entropy

MDPI

Article

Coarse-Graining and the Blackwell Order

Johannes Rauh [1,*], Pradeep Kr. Banerjee [1], Eckehard Olbrich [1], Jürgen Jost [1], Nils Bertschinger [2] and David Wolpert [3,4]

[1] Max Planck Institute for Mathematics in the Sciences, 04103 Leipzig, Germany; pradeep@mis.mpg.de (P.K.B.); olbrich@mis.mpg.de (E.O.); jjost@mis.mpg.de (J.J.)
[2] Frankfurt Institute for Advanced Studies, 60438 Frankfurt, Germany; bertschinger@fias.uni-frankfurt.de
[3] Santa Fe Institute, Santa Fe, NM 87501, USA; dhw@santafe.edu
[4] Massachusetts Institute of Technology, Cambridge, MA 02139, USA
* Correspondence: jrauh@mis.mpg.de

Received: 15 June 2017; Accepted: 20 September 2017; Published: 6 October 2017

Abstract: Suppose we have a pair of information channels, κ_1, κ_2, with a common input. The Blackwell order is a partial order over channels that compares κ_1 and κ_2 by the maximal expected utility an agent can obtain when decisions are based on the channel outputs. Equivalently, κ_1 is said to be Blackwell-inferior to κ_2 if and only if κ_1 can be constructed by garbling the output of κ_2. A related partial order stipulates that κ_2 is more capable than κ_1 if the mutual information between the input and output is larger for κ_2 than for κ_1 for any distribution over inputs. A Blackwell-inferior channel is necessarily less capable. However, examples are known where κ_1 is less capable than κ_2 but not Blackwell-inferior. We show that this may even happen when κ_1 is constructed by coarse-graining the inputs of κ_2. Such a coarse-graining is a special kind of "pre-garbling" of the channel inputs. This example directly establishes that the expected value of the shared utility function for the coarse-grained channel is larger than it is for the non-coarse-grained channel. This contradicts the intuition that coarse-graining can only destroy information and lead to inferior channels. We also discuss our results in the context of information decompositions.

Keywords: Channel preorders; Blackwell order; degradation order; garbling; more capable; coarse-graining

MSC: 62C05; 62B15; 94A15

1. Introduction

Suppose we are given the choice of two channels that both provide information about the same random variable, and that we want to make a decision based on the channel outputs. Suppose that our utility function depends on the joint value of the input to the channel and our resultant decision based on the channel outputs. Suppose as well that we know the precise conditional distributions defining the channels, and the distribution over channel inputs. Which channel should we choose? The answer to this question depends on the choice of our utility function as well as on the details of the channels and the input distribution. So, for example, without specifying how we will use the channels, in general we cannot just compare their information capacities to choose between them.

Nonetheless, for certain pairs of channels we can make our choice, even without knowing the utility functions or the distribution over inputs. Let us represent the two channels by two (column) stochastic matrices κ_1 and κ_2, respectively. Then, if there exists another stochastic matrix λ such that $\kappa_1 = \lambda \cdot \kappa_2$, there is never any reason to strictly prefer κ_1; if we choose κ_2, we can always make our decision by chaining the output of κ_2 through the channel λ and then using the same decision function we would have used had we chosen κ_1. This simple argument shows that whatever the three stochastic matrices are and whatever the decision rule we would use if we chose channel κ_1, we can always get the same expected utility by instead choosing channel κ_2 with an appropriate decision rule. In this kind of situation, where $\kappa_1 = \lambda \cdot \kappa_2$, we say that κ_1 is a *garbling* (or *degradation*) of κ_2. It is much more difficult to prove that the converse also holds true:

Theorem 1. *(Blackwell's theorem [1]) Let κ_1, κ_2 be two stochastic matrices representing two channels with the same input alphabet. Then the following two conditions are equivalent:*

1. *When the agent chooses κ_2 (and uses the decision rule that is optimal for κ_2), her expected utility is always at least as big as the expected utility when she chooses κ_1 (and uses the optimal decision rule for κ_1), independent of the utility function and the distribution of the input S.*
2. *κ_1 is a garbling of κ_2.*

Blackwell formulated his result in terms of a statistical decision maker who reacts to the outcome of a *statistical experiment*. We prefer to speak of a decision problem instead of a statistical experiment. See [2,3] for an overview.

Blackwell's theorem motivates looking at the following partial order over channels κ_1, κ_2 with a common input alphabet:

$$\kappa_1 \leq \kappa_2 \quad :\Longleftrightarrow \begin{cases} \text{one of the two statements} \\ \text{in Blackwell's theorem holds true.} \end{cases}$$

We call this partial order the *Blackwell order* (this partial order is called *degradation order* by other authors [4,5]). If $\kappa_1 \leq \kappa_2$, then κ_1 is said to be Blackwell-inferior to κ_2. Strictly speaking, the Blackwell order is only a preorder, since there are channels $\kappa_1 \neq \kappa_2$ that satisfy $\kappa_1 \leq \kappa_2 \leq \kappa_1$ (when κ_1 arises from κ_2 by permuting the output alphabet). However, for our purposes, such channels can be considered as equivalent. We write $\kappa_1 < \kappa_2$ if $\kappa_1 \leq \kappa_2$ and $\kappa_1 \not\geq \kappa_2$. By Blackwell's theorem, this implies that κ_2 performs at least as well as κ_1 in any decision problem and that there exist decision problems in which κ_2 outperforms κ_1.

For a given distribution of S, we can also compare κ_1 and κ_2 by comparing the two mutual informations $I(S;X_1)$, $I(S;X_2)$ between the common input S and the channel outputs X_1 and X_2. The data processing inequality shows that $\kappa_2 \geq \kappa_1$ implies $I(S;X_2) \geq I(S;X_1)$. However, the converse implication does not hold. The intuitive reason is that for the Blackwell order, not only the amount of information is important. Rather, the question is how much of the information that κ_1 or κ_2 preserve is relevant for a given fixed decision problem (that is, a given fixed utility function).

Given two channels κ_1, κ_2, suppose that $I(S;X_2) \geq I(S;X_1)$ for all distributions of S. In this case, we say that κ_2 is *more capable* than κ_1. Does this imply that $\kappa_1 \leq \kappa_2$? The answer is known to be negative in general [6]. In Proposition 2 we introduce a new surprising example of this phenomenon with a particular structure. In fact, in this example, κ_1 is a Markov approximation of κ_2 by a deterministic function, in the following sense: Consider another random variable $f(S)$ that arises from S by applying a (deterministic) function f. Given two random variables S, X, denote by $X \leftarrow S$ the channel defined by the conditional probabilities $P_{X|S}(x|s)$, and let $\kappa_2 := (X \leftarrow S)$ and $\kappa_1 := (X \leftarrow f(S)) \cdot (f(S) \leftarrow S)$. Thus, κ_1 can be interpreted as first replacing S by $f(S)$ and then sampling X according to the conditional distribution $P_{X|S}(x|f(s))$. Which channel is superior? Using the data processing inequality, it is easy to see that κ_1 is less capable than κ_2. However, as Proposition 2 shows, in general $\kappa_1 \not\leq \kappa_2$.

We call κ_1 a Markov approximation, because the output of κ_1 is independent of the input S given $f(S)$. The channel κ_1 can also be obtained from κ_2 by "pre-garbling" (Lemma 3); that is, there is another stochastic matrix λ^f that satisfies $\kappa_1 = \kappa_2 \cdot \lambda^f$. It is known that pre-garbling may improve the performance of a channel (but not its capacity) as we recall in Section 2. What may be surprising is that this can happen for pre-garblings of the form λ^f, which have the effect of coarse-graining according to f.

The fact that the more capable preorder does not imply the Blackwell order shows that "Shannon information," as captured by the mutual information, is not the same as "Blackwell information," as needed for the Blackwell decision problems. Indeed, our example explicitly shows that even though coarse-graining always reduces Shannon information, it need not reduce Blackwell information. Finally, let us mention that there are further ways of comparing channels (or stochastic matrices); see [5] for an overview.

Proposition 2 builds upon another effect that we find paradoxical: Namely, there exist random variables S, X_1, X_2 and there exists a function $f : \mathcal{S} \to \mathcal{S}'$ from the support of S to a finite set S' such that the following holds:

1. S and X_1 are independent given $f(S)$.
2. $(X_1 \leftarrow f(S)) \;\le\; (X_2 \leftarrow f(S))$.
3. $(X_1 \leftarrow S) \;\not\le\; (X_2 \leftarrow S)$.

Statement (1) says that everything X_1 knows about S, it knows through $f(S)$. Statement (2) says that X_2 knows more about $f(S)$ than X_1. Still, (3) says that we cannot conclude that X_2 knows more about S than X_1. The paradox illustrates that it is difficult to formalize what it means to "know more."

Understanding the Blackwell order is an important aspect of understanding information decompositions; that is, the quest to find new information measures that separate different aspects of the mutual information $I(S; X_1, \ldots, X_k)$ of k random variables $X_1, \ldots, X_k$ and a target variable S (see the other contributions of this special issue and references therein). In particular, [7] argues that the Blackwell order provides a natural criterion when a variable X_1 has unique information about S with respect to X_2. We hope that the examples we present here are useful in developing intuition on how information can be shared among random variables and how it behaves when applying a deterministic function, such as a coarse-graining. Further implications of our examples on information decompositions are discussed in [8]. In the converse direction, information decomposition measures (such as measures of unique information) can be used to study the Blackwell order and deviations from the Blackwell order. We illustrate this idea in Example 4.

The remainder of this work is organized as follows: In Section 2, we recall how pre-garbling can be used to improve the performance of a channel. We also show that the pre-garbled channel will always be less capable and that simultaneous pre-garbling of both channels preserves the Blackwell order. In Section 3, we state a few properties of the Blackwell order, and we explain why we find these properties counter-intuitive and paradoxical. In particular, we show that coarse-graining the input can improve the performance of a channel. Section 4 contains a detailed discussion of an example that illustrates these properties. In Section 5 we use the unqiue information measure from [7], which has properties similar to the Le Cam's deficiency, to illustrate deviations from the Blackwell relation.

2. Pre-Garbling

As discussed above (and as made formal in Blackwell's theorem (Theorem 1)), garbling the output of a channel ("post-garbling") never increases the quality of a channel. On the other hand, garbling the input of a channel ("pre-garbling") may increase the performance of a channel, as the following example shows.

Example 1. *Suppose that an agent can choose an action from a finite set $\mathcal{A}$. She then receives a utility $u(a,s)$ that depends both on the chosen action $a \in \mathcal{A}$ and on the value s of a random variable S. Consider the channels*

$$\kappa_1 = \begin{pmatrix} 0.9 & 0 \\ 0.1 & 1 \end{pmatrix} \text{ and } \kappa_2 = \kappa_1 \cdot \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0.9 \\ 1 & 0.1 \end{pmatrix},$$

and the utility function

s	*0*	*0*	*1*	*1*
a	*0*	*1*	*0*	*1*
$u(s,a)$	*2*	*0*	*0*	*1*

For uniform input, the optimal decision rule for κ_1 *is*

$$a(0) = 0,\ a(1) = 1$$

and the opposite

$$a(0) = 1,\ a(1) = 0$$

for κ_2*. The expected utility with* κ_1 *is 1.4, while using* κ_2*, it is slightly higher (1.45).*

It is also not difficult to check that neither of the two channels is a garbling of the other (cf. Propsition 3.22 in [5]).

The intuitive reason for the difference in the expected utilities is that the channel κ_2 transmits one of the states without noise and the other state with noise. With a convenient pre-processing, it is possible to make sure that the relevant information for choosing an action and for optimizing expected utility is transmitted with less noise.

Note the symmetry of the example: each of the two channels arises from the other by a convenient pre-processing, since the pre-processing is invertible. Hence, the two channels are not comparable by the Blackwell order. In contrast, two channels that only differ by an invertible garbling of the output are equivalent with respect to the Blackwell order.

The pre-garbling in Example 1 is invertible, and so it is more aptly described as a pre-processing. In general, though, pure pre-garbling and pure pre-processing are not easily distinguishable, and it is easy to perturb Example 1 by adding noise without changing the conclusion. In Section 3, we will present an example in which the pre-garbling consists of coarse-graining. It is much more difficult to understand how coarse-graining can be used as sensible pre-processing.

Even though pre-garbling can make a channel better (or, more precisely, more suited for a particular decision problem at hand), pre-garbling cannot invert the Blackwell order:

Lemma 1. *If* $\kappa_1 < \kappa_2 \cdot \lambda$*, then* $\kappa_1 \not\geq \kappa_2$*.*

Proof. Suppose that $\kappa_1 < \kappa_2 \cdot \lambda$. Then the capacity of κ_1 is less than the capacity of $\kappa_2 \cdot \lambda$, which is bounded by the capacity of κ_2. Therefore, the capacity of κ_1 is less than the capacity of κ_2. □

Additionally, it follows directly from Blackwell's theorem that

$$\kappa_1 \leq \kappa_2 \text{ implies } \kappa_1 \cdot \lambda \leq \kappa_2 \cdot \lambda$$

for any channel λ, where the input and output alphabets of λ equal the input alphabet of κ_1, κ_2. Thus, pre-garbling preserves the Blackwell order when applied to both channels simultaneously.

Finally, let us remark that certain kinds of simultaneous pre-garbling can also be "hidden" in the utility function; namely, in Blackwell's theorem, it is not necessary to vary the distribution of $\mathcal{S}$ as long as the support of the (fixed) input distribution has full support S (that is, every state of the input alphabet of κ_1 and κ_2 appears with positive probability). In this setting, it suffices to look only at different utility functions. When the input distribution is fixed, it is more convenient to think in terms of random variables instead of channels, which slightly changes the interpretation of the decision problem. Suppose we are given random variables S, X_1, X_2 and a utility function $u(a,s)$ depending on the value of S and an action $a \in \mathcal{A}$ as above. If we cannot look at both X_1 and X_2, should we choose to look at X_1 or at X_2 to make our decision?

Theorem 2. *(Blackwell's theorem for random variables [7]) The following two conditions are equivalent:*

1. *Under the optimal decision rule, when the agent chooses X_2, her expected utility is always at least as large as the expected utility when she chooses X_1, independent of the utility function.*
2. $(X_1 \leftarrow S) \leq (X_2 \leftarrow S)$.

3. Pre-Garbling by Coarse-Graining

In this section we present a few counter-intuitive properties of the Blackwell order.

Proposition 1. *There exist random variables S, X_1, X_2 and a function $f : \mathcal{S} \to \mathcal{S}$ from the support of S to a finite set S' such that the following holds:*

1. *S and X_1 are independent given $f(S)$.*
2. $(X_1 \leftarrow f(S)) \ < \ (X_2 \leftarrow f(S))$.
3. $(X_1 \leftarrow S) \ \not\leq \ (X_2 \leftarrow S)$.

This result may at first seem paradoxical. After all, property (3) implies that there exists a decision problem involving S for which it is better to use X_1 than X_2. Property (1) implies that any information that X_1 has about S is contained in X_1's information about $f(S)$. One would therefore expect that, from the viewpoint of X_1, any decision problem in which the task is to predict S and to react on S looks like a decision problem in which the task is to react to $f(S)$. But property (2) implies that for such a decision problem, it may in fact be better to look at X_2.

Proof of Proposition 1. The proof is by Example 2, which will be given in Section 4. This example satisfies

1. S and X_1 are independent given $f(S)$.
2. $(X_1 \leftarrow f(S)) \ \leq \ (X_2 \leftarrow f(S))$.
3. $(X_1 \leftarrow S) \ \not\leq \ (X_2 \leftarrow S)$.

It remains to show that it is also possible to achieve the strict relation $(X_1 \leftarrow f(S)) < (X_2 \leftarrow f(S))$ in the second statement. This can easily be done by adding a small garbling to the channel $X_1 \leftarrow f(S)$ (e.g., by adding a binary symmetric channel with sufficiently small noise parameter ϵ). This ensures $(X_1 \leftarrow f(S)) < (X_2 \leftarrow f(S))$, and if the garbling is small enough, this does not destroy the property $(X_1 \leftarrow S) \ \not\leq \ (X_2 \leftarrow S)$. □

The example from Proposition 1 also leads to the following paradoxical property:

Proposition 2. *There exist random variables S, X and there exists a function $f : \mathcal{S} \to \mathcal{S}'$ from the support of S to a finite set S' such that the following holds:*

$$(X \leftarrow f(S)) \cdot (f(S) \leftarrow S) \ \not\leq \ X \leftarrow S.$$

Let us again give a heuristic argument for why we find this property paradoxical. Namely, the combined channel $(X \leftarrow f(S)) \cdot (f(S) \leftarrow S)$ can be seen as a Markov chain approximation of the direct channel $X \leftarrow S$ that corresponds to replacing the conditional distribution

$$P_{X|S}(x|s) = \sum_{f(s)} P_{X|Sf(S)}(x|s, f(s)) P_{f(S)|S}(f(s)|s).$$

by

$$\sum_{f(s)} P_{X|f(S)}(x|f(s)) P_{f(S)|S}(f(s)|s).$$

Proposition 2 together with Blackwell's theorem states that there exist situations where this approximation is better than the correct channel.

Proof of Proposition 2. Let S, X_1, X_2 be as in Example 2 in Section 4 that also proves Proposition 1, and let $X = X_2$. In that example, the two channels $X_1 \leftarrow f(S)$ and $X_2 \leftarrow f(S)$ are equal. Moreover, X_1 and S are independent given $f(S)$. Thus, $(X \leftarrow f(S)) \cdot (f(S) \leftarrow S) = (X_1 \leftarrow S)$. Therefore, the statement follows from $(X_1 \leftarrow S) \not\preceq (X_2 \leftarrow S)$. □

On the other hand, the channel $(X \leftarrow f(S)) \cdot (f(S) \leftarrow S)$ is always less capable than $X \leftarrow S$:

Lemma 2. *For any random variables S, X, and function $f : \mathcal{S} \to \mathcal{S}$, the channel $(X \leftarrow f(S)) \cdot (f(S) \leftarrow S)$ is less capable than $X \leftarrow S$.*

Proof. For any distribution of S, let X' be the output of the channel $(X \leftarrow f(S)) \cdot (f(S) \leftarrow S)$. Then, X' is independent of S given $f(S)$. On the other hand, since f is a deterministic function, X' is independent of $f(S)$ given S. Together, this implies $I(S; X') = I(f(S); X')$. Using the fact that the joint distributions of $(X, f(S))$ and $(X', f(S))$ are identical and applying the data processing inequality gives

$$I(S; X') = I(f(S); X') = I(f(S); X) \leq I(S; X). \quad \square$$

The setting of Proposition 2 can also be understood as a specific kind of pre-garbling. Namely, consider the channel λ^f defined by

$$\lambda^f_{s',s} := P_{S|f(S)}(s'|f(s)).$$

The effect of this channel can be characterized as a randomization of the input: the precise value of S is forgotten, and only the value of $f(S)$ is preserved. Then, a new value s' is sampled for S according to the conditional distribution of S given $f(S)$.

Lemma 3. $(X \leftarrow f(S)) \cdot (f(S) \leftarrow S) = (X \leftarrow S) \cdot \lambda^f$.

Proof. $\displaystyle\sum_{s_1} P_{X|S}(x|s_1) P_{S|f(S)}(s_1|f(s)) = \sum_{s_1,t} P_{X|S}(x|s_1) P_{S|f(S)}(s_1|t) P_{f(S)|S}(t|s)$

$$= \sum_{t} P_{X|f(S)}(x|t) P_{f(S)|S}(t|s),$$

where we have used that $X - S - f(S)$ forms a Markov chain. □

While it is easy to understand that pre-garbling can be advantageous in general (since it can work as preprocessing), we find it surprising that this can also happen in the case where the pre-garbling is done in terms of a function f; that is, in terms of a channel λ^f that does coarse-graining.

4. Examples

Example 2. *Consider the joint distribution*

$f(s)$	s	x_1	x_2	$P_{f(S)SX_1X_2}$
0	0	0	0	1/4
0	1	0	1	1/4
0	0	1	0	1/8
0	1	1	0	1/8
1	2	1	1	1/4

and the function f: {0, 1, 2} → {0, 1} with f(0) = f(1) = 0 and f(2) = 1. Then, X_1 and X_2 are independent uniform binary random variables, and $f(S) = \text{AND}(X_1, X_2)$. By symmetry, the joint distributions of the pairs

$(f(S), X_1)$ *and* $(f(S), X_2)$ *are identical, and so the two channels* $X_1 \leftarrow f(S)$ *and* $X_2 \leftarrow f(S)$ *are identical. In particular,* $(X_1 \leftarrow f(S)) \leq (X_2 \leftarrow f(S))$.

On the other hand, consider the utility function

s	a	$u(s,a)$
0	0	0
0	1	0
1	0	1
1	1	0
2	0	0
2	1	1

To compute the optimal decision rule, let us look at the conditional distributions:

s	x_1	$P_{S\|X_1}(s\|x_1)$
0	0	1/2
1	0	1/2
0	1	1/4
1	1	1/4
2	1	1/2

s	x_2	$P_{S\|X_2}(s\|x_2)$
0	0	3/4
1	0	1/4
0	1	0
1	1	1/2
2	1	1/2

The optimal decision rule forX_1 *is a(0) = 0, a(1) = 1, with expected utility*

$$u_{X_1} := 1/2 \cdot 1/2 + 1/2 \cdot 1/2 = 1/2.$$

The optimal decision rule for X_2 *is a(0) = 0, a(1) ∈ {0, 1} (this is not unique in this case), with expected utility*

$$u_{X_2} := 1/2 \cdot 1/4 + 1/2 \cdot 1/2 = 3/8 < 1/2.$$

How can we understand this example? Some observations:

- It is easy to see that X_2 has more irrelevant information than X_1: namely, X_2 can determine relatively precisely when $S = 0$. However, since $S = 0$ gives no utility independent of the action, this information is not relevant. It is more difficult to understand why X_2 has less relevant information than X_1. Surprisingly, X_1 can determine more precisely when $S = 1$: if $S = 1$, then X_1 "detects this" (in the sense that X_1 chooses action 0) with probability $2/3$. For X_2, the same probability is only $1/3$.
- The conditional entropies of S given X_2 are smaller than the conditional entropies of S given X_1:

$$\begin{aligned} &H(S|X_1 = 0) = \log(2), && H(S|X_1 = 1) = \tfrac{3}{2}\log(2), \\ &H(S|X_2 = 0) = 2\log(2) - \tfrac{3}{2}\log(3) \approx 0.4150375\log(2), && H(S|X_2 = 1) = \log(2). \end{aligned}$$

- One can see in which sense $f(S)$ captures the relevant information for X_1, and indeed for the whole decision problem: knowing $f(S)$ is completely sufficient in order to receive the maximal utility for each state of S. However, when information is incomplete, it matters how the information about the different states of S is mixed, and two variables X_1, X_2 that have the same joint distribution with $f(S)$ may perform differently. It is somewhat surprising that it is the random variable that has less information about S and that is conditionally independent of S given $f(S)$ which actually performs better.

Example 2 is different from the pre-garbling Example 1 discussed in Section 2. In the latter, both channels had the same amount of information (mutual information) about S, but for the given decision problem the information provided by κ_2 was more relevant than the information provided by κ_1. The first difference in Example 2 is that X_1 has less mutual information about S than X_2 (Lemma 2). Moreover, both channels are identical with respect to $f(S)$; i.e., they provide the same information about $f(S)$, and for X_1 it is the only information it has about S. So, one could argue that X_2 has additional information that does not help, but decreases the expected utility instead.

We give another example which shows that X_2 can also be chosen as a deterministic function of S.

Example 3. *Consider the joint distribution*

$f(s)$	s	x_1	x_2	$P_{f(S)SX_1X_2}$
0	0	0	0	1/6
0	0	1	0	1/6
0	1	0	1	1/6
0	1	1	1	1/6
1	2	1	1	1/3

The function f is as above, but now also X_2 is a function of S. Again, the two channels $X_1 \leftarrow f(S)$ and $X_2 \leftarrow f(S)$ are identical, and X_1 is independent of S given $f(S)$. Consider the utility function

s	a	$u(s,a)$
0	0	0
0	1	0
1	0	0
1	1	1
2	0	0
2	1	-1

One can show that it is optimal for an agent who relies on X_2 to always choose action 0, which brings no reward (and no loss). However, when the agent knows that X_1 is zero, he may safely choose action 1 and has a positive probability of receiving a positive reward.

To add another interpretation to the last example, we visualize the situation in the following Bayesian network:

$$X \leftarrow S \rightarrow f(S) \rightarrow X',$$

where, as in Proposition 2 and its proof, we let $X = X_2$, and we consider $X' = X_1$ as an approximation of X. Then, S denotes the state of the system that we are interested in, and X denotes a given set of observables of interest. $f(S)$ can be considered as a "proxy" in situations where it is difficult to observe X directly. For example, in neuroimaging, instead of directly measuring the neural activity X, one might look at an MRI signal $f(S)$. In economic and social sciences, monetary measures like the GDP are used as a proxy for prosperity.

A decision problem can always be considered as a classification problem defined by the utility $u(s,a)$ by considering the optimal action as the class label of state S. Proposition 2 now says that there exist S, X, $f(S)$, and a classification problem $u(s,a)$, such that the approximated features X' (simulated from $f(S)$) allow for a better classification (higher utility) than the original features X.

In such a situation, looking at $f(S)$ will always be better than looking at either X or X'. Thus, the paradox will only play a role in situations where it is not possible to base the decision on $f(S)$ directly. For example, $f(S)$ might still be too large, or X might have a more natural interpretation, making it easier to interpret for the decision taker. However, when it is better to base a decision on a proxy rather than directly on the observable of interest, this interpretation may be erroneous.

5. Information Decomposition and Le Cam Deficiency

Given two channels κ_1, κ_2, how can one decide whether or not $\kappa_1 \leq \kappa_2$? The easiest way is to check whether the equation $\kappa_1 = \lambda \cdot \kappa_2$ has a solution λ that is a stochastic matrix. In the finite alphabet case, this amounts to checking the feasibility of a linear program, which is considered computationally easy. However, when the feasibility check returns a negative result, this approach does not give any more information (e.g., how far κ_1 is away from being a garbling of κ_2). A function that quantifies how far κ_1 is from being a garbling of κ_2 is given by the *(Le Cam) deficiency* and its various generalizations [9]. Another such function is given by UI defined in [7] that accounts for the fact that the channels we consider are of the form $\kappa_1 = (X_1 \leftarrow S)$ and $\kappa_2 = (X_2 \leftarrow S)$; that is, they are derived from conditional distributions of random variables. In contrast to the deficiencies, UI depends on the input distribution to these channels.

Let $P_{SX_1X_2}$ be a joint distribution of S and the outputs X_1 and X_2. Let Δ_P be the set of all joint distributions of the random variables S, X_1, X_2 (with the same alphabets) that are compatible with the marginal distributions of $P_{SX_1X_2}$ for the pairs (S, X_1) and (S, X_2); i.e.,

$$\Delta_P := \left\{ Q_{SX_1X_2} \in \Delta : Q_{SX_1} = P_{SX_1}, Q_{SX_2} = P_{SX_2} \right\}.$$

In other words, Δ_P consists of all joint distributions that are compatible with κ_1 and κ_2 and that have the same distribution for S as $P_{SX_1X_2}$. Consider the function

$$UI(S; X_1 \backslash X_2) := \min_{Q \in \Delta_P} I_Q(S; X_1 | X_2),$$

where I_Q denotes the conditional mutual information evaluated with respect to the the joint distribution Q. This function has the following property: $UI(S; X_1 \backslash X_2) = 0$ if and only if $\kappa_1 \leq \kappa_2$ [7]. Computing UI is a convex optimization problem. However, the condition number can be very bad, which makes the problem difficult in practice.

UI is interpreted in [7] as a measure of the *unique* information that X_1 conveys about S (with respect to X_2). So, for instance, with this interpretation Example 2 can be summarized as follows: neither X_1 nor X_2 has unique information about $f(S)$. However, both variables have unique information about S, although X_1 is conditionally independent of S given $f(S)$ and thus—in contrast to X_2—contains no "additional" information about S. We now apply UI to a parameterized version of the AND gate in Example 2.

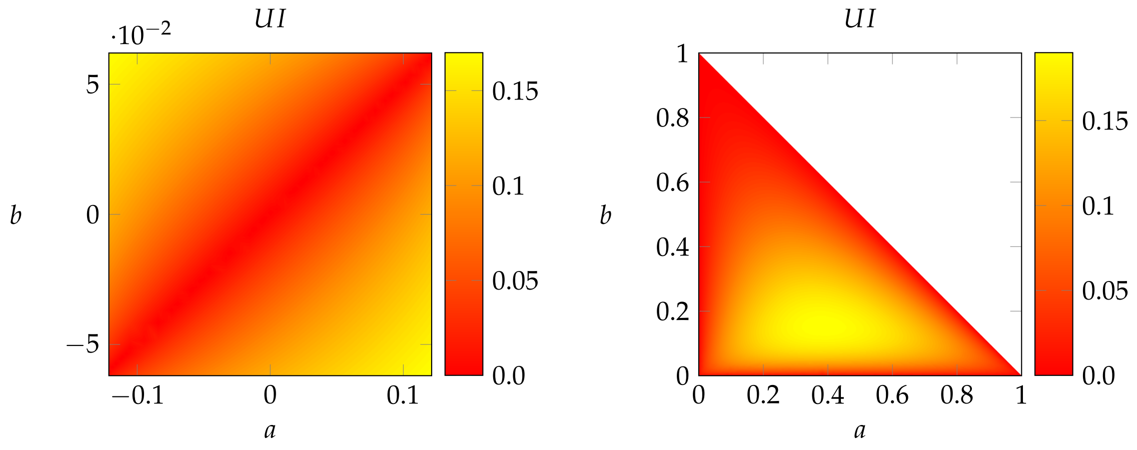
Example 4. *Figure 1a shows a heat map of UI computed on the set of all distributions of the form*

$f(s)$	s	x_1	x_2	$P_{f(S)SX_1X_2}$
0	0	0	0	$1/8 + 2b$
0	1	0	0	$1/8 - 2b$
0	0	0	1	$1/8 + a$
0	1	0	1	$1/8 - a$
0	0	1	0	$1/8 + a/2 + b$
0	1	1	0	$1/8 - a/2 - b$
1	2	1	1	$1/4$

where $-1/8 \leq a \leq 1/8$ and $-1/16 \leq b \leq 1/16$. This is the set of distributions of S, X_1, X_2 that satisfy the following constraints:

1. *X_1, X_2 are independent;*
2. *$f(S) = \text{AND}(X_1, X_2)$, where f is as in Example 2; and*
3. *X_1 is independent of S given $f(S)$.*

Along the secondary diagonal $b = a/2$, the marginal distributions of the pairs (S, X_1) and (S, X_2) are identical. In such a situation, the channels $(X_1 \leftarrow S)$ and $(X_2 \leftarrow S)$ are Blackwell-equivalent, and so UI vanishes. Further away from the diagonal, the marginal distributions differ, and UI grows. The maximum value is achieved at the corners for $(a, b) = (-1/8, 1/16), (1/8, -1/16)$. At the upper left corner $(a, b) = \pm(-1/8, 1/16)$, we recover Example 2.



(**a**)

(**b**)

Figure 1. Heatmaps for the function UI in (**a**) Example 4, and (**b**) Example 5.

Example 5. *Figure 1b shows a heat map of UI computed on the set of all distributions of the form*

$f(s)$	s	x_1	x_2	$P_{f(S)SX_1X_2}$
0	*0*	*0*	*0*	$a^2/(a+b)$
0	*0*	*1*	*0*	$ab/(a+b)$
0	*1*	*0*	*1*	$ab/(a+b)$
0	*1*	*1*	*1*	$b^2/(a+b)$
1	*2*	*1*	*1*	*1 - a - b*

where $a, b \geq 0$ *and* $a + b \leq 1$*. This extends Example 3, which is recovered for* $a = b = 1/3$*. This is the set of distributions of* S, X_1, X_2 *that satisfy the following constraints:*

1. X_2 *is a function of S, where the function is as in Example 3.*
2. X_1 *is independent of S given* $f(S)$*.*
3. *The channels* $X_1 \leftarrow f(S)$ *and* $X_2 \leftarrow f(S)$ *are identical.*

Acknowledgments: We thank the participants of the PID workshop at FIAS in Frankfurt in December 2016 for many stimulating discussions on this subject. Nils Bertschinger thanks h.c. Maucher for funding his position.

Author Contributions: The research was initiated by J.R. and carried out by all authors. Computer experiments to find and analyze the examples were done by P.K.B. D.W. simplified Example 1. J.J. and N.B. added interpretation. The manuscript was written by J.R., P.K.B., E.O. and D.W. Nobody played the synthesizer. All authors have read and approved the final manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Blackwell, D. Equivalent Comparisons of Experiments. *Ann. Math. Stat.* **1953**, *24*, 265–272.
2. Torgersen, E. *Comparison of Statistical Experiments*; Cambridge University Press: New York, NY, USA, 1991.
3. Le Cam, L. Comparison of Experiments—A Short Review. *Stat. Probab. Game Theory* **1996**, *30*, 127–138.
4. Bergmans, P. Random coding theorem for broadcast channels with degraded components. *IEEE Trans. Inf. Theory* **1973**, *19*, 197–207.
5. Cohen, J.; Kemperman, J.; Zbăganu, G. *Comparisons of Stochastic Matrices with Applications in Information Theory, Statistics, Economics, and Population Sciences*; Birkhäuser: Boston, MA, USA, 1998.
6. Körner, J.; Marton, K. Comparison of two noisy channels. In *Topics in Information Theory*; Colloquia Mathematica Societatis János Bolyai: Keszthely, Hungary, 1975; Volume 16, pp. 411–423.
7. Bertschinger, N.; Rauh, J.; Olbrich, E.; Jost, J.; Ay, N. Quantifying unique information. *Entropy* **2014**, *16*, 2161–2183.

8. Rauh, J.; Banerjee, P.K.; Olbrich, E.; Jost, J.; Bertschinger, N. On extractable shared information. *arXiv* **2017**, arXiv:1701.07805.
9. Raginsky, M. Shannon meets Blackwell and Le Cam: Channels, codes, and statistical experiments. In Proceedings of the 2011 IEEE International Symposium on Information Theory Proceedings, St. Petersburg, Russia, 31 July–5 August 2011; pp. 1220–1224.