Directed Data-Processing Inequalities for Systems with Feedback

Derpich, Milan; Østergaard, Jan

Published in:
Entropy

DOI (link to publication from Publisher):
10.3390/e23050533

Publication date:
2021

Document Version
Publisher's PDF, also known as Version of record

entropy

MDPI

Article

Directed Data-Processing Inequalities for Systems with Feedback

Milan S. Derpich [1,*] **and Jan Østergaard** [2,*]

[1] Department of Electronic Engineering, Universidad Técnica Federico Santa María, Av. España 1680, Valparaíso 2390123, Chile
[2] Department of Electronic Systems, Aalborg University, 9220 Aalborg, Denmark
* Correspondence: milan.derpich@usm.cl (M.S.D.); jo@es.aau.dk (J.Ø.)

Abstract: We present novel data-processing inequalities relating the mutual information and the directed information in systems with feedback. The internal deterministic blocks within such systems are restricted only to be causal mappings, but are allowed to be non-linear and time varying, and randomized by their own external random input, can yield any stochastic mapping. These randomized blocks can for example represent source encoders, decoders, or even communication channels. Moreover, the involved signals can be arbitrarily distributed. Our first main result relates mutual and directed information and can be interpreted as a law of conservation of information flow. Our second main result is a pair of data-processing inequalities (one the conditional version of the other) between nested pairs of random sequences entirely within the closed loop. Our third main result introduces and characterizes the notion of in-the-loop (ITL) transmission rate for channel coding scenarios in which the messages are internal to the loop. Interestingly, in this case the conventional notions of transmission rate associated with the entropy of the messages and of channel capacity based on maximizing the mutual information between the messages and the output turn out to be inadequate. Instead, as we show, the ITL transmission rate is the unique notion of rate for which a channel code attains zero error probability if and only if such an ITL rate does not exceed the corresponding directed information rate from messages to decoded messages. We apply our data-processing inequalities to show that the supremum of achievable (in the usual channel coding sense) ITL transmission rates is upper bounded by the supremum of the directed information rate across the communication channel. Moreover, we present an example in which this upper bound is attained. Finally, we further illustrate the applicability of our results by discussing how they make possible the generalization of two fundamental inequalities known in networked control literature.

Keywords: data-processing inequality; directed information; mutual information; networked control; feedback capacity

Citation: Derpich, M.S.; Østergaard, J. Directed Data-Processing Inequalities for Systems with Feedback. *Entropy* **2021**, *23*, 533. https://doi.org/10.3390/e23050533

Academic Editor: José A. Tenreiro Machado

Received: 24 March 2021
Accepted: 21 April 2021
Published: 26 April 2021

1. Introduction

The data-processing inequality states that if x, y, z are random variables such that x and z become independent when conditioning upon y, then:

$$I(x;y) \geq I(x;z) \tag{1}$$
$$I(y;z) \geq I(x;z), \tag{2}$$

where $I(x;y)$ denotes the mutual information between x and y [1] (p. 252) (a definition of mutual information is provided in Section 2.2 below). Among its many uses, the data-processing inequality plays a key role in the proof of the converse part (i.e., outer bounds) in rate-distortion [1–5] (p. 317), channel capacity [1,6–8] (pp. 208, 217, 540 and 566), and joint source-channel coding theorems [1,9–11] (p. 221), and has recently been extended to von Newmann algebras, which have applications in quantum field theory (see [12] and the references therein).

It is well known that mutual information has an important limitation in systems with feedback, such as the one shown in Figure 1a. In this system, p, q, r, s, e, u, x, and y are random sequences, and the blocks $\mathcal{S}_1, \ldots, \mathcal{S}_4$ are deterministic causal mappings with an added delay of at least one sample. These blocks, randomized by their exogenous random inputs p, q, r, s, may yield any causal stochastic dynamic mappings. As pointed out in [13], for sequences inside the loop, such as x and y, $I(\mathrm{x};\mathrm{y})$ does not distinguish the probabilistic interdependence produced by the effect x has on y from that stemming from the influence of y on x. This limitation motivated the introduction of the directed information in [13]. This notion assesses the amount of information that causally "flows" from a given random and ordered sequence to another. For this reason, it has increasingly found use in diverse applications, including characterizing the capacity of channels with feedback [13–16], the rate distortion function under causality constraints [5], establishing some of the fundamental limitations in networked control [17–23], determining causal relationships in neural networks [24], and portfolio theory and hypothesis testing [25], to name a few.

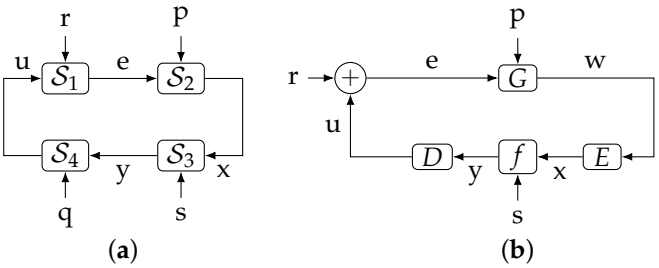

Figure 1. (**a**): The general system considered in this work. (**b**): A special case of (**a**), corresponding to the closed-loop system studied in [18].

The directed information from a randomsequence x^k to a random sequence y^k is defined as:

$$I(\mathrm{x}^k \to \mathrm{y}^k) \triangleq \sum_{i=1}^{k} I(\mathrm{y}(i); \mathrm{x}^i \,|\, \mathrm{y}^{i-1}), \tag{3}$$

where the notation x^i represents the sequence $\mathrm{x}(1), \mathrm{x}(2), \ldots, \mathrm{x}(i)$ and $I(\mathrm{x};\mathrm{y} \,|\, \mathrm{z})$ is the mutual information between x and y conditioned on (or given) z (hereafter we use non-italic letters, such as x, for random variables, denoting a particular realization by the corresponding italic character, x). The causality inherent in this definition becomes evident when comparing it with the mutual information between x^k and y^k, given by $I(\mathrm{x}^k;\mathrm{y}^k) = \sum_{i=1}^{k} I(\mathrm{y}(i); \mathrm{x}^k \,|\, \mathrm{y}^{i-1})$. In the latter sum, what matters is the amount of information about the *entire* sequence x^k present in $\mathrm{y}(i)$, given the past values y^{i-1}. By contrast, in the conditional mutual information in the sum of (3), only the past and current values of x^k are considered, that is, x^i. Thus, $I(\mathrm{x}^k \to \mathrm{y}^k)$ represents the amount of information causally conveyed from x^k to y^k. A related notion is the causally conditioned directed information introduced in [14], defined as:

$$I(\mathrm{x}^k \to \mathrm{y}^k \parallel \mathrm{q}^k) \triangleq \sum_{i=1}^{k} I(\mathrm{y}(i); \mathrm{x}^i \,|\, \mathrm{y}^{i-1}, \mathrm{q}^i). \tag{4}$$

In this paper, we derive inequalities involving directed and mutual information within feedback systems. For this purpose, we consider the general feedback system shown in Figure 1a. In this diagram, the blocks $\mathcal{S}_1, \ldots, \mathcal{S}_4$ represent possibly non-linear and time-varying causal discrete-time systems such that the total delay of the loop is at least one sample. These blocks can model, for example, source encoders, decoders or even communication channels. In the same figure, r, p, s, q are exogenous random signals (scalars, vectors, or sequences), which could represent, for example, any combination of disturbances, noises,

random initial states, or side information. We note that any of these exogenous signals, in combination with their corresponding deterministic mapping $\mathcal{S}_i$, can also yield any desired stochastic causal mapping (for example, a noisy communication channel, a zero-delay source coder or decoder, or a causal dynamic system with disturbances and a random initial state).

1.1. Main Contributions

1. Our first main contribution is the following theorem. It states a fundamental result, which relates the directed information between two signals within a feedback loop, say x and y, to the mutual information between an external set of signals and y:

   Theorem 1. *In the system shown in Figure 1a, it holds that:*

   $$I(\mathrm{x}^k \to \mathrm{y}^k) = I(\mathrm{q}^k, \mathrm{r}^k, \mathrm{p}^k \to \mathrm{y}^k) - I(\mathrm{q}^k, \mathrm{r}^k, \mathrm{p}^k \to \mathrm{y}^k \parallel \mathrm{x}^k) \leq I(\mathrm{p}^k, \mathrm{q}^k, \mathrm{r}^k; \mathrm{y}^k), \quad \forall k \in \mathbb{N}, \tag{5}$$

   with equality achieved if s *is independent of* $(\mathrm{p}, \mathrm{q}, \mathrm{r})$.

   The proof is in Section 3. This fundamental result, which for the cases in which $\mathrm{s} \perp\!\!\!\perp (\mathrm{p}, \mathrm{q}, \mathrm{r})$ can be understood as a *law of conservation of information flow*, is illustrated in Figure 2. (Here, and in the sequel, we use the notation $\mathrm{x} \perp\!\!\!\perp \mathrm{y}$ to mean "x is independent of y".) For such a cases, the information causally conveyed from x to y equals the information flow from $(\mathrm{q}, \mathrm{r}, \mathrm{p})$ to y. When $(\mathrm{p}, \mathrm{q}, \mathrm{r})$ are not independent of s, part of the mutual information between $(\mathrm{p}, \mathrm{q}, \mathrm{r})$ and y (corresponding to the term $I(\mathrm{q}^k, \mathrm{r}^k, \mathrm{p}^k \to \mathrm{y}^k \parallel \mathrm{x}^k)$) can be thought of as being "leaked" through s, thus bypassing the forward link from x to y. This provides an intuitive interpretation for (5).

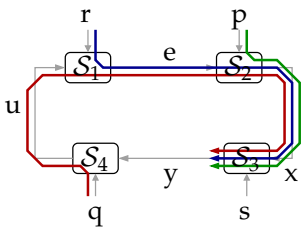

Figure 2. The flow of information between exogenous signals $(\mathrm{p}, \mathrm{q}, \mathrm{r})$ and the internal signal y equals the directed information from x^k to y^k when $\mathrm{s} \perp\!\!\!\perp (\mathrm{p}, \mathrm{q}, \mathrm{r})$.

   Remark 1. *Theorem 1 implies that $I(\mathrm{x}^k \to \mathrm{y}^k)$ is only a part of (or at most equal to) the information "flow" between all the exogenous signals entering the loop outside the link* $\mathrm{x} \to \mathrm{y}$ *(namely* $(\mathrm{q}, \mathrm{r}, \mathrm{p})$*), and* y*. In particular, if* $(\mathrm{p}, \mathrm{q}, \mathrm{r})$ *were deterministic, then* $I(\mathrm{x}^k \to \mathrm{y}^k) = 0$*, regardless of the blocks* $\mathcal{S}_1, \ldots, \mathcal{S}_4$ *and irrespective of the nature of* s*.*

2. Our second main result is the following theorem, which relates directed information involving four different sequences internal to the loop. The proof is in Appendix A on page 19.

   Theorem 2 (Full Closed-Loop Directed Data-Processing Inequality)**.** *Consider the system shown in Figure 1a.*

   (a) *If* $(\mathrm{q}, \mathrm{s}) \perp\!\!\!\perp (\mathrm{r}, \mathrm{p})$ *and* $\mathrm{q} \perp\!\!\!\perp \mathrm{s}$*, or if* $(\mathrm{p}, \mathrm{s}) \perp\!\!\!\perp (\mathrm{r}, \mathrm{q})$ *and* $\mathrm{p} \perp\!\!\!\perp \mathrm{s}$*, then:*

   $$I(\mathrm{x}^k \to \mathrm{y}^k) \geq I(\mathrm{e}^k \to \mathrm{u}^k). \tag{6}$$

   (b) *If* $(\mathrm{q}, \mathrm{s}) \perp\!\!\!\perp (\mathrm{r}, \mathrm{p})$ *and* $\mathrm{q}_{i+1}^k \leftrightarrow \mathrm{q}^i \leftrightarrow \mathrm{s}^i$ *for* $i = 1, 2, \ldots, k-1$*, then :*

   $$I(\mathrm{x}^k \to \mathrm{y}^k \parallel \mathrm{q}^k) \geq I(\mathrm{e}^k \to \mathrm{u}^k). \tag{7}$$

(The Markov chain notation $\mathrm{t} \leftrightarrow \mathrm{v} \leftrightarrow \mathrm{w}$ means "t and w are independent when v is given".) To the best of our knowledge, Theorem 2 is the first result available in the literature providing a lower bound to the gap between two instances of nested directed information, involving four different signals inside the feedback loop. This result can be seen as the first full extension of the open-loop (traditional) data-processing inequality, to arbitrary closed-loop scenarios. (Notice that there is no need to consider systems with more than four mappings, since all external signals entering the loop between a given pair of internal signals can be regarded as exogenous inputs to a single equivalent deterministic mapping.)

3. Our third main contribution is introducing the notion of *in-the-loop* (ITL) transmission rate (in Section 6) for the (seldom considered) channel-coding scenario in which the messages to be transmitted and the communication channel are internal to a feedback loop. We show that the supremum of the directed information rate across such a channel upper bounds the achievable ITL transmission rates. Moreover, we present an example in which this upper bound is attainable. This gives further operational meaning to the directed information rate in closed-loop scenarios.
4. Finally, we provide additional examples of the applicability of our results by discussing how they allow one to obtain the generalizations of two fundamental inequalities known in networked control literature. The first one appears in [18] (Lemma 4.1) and is written in (12) below. This generalization is a consequence of Theorem 4 and is discussed in Remarks 3 and 5 below. The second generalization applies to [20] (Theorem 4.1) and is described on page 6 below. It is an application of Theorem 2 that has just been carried out by the authors in [26], which is all the more important since, as we also reveal in that note, there is a flaw in the proof of [20] (Theorem 4.1).

A key ingredient in proving most of our theorems is provided by Lemma 1, stated in Section 2.5. It allows one to rigorously establish some of the non-trivial conditional independencies that arise in a feedback loop with several (possibly stochastic) dynamic systems.

Put together, the law of conservation of information flow from Theorem 1, our extension of the data processing inequality to general feedback systems from Theorem 2, and our other results constitute both a conceptual framework and a toolbox for addressing information flow problems in feedback systems. We are convinced that this contribution will be instrumental in establishing new results on, e.g., rate-distortion and channel capacity problems with feedback.

The literature review presented next will allow the reader to further assess the novelty and relevance of our results.

1.2. Existing Related Results

There exist several results characterizing the relationship between $I(\mathrm{x}^k \to \mathrm{y}^k)$ and $I(\mathrm{x}^k; \mathrm{y}^k)$. First, it is well known that $I(\mathrm{x}^k \to \mathrm{y}^k) \leq I(\mathrm{x}^k; \mathrm{y}^k)$, with equality if and only if y^k is causally related to x^k [13]. A conservation law of mutual and directed information has been found in [27], which asserts that $I(\mathrm{x}^k \to \mathrm{y}^k) + I(0 * \mathrm{y}^{k-1} \to \mathrm{x}^k) = I(\mathrm{x}^k; \mathrm{y}^k)$, where $0 * \mathrm{y}^{k-1}$ denotes the concatenation $0, \mathrm{y}(1), \ldots, \mathrm{y}^{k-1}$.

Given its prominence in settings involving feedback, it is perhaps in these scenarios where the directed information becomes most important. For instance, the directed information has been instrumental in characterizing the capacity of channels with feedback (see, e.g., [15,16,28] and the references therein), as well as the rate-distortion function in setups involving feedback [5,20–22,29].

For the simple case in which all the systems $\{\mathcal{S}_i\}_{i=1}^4$ are linear time invariant (LTI) and stable, and assuming $\mathrm{p}, \mathrm{x}, \mathrm{q} = 0$ (deterministically), it was shown in [30] that $I(\mathrm{r}^k \to \mathrm{e}^k)$ does not depend on whether there is feedback from e to u or not.

Inequalities between mutual and directed information in a less restricted setup, shown in Figure 1b, have been found in [18,19]. In that setting (a networked-control system), G is a strictly causal LTI dynamic system having (vector) state sequence $\{\mathbf{x}(i)\}_{i=0}^{\infty}$, with $\mathrm{p} \triangleq \mathbf{x}(0)$ being the random initial state in its state-space representation. The external signal r (which

could correspond to a disturbance) is statistically independent of s, the latter corresponding to, for example, side information or channel noise. Both are also statistically independent of p.

The blocks labeled E, D, and f correspond to an encoder, a decoder, and a channel, respectively, all of which are causal. The channel f maps s^k and x^k to $\mathrm{y}(k)$ in a possibly time-varying manner, i.e., $\mathrm{y}(k) = f(k, \mathrm{x}^k, \mathrm{s}^k)$. Similarly, the concatenation of the encoder, the channel and the decoder, maps s^k and w^k to $\mathrm{u}(k)$ as a possibly time-dependent function $\mathrm{u}(k) = \psi(k, \mathrm{w}^k, \mathrm{s}^k)$. Under these assumptions, the following fundamental result was shown in [19] (Lemma 5.1):

$$I(\mathrm{r}^k, \mathrm{p}\,;\, \mathrm{u}^k) \geq I(\mathrm{r}^k; \mathrm{u}^k) + I(\mathrm{p}; \mathrm{e}^k). \tag{8}$$

By further assuming in [19] that the decoder D in Figure 1b is deterministic, the following Markov chain naturally holds,

$$(\mathrm{p}, \mathrm{r}^k) \longleftrightarrow \mathrm{y}^k \longleftrightarrow \mathrm{u}^k, \tag{9}$$

leading directly to:

$$I(\mathrm{r}^k, \mathrm{p}\,;\, \mathrm{y}^k) \geq I(\mathrm{r}^k; \mathrm{u}^k) + I(\mathrm{p}; \mathrm{e}^k), \tag{10}$$

which is found in the proof of [19] (Corollary 5.3). The deterministic nature of the decoder D played a crucial role in the proof of this result, since otherwise the Markov chain (9) does not hold, in general, due to the feedback from u to y.

Notice that both (8) and (10) provide lower bounds to mutual information as the sum of two mutual information terms, each of them relating a signal *external* to the loop (such as p, r^k) to a signal *internal* to the loop (such as u^k or y^k). Instead, the inequality:

$$I(\mathrm{x}^k \to \mathrm{y}^k) \geq I(\mathrm{r}^k; \mathrm{y}^k), \tag{11}$$

which holds for the system in Figure 1a and appears in [13] (Theorem 3) (and rediscovered later in [17] (Lemma 4.8.1)), involves the directed information between two internal signals and the mutual information between the second of these and an external sequence.

Remark 2. *By using* (22), $I(\mathrm{p}^k, \mathrm{q}^k, \mathrm{r}^k; \mathrm{y}^k) = I(\mathrm{r}^k; \mathrm{y}^k) + I(\mathrm{p}^k, \mathrm{q}^k; \mathrm{y}^k \,|\, \mathrm{r}^k)$. *Then, applying Theorem 1, we recover* (11), *whenever* $\mathrm{s} \perp\!\!\!\perp (\mathrm{q}, \mathrm{r}, \mathrm{p})$. *Thus, [13,17] (Theorem 3) and (Lemma 4.8.1)) can be obtained as a corollary of Theorem 1.*

A related bound, similar to (10) but involving information rates and with the leftmost mutual information replaced by the directed information from x^k to y^k (which are two signals internal to the loop), has been obtained in [18] (Lemma 4.1) for the networked control system of Figure 1b:

$$\bar{I}(\mathrm{x} \to \mathrm{y}) \geq \bar{I}(\mathrm{r}; \mathrm{u}) + \lim_{k \to \infty} \frac{I(\mathrm{p}; \mathrm{e}^k)}{k}, \tag{12}$$

with $\bar{I}(\mathrm{x} \to \mathrm{y}) \triangleq \lim_{k\to\infty} \frac{1}{k} I(\mathrm{x}^k \to \mathrm{y}^k)$ and $\bar{I}(\mathrm{r}; \mathrm{u}) \triangleq \lim_{k\to\infty} \frac{1}{k} I(\mathrm{r}^k; \mathrm{u}^k)$, provided $\sup_{i\geq 0} \mathrm{E}\big[\mathbf{x}(i)^T \mathbf{x}(i)\big] < \infty$. This result relies on three assumptions: (a) that the channel f is memory-less and satisfies a "conditional invertibility" property, (b) a finite-memory condition, and (c) a fading-memory condition, the latter two related to the decoder D (see Figure 1).

It is worth noting that, as defined in [18], these assumptions upon D exclude the use of side information by the decoder and/or the possibility of D being affected by random noise or having a random internal state that is non-observable (please see [18] for a detailed description of these assumptions).

Remark 3. *In Section 4 we present Theorem 4, which yields (12) as a special case, but for the general system of Figure 1a and with no other assumption than mutual independence between* $\mathrm{r}, \mathrm{p}, \mathrm{q}, \mathrm{s}$. *Moreover, since with this independence condition Theorem 1 yields* $I(\mathrm{x}^k \to \mathrm{u}^k) = I(\mathrm{r}^k, \mathrm{p}^k; \mathrm{u}^k)$, *the same happens with (8).*

The inequality (11) has been extended in [16] (Theorem 1), for the case of discrete-valued random variables and assuming $\mathrm{s} \perp\!\!\!\perp (\mathrm{r}, \mathrm{p}, \mathrm{q})$, as the following identity (written in terms of the signals and setup shown in Figure 1a):

$$I(\mathrm{x}^k \to \mathrm{y}^k) = I(\mathrm{p}^k, \mathrm{y}^k) + I(\mathrm{x}^k \to \mathrm{y}^k \,|\, \mathrm{p}^k). \tag{13}$$

Letting $\mathrm{q} = \mathrm{s}$ in Figure 1a and with the additional assumption that $(\mathrm{p}, \mathrm{s}) \perp\!\!\!\perp \mathrm{q}$, it was also shown in [16] (Theorem 1) that:

$$I(\mathrm{x}^k \to \mathrm{y}^k) = I(\mathrm{p}^k; \mathrm{y}^k) + I(\mathrm{q}^{k-1}; \mathrm{y}^k) + I(\mathrm{p}^k; \mathrm{q}^{k-1} \,|\, \mathrm{y}^k), \tag{14}$$

for the cases in which $\mathrm{u}(i) = \mathrm{y}(i) + \mathrm{q}(i)$ (i.e., when the concatenation of $\mathcal{S}_4$ and $\mathcal{S}_1$ corresponds to a summing node). In [16], (13) and (14) play important roles in characterizing the capacity of channels with noisy feedback.

To the best of our knowledge, (8), (10), (11)–(14) are the only results available in the literature that lower bound the difference between internal-to-internal directed information and external-to-internal mutual information. There exist even fewer published results in relation to inequalities between two directed information terms involving only signals internal to the loop. To the best of our knowledge, the only inequality of this type in the literature is the one found in the proof of Theorem 4.1 of [20]. The latter takes the form of a (conditional) data-processing inequality for directed information in closed-loop systems, and states that:

$$I(\mathrm{x}^k \to \mathrm{y}^k \parallel \mathrm{q}^k) \geq I(\mathrm{x}^k \to \mathrm{u}^k), \tag{15}$$

provided: $\mathrm{q} \perp\!\!\!\perp (\mathrm{r}, \mathrm{p})$ and if $\mathcal{S}_4$ is such that y^i is a function of $(\mathrm{u}^i, \mathrm{q}^i)$ (i.e., if $\mathcal{S}_4$ is conditionally invertible) $\forall i$.

Inequality (15) plays a crucial role in [20], since it allows [20] (Thm. 4.1) to lower bound the average data rate across a digital error-free channel by a directed information. The setup considered in that theorem is shown in Figure 3, where $\mathcal{F}$ is a plant, and $\mathcal{E}$, $\mathcal{D}$ are the (source) encoder and decoder, respectively. In this figure, the variables that have been adapted to match those in Figure 4a ($\mathrm{r}, \mathrm{p}, \mathrm{x}$ correspond to disturbance, initial state, and plant output, respectively). Assuming $(\mathrm{r}, \mathrm{p}) \perp\!\!\!\perp (\mathrm{q}, \mathrm{s})$ and a conditionally invertible decoder, and letting $R(i)$ be the expected length (in bits) necessary for a binary representation of $\mathrm{y}(i)$ given q^i, it states that $\frac{1}{k}\sum_{i=1}^{k} R(i) \geq \frac{1}{k} I(\mathrm{x}^k \to \mathrm{u}^k)$, $k = 1, 2, \ldots$. This is a key result, because, combined with [20] (Equation (9)), it yields:

$$\frac{1}{k} I(\mathrm{x}^k \to \mathrm{u}^k) \leq \frac{1}{k} \sum_{i=1}^{k} R(i) \leq \frac{1}{k} I(\mathrm{x}^k \to \mathrm{u}^k) + 1 \quad \text{[bits/sample]}, \quad k = 1, 2, \ldots. \tag{16}$$

This result highlights the operational meaning of the directed information as a lower bound (tight to within one bit) to the data rate of any given source code in a closed-loop system. This fact has been a crucial ingredient in characterizing the best rate performance achievable in Gaussian linear quadratic networked control [23,31], demonstrating the relevance of directed data-processing inequalities.

Unfortunately, as we will reveal in [26], the proof of [20] (Theorem 4.1) turns out to be invalid, since it relies upon [20] (Lemma 4.2), whose first claim does not hold. In [26] we use Theorem 2 to prove Theorem 4.1 of [20] without requiring a conditionally invertible decoder. This further illustrates the applicability of our results.

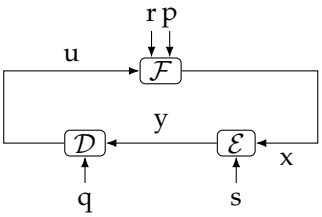

Figure 3. The networked control system considered in [20] (Figure 2), slightly simplified. The variables $\mathrm{r}, \mathrm{p}, \mathrm{x}, \mathrm{y}$ correspond to d, x_o, y, s in [20], respectively.

In [23] (Lemma 1) another data-processing inequality is stated, which for the system in Figure 1a is equivalent to:

$$I(\mathrm{x}^k \to \mathrm{y}^k \,\|\, \mathrm{u}_+^{k-1}) \geq I(\mathrm{x}^k \to \mathrm{u}^k), \quad k = 1, 2, \ldots \tag{17}$$

where: $I(\mathrm{x}^k \to \mathrm{y}^k \,\|\, \mathrm{u}_+^{k-1}) \triangleq \sum_{i=1}^k I(\mathrm{x}^i; \mathrm{y}(i) | \, \mathrm{y}^{i-1}, \mathrm{u}^{i-1})$. However, in [23] the blocks $\mathcal{S}_3$, $\mathcal{S}_4$ are defined implicitly, writing instead their input–output relation as collections of stochastic kernels $\mathbb{P}(\mathrm{y}(i) | \, \mathrm{x}^i, \mathrm{y}^{i-1})$, $\mathbb{P}(\mathrm{u}(i) | \, \mathrm{y}^i, \mathrm{u}^{i-1})$, $i = 1, 2, \ldots$. The notation $\mathbb{P}(\mathrm{t} | \mathrm{v}, \mathrm{w})$ is to be understood as the conditional distribution of t given (v, w). Crucially, this entails the implicit assumption that given x^i and y^{i-1}, $\mathrm{y}(i)$ is independent of every other signal in the system (and likewise for $\mathbb{P}(\mathrm{u}(i) | \, \mathrm{y}^i, \mathrm{u}^{i-1})$). In the representation of Figure 1a, this corresponds to assuming $\mathrm{q} \perp\!\!\!\perp \mathrm{s}$ and $(\mathrm{q}, \mathrm{s}) \perp\!\!\!\perp (\mathrm{r}, \mathrm{p})$.

Remark 4. *The conditioning on the side information* q *in both Theorem 2 and [20] (Theorem 4.1) is motivated by the use of* entropy coded subtractively dithered quantization *(ECSDQ) in obtaining the upper bound in* (16)*. For such a scenario, the sequences* q *and* s *are identical and correspond to the dither signal, which is independent of* r, p*. This satisfies the requirements of* (6) *in Theorem 2 and of [20] (Theorem 4.1), but not the assumption that* $\mathrm{q} \perp\!\!\!\perp \mathrm{s}$ *and* $(\mathrm{q}, \mathrm{s}) \perp\!\!\!\perp (\mathrm{r}, \mathrm{p})$ *implicit in [23] (Lemma 1), which yields* (17)*. In spite of this, Lemma 1 of [23] is used in that paper to prove the lower bound in [23] (Equation (8)), an analogue of* (16) *which also considers the use of ECSDQ for the rate term and its upper bound.*

1.3. Outline of the Paper

The remainder of the paper continues with some preliminary definitions and results in Section 2, the last of which is Lemma 2, a crucial tool for proving most of our theorems. Then follows the proof of Theorem 1 in Section 3. Section 4 presents additional inequalities relating mutual information between external–internal signal pairs, and directed information from one internal signal to another internal signal. These results can be seen as extensions or consequences of Theorem 1. Then we develop in Section 5 inequalities between two nested directed information expressions. Such results are the precursors of Theorem 2 and, as such, play a key role in its proof (which opens Appendix A). The notions and results associated with in-the-loop channel coding are developed in Section 6. The main conclusions of this work are presented in Section 7. Appendix A provides the proofs that are not written right after their corresponding theorems.

An earlier version of this work was made publicly available on arxiv.org [32] and, as such, it was cited in [23,31,33–35].

2. Preliminaries

2.1. Notation

The set of natural numbers is denoted $\mathbb{N}$. Random variables are denoted using non-italic characters, such as x. We write x^i to represent the sequence $\mathrm{x}(1), \mathrm{x}(2), \ldots, \mathrm{x}(i)$. We write $\mathrm{x} \perp\!\!\!\perp \mathrm{y}$ to express that x and y are independent. We use $\Pr\{\text{“outcome”}\}$ to denote the probability of a specific outcome of one or more random variables. For example, $\Pr\{\mathrm{x} = x, \mathrm{y} \in \mathcal{Y}\}$ is the probability that $\mathrm{x} = 1$ and y is in a given set $\mathcal{Y}$. Likewise, $\Pr\{\text{“outcome”} | \text{“outcome 2”}\}$ is the conditional probability of “outcome 1” given “out-

come 2″. The Markov-chain notation $\mathsf{x} \leftrightarrow \mathsf{y} \leftrightarrow \mathsf{z}$ means $\Pr\{\mathsf{x} \in \mathcal{X}, \mathsf{z} \in \mathcal{Z} | \, \mathsf{y} \in \mathcal{Y}\} = \Pr\{\mathsf{x} \in \mathcal{X} | \, \mathsf{y} \in \mathcal{Y}\} \Pr\{\mathsf{z} \in \mathcal{Z} | \, \mathsf{y} \in \mathcal{Y}\}$, for every choice of the sets $\mathcal{X}$, $\mathcal{Y}$, $\mathcal{Z}$ in the event spaces of x, y, and z, respectively. For two probability measures μ, ν on a common event space $\mathscr{U}$ the notation $\mu \ll \nu$ means that μ is absolutely continuous with respect to ν, i.e., that $\forall \mathcal{U} \in \mathscr{U} : \nu(\mathcal{U}) = 0 \Rightarrow \mu(\mathcal{U}) = 0$.

2.2. Mutual Information

Let $(\Omega, \mathcal{F}, P)$ be a probability space, and $(\mathcal{X}, \mathcal{F}_\mathcal{X})$ and $(\mathcal{Y}, \mathcal{F}_\mathcal{Y})$ be measurable spaces, and consider the random variables $\mathsf{x} : \Omega \to \mathcal{X}$, $\mathsf{y} : \Omega \to \mathcal{Y}$. Define $\mathcal{M} \triangleq \mathcal{F}_\mathcal{X} \otimes \mathcal{F}_\mathcal{Y}$, i.e, the σ-algebra generated by the rectangles $\{A \times B : A \in \mathcal{X}, B \in \mathcal{Y}\}$. Consider a probability space $(\mathcal{X} \times \mathcal{Y}, \mathcal{M}, m)$ where m is the (joint) distribution of (x, y), i.e, $m = P \circ (\mathsf{x}, \mathsf{y})^{-1}$.

Denote the marginal probability distributions of x and y by μ and ν, respectively, where:

$$\mu(A) = m(A \times \mathcal{Y}), \quad A \in \mathcal{F}_\mathcal{X} \tag{18}$$

$$\nu(B) = m(\mathcal{X} \times B), \quad B \in \mathcal{F}_\mathcal{Y} \tag{19}$$

Define the product measure $\pi \triangleq \mu \times \nu$ on $(\mathcal{X} \times \mathcal{Y}, \mathcal{M})$.

Definition 1. *With the above definitions, the mutual information between* x *and* y *is defined as:*

$$I(\mathsf{x};\mathsf{y}) \triangleq \int \log\left(\frac{dm}{d\pi}\right) dm, \tag{20}$$

where $\frac{dm}{d\pi}$ *is the Radon–Nikodym derivative of* m *with respect to* π *[36].*

An ensemble of random variables has a standard alphabet if it takes values from a set $\mathcal{A}$ in a standard measurable space $(\mathcal{A}, \mathcal{F}_\mathcal{A})$ [37] [Section 1.4] and its probability measure is defined on $\mathcal{F}_\mathcal{A}$. For our purposes, it suffices to say that standard alphabets include discrete spaces, the real line, Euclidean vector spaces, and Polish spaces (i.e., complete separable metric spaces) [38].

Lemma 1 (Chain Rule of Mutual Information from [37] (Corollary 7.14))**.** *Suppose* $\mathsf{x}, \mathsf{y}, \mathsf{z}$ *are random variables with standard alphabets and with joint distribution* P_{xyz}*. Suppose also that there exists a product distribution* $M_{\mathsf{xyz}} = M_\mathsf{x} \times M_{\mathsf{yz}}$ *such that* $M_{\mathsf{xyz}} \gg P_{\mathsf{xyz}}$*. (This is true, for example, if* $P_\mathsf{x} \times P_{\mathsf{yz}} \gg P_{\mathsf{xyz}}$*.) Then:*

$$I(\mathsf{x};\mathsf{y},\mathsf{z}) = I(\mathsf{x};\mathsf{y}) + I(\mathsf{x};\mathsf{z} \, | \, \mathsf{y}). \tag{21}$$

From [37] (Lemma 7.4 and Equation 7.28), we have that $I(\mathsf{x};\mathsf{y},\mathsf{z}) < \infty \Rightarrow P_\mathsf{x} \times P_{\mathsf{yz}} \gg P_{\mathsf{xyz}}$, and thus (21) also holds if $I(\mathsf{x};\mathsf{y},\mathsf{z})$ is finite.

The conditional version of the chain rule of mutual information [39] (see also [37] (Corollary 2.5.1)) will be extensively utilized in the proofs of our results:

$$I(\mathsf{t},\mathsf{v};\mathsf{w}|\mathsf{z}) = I(\mathsf{v};\mathsf{w}|\mathsf{z}) + I(\mathsf{t};\mathsf{w}|\mathsf{v},\mathsf{z}). \tag{22}$$

For discrete random variables x, y, taking values from the sets $\mathcal{X}, \mathcal{Y}$, respectively, the entropy of x is defined as:

$$H(\mathsf{x}) \triangleq -\sum_{x \in \mathcal{X}} \Pr\{\mathsf{x} = x\} \log(\Pr\{\mathsf{x} = x\}) \tag{23}$$

and the conditional entropy of x given y is defined as:

$$H(\mathsf{x} \, | \, \mathsf{y}) \triangleq -\sum_{x \in \mathcal{X}, y \in \mathcal{Y}} \Pr\{\mathsf{x} = x, \mathsf{y} = y\} \log(\Pr\{\mathsf{x} = x | \, \mathsf{y} = y\}) \tag{24}$$

The entropy satisfies the chain rule:

$$H(\mathrm{x},\mathrm{y}) = H(\mathrm{x}) + H(\mathrm{y}\,|\,\mathrm{x}) = H(\mathrm{y}) + H(\mathrm{x}\,|\,\mathrm{y}) \tag{25}$$

and is related to the mutual information as:

$$I(\mathrm{x};\mathrm{y}) = H(\mathrm{x}) + H(\mathrm{y}) - H(\mathrm{x},\mathrm{y}) = H(\mathrm{x}) - H(\mathrm{x}\,|\,\mathrm{y}) = H(\mathrm{y}) - H(\mathrm{y}\,|\,\mathrm{x}). \tag{26}$$

2.3. System Description

We begin by providing a formal description of the systems labeled $\mathcal{S}_1 \dots \mathcal{S}_4$ in Figure 1a. Their input–output relationships are given by the possibly-varying deterministic mappings (For notational simplicity, we omit writing their time dependency explicitly):

$$\mathrm{e}(i) = \mathcal{S}_1(\mathrm{u}^{i-d_1(i)}, \mathrm{r}^i), \tag{27a}$$
$$\mathrm{x}(i) = \mathcal{S}_2(\mathrm{e}^{i-d_2(i)}, \mathrm{p}^i), \tag{27b}$$
$$\mathrm{y}(i) = \mathcal{S}_3(\mathrm{x}^{i-d_3(i)}, \mathrm{s}^i), \tag{27c}$$
$$\mathrm{u}(i) = \mathcal{S}_4(\mathrm{y}^{i-d_4(i)}, \mathrm{q}^i), \tag{27d}$$

where $\mathrm{r}, \mathrm{p}, \mathrm{s}, \mathrm{q}$ are exogenous random signals and the (possibly time-varying) delays $d_1, d_2, d_3, d_4 \in \{0, 1, \dots\}$ are such that:

$$d_1(k) + d_2(k) + d_3(k) + d_4(k) \geq 1, \quad \forall k \in \mathbb{N}.$$

That is, the concatenation of $\mathcal{S}_1, \dots, \mathcal{S}_4$ has a delay of at least one sample. For every $i \in \{1, \dots, k\}$, $\mathrm{r}(i) \in \mathbb{R}^{n_\mathrm{r}(i)}$, i.e., $\mathrm{r}(i)$ is a real random vector whose dimension is given by some function $n_\mathrm{r} : \{1, \dots, k\} \to \mathbb{N}$. The other sequences $(\mathrm{q}, \mathrm{p}, \mathrm{s}, \mathrm{x}, \mathrm{y}, \mathrm{u})$ are defined likewise.

2.4. A Necessary Modification of the Definition of Directed Information

As stated in [13], the directed information (as defined in (3)) is a more meaningful measure of the flow of information between x^k and y^k than the conventional mutual information $I(\mathrm{x}^k;\mathrm{y}^k) = \sum_{i=1}^k I(\mathrm{y}(i);\mathrm{x}^k\,|\,\mathrm{y}^{i-1})$ when there exists causal feedback from y to x. In particular, if x^k and y^k are discrete-valued sequences, the input and output, respectively, of a forward channel, and if there exists *strictly causal* perfect feedback, so that $\mathrm{x}(i) = \mathrm{y}(i-1)$ (a scenario utilized in [13] as part of an argument in favor of the directed information), then the mutual information becomes:

$$\begin{aligned} I(\mathrm{x}^k;\mathrm{y}^k) &= H(\mathrm{y}^k) - H(\mathrm{y}^k\,|\,\mathrm{x}^k) = H(\mathrm{y}^k) - H(\mathrm{y}^k\,|\,\mathrm{y}^{k-1}) = H(\mathrm{y}^k) - H(\mathrm{y}(k)|\,\mathrm{y}^{k-1}) \\ &= H(\mathrm{y}^{k-1}). \end{aligned}$$

Thus, when strictly causal feedback is present, $I(\mathrm{x}^k;\mathrm{y}^k)$ fails to account for how much information about x^k has been conveyed to y^k through the forward channel that lies between them.

It is important to note that in [13] (as well as in many works concerned with communications), the forward channel is instantaneous, i.e., it has no delay. Therefore, if a feedback channel is utilized, then this feedback channel must have a delay of at least one sample, as in the example above. However, when studying the system in Figure 1a, we may need to evaluate the directed information between signals x^k and y^k which are, respectively, the input and output of a *strictly casual* forward channel (i.e., with a delay of at least one sample), whose output is instantaneously fed back to its input. In such a case, if one further assumes perfect feedback and sets $\mathrm{x}(i) = \mathrm{y}(i)$, then, in the same spirit as before,

$$I(\mathrm{x}^k \to \mathrm{y}^k) = \sum_{i=1}^k I(\mathrm{y}(i);\mathrm{x}^i\,|\,\mathrm{y}^{i-1}) = \sum_{i=1}^k \Big[H(\mathrm{y}(i)|\,\mathrm{y}^{i-1}) - H(\mathrm{y}(i)|\,\mathrm{x}^i,\mathrm{y}^{i-1})\Big] = H(\mathrm{y}^k).$$

As one can see, Massey's definition of directed information ceases to be meaningful if instantaneous feedback is utilized.

It is natural to solve this problem by recalling that, in the latter example, the forward channel had a delay, say d, greater than one sample. Therefore, if we are interested in measuring how much of the information in $\mathrm{y}(i)$, not present in y^{i-1}, was conveyed from x^i through the forward channel, we should look at the mutual information $I(\mathrm{y}(i);\mathrm{x}^{i-d}\,|\,\mathrm{y}^{i-1})$, because only the input samples x^{i-d} can have an influence on $\mathrm{y}(i)$. For this reason, we introduce the following, modified notion of directed information.

Definition 2 (Directed Information with Forward Delay)**.** *In this paper, the directed information from* x^k *to* y^k *through a forward channel with a non-negative time varying delay of* $d_{xy}(i)$ *samples is defined as:*

$$I(\mathrm{x}^k \to \mathrm{y}^k) \triangleq \sum_{i=1}^{k} I(\mathrm{y}(i);\mathrm{x}^{i-d_{xy}(i)}\,|\,\mathrm{y}^{i-1}). \tag{28}$$

For a zero-delay forward channel, the latter definition coincides with Massey's [13].

Likewise, we adapt the definition of causally-conditioned directed information to the definition:

$$I(\mathrm{x}^k \to \mathrm{y}^k \parallel \mathrm{e}^k) \triangleq \sum_{i=1}^{k} I(\mathrm{y}(i);\mathrm{x}^{i-d_{xy}(i)}\,|\,\mathrm{y}^{i-1},\mathrm{e}^i).$$

where, as before, $d_{xy}(i)$ is the delay from x to $\mathrm{y}(i)$.

2.5. A Fundamental Lemma

The following result is an essential ingredient in the proof of most of our theorems:

Lemma 2. *In the system shown in Figure 4, the exogenous signals* r,q *are mutually independent and* $\mathcal{S}_1,\mathcal{S}_2$ *are deterministic (possibly time-varying) causal measurable functions characterized by* $\mathrm{y}^i=\mathcal{S}_1(\mathrm{r}^i,\mathrm{u}^i)$, $\mathrm{u}^i=\mathcal{S}_2(\mathrm{q}^i,\mathrm{y}^{i-1})$, $\forall i\in\{1,\dots\}$, *with* $\mathrm{y}_0=y_0$ *(deterministic). For this system, and for every* $0\le j\le i\le k$ *such that* $i-j\le 1$ *and* $i\ge 1$, *the following Markov chain holds:*

$$\mathrm{r}^k \longleftrightarrow (\mathrm{u}^i,\mathrm{y}^j) \longleftrightarrow \mathrm{q}^k, \quad \forall k\in\mathbb{N}. \tag{29}$$

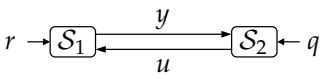

Figure 4. Two arbitrary causal systems $\mathcal{S}_1,\mathcal{S}_2$ interconnected in a feedback loop. The exogenous signals r,q are mutually independent.

Proof. Let $\mathscr{R},\mathscr{Q},\mathscr{U},\mathscr{Y}$ be the event spaces of $\mathrm{r}^k,\mathrm{q}^k,\mathrm{u}^i,\mathrm{y}^j$, respectively. Since $\mathrm{y}^j=\mathcal{S}_1(\mathrm{r}^j,\mathrm{u}^j)$ and $\mathrm{u}^i=\mathcal{S}_2(\mathrm{q}^i,\mathrm{y}^{i-1})$ are deterministic measurable functions, it follows that for every possible pair of events $\mathcal{U}\in\mathscr{U}$, $\mathcal{Y}\in\mathscr{Y}$, the preimage sets $\mathcal{R}_{\mathcal{U},\mathcal{Y}}\triangleq\{r^k:\mathcal{S}_1(r^j,u^j)\in\mathcal{Y},u^i\in\mathcal{U}\}$ and $\mathcal{Q}_{\mathcal{U},\mathcal{Y}}\triangleq\{q^k:\mathcal{S}_2(q^i,y^{i-1})\in\mathcal{U},y^j\in\mathcal{Y}\}$ are also deterministic and belong to $\mathscr{R}$ and $\mathscr{Q}$, respectively. Thus, $(\mathrm{u}^i,\mathrm{y}^j)\in\mathcal{U}\times\mathcal{Y}\iff(\mathrm{r}^k\in\mathcal{R}_{\mathcal{U},\mathcal{Y}},\mathrm{q}^k\in\mathcal{Q}_{\mathcal{U},\mathcal{Y}})$. This means that for every pair of events $R\in\mathscr{R},Q\in\mathscr{Q}$,

$$
\begin{aligned}
&\Pr\{\mathrm{r}^k \in R, \mathrm{q}^k \in Q | \, \mathrm{y}^j \in \mathcal{Y}, \mathrm{u}^i \in \mathcal{U}\} \\
&\quad \overset{(a)}{=} \Pr\{\mathrm{r}^k \in R, \mathrm{q}^k \in Q | \, \mathrm{r}^k \in \mathcal{R}_{\mathcal{U},\mathcal{Y}}\,, \, \mathrm{q}^k \in \mathcal{Q}_{\mathcal{U},\mathcal{Y}}\} \\
&\quad \overset{(b)}{=} \frac{\Pr\{\mathrm{r}^k \in R \cap \mathcal{R}_{\mathcal{U},\mathcal{Y}}\,, \, \mathrm{q}^k \in Q \cap \mathcal{Q}_{\mathcal{U},\mathcal{Y}}\}}{\Pr\{\mathrm{r}^k \in \mathcal{R}_{\mathcal{U},\mathcal{Y}}\,, \, \mathrm{q}^k \in \mathcal{Q}_{\mathcal{U},\mathcal{Y}}\}} \\
&\quad \overset{(c)}{=} \frac{\Pr\{\mathrm{r}^k \in R \cap \mathcal{R}_{\mathcal{U},\mathcal{Y}}\}}{\Pr\{\mathrm{r}^k \in \mathcal{R}_{\mathcal{U},\mathcal{Y}}\}} \cdot \frac{\Pr\{\mathrm{q}^k \in Q \cap \mathcal{Q}_{\mathcal{U},\mathcal{Y}}\}}{\Pr\{\mathrm{q}^k \in \mathcal{Q}_{\mathcal{U},\mathcal{Y}}\}} \\
&\quad = \frac{\Pr\{\mathrm{r}^k \in R \cap \mathcal{R}_{\mathcal{U},\mathcal{Y}}\} \Pr\{\mathrm{q}^k \in \mathcal{Q}_{\mathcal{U},\mathcal{Y}}\}}{\Pr\{\mathrm{r}^k \in \mathcal{R}_{\mathcal{U},\mathcal{Y}}\} \Pr\{\mathrm{q}^k \in \mathcal{Q}_{\mathcal{U},\mathcal{Y}}\}} \cdot \frac{\Pr\{\mathrm{q}^k \in Q \cap \mathcal{Q}_{\mathcal{U},\mathcal{Y}}\} \Pr\{\mathrm{r}^k \in \mathcal{R}_{\mathcal{U},\mathcal{Y}}\}}{\Pr\{\mathrm{q}^k \in \mathcal{Q}_{\mathcal{U},\mathcal{Y}}\} \Pr\{\mathrm{r}^k \in \mathcal{R}_{\mathcal{U},\mathcal{Y}}\}} \\
&\quad \overset{(d)}{=} \frac{\Pr\{\mathrm{r}^k \in R \cap \mathcal{R}_{\mathcal{U},\mathcal{Y}}\,, \, \mathrm{q}^k \in \mathcal{Q}_{\mathcal{U},\mathcal{Y}}\}}{\Pr\{\mathrm{r}^k \in \mathcal{R}_{\mathcal{U},\mathcal{Y}}\,, \, \mathrm{q}^k \in \mathcal{Q}_{\mathcal{U},\mathcal{Y}}\}} \cdot \frac{\Pr\{\mathrm{q}^k \in Q \cap \mathcal{Q}_{\mathcal{U},\mathcal{Y}}\,, \, \mathrm{r}^k \in \mathcal{R}_{\mathcal{U},\mathcal{Y}}\}}{\Pr\{\mathrm{q}^k \in \mathcal{Q}_{\mathcal{U},\mathcal{Y}}\,, \, \mathrm{r}^k \in \mathcal{R}_{\mathcal{U},\mathcal{Y}}\}} \\
&\quad \overset{(e)}{=} \Pr\{\mathrm{r}^k \in R \mid \, \mathrm{r}^k \in \mathcal{R}_{\mathcal{U},\mathcal{Y}}\,, \, \mathrm{q}^k \in \mathcal{Q}_{\mathcal{U},\mathcal{Y}}\} \cdot \Pr\{\mathrm{q}^k \in Q \mid \, \mathrm{q}^k \, \mathcal{Q}_{\mathcal{U},\mathcal{Y}}\,, \, \mathrm{r}^k \in \mathcal{R}_{\mathcal{U},\mathcal{Y}}\} \\
&\quad \overset{(f)}{=} \Pr\{\mathrm{r}^k \in R \mid \, \mathrm{y}^j \in \mathcal{Y}\,, \, \mathrm{u}^i \in \mathcal{U}\} \cdot \Pr\{\mathrm{q}^k \in Q \mid \, \mathrm{y}^j \in \mathcal{Y}, \mathrm{u}^i \in \mathcal{U}\}
\end{aligned}
$$

where (a) and (f) follow because of the equivalence between the events $(\mathrm{y}^j \in \mathcal{Y}, \mathrm{u}^i \in \mathcal{U})$ and $(\mathrm{r}^k \in \mathcal{R}_{\mathcal{U},\mathcal{Y}}, \mathrm{q}^k \in \mathcal{Q}_{\mathcal{U},\mathcal{Y}})$, (b) and (e) follow from Bayes rule, and (c) and (d) are true because $\mathrm{r}^k \perp\!\!\!\perp \mathrm{q}^k$. This completes the proof. □

3. Proof of Theorem 1

It is clear from Figure 1a and from (27) that the relationship between r, p, q, s, x, and y can be represented by the diagram shown in Figure 5.

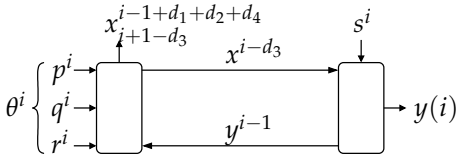

Figure 5. Representation of the system of Figure 1b highlighting the dependency between p, q, r, s, x, and y. The dependency on i of the delays $d_1(i), \ldots, d_4(i)$ is omitted for clarity.

From this diagram and Lemma 2 it follows that if s is independent of $(\mathrm{r}, \mathrm{p}, \mathrm{q})$, then the following Markov chain holds:

$$\mathrm{y}(i) \longleftrightarrow (\mathrm{x}^{i-d_3(i)}, \mathrm{y}^{i-1}) \longleftrightarrow (\mathrm{p}^i, \mathrm{q}^i, \mathrm{r}^i). \tag{30}$$

Denoting the triad of exogenous signals $\mathrm{p}^k, \mathrm{q}^k, \mathrm{r}^k$ by:

$$\theta^k \triangleq (\mathrm{p}^k, \mathrm{q}^k, \mathrm{r}^k), \tag{31}$$

we have the following:

$$
\begin{aligned}
I(\mathrm{x}^k \to \mathrm{y}^i) &= \sum_{i=1}^{k} I(\mathrm{y}(i); \mathrm{x}^{i-d_3(i)} \,|\, \mathrm{y}^{i-1}) \\
&\overset{(22)}{=} \sum_{i=1}^{k} \Big[I(\theta^i, \mathrm{x}^{i-d_3(i)}; \mathrm{y}(i) | \, \mathrm{y}^{i-1}) - I(\theta^i; \mathrm{y}(i) | \, \mathrm{x}^{i-d_3(i)}, \mathrm{y}^{i-1}) \Big] \\
&\overset{(a)}{=} \sum_{i=1}^{k} \Big[I(\theta^i; \mathrm{y}(i) | \, \mathrm{y}^{i-1}) - I(\theta^i; \mathrm{y}(i) | \, \mathrm{x}^{i-d_3(i)}, \mathrm{y}^{i-1}) \Big] \qquad (32a) \\
&\overset{(b)}{\leq} \sum_{i=1}^{k} I(\theta^i; \mathrm{y}(i) | \, \mathrm{y}^{i-1}) \overset{(c)}{\leq} \sum_{i=1}^{k} I(\theta^k; \mathrm{y}(i) | \, \mathrm{y}^{i-1}) \qquad (32b) \\
&= I(\theta^k; \mathrm{y}^k). \qquad (32c)
\end{aligned}
$$

In the above, (a) follows from the fact that, if y^{i-1} is known, and then $\mathrm{x}^{i-d_3(i)}$ is a deterministic function of θ^i. The resulting sums on the right-hand side of (32a) correspond to $I(\mathrm{q}^k, \mathrm{r}^k, \mathrm{p}^k \to \mathrm{y}^k) - I(\mathrm{q}^k, \mathrm{r}^k, \mathrm{p}^k \to \mathrm{y}^k \parallel \mathrm{x}^k)$, thereby proving the first part of the theorem, i.e., the equality in (5). In turn, (b) stems from the non-negativity of mutual information turn into equality if $\mathrm{s} \perp\!\!\!\perp (\mathrm{r}, \mathrm{p}, \mathrm{q})$, as a direct consequence of the Markov chain in (30). Finally, equality holds in (c) if $\mathrm{s} \perp\!\!\!\perp (\mathrm{q}, \mathrm{r}, \mathrm{p})$, since y depends causally upon θ. This shows that equality in (5) is achieved if $\mathrm{s} \perp\!\!\!\perp (\mathrm{q}, \mathrm{r}, \mathrm{p})$, completing the proof.

4. Relationships between Mutual and Directed Information

The following result provides an inequality relating $I(\mathrm{x}^k \to \mathrm{y}^k)$ with the separate flows of information $I(\mathrm{r}^k; \mathrm{y}^k)$ and $I(\mathrm{p}^k, \mathrm{q}^k \,;\, \mathrm{y}^k)$.

Theorem 3. *For the system shown in Figure 1a, if* $\mathrm{s} \perp\!\!\!\perp (\mathrm{p}, \mathrm{q}, \mathrm{r})$ *and* $\mathrm{r}^k \perp\!\!\!\perp (\mathrm{p}^k, \mathrm{q}^k)$*, then:*

$$
I(\mathrm{x}^k \to \mathrm{y}^k) \geq I(\mathrm{r}^k; \mathrm{y}^k) + I(\mathrm{p}^k, \mathrm{q}^k \,;\, \mathrm{y}^k). \qquad (33)
$$

with equality if and only if the Markov chain $(\mathrm{p}^k, \mathrm{q}^k) \leftrightarrow \mathrm{y}^k \leftrightarrow \mathrm{r}^k$ *holds.*

Theorem 3 shows that, provided $(\mathrm{p}, \mathrm{q}, \mathrm{r}) \perp\!\!\!\perp \mathrm{s}$, $I(\mathrm{x}^k \to \mathrm{y}^k)$ is lower bounded by the sum of the individual flows from all the subsets in any given partition of $(\mathrm{p}^k, \mathrm{q}^k, \mathrm{r}^k)$, to y^k, provided these subsets are mutually independent. Indeed, both Theorems 1 and 3 can be generalized for any appropriate choice of external and internal signals. More precisely, let Θ be the set of all external signals in a feedback system. Let α and β be two internal signals in the loop. Define $\Theta_{\alpha,\beta} \subset \Theta$ as the set of exogenous signals that are introduced to the loop at every subsystem $\mathcal{S}_i$ that lies in the path going from α to β. Thus, for any $\rho \in \Theta \setminus \Theta_{\alpha,\beta}$, if $\Theta_{\alpha,\beta} \perp\!\!\!\perp \Theta \setminus \Theta_{\alpha,\beta}$, we have that (5) and (33) become:

$$
I(\alpha \to \beta) = I(\Theta \setminus \{\Theta_{\alpha,\beta}\}; \beta), \qquad (34)
$$

$$
I(\alpha \to \beta) - I(\rho; \beta) \geq I(\Theta \setminus \{\rho \cup \Theta_{\alpha,\beta}\}; \beta), \qquad (35)
$$

respectively.

To finish this section, we present a stronger, non-asymptotic version of inequality (12):

Theorem 4. *In the system shown in Figure 1a, if* $(\mathrm{r}, \mathrm{p}, \mathrm{q}, \mathrm{s})$ *are mutually independent, then:*

$$
I(\mathrm{x}^k \to \mathrm{y}^k) = I(\mathrm{r}^k; \mathrm{u}^k) + I(\mathrm{p}^k; \mathrm{e}^k) + I(\mathrm{q}^k; \mathrm{y}^k) + I(\mathrm{p}^k; \mathrm{u}^k \,|\, \mathrm{e}^k) + I(\mathrm{r}^k, \mathrm{p}^k; \mathrm{y}^k \,|\, \mathrm{u}^k). \qquad (36)
$$

Remark 5. *As anticipated, Theorem 4 can be seen as an extension of* (12) *to the more general setup shown in Figure 1a, where the assumptions made in [18] (Lemma 4.1) do not need to hold. In particular, letting the decoder* D *and* p *in Figure 1b correspond to* $\mathcal{S}_4$ *and* p^k *in Figure 1a,*

respectively, we see that inequality (12) *holds even if the channel f has memory or D and E have independent initial states, or if the internal state of D is not observable [40].*

Theorem 4 also admits an interpretation in terms of information flows. This can be appreciated in the diagram shown in Figure 6, which depicts the individual full-turn flows (around the entire feedback loop) stemming from q, r, and p. Theorem 4 states that the sum of these individual flows is a lower bound for the directed information from x to y, provided $\mathrm{q}, \mathrm{r}, \mathrm{p}, \mathrm{s}$ are independent.

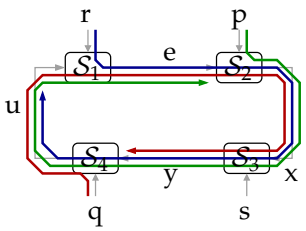

Figure 6. A representation of the three first information flows on the right-hand side of (36).

5. Relationships between Nested Directed Information

This section presents three closed-loop versions of the data-processing inequality *relating two directed information terms*, both between pairs of signals *internal* to the loop. As already mentioned in Section 1, to the best of our knowledge, the first inequality of this type to appear in the literature is the one in Theorem 4.1 in [20] (see (15)). Recall that the latter result stated that $I(\mathrm{x}^k \to \mathrm{y}^k \,\|\, \mathrm{q}^k) \geq I(\mathrm{x}^k \to \mathrm{u}^k)$, requiring $\mathcal{S}_4$ to be such that y^i is a deterministic function of $(\mathrm{u}^i, \mathrm{q}^i)$ and that $\mathrm{q} \perp\!\!\!\perp (\mathrm{r}, \mathrm{p})$. The following result presents another inequality that also relates two nested directed information terms, namely, $I(\mathrm{x}^k \to \mathrm{y}^k)$ and $I(\mathrm{e}^k \to \mathrm{y}^k)$, but requiring only that $\mathrm{s} \perp\!\!\!\perp (\mathrm{q}, \mathrm{r}, \mathrm{p})$.

Theorem 5. *For the closed-loop system in Figure 1b, if* $(\mathrm{q}, \mathrm{r}, \mathrm{p}) \perp\!\!\!\perp \mathrm{s}$, *then:*

$$I(\mathrm{x}^k \to \mathrm{y}^k) \geq I(\mathrm{e}^k \to \mathrm{y}^k). \tag{37}$$

Notice that Theorem 5 does not require p to be independent of r or q. This may seem counterintuitive upon noting that p enters the loop between the link from e to x.

The following theorem is an identity between two directed information terms involving only internal signals. It can also be seen as a complement to Theorem 5, since it can be directly applied to establish the relationship between $I(\mathrm{e}^k \to \mathrm{y}^k)$ and $I(\mathrm{e}^k \to \mathrm{u}^k)$.

Theorem 6. *For the system shown in Figure 1a, if* $\mathrm{s} \perp\!\!\!\perp (\mathrm{q}, \mathrm{r}, \mathrm{p})$, *then:*

$$I(\mathrm{x}^k \to \mathrm{y}^k) \geq I(\mathrm{x}^k \to \mathrm{u}^k) + I(\mathrm{q}^k ; \mathrm{y}^k) + I(\mathrm{r}^k, \mathrm{p}^k ; \mathrm{y}^k \,|\, \mathrm{u}^k) + I(\mathrm{q}^k ; \mathrm{r}^k \,|\, \mathrm{u}^k, \mathrm{y}^k), \tag{38}$$

with equality if, in addition, $\mathrm{q} \perp\!\!\!\perp (\mathrm{r}, \mathrm{p})$. *In the latter case, it holds that:*

$$I(\mathrm{x}^k \to \mathrm{y}^k) = I(\mathrm{x}^k \to \mathrm{u}^k) + I(\mathrm{q}^k ; \mathrm{y}^k) + I(\mathrm{r}^k, \mathrm{p}^k ; \mathrm{y}^k \,|\, \mathrm{u}^k). \tag{39}$$

Notice that by requiring additional independence conditions upon the exogenous signals (specifically, $\mathrm{q} \perp\!\!\!\perp \mathrm{s}$), Theorem 6 (and, in particular, (39)) yields:

$$I(\mathrm{x}^k \to \mathrm{y}^k) \geq I(\mathrm{x}^k \to \mathrm{u}^k), \tag{40}$$

which strengthens the inequality in [20] (Theorem 4.1) (stated above in (15)). More precisely, (40) does not require conditioning one of the directed information terms and holds irrespective of the invertibility of the mappings in the loop.

6. Giving Operational Meaning to the Directed Information: In-the-Loop Channel Coding

In this section we introduce the notions of in-the-loop transmission rate and capacity and show that they are related by the directed information rate across the channel in the same feedback loop. This provides another example to illustrate the applicability of Theorems 1 and 2 and also provides further operational meaning to the directed information rate.

Consider the scheme shown in Figure 7, and suppose $\mathcal{C}$ is a noisy communication channel. Let $\mathcal{E}$ and $\mathcal{D}$ be the channel encoder and decoder, respectively, with r and p being side information sequences causally and independently available to each of them such that $(\mathrm{r},\mathrm{p}) \perp\!\!\!\perp (\mathrm{s},\mathrm{q})$. This means that, for $k = 1, 2, \ldots, n$,

$$(\mathrm{p}_1^n, \mathrm{r}_{k+1}^n) \leftrightarrow \mathrm{r}_1^k \leftrightarrow (\mathrm{w}_1^{k+1}, \mathrm{x}_1^k, \mathrm{y}_0^k). \tag{41}$$

$$\mathrm{p}_{k+1}^n \leftrightarrow \mathrm{p}_1^k \leftrightarrow \mathrm{r}_1^k \tag{42}$$

$$\mathrm{p}_{k+1}^n \leftrightarrow \mathrm{p}_1^k \leftrightarrow (\mathrm{w}_1^{k+1}, \mathrm{x}_1^k, \mathrm{y}_0^k). \tag{43}$$

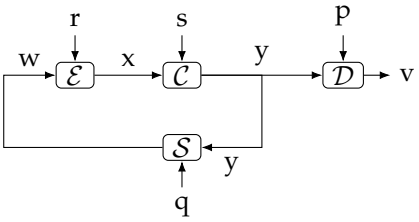

Figure 7. A communication feedback system in which the messages and the channel output are within the loop.

A crucial aspect of this scenario is the fact that the messages $\mathrm{w}_1^n, \mathrm{w}_{n+1}^{2n}, \ldots$ to be encoded are contained in the sequence w, a signal internal to the loop; they can be regarded as a corrupted version of the decoded messages, which comprise the sequence v. This is a key difference with respect to the available literature on feedback capacity, where, to the best of the authors' knowledge, the messages are exogenous and the feedback signal only helps in the encoding task (exceptions can be found in some papers on networked control which consider in-the-loop channel coding, such as, e.g., [41,42]). In Figure 7, the latter standard scenario corresponds to encoding the sequence r.

The fact that the messages to be encoded bear information from the decoded message symbols requires one to redefine the notion of information transmission rate commonly used in the standard scenario. To see this, let $\mathrm{w}(k) \in \mathcal{W}$, $k = 1, 2, \ldots$, for some finite alphabet $\mathcal{W}$ of cardinality $|\mathcal{W}|$, and notice that the transmission rate definitions $\log(|\mathcal{W}|)$ and $H(\mathrm{w}_1^n)/n$ are unsatisfactory if $\mathrm{w}(k) = \mathrm{y}(k-1)$, $k = 1, 2, \ldots$, i.e., if the messages to be transmitted are already available at the decoder (more generally, if there is no randomness in the feedback path). This suggests that a suitable notion of transmission rate for this scenario should exclude information that is already known by the receiver .

In view of the above, we propose the following notion of transmission rate for the case in which the messages to be transmitted are in the loop:

Definition 3. *For the system described in Figure 1, the in-the-loop (ITL) transmission rate is defined as:*

$$R_{ITL}^n \triangleq \frac{1}{n} \sum_{k=1}^{n} H(\mathrm{w}(k) | \, \mathrm{w}_1^{k-1}, \mathrm{y}_0^{k-1}, \mathrm{p}_1^k). \tag{44}$$

The meaning of the ITL transmission rate is further elucidated by considering the following scenarios:

1. If the feedback channel is deterministic, then $\mathrm{w}(k)$ is a deterministic function of y_0^{k-1} and thus $R^n_{\mathrm{ITL}} = 0$, as desired.
2. If the (forward) communication channel is noiseless, then at each time $k-1$, we have $\mathrm{y}_1^{k-1} = \mathrm{w}_1^{k-1}$. Therefore $R^n_{\mathrm{ITL}} = H(\mathrm{w}_1^n \,|\, \mathrm{y}_0, \mathrm{p}_1^n)/n$. Again, if the feedback channel is deterministic, the ITL transmission rate is zero.
3. In the absence of feedback, $R^n_{\mathrm{ITL}} = \frac{1}{n} H(\mathrm{w}_1^n)$, recovering the notion of transmission rate of the case in which the messages are exogenous to the loop.

Thus, R^n_{ITL} can be interpreted as the sum of the information the encoder attempts to transmit at each sample time (expressed by the conditional entropy $H(\mathrm{w}(k)|\, \mathrm{w}_1^{k-1}, \mathrm{y}_0^{k-1}, \mathrm{p}_1^k)$) that is novel for the transmitter (because of the conditioning on w_1^{k-1}) and novel for the receiver (due to the conditioning on $\mathrm{y}_0^{k-1}, \mathrm{p}_1^k$).

Theorem 7. *Consider the setup depicted in Figure 7, where $\mathcal{E}$ and $\mathcal{D}$ are the channel encoder and decoder, respectively, and $\mathcal{C}$ is the communication channel. Suppose the message and side-information samples $\mathrm{w}(k) \in \mathcal{W}$, $\mathrm{r}(k) \in \mathcal{R}$, $k = 1, 2, \ldots$, respectively, where $\mathcal{W}$ and $\mathcal{R}$ are finite alphabets. Define the binary random variable e_n to equal 1 if $\mathrm{v}_1^n \neq \mathrm{w}_1^n$ and 0 otherwise. Then, for every $n \in \mathbb{N}$,*

$$R^n_{ITL} \geq I(\mathrm{w}_1^n \to \mathrm{y}_0^n \,\|\, \mathrm{p}_1^n), \tag{45}$$

with equality if and only if $H(\mathrm{w}_1^n \,|\, \mathrm{y}_0^n, \mathrm{p}_1^n) = 0$. Moreover,

$$\Pr\{\mathrm{e}_n = 1\} = \frac{R^n_{ITL} - \frac{1}{n} I(\mathrm{w}_1^n \to \mathrm{y}_0^n \,\|\, \mathrm{p}_1^n) - \frac{1}{n} H(\mathrm{e}_n \,|\, \mathrm{y}_0^n, \mathrm{p}_1^n)}{\frac{1}{n} H(\mathrm{w}_1^n \,|\, \mathrm{y}_0^n, \mathrm{p}_1^n, \mathrm{e}_n = 1)} \tag{46}$$

$$\geq \frac{R^n_{ITL} - \frac{1}{n} I(\mathrm{w}_1^n \to \mathrm{y}_0^n \,\|\, \mathrm{p}_1^n) - 1/n}{\log_2(|\mathcal{W}|)} \tag{47}$$

Proof. Recall that:

$$I(\mathrm{w}_1^n \to \mathrm{y}_0^n \,\|\, \mathrm{p}_1^n) = \sum_{k=1}^n I(\mathrm{w}_1^k; \mathrm{y}(k) |\, \mathrm{y}_0^{k-1}, \mathrm{p}_1^k) = \sum_{k=1}^n H(\mathrm{w}_1^k \,|\, \mathrm{y}_0^{k-1}, \mathrm{p}_1^k) - \sum_{k=1}^n H(\mathrm{w}_1^k \,|\, \mathrm{y}_0^k, \mathrm{p}_1^k) \tag{48}$$

On the other hand,

$$nR^n_{\mathrm{ITL}} = \sum_{k=1}^n H(\mathrm{w}(k) |\, \mathrm{w}_1^{k-1}, \mathrm{y}_0^{k-1}, \mathrm{p}_1^k) \overset{(\mathrm{cr})}{=} \sum_{k=1}^n H(\mathrm{w}_1^k \,|\, \mathrm{y}_0^{k-1}, \mathrm{p}_1^k) - \sum_{k=2}^n H(\mathrm{w}_1^{k-1} \,|\, \mathrm{y}_0^{k-1}, \mathrm{p}_1^k) \tag{49}$$

$$\overset{(48)}{=} \sum_{k=1}^n H(\mathrm{w}_1^k \,|\, \mathrm{y}_0^k, \mathrm{p}_1^k) - \sum_{k=2}^n H(\mathrm{w}_1^{k-1} \,|\, \mathrm{y}_0^{k-1}, \mathrm{p}_1^k) + I(\mathrm{w}_1^n \to \mathrm{y}_0^n \,\|\, \mathrm{p}_1^n) \tag{50}$$

$$\overset{(43)}{=} H(\mathrm{w}_1^n \,|\, \mathrm{y}_0^n, \mathrm{p}_1^n) + I(\mathrm{w}_1^n \to \mathrm{y}_0^n \,\|\, \mathrm{p}_1^n), \tag{51}$$

where the equality (cr) follows from the chain rule of entropy. This proves the first part of the theorem.

Let us now re-derive the first steps leading to Fano's inequality, to include the side-information p_1^n and to verify that it is not affected by the fact that w and y are within the loop.

$$H(\mathrm{w}_1^n \,|\, \mathrm{y}_0^n, \mathrm{p}_1^n) \overset{(\mathrm{cr})}{=} H(\mathrm{w}_1^n, \mathrm{e}_n \,|\, \mathrm{y}_0^n, \mathrm{p}_1^n) - H(\mathrm{e}_n \,|\, \mathrm{y}_0^n, \mathrm{p}_1^n, \mathrm{w}_1^n) \tag{52}$$

$$\overset{(a)}{=} H(\mathrm{e}_n \,|\, \mathrm{y}_0^n, \mathrm{p}_1^n) + H(\mathrm{w}_1^n \,|\, \mathrm{y}_0^n, \mathrm{p}_1^n, \mathrm{e}_n) \tag{53}$$

$$\overset{(b)}{=} H(\mathrm{e}_n \,|\, \mathrm{y}_0^n, \mathrm{p}_1^n) + H(\mathrm{w}_1^n \,|\, \mathrm{y}_0^n, \mathrm{p}_1^n, \mathrm{e}_n = 1) \Pr\{\mathrm{e}_n = 1\}, \tag{54}$$

where the equality (cr) follows from the chain rule of entropy and (a) holds because $H(\mathrm{e}_n \,|\, \mathrm{y}_0^n, \mathrm{p}_1^n, \mathrm{w}_1^n) = 0$ and from the chain rule, while (b) is because $H(\mathrm{w}_1^n \,|\, \mathrm{y}_0^n, \mathrm{p}_1^n, \mathrm{e}_n = 0) = 0$.

Substituting this into (51),

$$nR_{\mathrm{ITL}}^n = H(\mathrm{e}_n \,|\, \mathrm{y}_0^n, \mathrm{p}_1^n) + H(\mathrm{w}_1^n \,|\, \mathrm{y}_0^n, \mathrm{p}_1^n, \mathrm{e}_n = 1)\Pr\{\mathrm{e}_n = 1\} + I(\mathrm{w}_1^n \to \mathrm{y}_0^n \,\|\, \mathrm{p}_1^n). \tag{55}$$

Noting that $H(\mathrm{e}_n \,|\, \mathrm{y}_0^n, \mathrm{p}_1^n) \leq 1$ and $H(\mathrm{w}_1^n \,|\, \mathrm{y}_0^n, \mathrm{p}_1^n, \mathrm{e}_n = 1) \leq n\log(|\mathcal{W}|)$ leads directly to (46), comparing the proof. □

Theorem 7 allows one to draw an additional interpretation of the ITL transmission rate. We extend first the identity of [27] to include causal conditioning by p_1^n:

$$I(\mathrm{w}_1^n; \mathrm{y}_0^n, \mathrm{p}_1^n) = I(\mathrm{w}_1^n; \mathrm{y}_n \,|\, \mathrm{y}_0^{n-1}, \mathrm{p}_1^n) + I(\mathrm{w}_n; \mathrm{y}_0^{n-1} \,|\, \mathrm{w}_1^{n-1}, \mathrm{p}_1^n) + I(\mathrm{w}_1^{n-1}; \mathrm{y}_0^{n-1}, \mathrm{p}_1^{n-1}) \tag{56}$$

$$= \sum_{k=1}^{n} I(\mathrm{w}_1^k; \mathrm{y}(k) | \, \mathrm{y}_0^{k-1}, \mathrm{p}_1^k) + \sum_{k=1}^{n} I(\mathrm{w}(k); \mathrm{y}_0^{k-1} \,|\, \mathrm{w}_1^{k-1}, \mathrm{p}_1^k) = I(\mathrm{w}_1^n \to \mathrm{y}_0^n \,\|\, \mathrm{p}_1^n) + I(\mathrm{y}_0^{n-1} \to \mathrm{w}_1^n \,\|\, \mathrm{p}_1^n), \tag{57}$$

where:

$$I(\mathrm{y}_0^{n-1} \to \mathrm{w}_1^n \,\|\, \mathrm{p}_1^n) \triangleq \sum_{k=1}^{n} I(\mathrm{w}(k); \mathrm{y}_0^{k-1} \,|\, \mathrm{w}_1^{k-1}, \mathrm{p}_1^k). \tag{58}$$

It readily follows from (56) that:

$$H(\mathrm{w}_1^n) - I(\mathrm{y}_0^{n-1} \to \mathrm{w}_1^n \,\|\, \mathrm{p}_1^n) = I(\mathrm{w}_1^n \to \mathrm{y}_0^n \,\|\, \mathrm{p}_1^n) + H(\mathrm{w}_1^n \,|\, \mathrm{y}_0^n, \mathrm{p}_1^n) \overset{(51)}{=} nR_{\mathrm{ITL}}^n. \tag{59}$$

Thus, the ITL transmission rate corresponds to the entropy rate of the messages having extracted from it the information flowing from the decoder input to the messages.

The main result of this section is the following theorem, which asserts that the supremum of achievable ITL transmission rates is upper bounded by the directed information across the communication channel.

Theorem 8. *Consider the setup depicted in Figure 7, where $\mathcal{E}$ and $\mathcal{D}$ are the channel encoder and decoder, respectively, and $\mathcal{C}$ is the communication channel. Then the supremum of achievable ITL transmission rates is upper bounded by the supremum of the directed information rate from* x *to* y *causally conditioned by* p_1^n.

Proof. The result follows directly from Theorems 2 and 7. □

Thus, the supremum of $\lim_{n\to\infty} I(\mathrm{x}_1^n \to \mathrm{y}_1^n \,\|\, \mathrm{p}_1^n)$ is an outer bound to the capacity region of ITL transmission rates.

In the following example, this bound is reachable.

Example 1. *Consider the case in which the forward channel $\mathcal{C}$ in Figure 7 is transparent, i.e.,* $\mathrm{y}(k) = \mathrm{x}(k)$ *for* $k = 0, 1, \ldots$, *as shown in Figure 8. Let* $\mathrm{y}(k) \in \{0, 1, 2, 3\}$, $k = 0, 1, \ldots$. *Let* $\mathrm{q}(0) = 1$ *(deterministically) and* $\mathrm{q}(1), \mathrm{q}(2), \ldots$ *be binary and i.i.d. with* $\Pr\{\mathrm{q}(k) = 1\} = \alpha = 0.9$. *The feedback channel $\mathcal{S}$ is defined by the following recursion:*

$$\mathrm{w}(k) = \begin{cases} \mathrm{q}(k) & , \text{ if } \mathrm{q}(k-1) = (\mathrm{y}(k-1) \mod 2) \\ (\mathrm{y}(k-1) \mod 2) & , \text{ if } \mathrm{q}(k-1) \neq (\mathrm{y}(k-1) \mod 2) \end{cases}, \quad k = 1, 2, \ldots \tag{60}$$

Thus, $\mathcal{S}$ outputs a new sample of q *iff the previous sample of* q *is matched by the previous sample* $\mod 2$ *of* y. *Otherwise, it lets* $\mathrm{y}(k-1) \mod 2$ *pass through.*

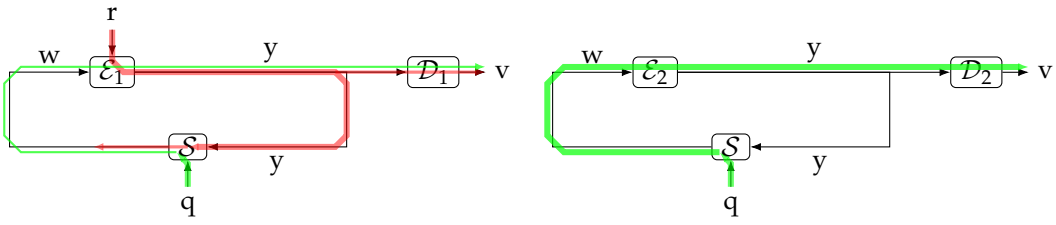

Figure 8. The feedback communication system considered in Example 1. In (**a**), encoder/decoder pair 1 yields a large mutual information between w and v by adding a backward information flow (in red) to the forward information flow (in green). The latter is only part of thee entropy rate of q. In (**b**), encoder/decoder pair 2 yields a smaller mutual information between w and v, but it corresponds to the greatest possible forward information flow, which coincides with the entropy rate of q. Thus, it is capacity achieving with respect to the ITL transmission rate.

Consider first the following encoder–decoder pair, designed with the aim of achieving zero-error communication while maximizing $H(\mathrm{w}_1^n)/n = I(\mathrm{w}_1^n;\mathrm{v}_1^n)$.

Encoder $\mathcal{E}_1$*: Let the side-information sequence* r *be binary i.i.d. and independent of* q*, with* $\Pr\{\mathrm{r}(k)=1\}=\beta$*, and:*

$$\mathrm{y}(0) = \mathrm{r}(0) \tag{61}$$

$$\mathrm{y}(k) = \begin{cases} \mathrm{r}(k) & \text{, if } \mathrm{w}(k) = (\mathrm{y}(k-1) \mod 2) \\ \mathrm{r}(k)+2 & \text{, if } \mathrm{w}(k) \neq (\mathrm{y}(k-1) \mod 2) \end{cases}, \quad k=1,2,\ldots \tag{62}$$

Decoder $\mathcal{D}_1$*:*

$$\mathrm{v}(k) = \begin{cases} \mathrm{y}(k-1) \mod 2 & \text{, if } \mathrm{y}(k) \leq 1 \\ (\mathrm{y}(k-1) \mod 2) \oplus 1 & \text{, if } \mathrm{y}(k) > 1 \end{cases} \tag{63}$$

where $\oplus$ *is the exclusive–or binary operator. With this choice,* $\mathrm{v}(k) = \mathrm{w}(k)$ *for* $k = 1,2,\ldots$*. In addition,*

$$\mathrm{w}(k) = \begin{cases} \mathrm{q}(k) & \text{, if } \mathrm{r}(k-1) = \mathrm{q}(k-1) \\ \mathrm{r}(k-1) & \text{, if } \mathrm{r}(k-1) \neq \mathrm{q}(k-1) \end{cases}, \quad k=1,2,\ldots \tag{64}$$

Therefore,

$$\Pr\{\mathrm{w}(1)=1\} = \alpha\beta. \tag{65}$$

and, for $k \geq 2$*,*

$$\Pr\{\mathrm{w}(k)=1\} = (\alpha^2 + (1-\alpha)^2)\beta + \alpha(1-\alpha) \tag{66}$$

Thus, and since $\alpha = 0.9$*, the entropy of each* $\mathrm{w}(k)$ *is maximized by* $\beta = 0.5055$*. However, encoder* $\mathcal{E}_1$ *makes the samples of* w *interdependent, so finding the value of* β *that maximizes* $H(\mathrm{w}_1^n)/n$ *(and thus* $I(\mathrm{w}_1^n;\mathrm{v}_1^n)$ *as well) is more involved, and that value does not need to be the same. We have found numerically that (for* $n = 22$*) the maximum of* $H(\mathrm{x}_1^n)/n = I(\mathrm{w}_1^n;\mathrm{v}_1^n)/n$ *is (approximately) 0.9941 [bits/sample], attained with* $\beta = 0.503$*, very close to the* β *which maximizes* $H(\mathrm{w}_1^n)/n$*.*

For later comparison, we also calculate the value of R_{ITL}^n *yielded by this choice of encoder:*

$$R_{ITL}^n \overset{(a)}{=} I(\mathrm{w}_1^n \to \mathrm{y}_1^n) \overset{Thm.\ 1}{=} I(\mathrm{q}_0^n;\mathrm{y}_0^n) = \sum_{k=0}^{n} (H(\mathrm{q}(k)|\,\mathrm{q}_0^{k-1}) - H(\mathrm{q}(k)|\,\mathrm{q}_0^{k-1},\mathrm{y}_0^n)), \tag{67}$$

where (a) *holds from Theorem 7 because* $H(\mathrm{w}_1^n \,|\, \mathrm{y}_0^n) = 0$. *Defining the binary random variables* $\mathrm{t}(k) \triangleq 1$ *when* $(\mathrm{y}(k) \mod 2) = \mathrm{q}(k)$ *and* 0 *otherwise, we obtain:*

$$H(\mathrm{q}(k)|\,\mathrm{q}_0^{k-1}, \mathrm{y}_0^n) = H(\mathrm{q}(k)|\,\mathrm{q}_0^{k-1}, \mathrm{y}_0^n, \mathrm{t}(k-1)) \tag{68}$$

$$= H(\mathrm{q}(k)|\,\mathrm{q}_0^{k-1}, \mathrm{y}_0^n, \mathrm{t}(k-1)=0)\Pr\{\mathrm{t}(k-1)=0\} + H(\mathrm{q}(k)|\,\mathrm{q}_0^{k-1}, \mathrm{y}_0^n, \mathrm{t}(k-1)=1)\Pr\{\mathrm{t}(k-1)=1\}$$

$$\overset{(64)}{=} H(\mathrm{q}(k))\Pr\{\mathrm{t}(k-1)=0\} + 0 \cdot \Pr\{\mathrm{t}(k-1)=1\} \tag{69}$$

Thus,

$$R_{ITL}^n = H(\mathrm{q}(k))(1 - \Pr\{\mathrm{t}(k-1)=0\}) = H(\mathrm{q}(k))(\alpha(1-\beta) + (1-\alpha)\beta) \tag{70}$$

$$= 0.469 \times 0.4976 = 0.2334 \quad \textit{[bits/sample]}, \tag{71}$$

using $\beta = 0.503$.

The second encoder/decoder pair is set to maximize R_{ITL}^n, *and is defined as follows:*

Encoder $\mathcal{E}_2$*:*

$$\mathrm{y}(k) = \begin{cases} 1 & , \textit{if } k = 0 \\ \mathrm{w}(k) & , \textit{if } k \geq 1 \end{cases} \tag{72}$$

Thus, zero-error communication is trivially attained with the simple decoding rule:

Decoder $\mathcal{D}_2$*:*

$$\mathrm{v}(k) = \mathrm{y}(k), \quad k \geq 1. \tag{73}$$

In addition, encoder $\mathcal{E}_2$ *yields* $\mathrm{w}(k) = \mathrm{q}(k)$, *for* $k \geq 1$. *Therefore,*

$$\frac{1}{n} I(\mathrm{w}_1^n \to \mathrm{y}_0^n) \overset{\textit{Thm. 7}}{=} R_{ITL}^n = \frac{1}{n} H(\mathrm{q}_1^n) = 0.469 \quad \textit{[bits/sample]} \tag{74}$$

As expected, encoder $\mathcal{E}_2$ *yields a higher* R_{ITL}^n *than encoder* $\mathcal{E}_1$. *More significant is the fact that encoder/decoder pair 2 achieves the in-the-loop capacity for this channel, since:*

$$\frac{1}{n} I(\mathrm{w}_1^n \to \mathrm{y}_0^n) \overset{(a)}{=} I(\mathrm{q}_1^n; \mathrm{y}_0^n) \leq H(\mathrm{q}_1^n) \tag{75}$$

The previous example illustrates an important fact that is closely related with the motivation behind the definition of R_{ITL}^n: maximizing the mutual information between the messages to be transmitted and the decoded messages (a leitmotif in traditional channel coding, wherein messages are generated outside the loop) is not suitable when messages are in the loop.

Indeed, (56) provides a mathematically precise meaning to the above observation. It reveals why maximizing $I(\mathrm{w}_1^n; \mathrm{y}_0^n, \mathrm{p}_1^n)$ does not necessarily mean maximizing $I(\mathrm{w}_1^n \to \mathrm{y}_0^n \,\|\, \mathrm{p}_1^n)$, since the former is the sum of backward and forward information flows (represented in green and red in Figure 8, respectively).

Finally, Theorems 7 and 8 imply that in the design of any encoder for in-the-loop messages, aiming to yield the joint probability distribution of channel input and output sequences that maximizes the directed information is of practical importance: it is necessary for achieving the highest "useful" transmission rate while minimizing the probability of error.

7. Concluding Remarks

The widely used data processing inequality does not hold for systems with feedback. In this work, we provided a very general *directed information* data processing inequality that is applicable to feedback systems. A key insight to be gained from this new inequality is that, for nested pairs of sequences, the further apart the signals in the feedback system are

from each other, the lower is the directed information between them (measuring distance from starting to finishing sequence and in the direction of cause and effect). Thus, post processing signals within a feedback loop cannot increase the information, which is similar to the open loop case. In order to obtain these results, we considered arbitrary causal systems that are interconnected in a feedback loop, with arbitrarily distributed signals. We were able to overcome the generally non-trivial dependencies between the signals in such a scenario by establishing a family of useful Markov chains that conditionally decouple the sequences in the system. These Markov chains are useful by themselves for studies involving interconnected systems. We further used the Markov chains to derive a number of fundamental information inequalities that are applicable to signals that are entirely within feedback loops or where some signals are inside and others outside the loop. With the use of these inequalities, we were able to show that the conventional notion of channel capacity is not adequate for *in-the-loop* communications. Instead, we provided the new notion of in-the-loop channel capacity, and described a special case where it was achievable. As an additional application of our results, we discussed how they allow one to generalize two known fundamental inequalities in networked control involving directed information. We are confident that our analysis provides useful insights to understand and think about information flows in single-loop feedback systems, and that our results will serve as a toolbox for research in, e.g., networked control systems or communications within a feedback loop.

There are several future research directions stemming from this work, from which we outline the following three:

1. Establishing whether (and under which conditions, if any) in the system of Figure 1, each of the following inequalities is true or false:
$$I(\mathrm{r}^k;\mathrm{e}^k) \geq I(\mathrm{r}^k;\mathrm{x}^k) \geq I(\mathrm{r}^k;\mathrm{y}^k) \geq I(\mathrm{r}^k;\mathrm{u}^k). \tag{76}$$
2. Extending Theorems 1 and 2 to scenarios with more than one feedback loop.
3. Exploring if tree codes [43] can be tailored to maximize the ITL data rate instead of the conventional data rate within a feedback loop. If such adaptation is possible, it would be interesting to assess how close to the ITL channel capacity such codes can perform.

Author Contributions: Both authors contributed to the derivation of the results presented in this paper and on its writing. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partially funded by ANID-CONICYT Basal fund FB0008.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

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

Appendix A. Proofs

Proof of Theorem 2. If $(\mathrm{q},\mathrm{s}) \perp\!\!\!\perp (\mathrm{r},\mathrm{p})$ and $\mathrm{q} \perp\!\!\!\perp \mathrm{s}$, then (6) follows by applying Theorem 5 and then Theorem 6. If $(\mathrm{p},\mathrm{s}) \perp\!\!\!\perp (\mathrm{r},\mathrm{q})$ and $\mathrm{p} \perp\!\!\!\perp \mathrm{s}$, then one arrives at (6) by applying Theorem 6 followed by Theorem 5.

To prove the second part, notice that:

$$I(\mathrm{x}^k \to \mathrm{y}^k \parallel \mathrm{q}^k) = I(\mathrm{x}^k \to \mathrm{y}^k \,|\, \mathrm{q}^k) \tag{A1}$$

which follows since $\mathrm{x}^{i-d_3(i)}, \mathrm{y}^i$ are deterministic functions of $(\mathrm{r}^k,\mathrm{p}^k,\mathrm{s}^i,\mathrm{q}^i)$ and $\mathrm{q}_{i+1}^k \leftrightarrow \mathrm{q}^i \leftrightarrow (\mathrm{r}^k,\mathrm{p}^k,\mathrm{s}^i)$, a Markov chain that results from combining $\mathrm{q}_{i+1}^k \leftrightarrow \mathrm{q}^i \leftrightarrow \mathrm{s}^i$ with $(\mathrm{q}^k,\mathrm{s}^k) \perp\!\!\!\perp (\mathrm{r}^k,\mathrm{p}^k)$.

On the other hand, the fact that $(\mathrm{r}^k, \mathrm{p}^k) \perp\!\!\!\perp (\mathrm{q}^k, \mathrm{s}^k)$ allows one to obtain from Theorem 1 that:

$$I(\mathrm{x}^k \to \mathrm{y}^k \,|\, \mathrm{q}^k) = I(\mathrm{r}^k, \mathrm{p}^k; \mathrm{y}^k \,|\, \mathrm{q}^k). \tag{A2}$$

But:

$$\begin{aligned} I(\mathrm{r}^k, \mathrm{p}^k; \mathrm{y}^k \,|\, \mathrm{q}^k) &\overset{(22)}{=} I(\mathrm{r}^k, \mathrm{p}^k\,;\, \mathrm{u}^k, \mathrm{y}^k \,|\, \mathrm{q}^k) - I(\mathrm{r}^k, \mathrm{p}^k; \mathrm{u}^k \,|\, \mathrm{q}^k, \mathrm{y}^k) \\ &\overset{(a)}{=} I(\mathrm{r}^k, \mathrm{p}^k\,;\, \mathrm{u}^k, \mathrm{y}^k \,|\, \mathrm{q}^k) \\ &\overset{(22)}{=} I(\mathrm{r}^k, \mathrm{p}^k\,;\, \mathrm{u}^k, \mathrm{y}^k, \mathrm{q}^k) - I(\mathrm{r}^k, \mathrm{p}^k\,;\, \mathrm{q}^k) \\ &\overset{(b)}{=} I(\mathrm{r}^k, \mathrm{p}^k\,;\, \mathrm{u}^k, \mathrm{y}^k, \mathrm{q}^k) \\ &\overset{(22)}{=} I(\mathrm{r}^k, \mathrm{p}^k\,;\, \mathrm{u}^k) + I(\mathrm{r}^k, \mathrm{p}^k; \mathrm{y}^k, \mathrm{q}^k \,|\, \mathrm{u}^k) \end{aligned} \tag{A3}$$

where (a) is due to the fact that u^k is a deterministic function of $\mathrm{q}^k, \mathrm{y}^k$. Equality (b) holds if and only if $(r, \mathrm{p}) \perp\!\!\!\perp \mathrm{q}$. The fact that $(\mathrm{r}^k, \mathrm{p}^k) \perp\!\!\!\perp (\mathrm{q}^k, \mathrm{s}^k)$ allows one to obtain from Theorem 1 that $I(\mathrm{x}^k \to \mathrm{u}^k) = I(\mathrm{r}^k, \mathrm{p}^k; \mathrm{u}^k)$. Substituting this into (A3) and then into (A2) and the latter into (A1), we obtain $I(\mathrm{x}^k \to \mathrm{y}^k \parallel \mathrm{q}^k) \geq I(\mathrm{x}^k \to \mathrm{u}^k)$, which combined with Theorem 5 yields (7). This completes the proof. □

Proof of Theorem 3. Apply the chain-rule identity (22) to the *right-hand side* (RHS) of (5) to obtain:

$$I(\theta^k; \mathrm{y}^k) = I(\mathrm{p}^k, \mathrm{q}^k, \mathrm{r}^k; \mathrm{y}^k) = I(\mathrm{p}^k, \mathrm{q}^k; \mathrm{y}^k \,|\, \mathrm{r}^k) + I(\mathrm{r}^k; \mathrm{y}^k). \tag{A4}$$

Now, applying (22) twice, one can express the term $I(\mathrm{p}^k, \mathrm{q}^k; \mathrm{y}^k \,|\, \mathrm{r}^k)$ as follows:

$$\begin{aligned} I(\mathrm{p}^k, \mathrm{q}^k; \mathrm{y}^k \,|\, \mathrm{r}^k) &= I(\mathrm{p}^k, \mathrm{q}^k\,;\, \mathrm{y}^k, \mathrm{r}^k) - I(\mathrm{p}^k, \mathrm{q}^k; \mathrm{r}^k) = I(\mathrm{p}^k, \mathrm{q}^k\,;\, \mathrm{y}^k, \mathrm{r}^k) \\ &= I(\mathrm{p}^k, \mathrm{q}^k; \mathrm{y}^k) + I(\mathrm{p}^k, \mathrm{q}^k; \mathrm{r}^k \,|\, \mathrm{y}^k), \end{aligned} \tag{A5}$$

where the second equality follows since $(\mathrm{p}^k, \mathrm{q}^k) \perp\!\!\!\perp \mathrm{r}^k$. The result then follows directly by combining (A5) with (A4) and (5). □

Proof of Theorem 4. Since $\mathrm{q} \perp\!\!\!\perp (\mathrm{r}, \mathrm{p}, \mathrm{s})$,

$$I(\mathrm{x}^k \to \mathrm{y}^k) \overset{(a)}{=} I(\mathrm{x}^k \to \mathrm{u}^k) + I(\mathrm{q}^k; \mathrm{y}^k) + I(\mathrm{r}^k, \mathrm{p}^k; \mathrm{y}^k \,|\, \mathrm{u}^k) \tag{A6}$$

$$\overset{(b)}{=} I(\mathrm{r}^k, \mathrm{p}^k; \mathrm{u}^k) + I(\mathrm{q}^k; \mathrm{y}^k) + I(\mathrm{r}^k, \mathrm{p}^k; \mathrm{y}^k \,|\, \mathrm{u}^k) \tag{A7}$$

$$\overset{(c)}{=} I(\mathrm{r}^k; \mathrm{u}^k) + I(\mathrm{p}^k; \mathrm{u}^k \,|\, \mathrm{r}^k) + I(\mathrm{q}^k; \mathrm{y}^k) + I(\mathrm{r}^k, \mathrm{p}^k; \mathrm{y}^k \,|\, \mathrm{u}^k), \tag{A8}$$

where (a) is due to Theorem 6, (b) follows from Theorem 1 and the fact that $(\mathrm{s}, \mathrm{q}) \perp\!\!\!\perp (\mathrm{r}, \mathrm{p})$, and (c) follows from the chain rule of mutual information. For the second term on the RHS of the last equation, we have:

$$I(\mathrm{p}^k;\mathrm{u}^k\,|\,\mathrm{r}^k) \overset{(a)}{=} I(\mathrm{p}^k;\mathrm{u}^k\,|\,\mathrm{r}^k) + I(\mathrm{p}^k;\mathrm{r}^k) = I(\mathrm{p}^k;\mathrm{r}^k,\mathrm{u}^k) \tag{A9}$$

$$\overset{(b)}{=} I(\mathrm{p}^k;\mathrm{r}^k,\mathrm{u}^k,\mathrm{e}^k) - I(\mathrm{p}^k;\mathrm{e}^k\,|\,\mathrm{r}^k,\mathrm{u}^k) \tag{A10}$$

$$\overset{(c)}{=} I(\mathrm{p}^k;\mathrm{r}^k,\mathrm{u}^k,\mathrm{e}^k) \tag{A11}$$

$$\overset{(d)}{=} I(\mathrm{p}^k;\mathrm{e}^k) + I(\mathrm{p}^k;\mathrm{r}^k,\mathrm{u}^k\,|\,\mathrm{e}^k) \tag{A12}$$

$$\overset{(e)}{=} I(\mathrm{p}^k;\mathrm{e}^k) + I(\mathrm{p}^k;\mathrm{u}^k\,|\,\mathrm{e}^k) + I(\mathrm{p}^k;\mathrm{r}^k\,|\,\mathrm{u}^k,\mathrm{e}^k) \tag{A13}$$

$$\overset{(f)}{=} I(\mathrm{p}^k;\mathrm{e}^k) + I(\mathrm{p}^k;\mathrm{u}^k\,|\,\mathrm{e}^k), \tag{A14}$$

where (a) holds since $\mathrm{r} \perp\!\!\!\perp \mathrm{p}$, (b), (d), and (e) stem from the chain rule of mutual information (22), and (c) is a consequence of the fact that $\mathrm{e}^k = \mathcal{S}_1(\mathrm{u}^{k-d_1(k)},\mathrm{r}^k)$. Finally, (f) is due to the Markov chain $\mathrm{r}^k \leftrightarrow (\mathrm{u}^k,\mathrm{e}^k) \leftrightarrow \mathrm{p}^k$, which holds because $\mathrm{r} \perp\!\!\!\perp (\mathrm{p},\mathrm{s},\mathrm{q})$ as a consequence of Lemma 2 in the Appendix (see also Figure 1a). Substitution of (A14) into (A8) yields (36), thereby completing the proof. □

Proof of Theorem 5. Since $(\mathrm{p},\mathrm{q},\mathrm{r}) \perp\!\!\!\perp \mathrm{s}$, we can apply (11) (where now (q,r) plays the role of r), and obtain

$$I(\mathrm{x}^k \to \mathrm{y}^k) \geq I(\mathrm{q}^k,\mathrm{r}^k;\mathrm{y}^k). \tag{A15}$$

Now, we apply Theorem 1, which gives

$$I(\mathrm{q}^k,\mathrm{r}^k;\mathrm{y}^k) \geq I(\mathrm{e}^k \to \mathrm{y}^k), \tag{A16}$$

completing the proof. □

Proof of Theorem 6. We have that:

$$I(\mathrm{x}^k \to \mathrm{y}^k) \overset{(a)}{=} I(\mathrm{r}^k,\mathrm{p}^k,\mathrm{q}^k\,;\,\mathrm{y}^k)$$

$$\overset{(22)}{=} I(\mathrm{q}^k\,;\,\mathrm{y}^k) + I(\mathrm{r}^k,\mathrm{p}^k\,;\,\mathrm{y}^k\,|\,\mathrm{q}^k) \tag{A17}$$

$$\overset{(A3)}{=} I(\mathrm{r}^k,\mathrm{p}^k\,;\,\mathrm{u}^k) + I(\mathrm{r}^k,\mathrm{p}^k;\mathrm{y}^k,\mathrm{q}^k\,|\,\mathrm{u}^k)$$

$$\overset{(22)}{=} I(\mathrm{q}^k\,;\,\mathrm{y}^k) + I(\mathrm{r}^k,\mathrm{p}^k\,;\,\mathrm{u}^k) + I(\mathrm{r}^k,\mathrm{p}^k;\mathrm{y}^k\,|\,\mathrm{u}^k) + I(\mathrm{r}^k,\mathrm{p}^k;\mathrm{q}^k\,|\,\mathrm{u}^k,\mathrm{y}^k) \tag{A18}$$

$$\overset{(b)}{\geq} I(\mathrm{q}^k\,;\,\mathrm{y}^k) + I(\mathrm{x}^k \to \mathrm{u}^k) + I(\mathrm{r}^k,\mathrm{p}^k;\mathrm{y}^k\,|\,\mathrm{u}^k) + I(\mathrm{r}^k,\mathrm{p}^k;\mathrm{q}^k\,|\,\mathrm{u}^k,\mathrm{y}^k) \tag{A19}$$

$$\overset{(c)}{\geq} I(\mathrm{q}^k\,;\,\mathrm{y}^k) + I(\mathrm{x}^k \to \mathrm{u}^k) + I(\mathrm{r}^k,\mathrm{p}^k;\mathrm{y}^k\,|\,\mathrm{u}^k), \tag{A20}$$

where (a) follows from Theorem 1 and the assumption $(\mathrm{r},\mathrm{p},\mathrm{q}) \perp\!\!\!\perp \mathrm{s}$, (b) is from Theorem 1, with equality iff $(\mathrm{q},\mathrm{s}) \perp\!\!\!\perp (\mathrm{r},\mathrm{p})$, and from Lemma 2 (in the Appendix), (c) turns into equality if $\mathrm{q} \perp\!\!\!\perp (\mathrm{r},\mathrm{p},\mathrm{s})$. This completes the proof. □

References

1. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*, 2nd ed.; Wiley-Interscience: Hoboken, NJ, USA, 2006.
2. Salek, S.; Cadamuro, D.; Kammerlander, P.; Wiesner, K. Quantum Rate-Distortion Coding of Relevant Information. *IEEE Trans. Inf. Theory* **2019**, *65*, 2603–2613. [CrossRef]
3. Lindenstrauss, E.; Tsukamoto, M. From Rate Distortion Theory to Metric Mean Dimension: Variational Principle. *IEEE Trans. Inf. Theory* **2018**, *64*, 3590–3609. [CrossRef]
4. Yang, Y.; Grover, P.; Kar, S. Rate Distortion for Lossy In-Network Linear Function Computation and Consensus: Distortion Accumulation and Sequential Reverse Water-Filling. *IEEE Trans. Inf. Theory* **2017**, *63*, 5179–5206. [CrossRef]
5. Derpich, M.S.; Østergaard, J. Improved upper bounds to the causal quadratic rate-distortion function for Gaussian stationary sources. *IEEE Trans. Inf. Theory* **2012**, *58*, 3131–3152. [CrossRef]

6. Ramakrishnan, N.; Iten, R.; Scholz, V.B.; Berta, M. Computing Quantum Channel Capacities. *IEEE Trans. Inf. Theory* **2021**, *67*, 946–960. [CrossRef]
7. Song, J.; Zhang, Q.; Kadhe, S.; Bakshi, M.; Jaggi, S. Stealthy Communication Over Adversarially Jammed Multipath Networks. *IEEE Trans. Inf. Theory* **2020**, *68*, 7473–7484. [CrossRef]
8. Makur, A. Coding Theorems for Noisy Permutation Channels. *IEEE Trans. Inf. Theory* **2020**, *66*, 6723–6748. [CrossRef]
9. Kostina, V.; Verdú, S. Lossy joint source-channel coding in the finite blocklength regime. *IEEE Trans. Inf. Theory* **2013**, *59*, 2545–2575. [CrossRef]
10. Huang, Y.; Narayanan, K.R. Joint Source-Channel Coding with Correlated Interference. *IEEE Trans. Commun.* **2012**, *60*, 1315–1327. [CrossRef]
11. Steinberg, Y.; Merhav, N. On hierarchical joint source-channel coding with degraded side information. *IEEE Trans. Inf. Theory* **2006**, *52*, 886–903. [CrossRef]
12. Hollands, S. Trace- and improved data processing inequalities for von Neumann algebras. *arXiv* **2021**, arXiv:2102.07479.
13. Massey, J.L. Causality, feedback and directed information. In Proceedings of the International Symposium on Information Theory and Its Applications, Honolulu, HI, USA, 27–30 November 1990; pp. 303–305.
14. Kramer, G. Directed Information for Channels with Feedback. Ph.D. Thesis, Swiss Federal Institute of Technology, Zürich, Switzerland, 1998.
15. Tatikonda, S.; Mitter, S. The Capacity of Channels With Feedback. *IEEE Trans. Inf. Theory* **2009**, *55*, 323–349. [CrossRef]
16. Li, C.; Elia, N. The Information Flow and Capacity of Channels with Noisy Feedback. *arXiv* **2011**, arXiv:1108.2815.
17. Tatikonda, S.C. Control Under Communication Constraints. Ph.D. Thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, USA, 2000.
18. Martins, N.C.; Dahleh, M.A. Fundamental limitations of performance in the presence of finite capacity feedback. In Proceedings of the 2005 American Control Conference, Portland, OR, USA, 8–10 June 2005.
19. Martins, N.; Dahleh, M. Feedback control in the presence of noisy Channels: "Bode-like" fundamental limitations of performance. *IEEE Trans. Autom. Control* **2008**, *53*, 1604–1615. [CrossRef]
20. Silva, E.I.; Derpich, M.S.; Østergaard, J. A framework for control system design subject to average data-rate constraints. *IEEE Trans. Autom. Control* **2011**, *56*, 1886–1899. [CrossRef]
21. Silva, E.I.; Derpich, M.S.; Østergaard, J. On the Minimal Average Data-Rate That Guarantees a Given Closed Loop Performance Level. In Proceedings of the 2nd IFAC Workshop on Distributed Estimation and Control in Networked Systems, NECSYS, Annecy, France, 13–14 September 2010; pp. 67–72.
22. Silva, E.I.; Derpich, M.S.; Østergaard, J. An achievable data-rate region subject to a stationary performance constraint for LTI plants. *IEEE Trans. Autom. Control* **2011**, *56*, 1968–1973. [CrossRef]
23. Tanaka, T.; Esfahani, P.M.; Mitter, S.K. LQG Control With Minimum Directed Information: Semidefinite Programming Approach. *IEEE Trans. Autom. Control* **2018**, *63*, 37–52. [CrossRef]
24. Quinn, C.; Coleman, T.; Kiyavash, N.; Hatsopoulos, N. Estimating the directed information to infer causal relationships in ensemble neural spike train recordings. *J. Comput. Neurosci.* **2011**, *30*, 17–44. [CrossRef]
25. Permuter, H.H.; Kim, Y.H.; Weissman, T. Interpretations of directed information in portfolio theory, data Compression, and hypothesis testing. *IEEE Trans. Inf. Theory* **2011**, *57*, 3248–3259. [CrossRef]
26. Derpich, M.S.; Østergaard, J. Comments on "A Framework for Control System Design Subject to Average Data-Rate Constraints". *arXiv* **2021**, arXiv:2103.12897 .
27. Massey, J.; Massey, P. Conservation of mutual and directed information. In Proceedings of the Proceedings. International Symposium on Information Theory, 2005. ISIT 2005, Adelaide, SA, Australia, 4–9 September 2005; pp. 157–158. [CrossRef]
28. Kim, Y.H.; Kim, Y.H. A Coding Theorem for a Class of Stationary Channels With Feedback. *IEEE Trans. Inf. Theory* **2008**, *54*, 1488–1499. [CrossRef]
29. Zamir, R.; Kochman, Y.; Erez, U. Achieving the Gaussian rate-distortion function by prediction. *IEEE Trans. Inf. Theory* **2008**, *54*, 3354–3364. [CrossRef]
30. Zhang, H.; Sun, Y.X. Directed information and mutual information in linear feedback tracking systems. In Proceedings of the 6-th World Congress on Intelligent Control and Automation, Dalian, China, 21–23 June 2006; pp. 723–727.
31. Silva, E.I.; Derpich, M.S.; Østergaard, J.; Encina, M.A. A characterization of the minimal average data rate that guarantees a given closed-lop performance level. *IEEE Trans. Autom. Control* **2016**, *61*, 2171–2186. [CrossRef]
32. Derpich, M.S.; Silva, E.I.; Østergaard, J. Fundamental Inequalities and Identities Involving Mutual and Directed Informations in Closed-Loop Systems. *arXiv* **2013**, arXiv:1301.6427.
33. Shahsavari Baboukani, P.; Graversen, C.; Alickovic, E.; Østergaard, J. Estimating Conditional Transfer Entropy in Time Series Using Mutual Information and Nonlinear Prediction. *Entropy* **2020**, *22*, 1124. [CrossRef] [PubMed]
34. Barforooshan, M.; Derpich, M.S.; Stavrou, P.A.; Ostergaard, J. The Effect of Time Delay on the Average Data Rate and Performance in Networked Control Systems. *IEEE Trans. Autom. Control* **2020**, 1. [CrossRef]
35. Baboukani, P.S.; Graversen, C.; Østergaard, J. Estimation of Directed Dependencies in Time Series Using Conditional Mutual Information and Non-linear Prediction. In Proceedings of the 2020 28th European Signal Processing Conference (EUSIPCO), Amsterdam, The Netherlands, 18–21 January 2021; pp. 2388–2392. [CrossRef]
36. Yeh, J. *Real Analysis*, 3rd ed.; World Scientific: Singapore, 2014.

37. Gray, R.M. *Entropy and Information Theory*, 2nd ed.; Science+Business Media, Springer: New York, NY, USA, 2011.
38. Gray, R.M. *Probability, Random Processes and Ergodic Properties*, 2nd ed.; Springer: Berlin/Heidelberg, Germany, 2009.
39. Yeung, R.W. *A First Course in Information Theory*; Springer: Berlin/Heidelberg, Germany, 2002.
40. Goodwin, G.C.; Graebe, S.; Salgado, M.E. *Control System Design*; Prentice Hall: Hoboken, NJ, USA, 2001 .
41. Sahai, A.; Mitter, S. The Necessity and Sufficiency of Anytime Capacity for Stabilization of a Linear System Over a Noisy Communication Link–Part I: Scalar Systems. *IEEE Trans. Inf. Theory* **2006**, *52*, 3369–3395. [CrossRef]
42. Khina, A.; Gårding, E.R.; Pettersson, G.M.; Kostina, V.; Hassibi, B. Control Over Gaussian Channels With and Without Source?Channel Separation. *IEEE Trans. Autom. Control* **2019**, *64*, 3690–3705. [CrossRef]
43. Khina, A.; Halbawi, W.; Hassibi, B. (Almost) practical tree codes. In Proceedings of the 2016 IEEE International Symposium on Information Theory, Barcelona, Spain, 10–15 July 2016; pp. 2404–2408. [CrossRef]