Coarse-graining and the Blackwell order

Johannes Rauh*, Pradeep Kr. Banerjee*, Eckehard Olbrich*, Jürgen Jost‡, Nils Bertschinger† and David Wolpert†§
*Max Planck Institute for Mathematics in the Sciences, Leipzig, Germany
†Frankfurt Institute for Advanced Studies, Frankfurt, Germany
‡Santa Fe Institute, Santa Fe, NM, USA
§MIT, Cambridge, MA, USA
dhw@santafe.edu

Abstract—Suppose we have a pair of information channels, κ₁, κ₂ with a common input. The Blackwell order is a partial order over channels that compares κ₁ and κ₂ by the maximal expected utility an agent can obtain when decisions are based on the outputs of κ₁ and κ₂. Equivalently, κ₁ is said to be Blackwell-inferior to κ₂ if and only if κ₁ can be constructed by garbling the output of κ₂. A related partial order stipulates that κ₂ is more capable than κ₁ if the mutual information between the input and output is larger for κ₂ than for κ₁ for any distribution over inputs. If one channel is Blackwell-inferior to another then it must be less capable. However examples are known where κ₁ is less capable than κ₂ even though it is not Blackwell-inferior.

We give a new example of this phenomenon in which κ₁ is constructed by coarse-graining the inputs of κ₂. 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 is surprising, as one might think that coarse-graining can only destroy information and lead to inferior channels.

I. 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 denote the two channels with κ₁ and κ₂, respectively. Then if there exists another stochastic matrix λ such that κ₁ = λ · κ₂, there is never any reason to strictly prefer κ₁. This is because if we choose κ₂, we can always make our decision by chaining the output of κ₂ through the channel λ and then using the same decision function we would have used had we chosen κ₁. This simple argument shows that whatever the three stochastic matrices are and whatever the decision rule we would use if we chose channel κ₁, we can always get the same expected utility (as we would get by choosing κ₁ and using the associated decision rule) by instead choosing channel κ₂ (for the appropriate associated decision rule). In this kind of situation, where κ₁ = λ · κ₂, we say that κ₁ is a garbling (or degradation) of κ₂. It is much more difficult to prove that the converse also holds true:

Theorem 1 (Blackwell’s theorem [11]). Let κ₁, κ₂ be two stochastic matrices representing two channels with the same input alphabet. Then the following two conditions are equivalent:

1) When the agent chooses κ₂ (and uses the decision rule that is optimal for κ₂), her expected utility is always at least as big as the expected utility when she chooses κ₁ (and uses the optimal decision rule for κ₁), independent of the utility function and the distribution of the input S.

2) κ₁ is a garbling of κ₂.

Blackwell’s theorem motivates looking at the following partial order over channels κ₁, κ₂ with a common input alphabet:

κ₁ ≤ κ₂ :⇔ one of the two statements
in Blackwell’s theorem holds true.

We call this partial order the Blackwell order. If κ₁ ≤ κ₂, then κ₁ is said to be Blackwell-inferior to κ₂. Strictly speaking, the Blackwell order is only a preorder, since there are channels κ₁ ≠ κ₂ that satisfy κ₁ ≤ κ₂ ≤ κ₁. However, for our purposes such channels can usually be considered as equivalent.

For a given distribution of S, we can also compare κ₁ and κ₂ by comparing the two mutual informations I(S; X₁), I(S; X₂) between the common input S and the channel outputs X₁ and X₂. The data processing inequality shows that κ₂ ≥ κ₁ implies I(S; X₂) ≥ I(S; X₁). 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 κ₁ or κ₂ preserve is relevant for a decision problem at hand.

Given two channels κ₁, κ₂, suppose that I(S; X₂) ≥ I(S; X₁) for all distributions of S. In this case, we say...
that $\kappa_2$ is more capable than $\kappa_1$. Conversely, $\kappa_1$ is less capable than $\kappa_2$. Does this imply that $\kappa_1 \preceq \kappa_2$? Surprisingly, specific examples are known where the answer is negative \[2\]. In Proposition \[6\] we introduce a new example of this phenomenon. In this example, $\kappa_1$ is a Markov approximation of $\kappa_2$ by a deterministic function, in the following sense: Consider a function $f : S \to S'$ from the support $S := \{s : P_S(s) = \Pr(S = s) > 0\}$ of $S$ to another set $S'$. Given two random variables $S, X$ denote by $X \leftarrow S$ the channel defined by the conditional probabilities $P_{X|S}(x|s)$ on $S$. Consider the two channels $\kappa_2 := (X \leftarrow S)$ and $\kappa_1 := (X \leftarrow f(S)) \cdot (f(S) \leftarrow S)$. Which channel is superior? Using the data processing inequality, it is easy to see that $\kappa_1$ is less capable than $\kappa_2$. However, as Proposition \[6\] shows, in general $\kappa_1 \not\preceq \kappa_2$.

We call $\kappa_1$ a Markov approximation, because the output of $\kappa_1$ is independent of the input $S$ given $f(S)$. $\kappa_1$ can also be obtained from $\kappa_2$ by “pre-garbling” (Lemma \[8\]): that is, $\kappa_1$ is obtained by composing another channel $\lambda'$ with $\kappa_2$. It is known that pre-garbling may improve the performance of a channel (but not its capacity!) as we recall in Section \[11\].

What may be surprising is that this can happen for pre-garbling distributions of the form $\lambda'$, 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 may not reduce Blackwell information.

Proposition \[6\] builds upon another effect that we find paradoxical: Namely, there exist random variables $S, X_1, X_2$ and there exists a function $f : S \to S'$ 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) $(X_1 \leftarrow f(S)) \not\preceq (X_2 \leftarrow f(S))$.
3) $(X_1 \leftarrow S) \preceq (X_2 \leftarrow S)$.

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.”

The remainder of this work is organized as follows: In Section \[11\] 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 \[13\] we state a few properties of the Blackwell order, and we explain why we find these properties counter-intuitive and paradoxical. In particular, we also show that coarse-graining the input can improve the performance of a channel. Section \[14\] contains a detailed discussion of an example that illustrates these properties. In the appendix, we briefly remark how to compute the Blackwell order and how to quantify deviations from the same.

II. PRE-GARBLING

As discussed above, Blackwell’s theorem states that 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 quality of a channel, as the following example shows.

Example 2. Suppose that an agent can choose an action from a finite set $A$. She then receives a utility $u(a, s)$ that depends both on the chosen action $a \in A$ and on the value $s$ of a random variable $S$. Consider the channels

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

and the utility function

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

For uniform input the optimal decision rule for $\kappa_1$ is

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

and the opposite

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

for $\kappa_2$. The expected utility with $\kappa_1$ is 1.4, while using $\kappa_2$, it is slightly higher, 1.45.

The intuitive reason for the difference in the expected utilities is that the channel $\kappa_2$ 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 which action has better 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, and 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 only with respect to the Blackwell order.

The pre-garbling in Example \[2\] 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 \[2\] by adding noise without changing the conclusion. In Section \[11\] 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 3. If $\kappa_1 \prec \kappa_2 \cdot \lambda$, then $\kappa_1 \npreceq \kappa_2$.

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

Also, 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 $\lambda$, where the input and output alphabets of $\lambda$ equal the input alphabet of $\kappa_1, \kappa_2$. 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 $S$, as long as the (fixed) input distribution has full support $S$ (that is, every state of the input alphabet of $\kappa_1$ and $\kappa_2$ 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 $a \in A$ as above. If we cannot look at both $X_1$ and $X_2$, should we rather look at $X_1$ or at $X_2$ to take our decision?

**Theorem 4** (Blackwell’s theorem for random variables). The following two conditions are equivalent:

1) Under the optimal decision rule, when the agent chooses $X_2$, her expected utility is at least as big as the expected utility when she chooses $X_1$, independent of the utility function.

2) $(X_1 \leftarrow S) \preceq (X_2 \leftarrow S)$.

**III. PRE-GARBLING BY COARSE-GRANING**

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

**Proposition 5.** There exist random variables $S, X_1, X_2$ and a function $f : S \to S'$ 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) $(X_1 \leftarrow f(S)) \prec (X_2 \leftarrow f(S))$.

3) $(X_1 \leftarrow S) \not\preceq (X_2 \leftarrow S)$.

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 $X_2$. Now, 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 is better to look at $X_2$.

**Proof of Proposition 5** The proof is by Example 9 which will be given in Section IV. This example satisfies

1) $S$ and $X_1$ are independent given $f(S)$.

2) $(X_1 \leftarrow f(S)) \preceq (X_2 \leftarrow f(S))$.

3) $(X_1 \leftarrow S) \not\preceq (X_2 \leftarrow S)$.

It only remains to show that it is possible to also achieve the strict relation $(X_1 \leftarrow f(S)) \prec (X_2 \leftarrow f(S))$ 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 $\epsilon$). This ensures $(X_1 \leftarrow f(S)) \prec (X_2 \leftarrow f(S))$, and if the garbling is small enough, this does not destroy the property $(X_1 \leftarrow S) \not\preceq (X_2 \leftarrow S)$.

The example from Proposition 5 also leads to the following paradoxical property:

**Proposition 6.** There exist random variables $S, X$ and there exists a function $f : S \to S'$ from the support of $S$ to a finite set $S'$ such that the following holds:

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

Let us again give a heuristic argument 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

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

by

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

Proposition 6 together with Blackwell’s theorem states that there are situations where this approximation is better than the direct channel.

**Proof of Proposition 6** Let $S, X_1, X_2$ be as in Example 9 in Section IV that also proves Proposition 5 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 less capable than $X \leftarrow S$:

**Lemma 7.** For any random variables $S, X$, and function $f : S \to S$, 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 again gives

$$I(S; X') = I(f(S); X') = I(f(S); X) \leq I(S; X).$$
The setting of Proposition \[\text{I}\] can also be understood as a specific kind of pre-garbling. Namely, consider the channel \(\lambda_f\) defined by

\[
\lambda_{f,s} = P_{S|f(S)}(s' | f(s)).
\]

The effect of this channel is the following: 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 8.** \((X \leftarrow f(S)) : (f(S) \leftarrow S) = (X \leftarrow S) \cdot \lambda_f\).

**Proof.**

\[
\sum_{s_1} P_{X|S}(x|s_1)P_{S|f(S)}(s_1 | f(s)) = \sum_{s_1,t} P_{X|S}(x|s_1)P_{S|f(S)}(s_1 | f(s))P_{f(S)|S}(t|s) = \sum_{t} P_{X|f(S)}(x|t)P_{f(S)|S}(t|s),
\]

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 \(\lambda_f\) that does coarse-graining.

### IV. Examples

**Example 9.** Consider the joint distribution

| \(s\) | \(x_1\) | \(x_2\) | \(P_{f(S)|S,X_1,X_2}\) |
|-------|--------|--------|------------------|
| 0     | 0      | 0      | 0.25             |
| 0     | 1      | 0      | 0.25             |
| 0     | 0      | 1      | 0.25             |
| 0     | 1      | 1      | 0.25             |
| 1     | 0      | 0      | 0.25             |
| 1     | 1      | 0      | 0.25             |
| 1     | 0      | 1      | 0.25             |
| 1     | 1      | 1      | 0.25             |

and the function \(f: \{0,1,2\} \rightarrow \{0,1\}\) 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 \(X_1 \leftarrow f(S) \preceq X_2 \leftarrow f(S)\).

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     | 0      | 0                |
| 2     | 1      | 1/2              |

The optimal decision rule for \(X_1\) 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 \(a(0) = 0, a(1) \in \{0,1\}\) (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.
- \(H(S|X_1 = 0) = \log(2), H(S|X_1 = 1) = \frac{3}{2} \log(2), H(S|X_2 = 0) = 2 \log(2) - \frac{3}{2} \log(3) \approx 0.4150375 \log(2), H(S|X_2 = 1) = \log(2)\). Thus, the conditional entropies of \(S\) given \(X_2\) are smaller than the conditional entropies of \(S\) given \(X_1\).
- 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 9 is different from the pre-garbling Example 2 discussed in Section III. 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 \(\kappa_2\) was more relevant than the information provide by \(\kappa_1\). The first difference in Example 9 is that \(X_1\) has less information about \(S\) than \(X_2\) (Lemma 7). Moreover, both channels are identical with respect to \(f(S)\), i.e. they are providing 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 though, but decreases the expected utility instead.

To visualize how far the channel \(X_1 \leftarrow S\) is from being a garbling of the channel \(X_2 \leftarrow S\), we use a function \(UI\)
The definition of $UI$ is briefly outlined in the Appendix. $UI$ is a nonnegative continuous function on the set of all joint distributions of $(S, X_1, X_2)$. $UI$ can be construed as the unique information that $X_1$ conveys about $S$ (with respect to $X_2$) [3]. So, for instance, in Example 9, while both channels have no unique information about $f(S)$, “adding information” about $S$ in one channel leads to having unique information about $S$ in both channels. We now apply our measure to a parameterized version of the AND gate in Example 9.

**Example 10.** Figure 1 shows a heat map of $UI$ computed on the set of all distributions $P_{SX_1X_2}$, the $(S, X_1, X_2)$-marginal of joint distributions of the form

$$f(s) = \begin{bmatrix} s & x_1 & x_2 \end{bmatrix} P_{f(S)SX_1X_2} = \begin{bmatrix} \frac{1}{8} + 2b & \frac{1}{8} - 2b \\ \frac{1}{8} + a & \frac{1}{8} - a \\ \frac{1}{8} + a/2 + b & \frac{1}{8} - a/2 - b \\ \frac{1}{4} \end{bmatrix}$$

where $-1/8 \leq a \leq 1/8$ and $-1/16 \leq b \leq 1/16$. One can show that this is the set of distributions of $S, X_1, X_2$ such that $f(S) = \text{AND}(X_1, X_2)$, where $f$ is as in Example 9. For $(a, b) = (-1/8, 1/16)$, we recover Example 9.

To generate the heat map in Figure 1, we randomly sampled a distribution $P_{SX_1X_2}$ from the set of all distributions of $(S, X_1, X_2)$ that are compatible with the chosen function $f$, with $P_{f(S)SX_1X_2}$, and which satisfies the condition $S - f(S) - X_1$. We then computed $UI$ to check whether $(X_1 \leftarrow S) \not\leq (X_2 \leftarrow S)$ holds or not. $UI$ vanishes along the secondary diagonal $(b = a/2)$ when the marginal distributions of the pairs $(S, X_1)$ and $(S, X_2)$ are identical. For all other tuples $(a, b)$, $UI$ is nonvanishing and achieves its maximum at the corners of the main diagonal for $(a, b) = (-1/8, 1/16)$, $(1/8, -1/16)$.

**REFERENCES**

[1] D. Blackwell, “Equivalent comparisons of experiments,” *The Annals of Mathematical Statistics*, vol. 24, no. 2, pp. 265–272, 1953.

[2] J. Körner and K. Marton, “Comparison of two noisy channels,” in *Topics in information theory*. Keszthely (Hungary): Colloquium Mathematica Societatis János Bolyai, 1975, vol. 16, pp. 411–423.

[3] N. Bertschinger, J. Rauh, E. Olbrich, J. Jost, and N. Ay, “Quantifying unique information,” *Entropy*, vol. 16, no. 4, pp. 2161–2183, 2014.

**APPENDIX**

**COMPUTING THE BLACKWELL ORDER**

Given two channels $\kappa_1, \kappa_2$, how can one decide whether or not $\kappa_1 \preceq \kappa_2$? The easiest way is to check whether the equation $\kappa_1 = \lambda \cdot \kappa_2$ has a solution $\lambda$ that is a stochastic matrix. In the finite alphabet case, this amounts to checking 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 $\kappa_1$ is away from being a garbling of $\kappa_2$. A measure for this can be constructed as follows: Let $P_{SX_1X_2}$ be a joint distribution of $S$ and the outputs $X_1$ and $X_2$ of $\kappa_1$ and $\kappa_2$ satisfying the constraints that $X_1 \leftarrow S = \kappa_1$ and $X_2 \leftarrow S = \kappa_2$ and that the support of $S$ equals the input alphabet $A$ of $\kappa_1$ and $\kappa_2$. Let $\Delta_P$ 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., $\Delta_P := \{Q_{SX_1X_2} \in \Delta : Q_{SX_1} = P_{SX_1}, Q_{SX_2} = P_{SX_2}, \forall (S, X_1, X_2) \in \Delta\}$. In other words, $\Delta_P$ consists of all joint distributions that are compatible with $\kappa_1$ and $\kappa_2$ and that have the same distribution for $S$ as $P_{SX_1X_2}$. Consider the function

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

where $I_Q$ denotes the conditional mutual information evaluated with respect to the the joint distribution $Q$. This function has the following property: $UI(S; X_1 \mid X_2) = 0$ if and only if $\kappa_1 \preceq \kappa_2$ [3]. Note that $UI$ depends not only on the two channels $\kappa_1, \kappa_2$, but also on the choice of the input distribution. Computing $UI$ is a convex optimization problem. However, the condition number can be very bad, which makes the problem difficult in practice.