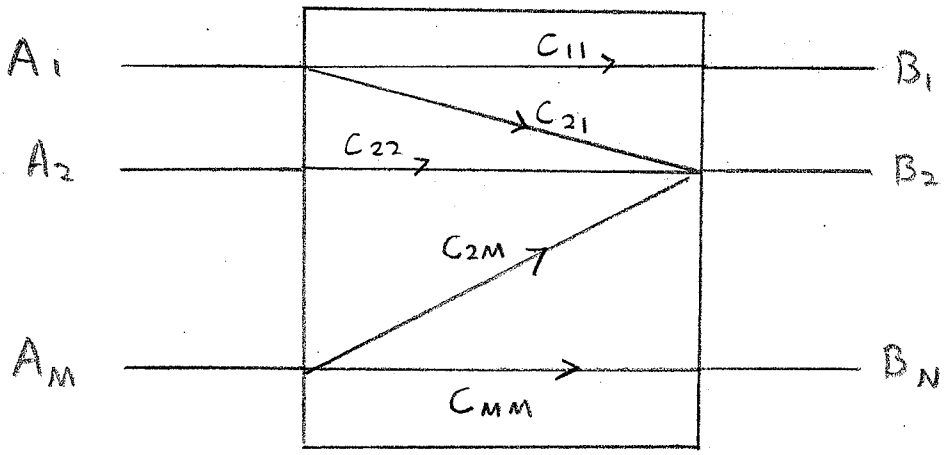
More on Inference

John Wyatt
3/31/04



Assume process is modeled by conditional probabilities

$$c_{ji} = p\left(B_j \mid A_i\right)$$

and the prior probabilities

$$p_i = p\left(A_i\right)$$

are known. We observe an output B_j and use that observation to make a decision as to what value $\hat{A}$ we think must have been the input A_i. We typically formulate in advance a *decision rule*

$$A_{\hat{i}} = d\left(B_j\right).$$

The decision rule is based on the conditional probabilities or transition probabilities c_{ji}, the probabilities $p_i = p(A_i)$ of the various inputs, and the costs of making various errors.

Various decision rules are possible, depending on the consequences of being wrong. In medical diagnosis, for example, where inference in this process model is quite important, the A_i's could be various possible disease states including "no disease" and the B_j's various possible outcomes of a set of laboratory tests. In this field the consequences of asserting "no disease" when a patient has a serious disease that responds well to early treatment are often vastly greater than the consequences of concluding a disease is suspected when the patient is well.

But in the simplest case we weight all errors equally, and this is typically the case for communications channels. When all errors are weighted equally, a good criterion for measuring performance is the overall probability of error

$$p_e = p\left(\hat{A} \neq A\right) = p\left\{\text{set of all } \left(A_i, B_j\right) : A_i \neq d\left(B_j\right)\right\}$$

Given the c_{ji}'s and p_i's, different decision rules will lead to different probabilities of error. It will come as no surprise that p_e is minimized by the simple decision rule

$$\hat{i} = d_{MAP}\left(B_j\right)$$

where $\hat{i}$ is the value of *i* that maximizes $p\left(A_i \mid B_j\right)$. Just pick the *most likely* A_i, given your observation.

This intuitively natural rule has the fancy (but logical) name *maximum a posteriori probability decision rule* (MAP) because

one chooses the A_i with the maximal probability after the data B_j has arrived.

Example

Your friend the trickster has two biased coins, A_1 and A_2. He pulls out each of them with probability ½. The probabilities of heads and tails for the two coins are

	$p(H)$	$p(T)$
A_1	.7	.3
A_2	.4	.6

The trickster lets you observe two tosses and asks you to guess which coin he tossed. Find the MAP decision rule that minimizes your probability of error.

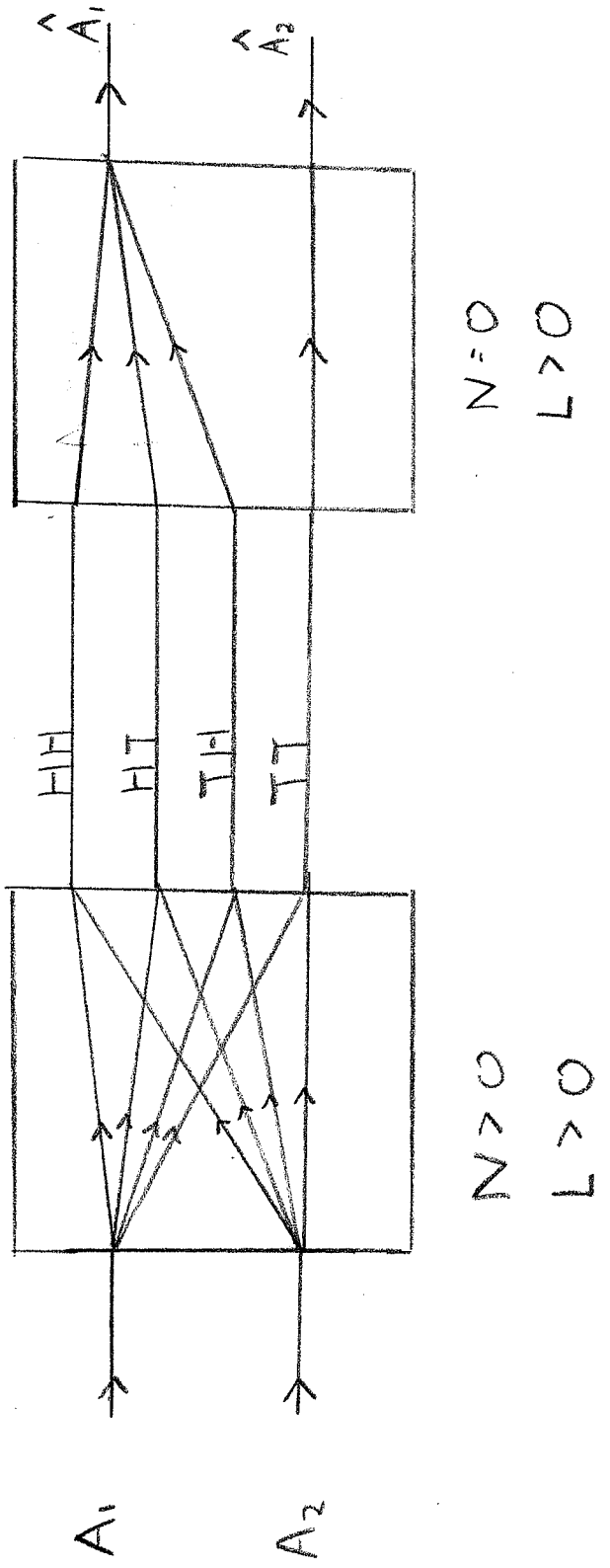
$P(A_1 \mid HH) =$
$P(A_2 \mid HH) =$ $\qquad d_{MAP}(HH) =$

$P(A_1 \mid HT) =$
$P(A_2 \mid HT) =$ $\qquad d_{MAP}(HT) =$

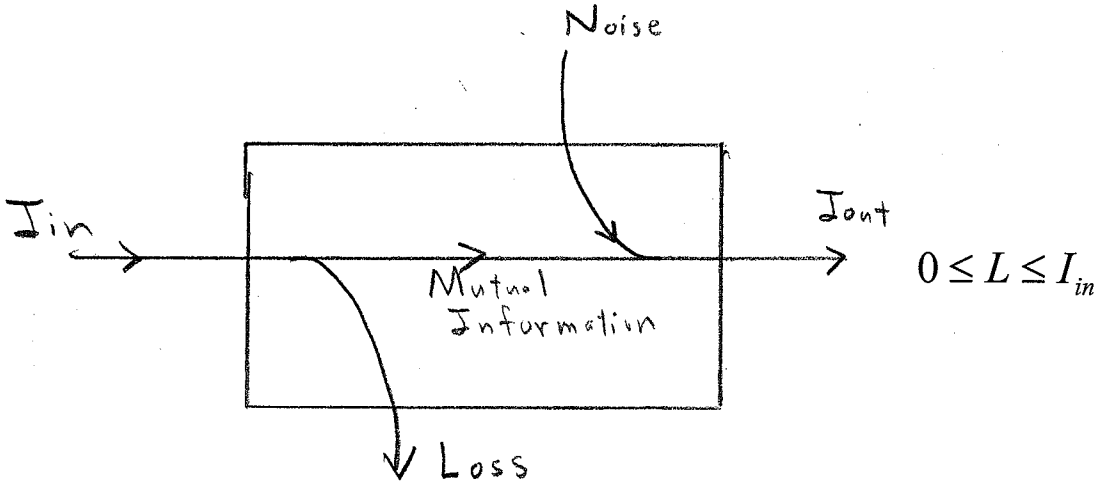
$P(A_1 \mid TH) =$
$P(A_2 \mid TH) =$ $\qquad d_{MAP}(TH) =$

$P(A_1 \mid TT) =$
$P(A_2 \mid TT) =$ $\qquad d_{MAP}(TT) =$

Notice that this entire system can be written as a cascade of individual processes:



What Role Do Information, Noise and Loss Play Here?



If $L = 0$, then $M = I_{in}$ and intuitively it seems we should be able to make perfect decisions with $p_e = 0$.

$$L = \sum_j p(B_j) \sum_i p(A_i \mid B_j) \log \frac{1}{\left(p(A_i \mid B_j)\right)}$$

$$L = 0 \Rightarrow \sum_i p(A_i \mid B_j) \log \frac{1}{p(A_i \mid B_j)} = 0, \text{ all } j \text{ with } p(B_j) > 0 \Rightarrow$$

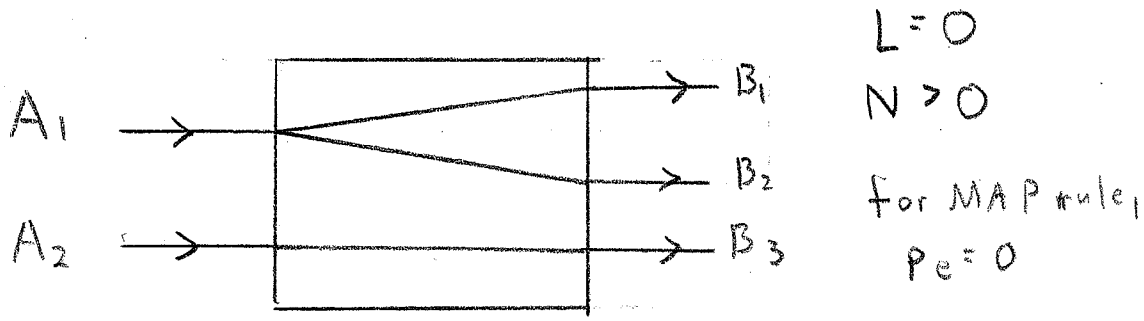
conditional entropy = 0 for each possible $B_j \Rightarrow$

for each B_j,

$$p(A_i \mid B_j) = 1 \text{ for some } i$$

$$p(A_i \mid B_j) = 0 \text{ for all other } i$$

You can exactly determine the input from the output and thus make perfect decisions. With $L = 0$, noise input N doesn't interfere with perfect inference.



At the other extreme, if $L = I_{in}$,

$$I_{in} - L = \sum_i p(A_i) \log \frac{1}{p(A_i)} - \sum_i \sum_j p(A_i, B_j) \log \frac{1}{p(A_i \mid B_j)} =$$

$$\sum_i \sum_j p(A_i, B_j) \frac{1}{p(A_i)} - \sum_i \sum_j p(A_i, B_j) \log \frac{1}{p(A_i \mid B_j)} =$$

$$\sum_i \sum_j p(A_i, B_j) \log \frac{p(A_i \mid B_j)}{p(A_i)} =$$

$$\sum_i \sum_j p(A_i, B_j) \log \frac{p(A_i, B_j)}{p(A_i) p(B_j)} =$$

$$\sum_i \sum_j p(A_i, B_j) \log \frac{1}{p(A_i) p(B_j)} - \sum_i \sum_j p(A_i, B_j) \log \frac{1}{p(A_i, B_j)} = 0.$$

But the Gibbs inequality says this difference is positive unless

$$p(A_i, B_j) = p(A_i)\,p(B_j), \text{ all } i, j$$

i.e.,

$$L = I_{in} \Rightarrow \text{ output } B \text{ and input } A \text{ are independent .}$$

In this case observing the output adds no useful information, and the optimum decision rule ignores the B_j's and always chooses $\hat{A}_i$ to maximize the input probability $p(A_i)$, i.e., it chooses the most likely input without regard to the output that may have occurred.

But what about the intermediate cases where

$$0 < L < I_{in} ?$$

While one intuitively expects p_e to generally rise as L rises, p_e for the optimal (MAP) decision rule cannot be found in the general case from knowledge of L (or even of L, I_{in} and N) alone.

Nonetheless an important relationship between p_e and L was found by Robert Fano, an emeritus professor of EECS at MIT. This bound plays an important role in the proof that transmission at rates beyond channel capacity has to produce errors. A key variable will be the entropy of the error

$$H_E = p_e \log\frac{1}{p_e} + (1-p_e)\log\frac{1}{(1-p_e)}$$

The Fano Bound (Two Input Case, M=2)

For every deterministic decision decision rule $A_{\hat{i}} = d\left(B_j\right)$

$$\boldsymbol{H_E \geq L}$$

Lets examine graphically what this means. Note that in general

$$0 \leq L \leq I_{in}$$

$$0 \leq H_E \leq 1 \text{ bit}$$

and for our case with two inputs

$$I_{in} \leq 1 \text{ bit.}$$

APPENDIX

The Fano Bound

For any process (i.e., memoryless channel) with M inputs A_i and N outputs B_j, and any decision rule $\hat{A}_i = d(B_j)$,

$$H_E + p_e \log(M-1) \geq L$$

Note that this reduces to the version in the lecture for M=2. A simpler but looser bound that is useful for large M can be found by noticing that

$$H_E \leq 1$$

$$\log(M-1) \leq \log M,$$

which gives

$$1 + p_e \log M \geq L,$$

i.e.,

$$p_e \geq \frac{L-1}{\log M}$$

Proof

For simplicity of notation, let $\overline{p_e} = 1 - p_e$.

$$H_E + p_e \log(M-1) = p_e \log\frac{1}{p_e} + \overline{p_e} \log\frac{1}{\overline{p_e}} + p_e \log(M-1) =$$

$$p_e \log \frac{(M-1)}{p_e} + \overline{p_e} \log \frac{1}{\overline{p_e}}$$

$$L = \sum_i \sum_j p\left(A_i, B_j\right) \log \frac{1}{p\left(A_i \mid B_j\right)}$$

We now consider separately the cases of no errors, where $A_i = d\left(B_j\right)$, and of errors $A_i \neq d\left(B_j\right)$ for the decision rule *d*, noting that

$$p_e = \sum_{\substack{i \; j \\ A_i \neq d\left(B_j\right)}} p\left(A_i, B_j\right)$$

$$\overline{p_e} = \sum_{\substack{i \; j \\ A_i \neq d\left(B_j\right)}} p\left(A_i, B_j\right) = \sum_j p\left(d\left(B_j\right), B_j\right)$$

$$H_E + p_e \log\left(M-1\right) =$$

$$\sum_{\substack{i \; j \\ A_i \neq d\left(B_j\right)}} p\left(A_i, B_j\right) \log \frac{M-1}{p_e} + \sum_j p\left(d\left(B_j\right), B_j\right) \log \frac{1}{\overline{p_e}}$$

Similarly,

$$L = \sum_{\substack{i \; j \\ A_i \neq d\left(B_j\right)}} p\left(A_i, B_j\right) \log \frac{1}{p\left(A_i \mid B_j\right)} + \sum_j p\left(d\left(B_j\right), B_j\right) \log \frac{1}{p\left(d\left(B_j\right) \mid B_j\right)}$$

Subtracting the first expression from the second,

$$L-\left(H_E+p_e\log(M-1)\right)=$$

$$\sum_{\substack{i\ j\\ A_i\neq d(B_j)}}\sum p\left(A_i,B_j\right)\log\frac{p_e}{(M-1)p\left(A_i\mid B_j\right)}+\sum_j p\left(d\left(B_j\right),B_j\right)\log\frac{\overline{p_e}}{p\left(d\left(B_j\right)\mid B_j\right)}\leq$$

(since $\ln x\leq[x-1]$ and therefore $\log_2 x\leq\frac{x-1}{\ln 2}$)

$$\frac{1}{\ln 2}\left\{\sum_{\substack{i\ j\\ A_i\neq d(B_j)}}\sum p\left(A_i,B_j\right)\left[\frac{p_e}{(M-1)p\left(A_i\mid B_j\right)}-1\right]+\sum_j p\left(d\left(B_j\right),B_j\right)\left[\frac{\overline{p_e}}{p\left(d\left(B_j\right)\mid B_j\right)}\right]\right\}=$$

$$\frac{1}{\ln 2}\left\{\sum_{\substack{i\ j\\ A_i\neq d(B_j)}}\sum\left(\frac{p_e p\left(B_j\right)}{(M-1)}-p\left(A_i,B_j\right)\right)+\sum_j\left(\overline{p_e}p\left(B_j\right)-p\left(d\left(B_j\right),B_j\right)\right)\right\}=$$

since there are $M-1$ values of A_i where $A_i\neq d\left(B_j\right)$

$$\frac{1}{\ln 2}\left\{\left((M-1)\frac{p_e}{(M-1)}-p_e\right)+\left(\overline{p_e}-\overline{p_e}\right)\right\}\leq 0,$$

i.e.,

$$H_E+p_e\log(M-1)\geq L.$$

■

Note that the inequality becomes an equality only when $\ln x = (x-1)$ in both sums, i.e., only in the symmetric case where for all i and j,

$$p(A_i \mid B_j) = \frac{p_e}{M-1}, \quad A_i \neq d(B_j)$$

$$p(A_i \mid B_j) = \overline{p_e}, \quad A_i \neq d(B_j).$$

Of course the second equation follows from the first, since the conditional probabilities must sum to 1.