Is the direction of greater Granger causal influence the same as the direction of information flow?

Praveen Venkatesh* and Pulkit Grover†
Electrical & Computer Engineering, and the Center for the Neural Basis of Cognition, Carnegie Mellon University
*vpraveen@cmu.edu, †pulkit@cmu.edu

***Abstract*— Granger causality is an established statistical measure of the "causal influence" that one stochastic process X has on another process Y. Along with its more recent generalization – Directed Information – Granger Causality has been used extensively in neuroscience, and in complex interconnected systems in general, to infer statistical causal influences. More recently, many works compare the Granger causality metrics along forward and reverse links (from X to Y and from Y to X), and interpret the direction of greater causal influence as the "direction of information flow". In this paper, we question whether the direction yielded by comparing Granger Causality or Directed Information along forward and reverse links is always the same as the direction of information flow. We explore this question using two simple theoretical experiments, in which the true direction of information flow (the "ground truth") is known by design. The experiments are based on a communication system with a feedback channel, and employ a strategy inspired by the work of Schalkwijk and Kailath. We show that in these experiments, the direction of information flow can be opposite to the direction of greater Granger causal influence or Directed Information. We also provide information-theoretic intuition for why such counterexamples are not surprising, and why Granger causality-based information-flow inferences will only get more tenuous in larger networks. We conclude that one must not use comparison/difference of Granger causality to infer the direction of information flow.**

I. Introduction

A. Motivation

This work is in large part motivated by a recent surge of interest in understanding neural circuits – the connectivity and dynamic activity of different regions of the brain – and how they give rise to behavior and experience. This is evidenced by the launching of the BRAIN initiative in the US and the Human Brain Project in Europe. To quote from *BRAIN 2025: A Scientific Vision*[1], we wish to "map connected neurons in local circuits and distributed brain systems, enabling an understanding of the relationship between neuronal structure and function", clearly indicating the move towards understanding (a) the connectivity and (b) the computational function of different brain regions. While the question of how the brain computes has been of immense interest for several decades, only recently have measurement techniques become sophisticated enough to be able to simultaneously record the activity of multiple neurons, or multiple neural populations.

In order to understand how the brain performs computations, it could be useful to first understand the directions of *information flow* in various parts of the brain (e.g. [1]–[5] etc.). In an effort to make headway on the goals of *BRAIN 2025*, several works use Granger causality (and less often, its information-theoretic generalization – Directed Information) to understand how this information flows (e.g. [6]–[8]), or to acquire directed maps of functional connectivity (e.g. [7]–[9]). For instance, in [6], Granger causal influences that are measured between somatosensory and motor sites are said to "support the idea that somatosensory feedback provides information to the sensorimotor system that is used to control motor output". This raises the question: do these directed connectivity maps, as determined by directional causal influence measures such as Granger causality, correctly identify the directions along which information flows in the brain?

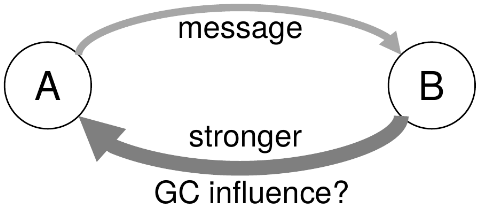

Fig. 1. Is the direction of stronger Granger-causal influence necessarily the same as the direction in which the message is flowing?

B. How Granger Causality is used in Neuroscience today

Several works have outlined the procedures involved in using Granger Causality to estimate causal influences in the brain ([6], [10]–[15]). Here, we briefly describe how Granger Causal influence is quantified, and how it is computed in these works.

Granger causality, as originally described by Granger [16], measures the level of causal influence that one process $\{X\}$ has on another process $\{Y\}$. The analysis compares the *error in predicting* the $\{Y\}$ process based on (i) simply the past of $\{Y\}$, and (ii) based on the past of both $\{X\}$ and $\{Y\}$[2]. The Granger causality metric is the ratio of these errors, encapsulating the *innovations* that the process $\{X\}$ *causally* supplies to the process $\{Y\}$. Many variants of Granger causality have also been developed, including a generalization – Directed Information (see [17]–[19]) – an

[1] The BRAIN Working Group's report to the Advisory Committee to the Director of the NIH

[2] A mathematical exposition of this process appears in Section III-A

information-theoretic quantity denoted by $I(\mathbf{X}^m \to \mathbf{Y}^m)$. These variants form possible alternatives for estimating the direction of causal influence, but Directed Information is a generalization of many of these metrics [18].

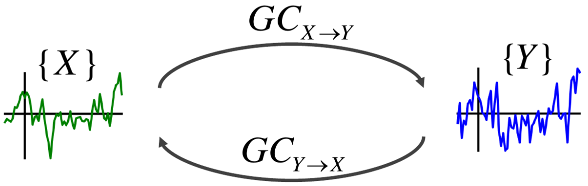

Fig. 2. In order to determine the direction of greater causal influence, the Granger Causality metrics in the forward and reverse directions are often compared [12].

In order to determine the direction of *greater* causal influence, the Granger Causality metric (ratio of residual variances) from $\{X\}$ to $\{Y\}$ is often *compared* to that from $\{Y\}$ to $\{X\}$. The direction of causal influence is then taken to be the direction with the greater Granger Causality metric (e.g. [6], [12], [19], see [12] for an understanding of what physical constraints motivate this comparison). Further, this direction of causal influence is interpreted to be the direction of information flow, which is the interpretation we question in this paper. We note here that Granger's original analysis does not compare this metric on forward and reverse links, and even the stronger notion of true causality (see remark 1 in section I-E) does not involve this comparison. However, this is commonly done in practice even in areas beyond neuroscience (e.g. [6], [12], [19], [20]).

We also note that many of these works use a spectral version of Granger Causality, that supplies this metric as a function of frequency. It then becomes possible to also determine the brain wave frequency at which these influences occur. However, we restrict our analysis to the simpler non-spectral version of Granger Causality, since it is sufficient for the purpose of our arguments.

While we accept that it might be possible to accurately estimate Granger Causal influence (provided measurements are taken suitably; see Section I-C) and that it could be useful in many situations (see Section V), interpreting the direction of greater Granger Causal influence as the direction of information flow can be erroneous, as we demonstrate in this paper.

C. A short survey of previous criticisms of Granger Causality

Several objections to the use of Granger Causality have been raised in the past. We give, here, a short overview of these and describe why our objection is novel, and possibly more fundamental in nature, at least in the context of its usage in neuroscience.

1) First, Granger Causality suffers from what we call the "hidden node problem". If two observed nodes receive causal influences from a third, latent node, then a causal influence may be detected between the observed nodes, even if they are independent of each other [21]. All nodes need to be observed, therefore, to avoid finding spurious influences.
2) Second, if the measurements from each node are differentially affected by noise, then the predicted direction of causal influence might be opposite to the true direction ([22], [23]). Measurements need to be relatively noiseless and precise in order to obtain the correct direction of causal influence.
3) Third, subsampling the processes can produce misleading Granger Causal relations [24]. Performing Granger Causal analysis on subsampled time series can lead one to miss the causal influence. If pre-processing involves subsampling, then this should be done with care.

It is important to note that the technical objections listed above are all deficiencies in or limitations of *measurement*. They indicate that incorrect Granger Causal influences may be estimated if there is some deficiency or limitation in the measurement procedure. These objections can be resolved by taking better measurements (by sampling more nodes, using sensors with higher signal-to-noise ratio, etc.).

Our objection, on the other hand, is more fundamental. *Even if* the measurements are made with infinite accuracy, and the regression coefficients associated with computing Granger Causality are precisely estimated, *and* the Granger Causality metric is perfectly computed (as is the case in our counter-examples), Granger Causal influence may *still* not yield the correct direction of information flow. To our knowledge, the argument that greater Granger causal influence can be opposite to the direction of information flow is a novel one. We believe that this argument is much more serious than previous objections, at least in the context of determining the directions of information flow in neuroscientific experiments, towards understanding the computational functions of brain regions.

A more serious objection has to do with the difference between statistical measures of causality (such as Granger Causality) and true causality, and whether or not our work simply alludes to this difference. Our principal argument is different, however, as we describe in remark 1 in Section I-E.

D. Our counter-examples and our main result

This paper considers two experiments (introduced in Section II) where a transmitter Tx wants to communicate a message to a receiver Rx in presence of a feedback channel (in one experiment, the feedback link is noiseless, while in the other it is noisy). We assume that the experimenter is able to record the transmissions of Tx and Rx using some probing mechanism. Provided with these measurements, the experimenter wants to estimate the direction of information flow, which in this context is the direction of flow of the message. Our results, derived in Section III and numerically illustrated in Section IV, show that the direction of information flow can be incorrectly inferred using both Granger causality and Directed Information. The first experiment considers the (unrealistic) case of communication across noiseless feedback channels. The second experiment allows for noise in the feedback channel. In both cases, linear strategies inspired by the scheme of Schalkwijk and Kailath [25] are used.

Our goal here is to bring out the point that whether Granger causality and Directed Information can be used to interpret the direction of information flow is an issue that can be, and perhaps should be, considered using thought experiments on simple communication problems where *information flow direction, and quantity, is already known*. If the direction of causal influence yielded by Granger causality or some other similar measure were to match the known direction of information flow, then that measure can be more confidently used in experiments. While our results strongly suggest that one needs to exercise care in interpreting Granger causality and Directed Information dominance as an indicator for the direction of information flow, there are several shortcomings that need to be addressed in order to understand the issue at depth. These shortcomings are discussed in detail in Section I-E.

We find it interesting to note that the mathematical machinery used in this paper amounts to routine arithmetic. Even simple counter-examples that do not employ difficult proof techniques are able to demonstrate our main result. This simplicity leads us to think that this counter-example is not very special, and that directions of stronger Granger Causal influence and information flow might have little to do with each other in more complex and/or noisy networks.

E. Possible objections to, and shortcomings of this work

A review of a previous conference submission of this paper had raised some objections to this work, which we discuss here to clarify our perspective. Further, our analysis has certain shortcomings, which we acknowledge. These will form the basis for future work.

1) Previous work has already noted that Granger Causal influence does not imply true causation [26]. This distinction is made rigorously by Judea Pearl [27], where he classifies Granger Causality as stemming from a "statistical" model, rather than from a "causal" model. The argument we make is very similar in spirit: we ask whether or not Granger Causality gives the correct direction of *information flow*. A question on the novelty of our work may therefore be raised: if our argument boils down to a restatement of Pearl's distinction, then this work has no new conceptual contribution.

   We make the case that our argument *is* novel in the following manner: in the systems we consider – the brain, as well as the communication system in our counter-example – causal influences exist in both directions. In our counter-example, for instance, the feedback communication algorithm that is employed involves transmissions from both Tx and Rx. The transmissions of each depend on what was transmitted by the other in the previous time instant. Causal influence and true causation, therefore, exist in both directions. The message, however, flows in only one direction: from the transmitter to the receiver. We ask whether or not the direction of *this* information flow can be discerned by comparing Granger causal influences in each direction. To this end, we give a concrete counter-example. This work is particularly relevant in the context of modern neuroscience, where such directions of information flow are desired in order to understand brain function.
2) Our analysis does not tackle an information source that evolves with time. Hence, our communication process (inspired from the scheme of Schalkwijk and Kailath [25]) is non-ergodic. Since Granger causality is really just relative errors in prediction of a process, in the presence or absence of knowledge of another process, we compute the obvious generalization of Granger causality to non-ergodic processes. Nevertheless, future work will address a situation with a linear dynamical system as the information source.
3) Our experiments restrict themselves to Gaussian noise for simplicity, but neural spiking and spike-rate models for spikes tend to be very different from those used here. This is a clear direction for future work.
4) The power and energy constraints are somewhat oversimplified to make the analysis simpler. This is for simplicity of exposition. A more general analysis is a simple extension.
5) We have also restricted ourselves to analyzing linear feedback communication strategies. In the presence of noise in the feedback link, linear communication strategies are known to be sub-optimal [28]. In order to make a water-tight argument, we would need to show that Granger Causality fails to correctly predict the direction of information flow, even when an optimal (non-linear) communication strategy is employed. This could be scope for future work. However, we do not expect results to change dramatically: when the feedback link is impaired by noise of only very low variance, the noiseless case should make for a good approximation of the system, and results should degrade gracefully, if there is any degradation at all.
6) We consider a simple point-to-point network. In general networks, this issue could be even more complex. However, since this issue shows up even for the simplest network, we feel that the problem will only be exacerbated when the network is large.

II. A SIMPLE FEEDBACK COMMUNICATION SCHEME

This section summarizes the analysis in the work of Schalkwijk and Kailath [25], and a simple (and previously known) extension to a noisy feedback case. It also establishes the model and the notation used in the paper.

The transmitter wants to convey a single zero-mean random number, Θ, having variance σ_θ^2, to the receiver. Θ could be obtained, for instance, by quantizing a bounded interval

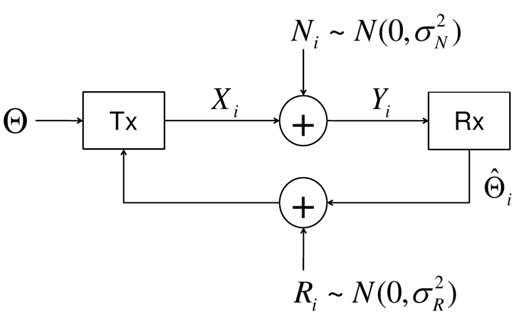

Fig. 3. A block-diagram representation of the communication system, describing the feedback channels and supplying notation for the variables used throughout the paper. Note that this diagram is the more general of the two cases discussed in sub-sections II-A and II-B, as it contains noise in the feedback link. The former, noiseless, case is equivalent to setting σ_R^2 to 0.

on the real line (e.g. $[-1,1]$), as is done in the scheme of Schalkwijk and Kailath [25]. The forward channel is an AWGN channel with noise variance σ_N^2. We will use simple linear communication strategies for a noiseless feedback channel, as well as an AWGN feedback channel with noise variance σ_R^2. In both cases, the estimators will be shown to be unbiased and consistent (the error mean is zero, and the error variance converges to zero).

A. Noiseless feedback: the Schalkwijk-Kailath strategy

In the first step, the transmitter sends[3] $X_1 = \Theta$, which the receiver receives with added noise. The receiver sends back an estimate of Θ over the feedback link. In all subsequent iterations, the transmitter sends the receiver the error in its latest estimate.

Therefore, in general, the transmitter sends

$$X_i = \Theta - \widehat{\Theta}_{i-1} \tag{1}$$

and the receiver receives

$$Y_i = X_i + N_i \tag{2}$$

where $N_i \sim \mathcal{N}(0,\sigma_N^2)$ iid. The receiver then estimates

$$\widehat{\Theta}_i = \widehat{\Theta}_{i-1} + \frac{Y_i}{i} \tag{3}$$

which results in:

$$\begin{aligned}
\widehat{\Theta}_i &= \widehat{\Theta}_{i-1} + \frac{X_i + N_i}{i} \\
&= \widehat{\Theta}_{i-1} + \frac{\Theta - \widehat{\Theta}_{i-1} + N_i}{i} \\
&= \frac{(i-1)\widehat{\Theta}_{i-1} + \Theta + N_i}{i} \\
i\widehat{\Theta}_i &= (i-1)\widehat{\Theta}_{i-1} + \Theta + N_i \\
&= (i-2)\widehat{\Theta}_{i-2} + \Theta + N_{i-1} + \Theta + N_i \\
&\ \ \vdots \\
&= i\Theta + \sum_{j=1}^{i} N_j \\
\widehat{\Theta}_i &= \Theta + \frac{1}{i}\sum_{j=1}^{i} N_j
\end{aligned} \tag{4}$$

Through this scheme, the estimate $\widehat{\Theta}_i$ is seen to converge to Θ in mean-square sense as $i \to \infty$:

$$\begin{aligned}
\mathbb{E}[\widehat{\Theta}_i] &= \mathbb{E}\left[\Theta + \frac{1}{i}\sum_{j=1}^{i} N_j\right] = \mathbb{E}[\Theta] + 0 \\
\mathbb{E}[(\widehat{\Theta}_i - \Theta)^2] &= \mathbb{E}\left[\left(\frac{1}{i}\sum_{j=1}^{i} N_j\right)^2\right] \\
&= \frac{1}{i^2}\mathbb{E}\left[\left(\sum_{j=1}^{i} N_j\right)^2\right] \\
&= \frac{i}{i^2}\mathbb{E}[N_1^2] = \frac{\sigma_N^2}{i} \xrightarrow{i\to\infty} 0.
\end{aligned}$$

B. Noisy feedback

In the presence of noise in the feedback link, restricting our attention to linear strategies, we can use a simple modification of the Schalkwijk-Kailath strategy, incorporating the feedback[4]. The receiver still simply transmits the estimate $\widehat{\Theta}_i$ based on the i-th forward channel output $Y_i = X_i + N_i$. The transmitter now receives corrupted versions $Z_i = \widehat{\Theta}_{i-1} + R_{i-1}$ of the receiver's transmissions. That is,

$$\text{Transmitter's transmissions: } X_i = \Theta - (\widehat{\Theta}_{i-1} + R_{i-1}) \tag{5}$$

$$\text{Channel outputs at the receiver: } Y_i = X_i + N_i \tag{6}$$

$$\text{Receiver's estimates \& transmissions: } \widehat{\Theta}_i = \widehat{\Theta}_{i-1} + \frac{Y_i}{i} \tag{7}$$

where R_{i-1} is the AWGN noise in the reverse link. $R_i \sim \mathcal{N}(0,\sigma_R^2)$ are iid. random variables.

Linear strategies are known to be suboptimal for this communication problem [28] (where Θ is a quantized random variable communicating a finite-rate message reliably), and for problems with non-classical information structures in general [29]. Nevertheless, we now show that the resulting estimates $\widehat{\Theta}_i$ still converge to Θ in mean-square sense:

$$\begin{aligned}
i\widehat{\Theta}_i &= i\widehat{\Theta}_{i-1} + X_i + N_i \\
&= i\widehat{\Theta}_{i-1} + \Theta - \widehat{\Theta}_{i-1} - R_{i-1} + N_i \\
&= (i-1)\widehat{\Theta}_{i-1} + \Theta - R_{i-1} + N_i \qquad (8) \\
&\overset{(a)}{=} i\Theta + \sum_{k=1}^{i} N_k - \sum_{k=1}^{i-1} R_k \\
\widehat{\Theta}_i &= \Theta + \frac{1}{i}\sum_{k=1}^{i} N_k - \frac{1}{i}\sum_{k=1}^{i-1} R_k \qquad (9)
\end{aligned}$$

[3] We assume that the power constraints are such that the scaling constant 'α' in [25] is 1.

[4] We note that implicitly, this strategy assumes an energy/SNR constraint on the feedback link. This is because the receiver simply sends back the estimate, which is shown to converge to the true value of Θ.

where (a) is obtained by expanding $\widehat{\Theta}_j$ recursively for $j = i-1, i-2, \ldots, 2$. Therefore, the error in estimating Θ converges to 0 in mean-square sense (*i.e.*, $\widehat{\Theta}_i$ is a consistent estimate of Θ_i even in the presence of noise on the feedback link).

III. Granger causality and directed information analyses for the strategies in Section II

A. Granger causality for the noiseless feedback case

In order to compute the Granger causality in the reverse direction (from the receiver to the transmitter), we model X_i as a linear function of its past.

$$X_i = \sum_{j=1}^{p} \alpha_j X_{i-j} + \epsilon_i \tag{10}$$

We then compute coefficients α_j such that the average error in fitting X_i is minimized. Note that α_j can themselves depend on i, since this is a non-stationary process (because the error $\Theta - \widehat{\Theta}_i = X_i$ converges to zero). We describe how these coefficients might be estimated in a more general setting, and justify using theoretically determined system parameters as regression coefficients in section III-D.

For now, we theoretically evaluate the system parameters. We start with equation (1) and manipulate terms to arrive at an equation bearing the required form of equation (10):

$$\begin{aligned} X_i &= \Theta - \widehat{\Theta}_{i-1} \\ &= \Theta - \big(\widehat{\Theta}_{i-2} + \frac{Y_{i-1}}{i-1}\big) \\ &= \Theta - (\Theta - X_{i-1}) - \frac{X_{i-1} + N_{i-1}}{i-1} \\ &= X_{i-1} - \frac{X_{i-1} + N_{i-1}}{i-1} \\ &= \frac{i-2}{i-1} X_{i-1} + \frac{N_{i-1}}{i-1} \end{aligned}$$

Therefore, $\alpha_1 = \frac{i-2}{i-1}$ and $\epsilon_i = \frac{N_{i-1}}{i-1}$, and hence $\mathrm{Var}(\epsilon_i) = \frac{\sigma_N^2}{(i-1)^2}$.

Next, we model X_i in terms of both the past of X and the past of $\widehat{\Theta}$:

$$X_i = \sum_{j=1}^{p} \alpha_j X_{i-j} + \sum_{j=1}^{p} \beta_j \widehat{\Theta}_{i-j} + \tilde{\epsilon}_i \tag{11}$$

We can manipulate equation (1) to bring it into the above form:

$$\begin{aligned} X_i &= \Theta - \widehat{\Theta}_{i-1} \\ &= X_{i-1} + \widehat{\Theta}_{i-2} - \widehat{\Theta}_{i-1} \end{aligned}$$

Since there is no noise expression here, the Granger causality ratio, $\mathrm{Var}(\epsilon_i)/\mathrm{Var}(\tilde{\epsilon}_i)$ goes to infinity.

In the forward direction, we do not explicitly compute the Granger causality ratio, but simply show that it is bounded strictly between 1 and ∞:

$$\widehat{\Theta}_i = \sum_{j=1}^{p} \alpha_j \widehat{\Theta}_{i-j} + \epsilon_i \tag{12}$$

$$\begin{aligned} \widehat{\Theta}_i &= \widehat{\Theta}_{i-1} + \frac{X_i + N_i}{i} \\ &= \widehat{\Theta}_{i-1} + \frac{\Theta - \widehat{\Theta}_{i-1} + N_i}{i} \\ &= \frac{i-1}{i}\widehat{\Theta}_{i-1} + \frac{\Theta}{i} + \frac{N_i}{i} \end{aligned}$$

which means that $0 < \frac{\sigma_N^2}{i^2} < \mathrm{Var}(\epsilon_i) < \frac{\sigma_N^2 + \sigma_\theta^2}{i^2} < \infty$, since the past of $\widehat{\Theta}$ cannot be used to explain N_i. Further, if we try to predict $\widehat{\Theta}_i$ from the previous $\widehat{\Theta}_{i-j}$ and X_{i-j}:

$$\widehat{\Theta}_i = \sum_{j=1}^{p} \alpha_j \widehat{\Theta}_{i-j} + \sum_{j=0}^{p-1} \beta_j X_{i-j} + \tilde{\epsilon}_i \tag{13}$$

we see that

$$\widehat{\Theta}_i = \widehat{\Theta}_{i-1} + \frac{X_i}{i} + \frac{N_i}{i}$$

so that $\mathrm{Var}(\tilde{\epsilon}_i) = \sigma_N^2/i^2$. The Granger causality ratio, $\mathrm{Var}(\epsilon_i)/\mathrm{Var}(\tilde{\epsilon}_i)$, in the forward direction is, therefore, finite.

The intuitive argument for why this is happening might go as follows: since the feedback link is noiseless, one can always perfectly predict the transmitted symbol from the past $\widehat{\Theta}$'s and the history of X. On the other hand, one can never perfectly predict $\widehat{\Theta}_i$ from the past X's and the history of $\widehat{\Theta}$.

B. Directed Information for the noiseless feedback case

Performing the Directed Information analysis for the scheme described above yields the same results. In order to ease the burden of computing Directed Information, we assume that Θ is normally distributed.

The directed information in the forward direction is computed as:

$$I(X^n \to \widehat{\Theta}^n) = \frac{1}{2}\log\left(1 + \frac{n\sigma_\theta^2}{\sigma_N^2}\right)$$

where n is the number of iterations of the Schalkwijk-Kailath algorithm. For a proof, refer appendix I, which appears in the full version of this paper [30].

In the reverse direction, the Directed Information is ∞.

$$\begin{aligned} I(0 * \widehat{\Theta}^{n-1} \to X^n) &= \sum_{i=0}^{n-1} I(X_{i+1}; \widehat{\Theta}^i | X^i) \\ I(X_{i+1}; \widehat{\Theta}^i | X^i) &= h(X_{i+1}|X^i) - h(X_{i+1}|X^i, \widehat{\Theta}^i) \end{aligned} \tag{14}$$

The first term in the equation above reduces to

$$\begin{aligned} h(X_{i+1}|X^i) &= h(\Theta - \widehat{\Theta}_i | X^i) \\ &= h\bigg(\Theta - \bigg(\widehat{\Theta}_{i-1} + \frac{X_i + N_i}{i}\bigg)\bigg| X^i\bigg) \\ &\overset{(a)}{=} h\bigg(\Theta - \bigg((\Theta - X_i) + \frac{N_i}{i}\bigg)\bigg| X^i\bigg) \\ &= h\bigg(\frac{N_i}{i}\bigg| X^i\bigg) \\ &= h(N_i) - \log(i) \\ &= \frac{1}{2}\log(2\pi e \sigma_N^2) - \log(i) \end{aligned}$$

where for (a), we have dropped X_i/i from the previous step, since it is conditioned over, and then written $\widehat{\Theta}_{i-1}$ as $(\Theta - X_i)$. On the other hand, the second term in equation (14) becomes

$$\begin{aligned} h(X_{i+1}|X^i, \widehat{\Theta}^i) &= h(\Theta - \widehat{\Theta}_i|X^i, \widehat{\Theta}^i) \\ &= h(\Theta|X^i, \widehat{\Theta}^i) \\ &\overset{(a)}{=} h(X_i + \widehat{\Theta}_{i-1}|X^i, \widehat{\Theta}^i) \\ &= h(0|X^i, \widehat{\Theta}^i) \\ &\overset{(b)}{=} -\infty \end{aligned}$$

where for (a) we have expressed Θ in terms of $\widehat{\Theta}_{i-1}$ and X_i and for (b), we have used the fact that the differential entropy of a constant (or equivalently, a Gaussian with zero variance) is negative infinity. This means that equation (14) becomes

$$\begin{aligned} I(X_{i+1}; \widehat{\Theta}^i|X^i) &= \infty \\ \Rightarrow I(0 * \widehat{\Theta}^{n-1} \to X^n) &= \infty \end{aligned}$$

C. Directed Information for the noisy feedback scenario

Since a noiseless feedback link is not realistic, we proceed to perform the same Directed Information calculations as above for the feedback link with additive white Gaussian noise of variance σ_R^2. While we could not derive simple closed form expressions for the Directed Information in the forward and reverse links, we were able to evaluate the expressions numerically. These are plotted in section IV.

The Directed Information in the forward direction can be written as

$$\begin{aligned} I(X^n \to \widehat{\Theta}^n) &= \sum_{i=1}^{n} I(\widehat{\Theta}_i; X^i|\widehat{\Theta}^{i-1}) \\ &= \sum_{i=1}^{n} h(\widehat{\Theta}_i|\widehat{\Theta}^{i-1}) - h(\widehat{\Theta}_i|\widehat{\Theta}^{i-1}, X^i) \qquad (15) \\ &= \sum_{i=1}^{n} \left(\frac{1}{2}\log(2\pi e \text{Var}[\Theta - R_{i-1} + N_i|\widehat{\Theta}^{i-1}]) \right. \\ &\qquad \left. - \frac{1}{2}\log(2\pi e \sigma_N^2) \right) \end{aligned}$$

For a derivation of this, see appendix II which appears in [30]. In the reverse direction,

$$\begin{aligned} I(0 * \widehat{\Theta}^{n-1} \to X^n) &= \sum_{i=0}^{n-1} I(X_{i+1}; \widehat{\Theta}^i|X^i) \\ &= \sum_{i=0}^{n-1} h(X_{i+1}|X^i) - h(X_{i+1}|X^i, \widehat{\Theta}^i) \qquad (16) \\ &= \frac{1}{2}\log\left(2\pi e \frac{\sigma_N^2 + \sigma_R^2}{\sigma_R^2}\right) \\ &+ \sum_{i=2}^{n-1} \left(\frac{1}{2}\log\left(2\pi e \text{Var}\left[R_{i-1} - \frac{N_i}{i} - R_i \middle| X^i\right]\right) \right. \\ &\qquad \left. - \frac{1}{2}\log(2\pi e \sigma_R^2) \right) \end{aligned}$$

For a derivation of this, see appendix III which appears in [30].

D. A note on estimating regression coefficients

The Granger Causality and Directed Information metrics are a function of the regression coefficients estimated from the data by fitting the models given by equations (10) and (11). In our analysis, we described the data-generation model: the algorithm inspired by the Schalkwijk and Kailath scheme for feedback communication. We then proceeded to use the system parameters of this model directly as regression coefficients in our Granger Causality computation. This could be construed as being erroneous: we ought to simulate the data generation, and estimate the regression coefficients from the generated data. This would better model the actions of the neuroscientist who seeks to perform Granger Causality analysis.

We justify our knowledge of the system parameters and their use as regression coefficients in the following manner: we assume that the regression coefficients can be accurately estimated from data, since neuroscientific experiments typically record the *same* processes multiple times – these are called "trials". The availability of multiple trials of the same process can be leveraged to estimate the system parameters accurately, even if the processes are non-stationary.

Suppose we record two non-stationary processes, $\{X_t\}_{t=1}^n$ and $\{Y_t\}_{t=1}^n$, for which we seek to compute the Granger Causality metrics. To this end, we must find coefficients $\alpha_j(t)$ and $\beta_j(t)$ to minimize the error in fitting the models given by equations (10) and (11). Note that α and β depend on t, since we assume the process is non-stationary. However, since we record the *same* process in each trial, the $\alpha_j(t)$ and $\beta_j(t)$ are constant across trials. Estimating them from data that has many trials is then a simple matter of linear regression.

For a given time instant t, the i^{th} trial is modeled as

$$X_t^{(i)} = \sum_{j=1}^{p} \alpha_j(t) X_{t-j}^{(i)} + \epsilon_t^{(i)}$$

Note that $\alpha_j(t)$ does not depend on i. Collecting variables across N trials, we can write the full model in vector form:

$$\begin{bmatrix} X_t^{(1)} \\ X_t^{(2)} \\ \vdots \\ X_t^{(N)} \end{bmatrix} = \begin{bmatrix} X_{t-1}^{(1)} & \cdots & X_{t-p}^{(1)} \\ X_{t-1}^{(2)} & \cdots & X_{t-p}^{(2)} \\ \vdots & \ddots & \vdots \\ X_{t-1}^{(N)} & \cdots & X_{t-p}^{(N)} \end{bmatrix} \begin{bmatrix} \alpha_1(t) \\ \alpha_2(t) \\ \vdots \\ \alpha_p(t) \end{bmatrix} + \begin{bmatrix} \epsilon_t^{(1)} \\ \epsilon_t^{(2)} \\ \vdots \\ \epsilon_t^{(N)} \end{bmatrix}$$

If we call the vector on the LHS $\mathbf{Y}$ and the matrix on the RHS $\mathbf{X}$, then the vector of $\alpha_j(t)$'s ($\alpha(\mathbf{t})$) can be estimated at time instant t using Ordinary Least Squares:

$$\hat{\alpha}(t) = (\mathbf{X^T X})^{-1}\mathbf{X^T Y}$$

This is an unbiased and consistent estimator for $\alpha(\mathbf{t})$. This analysis can be trivially extended to the model described by equation (11). With a sufficiently large number of trials, therefore, the system parameters (to be used as regression coefficients in the Granger Causality analysis) can be estimated to arbitrarily high accuracy.

It should be noted that we have restricted ourselves to an analysis of linear strategies: the channel, the extended Schalkwijk and Kailath scheme to a noisy feedback link, and the proposed regression model are all linear.

As a final remark, we note that estimating the regression coefficients accurately is a conservative assumption on our part. As mentioned at the end of section I-C, we see that *despite* being computed accurately, the metrics of Granger Causality and Directed Information incorrectly estimate the direction of information flow. A rigorous analysis would warrant the computation of these coefficients in simulation, for a finite number of trials. It is our belief, however, that our result is unlikely to degrade if the coefficients are not estimated perfectly. Future work will address this matter in greater depth.

IV. Numerical results demonstrating that Granger Causal influence can be opposite in direction to information flow

In the noiseless feedback case, the directed information on the forward link is finite, while that on the feedback link is infinite (refer sections III-A and III-B). However, for completeness, we examine the case when noise is present in the feedback link, as is illustrated in Fig. 4, through numerical calculation of expressions in the last section. For cases where the noise variance of the feedback link (σ_R^2) is moderately smaller than feedforward noise variance (σ_N^2), we observe that the Directed Information in the direction of the reverse link can dominate that in the direction of the forward link. With sufficiently many (albeit sometimes a large number of) iterations, the Directed Information in forward direction starts to dominate. However, the point at which this happens depends on the (often unknown) ratio of noise in these links.

V. Conclusions and discussions

We demonstrate, by means of a concrete counter-example, that the direction predicted by causal influence metrics such as Granger Causality and Directed Information can be opposite to the true direction of information flow. There are, however, several shortcomings to our analysis, which we list in section I-E. We seek to address many of these shortcomings in future work.

It might appear that we make a circular argument while computing the Granger Causal influences in our counter-example, since we supply the model for the stochastic process and use the system parameters of the model directly to compute the Granger Causality metric. However, as we state in Section III-D, we assume that the regression coefficients of the Autoregressive model can be exactly estimated (even if the AR process is non-stationary), and discuss how this might be achieved with the help of multiple trials.

As a final remark, we emphasize that this work only demonstrates the error in *interpreting* the direction of causal

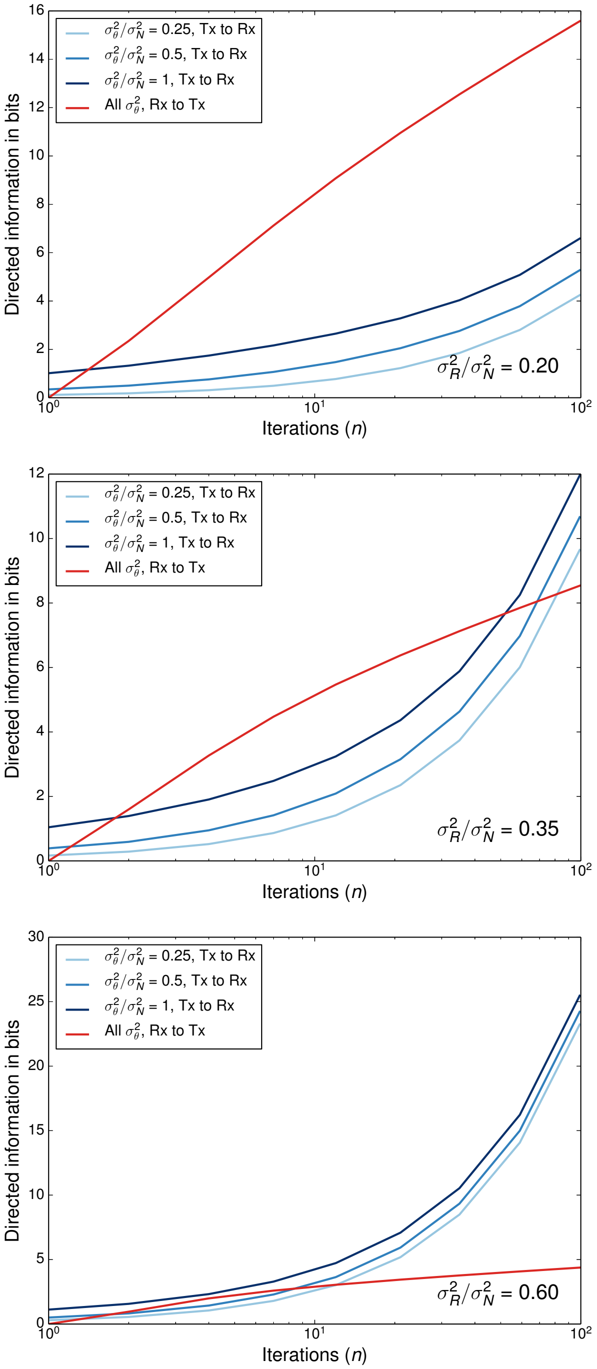

Fig. 4. Plots for forward and backward directed information computations for $\sigma_R^2/\sigma_N^2 = 0.2$ (top), 0.35 (center) and 0.6 (bottom). In each plot, curves for directed information in both directions are illustrated for ratios $\sigma_\theta^2/\sigma_N^2 = 0.25, 0.5,$ and 1. The x-axis is the number of iterations of message-passing between the transmitter and the receiver. For cases when feedback noise variance σ_R^2 is moderately smaller than feedforward noise variance σ_N^2, directed information in the reverse link can dominate that in the forward link. With sufficiently many (albeit sometimes large, as illustrated in the top figure) iterations, directed information in forward information starts to dominate. However, the point at which this happens depends on the (often unknown) ratio of noise in these links.

influence as the direction of *information flow*. We do *not* seek to invalidate much of the neuroscientific work that has been done in this direction; we merely caution against making (what might be construed as hopeful) extrapolations from causal influences to information flows.

We do not seek to understate the importance of determining causal influences in the brain; understanding causal influence itself may have a great deal of benefit. For instance, we might seek to understand the spread of activity in the brain during an epileptic seizure – in such applications, we are not concerned with how information is being transferred through the neural circuitry; we only seek to determine the source of the activity for the purpose of surgical intervention.

Acknowledgments

The authors would like to thank Nicola Elia for discussions that suggested this possible misunderstanding in use of Granger causality and Directed Information in computational systems. We also thank Rob Kass, Bruno Sinopoli, Momin Malik and Kun Zhang for useful discussions. Finally, we thank the reviewers of a previous conference submission of this work for their insightful remarks and comments. We thank the CMU BrainHUB initiative for support through a Proseed-BrainHUB seed grant. Pulkit Grover was supported through a CMU startup grant, and Praveen Venkatesh through a CMU CIT Dean's fellowship and a CMU Presidential Fellowship.

References

[1] K. J. Blinowska, R. Kuś, and M. Kamiński, "Granger causality and information flow in multivariate processes," *Physical Review E*, vol. 70, no. 5, p. 050902, 2004.

[2] M. Dhamala, G. Rangarajan, and M. Ding, "Analyzing information flow in brain networks with nonparametric Granger causality," *NeuroImage*, vol. 41, no. 2, pp. 354–362, 2008.

[3] G. Nolte, A. Ziehe, V. V. Nikulin, A. Schlögl, N. Krämer, T. Brismar, and K.-R. Müller, "Robustly estimating the flow direction of information in complex physical systems," *Physical review letters*, vol. 100, no. 23, p. 234101, 2008.

[4] A. Korzeniewska, M. Mańczak, M. Kamiński, K. J. Blinowska, and S. Kasicki, "Determination of information flow direction among brain structures by a modified directed transfer function (dDTF) method," *Journal of neuroscience methods*, vol. 125, no. 1, pp. 195–207, 2003.

[5] M. B. Schippers, A. Roebroeck, R. Renken, L. Nanetti, and C. Keysers, "Mapping the information flow from one brain to another during gestural communication," *Proceedings of the National Academy of Sciences*, vol. 107, no. 20, pp. 9388–9393, 2010.

[6] A. Brovelli, M. Ding, A. Ledberg, Y. Chen, R. Nakamura, and S. L. Bressler, "Beta oscillations in a large-scale sensorimotor cortical network: directional influences revealed by Granger causality," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 26, pp. 9849–9854, 2004.

[7] R. Goebel, A. Roebroeck, D.-S. Kim, and E. Formisano, "Investigating directed cortical interactions in time-resolved fMRI data using vector autoregressive modeling and Granger causality mapping," *Magnetic resonance imaging*, vol. 21, no. 10, pp. 1251–1261, 2003.

[8] G. Deshpande, X. Hu, R. Stilla, and K. Sathian, "Effective connectivity during haptic perception: a study using Granger causality analysis of functional magnetic resonance imaging data," *Neuroimage*, vol. 40, no. 4, pp. 1807–1814, 2008.

[9] K. Friston, R. Moran, and A. K. Seth, "Analysing connectivity with Granger causality and dynamic causal modelling," *Current opinion in neurobiology*, vol. 23, no. 2, pp. 172–178, 2013.

[10] C. Bernasconi, A. von Stein, C. Chiang, and P. KoÈnig, "Bi-directional interactions between visual areas in the awake behaving cat," *Neuroreport*, vol. 11, no. 4, pp. 689–692, 2000.

[11] M. Ding, S. L. Bressler, W. Yang, and H. Liang, "Short-window spectral analysis of cortical event-related potentials by adaptive multivariate autoregressive modeling: data preprocessing, model validation, and variability assessment," *Biological cybernetics*, vol. 83, no. 1, pp. 35–45, 2000.

[12] A. Roebroeck, E. Formisano, and R. Goebel, "Mapping directed influence over the brain using Granger causality and fMRI," *Neuroimage*, vol. 25, no. 1, pp. 230–242, 2005.

[13] S. L. Bressler and A. K. Seth, "Wiener–Granger causality: a well established methodology," *Neuroimage*, vol. 58, no. 2, pp. 323–329, 2011.

[14] L. Barnett and A. K. Seth, "The MVGC multivariate Granger causality toolbox: a new approach to Granger-causal inference," *Journal of neuroscience methods*, vol. 223, pp. 50–68, 2014.

[15] M. Ding, Y. Chen, and S. L. Bressler, "17 Granger causality: basic theory and application to neuroscience," *Handbook of time series analysis: recent theoretical developments and applications*, p. 437, 2006.

[16] C. W. Granger, "Investigating causal relations by econometric models and cross-spectral methods," *Econometrica: Journal of the Econometric Society*, pp. 424–438, 1969.

[17] J. Massey, "Causality, feedback and directed information," in *Proc. Int. Symp. Inf. Theory Applic.(ISITA-90)*. Citeseer, 1990, pp. 303–305.

[18] C. J. Quinn, T. P. Coleman, N. Kiyavash, and N. G. Hatsopoulos, "Estimating the directed information to infer causal relationships in ensemble neural spike train recordings," *Journal of computational neuroscience*, vol. 30, no. 1, pp. 17–44, 2011.

[19] J. Jiao, H. H. Permuter, L. Zhao, Y.-H. Kim, and T. Weissman, "Universal estimation of directed information," *Information Theory, IEEE Transactions on*, vol. 59, no. 10, pp. 6220–6242, 2013.

[20] W. Hesse, E. Möller, M. Arnold, and B. Schack, "The use of time-variant EEG Granger causality for inspecting directed interdependencies of neural assemblies," *Journal of neuroscience methods*, vol. 124, no. 1, pp. 27–44, 2003.

[21] J. Pearl, *Causality*. Cambridge university press, 2009, pp. 54–57.

[22] H. Nalatore, M. Ding, and G. Rangarajan, "Mitigating the effects of measurement noise on granger causality," *Physical Review E*, vol. 75, no. 3, p. 031123, 2007.

[23] J. Andersson, "Testing for granger causality in the presence of measurement errors," *Economics Bulletin*, 2005.

[24] M. Gong, S. UTS, K. Zhang, M. DE, B. Schölkopf, D. Tao, and P. Geiger, "Discovering temporal causal relations from subsampled data," in *Proceedings of The 32nd International Conference on Machine Learning*, 2015, pp. 1898–1906.

[25] J. Schalkwijk and T. Kailath, "A coding scheme for additive noise channels with feedback–i: No bandwidth constraint," *Information Theory, IEEE Transactions on*, vol. 12, no. 2, pp. 172–182, 1966.

[26] C. W. Granger, "Testing for causality: a personal viewpoint," *Journal of Economic Dynamics and control*, vol. 2, pp. 329–352, 1980.

[27] J. Pearl, *Causality*. Cambridge university press, 2009, p. 39.

[28] Y.-H. Kim, A. Lapidoth, and T. Weissman, "The Gaussian channel with noisy feedback," in *Information Theory, IEEE International Symposium on*. IEEE, 2007, pp. 1416–1420.

[29] H. S. Witsenhausen, "A counterexample in stochastic optimum control," *SIAM Journal on Control*, vol. 6, no. 1, pp. 131–147, 1968.

[30] *Full version of this paper, including appendices*. [Online]. Available: http://tinyurl.com/pulkitgrover/files/Allerton15Granger.pdf

APPENDIX I
DIRECTED INFORMATION IN THE FORWARD DIRECTION, NOISELESS FEEDBACK

The Directed Information in the forward direction is computed as:

$$\begin{aligned} I(X^n \to \widehat{\Theta}^n) &= \sum_{i=1}^{n} I(\widehat{\Theta}_i; X^i|\widehat{\Theta}^{i-1}) \\ &= \sum_{i=1}^{n} h(\widehat{\Theta}_i|\widehat{\Theta}^{i-1}) - h(\widehat{\Theta}_i|\widehat{\Theta}^{i-1}, X^i) \quad (17) \end{aligned}$$

Taking the first term in (17),

$$\begin{aligned} h(\widehat{\Theta}_i|\widehat{\Theta}^{i-1}) &= h\left(\widehat{\Theta}_{i-1} + \frac{X_i + N_i}{i}\middle|\widehat{\Theta}^{i-1}\right) \\ &= h(\Theta - \widehat{\Theta}_{i-1} + N_i|\widehat{\Theta}^{i-1}) - \log(i) \\ &= h(\Theta + N_i|\widehat{\Theta}^{i-1}) - \log(i) \\ &= h(\Theta + N_i|\widehat{\Theta}_{i-1}) - \log(i) \end{aligned}$$

where we have dropped the conditioning on all except $\widehat{\Theta}_{i-1}$ in the last step. Define $U = \Theta + N_i$ and $V = \widehat{\Theta}_{i-1}$. Since all variables are Gaussian, it suffices to find the variance of the conditional distribution $U|V$.

$$\begin{aligned} \mathbb{E}[U] = 0, \mathbb{E}[V] = 0, \text{Var}[U] = \sigma_\theta^2 + \sigma_N^2, \text{Var}[V] &= \sigma_\theta^2 + \frac{\sigma_N^2}{i-1} \\ \text{Cov}[U,V] &= \mathbb{E}[UV] - \mathbb{E}[U]\mathbb{E}[V] \\ &= \sigma_\theta^2 \\ \rho^2 &= \frac{\sigma_\theta^4}{(\sigma_\theta^2 + \sigma_N^2)(\sigma_\theta^2 + \frac{\sigma_N^2}{(i-1)})} \end{aligned}$$

$$U|V = v \ \sim \ \mathcal{N}\left(\sqrt{\frac{\sigma_\theta^2 + \sigma_N^2}{\sigma_\theta^2 + \frac{\sigma_N^2}{i-1}}}\rho v, (1-\rho^2)(\sigma_\theta^2 + \sigma_N^2)\right)$$

Hence, the entropy of the conditional distribution is

$$h(U|V = v) = \frac{1}{2}\log(2\pi e(1-\rho^2)(\sigma_\theta^2 + \sigma_N^2)) \overset{(a)}{=} h(U|V)$$

where (a) follows because the conditional entropy is independent of v. Thus,

$$h(\widehat{\Theta}_i|\widehat{\Theta}^{i-1}) = \frac{1}{2}\log\left(2\pi e\frac{\sigma_N^2(i\sigma_\theta^2 + \sigma_N^2)}{((i-1)\sigma_\theta^2 + \sigma_N^2)}\right) - \log(i) \quad (18)$$

The next term in equation (17) is

$$\begin{aligned} h(\widehat{\Theta}_i|\widehat{\Theta}^{i-1}, X^i) &= h\left(\frac{N_i}{i}\middle|\widehat{\Theta}^{i-1}, X^i\right) \\ &= h(N_i) - \log(i) \\ &= \frac{1}{2}\log(2\pi e\sigma_N^2) - \log(i) \quad (19) \end{aligned}$$

Putting equations (18) and (19) together, we can compute the forward Directed Information:

$$\begin{aligned} I(\widehat{\Theta}_i; X^i|\widehat{\Theta}^{i-1}) &= h(\widehat{\Theta}_i|\widehat{\Theta}^{i-1}) - h(\widehat{\Theta}_i|\widehat{\Theta}^{i-1}, X^i) \\ &= \frac{1}{2}\log\left(\frac{i\sigma_\theta^2 + \sigma_N^2}{(i-1)\sigma_\theta^2 + \sigma_N^2}\right) \\ I(X^n \to \widehat{\Theta}^n) &= \sum_{i=1}^{n} I(\widehat{\Theta}_i; X^i|\widehat{\Theta}^{i-1}) \\ &\overset{(a)}{=} \frac{1}{2}\log\left(\frac{n\sigma_\theta^2 + \sigma_N^2}{\sigma_N^2}\right) \\ &= \frac{1}{2}\log\left(1 + \frac{n\sigma_\theta^2}{\sigma_N^2}\right) \end{aligned}$$

where (a) follows through by expanding out the product inside the logarithms and canceling terms. Clearly, this value is finite.

APPENDIX II
DIRECTED INFORMATION IN THE FORWARD DIRECTION, WITH NOISY FEEDBACK

$$\begin{aligned} I(X^n \to \widehat{\Theta}^n) &= \sum_{i=1}^{n} I(\widehat{\Theta}_i; X^i|\widehat{\Theta}^{i-1}) \\ &= \sum_{i=1}^{n} h(\widehat{\Theta}_i|\widehat{\Theta}^{i-1}) - h(\widehat{\Theta}_i|\widehat{\Theta}^{i-1}, X^i) \end{aligned}$$

Taking the first of the two terms in the above expression,

$$\begin{aligned} h(\widehat{\Theta}_i|\widehat{\Theta}^{i-1}) &\overset{(a)}{=} h(\widehat{\Theta}_{i-1}\frac{i-1}{i} + \frac{\Theta}{i} - \frac{R_{i-1}}{i} + \frac{N_i}{i}|\widehat{\Theta}^{i-1}) \\ &= h(\Theta - R_{i-1} + N_i|\widehat{\Theta}^{i-1}) - \log(i) \end{aligned}$$

where for (a) we have used equation (8). The Markov property no longer holds in this case, but we proceed in the same manner. We define $U = \Theta - R_{i-1} + N_i$ and $\underline{\text{V}} = \widehat{\Theta}^{i-1}$. Recalling equation (9), for $j \in \{1, \dots i-1\}$, $p \in \{1, \dots i-1\}$ and $q \in \{1, \dots i-1\}$, we have

$$\begin{aligned} \mathbb{E}[U] &= 0, \ \mathbb{E}[\widehat{\Theta}_j] = 0, \\ \mathbb{E}[U^2] &= \sigma_\theta^2 + \sigma_R^2 + \sigma_N^2, \ \mathbb{E}[U\widehat{\Theta}_j] = \sigma_\theta^2 \\ \mathbb{E}[\widehat{\Theta}_p\widehat{\Theta}_q] &= \mathbb{E}\Bigg[\left(\Theta + \frac{1}{p}\sum_{k=1}^{p} N_k - \frac{1}{p}\sum_{k=1}^{p-1} R_k\right) \\ &\qquad \left(\Theta + \frac{1}{q}\sum_{k=1}^{q} N_k - \frac{1}{q}\sum_{k=1}^{q-1} R_k\right)\Bigg] \\ &= \sigma_\theta^2 + \frac{\min\{p,q\}}{pq}\sigma_N^2 + \frac{\min\{p-1,q-1\}}{pq}\sigma_R^2 \\ \text{Var}[U|\widehat{\Theta}^{i-1}] &= \mathbb{E}[U^2] - \mathbb{E}[U\underline{\text{V}}]\mathbb{E}[\underline{\text{V}}\underline{\text{V}}^T]^{-1}\mathbb{E}[\underline{\text{V}}U] \\ h(U|\widehat{\Theta}^{i-1}) &= \frac{1}{2}\log(2\pi e\text{Var}[U|\widehat{\Theta}^{i-1}]) \quad (20) \end{aligned}$$

We can not derive a simple closed form for this expression, but we have computed it numerically for the plots in Section IV. The second term in equation (15) is

$$
\begin{aligned}
h(\widehat{\Theta}_i|\widehat{\Theta}^{i-1},X^i) &= h\left(\widehat{\Theta}_{i-1}+\frac{X_i+N_i}{i}\middle|\widehat{\Theta}^{i-1},X^i\right)\\
&= h(N_i|\widehat{\Theta}^{i-1},X^i)-\log(i)\\
&= \frac{1}{2}\log(2\pi e\sigma_N^2)-\log(i) \qquad (21)
\end{aligned}
$$

From equations (20) and (21), we compute the forward-directed information as depicted in Section IV, for different values of σ_θ^2.

Appendix III
Directed Information in the reverse direction, with noisy feedback

First, we derive an expression for X_i, which we will use later.

$$
\begin{aligned}
X_i &= \Theta-\widehat{\Theta}_{i-1}-R_{i-1}\\
&= \Theta-\left(\Theta+\frac{1}{i-1}\sum_{k=1}^{i-1}N_k-\frac{1}{i-1}\sum_{k=1}^{i-2}R_k\right)-R_{i-1}\\
&= \frac{1}{i-1}\sum_{k=1}^{i-2}R_k-\frac{1}{i-1}\sum_{k=1}^{i-1}N_k-R_i \qquad (22)
\end{aligned}
$$

$$
\begin{aligned}
I(0*\widehat{\Theta}^{n-1}\to X^n) &= \sum_{i=0}^{n-1}I(X_{i+1};\widehat{\Theta}^i|X^i)\\
&= \sum_{i=0}^{n-1}h(X_{i+1}|X^i)-h(X_{i+1}|X^i,\widehat{\Theta}^i)
\end{aligned}
\qquad (23)
$$

Taking the first term inside the summation,

$$
\begin{aligned}
h(X_{i+1}|X^i) &= h(\Theta-\widehat{\Theta}_i-R_i|X^i)\\
&\overset{(a)}{=} h\left(\left(-\widehat{\Theta}_{i-1}-\frac{X_i+N_i}{i}\right)-R_i\middle|X^i\right)\\
&\overset{(b)}{=} h\left(\left(X_i-\Theta+R_{i-1}-\frac{N_i}{i}\right)-R_i\middle|X^i\right)\\
&\overset{(c)}{=} h\left(R_{i-1}-\frac{N_i}{i}-R_i\middle|X^i\right)
\end{aligned}
$$

where in (a) above, we have dropped $\Theta = X_1$, in (b) we have re-expressed $\widehat{\Theta}_{i-1}$ in terms of X_i, Θ and R_{i-1}, and in (c) we have dropped X_i and Θ again. As before, define $U = R_{i-1}-\frac{N_i}{i}-R_i$, so that

$$
\mathbb{E}[U]=0,\ \ \mathbb{E}[X_j]=0,\ \ \mathbb{E}[U^2]=2\sigma_R^2+\frac{\sigma_N^2}{i^2}
$$

For $i \geq 3$ and $j \in \{3,\dots i\}$, we can use equation (22) to see that

$$
\begin{aligned}
\mathbb{E}[UX_j] &= \mathbb{E}\Bigg[\left(R_{i-1}-\frac{N_i}{i}-R_i\right)\\
&\qquad\left(\frac{1}{j-1}\sum_{k=1}^{j-2}R_k-\frac{1}{j-1}\sum_{k=1}^{j-1}N_k-R_{j-1}\right)\Bigg]\\
&= -\mathbb{E}[R_{i-1}R_{j-1}] = -\sigma_R^2\delta_{ij}\\
\mathbb{E}[UX_1] &= \mathbb{E}\left[\left(R_{i-1}-\frac{N_i}{i}-R_i\right)\Theta\right]=0\\
\mathbb{E}[UX_2] &= \mathbb{E}\left[\left(R_{i-1}-\frac{N_i}{i}-R_i\right)(\Theta-(\Theta+N_i)-R_1)\right]=0\\
\mathbb{E}[X_pX_q] &= \mathbb{E}\Bigg[\left(\frac{1}{p-1}\sum_{k=1}^{p-2}R_k-\frac{1}{p-1}\sum_{k=1}^{p-1}N_k-R_{p-1}\right)\\
&\qquad\left(\frac{1}{q-1}\sum_{k=1}^{q-2}R_k-\frac{1}{q-1}\sum_{k=1}^{q-1}N_k-R_{q-1}\right)\Bigg]\\
&= \frac{\min\{p-2,q-2\}}{(p-1)(q-1)}\sigma_R^2\\
&\quad+\frac{\min\{p-1,q-1\}}{(p-1)(q-1)}\sigma_N^2\\
&\quad+\sigma_R^2\delta_{pq}+\frac{1}{p-1}\sigma_R^2\mathbb{I}_{p>q}+\frac{1}{q-1}\sigma_R^2\mathbb{I}_{q>p}\\
\mathbb{E}[X_1X_1] &= \mathbb{E}[\Theta^2]=\sigma_\theta^2\\
\mathbb{E}[X_1X_j] &= \mathbb{E}\Bigg[\Theta\Bigg(\frac{1}{j-1}\sum_{k=1}^{j-2}R_k\\
&\qquad-\frac{1}{j-1}\sum_{k=1}^{j-1}N_k-R_{j-1}\Bigg)\Bigg]=0\\
\text{Var}[U|X^i] &= \mathbb{E}[U^2]-\mathbb{E}[UX^i]\mathbb{E}[X^i{X^i}^T]^{-1}\mathbb{E}[X^iU]\\
h(U|X^i) &= \frac{1}{2}\log(2\pi e\text{Var}[U|X^i]) \qquad (24)
\end{aligned}
$$

The above argument can be extended to $i = 2$ by letting the final index of the summation term be smaller than the starting index, implying that the whole summation term is simply dropped. The special cases of $i = 0$ and $i = 1$ are handled at the end. The second term from equation (16) becomes

$$
\begin{aligned}
h(X_{i+1}|X^i,\widehat{\Theta}^i) &= h(\Theta-\widehat{\Theta}_i-R_i|X^i,\widehat{\Theta}^i)\\
&= \frac{1}{2}\log(2\pi e\sigma_R^2) \qquad (25)
\end{aligned}
$$

because $X_1 = \Theta$ and because R_i is independent of all the X^i and $\widehat{\Theta}^i$. For the special cases of $i = 0$ and $i = 1$, we solve for the value of mutual information explicitly:

$$
\begin{aligned}
i=0: &\quad I(X_1;\widehat{\Theta}^0|X^0) &&= h(X_1)-h(X_1)=0\\
i=1: &\quad I(X_2;\widehat{\Theta}^1|X^1) &&= h(X_2|X_1)-h(X_2|X_1,\widehat{\Theta}_1)\\
&\quad h(X_2|X_1) &&= h(\Theta-\widehat{\Theta}_1-R_1|X^1)\\
& &&= h(-\Theta-N_1-R_1|X_1)\\
& &&= h(-N_1-R_1)\\
& &&= \frac{1}{2}\log(2\pi e(\sigma_N^2+\sigma_R^2))
\end{aligned}
$$

$$\begin{aligned} h(X_2|X_1,\widehat{\Theta}_1) &= h(-R_1) = \frac{1}{2}\log(2\pi e\sigma_R^2) \\ I(X_2;\widehat{\Theta}^1|X^1) &= \frac{1}{2}\log\left(2\pi e\frac{\sigma_N^2+\sigma_R^2}{\sigma_R^2}\right) \end{aligned}$$

From equations (24) and (25), along with the two special cases above, we can now compute the reverse-directed information plotted in Section IV.