Information Theory and Related Fileds

Lecture 3: Channel Coding

Lei Yu

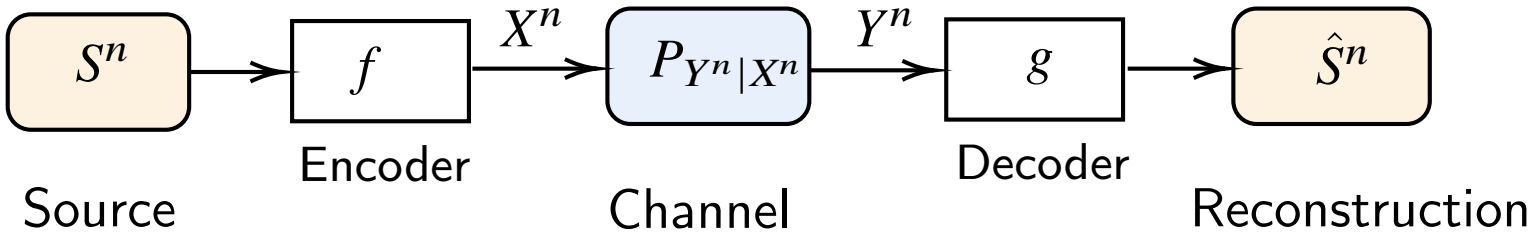
Nankai University

Online Short Course at Beijing Normal University

Outline

Outline

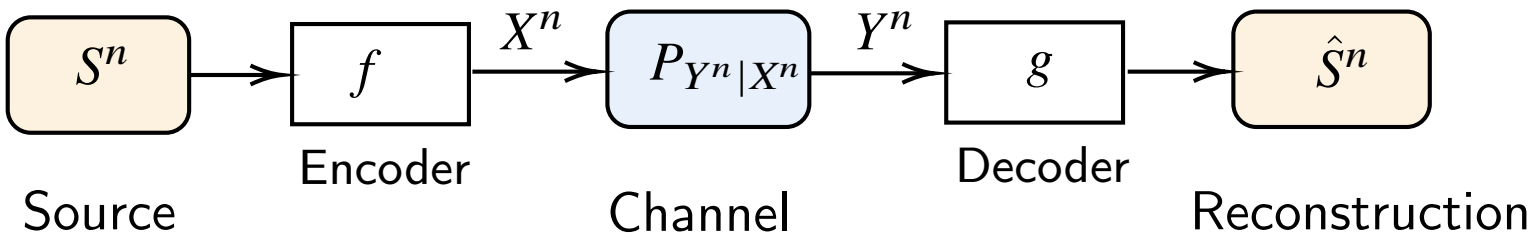
Recall: Source-Channel Coding Theorem



Theorem ([Shannon'48])

Consider discrete memoryless source S and discrete memoryless channel $P_{Y|X}$. There is a sequence of encoder-decoder pairs (f_n, g_n) such that $\mathbb{P}(S^n \neq \hat{S}^n) \to 0$ (as $n \to \infty$) if $H(S) < C(P_{Y|X})$, and only if $H(S) \leq C(P_{Y|X})$.

Recall: Source-Channel Coding Theorem

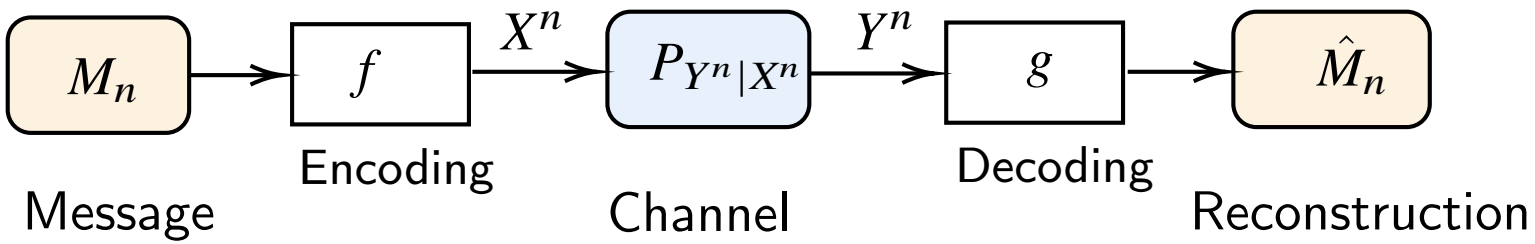


Theorem ([Shannon'48])

Consider discrete memoryless source S and discrete memoryless channel $P_{Y|X}$. There is a sequence of encoder-decoder pairs (f_n, g_n) such that $\mathbb{P}(S^n \neq \hat{S}^n) \to 0$ (as $n \to \infty$) if $H(S) < C(P_{Y|X})$, and only if $H(S) \leq C(P_{Y|X})$.

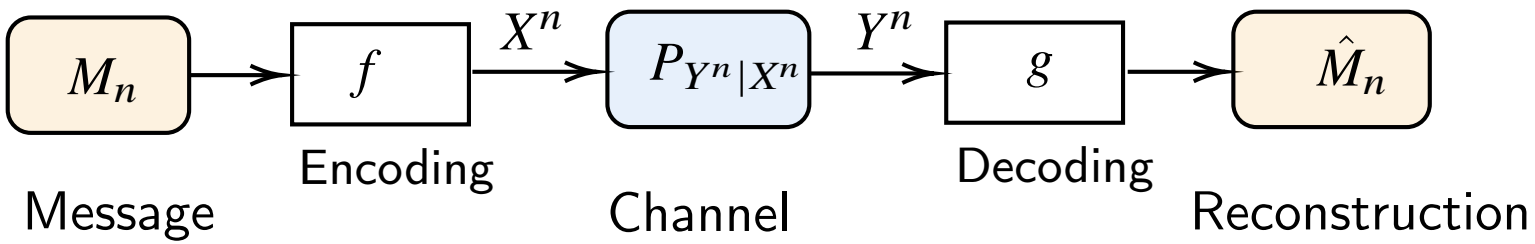
- Source coding (with noiseless rate-R channel) was proven in last lecture

Another Special Case: Channel Coding



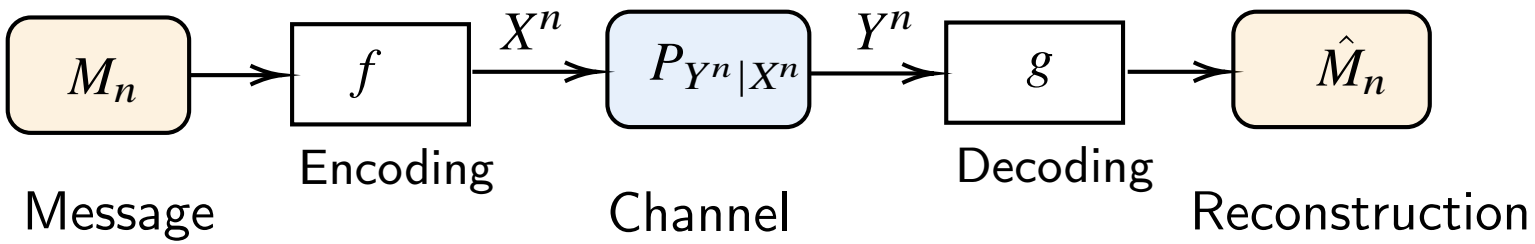
- A rate-R message (or a uniform source) is a r.v. $M \sim \mathrm{Unif}[2^{nR}]$.

Another Special Case: Channel Coding



- A rate-R message (or a uniform source) is a r.v. $M \sim \mathrm{Unif}[2^{nR}]$.
- For this case, $f : [2^{nR}] \to \mathcal{X}^n$ and $g : \mathcal{Y}^n \to [2^{nR}]$ are respectively also called channel encoder and channel decoder, and R is also called the rate of (f, g).

Another Special Case: Channel Coding



- A rate-R message (or a uniform source) is a r.v. $M \sim \mathrm{Unif}[2^{nR}]$.
- For this case, $f : [2^{nR}] \to \mathcal{X}^n$ and $g : \mathcal{Y}^n \to [2^{nR}]$ are respectively also called channel encoder and channel decoder, and R is also called the rate of (f, g).
- Essence of channel coding: Resist to noise interference in order to transmit the rate-R message almost with no error to the receiver, i.e., $\mathbb{P}(M \neq \hat{M}) \to 0$.
  - In other words, convert the noisy channel to an almost noiseless one

Channel Coding Theorem

As a special case of the source-channel coding theorem, Shannon showed:

Channel Coding Theorem

As a special case of the source-channel coding theorem, Shannon showed:

Theorem (Channel Coding [Shannon'48])

Consider a rate-R message and a discrete memoryless channel $P_{Y|X}$. There is a sequence of encoder-decoder pairs (f_n, g_n) such that $\mathbb{P}(M \neq \hat{M}) \to 0$ (as $n \to \infty$) if $R < C(P_{Y|X})$, and only if $R \leq C(P_{Y|X})$, where

$$C(P_{Y|X}) := \max_{P_X} I(X;Y)$$

is called channel capacity.

Channel Coding Theorem

As a special case of the source-channel coding theorem, Shannon showed:

Theorem (Channel Coding [Shannon'48])

Consider a rate-R message and a discrete memoryless channel $P_{Y|X}$. There is a sequence of encoder-decoder pairs (f_n, g_n) such that $\mathbb{P}(M \neq \hat{M}) \to 0$ (as $n \to \infty$) if $R < C(P_{Y|X})$, and only if $R \leq C(P_{Y|X})$, where

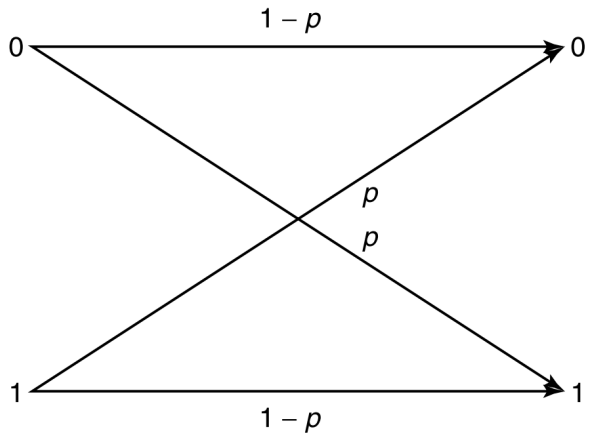
$$C(P_{Y|X}) := \max_{P_X} I(X;Y)$$

is called channel capacity.

We will prove this theorem in this lecture.

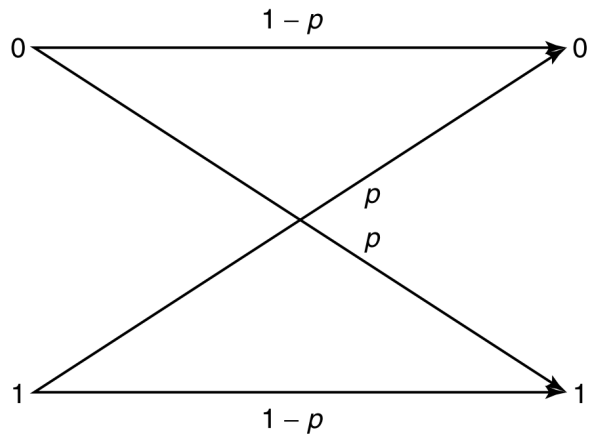
Example

Binary Symmetric Channel $\mathrm{BSC}(p)$: $Y = X \oplus W$ with $W \sim \mathrm{Bern}(p)$ independent of X where $\oplus$ is XOR.



Example

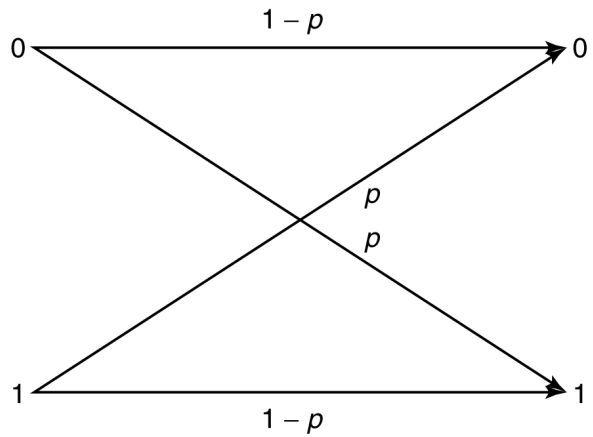
Binary Symmetric Channel $\mathrm{BSC}(p)$: $Y = X \oplus W$ with $W \sim \mathrm{Bern}(p)$ independent of X where $\oplus$ is XOR.



- Any input distribution must be a Bernoulli distribution.
- If $X \sim \mathrm{Bern}(q)$, then $Y \sim \mathrm{Bern}(p * q)$ with $p * q := p(1-q) + q(1-p)$ denoting the binary convolution.

Example

Binary Symmetric Channel $\mathrm{BSC}(p)$: $Y = X \oplus W$ with $W \sim \mathrm{Bern}(p)$ independent of X where $\oplus$ is XOR.



- Any input distribution must be a Bernoulli distribution.
- If $X \sim \mathrm{Bern}(q)$, then $Y \sim \mathrm{Bern}(p * q)$ with $p * q := p(1-q) + q(1-p)$ denoting the binary convolution.
- For this input X,

$$I(X;Y) = H(Y) - H(Y|X) = H_2(p * q) - H_2(p)$$

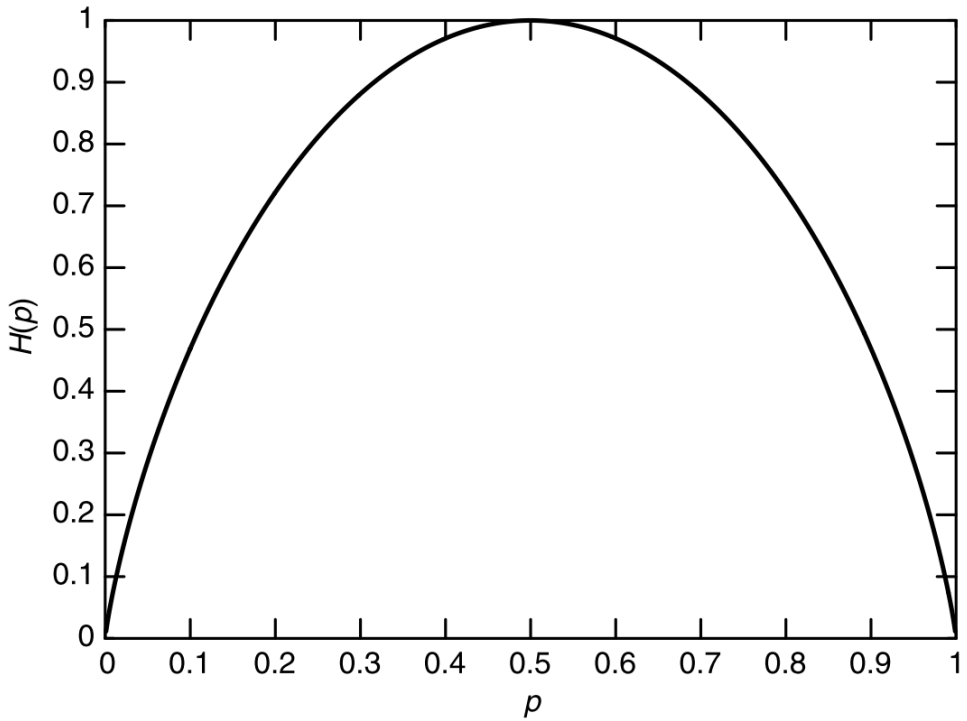
where $H_2(p) := p \log \frac{1}{p} + (1-p) \log \frac{1}{1-p}$ is the binary entropy function.

Example (cont.)

So, we have

$$C(\mathrm{BSC}(p)) = \max_{q \in [0,1]} H_2(p * q) - H_2(p)$$

$$= 1 - H_2(p), \qquad \text{optimal } q = \frac{1}{2}$$



Outline

Achievability Part ("If" Part)

Definition

The (weakly) joint typical set $\mathcal{A}_\epsilon^{(n)}(P_{XY})$ (or shortly, $\mathcal{A}_\epsilon^{(n)}$) with respect to P_{XY} is the set of $(x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n$ such that

$$\left| -\frac{1}{n} \log P_X^{\otimes n}(x^n) - H(X) \right| \le \epsilon$$

$$\left| -\frac{1}{n} \log P_Y^{\otimes n}(y^n) - H(Y) \right| \le \epsilon$$

$$\left| -\frac{1}{n} \log P_{XY}^{\otimes n}(x^n, y^n) - H(X,Y) \right| \le \epsilon.$$

Properties of Joint Typical Sets

Fact: [Cover–Thomas' book]

1. (Joint AEP) $P_{XY}^{\otimes n}(\mathcal{A}_\epsilon^{(n)}) \to 1$ as $n \to \infty$.
2. $|\mathcal{A}_\epsilon^{(n)}| \le 2^{n(H(X,Y)+\epsilon)}$.
3. $|\mathcal{A}_\epsilon^{(n)}| \ge (1-\epsilon)2^{n(H(X,Y)-\epsilon)}$ for sufficiently large n.
4. Let $(X'^n, Y'^n) \sim P_X^{\otimes n} \otimes P_Y^{\otimes n}$. For sufficiently large n,

$$(1-\epsilon)2^{-n(I(X;Y)+3\epsilon)} \le \mathbb{P}\left\{(X'^n, Y'^n) \in \mathcal{A}_\epsilon^{(n)}\right\} \le 2^{-n(I(X;Y)-3\epsilon)}.$$

Coding Scheme

- Let P_X attain $C = \max_{P_X} I(X;Y)$. Let R be any number $< I(X;Y) - 3\epsilon = C - 3\epsilon$.

Coding Scheme

- Let P_X attain $C = \max_{P_X} I(X;Y)$. Let R be any number $< I(X;Y) - 3\epsilon = C - 3\epsilon$.
- **Generation of codebook:** Randomly generate 2^{nR} sequences (codewords) X^n drawn i.i.d. $\sim P_X^{\otimes n}$. Index them by $i \in [2^{nR}]$. Denote (random) codebook $\mathcal{C}_n := \left\{X^n(1), X^n(2), ..., X^n(2^{nR})\right\}$. Reveal this codebook to the encoder and decoder.

Coding Scheme

- Let P_X attain $C = \max_{P_X} I(X;Y)$. Let R be any number $< I(X;Y) - 3\epsilon = C - 3\epsilon$.
- **Generation of codebook:** Randomly generate 2^{nR} sequences (codewords) X^n drawn i.i.d. $\sim P_X^{\otimes n}$. Index them by $i \in [2^{nR}]$. Denote (random) codebook $\mathcal{C}_n := \left\{X^n(1), X^n(2), ..., X^n(2^{nR})\right\}$. Reveal this codebook to the encoder and decoder.
- **Encoding:** Given $M_n = m$, encode M_n by $X^n(m)$.

Coding Scheme

- Let P_X attain $C = \max_{P_X} I(X;Y)$. Let R be any number $< I(X;Y) - 3\epsilon = C - 3\epsilon$.
- **Generation of codebook:** Randomly generate 2^{nR} sequences (codewords) X^n drawn i.i.d. $\sim P_X^{\otimes n}$. Index them by $i \in [2^{nR}]$. Denote (random) codebook $C_n := \left\{X^n(1), X^n(2), ..., X^n(2^{nR})\right\}$. Reveal this codebook to the encoder and decoder.
- **Encoding:** Given $M_n = m$, encode M_n by $X^n(m)$.
- **Decoding:** Declare the reconstruction of M_n as $\hat{m}$ if $(X^n(\hat{m}), Y^n) \in \mathcal{A}_\epsilon^{(n)}$ and $(X^n(\hat{m}'), Y^n) \notin \mathcal{A}_\epsilon^{(n)}, \forall \hat{m}' \neq \hat{m}$. If there is no such $\hat{m}$ or more than one such $\hat{m}$, let $\hat{m} = 1$.

Calculation of Probability of Error

Lemma ([Cover–Thomas' book])

If $R < I(X;Y) - 3\epsilon$, *then* $\mathbb{P}_{M_n, \mathcal{C}_n}(M_n \neq \hat{M}_n) \to 0$ *as* $n \to 0$.

Calculation of Probability of Error

Lemma ([Cover–Thomas' book])

If $R < I(X;Y) - 3\epsilon$, *then* $\mathbb{P}_{M_n, \mathcal{C}_n}(M_n \neq \hat{M}_n) \to 0$ *as* $n \to 0$.

Intuition behind this lemma:

- Given $M_n = m$, by the joint AEP, $\mathbb{P}\left\{(X^n(m), Y^n) \in \mathcal{A}_\epsilon^{(n)}\right\} \to 1$. That is, with high probability (w.h.p.), $X^n(m)$ is jointly typical with Y^n.

Calculation of Probability of Error

Lemma ([Cover–Thomas' book])

If $R < I(X;Y) - 3\epsilon$, *then* $\mathbb{P}_{M_n, \mathcal{C}_n}(M_n \neq \hat{M}_n) \to 0$ *as* $n \to 0$.

Intuition behind this lemma:

- Given $M_n = m$, by the joint AEP, $\mathbb{P}\left\{(X^n(m), Y^n) \in \mathcal{A}_\epsilon^{(n)}\right\} \to 1$. That is, with high probability (w.h.p.), $X^n(m)$ is jointly typical with Y^n.
- Y^n is independent of $X^n(m'), m' \neq m$, since Y^n only depends on $X^n(m)$ and $X^n(m)$ is independent of $X^n(m'), m' \neq m$.
  - So, $\mathbb{P}\left\{(X^n(m'), Y^n) \in \mathcal{A}_\epsilon^{(n)}\right\} \approx 2^{-nI(X;Y)}$ for all $m' \neq m$.
  - The (expected) total number of codewords $X^n(m'), m' \neq m$ jointly typical with Y^n is $2^{n(R-I(X;Y))}$ which is exponentially small when $R < I(X;Y)$. That is, w.h.p., $X^n(m'), m' \neq m$ are not jointly typical with Y^n.

Calculation of Probability of Error

Lemma ([Cover–Thomas' book])

If $R < I(X;Y) - 3\epsilon$, *then* $\mathbb{P}_{M_n, \mathcal{C}_n}(M_n \neq \hat{M}_n) \to 0$ *as* $n \to 0$.

Intuition behind this lemma:

- Given $M_n = m$, by the joint AEP, $\mathbb{P}\left\{(X^n(m), Y^n) \in \mathcal{A}_\epsilon^{(n)}\right\} \to 1$. That is, with high probability (w.h.p.), $X^n(m)$ is jointly typical with Y^n.
- Y^n is independent of $X^n(m'), m' \neq m$, since Y^n only depends on $X^n(m)$ and $X^n(m)$ is independent of $X^n(m'), m' \neq m$.
  - So, $\mathbb{P}\left\{(X^n(m'), Y^n) \in \mathcal{A}_\epsilon^{(n)}\right\} \approx 2^{-nI(X;Y)}$ for all $m' \neq m$.
  - The (expected) total number of codewords $X^n(m'), m' \neq m$ jointly typical with Y^n is $2^{n(R-I(X;Y))}$ which is exponentially small when $R < I(X;Y)$. That is, w.h.p., $X^n(m'), m' \neq m$ are not jointly typical with Y^n.
- So, w.h.p., $X^n(m)$ is the unique codeword jointly typical with Y^n.

Calculation of Probability of Error

Lemma ([Cover–Thomas' book])

If $R < I(X;Y) - 3\epsilon$, *then* $\mathbb{P}_{M_n, \mathcal{C}_n}(M_n \neq \hat{M}_n) \to 0$ *as* $n \to 0$.

Intuition behind this lemma:

- Given $M_n = m$, by the joint AEP, $\mathbb{P}\left\{(X^n(m), Y^n) \in \mathcal{A}_\epsilon^{(n)}\right\} \to 1$. That is, with high probability (w.h.p.), $X^n(m)$ is jointly typical with Y^n.
- Y^n is independent of $X^n(m'), m' \neq m$, since Y^n only depends on $X^n(m)$ and $X^n(m)$ is independent of $X^n(m'), m' \neq m$.
  - So, $\mathbb{P}\left\{(X^n(m'), Y^n) \in \mathcal{A}_\epsilon^{(n)}\right\} \approx 2^{-nI(X;Y)}$ for all $m' \neq m$.
  - The (expected) total number of codewords $X^n(m'), m' \neq m$ jointly typical with Y^n is $2^{n(R-I(X;Y))}$ which is exponentially small when $R < I(X;Y)$. That is, w.h.p., $X^n(m'), m' \neq m$ are not jointly typical with Y^n.
- So, w.h.p., $X^n(m)$ is the unique codeword jointly typical with Y^n.

Removing randomness of codebook: The lemma above holds on average over $\mathcal{C}_n$, there must exist a sequence of fixed c_n such that $\mathbb{P}_{M_n}(M_n \neq \hat{M}_n | \mathcal{C}_n = c_n) \to 0$.

Converse Part ("Only If" Part)

Recall that:

Lemma (Fano's inequality [Cover–Thomas' book])

Given two random variables X and Y, let $\hat{X} = g(Y)$ be any estimator of X given Y and let $\epsilon = \mathbb{P}(X \neq \hat{X})$ be the probability of error. Then

$$H(X|Y) \leq H(X|\hat{X}) \leq H_2(\epsilon) + \epsilon \log |\mathcal{X}|.$$

This inequality can be weakened to

$$H(X|Y) \leq H(X|\hat{X}) \leq 1 + \epsilon \log |\mathcal{X}|.$$

Converse Part (cont.)

Proof of Converse: For a pair of rate-R encoder-decoder (f_n, g_n), denote $X^n = f_n(M_n)$ and $\hat{M}_n = g_n(Y^n)$. Obviously, $M_n \leftrightarrow X^n \leftrightarrow Y^n \leftrightarrow \hat{M}_n$. Denote $\epsilon_n = \mathbb{P}(M_n \neq \hat{M}_n)$. We then have

$$
\begin{aligned}
\log 2^{nR} &= H(M_n) && \text{uniform r.v.} \\
&= H(M_n|\hat{M}_n) + I(M_n; \hat{M}_n) \\
&\leq 1 + \epsilon_n \log 2^{nR} + I(M_n; \hat{M}_n) && \text{Fano's inequality} \\
&\leq 1 + \epsilon_n nR + I(X^n; Y^n) && \text{DPI} \\
&= 1 + \epsilon_n nR + \sum_{i=1}^{n} I(Y_i; X^n|Y^{i-1}) && \text{chain rule} \\
&= 1 + \epsilon_n nR + \sum_{i=1}^{n} H(Y_i|Y^{i-1}) - H(Y_i|X^n, Y^{i-1}) \\
&\leq 1 + \epsilon_n nR + \sum_{i=1}^{n} H(Y_i) - H(Y_i|X_i) && \begin{cases} \text{conditioning reduces entropy} \\ \& \text{ memorylessness of channel} \end{cases}
\end{aligned}
$$

Converse Part (cont.)

$$
\begin{aligned}
nR &\le 1 + \epsilon_n nR + \sum_{i=1}^{n} H(Y_i) - H(Y_i|X_i) && \text{copy from last slide} \\
&= 1 + \epsilon_n nR + \sum_{i=1}^{n} I(X_i; Y_i) \\
&\le 1 + \epsilon_n nR + nC && \text{definition of capacity}
\end{aligned}
$$

Converse Part (cont.)

$$
\begin{aligned}
nR &\le 1 + \epsilon_n nR + \sum_{i=1}^{n} H(Y_i) - H(Y_i|X_i) && \text{copy from last slide} \\
&= 1 + \epsilon_n nR + \sum_{i=1}^{n} I(X_i; Y_i) \\
&\le 1 + \epsilon_n nR + nC && \text{definition of capacity}
\end{aligned}
$$

So,

$$R \le C + \frac{1}{n} + \epsilon_n R.$$

Since $\epsilon_n \to 0$ as $n \to \infty$, taking $\lim_{n\to\infty}$, we then have

$$R \le C.$$

Outline

Transmitting Power is Usually Limited

- Transmitting power is usually limited, especially when the sender is a mobile device.

Transmitting Power is Usually Limited

- Transmitting power is usually limited, especially when the sender is a mobile device.
- A cost function is a mapping

$$c : \mathcal{X} \to [0, \infty)$$

Transmitting Power is Usually Limited

- Transmitting power is usually limited, especially when the sender is a mobile device.
- A cost function is a mapping

$$c : \mathcal{X} \to [0, \infty)$$

- Example: Squared-cost (or power) function $c(x) = x^2$

Transmitting Power is Usually Limited

- Transmitting power is usually limited, especially when the sender is a mobile device.
- A cost function is a mapping

$$c : \mathcal{X} \to [0, \infty)$$

- Example: Squared-cost (or power) function $c(x) = x^2$
- The cost of a sequence x^n is defined by

$$c(x^n) = \frac{1}{n} \sum_{i=1}^{n} c(x_i).$$

Channel Coding with Input Constraint

- A rate-cost pair (R, P) is said to be achievable if there exists a sequence of rate-R encoder and decoder (f_n, g_n) such that $\mathbb{P}(M_n \neq \hat{M}_n) \to 0$ and

$$c(X^n) \leq P \text{ a.s.}$$

Channel Coding with Input Constraint

- A rate-cost pair (R, P) is said to be achievable if there exists a sequence of rate-R encoder and decoder (f_n, g_n) such that $\mathbb{P}(M_n \neq \hat{M}_n) \to 0$ and

$$c(X^n) \leq P \text{ a.s.}$$

- The (operational) capacity-cost function $C_{\text{op}}(P)$ is the supremum of rates R such that (R, P) is achievable.

Channel Coding Theorem

Theorem (Channel Coding with Input Constraint [Shannon'48])

Consider a memoryless channel $P_{Y|X}$ and a cost function c. Then,

$$C_{\mathrm{op}}(P) = C(P) := \max_{P_X : \mathbb{E}c(X) \le P} I(X;Y).$$

Channel Coding Theorem

Theorem (Channel Coding with Input Constraint [Shannon'48])

Consider a memoryless channel $P_{Y|X}$ and a cost function c. Then,

$$C_{\mathrm{op}}(P) = C(P) := \max_{P_X : \mathbb{E}c(X) \le P} I(X;Y).$$

Proof is similar to the one for channel coding without input constraint.

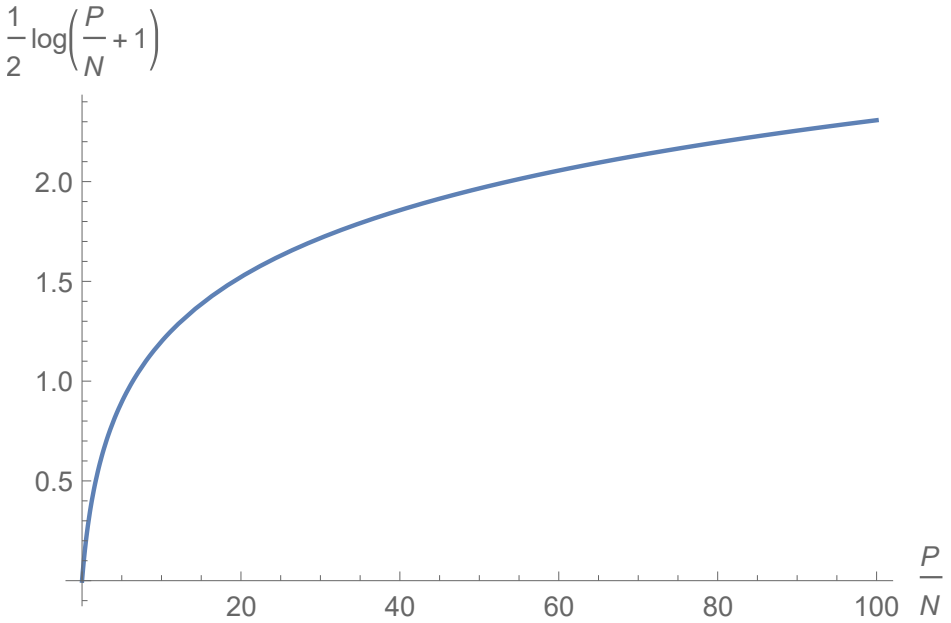
Example: Gaussian Channel with Power Constraint

Fact

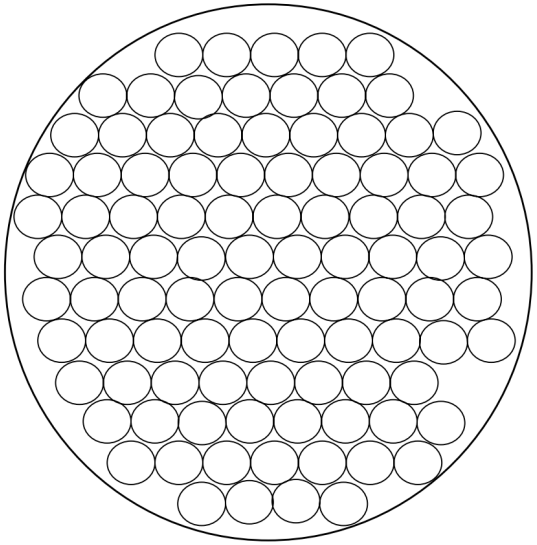
The capacity-cost function for a Gaussian channel $Y = X + Z$, $Z \sim \mathcal{N}(0, N)$ *with power constraint is given by*

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

where the optimal P_X *attaining* $C(P)$ *is* $P_X = \mathcal{N}(0, P)$.

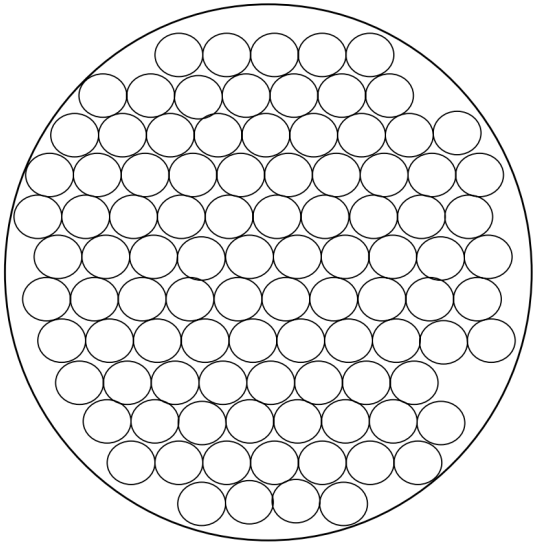


Intuition for Gaussian Channel



- By the AEP, the Gaussian noise Z^n is concentrated on a thin spherical shell (or a ball) of radius around $\sqrt{nN}$
- These balls are scattered throughout the space of received vectors. The received vectors have energy no greater than $n(P+N)$, so they lie in a ball of radius $\sqrt{n(P+N)}$.

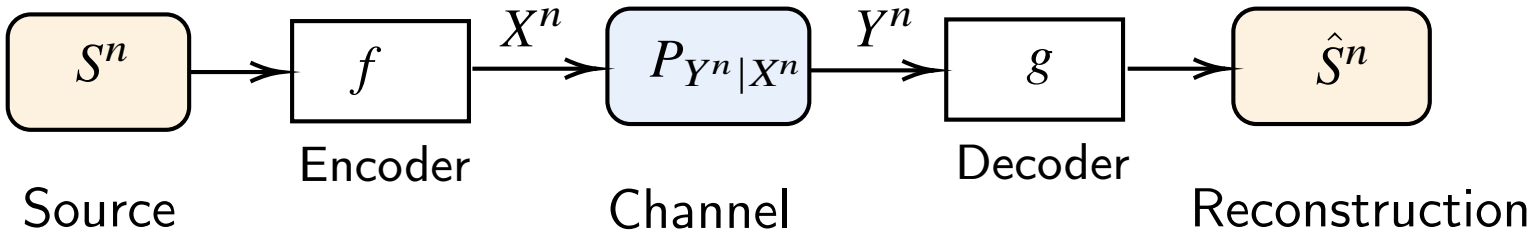
Intuition for Gaussian Channel



- By the AEP, the Gaussian noise Z^n is concentrated on a thin spherical shell (or a ball) of radius around $\sqrt{nN}$
- These balls are scattered throughout the space of received vectors. The received vectors have energy no greater than $n(P+N)$, so they lie in a ball of radius $\sqrt{n(P+N)}$.
- Pack a radius-$\sqrt{n(P+N)}$ ball by radius-$\sqrt{nN}$ balls: The number of small balls packed is at most $\frac{\mathrm{Vol}(\mathrm{Ball}_{\sqrt{n(P+N)}})}{\mathrm{Vol}(\mathrm{Ball}_{\sqrt{nN}})} = \frac{\left(\sqrt{n(P+N)}\right)^n}{\left(\sqrt{nN}\right)^n} = 2^{n\cdot\frac{1}{2}\log\left(1+\frac{P}{N}\right)}$

Outline

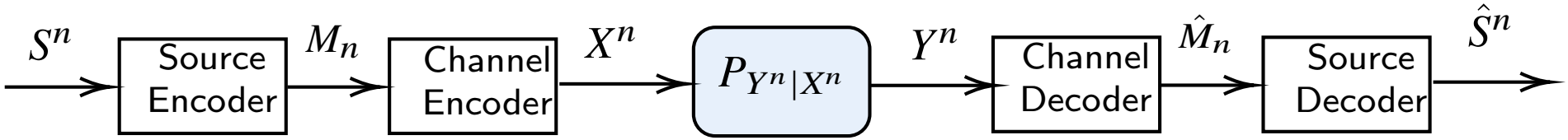
Turn Back to Source-Channel Coding



Theorem ([Shannon'48])

Consider discrete memoryless source S and discrete memoryless channel $P_{Y|X}$. There is a sequence of encoder-decoder pairs (f_n, g_n) such that $\mathbb{P}(S^n \neq \hat{S}^n) \to 0$ (as $n \to \infty$) if $H(S) < C(P_{Y|X})$, and only if $H(S) \leq C(P_{Y|X})$.

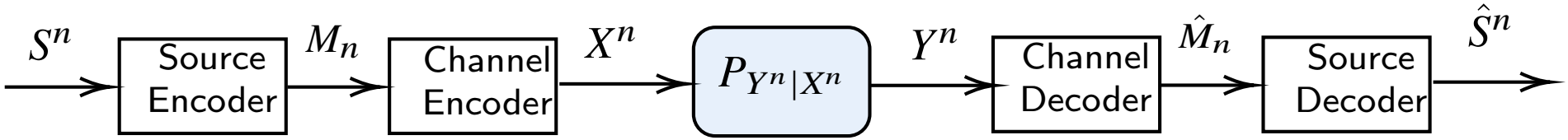
Proof



Achievability: Separate source and channel coding.

- The rate R of M_n (i.e., the source coding rate and also channel coding rate) is chosen such that $H(S) < R < C(P_{Y|X})$

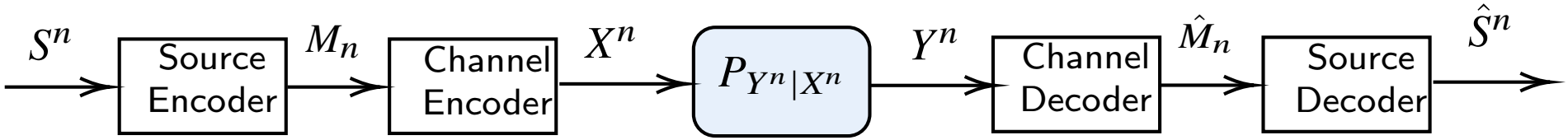
Proof



Achievability: Separate source and channel coding.

- The rate R of M_n (i.e., the source coding rate and also channel coding rate) is chosen such that $H(S) < R < C(P_{Y|X})$
- By channel coding theorem, $\mathbb{P}(\text{channel coding error}) \to 0$ (since $R < C(P_{Y|X})$)
- By source coding theorem, $\mathbb{P}(\text{source coding error}) \to 0$ (since $H(S) < R$)

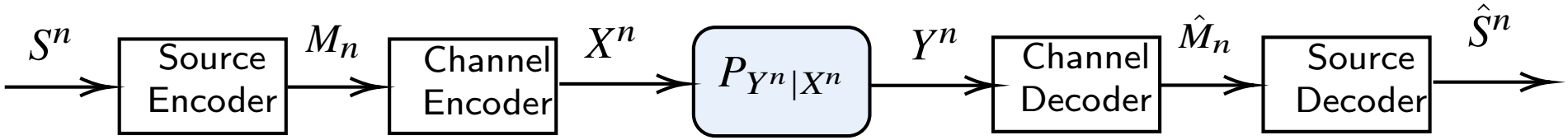
Proof



Achievability: Separate source and channel coding.

- The rate R of M_n (i.e., the source coding rate and also channel coding rate) is chosen such that $H(S) < R < C(P_{Y|X})$
- By channel coding theorem, $\mathbb{P}(\text{channel coding error}) \to 0$ (since $R < C(P_{Y|X})$)
- By source coding theorem, $\mathbb{P}(\text{source coding error}) \to 0$ (since $H(S) < R$)
- When no source and channel coding errors, we have $S^n = \hat{S}^n$. So, $\mathbb{P}(S^n \neq \hat{S}^n) \to 0$.

Proof



Achievability: Separate source and channel coding.

- The rate R of M_n (i.e., the source coding rate and also channel coding rate) is chosen such that $H(S) < R < C(P_{Y|X})$
- By channel coding theorem, $\mathbb{P}(\text{channel coding error}) \to 0$ (since $R < C(P_{Y|X})$)
- By source coding theorem, $\mathbb{P}(\text{source coding error}) \to 0$ (since $H(S) < R$)
- When no source and channel coding errors, we have $S^n = \hat{S}^n$. So, $\mathbb{P}(S^n \neq \hat{S}^n) \to 0$.

Converse: Similar to the converse proof for channel coding

Lossy Source-Channel Coding

Theorem ([Shannon'48])

Consider a memoryless source S with bounded distortion function d, and a memoryless channel $P_{Y|X}$ with cost function c. There is a sequence of encoder-decoder pairs (f_n, g_n) such that

$$\limsup_{n\to\infty} \mathbb{E}d(S^n, \hat{S}^n) \le D,$$

$$c(X^n) \le P \text{ a.s.},$$

if $R(D) < C(P)$, and only if $R(D) \le C(P)$. Recall that $R(D)$ is the rate-distortion function and $C(P)$ is the capacity-cost function.

Proof is similar to the lossless source-channel coding theorem.

Outline

Key Ideas in Information Theory

- **Theoretical limits**:
  - Simplify a practical problem and ask for the theoretically optimal solution, even if this solution is unrealistic.
  - Information theory teaches us how to find a desirable solution, rather than gives us the desirable solution directly.
  - Information theory gains our insights into the practical problem, especially on how far is the optimal performance from us.

Key Ideas in Information Theory

- **Theoretical limits**:
  - Simplify a practical problem and ask for the theoretically optimal solution, even if this solution is unrealistic.
  - Information theory teaches us how to find a desirable solution, rather than gives us the desirable solution directly.
  - Information theory gains our insights into the practical problem, especially on how far is the optimal performance from us.
- **AEP** (or **typical sets**): i.i.d. r.v.'s are approximately uniformly distributed on a small set (i.e., typical set).

Key Ideas in Information Theory

- **Theoretical limits**:
  - Simplify a practical problem and ask for the theoretically optimal solution, even if this solution is unrealistic.
  - Information theory teaches us how to find a desirable solution, rather than gives us the desirable solution directly.
  - Information theory gains our insights into the practical problem, especially on how far is the optimal performance from us.
- **AEP** (or **typical sets**): i.i.d. r.v.'s are approximately uniformly distributed on a small set (i.e., typical set).
- **DPI**: Various DPI's have many applications in deriving bounds for many other problems, e.g., algorithms, statistical estimation, learning, etc.

Broadcast Channel

Network information theory: Communication systems with multi-senders and/or multi-reviewers

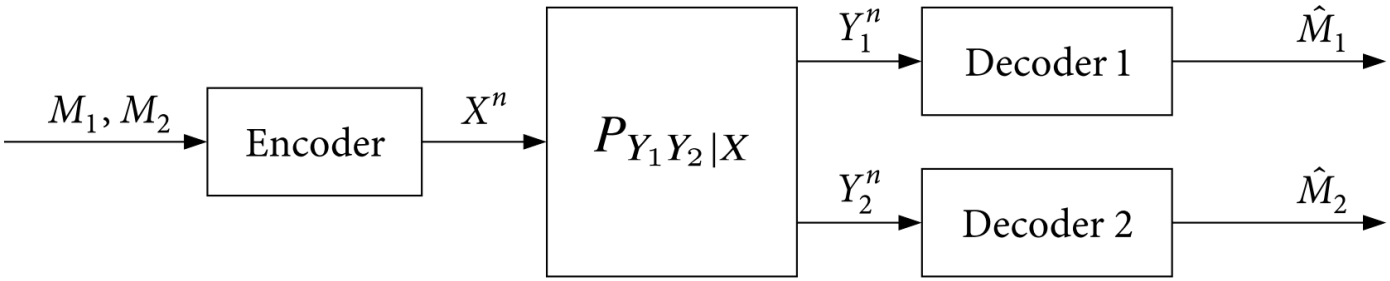
- There are many, many works in the literature on the network information theory; see [El Gamal–Kim's book]
- Unfortunately, there are also many, many problems that are still open

Broadcast Channel

Network information theory: Communication systems with multi-senders and/or multi-reviewers

- There are many, many works in the literature on the network information theory; see [El Gamal–Kim's book]
- Unfortunately, there are also many, many problems that are still open

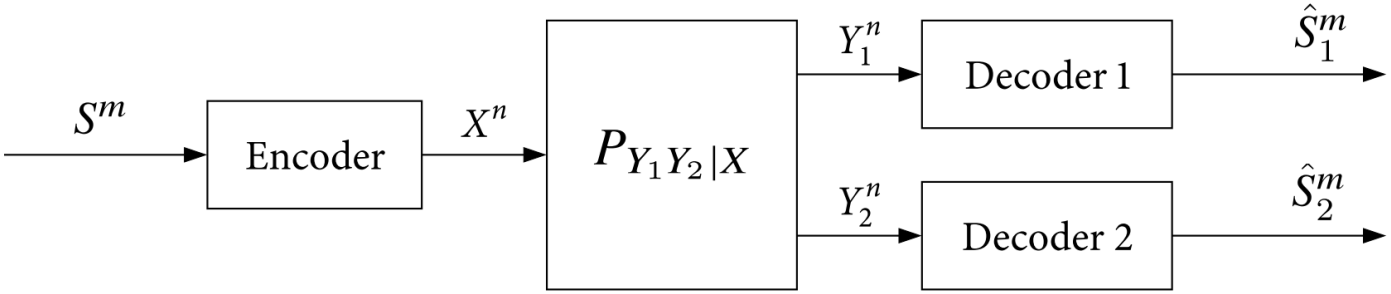
Open Problem 1: the capacity region of broadcast channel is still unknown.



Gaussian Source Broadcast

Open Problem 2: the distortion region for transmitting a Gaussian source over a Gaussian broadcast channel is still unknown when $m \neq n$.

- Interestingly, when $m = n$, symbol-by-symbol mapping is optimal, i.e., $X = \alpha S,\ \hat{S}_i = \beta_i Y_i, i = 1, 2$



Second-order Coding Rate

Theorem ([Strassen'62])

Suppose that a discrete memoryless source S^n can be transmitted over a noiseless channel $M \mapsto M \in [2^b]$ within error probability ϵ (i.e., $\mathbb{P}(S^n \neq \hat{S}^n) \leq \epsilon$). Then, given $\epsilon \in (0,1)$, the minimum b (number of bits) is

$$b^* = \begin{cases} nH(X) - \sqrt{nV(X)}\Phi^{-1}(\epsilon) - \frac{1}{2}\log n + O(1), & V(X) > 0 \\ nH(X) + O(1), & V(X) = 0 \end{cases}.$$

Here, $V(X)$ is the dispersion of the source (i.e., the variance of the self-information), which is given by

$$V(X) := \mathrm{Var}\left(\log \frac{1}{P_X(X)}\right) = \sum_x P_X(x)\left(\log \frac{1}{P_X(x)} - H(X)\right)^2$$

Second-order Coding Rate (cont.)

Theorem ([Polyanskiy'10, Tomamichel–Tan'13])

Suppose that a message $M \sim \mathrm{Unif}[2^b]$ can be transmitted over a discrete memoryless channel $P_{Y|X}$ within error probability ϵ (i.e., $\mathbb{P}(M \neq \hat{M}) \leq \epsilon$). If given $\epsilon \in (0,1)$, $P_{Y|X}$ is non-singular (i.e., for some (x, x', y), $P_{Y|X}(y|x)$ and $P_{Y|X}(y|x')$ are positive but not equal) and satisfies $V_\epsilon > 0$, then the maximum b (number of bits) is

$$b^* = nC + \sqrt{nV_\epsilon}\Phi^{-1}(\epsilon) + \frac{1}{2}\log n + O(1).$$

Here, V_ϵ is the ϵ-dispersion of the channel, which is given by

$$V_\epsilon := \begin{cases} \min_{P_X: I(X;Y)=C} V(X;Y), & \epsilon < \frac{1}{2} \\ \max_{P_X: I(X;Y)=C} V(X;Y), & \epsilon \geq \frac{1}{2} \end{cases}$$

where

$$V(X;Y) := \sum_{x,y} P_{XY}(x,y)\left(\log \frac{P_{XY}(x,y)}{P_X(x)P_Y(y)} - I(X;Y)\right)^2.$$

Sphere Packing and Covering (Zero-error Coding)

- Gaussian source coding is an approximate sphere covering problem, and Gaussian channel coding is an approximate sphere packing problem.
- Here "approximation" means the volume of uncovered region (for source coding) or overlapped region (for channel coding) is negligible, compared with the volume of the big sphere.

Sphere Packing and Covering (Zero-error Coding)

- Gaussian source coding is an approximate sphere covering problem, and Gaussian channel coding is an approximate sphere packing problem.
- Here "approximation" means the volume of uncovered region (for source coding) or overlapped region (for channel coding) is negligible, compared with the volume of the big sphere.
- When no uncovered region is required, the rate of minimum number of small spheres required is unchanged (i.e., $\log \frac{b}{a}$ with b, a resp. radii of big and small spheres).

Sphere Packing and Covering (Zero-error Coding)

- Gaussian source coding is an approximate sphere covering problem, and Gaussian channel coding is an approximate sphere packing problem.
- Here "approximation" means the volume of uncovered region (for source coding) or overlapped region (for channel coding) is negligible, compared with the volume of the big sphere.
- When no uncovered region is required, the rate of minimum number of small spheres required is unchanged (i.e., $\log \frac{b}{a}$ with b, a resp. radii of big and small spheres).
- When no overlapped region is required, the rate of maximum number of small spheres is no longer $\log \frac{b}{a}$.

Sphere Packing (cont.)

What is the densest sphere packing in the n-dimensional Euclidean space?

Sphere Packing (cont.)

What is the densest sphere packing in the n-dimensional Euclidean space?

- For $n = 3$, hexagonal close packing was conjectured to be the densest by Kepler in 1611, which was solved by Hales in 1998 using a computer-aided proof.

Sphere Packing (cont.)

What is the densest sphere packing in the n-dimensional Euclidean space?

- For $n = 3$, hexagonal close packing was conjectured to be the densest by Kepler in 1611, which was solved by Hales in 1998 using a computer-aided proof.
- For $n = 8$, E8 lattice packing was proven to be the densest by Viazovska (Fields Medal 2022) in 2016 (by a linear programming method)

Sphere Packing (cont.)

What is the densest sphere packing in the n-dimensional Euclidean space?

- For $n = 3$, hexagonal close packing was conjectured to be the densest by Kepler in 1611, which was solved by Hales in 1998 using a computer-aided proof.
- For $n = 8$, E8 lattice packing was proven to be the densest by Viazovska (Fields Medal 2022) in 2016 (by a linear programming method)
- For $n = 24$, Leech lattice packing was proven to be the densest by Viazovska and collaborators in 2017

Sphere Packing (cont.)

What is the densest sphere packing in the n-dimensional Euclidean space?

- For $n = 3$, hexagonal close packing was conjectured to be the densest by Kepler in 1611, which was solved by Hales in 1998 using a computer-aided proof.
- For $n = 8$, E8 lattice packing was proven to be the densest by Viazovska (Fields Medal 2022) in 2016 (by a linear programming method)
- For $n = 24$, Leech lattice packing was proven to be the densest by Viazovska and collaborators in 2017
- The asymptotic rate is still open (the best known bound $0.5990 \leq \log \frac{b}{a} - R^* \leq 1$ with R^* denoting the asymptotic rate)

Sphere Packing (cont.)

What is the densest sphere packing in the n-dimensional Euclidean space?

- For $n = 3$, hexagonal close packing was conjectured to be the densest by Kepler in 1611, which was solved by Hales in 1998 using a computer-aided proof.
- For $n = 8$, E8 lattice packing was proven to be the densest by Viazovska (Fields Medal 2022) in 2016 (by a linear programming method)
- For $n = 24$, Leech lattice packing was proven to be the densest by Viazovska and collaborators in 2017
- The asymptotic rate is still open (the best known bound $0.5990 \leq \log \frac{b}{a} - R^* \leq 1$ with R^* denoting the asymptotic rate)

What is the asymptotic rate of the densest sphere packing in the n-dimensional Hamming space? — known as the coding problem, but also still open

References

1. C. E. Shannon. A mathematical theory of communication. The Bell Systems Technical Journal, 27:379–423, 1948.
2. (Suit for Beginners) T. M. Cover and J. A. Thomas. Elements of Information Theory. Wiley-Interscience, 2nd edition, 2006.
3. (Suit for Beginners, Method of Types) A. El Gamal and Y.-H. Kim. Network Information Theory. Cambridge University Press, Cambridge, U.K., 2012.
4. (Method of Types) I. Csiszár and J. Körner. Information Theory: Coding Theorems for Discrete Memoryless Systems. Cambridge University Press, 2011.
5. (Information-Spectrum Method) T. S. Han. Information-Spectrum Methods in Information Theory. Springer Berlin Heidelberg, Feb 2003.
6. (Second-order Coding) Vincent Y. F., Tan. Asymptotic estimates in information theory with non-vanishing error probabilities." Foundations and Trends in Communications and Information Theory, 2014.

Thank you!