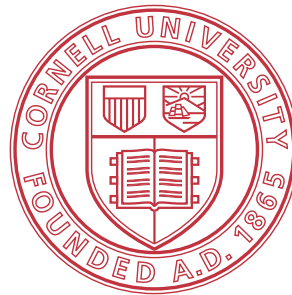


Some Network Information Theory Problems (as Understood by a CS Student)

Sergio D. Servetto

School of Electrical and Computer Engineering
Cornell University



(Presented at the Department of Computer Science, University of Illinois, Urbana-Champaign; February 9th, 2007.)

Acknowledgements

- João Barros (Dept. of Computer Science, U. Porto, Portugal).
- Ying Li, Mung Chiang, Sergio Verdú (Princeton/EE).
- Sources of support:
 - *Fundamental Performance Limits of Large-Scale Sensor Networks.*
NSF CAREER award CCR-0238271.
 - *The Reachback Channel in Wireless Sensor Networks.*
NSF SENSORS grant CCR-0330059. PI, joint with T. Berger, L. Tong, S. Wicker.
 - *Self-Configuring Sensor Networks for Disaster Prevention, Mitigation and Recovery.* NSF ITR grant ANR-0325556. Co-PI, joint with Cornell ECE, CEE and Economics faculty, and staff at NYS Wadsworth Labs.

Outline

- Reliable Communication over a Single Discrete Memoryless Noisy Channel
- Reliable Communication over a Network of Discrete Memoryless Noisy Channels
- Distributed Compression of Dependent Sources
- Summary and Conclusions

Outline

- Reliable Communication over a Single Discrete Memoryless Noisy Channel
 - Basic Information Measures, Typical Sequences and Zero-Error Data Compression
 - Channel Capacity and the Source/Channel Separation Theorem
 - Data Compression Subject to a Fidelity Criterion
 - On Why We Consider This To Be A Solved Problem
- Reliable Communication over a Network of Discrete Memoryless Noisy Channels
- Distributed Compression of Dependent Sources
- Summary and Conclusions

The Basic Questions of Information Theory

Information theory deals with two fundamental questions in communications:

- Q: What is the ultimate limit to which information can be compressed?

A: *The source entropy, \mathcal{H} .*

- Q: What is the ultimate rate at which information can be transmitted?

A: *The channel capacity, C .*

In this first part, we will try to understand what these results mean, and the rudiments of how they are proved in a purely discrete setup.

C. E. Shannon. *A Mathematical Theory of Communication*. Bell Syst. Tech. J., 27:379-423&623-656, 1948.

Available from <http://www.cparity.com/it/demo/external/shannon.pdf>.

Basic Information Measures: Entropy

Given:

- $\mathcal{X} = \{x_1, \dots, x_N\}$ – a finite alphabet.
- X – random variable taking values in \mathcal{X} .
- $p(x) \triangleq P(X = x)$ ($x \in \mathcal{X}$) – a probability mass function.

the *entropy* \mathcal{H}_X of a random variable X is defined as

$$\mathcal{H}_X(X) \triangleq \sum_{x \in \mathcal{X}} p(x) \log(1/p(x)) = E_{p(x)} \log(1/p(X)).$$

\mathcal{H}_X is a measure of uncertainty about the outcomes of X .

Basic Information Measures: Conditional Entropy

Given:

- $\mathcal{X} = \{x_1, \dots, x_N\}$ and $\mathcal{Y} = \{y_1, \dots, y_M\}$ – two finite alphabets.
- X and Y – random variables taking values in \mathcal{X} and \mathcal{Y} .
- $p(xy) \triangleq P(X = x \wedge Y = y) ((xy) \in \mathcal{X} \times \mathcal{Y})$ – a *joint* pmf.

the *conditional entropy* $\mathcal{H}_{X|Y}$ is defined as

$$\mathcal{H}_{X|Y}(X|Y) \triangleq \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(xy) \log(1/p(x|y)) = E_{p(xy)} \log(1/p(X|Y)).$$

$\mathcal{H}_{X|Y}$ is a measure of the uncertainty about the outcomes of X in the presence of some related information Y .

Basic Information Measures: Mutual Information

Given:

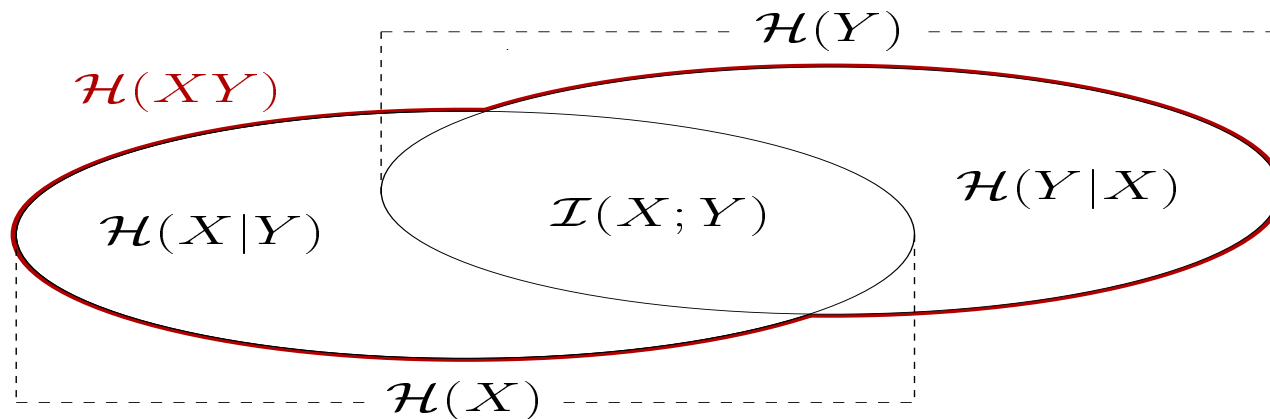
- $\mathcal{X} = \{x_1, \dots, x_N\}$ and $\mathcal{Y} = \{y_1, \dots, y_M\}$ – two finite alphabets.
- X and Y – random variables taking values in \mathcal{X} and \mathcal{Y} .
- $p(xy) \triangleq P(X = x \wedge Y = y) ((xy) \in \mathcal{X} \times \mathcal{Y})$ – a *joint* pmf.

The *mutual information* \mathcal{I} is defined as

$$\mathcal{I}(X \wedge Y) \triangleq \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(xy) \log \left(\frac{p(xy)}{p(x)p(y)} \right) = E_{p(xy)} \log (p(XY)/p(X)p(Y)).$$

\mathcal{I} is a measure of the amount of common information information in X and Y , or equivalently, of “how far” X and Y are from being independent.

“Sand on Multiple Balances, Uncertainty = Weight”



- $\mathcal{H}(XY) = \mathcal{H}(X) + \mathcal{H}(Y|X)$ – uncertainty in XY = uncertainty in X plus leftover
- $\mathcal{I}(X \wedge Y) = \mathcal{H}(X) - \mathcal{H}(X|Y)$ – mutual information = reduction in uncertainty
- $\mathcal{I}(X \wedge X) = \mathcal{H}(X) - \mathcal{H}(X|X) = \mathcal{H}(X)$ – relationship with entropy
- $\mathcal{I}(X \wedge Y) = \mathcal{H}(X) - \mathcal{H}(X|Y) = \mathcal{H}(X) - \mathcal{H}(X) = 0$ – for X and Y independent

Typical Sequences – Informally

- Suppose your favorite football team goes undefeated in a season of n games, and that results are independent from game to game.
- X_i : outcome (won/lost) of the i -th game in the season – undefeated means $P(X_i = \text{won}) = 1$.
- When sampling iid, there is *only one* sequence (out of 2^n possible ones) that we will ever see: $[\text{won}, \text{won}, \dots, \text{won}]$.

*We would like to discriminate among sequences we are likely to encounter when sampling iid, and sequences that will most likely never show up – those we encounter we will call **typical** sequences.*

Typical Sequences – A Bit More Formally

Given:

- $\mathcal{X} = \{x_1, \dots, x_N\}$ – a finite alphabet.
- X – random variable taking values in \mathcal{X} .
- $p(x) \triangleq P(X = x)$ ($x \in \mathcal{X}$) – a probability mass function.
- A block length n .

The set of *typical sequences* $A^{(n)}(X)$ is defined by

$$A^{(n)}(X) = \left\{ \mathbf{x} \in \mathcal{X}^n : -\frac{1}{n} \sum_{i=1}^n \log(p(x_i)) \approx \mathcal{H}(X) \right\}.$$

$A^{(n)}(X)$: *set of sequences with empirical entropy “close” to the true entropy.*

Properties of Typical Sets

Two fundamental properties of typical sets:

- High probability: for block lengths n that are large enough,

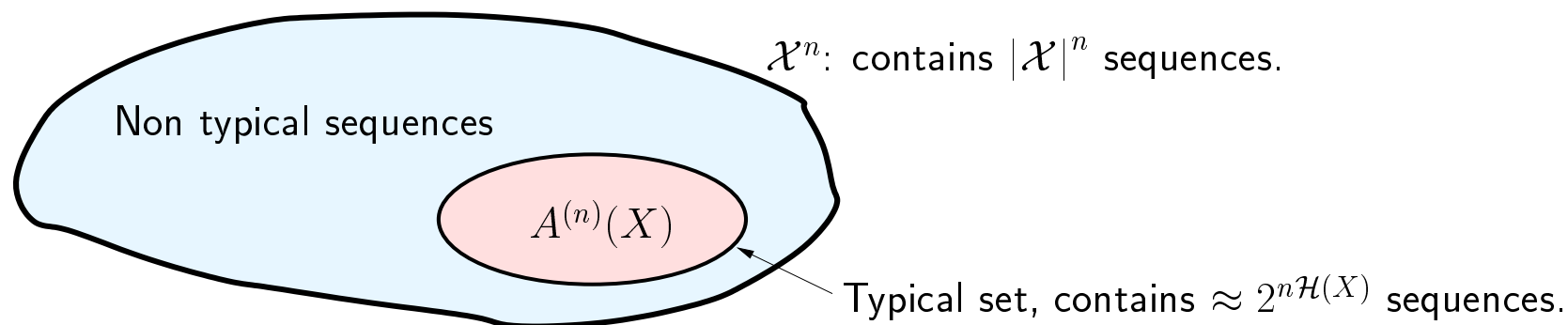
$$P\left(A^{(n)}(X)\right) \approx 1.$$

- Uniform distribution: again for block lengths n that are large enough,

$$\left|A^{(n)}(X)\right| \approx 2^{n\mathcal{H}(X)} \quad P\left(\mathbf{x} \mid \mathbf{x} \in A^{(n)}(X)\right) \approx 2^{-n\mathcal{H}(X)}$$

We say these are fundamental properties because we will be able to use them to prove Shannon's first theorem – the fact that sources can be compressed down to their entropy.

Data Compression I: Variable-Length Codes



To compress the source of information X , form blocks of length n , and then:

- If a sequence $\mathbf{x} \in A^{(n)}(X)$, use $\frac{1}{n} \log_2 |A^{(n)}(X)| \approx \mathcal{H}(X)$ bits/symbol.
- If not, use $\log_2 |\mathcal{X}|$ bits/symbol.

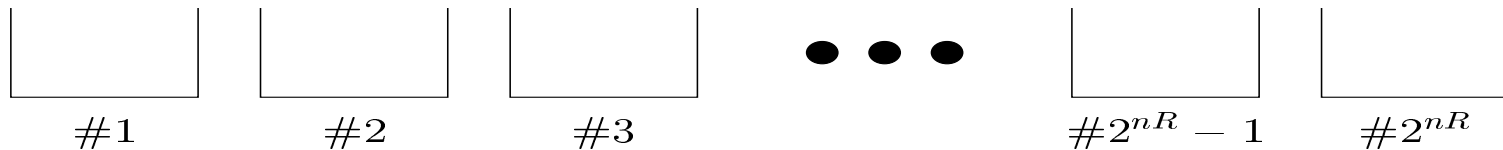
Therefore, the average code length is $\approx \mathcal{H}(X)$.

Data Compression II: Random Binning

A proof of the source coding theorem based on random binning:

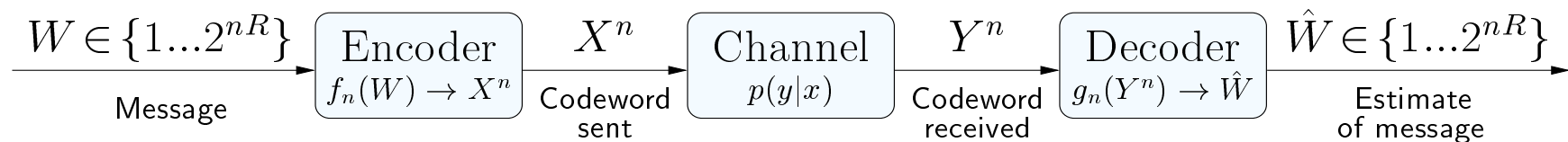
- Randomly place all sequences in \mathcal{X}^n into 2^{nR} bins.
- Reveal bin assignments to both encoder and decoder.
- To encode: given a sequence X^n , send bin index that contains it.
- To decode: look inside bin from encoder for a *typical* sequence.

Key: if $R > \mathcal{H}(X)$, there are exponentially many more bins than typical sequences... so almost all bins contain zero or one such sequences.



Reliable Communication over Unreliable Channels

A model for a system that transfers information:



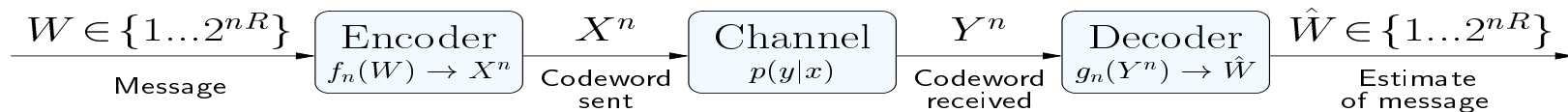
- Fix a block length n , and choose a message W (out of 2^{nR} choices).
- Based on W , the encoder chooses a codeword X^n to transmit.
- The channel produces a *noisy* version of X^n , that we call Y^n .
- Based on Y^n , the decoder forms an estimate \hat{W} of the original W .

Q: How many different W 's can we transmit such that $P(W \neq \hat{W}) \approx 0$?

Single-Letter Characterization of Channel Capacity

- Answer 1 (difficult):

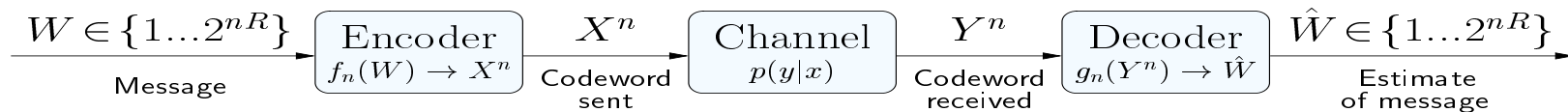
Largest value of R (denoted C_{op}) for which there exist an encoder f_n and a decoder g_n such that $P(W \neq \hat{W}) \approx 0$, for all n large enough.



Single-Letter Characterization of Channel Capacity

- Answer 1 (difficult):

Largest value of R (denoted C_{op}) for which there exist an encoder f_n and a decoder g_n such that $P(W \neq \hat{W}) \approx 0$, for all n large enough.



- Answer 2 (easier):

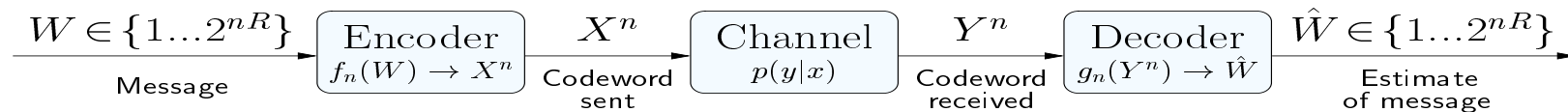
$$C_{\text{inf}} = \max_{p(x)} \mathcal{I}(X \wedge Y).$$

Note: $p(y|x)$ is the channel, $p(x)$ is the distribution of its inputs. Then we have $p(xy) = p(x)p(y|x)$, and hence can compute $\mathcal{I}(X \wedge Y)$.

Single-Letter Characterization of Channel Capacity

- Answer 1 (difficult):

Largest value of R (denoted C_{op}) for which there exist an encoder f_n and a decoder g_n such that $P(W \neq \hat{W}) \approx 0$, for all n large enough.



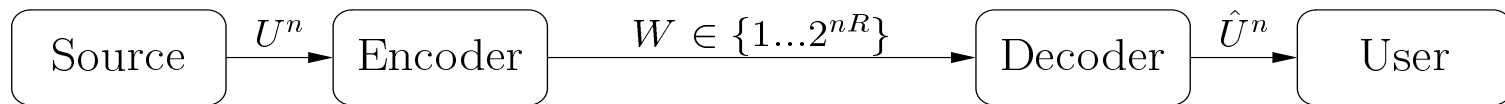
- Answer 2 (easier):
$$C_{\text{inf}} = \max_{p(x)} \mathcal{I}(X \wedge Y).$$

Note: $p(y|x)$ is the channel, $p(x)$ is the distribution of its inputs. Then we have $p(xy) = p(x)p(y|x)$, and hence can compute $\mathcal{I}(X \wedge Y)$.

Most important theorem of information theory: $C_{op} = C_{\text{inf}}$!

Summary: Data Compression and Channel Capacity

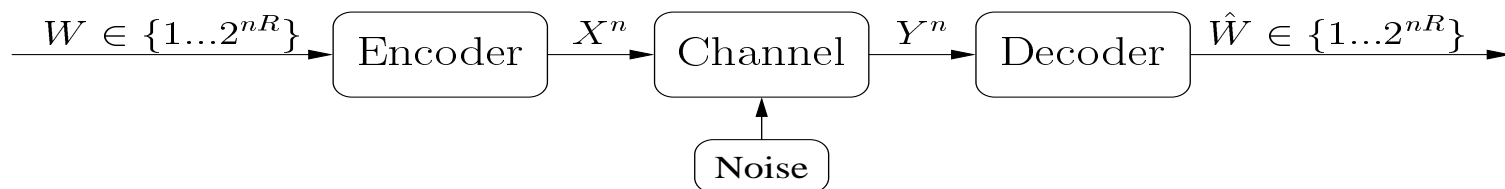
- The data compression theorem:



\hat{U}^n is a good encoding of U^n if and only if $R > \mathcal{H}(U)$.

(i.e., can compress the source into bits, irrespective of the channel.)

- The channel capacity theorem:



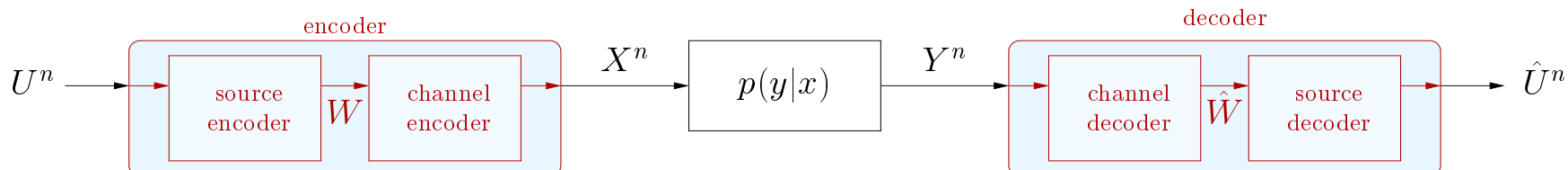
\hat{W}^n is a good estimate of W^n if and only if $R < \mathcal{I}(X \wedge Y)$.

(i.e., can transmit bits over the channel, irrespective of the source.)

The Joint Source/Channel Coding Theorem

We can combine the previous two theorems to obtain the *Joint Source/Channel Coding Theorem*:

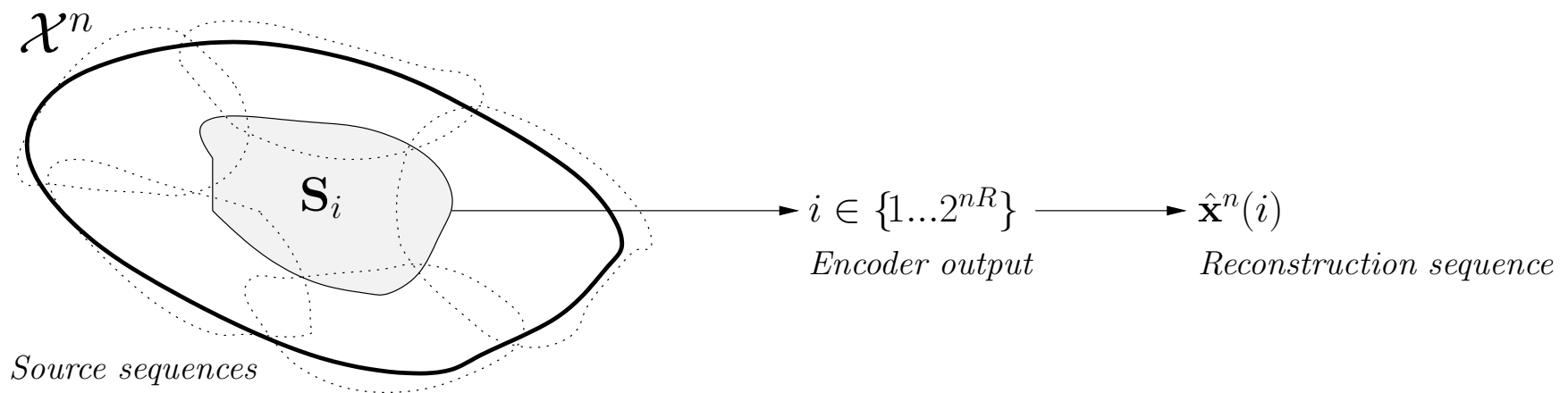
The random variable U can be communicated reliably over the channel $p(y|x)$ if and only if $\mathcal{H}(U) < \mathcal{I}(X \wedge Y)$.



*Because of the fact that it is possible to design the source and channel encoder/decoder independently of each other, this result is also known as the **separation principle**.*

Data Compression Subject to a Fidelity Criterion

What if $\mathcal{H}(X) > \mathcal{C}$? Either change the channel, or else put the source on a diet...

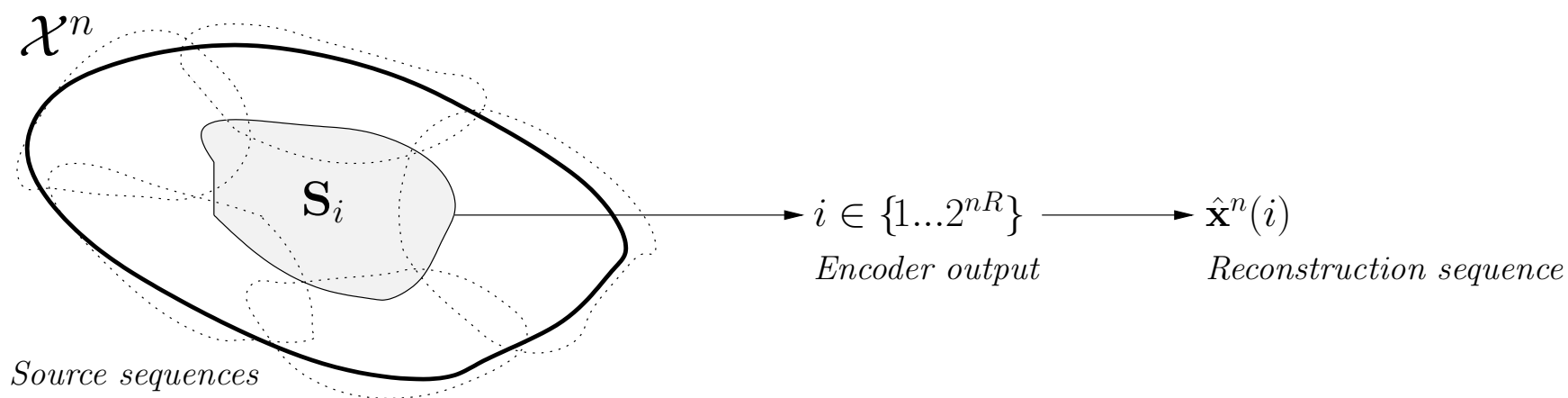


- $\mathcal{R}(D)$: minimum number of bits to achieve distortion D .

-

Data Compression Subject to a Fidelity Criterion

What if $\mathcal{H}(X) > \mathcal{C}$? Either change the channel, or else put the source on a diet...



- $\mathcal{R}(D)$: minimum number of bits to achieve distortion D .

- A computable form:
$$\mathcal{R}(D) = \min_{p(\hat{x}|x): E(X, \hat{X}) \leq D} \mathcal{I}(X \wedge \hat{X}).$$

($p(\hat{x}|x)$ is chosen to satisfy the distortion constraint, $p(x)$ is the source distribution.)

Communication of DMSs over DMCs: a Solved Problem

- Channel capacity: pack “spheres” $A_\epsilon^n(Y|\mathbf{x}^n)$ into $A_\epsilon^n(Y)$.
- Rate distortion: cover $A_\epsilon^n(X)$ with “spheres” $A_\epsilon^n(X|\hat{\mathbf{x}}^n)$.
- Reduce both problems to tractable optimizations:
 - $\mathcal{I}(X \wedge Y)$ is concave in $p(x)$: maximization is “easy.”
 - $\mathcal{I}(X \wedge \hat{X})$ is convex in $p(\hat{x}|x)$, with a linear constraint: minimization also “easy.”
- The Blahut-Arimoto algorithm: an efficient numerical method to solve these optimization problems.

Communication of DMSs over DMCs: a Solved Problem

- Channel capacity: pack “spheres” $A_\epsilon^n(Y|\mathbf{x}^n)$ into $A_\epsilon^n(Y)$.
- Rate distortion: cover $A_\epsilon^n(X)$ with “spheres” $A_\epsilon^n(X|\hat{\mathbf{x}}^n)$.
- Reduce both problems to tractable optimizations:
 - $\mathcal{I}(X \wedge Y)$ is concave in $p(x)$: maximization is “easy.”
 - $\mathcal{I}(X \wedge \hat{X})$ is convex in $p(\hat{x}|x)$, with a linear constraint: minimization also “easy.”
- The Blahut-Arimoto algorithm: an efficient numerical method to solve these optimization problems.

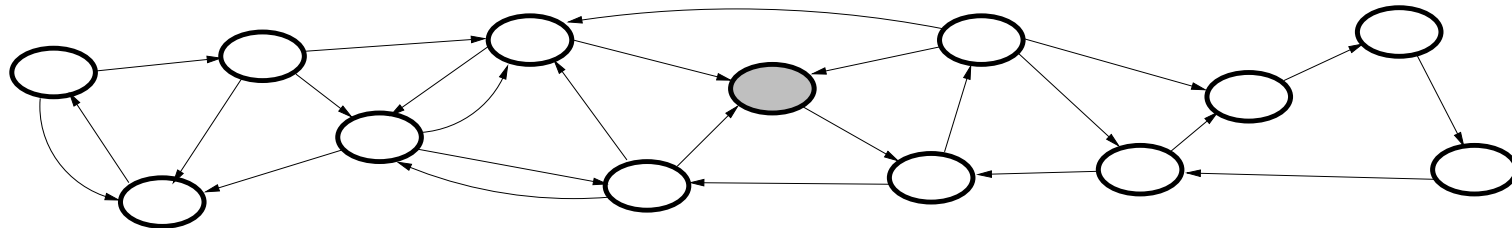
Complete and computable answers. Now, with networks...

Outline

- Reliable Communication over a Single Discrete Memoryless Noisy Channel
- Reliable Communication over a Network of Discrete Memoryless Noisy Channels
 - Data Collection Over a Graph of DMCs
 - Illustration for a Three-Node Network
 - Computability Issues
- Distributed Compression of Dependent Sources
- Summary and Conclusions

Collecting Dependent Data Over a Graph of DMCs

- $M + 1$ nodes $v_0 v_1 \dots v_M$, correlated observations U_i at each v_i .
- For each node pair (v_i, v_j) , a DMC $(\mathcal{X}_{ij}, p(y_{ij}|x_{ij}), \mathcal{Y}_{ij})$, of capacity C_{ij} .
- A special node v_0 , that must reconstruct $U_1 \dots U_M$.



Interference suppressed at the MAC layer – received signals depend only on signals transmitted by one node and on channel noise.

Goal: determine single-letter conditions under which it is possible to reconstruct $U_1 \dots U_M$ at v_0 with low probability of error.

Main Result

There exist codes for this problem with vanishing probability of error *if and only if*, for all non-empty subsets $S \subseteq \{1 \dots M\}$,

$$H(U_S|U_{S^c}) < \sum_{i \in S, j \in S^c} C_{ij}.$$

Furthermore, in a capacity achieving code:

- Source and channel coding are performed separately.
- Decode+Forward is employed, with optimal routes obtained from a suitably defined network flow problem.

J. Barros, S. D. Servetto. *Network Information Flow with Correlated Sources*. IEEE Trans. Inform. Theory, 52(1):155-170, 2006.

Illustration for a Three-Node Network

The main result, specialized to a network with three nodes:

$$H(U_1|U_2U_0) < C_{10} + C_{12}$$

$$H(U_2|U_1U_0) < C_{20} + C_{21}$$

$$H(U_1U_2|U_0) < C_{10} + C_{20}$$

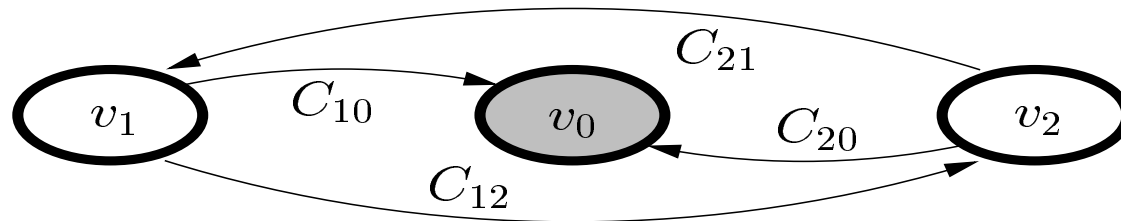
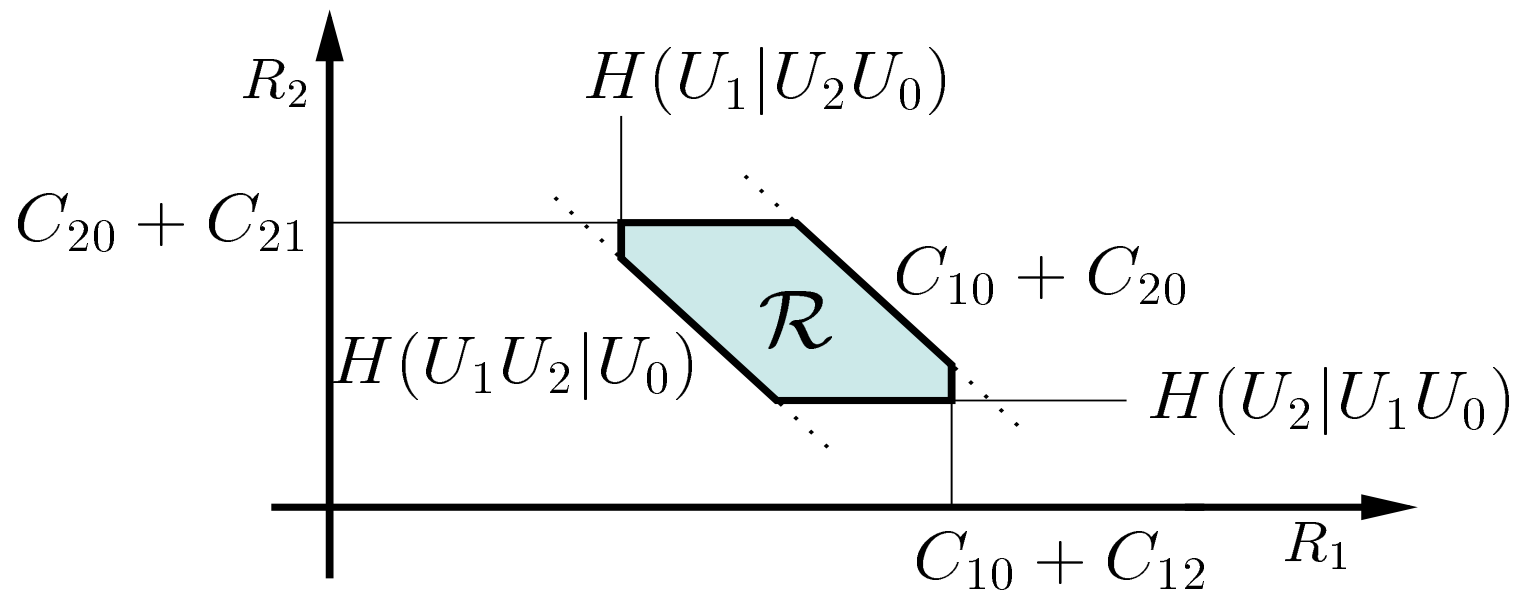


Illustration for a Three-Node Network

The polytope \mathcal{R} that generalizes the condition " $\mathcal{H} < \mathcal{C}$:"



Reliable communication is possible if and only if $\mathcal{R} \neq \emptyset$.

Computability Issues

The number of faces in \mathcal{R} grows exponentially in network size... *computable???*

Computability Issues

The number of faces in \mathcal{R} grows exponentially in network size... *computable???*

- Yes: matroids save the day.
- $\mathcal{R} = \mathcal{P}_1 \cap \mathcal{P}_2$:
 - $f(S) = \sum_{i \in S, j \in S^c} C_{ij}$ is submodular. So $R(S) \leq f(S)$ defines a polymatroid \mathcal{P}_1 .
 - $g(S) = H(U_S | U_{S^c})$ is supermodular. So $R(S) \geq g(S)$ defines a contrapolymatroid \mathcal{P}_2 .
 - Non-emptiness of $\mathcal{P}_1 \cap \mathcal{P}_2$ can be solved in polynomial time.
- Can use this to formulate network design problems.

Computability Issues

The number of faces in \mathcal{R} grows exponentially in network size... *computable???*

- Yes: matroids save the day.
- $\mathcal{R} = \mathcal{P}_1 \cap \mathcal{P}_2$:
 - $f(S) = \sum_{i \in S, j \in S^c} C_{ij}$ is submodular. So $R(S) \leq f(S)$ defines a polymatroid \mathcal{P}_1 .
 - $g(S) = H(U_S | U_{S^c})$ is supermodular. So $R(S) \geq g(S)$ defines a contrapoly matroid \mathcal{P}_2 .
 - Non-emptiness of $\mathcal{P}_1 \cap \mathcal{P}_2$ can be solved in polynomial time.
- Can use this to formulate network design problems.

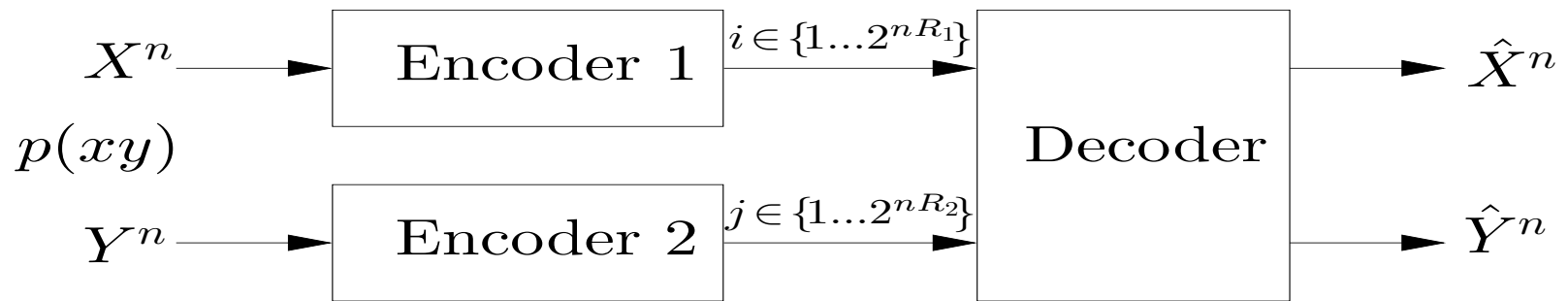
Bottom line: computability issues far more elaborate here. And central to being able to declare the problem to be “solved.”

Outline

- Reliable Communication over a Single Discrete Memoryless Noisy Channel
- Reliable Communication over a Network of Discrete Memoryless Noisy Channels
- Distributed Compression of Dependent Sources
 - Multiterminal Source Coding with Two Encoders
 - Structured Covers of the Product Source
- Summary and Conclusions

Multiterminal Source Coding with Two Encoders

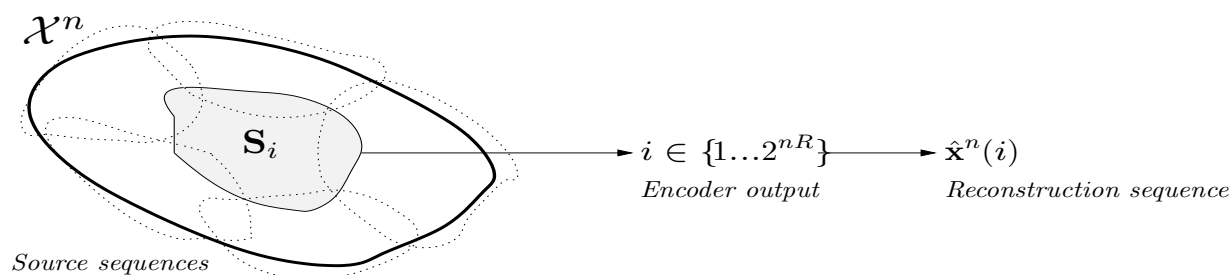
Same rate-distortion problem as before, now with two encoders:



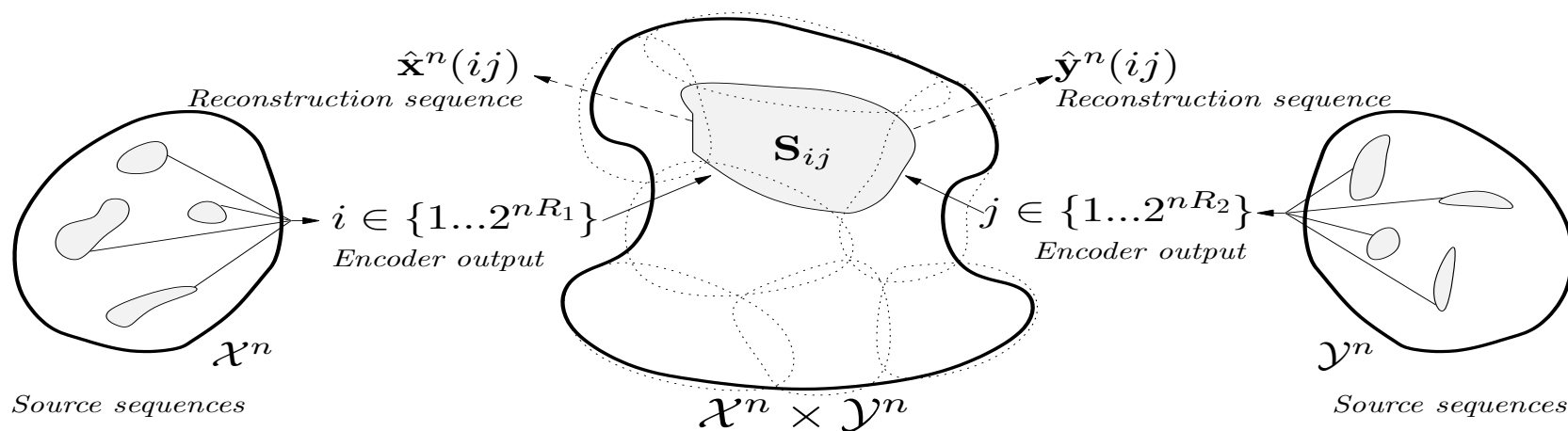
- In the classical rate-distortion problem, solution relied on being able to give a simple description for size of the largest elements that can be used to cover $A_\epsilon^n(X)$...
- ... but such a description is much harder to come by in this problem!

Structured Covers of the Product Source

In the classical problem: $S_i \approx A_\epsilon^n(X|\hat{x}^n(i))$



In this problem: ??? (an entire talk only on this...)



Outline

- Reliable Communication over a Single Discrete Memoryless Noisy Channel
- Reliable Communication over a Network of Discrete Memoryless Noisy Channels
- Distributed Compression of Dependent Sources
- **Summary and Conclusions**

Conclusions

- Information theory started out of Shannon's wish to develop a theory that would explain the fundamental limits of communication systems. Main goal: finding *computable* descriptions of certain combinatorial objects.

-

-

Conclusions

- Information theory started out of Shannon's wish to develop a theory that would explain the fundamental limits of communication systems. Main goal: finding *computable* descriptions of certain combinatorial objects.
- For as long as point-to-point systems were involved, there was a "division of labor:"
 - Design codes and algorithms (LZ, Hamming, arithmetic coding, BCH, ...).
 - Design modems and hardware.

Challenges remain in the theory of point-to-point channels (fading, memory, feedback, ...), but *a lot* of progress has been made.

-

Conclusions

- Information theory started out of Shannon's wish to develop a theory that would explain the fundamental limits of communication systems. Main goal: finding *computable* descriptions of certain combinatorial objects.
- For as long as point-to-point systems were involved, there was a "division of labor:"
 - Design codes and algorithms (LZ, Hamming, arithmetic coding, BCH, ...).
 - Design modems and hardware.

Challenges remain in the theory of point-to-point channels (fading, memory, feedback, ...), but *a lot* of progress has been made.

- In networks, that division of labor is much less clear/effective: computational issues are **the** main challenge to overcome in the construction of the theory.

Conclusions

- Information theory started out of Shannon's wish to develop a theory that would explain the fundamental limits of communication systems. Main goal: finding *computable* descriptions of certain combinatorial objects.
- For as long as point-to-point systems were involved, there was a "division of labor:"
 - Design codes and algorithms (LZ, Hamming, arithmetic coding, BCH, ...).
 - Design modems and hardware.

Challenges remain in the theory of point-to-point channels (fading, memory, feedback, ...), but *a lot* of progress has been made.

- In networks, that division of labor is much less clear/effective: computational issues are **the** main challenge to overcome in the construction of the theory.

Bottom line: who could be better equipped than us (CS people) to tackle these questions? I am certainly trying myself...

Main Corollary...



<http://cn.ece.cornell.edu/>