A review of special topics on

Coding for Collaborative Estimation on Distributed Networks

Emmanuel Oyekanlu

Adaptive Signal Processing and Information Theory Research Group
Drexel University, Philadelphia
Aug. 2, 2012

OUTLINE

Definitions:

- Source Code
- Expected Length of a source code
- Length of a codeword
- Variable/fixed length coding
- Example: Huffman coding
- Lossless coding
- Distortion
- Worst case length of a codeword

Paper Review

First Paper

"The Zero-Error Side Information Problem and Chromatic Numbers", H. S. Witsenhausen

Definitions:

- Graph
- Path
-Chromatic Number
- Characteristic Graph
- Bipartite Graph
- Graph Entopy

Second Paper

"Coding for Computing", A. Orlitsky, R. Roche

Definition of Terms

(1) Source Code: A source code C for a random variable X is a mapping from χ , the range of X, to D* , the set of finite-length strings of symbols from a D-ary alphabet. Let C*(x)* denote the codeword which is a binary sequence corresponding to *x* and let *l*(*x*) denote the length of C*(x)*; *[1]*

Example: C(red) = 00, C(blue) = 11 is a source code for χ = {red, blue} with alphabet D = {0, 1}.

[1] T. Cover, J. Thomas; Element of Information Theory, pg 104

(3) Length of a Codeword: The length λ_j of a codeword $C(A_j)$ is the number of bits of this codeword.

Example: Here is a code for a three symbol alphabet {a, b, c}, that is, A1 = a, A2 = b, A3 = c.

$C(\alpha) = 0$, $C(\beta) = 00111$, $C(\gamma) = 0$. The lengths of $\lambda_1 = 1$, $\lambda_2 = 5$ and $\lambda_3 = 1$ respectively.

(2) The Expected length L(C) of a source code C(x) for a random variable x with probability mass function p(x) is given by

$$L(c) = \sum_{x \in \chi} p(x) l(x)$$

Where $l(x)$ is the length of the codeword associated with x. Without loss of generality, we can assume that the D-ary alphabet is

$$D = \{0, 1, \ldots\ldots, D\text{-}1\}$$

Example: Let x be a random variable with the following distribution and codeword assignment:

$P(X=1)= 1/2$, codeword $C(1) = 0$

$P(X=2)= 1/4$, codeword $C(2) = 10$

$P(X=3)= 1/8$, codeword $C(3) = 110$

$P(X=4)= 1/8$, codeword $C(4) = 111$.

In this example, the expected length $L(C) = El(X)$ of this code is also 1.75 bits

(4)Fixed Length Code: A fixed length code is a code such that $\lambda_i = \lambda_j$ for all i, j. This means that all codewords have the same length (number of bits)

Example: $C(\alpha) = 00$, $C(\beta) = 01$, $C(\gamma)= 10$

Unicode and ASCII are fixed length codes since all characters requires same amount of storage: 16bits and 8 bits respectively.

(5)Variable length Code: It is a code that is not a fixed length code. Variable length code may give different lengths to codewords.

Example: $C(\alpha) = 0$, $C(\beta) = 10$, $C(\gamma)= 11$

Variable length codes can allow sources to be compressed and decompressed with zero error (probability of error is exactly zero when $N \geq 1$ i. e. $(P[(X^N, Y^N) \neq (\hat{X}^N, \hat{Y}^N)] = 0$ when $N \geq 1$) (Example: Huffman- details later).

With fixed length coding, data compression is only possible for large blocks of data and any compression beyond the logarithm of the total number of possibilities comes with a finite probability of failure. Error free compression $(P[(X^N \neq \hat{X}^N] = 0)$ for fixed length codes requires that $R \geq \log|\chi|$; [2]

[2] A.El Gamal, Young-Han Kim; Network Information Theory, pg 70, pp 3

When the source symbols are not equally probable, an efficient encoding method is to use variable-length code words. A good example is Morse code. In Morse code, the letter that occur more frequently are assigned short code words and letters less frequently are assigned long code words.

Problem? How to devise a method for selecting and assigning the codewords to source letters.

Another good example mentioned earlier is Huffman Coding.

A Reminder on Huffman Coding Algorithm:

Huffman (1952) devised a variable-length encoding algorithm, based on the source letter probabilities P(x_i), $I = 1, 2, \ldots\ldots L$.

This algorithm is optimum in the sense that the average number of binary digits required to represent the source symbols is a minimum, subject to the constraint that the code words satisfy the prefix conditions. (a little reminder about prefix conditions: A prefix or instantaneously decodable means no codeword is a prefix of another [3]. Recall the Krapft inequality.

[3] Walsh M.J, "Multiterminal Information Theory", Lecture Notes, Drexel University, Spring Quarter, 2012

(6)Lossless Coding: This is coding with vanishingly small error. It is the original distributed source coding idea as introduced by Slepian and Wolf. It requires that:

$$(P[(X^N , Y^N)\neq(\hat{X}^N,\hat{Y}^N)] \rightarrow 0 \text{ when } N\rightarrow\infty \;)$$

Now compare with zero error coding:

$$(P[(X^N , Y^N)\neq(\hat{X}^N,\hat{Y}^N)] = 0 \text{ when } N \geq 1 \;)$$

[3, 5]

[5] Dragotti P, Gastpar M, Distributed Source Coding, 2009, pg90

Distortion: The distortion $d(\mathrm{x}, \hat{X})$ is a measure of the cost of representing the symbol x by $\hat{X}$. [7]

The distortion between sequences x and $\hat{X}$ is defined by

$$d(\mathrm{x}, \hat{X}) = 1/n \sum_{i=1}^{n} d(Xi, \hat{Xi})$$

The distortion for a sequence is the average of the per symbol distortion of the elements of the sequence.

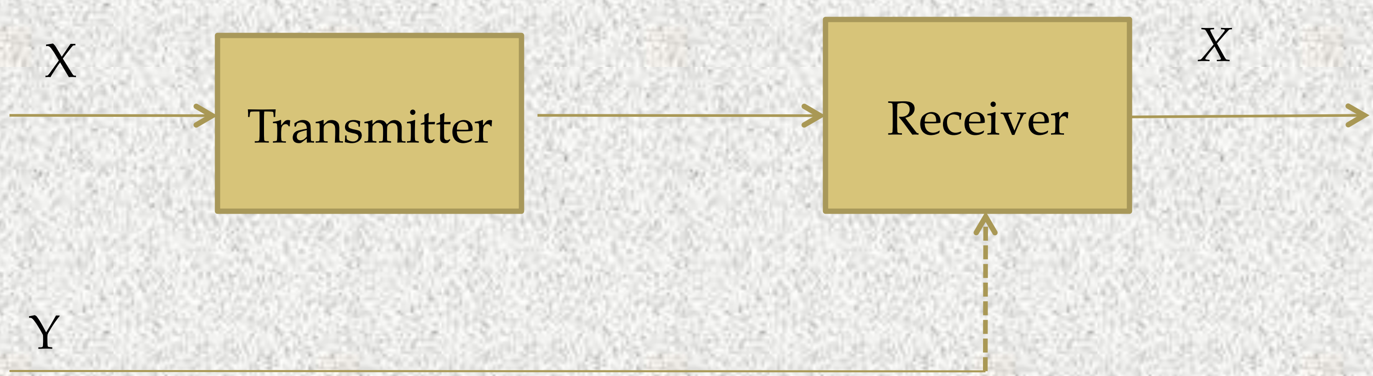
[7] T. Cover, J. Thomas; Element of Information Theory, pg 304-305

Worst Case Length of a Codeword: This is length of the longest codeword.

$$\max_{i} \lambda i$$

Review:

The Zero-Error Side Information Problem and Chromatic Numbers (H. S. Witsenhausen); IEEE 1976.



Witsenhausen attempts to know the number of bits that transmitter Px must transmit in the worst case in order for Py to decode x without error.

A discrete random variable X is to be transmitted by means of a discrete signal. The receiver has prior knowledge of a discrete random variable Y jointly distributed with X. The probability of error must be exactly zero, and the problem is to minimize the signal's alphabet size.

$$\lambda_{max}(\alpha) = \max_i \lambda i\,(\alpha)$$

$$\alpha \mid \quad \min \lambda_{max}(\alpha)$$

$$P[x \neq \hat{x}] = 0$$

We can begin the discourse by defining a construct useful in formulating all results

Graph

$G = (V,E)$ where V is a set of nodes(vertices). E is a set of edges (links).

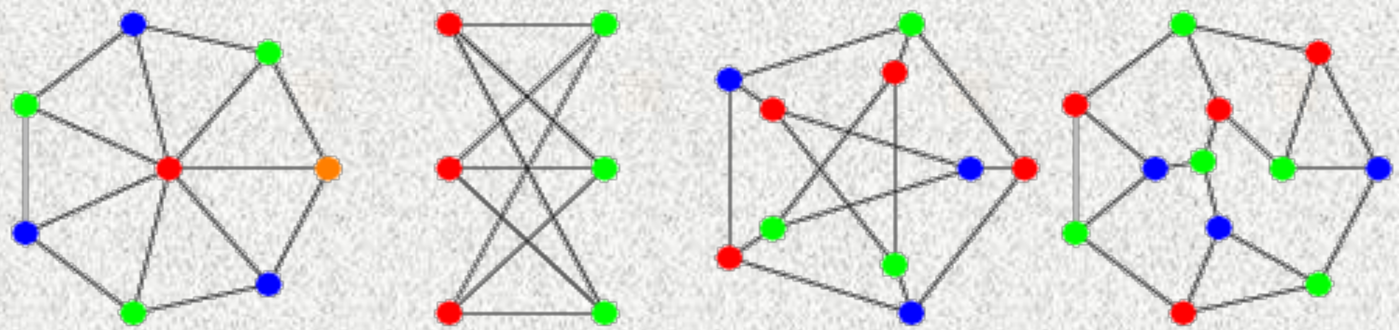
A network can be modeled as a directed acyclic graph $G = (V, E)$. There is a finite set of vertices or nodes V, and collection of edges e ϵ E which are ordered pairs of vertices $e = (v_1, v_2)$, $v_1, v_2 \epsilon V$. We call vertex v_1 the tail of edge $e = (v_1, v_2)$ and the vertex v_2 the head of edge e. A sequence of vertices

e_1 , e_2, ……..,e_k such that the head of edge e_n is the tail of the next edge e_{n+1} is called a directed path in the graph G. A directed path with the property that the tail of e_1 is the head of edge e_k is called a cycle, and the graph is *acyclic* if it has no cycles [3].

[3] Walsh M.J, "Multiterminal Information Theory", Lecture Notes, Drexel University, Spring Quarter, 2012

Vertex Coloring in a Graph

A vertex coloring is an assignment of labels or colors to each vertex of a graph such that no edge connects two identically colored vertices. The most common type of vertex coloring seeks to minimize the number of colors for a given graph [5]. Such a coloring is known as a *minimum vertex coloring*. Example is shown below:

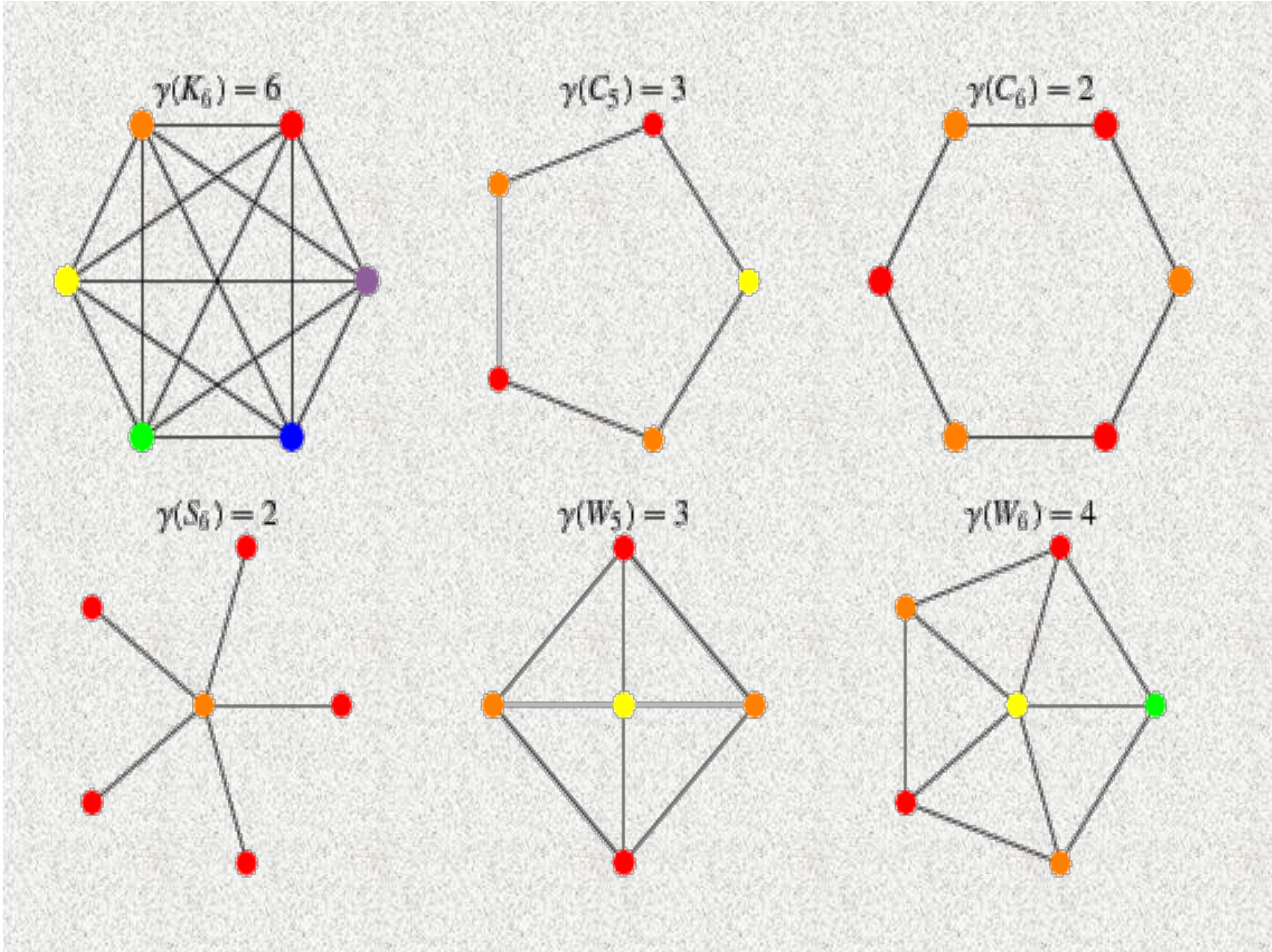


[5]http://mathworld.wolfram.com/VertexColoring.html (accessed on 08/04/2012)

Chromatic Number of a Graph

The chromatic number of a graph G is the smallest number of colors $y(G)$ needed to color the vertices of G so that no two adjacent vertices share the same color.

Consider the examples below for graphs with chromatic numbers: 6, 3, 2, 2, 3, 4


$\gamma(K_6) = 6$
$\gamma(C_5) = 3$
$\gamma(C_6) = 2$
$\gamma(S_6) = 2$
$\gamma(W_5) = 3$
$\gamma(W_6) = 4$

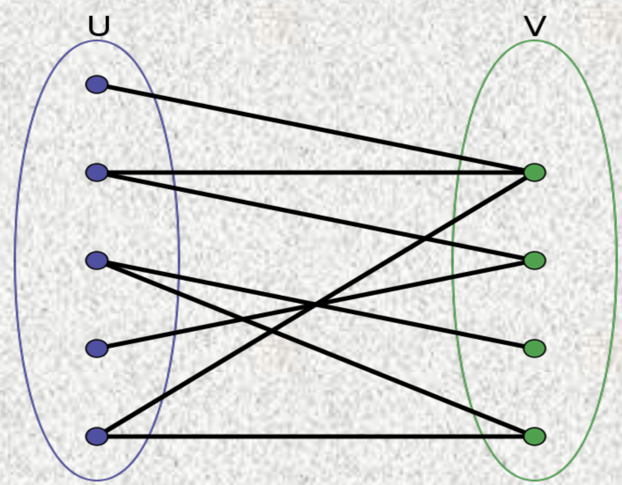
Characteristic Graph

The characteristic graph $G_x = (V_x , E_x)$ of X with respect to Y, $p(x , y)$, and $f(x , y)$ is defined as: $V_x = \chi$, and $(x_1 , x_2) \epsilon \chi^2$ *is in* E_x *if* there exist a y $\epsilon \gamma$ such that $p(x_1 , y)p(x_2 , y) > 0$ and $f(x_1 , y) \neq f(x_2 , y)$ *[8]*

[8] Doshi et'al "Functional Compression Through Graph Coloring", IEEE Transaction on Information Theory, Vol. 56, No 8. Aug., 2010

Bipartite Graph

A graph is bipartite if its vertex set can be partitioned into two disjoint sets U and V such that each edge has one vertex in U and one vertex in V. *[9]*



Graph Entropy

Given a graph G = (V , E) and a distribution on the vertices, the graph entropy could be defined as:

$$H_G(X) = \min_{x \,\epsilon\, w \,\epsilon\, \Gamma(G)} I(W ; X)$$

where Γ(G) is the set of all independent sets of G. The notation x ϵ w ϵ Γ(G) means we are minimizing over all distributions *p(w, x)* such that *p(w , x) > 0* implies *x ϵ w where w* is an independent set of the graph G [10].

[10] J. Körner, "Coding of an Information Source Having Ambiguous Alphabet and the Entropy of Graphs", 6th Prague Conference on Information Theory, 1973, pp. 411-425.

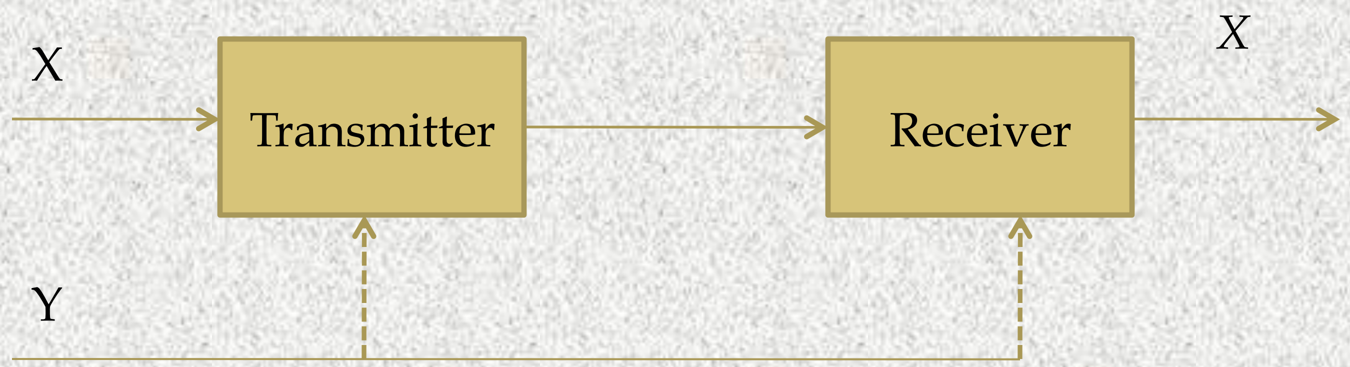
The Witsenhausen problem of Zero-error side information and chromatic numbers involves transmitting X to a receiver which has knowledge of Y (the side information) by means of a discrete signal taking as few values as possible.

No Errors are allowed.

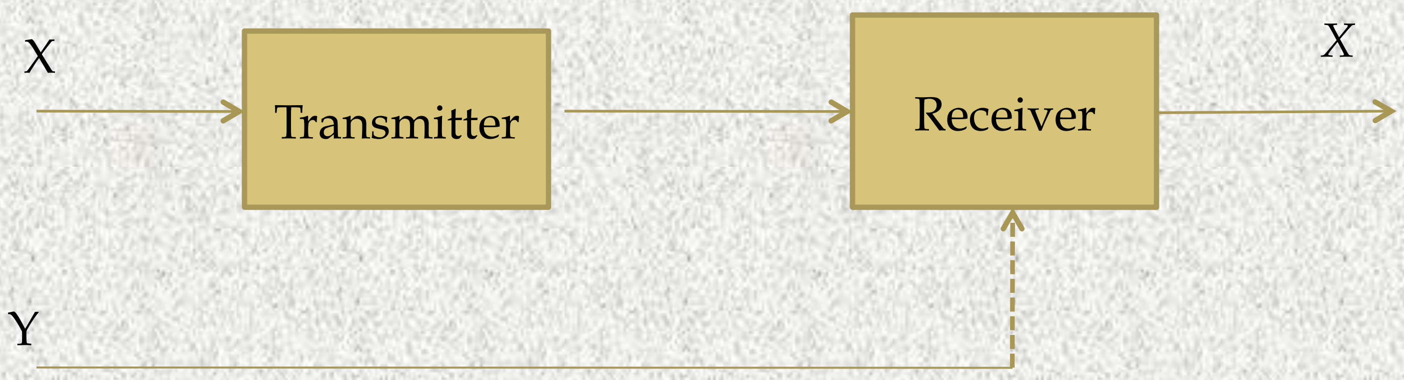
The worst case length of the code must be considered.

There are two cases depending on whether the transmitter has or does not have access to Y. Also, both the single sample situation and the case of block coding over independent repetitions of n pairs (X , Y) need consideration.

Case 1: When Y is available at both Transmitter and Receiver



Case 2: When Y is available only at the receiver



Let X and Y be two discrete random variables with joint distributions P{X = i, Y = j} = p_{ij}, $i = 1, \ldots r$; $j = 1, \ldots m$.
It may be assumed without loss of generality that all marginal probabilities are positive.

Matrix p_{ij} can be specified by a bipartite graph B_{XY} in which two sets of vertices correspond to the alphabets of X and Y respectively, and X-vertex i is joined to Y-vertex j, if and only if pij is positive.

If independent pairs (X_1, Y_1) and (X_2, Y_2) are considered, then the bipartite graph $B(X_1, X_2)(Y_1, Y_2)$ for the pair (X_1, X_2) versus (Y_1, Y_2) is the product of the bipartite graphs BX_1Y_1 and BX_2Y_2. Vertex (X_1X_2) is joined to (y_1y_2) iff X_1 was joined to Y_1 and X_2 to Y_2 in the factor graphs.

From the bipartite graph B_{XY}, a graph G_x is derived in the following way: the vertices of G_x correspond to the X alphabet, x_1 is joined to x_2 by an edge iff there is a vertex y in B_{XY} joined to both x_1 and x_2. In other words, x_1 is joined to x_2 when for some y, $p_{x_1 y} p_{x_2 y} > 0$.

The product of n copies of G_x will be denoted

$$G_X^n$$

When Side information is at both ends

The case where the transmitter also has access to Y is essentially trivial.

Witsenhausen's Proposition

When Y is also known at the transmitter, the minimum signal alphabet size, for encoding a sequence of n independent pairs with $n \geq 1$, is k^n *where* k is the maximum degree of any Y-vertex in B_{XY}

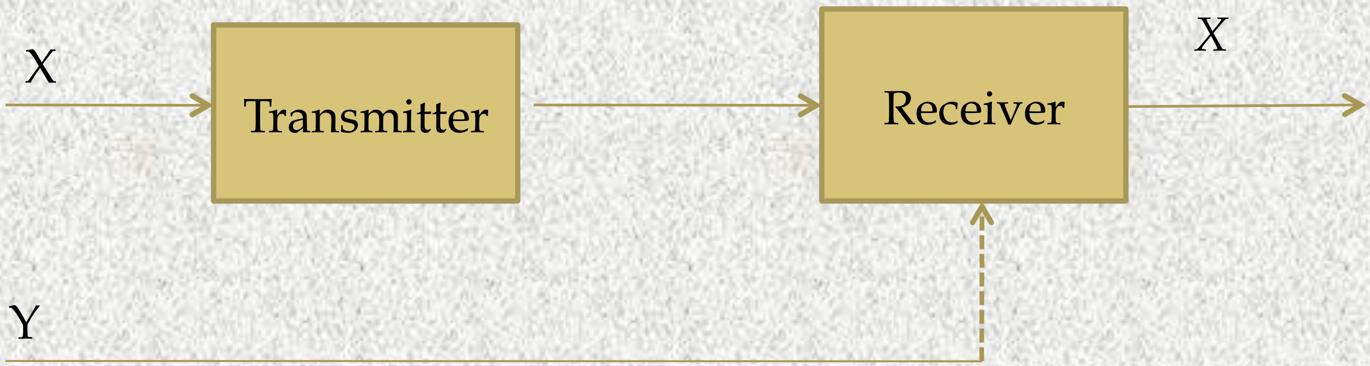
Proof:

For $n = 1$, use the code obtained by labeling the edges at each Y-vertex with distinct elements of $\{1, \ldots\ldots, k\}$. This permits the receiver to determine X from known Y value and the label. If fewer than k signals are used, then, at the vertex achieving degree k, at least two edges are assigned the same signal and hence will create an ambiguity at the receiver. For $n > 1$, the same argument applies to the graph $(B_{XY})^n$ obtained by the product of n copies of B_{XY}, then (y, y, ….., y) achieves k^n, the maximum degree for the product graph.

Witsenhausen's Conclusion

In this case, there is no saving in block encoding

When side information is only known at the Receiver



If the side information is not available at the transmitter and independent sequences of length n is considered, then the problem is to choose a function f such that the signal $Z = f(X_1, \ldots\ldots X_n) = g(Y_1, \ldots\ldots\ldots Y_n, Z)$

Witsenhausen's Proposition

When Y is not known at the transmitter, the minimum signal alphabet size, for encoding a sequence of n independent pairs with $n \geq 1$, is the chromatic number γ (G_X^n) of the product of n copies of the graph G_x

Proof:

First consider the single sample case, n = 1. To each vertex of G_x is assigned the value of *Z (color)* that f takes for the value of X corresponding to the vertex. If x_1 and $_{x2}$ are adjacent, then, for some y, both (x_1 , y) and (x_2 , y) have positive probability and would be undistinguishable to the receiver if x_1 and x_2 had the same color. Conversely, if *f* defines a coloring in which adjacent vertices always have different colors, then for any y, all x_i for which x_i , y has positive probability have different colors because these x_i form a clique (complete subgraph) of Gx by definition of the derived graph. Hence, any such coloring defines an *f* for which unambiguous decoding is possible. The alphabet size of *Z* is the number of colors and its minimum is known as the chromatic number γ (G_X^n)

Witsenhausen considered a characteristic graph with vertices equal to the support of the random variable X and the edge set defined such that x and x′ have an edge f(x) ≠ f(x′) when both x and x′ are jointly probable with y.

He showed that the chromatic number γ (G_X^n) is the minimum signal alphabet size for encoding x such that it could be known at receiver P_y

Witsenhausen's Conclusion

If Y is not known at the transmitter, then the problem is equivalent to the chromatic number problem for graphs, and the block coding may produce savings.

The savings produced however is the worst case length savings. Which is a little bit higher than the expected length of the code.