Mutual Information, Synergy and Some Curious Phenomena for Simple Channels

I. Kontoyiannis
Div of Applied Mathematics
& Dpt of Computer Science
Brown University
Providence, RI 02912, USA
Email: yiannis@dam.brown.edu

B. Lucena
Division of Computer Science
UC-Berkeley, Soda Hall
Berkeley, CA 94720, USA
Email: lucena@cs.berkeley.edu

***Abstract*— Suppose we are allowed to observe two equally noisy versions of some signal X, where the level of the noise is fixed. We are given a choice: We can either observe two independent noisy versions of X, or two correlated ones. We show that, contrary to what classical statistical intuition suggests, it is often the case that correlated data is more valuable than independent data. We investigate this phenomenon in a variety of contexts, give numerous examples for standard families of channels and present general sufficient conditions for deciding this dilemma. One of these conditions draws an interesting connection with the information-theoretic notion of "synergy," which has received a lot of attention in the neuroscience literature recently.**

I. Introduction

The following examples motivate much of our discussion.

A broadcast channel. Consider the problem of communicating a message to two remote receivers. We are given two options. Either send the message to two independent intermediaries and have each of them relay the message to one of the receivers, or send the message to only one intermediary and have her re-send the message to the two receivers in two independent transmissions. Assuming that the two receivers are allowed to cooperate, which option is more efficient? Although intuition perhaps suggests that it is preferable to use two independent intermediaries so that we are not "stuck" with the noise incurred in the first transmission, in many cases this intuition turns out to be wrong.

A sampling problem. Suppose we want to test for the presence of a disease in a certain population, and we do so by randomly selecting and testing members of the population. Assuming the test is imperfect (there is a certain chance we might get a false positive or a false negative), is it better to test one person twice, or two people, once each? Again, although it may seem more reasonable to test two people independently, we find that the opposite is often true.

These questions are formalized as follows; see Fig. 1. We have two sets of conditional distributions, or channels, $P = (P(y|x))$ and $Q = (Q(z|y))$ on the same alphabet A.

Scenario 1. Independent observations. Suppose X has a given distribution $P_X(x)$ on A, let Y_1, Y_2 be conditionally independent given X, each with distribution $P(y|X)$, and let Z_1 and Z_2 be distributed according to $Q(z|Y_1)$ and $Q(z|Y_2)$, independently of the remaining variables. In this case, Z_1 and Z_2 are two (conditionally) independent, noisy observations of the variable of interest X.

Scenario 2. Correlated observations. Given X as before, let Y be distributed according to $P(y|X)$, and let W_1, W_2 be conditionally independent of X and of one another given Y, each of them distributed according to $Q(z|Y)$. In this case the two noisy observations W_1 and W_2 are typically *not* conditionally independent given X.

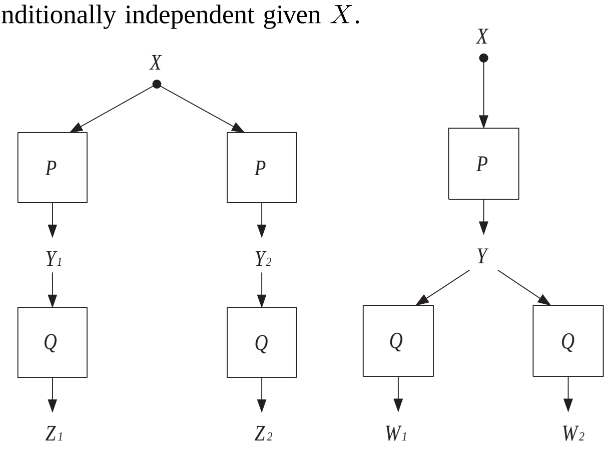

Fig. 1. Independent observations (left) vs. correlated observations (right).

The joint distribution of (X, Z_1) is of course the same as the joint distribution of (X, W_1), so there is no difference in observing Z_1 or W_1. But the joint distribution of (X, Z_1, Z_2) is different from that of (X, W_1, W_2), so the amount of information we obtain from the second observation is different in these two scenarios.

In this note we consider the following question: For a given pair of channels P and Q and for a fixed "source" distribution P_X, we can choose between the two independent observations (Z_1, Z_2) or the two correlated observations (W_1, W_2). If our goal is to maximize the mutual information between X and our observations, which scenario should we pick?

At first glance, it is perhaps natural to expect that the two (conditionally) independent observations would always be better. The first goal of this work is to show that this intuition is *often* incorrect. We exhibit simple and very natural examples for which correlated observations provide an advantage over independent ones. Our second goal is to investigate the underlying reasons for this phenomenon. For a variety of channels of interest we derive simple sufficient conditions, which guarantee that one or the other scenario is preferable. Moreover, in attempting to understand the "surprising" phenomenon where the second scenario is better – when the second *correlated* observation is *more valuable* than the second *independent* observation – we draw a connection with the statistical phenomenon of "synergy" identified in neurobiological studies of the brain.

Of course one could choose a figure of merit other than mutual information – we could ask which if the two scenarios yields a channel with a higher capacity, or we could compare the different probabilities of error when estimating X based on the two different types of observations. We pursue these and other related questions in a longer version of this work.

Throughout the paper when we say that "correlated observations are better than independent ones" we mean that $I(X;W_1,W_2) > I(X;Z_1,Z_2)$. All logarithms are natural logarithms with base e, and all familiar information-theoretic quantities are therefore expressed in nats. All proofs are omitted here; complete arguments will be presented in the longer version of this paper.

In Section II we look at the simplest case of binary channels. In Section III we consider a more general class of finite-alphabet channels called Potts channels, and in Section IV we offer some general results valid for arbitrary channels. There we introduce the concept of "synergy," we discuss its relevance in this setting, and we show that the existence of synergy is a sufficient condition for our "surprising" phenomenon – for the correlated observations (W_1,W_2) to carry more information about X than the independent observations (Z_1,Z_2). In Section V we give general results on the presence or absence of synergy for Gaussian channels.

Note that various related issues have been examined in the literature. The idea of synergy implicitly appears in [6][3][4][9]; for connections with neuroscience see [1][8]; the relationship of some related statistical issues with von Neumann's problem of computation in noisy circuits is examined in [2]; some analogous problems to those treated here for large tree networks are considered in [7].

II. Binary Channels

Example 1. Suppose P is the Z-channel with parameter $\delta \in (0,1)$, denoted $Z(\delta)$, and Q is the binary symmetric channel with parameter $\epsilon \in (0,1/2)$, denoted BSC(ϵ); see Figure 2.

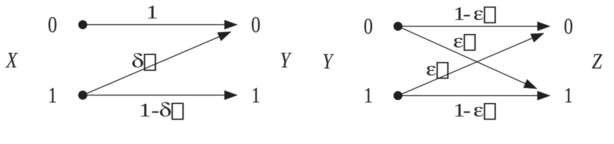

Fig. 2. The $Z(\delta)$ channel and the BSC(ϵ) channel.

This setting can be interpreted as a simple model for the sampling problem described in the Introduction: The root variable X is equal to 1 if a disease is present in a certain population, and it is 0 otherwise. Suppose our prior understanding is that X is Bernoulli(p), i.e., the prior probability that the disease is not present is $(1-p)$, and the prior probability that it is present is p, in which case a proportion $(1-\delta) > 0$ of the people are infected. If the disease is not present then we will certainly pick a healthy individual, otherwise the probability we will pick an infected individual is $(1-\delta)$. Moreover, we suppose that in testing the selected individuals, the probability of either a false positive or a false negative is ϵ.

With the (fairly realistic) parameters $p = 1/2$, $(1-\delta) = .02$ and $\epsilon = .1$, direct calculation shows that testing one individual twice is more valuable than testing two people, once each, so that correlated observations are better than independent ones, or, formally, $I(X;W_1,W_2) > I(X;Z_1,Z_2)$.

In fact, as we show next, this phenomenon occurs for a wide range of parameter values.

Proposition 1. (Binary Asymmetric Channels) Suppose that $X \sim$ Bernoulli(p), P is the $Z(\delta)$ channel and Q is the BSC(ϵ) channel, where the parameters are in the range $p \in (0,1)$, $\delta \in (0,1)$, $\epsilon \in (0,1/2)$. For all $p \in (0,1)$:

(i) For all $\epsilon \in (0,1/2)$ there exists δ small enough such that independent observations are better, i.e. there exists δ^* such that $\delta \in (0,\delta^*)$ implies $I(X;W_1,W_2) < I(X;Z_1,Z_2)$.

(ii) For all $\epsilon \in (0,1/2)$ there exists δ large enough so that correlated observations are better, i.e., there exists δ^* such that $\delta \in (\delta^*,1)$, implies $I(X;W_1,W_2) > I(X;Z_1,Z_2)$.

(iii) For any $\delta \in (0,1)$ there exists ϵ small enough such that independent observations are better, i.e., there exists ϵ^* such that $\epsilon < \epsilon^*$ implies $I(X;W_1,W_2) < I(X;Z_1,Z_2)$.

(iv) For $\delta > \frac{2p}{2p+1}$, correlated observations are better for large enough ϵ, and for $\delta < \frac{2p}{2p+1}$ independent observations are better for large enough ϵ: For $\delta > \frac{2p}{2p+1}$, there exists ϵ_1 such that $\epsilon > \epsilon_1$ implies $I(X;W_1,W_2) > I(X;Z_1,Z_2)$, whereas, for $\delta < \frac{2p}{2p+1}$ there exists ϵ_2 such that $\epsilon > \epsilon_2$ implies $I(X;W_1,W_2) < I(X;Z_1,Z_2)$.

The following three diagrams show three numerical examples that illustrate the result of the proposition quantitatively.

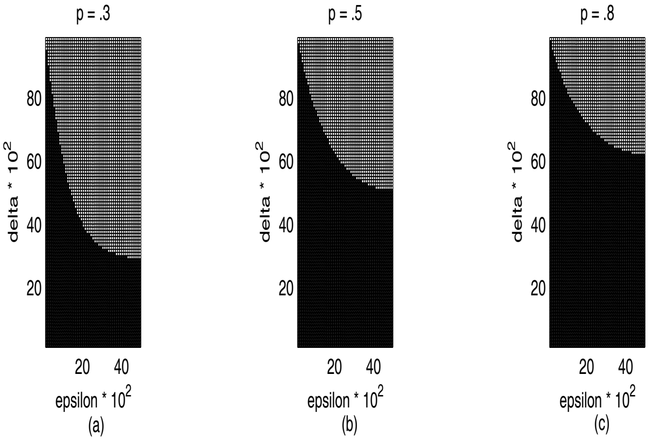

Fig. 3. Lighter color indicates regions in the δ-ϵ plane where the correlated observations are more informative for three different values of p.

The above proposition says that, in the particular case where P is $Z(\delta)$ and Q is BSC(ϵ), correlated observations are better than independent ones only when δ is large enough, that is, when the first channel P is highly non-symmetric. On the other extreme, in the following proposition we show that when P and Q are both BSCs then correlated observations are never better than independent ones.

Proposition 2. (Binary Symmetric Channels) Suppose that $X \sim$ Bernoulli(p), and P and Q are BSCs with parameters ϵ_1 and ϵ_2, respectively. Then for any choice of the parameters $p \in (0,1)$, $\epsilon_1 \in (0,1/2)$, $\epsilon_2 \in (0,1/2)$, independent observations are always at least as good as correlated ones, i.e., $I(X;Z_1,Z_2) \geq I(X;W_1,W_2)$.

Roughly speaking, Proposition 1 says that the "surprising" phenomenon (correlated observations being better than independent ones) only occurs when the first channel P is highly non-symmetric, and Proposition 2 says that it *never* happens for symmetric channels. This may seem to suggest that the surprise is caused by this lack of symmetry, but as we show in the following section, when the alphabet is not binary, we can still get the surprise with perfectly symmetric channels.

III. The Potts Channel

A natural generalization of the BSC to a general finite alphabet A with $m \geq 3$ elements is the Potts(α) channel defined by the conditional probabilities $W(i|i) = \alpha$ for all $i \in A$, and $W(j|i) = \beta = \frac{1-\alpha}{m-1}$ for all $i \neq j \in A$, where the parameter α is in $(0,1)$; see Figure 4.

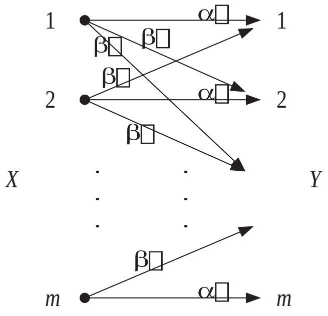

Fig. 4. The Potts channel with parameter α on an alphabet of size m.

Proposition 3. (Potts Channels) If the root X is uniformly distributed and both P and Q are Potts channels with the same parameter α, then there exist $0 < \alpha_* < \alpha^* < 1$ such that:

(i) For all $\alpha > \alpha^*$, independent observations are better that correlated ones, i.e., $I(X;Z_1,Z_2) > I(X;W_1,W_2)$.

(ii) For all $\alpha < \alpha_*$, correlated observations are better that independent ones, i.e., $I(X;Z_1,Z_2) < I(X;W_1,W_2)$.

IV. Arbitrary Channels and Synergy

Next we investigate general conditions, under which the surprising phenomenon of getting more information from correlated observations than from independent ones occurs.

Since the distributions of (X,Z_1) and of (X,W_1) are the same, the amount of information obtained from the first sample is the same in both scenarios, $I(X;Z_1) = I(X;W_1)$. So the surprise will occur only if the second correlated sample is more valuable than the second independent one, namely if $I(X;W_2|W_1) > I(X;Z_2|Z_1)$. But since Z_1 and Z_2 are conditionally independent given X, we always have

$$I(X;Z_2|Z_1) \leq I(X;Z_2) = I(X;W_2),$$

and therefore a sufficient condition for the surprise is that

$$I(X;W_2|W_1) > I(X;W_2).$$

This says that the information $I(X;W_2|W_1)$ we obtain about X from the second observation W_2 given W_1, is *greater* than the information $I(X;W_2) = I(X;W_1)$ obtained from the first observation; in other words, W_2 not only contains no "redundant" information, but it somehow "collaborates" with W_1 in order to give extra information about X. This idea is formalized in the following definition.

Definition. Consider three jointly distributed random variables (X,V_1,V_2). We define the *synergy* between them as,

$$S(X,V_1,V_2) = I(X;V_2|V_1) - I(X;V_2),$$

and we say that the random variables X, V_1, V_2 are *synergetic* whenever $S(X,V_1,V_2) > 0$.

The concept and the term synergy as defined here are borrowed from the neuroscience literature; see, e.g., [8]. There, X represents some kind of sensory stimulus, for example an image, and the observations V_1, V_2 represent the response of two neurons in the brain, for example in the primary visual cortex. In [8] the amount of synergy between V_1 and V_2 is defined as the amount of information (V_1,V_2) jointly provide about X, minus the sum of the amounts of information they provide individually:

$$S(X,V_1,V_2) = I(X;V_1,V_2) - I(X;V_1) - I(X;V_2). \qquad (1)$$

A simple application of the chain rule for mutual information shows that this is equivalent to our definition. Similarly, by the chain rule we also have that the synergy is circularly symmetric in its three arguments. Alternatively, the synergy can be expressed in terms of entropies as:

$$\begin{aligned} S(X,V_1,V_2) \;=\; & -H(X,V_1,V_2) \\ & +H(X,V_1) + H(X,V_2) + H(V_1,V_2) \\ & -H(X) - H(V_1) - H(V_2). \qquad (2) \end{aligned}$$

Proposition 4. (General Channels and Synergy) Let X have an arbitrary distribution P_X, and P, Q be two arbitrary finite-alphabet channels. If the synergy $S(X,W_1,W_2)$ is positive, then the correlated observations are better than the independent ones, $I(X;W_1,W_2) > I(X;Z_1,Z_2)$. More generally, the correlated observations are better than the independent ones if and only if $S(X,W_1,W_2) > -I(Z_1;Z_2)$.

Although the statement above is given for finite-alphabet channels, the same result holds channels on arbitrary alphabets. Motivated by Proposition 4, we now look at three simple examples, and determine conditions under which synergy does or does not occur.

Example 2. Maximal Binary Synergy. Consider three binary variables (X, V_1, V_2), where we think of V_1 and V_2 as noisy observations of X and we assume that (X, V_1) has the same distribution as (X, V_2). This covers all the scenarios considered in Section II. Under what conditions is the synergy $S(X, V_1, V_2)$ maximized? We have,

$$\begin{aligned} S(X, V_1, V_2) &= I(V_1; V_2|X) - I(V_1; V_2) \\ &= H(V_1|X) - H(V_1|X, V_2) - I(V_1; V_2) \\ &\leq 1, \end{aligned}$$

with equality if and only if $H(V_1|X) = 1$ and $H(V_1|X, V_2) = I(V_1; V_2) = 0$, that is, if and only if the three variables are pairwise independent, they all have Bernoulli(1/2) distribution, and any one of them is a deterministic function of the other two. This can be realized in essentially only one way: X and V_1 are independent Bernoulli(1/2) random variables, and V_2 is their sum modulo 2. In that case the maximal synergy is also intuitively obvious: The first observation is independent of X and hence entirely useless, but the second one is maximally useful (given the first), as it tells us the value of X exactly.

Example 3. Frustration in a Simple Spin System. Matsuda in [5] presented a simple example which exhibits an interesting connection between the notion of synergy and the phenomenon of "frustration" in a physical system. In our notation, let X, V_1, V_2 denote three random variables with values in $A = \{+1, -1\}$. Physically, these represent the directions of the spins of three different particles. Assume that their joint distribution is given by the Gibbs measure

$$\Pr(X = x, V_1 = v_1, V_2 = v_2) = \frac{1}{Z}\, e^{\alpha(xv_1 + xv_2 + v_1v_2)},$$

where $Z = Z(\alpha) = 2e^{3\alpha} + 6e^{-\alpha}$ is the normalizing constant, and α is a parameter related to the temperature of the system and the strength of the interaction between the particles. When α is positive, then the preferred states of the system – i.e., the triplets (x, v_1, v_2) that have higher probability – are those in which each pair of particles has the same spin, namely $x = v_1 = v_2 = +1$ and $x = v_1 = v_2 = -1$. Similarly, when α is negative the preferred states are those in which the spins in each pair are different; but this is of course impossible to achieve for all three pairs simultaneously. In physics, this phenomenon where the preferred local configurations are incompatible with the global state of the system is referred to as "frustration."

A cumbersome but straightforward calculation shows that the synergy $S(X, V_1, V_2)$ can be calculated explicitly to be,

$$2\log(Z(\alpha)/8) + \frac{6(e^{3\alpha} + e^{-\alpha})}{Z(\alpha)} \log\left[\frac{2e^{\alpha}}{e^{3\alpha} + e^{-\alpha}}\right],$$

which is plotted as a function of α in Figure 5. We observe that the synergy is positive exactly when the system is frustrated, i.e., when α is negative.

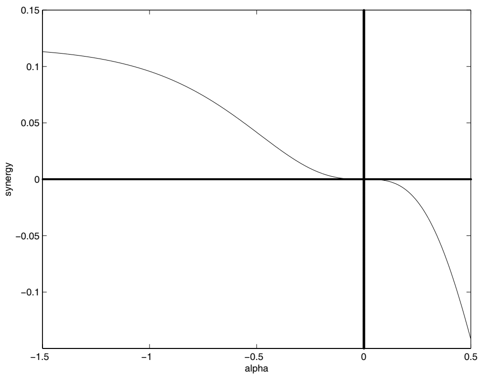

Fig. 5. The synergy of the spin system in Example 3 as a function of α.

Example 4. Gaussian Additive Noise Channels. Suppose X is a Gaussian signal and V_1, V_2 are observations obtained through additive Gaussian noise channels. Specifically, we assume that the total noise power between X and each V_i is fixed at some value N, but that we can control the degree of the dependence between the two observations,

$$\begin{aligned} V_1 &= X + rZ_0 + sZ_1 \\ V_2 &= X + rZ_0 + sZ_2, \end{aligned}$$

where Z_0, Z_1, Z_2 are independent $N(0, N)$ variables which are also independent of $X \sim N(0, 1)$, the parameter $r \in [0, 1]$ is in our control, and s is chosen so that the total noise power stays constant, i.e., $r^2 + s^2 = 1$. See Figure 6.

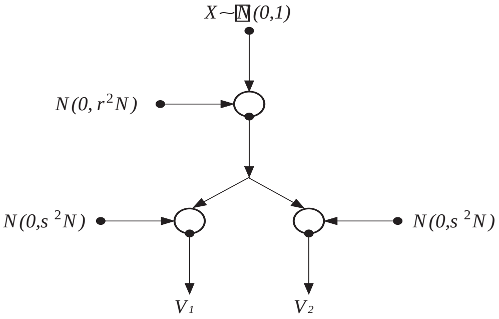

Fig. 6. The additive Gaussian noise setting of Example 3.

How should we choose the parameter r in order to maximize the information $I(X; V_1, V_2)$? Taking $r = 0$ corresponds to independent observations, taking $0 < r < 1$ gives correlated observations as in the second scenario, and the extreme case $r = 1$ gives the same observation twice $V_1 = V_2$.

The covariance matrix of (X, V_1, V_2) is easily read off of the above description, and the mutual information $I(X; V_1, V_2)$ can be calculated explicitly to be,

$$I(X; V_1, V_2) = \frac{1}{2}\log\left[1 + \frac{2}{N(1 + r^2)}\right],$$

which is obviously decreasing in r. Therefore the maximum is achieved at $r = 0$, meaning that here independent observations are *always* better. Therefore, Proposition 4 implies that we *never* have positive synergy for any r.

V. General Gaussian Channels

Example 5. The General Symmetric Gaussian Case. Suppose (X, V_1, V_2) are arbitrary (jointly) Gaussian random variables such that (X, V_1) and (X, V_2) have the same distribution. When is the synergy $S(X, V_1, V_2)$ maximized? As we will see, the synergy is positive if and only if the correlation between the two observations is either negative or positive but small enough.

Without loss of generality we may take all three random variables to have zero mean. In the most general case, the covariance matrix of (X, V_1, V_2) can be written in the form,

$$K = \begin{pmatrix} \sigma^2 & \alpha & \alpha \\ \alpha & \tau^2 & \beta \\ \alpha & \beta & \tau^2 \end{pmatrix}$$

for arbitrary positive variances σ^2, τ^2, and for $\alpha \in (-\sigma\tau, \sigma\tau)$, $\beta \in (-\tau^2, \tau^2)$. In order to ensure that K is positive definite (so that it can be a legitimate covariance matrix) we also need to restrict the parameter values so that $\det(K) > 0$, which reduces to the relation,

$$\tau^2 + \beta > \frac{2\alpha^2}{\sigma^2}. \qquad (3)$$

Using the expression in (2), where the entropies now are interpreted as differential entropies, the synergy can be evaluated by a straightforward calculation,

$$S(X, V_1, V_2) = \frac{1}{2} \log \left\{ \frac{(\tau^2 + \beta)(\sigma^2\tau^2 - \alpha^2)^2}{\sigma^2\tau^4[\sigma^2(\tau^2 + \beta) - 2\alpha^2]} \right\}.$$

Solving the inequality $S(X, V_1, V_2) > 0$ we obtain that we have synergy if and only if

$$\beta < \frac{\alpha^2\tau^2}{2\sigma^2\tau^2 - \alpha^2}, \qquad (4)$$

which means that we have synergy if and only if β is negative (but still not "too" negative, subject to the constraint (3)) or small enough according to (4).

Example 6. A Gaussian Example with Asymmetric Observations. We take X and the observations V_1, V_2 to be jointly Gaussian, all with unit variances and zero means. We assume that the correlation between the two observations remains fixed, but we let the correlation between X and each V_i vary in such a way that it is split between the two:

$$E(XV_1) = \lambda\rho \qquad \text{and} \qquad E(XV_2) = (1 - \lambda)\rho.$$

We also assume that $\lambda \in (0, 1)$ is a parameter we can control, that $\rho \in (-1, 1)$ is fixed, and we define $\overline{\rho} > 0$ by $\rho^2 + \overline{\rho}^2 = 1$. By symmetry, we can restrict attention to the range $\lambda \in (0, 1/2]$.

Our question here is to see whether the asymmetry in the two observations (corresponding to values of $\lambda \neq 1/2$) increases the synergy or not. Before proceeding with the answer we look at the two extreme points.

For $\lambda = 0$ we see that X and V_1 are independent $N(0, 1)$ variables, and their joint distribution with V_2 can be described by $V_2 = \rho X + \overline{\rho} V_1$. In this case the mutual information $I(X; V_2)$ is some easy to calculate finite number, whereas the conditional mutual information $I(X; V_2|V_1)$ is infinite, because X and V_2 are deterministically related given V_1. Therefore the synergy

$$S(X, V_1, V_2) = I(X; V_2|V_1) - I(X; V_2) = \infty.$$

At the other extreme, when $\lambda = 1/2$ we have a symmetric distribution as in the previous example, with $\sigma^2 = \tau^2 = 1$, $\alpha = \rho/2$ and $\beta = \overline{\rho}$. Therefore, here we only have synergy when $\beta = \overline{\rho}$ is negative or small enough, i.e., when the correlation between the two observations is either negative or small enough. Specifically, substituting the above parameter values in condition (4) we see that we have symmetry if and only if $\overline{\rho}$ satisfies $\overline{\rho}^3 + \overline{\rho}^2 + 7\overline{\rho} < 1$, i.e., if and only if $-1 < \overline{\rho} < 0.13968\ldots$.

The fact that the synergy is infinite for $\lambda = 0$ and reduces to a reasonable value (which may or may not be positive) at $\lambda = 1/2$ suggests that perhaps the asymmetry in some way "helps" the synergy, and that the synergy may in fact be decreasing with λ. As it turns out, this is "almost" true:

Proposition 5. (Asymmetric Gaussian Observations) Suppose that (X, V_1, V_2) are jointly Gaussian as described above. Let

$$\epsilon = 1 - \sqrt{8(\sqrt{5/4} - 1)} \approx 0.02826.$$

(i) If $\rho \in (-1, 1 - \epsilon)$ then the synergy is decreasing in λ for all $\lambda \in (0, 1/2)$.

(ii) If $\rho \in (1 - \epsilon, 1)$ then the synergy is decreasing for $\lambda \in (0, \lambda^*)$ and increasing for $\lambda \in (\lambda^*, 1/2)$, where

$$\lambda^* = \lambda^*(\rho) = \frac{1}{2}\left[1 - \sqrt{1 - \frac{4\overline{\rho}}{\rho^2}}\right].$$

Acknowledgments

I.K. was supported in part by a Sloan Foundation Research Fellowship and by NSF grant #00-73378-CCR. B.L. was supported in part by an NSF Mathematical Sciences Postdoctoral Research Fellowship.

References

[1] N. Brenner, Strong S.P., R. Koberle, W. Bialek, and R.R. de Ruyter van Steveninck. Synergy in a neural code. *Neural Comput.*, 12(7):1531–1552, 2000.

[2] W. Evans, C. Kenyon, Y. Peres, and L.J. Schulman. Broadcasting on trees and the Ising model. *Ann. Appl. Probab.*, 10(2):410–433, 2000.

[3] R.M. Fano. *Transmission of information: A statistical theory of communications.* The M.I.T. Press, Cambridge, Mass., 1961.

[4] T.S. Han. Nonnegative entropy measures of multivariate symmetric correlations. *Inform. and Control*, 36(2):133–156, 1978.

[5] H. Matsuda. Physical nature of higher-order mutual information: Intrinsic correlations and frustration. *Phys. Rev. E*, 62(3, Part A):3096–3102, 2000.

[6] W.J. McGill. Multivariate information transmission. *Trans. I.R.E.*, PGIT-4:93–111, 1954.

[7] E. Mossel. Reconstruction on trees: beating the second eigenvalue. *Ann. Appl. Probab.*, 11(1):285–300, 2001.

[8] E. Schneidman, W. Bialek, and M.J. II Berry. Synergy, redundancy, and independence in population codes. *Journal of Neuroscience*, 23(37):11539–11553, 2003.

[9] R.W. Yeung. A new outlook on Shannon's information measures. *IEEE Trans. Inform. Theory*, 37(3, part 1):466–474, 1991.