

# Information Theory Tutorial

## Review of ‘To code or not to code - Lossy Source Channel Communication revisited’

Mahesh Mahadevan

November 15, 2005

### 1 Introduction

In this tutorial, the paper ‘To code or not to code - Lossy Source Channel Communication revisited’ [1] is reviewed. This paper investigates the conditions under which a source-channel communication system is optimal. To test the optimality of a communication system, it is sufficient to measure the average cost and the average distortion, and check if these quantities lie on the optimal cost-distortion tradeoff curve. The separation principle introduced by Shannon in his classic paper [2] states that is optimal to split the source compression and channel coding into two successive stages. Optimality is generally achieved by means (asymptotically) long codewords.

However, the use of long codewords is not necessary. There exist cases for which uncoded transmission is optimal. For example, if a uniform binary source is plugged into a binary symmetric channel, an optimal communication system results (if the distortion is measured in terms of Hamming distance). The same holds if a gaussian source is connected to the input of a gaussian channel. The resulting communication system is optimal in the sense of mean square distortion. In these cases, we observe that optimality results from the ‘probabilistic matching’ of the source and channel statistics. The authors propose that this matching is the fundamental reason that the system achieves optimality and extend this idea to present a basis for communication systems that are optimal. The authors also look at cases of nonergodic and multiuser channels for which the source-channel coding approach gives better performance than the separation principle approach.

### 2 Relevant Definitions

Before we proceed, the basic elements of a communication system are explicitly defined and properties are stated, along with the condition for the *optimality* of a communication system.

*Definition(source)*: A discrete memoryless source  $(p_S, d)$  is defined by an input PMF  $p_S(s)$  drawn from an alphabet  $\mathcal{S}$  and a distortion measure  $d(s, \hat{s})$  (which implies a reconstruction alphabet  $\hat{\mathcal{S}}$  in which the source is reconstructed.)

For this source, we can define a *Rate – Distortion function*

$$R(D) = \min_{p_{\hat{\mathcal{S}}|\mathcal{S}}: E d(S, \hat{S}) \leq D} I(S; \hat{S}). \quad (1)$$

*Definition(channel)*: A discrete memoryless channel  $(p_{Y|X}, \rho)$  is defined by a conditional PMF  $p_{Y|X}(y|x)$  from an input alphabet  $\mathcal{X}$  to an output alphabet  $\mathcal{Y}$  and an input cost function  $\rho(x)$ .

For this source, we can define a *Capacity – Costfunction*

$$C(P) = \max_{p_X: E\rho(X) \leq P} I(X; Y). \quad (2)$$

When there is no cost constraint, we get the unconstrained channel capacity

$$C_0 = \max_{p_X} I(X; Y) \quad (3)$$

*Definition(Source – Channel code):* A Source-Channel code is specified by an encoding function  $F : \mathcal{S}^k \rightarrow \mathcal{X}^m$  and a decoding function  $G : \mathcal{Y}^m \rightarrow \hat{\mathcal{S}}^k$ . The rate  $\kappa = k/m$  is also a parameter of the code. (Note that  $\kappa$  is one of the parameters of the problem statement and is not a parameter to be optimized over.)

The average input cost  $\Gamma$  and the average distortion  $\Delta$  are two parameters that define the performance of the system.

$$\Gamma = \frac{1}{m} \sum_{i=1}^m E\rho(X_i) \quad (4)$$

$$\Delta = \frac{1}{k} \sum_{i=1}^k Ed(S_i, \hat{S}_i) \quad (5)$$

In this way, there are four pairs of entities in a communication system - the source and the channel, the encoder and decoder, the cost function and the distortion, the cost-distortion pair  $(\Gamma, \Delta)$ . These quantities are not all independent of each other, in fact once the first three pairs are specified, the fourth pair is completely determined. The pair  $(\Gamma, \Delta)$  is called the operating point of the system. It is easy to see that there exists a tradeoff between the two parameters i.e in general, to reduce the distortion, we need to increase the average input cost and vice versa. This leads to the definition of the *optimality* of a communication system.

*Definition(Optimality) :* A source channel code  $(F, G)$  of rate  $\kappa$  is optimal for transmission of a source  $(p_S, d)$  across a channel  $(p_{Y|X}, \rho)$  if

- a)  $\Delta$  for  $(F, G)$  is less than that of any other code of rate  $\kappa$  for a given average cost  $\Gamma$ .
- b)  $\Gamma$  for  $(F, G)$  is less than that of any other code of rate  $\kappa$  for a given average distortion  $\Delta$ .

This is equivalent to the alternative definition:

**Lemma 1** *A source channel code  $(F, G)$  is optimal for a source  $(p_S, d)$  and a channel  $(p_{Y|X}(Y|X), \rho)$  iff*

- 1) i)  $\kappa R(\Delta) = C(\Gamma)$
- ii) *neither can  $\Delta$  be lowered without changing  $R(\Delta)$  nor can  $\Gamma$  be lowered without changing  $C(\Gamma)$*
- 2)

$$\Delta = \min_D \{D : R(D) = \max_{D'} R(D')\} \text{ and}$$

$$\Gamma = \min_P \{P : C(P) = \min_{P'} C(P')\}$$

The alternative definition is a result of [2]. For case 1), by combining the rate-distortion theorem and capacity-cost theorem and the data-processing inequality, we see that  $\kappa R(\Delta) \leq C(\Gamma)$ . However, if equality in the previous condition holds it is not sufficient for optimality as it may be possible to reduce  $\Delta$  without changing  $R(\Delta)$  or reduce  $\Gamma$  without changing  $C(\Gamma)$ . Hence the second condition iii). The condition in Case 2) holds when  $\kappa R(\Delta) = C(\Gamma)$  is not *feasible*. If this is true, there does not exist any tradeoff between  $\Delta$  and  $\Gamma$ , but only one operating point .)

### 3 Single letter codes

Next, the authors look at the case when single letter codes perform optimally. Specifically, necessary and sufficient conditions are derived for the existence of single letter codes that are optimal.

For single letter codes,  $F : \mathcal{S} \rightarrow \mathcal{X}$  and  $G : \mathcal{Y} \rightarrow \hat{\mathcal{S}}$ . So,  $\kappa = 1$  and the blocklength is also 1. Single letter codes are easy to implement in terms of encoding and decoding and also operate at zero delay.

For a system deploying a single letter code, the following inequalities hold

$$\begin{aligned} R(\Delta) &= \min_{q_{\hat{S}|S}: Ed(s, \hat{s}) \leq \Delta} I(S; \hat{S}) \\ &\leq I(S; \hat{S}) \leq I(X; Y) \leq \max_{q_X: E\rho(x) \leq \Gamma} I(X; Y) \\ &= C(\Gamma) \end{aligned} \tag{6}$$

These can be derived as a result of the definition of rate-distortion function and cost-capacity function and the data processing inequality. These lead us to the following lemma:

**Lemma 2** *The condition  $R(\Delta) = C(\Gamma)$  holds iff these three conditions are satisfied:*

- a) *The distribution  $p_X$  of  $X$  achieves capacity on the channel with  $E\rho(x) = \Gamma$*
- b) *The distribution  $p_{\hat{S}|S}$  of  $\hat{S}|S$  achieves rate-distortion for the source with  $Ed(s, \hat{s}) = \Delta$*
- c)  *$f(\cdot)$  and  $g(\cdot)$  are 'information lossless' mappings.*

There are various ways in which these conditions can be verified. One particular method is to compute the capacity-cost function  $C(\cdot)$  of the channel and evaluate it at  $\Gamma$ . However, closed form expressions for the constrained capacity exist only for a small set of channels. We face the same problem if we try to calculate the rate-distortion function at  $\Delta$ . The correct approach to tackle this problem is to turn it on its head. For any distribution  $p_X$  at the input of the channel, there exists a cost function  $\rho'$  such that that the distribution  $p_X$  achieves the capacity of the channel under the cost constraint  $\rho'$ . Analogously, there exists a distortion function  $d'$  for a conditional distribution  $p_{\hat{S}|S}$ , such that the rate distortion function is achieved under the distortion constraint  $d'$ . Explicit formulations for the cost function  $\rho$  and the distortion function  $d$  are given in lemmas 3 and 4.

**Lemma 3** *For fixed source distribution  $p_S$ , single letter encoder  $f$  and channel conditional distribution  $p_{Y|X}$ ,*

- i) *for  $I(X; Y) < C_0$ , the first condition of lemma 2 is satisfied iff the input cost function satisfies*

$$\begin{aligned} \rho(x) &= c_1 D(p_{Y|X}(\cdot|x) || p_Y(\cdot)) + \rho_0, \text{ if } p(x) > 0 \\ &\geq c_1 D(p_{Y|X}(\cdot|x) || p_Y(\cdot)) + \rho_0, \text{ if } p(x) = 0 \end{aligned}$$

- ii) *if  $I(X; Y) = C_0$ , the first condition of lemma 2 is satisfied for all  $\rho(x)$ .*

**Lemma 4** *For fixed source distribution  $p_S$ , single letter encoder  $f$ , single letter decoder  $g$  and channel conditional distribution  $p_{Y|X}$ ,*

- i) *for  $I(S; \hat{S}) > 0$ , the second condition of lemma 2 is satisfied iff the distortion function satisfies*

$$d(s, \hat{s}) = -c_2 \log_2 p(s|\hat{s}) + d_0(s) \tag{7}$$

- ii) *if  $I(S; \hat{S}) = 0$ , the second condition of lemma 2 is satisfied for any  $d(s, \hat{s})$ .*

In these equations,  $c_0, c_1, \rho_0$  are positive constants,  $d_0(s)$  is an arbitrary function.  $D(\cdot||\cdot)$  is the Kullback-Leibler distance between two distributions. The proofs of these statements occur as

problems in [[3]p 147].

Consider an input distribution  $p_X$  at the input of the channel. For a particular cost function  $\rho$ , calculate the capacity of the channel. The distribution  $p_X$  may not achieve this capacity. So, now the cost function is found, such that in the set of all distributions that satisfy the cost constraint,  $p_X$  maximizes the mutual information. Lemma 3 says that the form of this cost function  $\rho$  is unique. Analogously in Lemma 4, a distortion function measure is found such that over the set of all distributions satisfying the distortion constraint,  $p_{\hat{S}|S}$  minimizes the mutual information  $I(\hat{S}; S)$ . Again, the form of this distortion function is unique.

In this way, the method to test if the condition  $R(\Delta) = C(\Gamma)$  is satisfied is easy. Once a source-channel code  $(F, G)$  is chosen, it induces distributions on  $p_X(x)$  and  $p_{\hat{S}|S}$ . From these the cost function  $\rho$  and the distortion measure  $d$  calculated from lemma 3 and lemma 4. If these are the same as the actual cost function of the channel and the distortion measure of the source respectively, then the system is optimal. Otherwise it is not optimal.

The condition  $R(\Delta) = C(\Gamma)$  is not sufficient to ensure the optimality of a communication system with single letter encoding. There is also the condition ii) of lemma 1, i.e neither can  $\Delta$  be lowered without changing  $R(\Delta)$  nor  $\Gamma$  can be lowered without changing  $C(\Gamma)$ . This condition comes into play only in these two cases -

- a) The curve of  $C(\Gamma)$  vs  $\Gamma$  is horizontal  $\rightarrow C(\Gamma) = C_0$
- b) The curve of  $R(\Delta)$  vs  $\Delta$  is horizontal  $\rightarrow R(\Delta) = 0$ .

Continuing this logic, we obtain the proposition,

**Proposition 1** *Suppose the transmission of the source  $(p_S, d)$  across the channel  $(p_{Y|X}, \rho)$  using a single letter code  $(f, g)$  satisfies  $R(\Delta) = C(\Gamma)$ .*

1)  $\Gamma$  cannot be lowered without changing  $C(\Gamma)$  iff

- a)  $I(X; Y) < C_0$ , or
- b)  $I(X; Y) = C_0$  and  $p_X$  is one of the distributions that achieves capacity at lowest cost.

2)  $\Delta$  cannot be lowered without changing  $R(\Delta)$  iff

- a)  $I(S; \hat{S}) > 0$ , or
- b)  $I(S; \hat{S}) = 0$  and  $p_{\hat{S}|S}$  is one of the distributions that has zero rate at lowest distortion.

By putting together the conditions we have derived, we can finally establish a simple criterion for the optimality of a communication system employing single-letter codes.

**Theorem 1** *For a source  $(p_S, d)$  and a channel  $(p_{Y|X}, \rho)$  such that  $R(\Delta) = C(\Gamma)$  is feasible, the optimality of a single letter coding scheme  $(f, g)$  is decided by the following statements -*

1. if  $I(S; \hat{S}) \neq I(X; Y)$ , then the system cannot be optimal.
2. if  $0 < I(S; \hat{S}) = I(X; Y) < C_0$ , the system is optimal iff  $\rho(x)$  and  $d(s, \hat{s})$  satisfy lemmas 3 and 4
3. if  $0 < I(S; \hat{S}) = I(X; Y) = C_0$  the system is optimal iff  $d(s, \hat{s})$  satisfies lemma 4 and  $p_X$  achieves  $C_0$  with minimum cost.
4. if  $0 = I(S; \hat{S}) = I(X; Y) < C_0$  the system is optimal iff  $\rho(x)$  satisfies lemma 3 and  $p_{\hat{S}|S}$  achieves zero rate with minimum average distortion.

This criterion is an optimality test only for single letter codes. However, this criterion can easily be extended to the case of any source-channel code because for appropriately extended alphabets,

any source-channel code can be viewed as a single letter code. For discrete memoryless sources and discrete memoryless channels(without feedback), the following results hold:

$$p(s^k) = \prod_{i=1}^k p(s_i) \quad (8)$$

$$p(y^m|x^m) = \prod_{i=1}^m p(y_i|x_i) \quad (9)$$

$$(10)$$

By extending the results of theorem 1 to the new, extended alphabets, we get the corollary-

**Corollary 1** For a source  $(p_{S^k}, d^{(k)})$ , and a channel  $(p_{Y^m|X^m}, \rho^{(m)})$ , suppose  $R(\Delta) = C(\Gamma)$  is feasible. Consider the transmission using a single letter source channel code  $F : S^k \rightarrow X^m, G : Y^m \rightarrow \hat{S}^k$  and suppose  $0 < I(S^k, \hat{S}^k) = I(X^m, Y^m) < C_0$

This system is optimal iff

$$\begin{aligned} \rho^{(m)}(x^m) &= c_1 D(p_{Y^m|X^m}(\cdot|x^m) || p_Y(\cdot)) + \rho_0, \text{ if } p(x^m) > 0 \\ &\geq c_1 D(p_{Y^m|X^m}(\cdot|x^m) || p_Y(\cdot)) + \rho_0, \text{ if } p(x^m) = 0 \\ d^k(s^k, \hat{s}^k) &= -c_2 \log_2 p(s^k | \hat{s}^k) + d_0(s) \end{aligned}$$

The last result demonstrates the concept of 'probabilistic matching'. Given the source and channel statistics, the above result gives the form of cost function and distortion measure for which the system is optimal. The goal of the code design can be said to choose the code such that the cost function and the distortion function from Corollary 1 are the closest match to the source distortion measure and the channel input cost function. This is where it is useful to increase the blocklength. By increasing the blocklength (and therefore the coding complexity) , it is possible to obtain a better match between the source and channel statistics and the cost and distortion functions.

It is interesting to find what code length is required to find the 'optimal' match. The authors consider this case in their paper. Specifically, for M -letter codes of rate 1, the following theorem can be proved.

**Theorem 2** Consider a source  $(p_S, d)$ , channel  $(p_{Y|X}, \rho)$ . Suppose that all alphabets are of the same size and the distortion measure has the property the matrix  $\{ 2^{-d(s, \hat{s})} \}$  is invertible and the channel probability matrix is invertible. Then there exists a finite block length code that performs optimally iff there exists a single-letter code which performs optimally for the same source-channel pair.

This theorem is proved in [4] . Though the restriction on  $\{ 2^{-d(s, \hat{s})} \}$  seems unusual, it is actually satisfied by most standard distortion measures like, for example, the squared-error distortion.

## 4 Extensions to Nonergodic and Multiuser Systems

The separation theorem [2] says that for point-to-point communication systems, performing the source coding and channel coding separately is optimal. However, this approach is sensitive to changes in source and channel statistics. For example, if channel degrades and the capacity of the channel goes below the rate at which information is transmitted over the channel, the error rate can

increase. Source - channel codes may offer a more graceful degradation of optimality with variation in channel parameters.

This can be seen in the following example. Consider a binary uniform source transmitting over a binary symmetric channel with parameter  $\epsilon$ . We can see that the single letter code with identity mapping is optimal for this case [eg 1 [1]]. For any value of  $\epsilon < 1/2$ , we see that this code is optimal. hence this code is universal for the transmission of a binary uniform source over a binary symmetric channel.

Instances of universality can be characterized by applying theorem 1 to a class of channels.

**Corollary 2** *The single letter code  $(f, g)$  is optimal for transmission of a source  $(p_S, d)$  over a class of channels  $\mathcal{W}$  iff for every channel  $i$  in  $\mathcal{W}$*

$$\begin{aligned} \rho^{(i)}(x) &= c_1^{(i)} D(p_{Y|X}^{(i)}(\cdot|x)||p_Y(\cdot)) + \rho_0^{(i)}, \text{ if } p(x) > 0 \\ &\geq c_1^{(i)} D(p_{Y|X}^{(i)}(\cdot|x)||p_Y(\cdot)) + \rho_0^{(i)}, \text{ if } p(x) = 0 \\ d(s, \hat{s}) &= -c_2^{(i)} \log_2(p^{(i)}(s|\hat{s})) + d_0^{(i)}(s) \end{aligned}$$

The authors also discuss a case in which separation theorem approach is inferior to the source-channel coding approach. The case considered is of single source gaussian broadcast channel with two users. We can see that uncoded transmission is optimal on each channel individually. In [pg 1155[1]] it is proved that the operating point achieved by uncoded transmission is superior to the curve achieved by a separation theorem approach.

## 5 Conclusion and other work

The separation theorem [2] states that for point-to-point ergodic communication systems, it is optimal to perform the source compression and channel coding separately. The system designed according to the separation principle achieves optimality by means of asymptotically long code-words, though this is not a requirement. Instead, the authors adopt the approach that optimality results from "probabilistic matching" of the source distribution and the channel statistics. This "matching" can occur independently of blocklength and indeed, the authors derive a necessary and sufficient condition for optimality of a source-channel communication system employing single-letter codes. Also, the separation principle is limited to ergodic point-to-point communication. However, source-channel codes perform optimally in certain nonergodic multi-user scenarios as in the example discussed in section 4. The authors have continued this approach and applied the principle of "probabilistic matching" to source-channel networks. Their results are presented in [4],[5] and [6].

## References

- [1] M Gastpar B. Rimoldi and M. Vetterli "To code or not to code", IEEE transactions on information theory vol 49 ,may 2003.
- [2] C.E Shannon "A Mathematical Theory of Communication" .
- [3] I. Csiszar and J. Korner ,Information theory : coding theory for discrete memoryless Sources.
- [4] "To code or not to code" , PhD dissertation, EPFL, Lausanne, Switzerland 2002.

- [5] "On the Capacity of large Gaussian Relay Networks", IEEE Transactions on Information Theory , submitted for publication.
- [6] M. Gastpar and M. Vetterli "On the Capacity of Wireless Networks: the Relay case", in Proc IEEE Infocom 2002, New York.