6.962 Week 3 Summary:

Communication with Side Information at the Transmitter

Presenter: Aaron Cohen

February 22, 2001

1 Introduction

This week we will be looking at how much information can be transmitted over a channel when side information about the channel is available to the transmitter. This problem is dual in many respects to the source coding with side information at the decoder that we discussed last week.

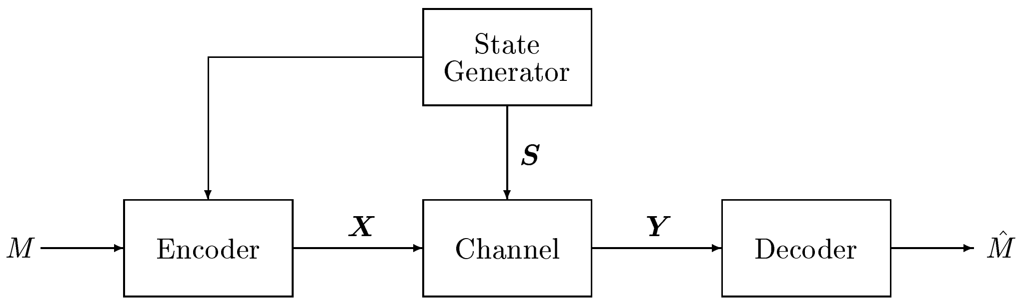

Figure 1: Communication with side information at the transmitter.

The basic problem that we will be looking at can be described as follows and is illustrated in Figure 1. The overall objective (as usual) is to reliably communicate a message $m \in \{1, \ldots, 2^{nR}\}$ from the encoder to the decoder using the channel n times. The largest rate R for which this is possible is called the capacity of the channel. The channel that we can use is described by a conditional probability distribution $p(y|x, s)$, where $x \in \mathcal{X}$ is the input to the channel, $s \in \mathcal{S}$ is the state of the channel, and $y \in \mathcal{Y}$ is the output of the channel. Initially, we will assume that the sets $\mathcal{X}$, $\mathcal{S}$, and $\mathcal{Y}$ are finite. The states of the channel are generated in an IID fashion using a given distribution $p(s)$. The encoder uses his knowledge of the channel state and the message M to be transmitted in order to produce a sequence of inputs to the channel $\boldsymbol{x} \in \mathcal{X}^n$. We will distinguish between two cases:

1. The ith channel input x_i can be a function of only the message and the channel states up to and including time i. In this case, we will say that the encoder has *causal* side information.

2. The channel input vector $\boldsymbol{x} \in \mathcal{X}^n$ can be a function of the message and the channel state vector $\boldsymbol{s} \in \mathcal{S}^n$. In this case, we will say that the encoder has *non-causal* side information.

Finally, the decoder uses the vector of outputs from the channel $\boldsymbol{y} \in \mathcal{Y}^n$ to produce an estimate the message $\hat{m}$.

Shannon [1] found the capacity when the encoder has causal side information and Gel'fand and Pinsker [2] found the capacity when the encoder has non-causal side information. We will discuss their results below in Section 3. An important example is when the state acts in an additive manner, as illustrated in Figure 2. Costa's writing on dirty paper [3] model is exactly this situation when

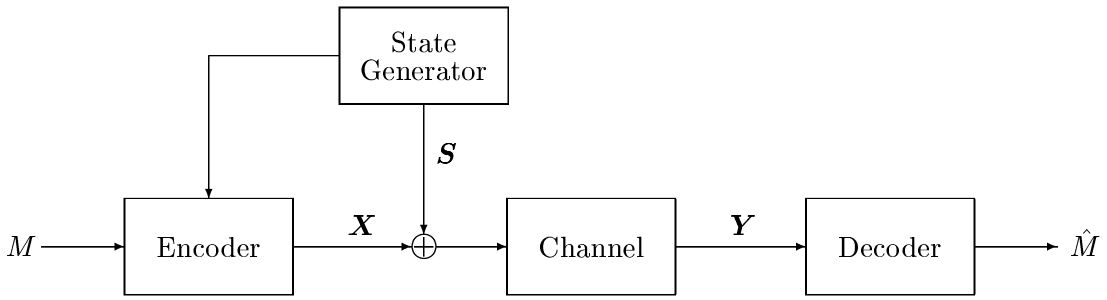

Figure 2: Example with additive state.

the state is IID Gaussian, the channel is AWGN, and the encoder has non-causal side information and has a limited amount of power. We will discuss his results and some extensions [4, 5, 6] in Section 4.

2 Related problems

A similar model to Figure 2 and writing on dirty paper arises in information embedding or watermarking. In that scenario, the state sequence is some data in which extra information needs to be embedded. For example, the data could be a Led Zeppelin song and the information to embed could be the ID number of the rightful owner of the song. The encoder can choose the vector $\boldsymbol{X}$

based on the message M and the *entire* song $\boldsymbol{S}$ (thus, the non-causal side information assumption makes sense here). He is constrained to make the version distributed to the public $\boldsymbol{S}+\boldsymbol{X}$ and the original version $\boldsymbol{S}$ similar in some sense, i.e., the "distortion" introduced by $\boldsymbol{X}$ should be small. If the purpose of the watermarking system is to make sure that the information can be recovered after the data has been transmitted over some known channel, then the analogy is complete. More generally, we might want to make sure that the information can be recovered from any transformation of the data that does not introduce too much additional distortion. Several researchers have studied aspects of this problem including Moulin and O'Sullivan [7, 8] (general capacity when ML decoder is available), Merhav [9] (error exponents), and my own work [4] (capacity for Gaussian data).

Many other variations on the problem described in Section 1 have been studied. The most obvious is to ask what happens when the state is known to the decoder or not known at all. However, these problems can be seen as simple discrete memoryless channels (at least when the state is generated IID). One interesting case studied by Heegard and El Gamal [10] is when there is only a rate R_e to describe the state sequence to the encoder and a rate R_d for the decoder (when R_e is unlimited and R_d is zero, then this case becomes the non-causal side information problem discussed above). Another way to modify the problem is to change the way the state sequence is generated. The first possibility is to make the state sequence a more general stochastic process (e.g., the Gilbert-Elliot channel). Other examples include when the state sequence can be arbitrary (the arbitrarily varying channel) or when a state is chosen randomly at the beginning and then remains fixed throughout the transmission (the compound channel). The past inputs could also affect the current state as in the intersymbol interference (ISI) channel. An excellent review of many of these channel models can be found in [11].

3 Capacity results

1. [1] The capacity with causal side information at the transmitter is given by

$$C_c = \max_{p(u), f: \mathcal{U}\times\mathcal{S}\mapsto\mathcal{X}} I(U;Y), \tag{1}$$

where U is an auxiliary random variable taking value in a set $\mathcal{U}$ with $|\mathcal{U}| \leq |\mathcal{Y}|$ and the joint distribution of the random variables S, U, X and Y is given by

$$p(s,u,x,y) = \begin{cases} p(s)p(u)p(y|x,s) & \text{if } x = f(u,s) \\ 0 & \text{otherwise} \end{cases}. \tag{2}$$

2. [2] The capacity with non-causal side information at the transmitter is given by

$$C_{nc} = \max_{p(u|s), f:\mathcal{U}\times\mathcal{S}\mapsto\mathcal{X}} I(U;Y) - I(U;S), \tag{3}$$

where $|\mathcal{U}| \leq |\mathcal{X}| + |\mathcal{S}|$ and

$$p(s,u,x,y) = \begin{cases} p(s)p(u|s)p(y|x,s) & \text{if } x = f(u,s) \\ 0 & \text{otherwise} \end{cases}. \tag{4}$$

As one would expect, $C_c \leq C_{nc}$, which one can see since the random variables U and S are independent (and hence $I(U;S) = 0$) under the joint distribution (2). Thus, we are maximizing the same expression in both (1) and (3), but we are maximizing over a larger set in the latter case.

Let us first examine the forward part in the causal case. We can think of the auxiliary random variable U as indexing a set of functions from $\mathcal{S}$ to $\mathcal{X}$, and each function generating a distribution on the output set $\mathcal{Y}$. There are $|\mathcal{X}|^{|\mathcal{S}|}$ possible functions, but we only need to use at most $|\mathcal{Y}|$ to achieve capacity, just as in any discrete memoryless channel. Once we have chosen a distribution, we generate codebooks in the usual way. That is, we generate 2^{nR} sequences $\boldsymbol{U}(1), \ldots, \boldsymbol{U}(2^{nR})$ independently and where each $\boldsymbol{U}(i)$ is a length-n IID sequence according to the distribution $p(u)$. At time j, the encoder takes the message m and the state s_j and produces the channel input $x_j = f(u_j(m), s_j)$, which clearly only uses the side information in a causal manner. A joint typicality decoder will find the correct message to be typical with high probability and find no other message jointly typical with high probability as long as $R < I(U;Y)$.

In the non-causal case, we need to use the binning technique discussed last week for source coding with side information. For each message m, generate a bin of 2^{nR_0} codewords $\boldsymbol{U}(m,1), \ldots \boldsymbol{U}(m, 2^{nR_0})$ in the same manner as above. Given the message m and the state sequence $\boldsymbol{s}$, the encoder finds the codeword in bin m that is jointly typical with $\boldsymbol{s}$, say $\boldsymbol{u}(m,j)$ (if no such j exists, then declare an error). The encoder then creates the input to the channel as $\boldsymbol{x} = f(\boldsymbol{u}(m, j), \boldsymbol{s})$. The decoder finds an $\hat{m}$ and a $\hat{j}$ such that $\boldsymbol{u}(\hat{m}, \hat{j})$ is jointly typical with the channel output sequence $\boldsymbol{y}$. The probability of encoding failure goes to zero as long as $R_0 > I(U; S)$ and the probability of a decoding failure goes to zero as long as $R + R_0 < I(U; Y)$. Thus, overall probability of error goes to zero as long as $R < I(U; Y) - I(U; S)$.

The converses are less interesting in that they do not tell us how to implement anything. However, we should investigate the difference between the two capacities. The crucial difference is that in the causal case[1] $(M, Y^{i-1}) \multimap (M, S^{i-1}) \multimap Y_i$. This follows critically from the causal nature of the side information. Letting $U(i) = (M, S^{i-1})$, we can write

$$
\begin{aligned}
nR - H(M|Y^n) &= H(M) - H(M|Y^n) \\
&= I(M; Y^n) \\
&= \sum_{i=1}^{n} I(M; Y_i|Y^{i-1}) \\
&\leq \sum_{i=1}^{n} I(M, Y^{i-1}; Y_i) \\
&\leq \sum_{i=1}^{n} I(M, S^{i-1}; Y_i) \\
&= \sum_{i=1}^{n} I(U(i); Y_i) \\
&\leq nC_c,
\end{aligned}
$$

where the second inequality follows from the data processing inequality using the above Markov chain and the final inequality follows since $U(i)$ is independent of S_i. If the message is recovered reliably, then $H(M|Y^n)$ must be small (Fano's inequality). Thus, the rate cannot be larger than C_c if the message is to be recovered reliably. Shannon [1] more precisely lower bounds the probability of error when a rate larger than C_c is used. The converse in the non-causal case follows is a similar but more messy manner.

[1]The notation $X \multimap Y \multimap Z$ indicates that the random variable X, Y and Z form a Markov chain.

3.1 Example

Here is an example given in [10]. There are three states for a channel with binary inputs and outputs, i.e., $\mathcal{S} = \{a, b, c\}$, $\mathcal{X} = \mathcal{Y} = \{0, 1\}$. In state a, the channel is a binary symmetric channel (BSC) with crossover probability ϵ. In state b, the output is always 0. In state c, the output is always 1. State a occurs with probability $1-p$ and states b and c each occur with probability $p/2$.

Let us first find the capacity in the causal case. Let $\mathcal{U} = \{u_0, u_1\}$ and let $f(u_i, s) = i$ for all $i \in \{0, 1\}$ and all $s \in \mathcal{S}$. If $U = u_0$, then $Y = 0$ with probability $(1-\epsilon)(1-p) + p/2$. If $U = u_1$, then $Y = 1$ with probability $(1-\epsilon)(1-p) + p/2$. The distribution on $\mathcal{Y}$ would be one of those two for any other function from $\mathcal{S}$ to $\mathcal{X}$, and thus we only need consider those two. The channel from $\mathcal{U}$ to $\mathcal{Y}$ is a BSC with crossover probability $p/2 + \epsilon(1-p)$ and thus $C_c = 1 - h(p/2 + \epsilon(1-p))$, where $h(\cdot)$ is the binary entropy function.

Let us now find the capacity in the non-causal case. We need a larger set $\mathcal{U} = \{u_0, u_1, u_2, u_3\}$. The function $f(u_i, s)$ will equal 0 if i is even and 1 is i is odd. Given $S = s$, let U be equiprobable between u_0 and u_1 if $s = a$, let $U = u_2$ if $s = b$, and let $U = u_3$ if $s = c$. To summarize, the four possible outcomes for the random variables S, U, and X are $(a, u_0, 0)$ with probability $(1-p)/2$, $(a, u_1, 1)$ with probability $(1-p)/2$, $(b, u_2, 0)$ with probability $p/2$, and $(c, u_3, 1)$ with probability $p/2$. With this joint distribution, $I(U; S) = H(U) - H(U|S) = 1 - (1-p)h(1/2) = p$. Furthermore, $I(U; Y) = H(Y) - H(Y|U) = 1 - (1-p)h(\epsilon)$. Thus, $I(U; Y) - I(U; S) = (1-p)(1-h(\epsilon))$. We cannot do better than this since this is the capacity with side information at both the encoder and decoder. Thus, $C_{nc} = (1-p)(1-h(\epsilon)) > C_c$. The capacity is larger in the non-causal case because we can find a codeword that matches up with the entire state sequence.

4 Writing on dirty paper

Costa's writing on dirty paper [3] was briefly described above and is summarized here in Figure 3. In this model, $\boldsymbol{S}$ is a sequence of IID $\mathcal{N}(0, Q)$ random variables, $\boldsymbol{Z}$ is a sequence of $\mathcal{N}(0, N)$ random variables, and $\boldsymbol{X}$ is constrained to have power P. Costa's main result is that $C_{nc} = \frac{1}{2} \log\left(1 + \frac{P}{N}\right)$ which is the *same* as if $\boldsymbol{S}$ were known at the decoder or if $\boldsymbol{S}$ were not present at all. Although the capacity expressions were only given for finite alphabets, we will compute C_{nc} anyway and check later that it is valid. We will then give some extensions of his basic model.

Let X be a $\mathcal{N}(0, P)$ random variable independent of S and let $U = X + \alpha S$. (We could have first generated U correlated with S and then made X a function of U and S as above.) Costa

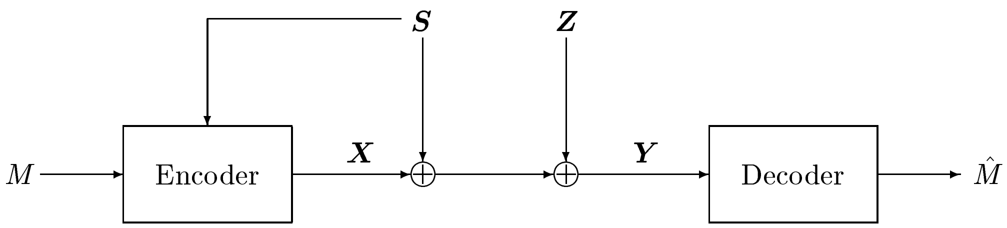

Figure 3: Writing on dirty paper.

computes $I(U;Y) - I(U;S)$ for this joint distribution and then optimizes over α. However, by first picking the optimal α, we can gain some additional insight. Let $\alpha^* = \frac{P}{P+N}$, which (from Costa) we know to be the optimal choice. Note that $\alpha^*(X+Z)$ is the MMSE estimate of X given $X+Z$, and in particular, $X - \alpha^*(X+Z)$ is independent of $X+Z$. Furthermore, $X - \alpha^*(X+Z)$ is independent of $Y = X + S + Z$ since they are jointly Gaussian and uncorrelated, but more generally since $X - \alpha^*(X+Z) -\!\!\circ\!\!- X + Z -\!\!\circ\!\!- X + S + Z$. We now compute

$$I(U;Y) - I(U;S) = \big(h(U) - h(U|Y)\big) - \big(h(U) - h(U|S)\big) = h(U|S) - h(U|Y).$$

Next,

$$h(U|S) = h(X + \alpha^* S|S) = h(X|S) = h(X),$$

where the last equality follows since X and S are independent. Also,

$$\begin{aligned} h(U|Y) &= h(X + \alpha^* S|Y) \\ &= h\big(X + \alpha^*(S - Y)|Y\big) \\ &= h\big(X - \alpha^*(X + Z)|Y\big) \\ &= h\big(X - \alpha^*(X + Z)\big) \\ &= h\big(X - \alpha^*(X + Z)|X + Z\big) \\ &= h(X|X + Z), \end{aligned}$$

where the fourth and fifth equalities follow from the above discussed independence. Combining these steps, we get that

$$\begin{aligned} I(U;Y)-I(U;S) &= h(U|S)-h(U|Y) \\ &= I(X;X+Z) \\ &= \frac{1}{2}\log\left(1+\frac{P}{N}\right). \end{aligned}$$

Since this is the capacity if $\boldsymbol{S}$ were not present at all, then this must be C_{nc}.

This new proof (given in [4]) allows us to extend Costa's result on when C_{nc} equals the capacity when $\boldsymbol{S}$ is not present at all. Indeed, the fact that $\boldsymbol{S}$ was Gaussian was not used anywhere in the proof, and thus $\boldsymbol{S}$ can have any (power-limited) distribution. A more general sufficient condition on the noise $\boldsymbol{Z}$ is if $\boldsymbol{X}$ has the capacity achieving distribution for the additive noise channel $\boldsymbol{Y}=\boldsymbol{X}+\boldsymbol{Z}$ and there exists a linear function $\alpha(\cdot)$ such that $\boldsymbol{X}-\alpha(\boldsymbol{X}+\boldsymbol{Z})$ is independent of $\boldsymbol{X}+\boldsymbol{Z}$. This is true if $\boldsymbol{Z}$ is a colored Gaussian process since the capacity achieving distribution is also Gaussian (with correlation matrix given by the waterfilling algorithm). A similar extension was given by Erez, Shamai and Zamir [6]. Brian Chen [5] previously showed that Costa's result is true if $\boldsymbol{S}$ and $\boldsymbol{Z}$ are both colored Gaussian.

No one has been able to compute C_c for this problem. (The causal version has been dubbed "writing on dirty tape," although if my wife asks, I don't know anything about any dirty tape.) However, some interesting asymptotic results were obtained in [6]. In particular, if we let $Q\to\infty$ and $N\to 0$, then $C_{nc}-C_c\to\frac{1}{2}\log\left(\frac{\pi e}{6}\right)$, which any student of Forney should recognize as the "shaping gain". The shaping gain also tells us the difference (asymptotically) in mean squared error between optimal uniform vector quantization versus uniform scalar quantization (both followed by entropy coding). Thus, it should not be surprising that this quantity gives the difference between optimal use of vector side information and optimal use of scalar side information.

References

[1] C. E. Shannon, "Channels with side information at the transmitter," *IBM Journal of Research and Development*, vol. 2, pp. 289–293, Oct. 1958.

[2] S. I. Gel'fand and M. S. Pinsker, "Coding for channel with random parameters," *Problems of Control and Information Theory*, vol. 9, no. 1, pp. 19–31, 1980.

[3] M. H. M. Costa, "Writing on dirty paper," *IEEE Trans. Inform. Theory*, vol. 29, pp. 439–441, May 1983.

[4] A. S. Cohen and A. Lapidoth, "The Gaussian watermarking game – Part I." Submitted to *IEEE Trans. Inform. Theory*, Jan. 2001.

[5] B. Chen, *Design and Analysis of Digital Watermarking, Information Embedding, and Data Hiding Systems.* PhD thesis, MIT, Cambridge, MA, 2000.

[6] U. Erez, S. Shamai, and R. Zamir, "Capacity and lattice-strategies for cancelling known interference," in *Proceedings of the Cornell Summer Workshop on Information Theory*, Aug. 2000.

[7] J. A. O'Sullivan, P. Moulin, and J. M. Ettinger, "Information theoretic analysis of steganography," in *Proc. of the Inter. Symposium on Info. Theory*, (Cambridge, MA), p. 297, 1998.

[8] P. Moulin and J. A. O'Sullivan, "Information-theoretic analysis of information hiding." Preprint, available at `http://www.ifp.uiuc.edu/~moulin/paper.html`, 1999.

[9] N. Merhav, "On random coding error exponents of watermarking systems," *IEEE Trans. Inform. Theory*, vol. 46, pp. 420–430, Mar. 2000.

[10] C. Heegard and A. A. El Gamal, "On the capacity of computer memory with defects," *IEEE Trans. Inform. Theory*, vol. 29, pp. 731–739, Sept. 1983.

[11] A. Lapidoth and P. Narayan, "Reliable communication under channel uncertainty," *IEEE Trans. Inform. Theory*, vol. 44, pp. 2148–2177, Oct. 1998.