i \sim P_X(x)$ then

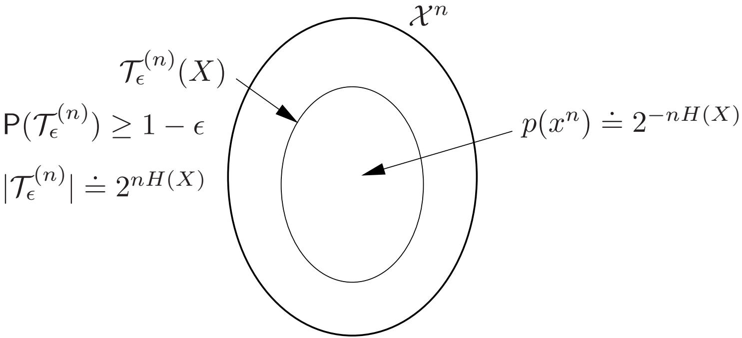
$$\lim_{n\to\infty} \mathbb{P}\left(X^n \in \mathcal{T}_{\epsilon}^{(n)}(X)\right) = 1$$

- Typical set *cardinality lower bound*:

$$\left|\mathcal{T}_{\epsilon}^{(n)}(X)\right| \geq (1-\epsilon)2^{n(H(X)-\delta(\epsilon))}$$

for sufficiently large n.

Intuitive Representation



Jointly Typical Sequences

- Let $x^n, y^n \in \mathcal{X}^n \times \mathcal{Y}^n$. The empirical joint pmf of (x^n, y^n) is defined as

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

Definition (Jointly typical set)

For a joint pmf $P_{XY}(x, y)$ and $\epsilon > 0$, the jointly *ϵ-typical set of sequence pairs* $(x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n$ is defined as

$$\mathcal{T}_\epsilon^{(n)}(X, Y) = \{(x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n : \\ |\pi(x, y|x^n, y^n) - P_{XY}(x, y)| \leq \epsilon P_{XY}(x, y), \quad \forall (x, y) \in \mathcal{X} \times \mathcal{Y}\}$$

Properties of the Jointly Typical Set

- Let (X^n, Y^n) be a jointly distributed, component-wise i.i.d., pair of random vectors with $(X_i, Y_i) \sim P_{XY}$ and let $(x^n, y^n) \in \mathcal{T}_\epsilon^{(n)}(X, Y)$, then the following properties hold:
  1. $x^n \in \mathcal{T}_\epsilon^{(n)}(X)$ and $y^n \in \mathcal{T}_\epsilon^{(n)}(Y)$.
  2. $P_{X^n Y^n}(x^n, y^n) \doteq 2^{-nH(X,Y)}$.
  3. $P_{X^n}(x^n) \doteq 2^{-nH(X)}$ and $P_{Y^n}(y^n) \doteq 2^{-nH(Y)}$.
  4. $P_{X^n|Y^n}(x^n|y^n) \doteq 2^{-nH(X|Y)}$ and $P_{Y^n|X^n}(y^n|x^n) \doteq 2^{-nH(Y|X)}$.

Conditional Typicality

Lemma (Conditional typicality lemma)

Let $\epsilon > \epsilon' > 0$. *For* $x^n \in \mathcal{T}_{\epsilon'}^{(n)}(X)$, *let* $Y^n \sim P_{Y^n|X^n}(y^n|x^n) = \prod_{i=1}^n P_{Y|X}(y_i|x_i)$. *Then*

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

- Let

$$\mathcal{T}_{\epsilon}^{(n)}(Y|x^n) = \left\{ y^n \in \mathcal{Y}^n : (x^n, y^n) \in \mathcal{T}_{\epsilon}^{(n)}(X,Y) \right\}$$

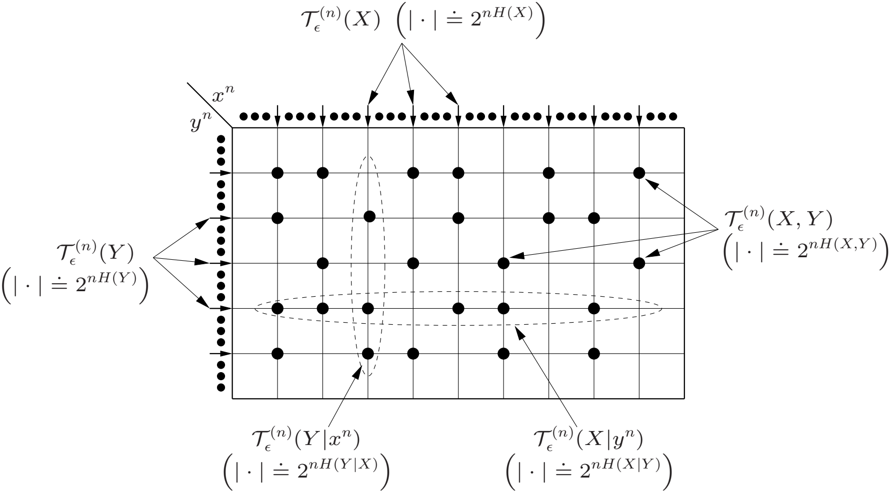
  Then

$$\left|\mathcal{T}_{\epsilon}^{(n)}(Y|x^n)\right| \leq 2^{n(H(Y|X)+\delta(\epsilon))}$$

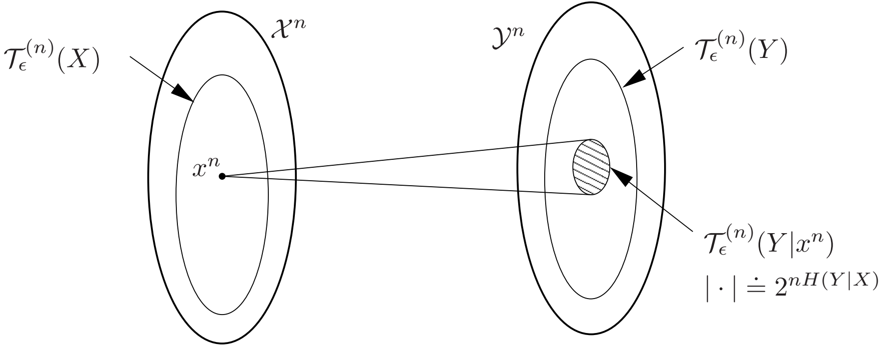
  and for sufficiently large n

$$\left|\mathcal{T}_{\epsilon}^{(n)}(Y|x^n)\right| \geq (1-\epsilon)2^{n(H(Y|X)-\delta(\epsilon))}$$

Useful Picture



Another Useful Picture



Jointly Typical Sets for Triplets

- For a pmf $P_{XYZ}(x, y, z)$ and $\epsilon > 0$, the jointly ϵ-typical set of sequences $(x^n, y^n, z^n) \in \mathcal{X}^n \times \mathcal{Y}^n \times \mathcal{Z}^n$ is defined as

$$\begin{aligned}\mathcal{T}_\epsilon^{(n)}(X, Y, Z) = \big\{ &(x^n, y^n, z^n) \in \mathcal{X}^n \times \mathcal{Y}^n \times \mathcal{Z}^n : \\ &|\pi(x, y, z | x^n, y^n, z^n) - P_{XYZ}(x, y, z)| \leq \epsilon P_{XYZ}(x, y, z) \\ &\forall (x, y, z) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{Z} \big\}\end{aligned}$$

- We can think of (X, Y, Z) as a "large" random variable, so that all the properties of the typical set seen before are inherited in this case too.

Joint Typicality Lemma

Lemma

Let $(U, X, Y) \sim P_{UXY}(u, x, y)$. *Then:*

1. *Fix two arbitrary sequences* $(u^n, x^n) \in \mathcal{U}^n \times \mathcal{X}^n$ *and let* $\tilde{Y}^n \sim \prod_{i=1}^n P_{Y|U}(y_i|u_i)$. *Hence*

$$\mathbb{P}\left((u^n, x^n, \tilde{Y}^n) \in \mathcal{T}_\epsilon^{(n)}(U, X, Y)\right) \leq 2^{-n(I(X;Y|U)-\delta(\epsilon))}.$$

2. *Let* $(\tilde{U}^n, \tilde{X}^n) \sim Q_{\tilde{U}^n\tilde{X}^n}(u^n, x^n)$ *(some arbitrary distribution), and let* $\tilde{Y}^n \sim \prod_{i=1}^n P_{Y|U}(y_i|u_i)$. *Hence*

$$\mathbb{P}\left((\tilde{U}^n, \tilde{X}^n, \tilde{Y}^n) \in \mathcal{T}_\epsilon^{(n)}(U, X, Y)\right) \leq 2^{-n(I(X;Y|U)-\delta(\epsilon))}.$$

3. *For any* $\epsilon > \epsilon' > 0$ *and sufficiently large* n, *if* $(u^n, x^n) \in \mathcal{T}_\epsilon^{(n)}(U, X)$ *and* $\tilde{Y}^n \sim \prod_{i=1}^n P_{Y|U}(y_i|u_i)$, *then*

$$\mathbb{P}\left((u^n, x^n, \tilde{Y}^n) \in \mathcal{T}_\epsilon^{(n)}(U, X, Y)\right) \geq (1-\epsilon)2^{-n(I(X;Y|U)+\delta(\epsilon))}.$$

Intuition beyond the Lemma

- A simpler case: Let $(X, Y) \sim P_{XY}(x, y)$ and consider $\tilde{Y}^n$ independent of X^n and distributed according to the product marginal pmf $\prod_{i=1}^{n} P_Y(y_i)$.
- The probability $\mathbb{P}((X^n, \tilde{Y}^n) \in \mathcal{T}_\epsilon^{(n)}(X, Y))$ is the probability that a randomly generated pair of sequences $\sim \prod_{i=1}^{n} P_X(x_i) P_Y(y_i)$ are jointly typical.
- We have $2^{nH(X)} \cdot 2^{nH(Y)}$ individually typical pairs, but only $2^{nH(XY)}$ of them are jointly typical. Since these sequences are approximately equiprobable, we have

$$\mathbb{P}((X^n, \tilde{Y^n}) \in \mathcal{T}_\epsilon^{(n)}(X, Y)) \approx \frac{2^{nH(XY)}}{2^{nH(X)} 2^{nH(Y)}} = 2^{-nI(X;Y)}$$

- In short, the mutual information is the exponent that determines the exponential decay of the probability that two independent sequences generated with the right marginal distributions "look like jointly typical".

Packing Lemma

Lemma (Packing Lemma)

Let $(U, X, Y) \sim P_{UXY}$. *Let* $(\tilde{U}^n, \tilde{Y}^n) \sim Q_{\tilde{U}^n \tilde{Y}^n}(u^n, y^n)$ *be a pair of arbitrarily distributed random sequences. Let* $X^n(m) : m \in \mathcal{A}$ *with* $|\mathcal{A}| \leq 2^{nR}$, *be a set of random sequences indexed by* m, *each distributed according to* $\prod_{i=1}^{n} P_{X|U}(x_i|\tilde{u}_i)$. *Assume that* $\{X^n(m) : m \in \mathcal{A}\}$ *are conditionally pairwise independent of* $\tilde{Y}^n$ *given* $\tilde{U}^n$ *(although they can be arbitrarily correlated among each other). Then, there exists* $\delta(\epsilon) \downarrow 0$ *as* $\epsilon \downarrow 0$ *such that for any* $\epsilon > 0$ *we have*

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

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

Proof of the Packing Lemma

- Define $\mathcal{E}_m = \{(\tilde{U}^n, X^n(m), \tilde{Y}^n) \in \mathcal{T}_\epsilon^{(n)}(U,X,Y)\}$. Then, from the Union Bound:

$$\mathbb{P}\left(\bigcup_{m\in\mathcal{A}} \mathcal{E}_m\right) \leq \sum_{m\in\mathcal{A}} \mathbb{P}(\mathcal{E}_m)$$

- Consider

$$\begin{aligned}
\mathbb{P}(\mathcal{E}_m) &= \mathbb{P}\left((\tilde{U}^n, X^n(m), \tilde{Y}^n) \in \mathcal{T}_\epsilon^{(n)}(U,X,Y)\right) \\
&= \sum_{(u^n,y^n)\in\mathcal{T}_\epsilon^{(n)}(U,X,Y)} Q_{\tilde{U}^n\tilde{Y}^n}(u^n,y^n)\mathbb{P}\left((u^n, X^n(m), y^n) \in \mathcal{T}_\epsilon^{(n)}(U,X,Y)|\tilde{U}^n = u^n\right) \\
&\leq \sum_{(u^n,y^n)\in\mathcal{T}_\epsilon^{(n)}(U,X,Y)} Q_{\tilde{U}^n\tilde{Y}^n}(u^n,y^n)2^{-n(I(X;Y|U)-\delta(\epsilon))} \quad \text{by JTL} \\
&\leq 2^{-n(I(X;Y|U)-\delta(\epsilon))}
\end{aligned}$$

- Summing over $m \in \mathcal{A}$ yields the desired result. □

Intuition beyond the Lemma

- A simpler case: Let $(X, Y) \sim P_{XY}(x, y)$ and consider $\tilde{X}^n$ independent of Y^n and distributed according to the product marginal pmf $\prod_{i=1}^n P_X(\tilde{x}_i)$.
- The probability $\mathbb{P}((\tilde{X}^n, Y^n) \in \mathcal{T}_\epsilon^{(n)}(X, Y))$ is the probability that a randomly generated pair of sequences $\sim \prod_{i=1}^n P_X(\tilde{x}_i) P_Y(y_i)$ are jointly typical.
- By the joint typicality lemma we have

$$\mathbb{P}((\tilde{X}^n, Y^n) \in \mathcal{T}_\epsilon^{(n)}(X, Y)) \approx \frac{2^{nH(XY)}}{2^{nH(X)} 2^{nH(Y)}} = 2^{-nI(X;Y)}$$

- When we have a set of size $|\mathcal{A}| = 2^{nR}$ of such vectors $\tilde{X}^n(m)$ and R is small enough, the probability that none of these vectors is jointly typical with X^n can be made arbitrarily large for sufficiently large n.

Outline

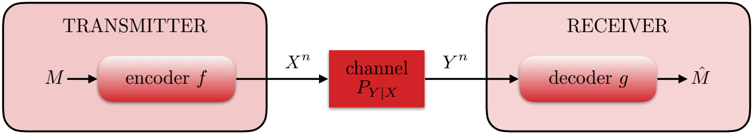
2. Fundamentals of Information Theory

1. **Information measures and basic inequalities**
    - Information quantities (entropy, divergence, mutual information)
    - Important properties (chain rule, conditioning reduces entropy, convexity/concavity)
    - Information inequalities (non-negativity, data processing inequality, Fano's inequality)

2. **Typicality**
    - Typical sequences and typical set
    - Joint and conditional typicality
    - Important properties and bounds
    - Packing lemma

3. **Point-to-point channel**
    - Formulation of point-to-point communication problem
    - Capacity (achievability, converse)

Point-to-Point Channel



Definition (Discrete Memoryless Channel)

A *discrete memoryless channel (DMC)* $(\mathcal{X}, P_{Y|X}, \mathcal{Y})$ is described by

- a finite input alphabet $\mathcal{X}$
- a finite output alphabet $\mathcal{Y}$
- and a conditional probability distribution $P_{Y|X}$

such that X denotes the channel input and Y the channel output respectively.

Examples

Example (Binary Symmetric Channel)

A *binary symmetric channel* $BSC(p)$ with cross-over probability $p \in [0,1]$ is a DMC $(\{0,1\}, P_{Y|X}, \{0,1\})$ characterized byt the transition probability matrix

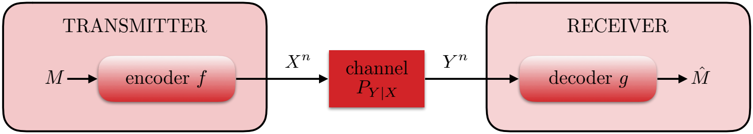
$$\begin{pmatrix} 1-p & p \\ p & 1-p \end{pmatrix}.$$

Example (Binary Erasure Channel)

A *binary erasure channel* $BEC(\epsilon)$ with erasure probability $\epsilon \in [0,1]$ is a DMC $(\{0,1\}, P_{Y|X}, \{0,?,1\})$ characterized byt the transition probability matrix

$$\begin{pmatrix} 1-\epsilon & \epsilon & 0 \\ 0 & \epsilon & 1-\epsilon \end{pmatrix}.$$

Channel Code



Definition (Code)

A $(2^{nR}, n)$ *code* $\mathcal{C}_n$ for a DMC $(\mathcal{X}, P_{Y|X}, \mathcal{Y})$ consists of

- a message set $\mathcal{M} = [1, 2^{nR}]$
- an encoding function $f : \mathcal{M} \to \mathcal{X}^n$ which maps a message m to a codeword x^n with n symbols
- a decoding function $g : \mathcal{Y}^n \to \mathcal{M} \cup \{?\}$ which maps a block of n channel outputs y^n to a message $\hat{m} \in \mathcal{M}$ or an error message $?$

The set of codewords $\{f(m) : m \in [1, 2^{nR}]\}$ is called the *codebook* of $\mathcal{C}_n$.

Achievable Rate and Capacity

- Messages are represented by a random variable M *uniformly distributed* over $\mathcal{M}$
- Rate of the code is defined as $\frac{1}{n}\log\lceil 2^{nR}\rceil$ in bits per channel use
- *Average probability of error* is defined as

$$\boldsymbol{P}_e(\mathcal{C}_n) = \mathbb{P}\left[\hat{M} \neq M \middle| \mathcal{C}_n\right]$$

Definition (Achievable Rate and Capacity)

A rate R is an *achievable rate* for the DMC $(\mathcal{X}, P_{Y|X}, \mathcal{Y})$ if there exists a sequence of $(2^{nR}, n)$ codes $\{\mathcal{C}_n\}_{n\geq 1}$ such that

$$\lim_{n\to\infty} \boldsymbol{P}_e(\mathcal{C}_n) = 0;$$

i.e., messages can be transmitted at a rate arbitrarily close to R and decoded with arbitrarily small probability of error. The *channel capacity* of the DMC is defined as

$$C = \sup\{R : R \text{ is an achievable rate}\}.$$

Achievable Rate and Capacity

- Messages are represented by a random variable M *uniformly distributed* over $\mathcal{M}$
- Rate of the code is defined as $\frac{1}{n}\log\lceil 2^{nR}\rceil$ in bits per channel use
- *Average probability of error* is defined as

$$\boldsymbol{P}_e(\mathcal{C}_n) = \mathbb{P}\left[\hat{M} \neq M \middle| \mathcal{C}_n\right]$$

Definition (Achievable Rate and Capacity)

A rate R is an *achievable rate* for the DMC $(\mathcal{X}, P_{Y|X}, \mathcal{Y})$ if there exists a sequence of $(2^{nR}, n)$ codes $\{\mathcal{C}_n\}_{n\geq 1}$ such that

$$\lim_{n\to\infty} \boldsymbol{P}_e(\mathcal{C}_n) = 0;$$

i.e., messages can be transmitted at a rate arbitrarily close to R and decoded with arbitrarily small probability of error. The *channel capacity* of the DMC is defined as

$$C = \sup\{R : R \text{ is an achievable rate}\}.$$

Remarks

- Typical goal of information theory is to characterize *achievable rates on the basis of information-theoretic quantities* that depend only on the given probability distributions and not on the block length n
- *Achievability proof* confirms the existence of codes for a class of achievable rates (also known as direct part)
- *Converse proof* asserts that codes with certain properties do not exist

➠ **Coding theorem = achievability + converse**

- *Formulation of the problem does not put any constraints either on the computational complexity or on the delay of the encoding and decoding procedures. In other words, the goal is to describe the* **fundamental limits of communications systems irrespective of their technological limitations**.

Remarks

- Typical goal of information theory is to characterize *achievable rates on the basis of information-theoretic quantities* that depend only on the given probability distributions and not on the block length n

- *Achievability proof* confirms the existence of codes for a class of achievable rates (also known as direct part)
- *Converse proof* asserts that codes with certain properties do not exist

➠ **Coding theorem = achievability + converse**

- *Formulation of the problem does not put any constraints either on the computational complexity or on the delay of the encoding and decoding procedures. In other words, the goal is to describe the* **fundamental limits of communications systems irrespective of their technological limitations**.

Remarks

- Typical goal of information theory is to characterize *achievable rates on the basis of information-theoretic quantities* that depend only on the given probability distributions and not on the block length n

- *Achievability proof* confirms the existence of codes for a class of achievable rates (also known as direct part)
- *Converse proof* asserts that codes with certain properties do not exist

➠ **Coding theorem = achievability + converse**

- *Formulation of the problem does not put any constraints either on the computational complexity or on the delay of the encoding and decoding procedures. In other words, the goal is to describe the* **fundamental limits of communications systems irrespective of their technological limitations**.

Random Coding Idea

- It is possible to prove the existence of codes without having to search for explicit code constructions
  - Construct random code by drawing the symbols of codewords independently at random according to a fixed probability distribution P_X on $\mathcal{X}$
  - If the average of the probability of error taken over all possible random codebooks goes to zero for n sufficiently large, then there exists a specific code such that the error probability goes to zero for n sufficiently large. This technique is referred to *random coding*.

Lemma (Selection Lemma)

Let $X_n \in \mathcal{X}_n$ be a random variable and let $\mathcal{F}$ be a finite set of functions $f : \mathcal{X}_n \to \Re^+$ such that $|\mathcal{F}|$ does not depend on n and

$$\mathbb{E}_{X_n}[f(X_n)] \leq \delta(n) \quad \forall f \in \mathcal{F}.$$

Then, there exists a specific realization $x_n \in \mathcal{X}_n$ such that

$$f(x_n) \leq \delta(n) \quad \forall f \in \mathcal{F}.$$

Channel Coding Theorem

Theorem (Channel Coding Theorem)

The capacity of a DMC $(\mathcal{X}, P_{Y|X}, \mathcal{Y})$ *is*

$$\boldsymbol{C = \max_{P_X} I(X;Y)}.$$

In other words, if $R < C$ *then* R *is an achievable rate and achievable rate must satisfy* $R \leq C$*.*

Achievability Proof

- Choose probability distribution P_X on $\mathcal{X}$ (w.l.o.g. such that $I(X;Y) > 0$)
- *Codebook construction:* Construct a codebook with $\lceil 2^{nR} \rceil$ codewords, labeled as $x^n(m)$ with $m \in [1, 2^{nR}]$, by generating the symbols $x_i(m)$ for $i \in [1, n]$ and $m \in [1, 2^{nR}]$ independently according to P_X. The codebook is revealed both to the encoder and to the decoder
- *Encoder f:* Given m, transmit $x^n(m)$
- *Decoder g:* Given y^n, output $\hat{m}$ if it is the unique message such that $(x^n(\hat{m}), y^n) \in \mathcal{T}_\epsilon^{(n)}(X,Y)$; otherwise output an error $?$
- Let C_n be the random variable that represents the randomly generated codebook $\mathcal{C}_n$

Achievability Proof (2)

- *Goal:* Construct coding scheme that achieves the rate $R < \max_{P_X} I(X;Y)$
- To do so, develop an upper bound for $\mathbb{E}[\boldsymbol{P}_e(C_n)]$

- Notice that

$$\begin{aligned}\mathbb{E}[\boldsymbol{P}_e(C_n)] &= \mathbb{E}_{C_n}\left[\mathbb{P}\left[M \neq \hat{M} \middle| C_n\right]\right] \\ &= \sum_{m \in \mathcal{M}} \mathbb{E}_{C_n}\left[\mathbb{P}\left[M \neq \hat{M} \middle| M = m, C_n\right]\right] P_M(m)\end{aligned}$$

- By symmetry of the random code construction, this is independent of m. Therefore, assume w.l.o.g. that message $m = 1$ has been sent

$$\mathbb{E}[\boldsymbol{P}_e(C_n)] = \mathbb{E}_{C_n}\left[\mathbb{P}\left[M \neq \hat{M} \middle| M = 1, C_n\right]\right]$$

Achievability Proof (2)

- *Goal:* Construct coding scheme that achieves the rate $R < \max_{P_X} I(X;Y)$
- To do so, develop an upper bound for $\mathbb{E}[\boldsymbol{P}_e(C_n)]$

- Notice that

$$\begin{aligned}\mathbb{E}[\boldsymbol{P}_e(C_n)] &= \mathbb{E}_{C_n}\left[\mathbb{P}\left[M \neq \hat{M} \middle| C_n\right]\right] \\ &= \sum_{m\in\mathcal{M}} \mathbb{E}_{C_n}\left[\mathbb{P}\left[M \neq \hat{M} \middle| M = m, C_n\right]\right] P_M(m)\end{aligned}$$

- By symmetry of the random code construction, this is independent of m. Therefore, assume w.l.o.g. that message $m = 1$ has been sent

$$\mathbb{E}[\boldsymbol{P}_e(C_n)] = \mathbb{E}_{C_n}\left[\mathbb{P}\left[M \neq \hat{M} \middle| M = 1, C_n\right]\right]$$

Achievability Proof (3)

- Notice that $\mathbb{E}[\boldsymbol{P}_e(C_n)]$ can be expressed in terms of the events

$$\mathcal{E}_i = \left\{(X^n(i), Y^n) \in \mathcal{T}_\epsilon^{(n)}(X,Y)\right\} \quad \text{for } i \in [1, 2^{nR}]$$

as $\mathbb{E}[\boldsymbol{P}_e(C_n)] = \mathbb{P}[\mathcal{E}_1^c \cup \bigcup_{i \neq 1} \mathcal{E}_i]$.

- By the *union bound*

$$\mathbb{E}[\boldsymbol{P}_e(C_n)] \leq \mathbb{P}[\mathcal{E}_1^c] + \sum_{i \neq 1} \mathbb{P}[\mathcal{E}_i] \tag{1}$$

- By the *AEP*

$$\mathbb{P}[\mathcal{E}_1^c] \leq \delta_\epsilon(n) \tag{2}$$

- Since Y^n is the output when $X^n(1)$ is transmitted and $X^n(1)$ is independent of $X^n(i)$ for $i \neq 1$, output Y^n is independent of $X^n(i)$ for $i \neq 1$ so that

$$\mathbb{P}[\mathcal{E}_i] \leq 2^{-n(I(X;Y) - \delta(\epsilon))} \quad \text{for } i \neq 1 \tag{3}$$

Achievability Proof (3)

- Notice that $\mathbb{E}[\boldsymbol{P}_e(C_n)]$ can be expressed in terms of the events

$$\mathcal{E}_i = \left\{(X^n(i), Y^n) \in \mathcal{T}_\epsilon^{(n)}(X,Y)\right\} \quad \text{for } i \in [1, 2^{nR}]$$

  as $\mathbb{E}[\boldsymbol{P}_e(C_n)] = \mathbb{P}[\mathcal{E}_1^c \cup \bigcup_{i\neq 1} \mathcal{E}_i]$.

- By the *union bound*

$$\mathbb{E}[\boldsymbol{P}_e(C_n)] \leq \mathbb{P}[\mathcal{E}_1^c] + \sum_{i\neq 1} \mathbb{P}[\mathcal{E}_i] \tag{1}$$

- By the *AEP*

$$\mathbb{P}[\mathcal{E}_1^c] \leq \delta_\epsilon(n) \tag{2}$$

- Since Y^n is the output when $X^n(1)$ is transmitted and $X^n(1)$ is independent of $X^n(i)$ for $i \neq 1$, output Y^n is independent of $X^n(i)$ for $i \neq 1$ so that

$$\mathbb{P}[\mathcal{E}_i] \leq 2^{-n(I(X;Y)-\delta(\epsilon))} \quad \text{for } i \neq 1 \tag{3}$$

Achievability Proof (3)

- Notice that $\mathbb{E}[\boldsymbol{P}_e(C_n)]$ can be expressed in terms of the events

$$\mathcal{E}_i = \left\{(X^n(i), Y^n) \in \mathcal{T}_\epsilon^{(n)}(X,Y)\right\} \quad \text{for } i \in [1, 2^{nR}]$$

as $\mathbb{E}[\boldsymbol{P}_e(C_n)] = \mathbb{P}[\mathcal{E}_1^c \cup \bigcup_{i\neq 1} \mathcal{E}_i]$.

- By the *union bound*

$$\mathbb{E}[\boldsymbol{P}_e(C_n)] \leq \mathbb{P}[\mathcal{E}_1^c] + \sum_{i\neq 1} \mathbb{P}[\mathcal{E}_i] \tag{1}$$

- By the *AEP*

$$\mathbb{P}[\mathcal{E}_1^c] \leq \delta_\epsilon(n) \tag{2}$$

- Since Y^n is the output when $X^n(1)$ is transmitted and $X^n(1)$ is independent of $X^n(i)$ for $i \neq 1$, output Y^n is independent of $X^n(i)$ for $i \neq 1$ so that

$$\mathbb{P}[\mathcal{E}_i] \leq 2^{-n(I(X;Y)-\delta(\epsilon))} \quad \text{for } i \neq 1 \tag{3}$$

Achievability Proof (4)

- Substituting (2) and (3) into (1) we obtain

$$\begin{aligned}\mathbb{E}[\boldsymbol{P}_e(C_n)] &\le \mathbb{P}[\mathcal{E}_1^c] + \sum_{i\neq 1} \mathbb{P}[\mathcal{E}_i] \\ &\le \delta_\epsilon(n) + \sum_{i\neq 1} 2^{-n(I(X;Y)-\delta(\epsilon))} \\ &\le \delta_\epsilon(n) + \lceil 2^{nR} \rceil 2^{-n(I(X;Y)-\delta(\epsilon))}\end{aligned}$$

➠ Thus, if we choose the rate R such that $R < I(X;Y) - \delta(\epsilon)$, then

$$\mathbb{E}[\boldsymbol{P}_e(C_n)] \le \delta_\epsilon(n)$$

- By applying the *selection lemma* to the random variable C_n and the function $\boldsymbol{P}_e$, we conclude that there exists a $(2^{nR}, n)$ code $\mathcal{C}_n$ such that $\boldsymbol{P}_e(\mathcal{C}_n) \le \delta_\epsilon(n)$. Since ϵ can be chosen arbitrarily small and since P_X is arbitrary, we conclude that all rates $R < \max_{P_X} I(X;Y)$ are achievable □

Converse Proof

- *Goal:* Show that *any* achievable rate must satisfy $R \leq \max_{P_X} I(X;Y)$ (no assumptions on the particular coding scheme)
- Let R be an achievable rate and let $\epsilon > 0$. For n sufficiently large, there exists a $(2^{nR}, n)$ code $\mathcal{C}_n$ such that

$$\frac{1}{n} H(M|\mathcal{C}_n) \geq R \quad \text{and} \quad \boldsymbol{P}_e(\mathcal{C}_n) \leq \delta(\epsilon)$$

- In the remainder we drop the conditioning on $\mathcal{C}_n$ to simplify notation
- By *Fano's inequality*, it holds

$$\frac{1}{n} H(M|Y^n) \leq \delta(\boldsymbol{P}_e(\mathcal{C}_n)) = \delta(\epsilon)$$

Converse Proof (2)

- Therefore,

$$
\begin{aligned}
R &\le \frac{1}{n}H(M) = \frac{1}{n}I(M;Y^n) + \frac{1}{n}H(M|Y^n) \\
&\le \frac{1}{n}I(M;Y^n) + \delta(\epsilon) && \text{(Fano's inequality)} \\
&\le \frac{1}{n}I(X^n;Y^n) + \delta(\epsilon) && \text{(data processing inequality on } M - X^n - Y^n\text{)} \\
&= \frac{1}{n}H(Y^n) - \frac{1}{n}H(Y^n|X^n) + \delta(\epsilon) \\
&= \frac{1}{n}\sum_{i=1}^{n}\Big(H(Y_i|Y^{i-1}) - H(Y_i|X_i)\Big) + \delta(\epsilon) && \text{(channel is memoryless)} \\
&\le \frac{1}{n}\sum_{i=1}^{n}(H(Y_i) - H(Y_i|X_i)) + \delta(\epsilon) && \text{(conditioning reduces entropy)} \\
&= \frac{1}{n}\sum_{i=1}^{n}I(X_i;Y_i) + \delta(\epsilon) \\
&\le \max_{P_X} I(X;Y) + \delta(\epsilon)
\end{aligned}
$$

- Since ϵ can be chosen arbitrarily small, we obtain $R \le \max_{P_X} I(X;Y)$ □

Converse Proof (2)

- Therefore,

$$
\begin{aligned}
R &\le \frac{1}{n}H(M) = \frac{1}{n}I(M;Y^n) + \frac{1}{n}H(M|Y^n) \\
&\le \frac{1}{n}I(M;Y^n) + \delta(\epsilon) && \text{(Fano's inequality)} \\
&\le \frac{1}{n}I(X^n;Y^n) + \delta(\epsilon) && \text{(data processing inequality on } M - X^n - Y^n\text{)} \\
&= \frac{1}{n}H(Y^n) - \frac{1}{n}H(Y^n|X^n) + \delta(\epsilon) \\
&= \frac{1}{n}\sum_{i=1}^{n}\Big(H(Y_i|Y^{i-1}) - H(Y_i|X_i)\Big) + \delta(\epsilon) && \text{(channel is memoryless)} \\
&\le \frac{1}{n}\sum_{i=1}^{n}(H(Y_i) - H(Y_i|X_i)) + \delta(\epsilon) && \text{(conditioning reduces entropy)} \\
&= \frac{1}{n}\sum_{i=1}^{n}I(X_i;Y_i) + \delta(\epsilon) \\
&\le \max_{P_X} I(X;Y) + \delta(\epsilon)
\end{aligned}
$$

- Since ϵ can be chosen arbitrarily small, we obtain $R \le \max_{P_X} I(X;Y)$ □

Converse Proof (2)

- Therefore,

$$\begin{aligned}
R &\le \frac{1}{n}H(M) = \frac{1}{n}I(M;Y^n) + \frac{1}{n}H(M|Y^n) \\
&\le \frac{1}{n}I(M;Y^n) + \delta(\epsilon) && \text{(Fano's inequality)} \\
&\le \frac{1}{n}I(X^n;Y^n) + \delta(\epsilon) && \text{(data processing inequality on } M - X^n - Y^n\text{)} \\
&= \frac{1}{n}H(Y^n) - \frac{1}{n}H(Y^n|X^n) + \delta(\epsilon) \\
&= \frac{1}{n}\sum_{i=1}^{n}\Big(H(Y_i|Y^{i-1}) - H(Y_i|X_i)\Big) + \delta(\epsilon) && \text{(channel is memoryless)} \\
&\le \frac{1}{n}\sum_{i=1}^{n}(H(Y_i) - H(Y_i|X_i)) + \delta(\epsilon) && \text{(conditioning reduces entropy)} \\
&= \frac{1}{n}\sum_{i=1}^{n}I(X_i;Y_i) + \delta(\epsilon) \\
&\le \max_{P_X} I(X;Y) + \delta(\epsilon)
\end{aligned}$$

- Since ϵ can be chosen arbitrarily small, we obtain $R \le \max_{P_X} I(X;Y)$ $\square$

Converse Proof (2)

- Therefore,

$$
\begin{aligned}
R &\le \frac{1}{n}H(M) = \frac{1}{n}I(M;Y^n) + \frac{1}{n}H(M|Y^n) \\
&\le \frac{1}{n}I(M;Y^n) + \delta(\epsilon) && \text{(Fano's inequality)} \\
&\le \frac{1}{n}I(X^n;Y^n) + \delta(\epsilon) && \text{(data processing inequality on } M - X^n - Y^n\text{)} \\
&= \frac{1}{n}H(Y^n) - \frac{1}{n}H(Y^n|X^n) + \delta(\epsilon) \\
&= \frac{1}{n}\sum_{i=1}^{n}\Big(H(Y_i|Y^{i-1}) - H(Y_i|X_i)\Big) + \delta(\epsilon) && \text{(channel is memoryless)} \\
&\le \frac{1}{n}\sum_{i=1}^{n}(H(Y_i) - H(Y_i|X_i)) + \delta(\epsilon) && \text{(conditioning reduces entropy)} \\
&= \frac{1}{n}\sum_{i=1}^{n} I(X_i;Y_i) + \delta(\epsilon) \\
&\le \max_{P_X} I(X;Y) + \delta(\epsilon)
\end{aligned}
$$

- Since ϵ can be chosen arbitrarily small, we obtain $R \le \max_{P_X} I(X;Y)$ □

Converse Proof (2)

- Therefore,

$$
\begin{aligned}
R &\le \frac{1}{n}H(M) = \frac{1}{n}I(M;Y^n) + \frac{1}{n}H(M|Y^n) \\
&\le \frac{1}{n}I(M;Y^n) + \delta(\epsilon) && \text{(Fano's inequality)} \\
&\le \frac{1}{n}I(X^n;Y^n) + \delta(\epsilon) && \text{(data processing inequality on } M - X^n - Y^n\text{)} \\
&= \frac{1}{n}H(Y^n) - \frac{1}{n}H(Y^n|X^n) + \delta(\epsilon) \\
&= \frac{1}{n}\sum_{i=1}^{n}\Big(H(Y_i|Y^{i-1}) - H(Y_i|X_i)\Big) + \delta(\epsilon) && \text{(channel is memoryless)} \\
&\le \frac{1}{n}\sum_{i=1}^{n}\left(H(Y_i) - H(Y_i|X_i)\right) + \delta(\epsilon) && \text{(conditioning reduces entropy)} \\
&= \frac{1}{n}\sum_{i=1}^{n} I(X_i;Y_i) + \delta(\epsilon) \\
&\le \max_{P_X} I(X;Y) + \delta(\epsilon)
\end{aligned}
$$

- Since ϵ can be chosen arbitrarily small, we obtain $R \le \max_{P_X} I(X;Y)$ $\square$

Converse Proof (2)

- Therefore,

$$
\begin{aligned}
R &\leq \frac{1}{n} H(M) = \frac{1}{n} I(M; Y^n) + \frac{1}{n} H(M|Y^n) \\
&\leq \frac{1}{n} I(M; Y^n) + \delta(\epsilon) && \text{(Fano's inequality)} \\
&\leq \frac{1}{n} I(X^n; Y^n) + \delta(\epsilon) && \text{(data processing inequality on } M - X^n - Y^n\text{)} \\
&= \frac{1}{n} H(Y^n) - \frac{1}{n} H(Y^n|X^n) + \delta(\epsilon) \\
&= \frac{1}{n} \sum_{i=1}^{n} \Big(H(Y_i|Y^{i-1}) - H(Y_i|X_i) \Big) + \delta(\epsilon) && \text{(channel is memoryless)} \\
&\leq \frac{1}{n} \sum_{i=1}^{n} \big(H(Y_i) - H(Y_i|X_i) \big) + \delta(\epsilon) && \text{(conditioning reduces entropy)} \\
&= \frac{1}{n} \sum_{i=1}^{n} I(X_i; Y_i) + \delta(\epsilon) \\
&\leq \max_{P_X} I(X; Y) + \delta(\epsilon)
\end{aligned}
$$

- Since ϵ can be chosen arbitrarily small, we obtain $R \leq \max_{P_X} I(X; Y)$ $\square$

Converse Proof (2)

- Therefore,

$$
\begin{aligned}
R &\le \frac{1}{n}H(M) = \frac{1}{n}I(M;Y^n) + \frac{1}{n}H(M|Y^n) \\
&\le \frac{1}{n}I(M;Y^n) + \delta(\epsilon) && \text{(Fano's inequality)} \\
&\le \frac{1}{n}I(X^n;Y^n) + \delta(\epsilon) && \text{(data processing inequality on } M - X^n - Y^n\text{)} \\
&= \frac{1}{n}H(Y^n) - \frac{1}{n}H(Y^n|X^n) + \delta(\epsilon) \\
&= \frac{1}{n}\sum_{i=1}^{n}\Big(H(Y_i|Y^{i-1}) - H(Y_i|X_i)\Big) + \delta(\epsilon) && \text{(channel is memoryless)} \\
&\le \frac{1}{n}\sum_{i=1}^{n}\left(H(Y_i) - H(Y_i|X_i)\right) + \delta(\epsilon) && \text{(conditioning reduces entropy)} \\
&= \frac{1}{n}\sum_{i=1}^{n} I(X_i;Y_i) + \delta(\epsilon) \\
&\le \max_{P_X} I(X;Y) + \delta(\epsilon)
\end{aligned}
$$

- Since ϵ can be chosen arbitrarily small, we obtain $R \le \max_{P_X} I(X;Y)$ □

Converse Proof (2)

- Therefore,

$$
\begin{aligned}
R &\le \frac{1}{n}H(M) = \frac{1}{n}I(M;Y^n) + \frac{1}{n}H(M|Y^n) \\
&\le \frac{1}{n}I(M;Y^n) + \delta(\epsilon) && \text{(Fano's inequality)} \\
&\le \frac{1}{n}I(X^n;Y^n) + \delta(\epsilon) && \text{(data processing inequality on } M - X^n - Y^n\text{)} \\
&= \frac{1}{n}H(Y^n) - \frac{1}{n}H(Y^n|X^n) + \delta(\epsilon) \\
&= \frac{1}{n}\sum_{i=1}^{n}\Big(H(Y_i|Y^{i-1}) - H(Y_i|X_i)\Big) + \delta(\epsilon) && \text{(channel is memoryless)} \\
&\le \frac{1}{n}\sum_{i=1}^{n}\left(H(Y_i) - H(Y_i|X_i)\right) + \delta(\epsilon) && \text{(conditioning reduces entropy)} \\
&= \frac{1}{n}\sum_{i=1}^{n} I(X_i;Y_i) + \delta(\epsilon) \\
&\le \max_{P_X} I(X;Y) + \delta(\epsilon)
\end{aligned}
$$

- Since ϵ can be chosen arbitrarily small, we obtain $R \le \max_{P_X} I(X;Y)$ □

Converse Proof (2)

- Therefore,

$$
\begin{aligned}
R &\le \frac{1}{n}H(M) = \frac{1}{n}I(M;Y^n) + \frac{1}{n}H(M|Y^n) \\
&\le \frac{1}{n}I(M;Y^n) + \delta(\epsilon) && \text{(Fano's inequality)} \\
&\le \frac{1}{n}I(X^n;Y^n) + \delta(\epsilon) && \text{(data processing inequality on } M - X^n - Y^n\text{)} \\
&= \frac{1}{n}H(Y^n) - \frac{1}{n}H(Y^n|X^n) + \delta(\epsilon) \\
&= \frac{1}{n}\sum_{i=1}^{n}\Big(H(Y_i|Y^{i-1}) - H(Y_i|X_i)\Big) + \delta(\epsilon) && \text{(channel is memoryless)} \\
&\le \frac{1}{n}\sum_{i=1}^{n}\left(H(Y_i) - H(Y_i|X_i)\right) + \delta(\epsilon) && \text{(conditioning reduces entropy)} \\
&= \frac{1}{n}\sum_{i=1}^{n} I(X_i;Y_i) + \delta(\epsilon) \\
&\le \max_{P_X} I(X;Y) + \delta(\epsilon)
\end{aligned}
$$

- Since ϵ can be chosen arbitrarily small, we obtain $R \le \max_{P_X} I(X;Y)$ $\square$