Exam, 2010-2011 Solutions

Q1. Constructing a balanced sequence containing three kinds of stimuli

Here we design a balanced cyclic sequence for three kinds of stimuli (labeled $\{0,1,2\}$*), in which every three-element sequence (except for the sequence* $\{0,0,0\}$*)occurs exactly once. We do this by extending the finite field* $\mathbb{Z}_3$ *to make a field of size 27,* $GF(3,3)$.

Analogous to $\mathbb{Z}_2$, $\mathbb{Z}_3$ *is the field containing* $\{0,1,2\}$*, with addition and multiplication defined (mod 3). The polynomial* $x^3+2x+1=0$ *has no solutions in* $\mathbb{Z}_3$*, so we add a formal quantity* ξ *to* $\mathbb{Z}_3$*, and assert that* $\xi^3+2\xi+1=0$*, and that* ξ *satisfies the associative, commutative, and distributive laws for addition and multiplication with itself and with the elements of* $\mathbb{Z}_3$.

A. Using $\xi^3+2\xi+1=0$, *express* ξ^r *in terms of* ξ^0, ξ^1,*and* ξ^2, *for* $r=1,...,26$. *Note that the sequence of coefficients of* ξ^0*, considered cyclically, has the property that ever sequence of three labels (except for the sequence* $\{0,0,0\}$*)occurs exactly once.*

Field operations are "mod 3", i.e,. we can replace -1 by +1, and 3 by 0, 4 by 1, etc.. So, for example, $\xi^3+2\xi+1=0$ implies $\xi^3=-2\xi-1=\xi+2$, and $2\xi^3=2\xi+4=2\xi+1$.
So, using the field properties and the equation satisfied by ξ, we find for example that
$\xi^4=\xi\cdot\xi^3=\xi(\xi+2)=\xi^2+2\xi$; $\xi^5=\xi\cdot\xi^4=\xi(\xi^2+2\xi)=\xi^3+2\xi^2=2\xi^2+\xi+2$.

Working similarly, the table of coefficients is:

	ξ^2	ξ^1	ξ^0
$\xi^0=$	0	0	1
$\xi^1=$	0	1	0
$\xi^2=$	1	0	0
$\xi^3=$	0	1	2
$\xi^4=$	1	2	0
$\xi^5=$	2	1	2
$\xi^6=$	1	1	1
$\xi^7=$	1	2	2
$\xi^8=$	2	0	2
$\xi^9=$	0	1	1
$\xi^{10}=$	1	1	0
$\xi^{11}=$	1	1	2
$\xi^{12}=$	1	0	2
$\xi^{13}=$	0	0	2
$\xi^{14}=$	0	2	0
$\xi^{15}=$	2	0	0
$\xi^{16}=$	0	2	1
$\xi^{17}=$	2	1	0
$\xi^{18}=$	1	2	1
$\xi^{19}=$	2	2	2
$\xi^{20}=$	2	1	1
$\xi^{21}=$	1	0	1
$\xi^{22}=$	0	2	2
$\xi^{23}=$	2	2	0
$\xi^{24}=$	2	2	1
$\xi^{25}=$	2	0	1
$\xi^{26}=$	0	0	1

B. Consider the sequence of the coefficients of ξ^0 that occur in the expansion of ξ^k, for $k = 0,...,25$. Note that the second half of the sequence can be obtained from the first half of the sequence by exchanging 1 and 2. Why must this be true?

As the above table shows, $\xi^{13} = 2$. So $\xi^{k+13} = 2\xi^k$, so the coefficient of ξ^0 in ξ^{k+13} is twice the coefficient of ξ^0 in ξ^k, which means that 1 is replaced by $2 \cdot 1 = 2$, and 2 is replaced by $2 \cdot 2 = 1$.

C. What is the size of the multiplicative group generated by ξ ? Show that for all elements g in this group, $g^{26} = 1$.

The group has 26 elements, $\xi^0 = 1, \xi, \xi^2, ..., \xi^{25}$, since the first time that the identity recurs in the above table is $\xi^{26} = 1$. (This in turn means that the above table contains all the nonzero elements of the field $GF(3,3)$, since this field has $3^3 = 27$ elements.

Since the order of every element of a finite group must divide the size of the group, the only possible orders are the factors of 26, namely, 1, 2, 13, and 26. To see that this implies that $g^{26} = 1$ for all group elements *g*, note that $g^2 = 1$, then $g^{26} = \left(g^2\right)^{13} = 1$, and if $g^{13} = 1$, then $g^{26} = \left(g^{13}\right)^2 = 1$ So in all cases, $g^{26} = 1$.

D. Which of the following maps are automorphisms of the multiplicative group (i.e., are 1-1 maps that preserve multiplication)? $U(g) = g^2$, $X(g) = g^3$, $Y(g) = g^5$

$U(g) = g^2$ cannot be an automorphism, since $\xi^{13} = 2$ but $U(\xi^{13}) = \xi^{26} = 1$,so U maps two different elements onto 1.

$X(g) = g^3$ is an automorphism; its inverse is $X \circ X$, since $X\left(X\left(X\left(g\right)\right)\right) = \left(\left(g^3\right)^3\right)^3 = g^{27} = g^{26+1} = g$

$Y(g) = g^5$ is an automorphism; its inverse is $Y \circ Y \circ Y$, since
$Y\left(Y\left(Y\left(Y\left(g\right)\right)\right)\right) = \left(\left(\left(g^5\right)^5\right)^5\right)^5 = g^{\left(5^4\right)} = g^{625} = g^{26 \cdot 24+1} = g$.

E. Which of the above maps are automorphisms of the field $GF(3,3)$ (i.e., are 1-1 maps that preserve multiplication and addition)?

$U(g) = g^2$ cannot be a field automorphism, since (as in D) it is not even an automorphism of the multiplicative group.

$X(g) = g^3$ is a field automorphism, since
$X(g+h) = (g+h)^3 = g^3 + 3g^2h + 3gh^2 + h^3 = g^3 + h^3 = X(g) + X(h)$.

$Y(g) = g^5$ is not a field automorphism. One way to see this is to try some random choices to see if $Y(g+h) = Y(g) + Y(h)$. A convenient way to do this is $g = 1$, $h = \xi$. From the above table, $1 + \xi = \xi^9$, so $Y(1+\xi) = Y(\xi^9) = \xi^{9\cdot5} = \xi^{45} = \xi^{26+19} = \xi^{19} = 2\xi^2 + 2\xi + 2$.
On the other hand, $Y(1) + Y(\xi) = 1 + \xi^5 = 1 + (2\xi^2 + \xi + 2) = 2\xi^2 + \xi$. So *Y* does not preserve addition.

Q2. Decomposing group representations

Recall that if we have a group G with a representation U_1 in V_1 and another representation U_2 in V_2, we can define a group representation $U_1 \otimes U_2$ on $V_1 \otimes V_2$ by $\left(U_{1,g} \otimes U_{2,g}\right)(v_1 \otimes v_2) = U_{1,g}(v_1) \otimes U_{2,g}(v_2)$. Here we will show that if V_1 and V_2 are the same (i.e., if $V_1 = V_2 = V$) and U_1 and U_2 are the same ($U_1 = U_2 = U$),we can decompose $U \otimes U$ into a symmetric and an antisymmetric component, and determine their characters. We will do this by decomposing $V \otimes V$ into $sym(V \otimes V)$ and $anti(V \otimes V)$, and seeing how each transformation in $U \otimes U$ acts in these components.

A. Recall that a linear transformation A on V acts on a typical element $sym(x \otimes y) = \frac{1}{2}\left(x \otimes y + y \otimes x\right)$ of $sym(V \otimes V)$ by $A\left(sym(x \otimes y)\right) = \frac{1}{2}\left(A(x) \otimes A(y) + A(y) \otimes A(x)\right) = sym(Ax \otimes Ay)$, and similarly for the action of A in $anti(V \otimes V)$. Given that A on V has distinct eigenvalues $\lambda_1, \ldots, \lambda_m$ and eigenvectors $v_1, \ldots, v_m$ (and that V has dimension m), find the eigenvalues and eigenvectors for the action of A in $sym(V \otimes V)$ and $anti(V \otimes V)$ and their traces, denoted $\mathrm{tr}(sym(A \otimes A))$ and $\mathrm{tr}(anti(A \otimes A))$.

A convenient basis for $sym(V \otimes V)$ is $sym(v_j \otimes v_k)$, for any pair *j*, *k* in $\{1, ..., m\}$ with $j \le k$. These are eigenvectors of *A*, since

$$A\left(sym(v_j \otimes v_k)\right) = \frac{A(v_j) \otimes A(v_j) + A(v_k) \otimes A(v_j)}{2} = \frac{\lambda_j\lambda_k\left(v_j \otimes v_k\right) + \lambda_k\lambda_j\left(v_k \otimes v_j\right)}{2} = \lambda_j\lambda_k sym(v_j \otimes v_k).$$

This calculation also shows that the eigenvalues are the products $\lambda_j\lambda_k$. We don't consider pairs *j*, *k* for which $k > j$, since these have already been counted: $sym(v_k \otimes v_j) = sym(v_j \otimes v_k)$.

A similar calculation holds for the action of *A* in $anti(V \otimes V)$, except that $anti(v_j \otimes v_k)$ is only a basis element (and an eigenvector) if $j \ne k$, since $anti(v_j \otimes v_j) = \frac{1}{2}\left(v_j \otimes v_j - v_j \otimes v_j\right) = 0$. We also don't consider pairs *j*, *k* for which $k > j$, since these have already been counted: $anti(v_k \otimes v_j) = -anti(v_j \otimes v_k)$.

Thus, the eigenvalues for the action of *A* in $sym(V \otimes V)$ are all pairwise products $\lambda_j\lambda_k$ with $j \le k$ (including λ_j^2)– a total of $\frac{m(m-1)}{2} + m = \frac{m(m+1)}{2}$, corresponding to the dimension of $sym(V \otimes V)$.

So $\mathrm{tr}(sym(A \otimes A)) = \sum_{1 \le j \le k}^{m} \lambda_j\lambda_k$.

The eigenvalues for the action of A in $anti(V \otimes V)$ are all pairwise products $\lambda_j \lambda_k$ with $j < k$ (excluding λ_j^2) – a total of $\frac{m(m-1)}{2}$ quantities, corresponding to the dimension of $anti(V \otimes V)$.

So $\text{tr}(anti(A \otimes A)) = \sum_{1 \le j < k}^{m} \lambda_j \lambda_k$.

B. Our next step is to express $\text{tr}(sym(A \otimes A))$ *(i.e., the trace of A acting in* $sym(V \otimes V)$*), and* $\text{tr}(anti(A \otimes A))$ *(i.e., the trace of A acting in* $anti(V \otimes V)$*), in terms of easier quantities. Given the same setup as above, find the trace of* $A \otimes A$ *(acting in*$V \otimes V$*) and the trace of* A^2 *(acting in V), and relate this to* $\text{tr}(sym(A \otimes A))$ *and* $\text{tr}(anti(A \otimes A))$.

To calculate $\text{tr}(A \otimes A)$ (the sum of its eigenvalues), we use the basis $v_j \otimes v_k$ (all *j* and *k* from 1 to *m*) for $V \otimes V$. The eigenvalue corresponding to $v_j \otimes v_k$ is $\lambda_j \lambda_k$, as $(A \otimes A)(v_j \otimes v_k) = (Av_j \otimes Av_k) = (\lambda_j v_j \otimes \lambda_k v_k) = \lambda_j \lambda_k (v_j \otimes v_k)$. So $\text{tr}(A \otimes A) = \sum_{j,k=1}^{m} \lambda_j \lambda_k = \left(\sum_{j=1}^{m} \lambda_j \right)^2 = (\text{tr}(A))^2$. This can be rephrased as

$$(\text{tr}(A))^2 = \sum_{j,k=1}^{m} \lambda_j \lambda_k = 2 \sum_{1 \le j < k}^{m} \lambda_j \lambda_k + \left(\sum_{j=1}^{m} \lambda_j \right)^2 .$$

To calculate $\text{tr}(A^2)$, we use the basis v_j for *V*. The eigenvalue corresponding to v_j is $\lambda_j{}^2$, as $A^2 v_j = A(Av_j) = A(\lambda_j v_j) = \lambda_j Av_j = \lambda_j{}^2 v_j$.

To express $\text{tr}(sym(A \otimes A))$ in terms of $\text{tr}(A \otimes A) = (\text{tr}(A))^2$ and $\text{tr}(A^2)$:

$$\text{tr}(sym(A \otimes A)) = \sum_{1 \le j \le k}^{m} \lambda_j \lambda_k = \sum_{1 \le j < k}^{m} \lambda_j \lambda_k + \sum_{j=1}^{m} \lambda_j{}^2 .$$

From $(\text{tr}(A))^2 = \sum_{j,k=1}^{m} \lambda_j \lambda_k = 2 \sum_{1 \le j < k}^{m} \lambda_j \lambda_k + \left(\sum_{j=1}^{m} \lambda_j \right)^2$, we have $\sum_{1 \le j < k}^{m} \lambda_j \lambda_k = \frac{1}{2} \left((\text{tr}(A))^2 - \left(\sum_{j=1}^{m} \lambda_j \right)^2 \right)$.

Therefore,

$$\text{tr}(sym(A \otimes A)) = \sum_{1 \le j < k}^{m} \lambda_j \lambda_k + \sum_{j=1}^{m} \lambda_j{}^2 = \frac{1}{2} \left((\text{tr}(A))^2 + \left(\sum_{j=1}^{m} \lambda_j \right)^2 \right) = \frac{1}{2} \left((\text{tr}(A))^2 + \text{tr}(A^2) \right),$$

and

$$\text{tr}(anti(A \otimes A)) = \sum_{1 \le j < k}^{m} \lambda_j \lambda_k = \frac{1}{2} \left((\text{tr}(A))^2 - \left(\sum_{j=1}^{m} \lambda_j \right)^2 \right) = \frac{1}{2} \left((\text{tr}(A))^2 - \text{tr}(A^2) \right).$$

C. Finally, recalling that the character is defined by $\chi_L(g) = \text{tr}(L_g)$*, express* $\chi_{sym(U \otimes U)}$ *and* $\chi_{anti(U \otimes U)}$ *in terms of* χ_U.

$$\chi_{sym(U\otimes U)}(g)=\mathrm{tr}\left(sym(U_g\otimes U_g)\right)=\frac{1}{2}\left(\left(\mathrm{tr}\left(U_g\right)\right)^2+\mathrm{tr}\left(\left(U_g\right)^2\right)\right)$$
$$=\frac{1}{2}\left(\left(\mathrm{tr}\left(U_g\right)\right)^2+\mathrm{tr}\left(U_{g^2}\right)\right)=\frac{1}{2}\left(\left(\chi_U(g)\right)^2+\chi_U(g^2)\right)$$

Similarly,

$$\chi_{anti(U\otimes U)}(g)=\mathrm{tr}\left(anti(U_g\otimes U_g)\right)=\frac{1}{2}\left(\left(\mathrm{tr}\left(U_g\right)\right)^2-\mathrm{tr}\left(\left(U_g\right)^2\right)\right)$$
$$=\frac{1}{2}\left(\left(\mathrm{tr}\left(U_g\right)\right)^2-\mathrm{tr}\left(U_{g^2}\right)\right)=\frac{1}{2}\left(\left(\chi_U(g)\right)^2-\chi_U(g^2)\right)$$

Q3: Mutual information: synergy, redundancy, etc.

Consider a stimulus S that is equally likely to have one of several values, $\{0,1,\ldots,N-1\}$, *and two neurons that are influenced by it. We only concern ourselves with snapshots, so the responses* R_1 *and* R_2 *can be considered to be binary* $\{0,1\}$. *In this scenario, we can calculate the information conveyed by each neuron alone:*
$I_1=H(S)+H(R_1)-H(S,R_1)$ *and* $I_2=H(S)+H(R_2)-H(S,R_2)$,
as well as the information conveyed by the entire "population",
$I_{1+2}=H(S)+H(R_1,R_2)-H(S,R_1,R_2)$.
Note that $H(S,R_1,R_2)$ *is the entropy of the table of* $4N$ *entries, listing the probability* $p(S,R_1,R_2)$ *of that each value of S is associated with one of the four outputs patterns that* R_1 *and* R_2 *can produce (i.e.,* $\{R_1=0,R_2=0\}$, $\{R_1=1,R_2=0\}$, $\{R_1=0,R_2=1\}$, $\{R_1=1,R_2=1\}$. *This table also determines the joint probabilities of S and each* R_i, *since* $p(S,R_1)=p(S,R_1,R_2=0)+p(S,R_1,R_2=1)$ *and similarly* $p(S,R_2)=p(S,R_1=0,R_2)+p(S,R_1=1,R_2)$.

Find an example of $p(S,R_1,R_2)$ *that illustrates each of the behaviors below, or, alternatively, show that the behavior is impossible. It may be work in terms of the stimulus probabilities* $p(S)$ *and the conditional probabilities* $p(R_1,R_2\,|\,S)$ *(the probability that particular values of* R_1 *and* R_2 *occur, given a value of S), noting that* $p(S,R_1,R_2)=p(R_1,R_2\,|\,S)p(S)$.

A. $I_1>0, I_2>0$, $I_{1+2}=I_1+I_2$ *(independent channels)*
Take $N=4$ and consider *S* to be a two-bit binary number ($0\to 00$, $1\to 01$, $2\to 10$, $3\to 11$), each with equal probability ($p(S)=1/4$). Have each neuron care only about one of the stimulus bits and ignore the other. Since the bits (the channels) are independent, the information in each channel should add.

To work this out as a simple example: we can have R_1 code the first bit with perfect reliability, and R_2 code the second bit with perfect reliability. So we expect that $I_1=I_2=1$, and $I_{1+2}=2$. To verify: if $S=s_2s_1$ (as bits), then we can take $p(R_1,R_2\,|\,S)=1$ if $R_1=s_1$ and $R_2=s_2$, and zero otherwise.

To calculate I_{1+2}: Express the stimulus-response relationship as a table

$p(R_1, R_2 \mid S)$	$R_2 = 0$ $R_1 = 0$	$R_2 = 0$ $R_1 = 1$	$R_2 = 1$ $R_1 = 0$	$R_2 = 1$ $R_1 = 1$
$S = 00$	1	0	0	0
$S = 01$	0	1	0	0
$S = 10$	0	0	1	0
$S = 11$	0	0	0	1

.

Since there are four equally likely inputs ($p(S) = 1/4$), the input entropy is $\log_2 4 = 2$. Since there are also four equally likely outputs, the output entropy is also $\log_2 4 = 2$. There are also 4 equally likely table entries, so the table entropy is $\log_2 4 = 2$. The information $I_{1+2} = 2 + 2 - 2 = 2$.

To calculate I_1 : Reduce the above table, by summing over values of R_2 :

$p(R_1 \mid S)$	$R_1 = 0$	$R_1 = 1$
$S = 00$	1	0
$S = 01$	0	1
$S = 10$	1	0
$S = 11$	0	1

.

The input entropy is $\log_2 4 = 2$. The output entropy is $\log_2 2 = 1$. There are 4 equally likely table entries, so the table entropy is $\log_2 4 = 2$. The information $I_1 = 2 + 1 - 2 = 1$.

The above is spelling things out in much more detail than necessary, and are all straightforward consequences of a basic property of information, namely, that for independent channels, information adds. The idea of course generalizes to neurons that are not fully reliable (i.e., $p(R_i = s_i) = p_i$); in this case, the table for I_{1+2} is

$p(R_1, R_2 \mid S)$	$R_2 = 0$ $R_1 = 0$	$R_2 = 0$ $R_1 = 1$	$R_2 = 1$ $R_1 = 0$	$R_2 = 1$ $R_1 = 1$
$S = 00$	$p_1 p_2$	$(1-p_1)p_2$	$p_1(1-p_2)$	$(1-p_1)(1-p_2)$
$S = 01$	$(1-p_1)p_2$	$p_1 p_2$	$(1-p_1)(1-p_2)$	$p_1(1-p_2)$
$S = 10$	$p_1(1-p_2)$	$(1-p_1)(1-p_2)$	$p_1 p_2$	$(1-p_1)p_2$
$S = 11$	$(1-p_1)(1-p_2)$	$p_1(1-p_2)$	$(1-p_1)p_2$	$p_1 p_2$

,

and the table for I_1 is

$p(R_1 \mid S)$	$R_1 = 0$	$R_1 = 1$
$S = 00$	p_1	$1-p_1$
$S = 01$	$1-p_1$	p_1
$S = 10$	p_1	$1-p_1$
$S = 11$	$1-p_1$	p_1

.

Here, $I_i = 1 + p_i \log_2 p_i + (1-p_i)\log_2(1-p_i)$, which is nonzero provided $p_i \neq 1/2$.

B. $I_1 > 0, I_2 > 0$, $I_{1+2} = \max(I_1, I_2)$ *(completely redundant channels)*

We can ensure that $I_{1+2} = I_1 = I_2$ by making R_1 and R_2 identical. To spell this out: Let *S* be binary, and take

$p(R_1, R_2 \mid S)$	$R_2=0$ $R_1=0$	$R_2=0$ $R_1=1$	$R_2=1$ $R_1=0$	$R_2=1$ $R_1=1$
$S=0$	p	0	0	$1-p$
$S=1$	$1-p$	0	0	p

.

That is, the only response configurations with nonzero probability are those for which $R_1 = R_2$, and, R_1 and R_2 have reliability $p(R = s_i) = p$.

The table for I_1 is

$p(R_1 \mid S)$	$R_1=0$	$R_1=1$
$S=0$	p	$1-p$
$S=1$	$1-p$	p

.

For these, $I_{1+2} = I_1 = I_2 = 1 + p\log_2 p + (1-p)\log_2(1-p)$.

C. $I_1 > 0, I_2 > 0$, $\max(I_1, I_2) < I_{1+2} < I_1 + I_2$ *(partially redundant channels)*

We can create partial redundancy by starting from A, and eliminating the any single input (for example, $S = 11$), so that now, if either neuron indicates a "1" input, then the other input must be 0 and the other neuron's activity is uninformative. So $N = 3$;
$p(S=00) = p(S=10) = p(S=01) = 1/3$.

For I_{1+2}, the table is

$p(R_1, R_2 \mid S)$	$R_2=0$ $R_1=0$	$R_2=0$ $R_1=1$	$R_2=1$ $R_1=0$	$R_2=1$ $R_1=1$
$S=00$	1	0	0	0
$S=01$	0	1	0	0
$S=10$	0	0	1	0

.

The input entropy, the output entropy, and the table entropy are all $\log_2 3$ (three equally likely possibilities for all), so $I_{1+2} = \log_2 3 + \log_2 3 - \log_2 3 = \log_2 3 \approx 1.585$.

For I_1,

$p(R_1 \mid S)$	$R_1=0$	$R_1=1$
$S=00$	1	0
$S=01$	0	1
$S=10$	1	0

.

The input entropy and the table entropy are still $\log_2 3$, but the output entropy is $h_3 = -2/3\log_2(2/3) - 1/3\log_2(1/3) \approx 0.918$. So $I_1 = \log_2 3 + h_3 - \log_2 3 = h_3$, and indeed, $h_3 = \max(I_1, I_2) < I_{1+2} < I_1 + I_2 = 2h_3$.

This of course also extends to less-than-reliable neurons, as above.

D. $I_{1+2} > I_1 + I_2$ *("synergistic" channels)*

We can set up a scenario in which neither neuron alone has any information about the stimulus, but the pair does. To do this, we caricature the idea that when $S=0$, the neurons are asynchronous, but when $S=1$, they are synchronized.

For I_{1+2}, the table is

$p(R_1,R_2 \mid S)$	$R_2=0$ $R_1=0$	$R_2=0$ $R_1=1$	$R_2=1$ $R_1=0$	$R_2=1$ $R_1=1$
$S=0$	0	1/2	1/2	0
$S=1$	1/2	0	0	1/2

.

The input entropy is 1, the output entropy is 2 (all possible outputs have probability 1/4), and the table entropy is 2. So $I_{1+2}=1+2-2=1$.

But for each neuron alone, the information is 0:

$p(R_i \mid S)$	$R_i=0$	$R_i=1$
$S=0$	1/2	1/2
$S=1$	1/2	1/2

,

since input entropy is 1, output entropy is 1, and table entropy is 2.

E. $I_{1+2} < \max(I_1, I_2)$ *("occluding" channels)*

This is impossible, because of the Data Processing Inequality. In detail: the response variable R_1 can be derived from the response variable (R_1, R_2) by ignoring R_2. Therefore, I_{1+2} cannot be less than I_1. Similarly, I_{1+2} cannot be less than than I_2. So $I_{1+2} < \max(I_1, I_2)$ is impossible.

F. $I_{1+2} < \min(I_1, I_2)$ *("strongly occluding" channels)*

This is impossible; it is a special case of E.

Q4. Mutual information with additive noise

Here we will determine the mutual information between a Gaussian stimulus s and a response r that are related by $r = gs + a$*, where g is a constant (the "gain), and a is an additive noise, uncorrelated with the input. For definiteness, we assume that the stimulus s has variance* V_S *and the noise term is drawn from a Gaussian with variance* V_A*, and that both have mean 0.*

A. For a Gaussian distribution of variance V, $p_V(x) = \frac{1}{\sqrt{2\pi V}} e^{-x^2/2V}$*, calculate the differential entropy.*

The differential entropy is $H_V = -\int_{-\infty}^{\infty} p_V(x) \log_2 p_V(x) dx$. Via standard steps:

$$H_V = -\int_{-\infty}^{\infty} p_V(x) \log_2 p_V(x) dx = -\frac{1}{\log(2)} \int_{-\infty}^{\infty} p_V(x) \ln p_V(x) dx = -\frac{1}{\log(2)} \int_{-\infty}^{\infty} p_V(x) \left(\ln \frac{1}{\sqrt{2\pi V}} - \frac{x^2}{2V} \right) dx .$$

The first term in parentheses is just a constant; its contribution to the integral can be calculated from $\int_{-\infty}^{\infty} p_V(x)dx = 1$ (since p_V is a probability distribution). The second term is proportional to x^2; its contribution can be calculated from $\int_{-\infty}^{\infty} x^2 p_V(x)dx = 1$ (since p_V has variance *V*). So,

$$H_V = -\int_{-\infty}^{\infty} p_V(x)\log_2 p_V(x)dx = -\frac{1}{\log 2}\left(\ln\frac{1}{\sqrt{2\pi V}} - \frac{1}{2}\right) = \frac{1}{2\log 2}\left(\ln(2\pi V)+1\right) = \frac{1}{2\log 2}\left(\ln V + \ln 2\pi e\right).$$

B. How is the output distributed (i.e., what is $p(r)$)? What is its differential entropy?

Since $r = gs + a$, it is a sum of two independent Gaussian components, both of mean 0. Therefore *r* is distributed as a Gaussian, of mean 0. Its variance is the sum of the variances of the two terms:

$$V_R = \left\langle r^2 \right\rangle = \left\langle (gs+a)^2 \right\rangle = \left\langle (gs)^2 + 2gsa + a^2 \right\rangle = g^2\left\langle s^2\right\rangle + 2g\left\langle sa\right\rangle + \left\langle a^2\right\rangle = g^2 V_S + V_A,$$

where $\langle sa\rangle = 0$ because we have hypothesized that they are independent.

Since *r* is distributed as a Gaussian with the above variance, its differential entropy is

$$H_R = \frac{1}{2\log 2}\left(\ln\left(g^2V_S + V_A\right) + \ln 2\pi e\right).$$

C. What is the distribution of the output, conditional on a particular value of the input, say, $s = s_0$? I.e., what is $p(r \mid s_0)$? What is its differential entropy?

With $s = s_0$ given, $r = gs_0 + a$, so $p(r \mid s_0)$ is a Gaussian with mean gs_0 and variance V_A. The mean does not affect the differential entropy, as it just translates the distribution. So

$$H_{R|s_0} = \frac{1}{2\log 2}\left(\ln V_A + \ln 2\pi e\right).$$

D. Calculate the mutual information between the input and the output.

We do this by comparing the unconditional entropy of the output, H_R, with the average condition entropy, $H_{R|s}$, i.e., $I(S,R) = H_R - \int_{-\infty}^{\infty} p(s)H_{R|s}ds$. Since $H_{R|s}$ is constant (part C), this reduces to $I(S,R) = H_R - H_{R|s_0}$, for any s_0. Thus,

$$I(S,R) = \frac{1}{2\log 2}\left(\ln\left(g^2V_S + V_A\right) + \ln 2\pi e\right) - \frac{1}{2\log 2}\left(\ln\left(V_A\right) + \ln 2\pi e\right) = \frac{1}{2\log 2}\ln\left(\frac{g^2V_S + V_A}{V_A}\right)$$

E. Calculate the "signal to-noise" ratio, namely, the ratio of the variance of the signal term, gs, to the variance of the noise term, a. Relate this to the mutual information.

$SNR = \dfrac{\left\langle (gs)^2\right\rangle}{\left\langle a^2\right\rangle} = g^2\dfrac{V_S}{V_A}$. So $I(S,R) = \dfrac{1}{2\log 2}\left(SNR + 1\right)$, a classic result.

F. Calculate the correlation coefficient between the input and the output, $C = \frac{\left\langle (s - \langle s \rangle)(r - \langle r \rangle) \right\rangle}{\sqrt{\left\langle (s - \langle s \rangle)^2 \right\rangle \left\langle (r - \langle r \rangle)^2 \right\rangle}}$. *Relate this to the mutual information.*

For the numerator of *C*: $\left\langle (s - \langle s \rangle)(r - \langle r \rangle) \right\rangle = \langle sr \rangle = \langle s(gs + a) \rangle = \left\langle gs^2 + sa \right\rangle = gV_S$.

So $C = \frac{gV_S}{\sqrt{V_S}\sqrt{V_R}} = \frac{gV_S}{\sqrt{V_S}\sqrt{g^2V_S + V_A}}$, from which $\frac{1}{C^2} = \frac{g^2{V_S}^2 + V_S V_A}{g^2{V_S}^2}$ and $\frac{1}{C^2} - 1 = \frac{V_A}{g^2V_S}$ and

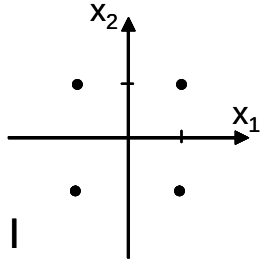
$$I(S,R) = \frac{1}{2\log 2}\ln\left(\frac{g^2V_S + V_A}{V_A}\right) = \frac{1}{2\log 2}\ln\left(1 + \frac{g^2V_S}{V_A}\right)$$

$$= \frac{1}{2\log 2}\ln\left(1 + \frac{1}{\frac{1}{C^2} - 1}\right) = \frac{1}{2\log 2}\ln\left(1 + \frac{C^2}{1 - C^2}\right) = \frac{1}{2\log 2}\ln\left(\frac{1}{1 - C^2}\right) = -\frac{1}{\log 2}\ln\sqrt{1 - C^2},$$

another classic result.

Q5. Toy examples of ICA

A. Take four data points, located at $(\pm 1, \pm 1)$, *(see diagram I) and consider its projection onto a unit vector* $v_\theta = (\cos\theta, \sin\theta)$, *to form a distribution of 4 points.*



I

How does the variance of this distribution depend on θ *? For which projections is it maximized (equivalently, what direction(s) would be selected by PCA?)*

The projected distribution consists of the four points at positions $x_i = \pm\cos\theta \pm \sin\theta$ along v_θ. Their mean is zero. So the variance is

$$V = \left\langle x^2 \right\rangle = \frac{1}{4}\left((\cos\theta + \sin\theta)^2 + (\cos\theta - \sin\theta)^2 + (-\cos\theta + \sin\theta)^2 + (-\cos\theta - \sin\theta)^2\right),$$

$$= \cos^2\theta + \sin^2\theta = 1$$

which is independent of θ. From the point of view of PCA, no directions are special (i.e., each accounts for the same amount of variance).

B. As in A, but determine how the kurtosis of the projected distribution depends on θ. *Recall, kurtosis is defined by* $\kappa = \frac{\left\langle \left(x - \langle x \rangle\right)^4 \right\rangle}{V^2} - 3$, *where* $V = \left\langle \left(x - \langle x \rangle\right)^2 \right\rangle$ *is the variance. For which directions is kurtosis largest?*

Since $\langle x \rangle = 0$ and $V = 1$,

$$\kappa = \langle x^4 \rangle - 3 = \frac{1}{4}\left((\cos\theta + \sin\theta)^4 + (\cos\theta - \sin\theta)^4 + (-\cos\theta + \sin\theta)^4 + (-\cos\theta - \sin\theta)^4 \right) - 3$$

$$= \cos^4\theta + 6\cos^2\theta \sin^2\theta + \sin^4\theta - 3$$

One could graph it (☹) or, note that

$$\kappa = \cos^4\theta + 6\cos^2\theta \sin^2\theta + \sin^4\theta - 3 = \left(\cos^2\theta + \sin^2\theta\right)^2 + 4\cos^2\theta \sin^2\theta - 3$$

$$= 4\cos^2\theta \sin^2\theta - 2 = \left(2\cos\theta \sin\theta\right)^2 - 2 = \left(\sin 2\theta\right)^2 - 2$$

So kurtosis is largest when $\left(\sin 2\theta\right)^2$ is largest, i.e., when $\sin 2\theta = \pm 1$, which happens when $\theta = \pi/2, 3\pi/2, \ldots$, i.e., when $\theta = \pi/4, 3\pi/4, \ldots$. Using maximum kurtosis as a criterion, ICA would identify the oblique axes as the unmixed coordinates.

C. As in A, but determine how the entropy of the projected distribution depends on θ. *For which projections is it minimized?*

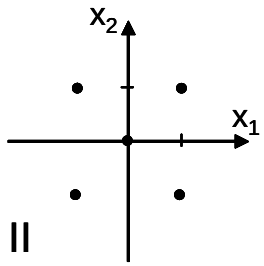
The projected distribution consists of the four points $\left(\pm\cos\theta \pm \sin\theta\right) v_\theta$. At all "typical" angles, i.e., angles that are not multiples of $\pi/4$, the four projection points are distinct. In these cases, each value $\left(\pm\cos\theta \pm \sin\theta\right)$ has probability $1/4$, and the distribution has entropy $-\frac{4}{4}\log_2(1/4) = 2$.

At atypical angles that correspond to the cardinal directions $\theta = 0, \pi/2, \pi, \ldots$, the points coincide in pairs, and each pair projects onto one of two values, ± 1. So they have probability $1/2$, and the distribution has entropy $-\frac{2}{2}\log_2(1/2) = 1$.

At atypical angles that correspond to the oblique directions $\theta = \pi/4, 3\pi/4, \ldots$, two of the points coincide (and project to 0), and two do not, and project to $\pm\sqrt{2}$. So the point 0 has probability $1/2$, and each of the points $\pm\sqrt{2}$ have probability $1/4$. The distribution has entropy $-\frac{2}{4}\log_2(1/4) - \frac{1}{2}\log_2(1/2) = 3/2$.

So the entropy is minimized at the cardinal angles, $\theta = 0, \pi/2, \pi, \ldots$.

Using minimum entropy as a criterion, ICA would identify the cardinal axes as the unmixed coordinates.



D. Same as in A, but take five data points, located at $(\pm 1, \pm 1)$ *and also the origin, (see diagram II) and consider its projection onto a unit vector* $v_\theta = (\cos\theta, \sin\theta)$*, to form a distribution of 5 points. How does the variance of this distribution depend on* θ*? For which projections is it maximized (equivalently, what direction(s) would be selected by PCA?)*

The projected distribution consists of five points, four at positions $x_i = \pm\cos\theta \pm \sin\theta$ and one at the origin. So the variance is

$$V = \langle x^2 \rangle = \frac{1}{5}\left((\cos\theta + \sin\theta)^2 + (\cos\theta - \sin\theta)^2 + (-\cos\theta + \sin\theta)^2 + (-\cos\theta - \sin\theta)^2 \right)$$
$$= \frac{4}{5}\cos^2\theta + \sin^2\theta = \frac{4}{5}$$

again independent of θ. So again, from the point of view of PCA, no directions are special.

E. As in B, but determine how the kurtosis of the projected distribution depends on θ*. For which directions is kurtosis largest?*

Since $\langle x \rangle = 0$ and $V = 4/5$,

$$\kappa = \frac{\langle x^4 \rangle}{(4/5)^2} - 3 = \left(\frac{5}{4}\right)^2 \frac{1}{5}\left((\cos\theta + \sin\theta)^4 + (\cos\theta - \sin\theta)^4 + (-\cos\theta + \sin\theta)^4 + (-\cos\theta - \sin\theta)^4 \right) - 3$$
$$= \frac{5}{4}\left(\cos^4\theta + 6\cos^2\theta\sin^2\theta + \sin^4\theta \right) - 3$$

This of course differs from the result in B only by a scale factor and an offset, so again the kurtosis is largest when $(\sin 2\theta)^2$ is largest, i.e., when $\theta = \pi/4, 3\pi/4, \ldots$. Using maximum kurtosis as a criterion, ICA would identify the oblique axes as the unmixed coordinates.

F. As in B, but determine how the entropy of the projected distribution depends on θ*. For which projections is it minimized?*

The projected distribution consists of the five points $(\pm\cos\theta \pm \sin\theta)v_\theta$ and the origin. At all "typical" angles, i.e., angles that are not multiples of $\pi/4$, the five projection points are distinct. In these cases, each has probability $1/5$, and the distribution has entropy $-\frac{5}{5}\log_2(1/5) = \log_2 5 \approx 2.322$.

At angles that correspond to the cardinal directions $\theta = 0, \pi/2, \pi, \ldots$, four of the points coincide in pairs, and there is one singleton. The two values, ± 1 (the results of the coincident pairs) have probability $2/5$, and the origin has probability $1/5$. So the distribution has entropy $-\frac{2}{5}\log_2(2/5) - \frac{2}{5}\log_2(2/5) - \frac{1}{5}\log_2(1/5) = \log_2 5 - \frac{4}{5} \approx 1.522$.

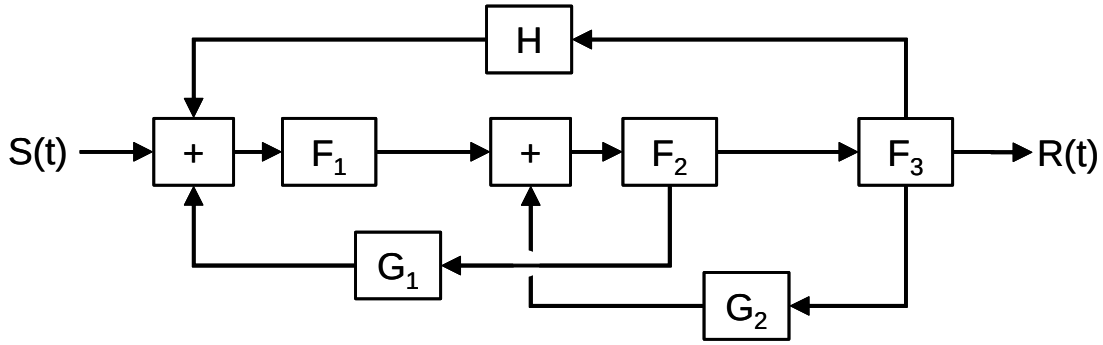
At angles that correspond to the oblique directions $\theta = \pi/4, 3\pi/4, \ldots$, three of the points coincide (and project to 0), and two do not, and project to $\pm\sqrt{2}$. So the point 0 has probability $3/5$, and each of the points $\pm\sqrt{2}$ have probability $1/5$. The distribution has entropy $-\frac{3}{5}\log_2(3/5) - \frac{1}{5}\log_2(1/5) - \frac{1}{5}\log_2(1/5) = \log_2 5 - \frac{3}{5}\log_2 3 \approx 1.371$.

So the entropy is minimized at the oblique angles, $\theta = \pi/4, 3\pi/4, \ldots$.

Using minimum entropy as a criterion, ICA would identify the oblique axes as the unmixed coordinates. Adding one point makes a difference for entropy, but not for the kurtosis.

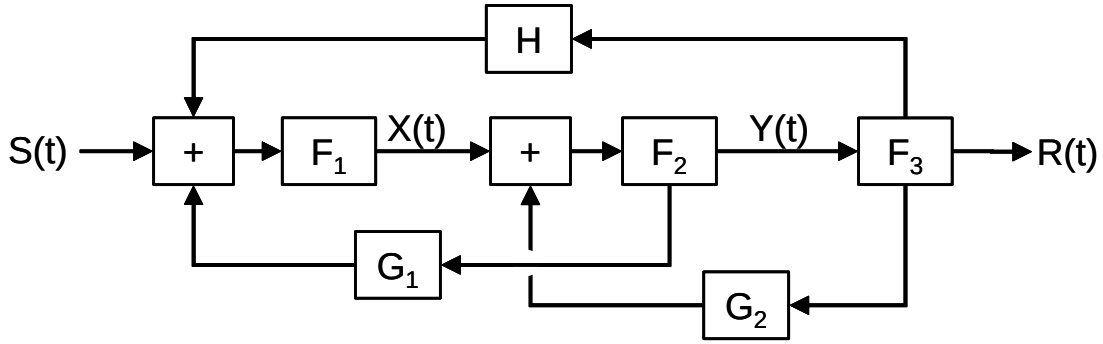
Q6. Linear systems and feedback

The diagram shows a linear system with input $S(t)$ and output $R(t)$, and the filters F_i, G_i, and H are linear filters with transfer functions $\tilde{F}_i(\omega)$, $\tilde{G}_i(\omega)$, and $\tilde{H}(\omega)$.



A. Find the Fourier transform $\tilde{S}(\omega)$ of $S(t)$ in terms of the Fourier transform $\tilde{R}(\omega)$ of $R(t)$.

Write $X(t)$ as the output of F_1 and $Y(t)$ as the output of F_2:



Then (omitting the ω-argument throughout, and using the standard rules for the transfer functions of combined linear systems), $\tilde{Y} = \tilde{F}_2\left(\tilde{X} + \tilde{F}_3\tilde{G}_2\tilde{Y}\right)$, from which $\tilde{Y} = \frac{\tilde{F}_2}{1 - \tilde{F}_2\tilde{F}_3\tilde{G}_2}\tilde{X}$.

Similarly, $\tilde{X} = \tilde{F}_1\left(\tilde{S} + \tilde{F}_2\tilde{G}_1\tilde{X} + \tilde{F}_3\tilde{H}\tilde{Y}\right)$, and, substituting the above equation for $\tilde{Y}$,

$\tilde{X} = \tilde{F}_1\left(\tilde{S} + \tilde{F}_2\tilde{G}_1\tilde{X} + \tilde{F}_3\tilde{H}\dfrac{\tilde{F}_2}{1-\tilde{F}_2\tilde{F}_3\tilde{G}_2}\tilde{X}\right)$, from which

$$\tilde{X} = \frac{\tilde{F}_1\tilde{S}}{1-\tilde{F}_1\tilde{F}_2\tilde{G}_1 - \tilde{F}_1\tilde{F}_3\tilde{H}\dfrac{\tilde{F}_2}{1-\tilde{F}_2\tilde{F}_3\tilde{G}_2}}.$$

Using $\tilde{Y} = \dfrac{\tilde{F}_2}{1-\tilde{F}_2\tilde{F}_3\tilde{G}_2}\tilde{X}$ for $\tilde{Y}$:

$$\tilde{Y} = \frac{\tilde{F}_1\tilde{F}_2\tilde{S}}{\left(1-\tilde{F}_1\tilde{F}_2\tilde{G}_1\right)\left(1-\tilde{F}_2\tilde{F}_3\tilde{G}_2\right)-\tilde{F}_1\tilde{F}_2\tilde{F}_3\tilde{H}}, \text{ so}$$

$$\tilde{R} = \tilde{F}_3\tilde{Y} = \frac{\tilde{F}_1\tilde{F}_2\tilde{F}_3}{\left(1-\tilde{F}_1\tilde{F}_2\tilde{G}_1\right)\left(1-\tilde{F}_2\tilde{F}_3\tilde{G}_2\right)-\tilde{F}_1\tilde{F}_2\tilde{F}_3\tilde{H}}\tilde{S}.$$

B. Write out the relationship between $\tilde{S}(\omega)$ and $\tilde{R}(\omega)$ when (a) the feedback between F_2 and F_1 is absent, (b) the feedback between F_3 and F_2 is absent, or (c) the feedback between F_3 and F_1 is absent.

This is readily done by setting the appropriate filter in the above circuit to zero:

(a) Setting $\tilde{G}_1 = 0$ yields $\tilde{R} = \dfrac{\tilde{F}_1\tilde{F}_2\tilde{F}_3}{1-\tilde{F}_2\tilde{F}_3\tilde{G}_2-\tilde{F}_1\tilde{F}_2\tilde{F}_3\tilde{H}}\tilde{S}$.

(b) Setting $\tilde{G}_2 = 0$ yields $\tilde{R} = \dfrac{\tilde{F}_1\tilde{F}_2\tilde{F}_3}{1-\tilde{F}_1\tilde{F}_2\tilde{G}_1-\tilde{F}_1\tilde{F}_2\tilde{F}_3\tilde{H}}\tilde{S}$.

(c) Setting $\tilde{H} = 0$ yields $\tilde{R} = \dfrac{\tilde{F}_1\tilde{F}_2\tilde{F}_3}{\left(1-\tilde{F}_1\tilde{F}_2\tilde{G}_1\right)\left(1-\tilde{F}_2\tilde{F}_3\tilde{G}_2\right)}\tilde{S}$.

C. Say that it is known that $F_1 = F_2 = F_3 = F$, and that $G_1 = G_2 = G$. Now consider the four configurations above (the full configuration, and the three configurations of part B, in which one of the three feedback filters has been removed). Do they all produce different outputs?

The transfer functions are given by:

Full system: $\dfrac{\tilde{R}}{\tilde{S}} = \dfrac{\tilde{F}^3}{\left(1-\tilde{F}^2\tilde{G}\right)^2-\tilde{F}^3\tilde{H}}$

$\tilde{G}_1$ removed: $\dfrac{\tilde{R}}{\tilde{S}} = \dfrac{\tilde{F}^3}{1-\tilde{F}^2\tilde{G}-\tilde{F}^3\tilde{H}}$.

$\tilde{G}_2$ removed: $\dfrac{\tilde{R}}{\tilde{S}} = \dfrac{\tilde{F}^3}{1-\tilde{F}^2\tilde{G}-\tilde{F}^3\tilde{H}}$, the same as $\tilde{G}_1$ removed.

$\tilde{H}$ removed: $\dfrac{\tilde{R}}{\tilde{S}} = \dfrac{\tilde{F}^3}{\left(1-\tilde{F}^2\tilde{G}\right)^2}$.