



LETTER

Fluctuation-response theorem for Kullback-Leibler divergences to quantify causation


To cite this article: Andrea Auconi *et al* 2021 *EPL* **135** 28002

View the [article online](#) for updates and enhancements.

You may also like

- [Estimating entropy production rates with first-passage processes](#)
Izaak Neri
- [Learning stochastic filtering](#)
Rahul O. Ramakrishnan, Andrea Auconi and Benjamin M. Friedrich
- [The fractional Kullback–Leibler divergence](#)
A Alexopoulos

Fluctuation-response theorem for Kullback-Leibler divergences to quantify causation

ANDREA AUCONI^{1(a)} , BENJAMIN M. FRIEDRICH^{1,2} and ANDREA GIANSANTI^{3,4}

¹ *cfaed, Technische Universität Dresden - 01069 Dresden, Germany*

² *Cluster of Excellence "Physics of Life" - 01307 Dresden, Germany*

³ *Dipartimento di Fisica, Sapienza Università di Roma - 00185 Rome, Italy*

⁴ *INFN, Sezione di Roma 1 - 00185 Rome, Italy*

received 22 February 2021; accepted in final form 18 June 2021

published online 29 September 2021

Abstract – We define a new measure of causation from a fluctuation-response theorem for Kullback-Leibler divergences, based on the information-theoretic cost of perturbations. This information response has both the invariance properties required for an information-theoretic measure and the physical interpretation of a propagation of perturbations. In linear systems, the information response reduces to the transfer entropy, providing a connection between Fisher and mutual information.

Copyright © 2021 EPLA

Introduction. – In the general framework of stochastic dynamical systems, the term *causation* refers to the influence that a variable x exerts over the dynamics of another variable y . Measures of causation find application in neuroscience [1], climate studies [2], cancer research [3], and finance [4]. However, a widely accepted quantitative definition of causation is still missing.

Causation manifests itself in two inseparable forms: information flow [5–8], and propagation of perturbations [9–12]. Ideally, a quantitative measure of causation should connect both perspectives.

Information flow is commonly quantified by the *transfer entropy* [13–17], that is the average conditional mutual information corresponding to the uncertainty reduction in forecasting the time evolution of y that is achieved upon knowledge of x . The mutual information is a special case of Kullback-Leibler (KL) divergence, a dimensionless measure of distinguishability between probability distributions [18]. As such, the transfer entropy abstracts from the underlying physics to give an invariant description in terms of the strength of probabilistic dependencies.

From the interventional point of view [9–12], causation is identified with how a perturbation applied to x propagates in the system to affect y . Although a direct perturbation of observables is unfeasible in most real-world situations, the fluctuation-response theorem establishes a

connection between the response to a small perturbation and the correlation of fluctuations in the natural (unperturbed) dynamics [19–22].

The fluctuation-response theorem considers the first-order expansion of the response with respect to the perturbation. The corresponding linear response coefficient has been suggested as a measure of causation [11,12]. However, it has the same physical units as y/x , and it can assume negative values; thus, is not directly related to any information-theoretic measure.

In stochastic dynamical systems with nonlinear interactions, perturbing x may not only affect the evolution of the expectation value of y , but it may also affect the evolution of the variance of y , and in fact its entire probability distribution. The KL divergence from the natural to the perturbed probability densities has recently been identified as the universal upper bound to the physical response of any observable relative to its natural fluctuations [23].

In this letter, we define a new measure of causation in the form of a linear response coefficient between KL divergences, which we would like to call *information response*. In particular, we consider the ratio of two KL divergences, one for the response and one for the perturbation, where the latter represents an information-theoretic cost of the perturbation. For small perturbations, we formulate a fluctuation-response theorem that expresses this ratio as a ratio of Fisher information.

^(a)E-mail: andrea.auconi@gmail.com (corresponding author)

In linear systems, this new information response reduces to the transfer entropy, which provides a connection between Fisher and mutual information, and thus a connection between fluctuation-response theory and information flows.

Kullback-Leibler (KL) divergence. – Consider two probability distributions $p(w)$ and $q(w)$ of a random variable w . The KL divergence from $q(w)$ to $p(w)$ is defined as

$$D[p(w)||q(w)] \equiv \int dw p(w) \ln \left(\frac{p(w)}{q(w)} \right); \quad (1)$$

it is not symmetric in its arguments, and non-negative. Importantly, it is *invariant* under invertible transformations $w \rightarrow w'$ [18], namely $D[p(w)||q(w)] = D[p(w')||q(w')]$.

The problem of causation. – Consider a stochastic system of n variables evolving with ergodic Markovian dynamics. Our goal is to *define* a quantitative measure of causation, *i.e.*, the influence that a variable x exerts over the dynamics of another variable y . We want this definition to have both the invariance property of KL divergences, and the physical interpretation of a propagation of perturbations.

Since the dynamics is ergodic, and therefore stationary, it suffices to consider the stochastic variables $x_0 \equiv x(t=0)$, $y_0 \equiv y(t=0)$ at $t=0$, and a time interval τ later $y_\tau \equiv y(t=\tau)$. To avoid cluttered notation, we will implicitly assume that the current values of the remaining $n-2$ variables are absorbed into y_0 , *e.g.*, $p(y_\tau | y_0) \equiv p(y_\tau | y_0, z_0)$. Conditioning on z_0 avoids confounding variables in z to introduce spurious causal links between x and y [24].

Local response divergence. – Let us consider the system at $t=0$ with steady-state distribution $p(x_0, y_0)$. We make an ideal measurement of its actual state (x_0, y_0) . Immediately after the measurement, we perturb the state by introducing a small displacement $\epsilon > 0$ of the variable x , namely $x_0 \Rightarrow x_0 + \epsilon$. If the effect of this perturbation propagates to y , then it is reflected in the KL divergence from the natural to the perturbed prediction

$$d_\tau^{x \rightarrow y}(x_0, y_0, \epsilon) \equiv D[p(y_\tau | x_0, y_0; x_0 \Rightarrow x_0 + \epsilon) || p(y_\tau | x_0, y_0)], \quad (2)$$

which is a function of the local condition (x_0, y_0) and the perturbation strength ϵ . This quantity was itself suggested as a causality quantifier in the intervention-effect framework of [25,26]. We name it local response divergence, and denote its ensemble average by $\langle d_\tau^{x \rightarrow y}(x_0, y_0, \epsilon) \rangle \equiv \int dx_0 dy_0 p(x_0, y_0) d_\tau^{x \rightarrow y}(x_0, y_0, \epsilon)$.

The concept of causation, interpreted in the framework of fluctuation-response theory, is only meaningful with respect to an arrow of time [27]. That means to postulate

that the perturbation cannot have effects at past times

$$p(y_\tau | x_0, y_0; x_0 \Rightarrow x_0 + \epsilon) \equiv \begin{cases} p(y_\tau | x_0 + \epsilon, y_0), & \text{for } \tau \geq 0, \\ p(y_\tau | x_0, y_0), & \text{for } \tau < 0. \end{cases} \quad (3)$$

In writing the conditional probability $p(y_\tau | x_0 + \epsilon, y_0)$, we implicitly assumed $p(x_0 + \epsilon, y_0) > 0$, meaning that the condition provoked by the perturbation is possible under the natural statistics. This implies that the response statistics can be predicted without actually perturbing the system, which is the main idea of fluctuation-response theory [19–22].

Information-theoretic cost. – The mean local response divergence $\langle d_\tau^{x \rightarrow y}(x_0, y_0, \epsilon) \rangle$, like any response function in fluctuation-response theory, is defined in relation to a perturbation, irrespective of how difficult it may be to perform this perturbation. Intuitively, we expect that it takes more effort to perturb those variables that fluctuate less. Therefore, we consider the KL divergence from the natural to the perturbed ensemble of conditions

$$c_x(\epsilon) \equiv D[p(x_0 - \epsilon, y_0) || p(x_0, y_0)], \quad (4)$$

to quantify the information-theoretic cost of perturbations, and call it *perturbation divergence*. Note that we defined the perturbation through the unperturbed density as $p(x_0 - \epsilon, y_0) \equiv p(x(t=0) + \epsilon = x_0, y(t=0) = y_0)$.

For example, for an underdamped Brownian particle, the perturbation divergence is equivalent to the average thermodynamic work required to perform an ϵ perturbation of its velocity, up to a factor being the temperature, see Supplementary Material [SupplementaryMaterial.pdf](#) (SM). For an equilibrium ensemble in a potential $U(x)$, with Boltzmann distribution $p(x) \sim \exp(-\beta U(x))$, the perturbation divergence is the average reversible work $c_x(\epsilon) = \beta \langle U(x + \epsilon) - U(x) \rangle$. Note that the definition of eq. (4) is general, and can be applied to more abstract models where thermodynamic quantities are not clearly identified.

Information response. – We introduce the information response as the ratio between mean local response divergence and perturbation divergence, in the limit of a small perturbation

$$\Gamma_\tau^{x \rightarrow y} \equiv \lim_{\epsilon \rightarrow 0} \frac{\langle d_\tau^{x \rightarrow y}(x_0, y_0, \epsilon) \rangle}{c_x(\epsilon)}. \quad (5)$$

We can interpret $\Gamma_\tau^{x \rightarrow y}$ as an information-theoretic linear response coefficient. This information response is our measure of $x \rightarrow y$ causation with respect to the timescale τ , see fig. 1. The time arrow requirement (eq. (3)) implies $\Gamma_\tau^{x \rightarrow y} = 0$ for $\tau < 0$.

Introducing the *local* information response $\gamma_\tau^{x \rightarrow y}(x_0, y_0) \equiv \lim_{\epsilon \rightarrow 0} d_\tau^{x \rightarrow y}(x_0, y_0, \epsilon) / c_x(\epsilon)$, we can equivalently write $\Gamma_\tau^{x \rightarrow y} = \langle \gamma_\tau^{x \rightarrow y}(x_0, y_0) \rangle$.

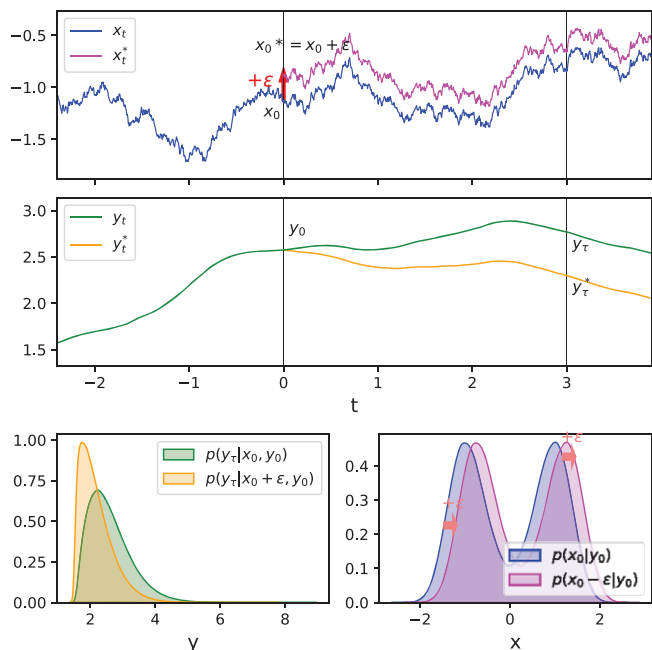


Fig. 1: Here we show, on a concrete example, the origin of the two KL divergences entering the information response of eq. (5). Top: response to the perturbation $x_0 \Rightarrow x_0 + \epsilon$ at the trajectory level. x_t^* (y_t^*) is the perturbed trajectory of x_t (y_t), for the same noise realization. Lower left panel: local response divergence $d_\tau^{x \rightarrow y}(x_0, y_0, \epsilon)$: change of predicted distribution of y_τ for the condition (x_0, y_0) for a timescale $\tau = 3$. Lower right panel: perturbation divergence $c_x(\epsilon)$: instantaneous displacement of the steady-state ensemble conditional to a particular y_0 . The dynamics follows the nonlinear stochastic model of eq. (17) with parameters $t_R = 10$, $q = 0.1$, $\alpha = 0.5$, $\beta = 0.2$, for a perturbation $\epsilon = 0.25$.

The information response in the form of eq. (5) inherently relies on the concept of controlled perturbations. We can reformulate it in purely observational form, in the spirit of the fluctuation-response theorem [19–22], provided $p(x_0, y_0, y_\tau)$ is sufficiently smooth.

Fisher information. – The one-parameter family $\{p(y_\tau | x_0, y_0)\}_{x_0}$ of probability densities parametrized by x_0 (for fixed y_0) can be equipped with a Riemannian metric having $d_\tau^{x \rightarrow y}(x_0, y_0, \epsilon)$ as squared line element. In fact, the leading-order term in the Taylor expansion of a KL divergence between probabilities that differ only by a small perturbation of a parameter is of second order, with coefficients known as Fisher information [18,28]. Explicitly, expanding the mean response divergence for $\tau > 0$, we obtain

$$\begin{aligned} \langle d_\tau^{x \rightarrow y}(x_0, y_0, \epsilon) \rangle = \\ -\frac{1}{2}\epsilon^2 \langle \partial_{x_0}^2 \ln p(y_\tau | x_0, y_0) \rangle + \mathcal{O}(\epsilon^3), \end{aligned} \quad (6)$$

where we used the interventional causality requirement (eq. (3)), and probability normalization. Similarly, for the

perturbation divergence we have

$$c_x(\epsilon) = -\frac{1}{2}\epsilon^2 \langle \partial_{x_0}^2 \ln p(x_0 | y_0) \rangle + \mathcal{O}(\epsilon^3). \quad (7)$$

Applying the Fisher information representation to the information response, for $\tau > 0$, we get

$$\Gamma_\tau^{x \rightarrow y} = \frac{\langle \partial_{x_0}^2 \ln p(y_\tau | x_0, y_0) \rangle}{\langle \partial_{x_0}^2 \ln p(x_0 | y_0) \rangle}, \quad (8)$$

that is the *fluctuation-response theorem* for KL divergences. For generalizations and a discussion of the connection with the classical fluctuation-response theorem see footnote ¹ and the SM. Equation (8) is the ratio of two second derivatives over the same physical variable x_0 , and it can be regarded as an application of L’Hôpital’s rule to eq. (5).

In general, Fisher information is not easily connected to Shannon entropy and mutual information [29]. Below, we show that for linear stochastic systems, the information response, which is a ratio of Fisher information (eq. (8)), is equivalent to the transfer entropy, a conditional form of mutual information.

Transfer entropy. – The most widely used measure of information flow is the conditional mutual information

$$T_\tau^{x \rightarrow y} \equiv \langle D[p(x_0, y_\tau | y_0) || p(x_0 | y_0)p(y_\tau | y_0)] \rangle, \quad (9)$$

which is generally called transfer entropy [13–17]. It is the average KL divergence from conditional independence of x_0 and y_τ given y_0 .

The transfer entropy is used in non-equilibrium thermodynamics of measurement-feedback systems, where it is related to work extraction and dissipation through fluctuation theorems [16,30,31]; in data science, causal network reconstruction from time series is based on statistical significance tests for the presence of transfer entropy [24].

If uncertainty is measured by the Shannon entropy $S[p(x)] = -\int dx p(x) \ln p(x)$, then the transfer entropy quantifies how much, on average, the uncertainty in predicting y_τ from y_0 decreases if we additionally get to know x_0 , $T_\tau^{x \rightarrow y} = \langle S[p(y_\tau | y_0)] - S[p(y_\tau | x_0, y_0)] \rangle$.

While the joint probability $p(x_0, y_0, y_\tau)$ contains all the physics of the interacting dynamics of x and y , the description in terms of the scalar transfer entropy $T_\tau^{x \rightarrow y}$ represents a form of coarse graining.

We introduce the local transfer entropy $t_\tau^{x \rightarrow y}(x_0, y_0) = D[p(y_\tau | x_0, y_0) || p(y_\tau | y_0)]$; thus for the (macroscopic) transfer entropy $T_\tau^{x \rightarrow y} = \langle t_\tau^{x \rightarrow y}(x_0, y_0) \rangle$.

We next show that $T_\tau^{x \rightarrow y}$ and $\Gamma_\tau^{x \rightarrow y}$ are intimately related for linear systems.

¹Equation (8) holds for a larger class of divergences beyond the KL divergence, because the Fisher information is the unique invariant metric [18].

Linear stochastic dynamics. – As example of application, we study the information response in Ornstein-Uhlenbeck (OU) processes [32], *i.e.*, linear stochastic systems of the type

$$\frac{d\xi_t^{(i)}}{dt} + \sum_{j=1}^n A_{ij}\xi_t^{(j)} = \eta_t^{(i)}, \quad (10)$$

where $\langle \eta_t^{(i)} \eta_{t'}^{(j)} \rangle = q_{ij}\delta(t-t')$ is the Gaussian white noise with symmetric and constant covariance matrix. For the system to be stationary, we require the eigenvalues of the interaction matrix A_{ij} to have positive real part. For our setting, we identify $x \equiv \xi^{(i)}$ and $y \equiv \xi^{(j)}$ for some particular (i, j) , and $z \equiv \{\xi^{(k)}\}_{k=1, \dots, n} \setminus \{\xi^{(i)}, \xi^{(j)}\}$ as the remaining variables. Here, probability densities are normal distributions, $p(y_\tau | x_0, y_0) = \mathcal{N}_{y_\tau}(\langle y_\tau | x_0, y_0 \rangle, \sigma_{y_\tau | x_0, y_0}^2)$, with mean $\langle y_\tau | x_0, y_0 \rangle$ and variance $\sigma_{y_\tau | x_0, y_0}^2 \equiv \langle y_\tau^2 | x_0, y_0 \rangle - \langle y_\tau | x_0, y_0 \rangle^2$, and similarly for $p(y_\tau | y_0)$ and $p(x_0 | y_0)$. Expectations depend linearly on the conditions, $\partial_{x_0} \langle y_\tau | x_0, y_0 \rangle = 0$, and variances are independent of them, $\partial_{x_0} \sigma_{y_\tau | x_0, y_0}^2 = 0$. Recall the implicit conditioning on the confounding variables z_0 through y_0 .

Applying these Gaussian properties to eq. (8), the information response becomes:

$$\Gamma_\tau^{x \rightarrow y} = \frac{(\partial_{x_0} \langle y_\tau | x_0, y_0 \rangle)^2 \sigma_{x_0 | y_0}^2}{\sigma_{y_\tau | x_0, y_0}^2}, \quad (11)$$

where $\partial_{x_0} \langle y_\tau | x_0, y_0 \rangle$ can be interpreted as the coefficient of x_0 in the linear regression for y_τ based on the predictors (x_0, y_0) , and $\sigma_{y_\tau | x_0, y_0}^2$ as its error variance. The variance $\sigma_{x_0 | y_0}^2$ quantifies the strength of the natural fluctuations of x_0 (variable to be perturbed) conditional on y_0 (other variables). In fact, the information-theoretic cost of the perturbation, $c_x(\epsilon) = \epsilon^2 \sigma_{x_0 | y_0}^{-2} + \mathcal{O}(\epsilon^3)$, is higher if x_0 and y_0 are more correlated.

In linear systems, the transfer entropy is equivalent to Granger causality [33]

$$T_\tau^{x \rightarrow y} = \ln \left(\frac{\sigma_{y_\tau | y_0}}{\sigma_{y_\tau | x_0, y_0}} \right), \quad (12)$$

as can be seen by substituting the Gaussian expressions for $p(y_\tau | x_0, y_0)$ and $p(y_\tau | y_0)$ into eq. (9).

The decrease in uncertainty in adding the predictor x_0 to the linear regression of y_τ based on y_0 reads

$$\sigma_{y_\tau | y_0}^2 - \sigma_{y_\tau | x_0, y_0}^2 = \sigma_{x_0 | y_0}^2 (\partial_{x_0} \langle y_\tau | x_0, y_0 \rangle)^2, \quad (13)$$

see the SM. Comparing eq. (11) with eq. (12) and using eq. (13), we obtain a non-trivial equivalence between information response and transfer entropy for OU processes,

$$\Gamma_\tau^{x \rightarrow y} = e^{2T_\tau^{x \rightarrow y}} - 1. \quad (14)$$

Remarkably, despite the equivalence of the macroscopic quantities $\Gamma_\tau^{x \rightarrow y}$ and $T_\tau^{x \rightarrow y}$, the corresponding local quantities are markedly different, see fig. 2. Interestingly, the

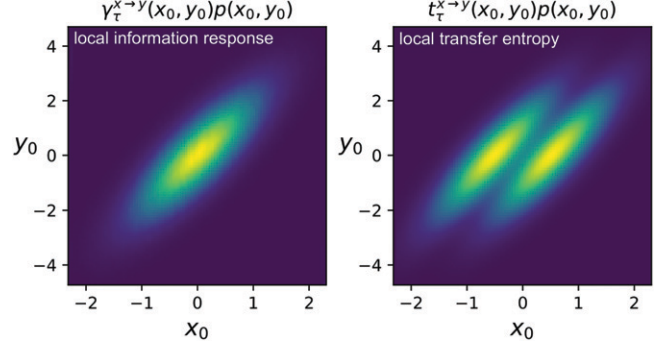


Fig. 2: Local information response (left) and local transfer entropy (right) are different, although their expectation values agree in linear systems. The model is the OU process of eq. (15) with parameters $t_R = 10$, $q = 0.1$, $\alpha = 0.5$, $\beta = 0.2$, observed with timescale $\tau = 3$.

same functional dependence on the transfer entropy given by eq. (14) is attained for linear systems also by the symmetrized transfer entropy [25,26], which is defined as in eq. (9) but using the symmetrized KL divergence [18].

In fig. 2, we show the local response divergence $\gamma_\tau^{x \rightarrow y}(x_0, y_0)$ and local transfer entropy $t_\tau^{x \rightarrow y}(x_0, y_0)$ for the hierarchical OU process of two variables

$$\begin{cases} \frac{dx}{dt} = -\frac{x}{t_R} + \eta_t, \\ \frac{dy}{dt} = \alpha x - \beta y, \end{cases} \quad (15)$$

with $\langle \eta_t \eta_{t'} \rangle = q\delta(t-t')$, and parameters $\alpha, \beta > 0$, $t_R > 0$, $q > 0$. This is possibly the simplest model of non-equilibrium stationary interacting dynamics with continuous variables [34]. However, the pattern of fig. 2 is qualitatively the same for any linear OU process. In fact, the perturbation $x_0 \Rightarrow x_0 + \epsilon$ shifts the prediction $p(y_\tau | x_0, y_0)$ by the same amount on the y -axis, $\epsilon \partial_{x_0} \langle y_\tau | x_0, y_0 \rangle$, independently of the condition (x_0, y_0) , without affecting the variance $\sigma_{y_\tau | x_0, y_0}^2$. Hence, $d_\tau^{x \rightarrow y}(x_0, y_0, \epsilon)$ is constant in space, and the local contribution only reflects the density $p(x_0, y_0)$, here a bivariate Gaussian. On the contrary, the KL divergence corresponding to the change of the prediction $p(y_\tau | y_0)$ into $p(y_\tau | x_0, y_0)$ given by the knowledge of x_0 , is strongly dependent on (x_0, y_0) . In fact, the local transfer entropy reads

$$t_\tau^{x \rightarrow y}(x_0, y_0) = T_\tau^{x \rightarrow y} + \frac{(\partial_{x_0} \langle y_\tau | x_0, y_0 \rangle)^2}{2\sigma_{y_\tau | y_0}^2} [(x_0 - \langle x_0 | y_0 \rangle)^2 - \sigma_{x_0 | y_0}^2], \quad (16)$$

see the SM. In particular, for likely values $x_0 \approx \langle x_0 | y_0 \rangle$, the divergence $t_\tau^{x \rightarrow y}(x_0, y_0)$ is smaller compared to the unlikely situations $x_0 \gg \langle x_0 | y_0 \rangle$ and $x_0 \ll \langle x_0 | y_0 \rangle$. Thus, when multiplied by the steady-state density $p(x_0, y_0)$, $t_\tau^{x \rightarrow y}(x_0, y_0)$ attains a bimodal shape.

Nonlinear example. – As a counter-example for the general validity of eq. (14) for nonlinear systems, consider the following nonlinear Langevin equation for two variables:

$$\begin{cases} \frac{dx}{dt} = -\frac{x}{t_R} + \eta_t, \\ \frac{dy}{dt} = \alpha x^2 - \beta y. \end{cases} \quad (17)$$

Numerical simulations (same parameters as for eq. (15)) show that eq. (14) is violated, see the SM for details. Hence, in general, the transfer entropy is not easily connected to the information response.

Ensemble information response. – Similar to the above, we can define an analogous information response at the ensemble level. From the same perturbation $x_0 \Rightarrow x_0 + \epsilon$, we consider the unconditional response divergence

$$d_{\tau}^{\widetilde{x \rightarrow y}}(\epsilon) \equiv D[p(y_{\tau} | x_0 \Rightarrow x_0 + \epsilon) | p(y_{\tau})], \quad (18)$$

i.e., we evaluate the response at the ensemble level, without knowledge of the measurement (x_0, y_0) ,

$$p(y_{\tau} | x_0 \Rightarrow x_0 + \epsilon) = \langle p(y_{\tau} | x_0, y_0; x_0 \Rightarrow x_0 + \epsilon) \rangle. \quad (19)$$

In general $d_{\tau}^{\widetilde{x \rightarrow y}}(\epsilon) \neq \langle d_{\tau}^{x \rightarrow y}(x_0, y_0, \epsilon) \rangle$.

We define the ensemble information response as

$$\begin{aligned} \Gamma_{\tau}^{\widetilde{x \rightarrow y}} &\equiv \lim_{\epsilon \rightarrow 0} \frac{d_{\tau}^{\widetilde{x \rightarrow y}}(\epsilon)}{c_x(\epsilon)} \\ &= -\frac{\langle \langle \partial_{x_0} \ln p(y_{\tau} | x_0, y_0) | y_{\tau} \rangle^2 \rangle}{\langle \partial_{x_0}^2 \ln p(x_0 | y_0) \rangle}, \end{aligned} \quad (20)$$

where the second line, valid only for $\tau > 0$, is the corresponding fluctuation-response theorem. A straightforward generalization to arbitrary perturbation profiles $\epsilon(x_0, y_0)$ is discussed in the SM. Note that we could write $d_{\tau}^{\widetilde{x \rightarrow y}}(\epsilon)$ through the Fisher information $\langle \partial_{\epsilon}^2 \ln p(y_{\tau} | x_0 + \epsilon, y_0) \rangle|_{\epsilon=0}$, but the partial derivative would be over the perturbation parameter ϵ , and we found it more natural to consider the self-prediction quantity $\langle \langle \partial_{x_0} \ln p(y_{\tau} | x_0, y_0) | y_{\tau} \rangle^2 \rangle$. See the SM for technical details on expectation brackets.

In linear systems, the ensemble information response takes the form

$$\Gamma_{\tau}^{\widetilde{x \rightarrow y}} = \Gamma_{\tau}^{x \rightarrow y} e^{-2I_{\tau}^{x,y,y}} = e^{-2I_{\tau}^{y,y}} (1 - e^{-2T_{\tau}^{x \rightarrow y}}), \quad (21)$$

where $I_{\tau}^{y,y} \equiv D[p(y_0, y_{\tau}) | p(y_0)p(y_{\tau})]$ is the mutual information between y_0 and y_{τ} , and $I_{\tau}^{x,y,y} = I_{\tau}^{x,y} + T_{\tau}^{x \rightarrow y}$ is the mutual information that the two predictors (x_0, y_0) together have on the output y_{τ} , see the SM.

From the non-negativity of information, we obtain the bound $0 \leq \Gamma_{\tau}^{\widetilde{x \rightarrow y}} \leq 1$. We see that $\Gamma_{\tau}^{\widetilde{x \rightarrow y}}$ increases with the transfer entropy $T_{\tau}^{x \rightarrow y}$, and decreases with the auto-correlation $I_{\tau}^{y,y}$. Since $I_{\tau}^{y,y}$ diverges for $\tau \rightarrow 0$ in continuous processes, the perturbation on the x ensemble takes

a finite time to fully propagate its effect to the y ensemble. Since time-lagged information vanishes for $\tau \rightarrow \infty$ in ergodic processes, ensembles relax asymptotically towards the steady state after a perturbation, and correspondingly the ensemble information response vanishes. This provides a trade-off shape for $\Gamma_{\tau}^{\widetilde{x \rightarrow y}}$ as a function of the timescale τ . Note the asymptotics $\Gamma_{\tau}^{\widetilde{x \rightarrow y}} / \Gamma_{\tau}^{x \rightarrow y} \rightarrow 1$ for $\tau \rightarrow \infty$, also resulting from ergodicity.

Discussion. – In this letter, we introduced a new measure of causation that has both the invariance properties required for an information-theoretic measure and the physical interpretation of a propagation of perturbations. It has the form of a linear response coefficient between Kullback-Leibler divergences, and it is based on the information-theoretic cost of perturbations. We would like to call it *information response*.

We study the behavior of the information response analytically in linear stochastic systems, and show that it reduces to the known transfer entropy in this case. This establishes a first connection between fluctuation-response theory and information flow, *i.e.*, the two main perspectives to the problem of causation at present. Additionally, it provides a new relation between Fisher and mutual information.

We suggest our information response for the design of new quantitative causal inference methods [24]. Its practical estimation on time series, as it is normally the case for information-theoretic measures, depends on the learnability of probability distributions from a finite amount of data [35,36].

We thank M. SCAZZOCCHIO for helpful discussions. AA is supported by the DFG through FR3429/3 to BMF; AA and BMF are supported through the Excellence Initiative by the German Federal and State Governments (Cluster of Excellence PoL EXC-2068).

REFERENCES

- [1] SETH ANIL K., BARRETT ADAM B. and BARNETT LIONEL, *J. Neurosci.*, **35** (2015) 3293.
- [2] RUNGE JAKOB, BATHIANY SEBASTIAN, BOLLT ERIK, CAMPS-VALLS GUSTAU, COUMOU DIM, DEYLE ETHAN, GLYMOUR CLARK, KRETSCHMER MARLENE, MAHECHA MIGUEL D., MUÑOZ-MARÍ JORDI *et al.*, *Nat. Commun.*, **10** (2019) 1.
- [3] LUZZATTO LUCIO and PANDOLFI PIER PAOLO, *New Engl. J. Med.*, **373** (2015) 84.
- [4] KWON OKYU and YANG J.-S., *EPL*, **82** (2008) 68003.
- [5] ITO SOSUKE and SAGAWA TAKAHIRO, *Phys. Rev. Lett.*, **111** (2013) 180603.
- [6] HOROWITZ JORDAN M. and ESPOSITO MASSIMILIANO, *Phys. Rev. X*, **4** (2014) 031015.
- [7] JAMES RYAN G., BARNETT NIX and CRUTCHFIELD JAMES P., *Phys. Rev. Lett.*, **116** (2016) 238701.

- [8] AUCONI ANDREA, GIANSAANTI ANDREA and KLIPP EDDA, *Phys. Rev. E*, **95** (2017) 042315.
- [9] PEARL JUDEA, *Causality* (Cambridge University Press) 2009.
- [10] JANZING DOMINIK, BALDUZZI DAVID, GROSSE-WENTRUP MORITZ, SCHÖLKOPF BERNHARD *et al.*, *Ann. Stat.*, **41** (2013) 2324.
- [11] AURELL ERIK and DEL FERRARO GINO, *J. Phys.: Conf. Ser.*, **699** (2016) 012002.
- [12] BALDOVIN MARCO, CECCONI FABIO and VULPIANI ANGELO, *Phys. Rev. Res.*, **2** (2020) 043436.
- [13] MASSEY JAMES, *Causality, feedback and directed information*, in *Proceedings of the International Symposium on Information Theory & Its Applications (ISITA-90)* (Cite-seer) 1990, pp. 303–305.
- [14] SCHREIBER THOMAS, *Phys. Rev. Lett.*, **85** (2000) 461.
- [15] AY NIHAT and POLANI DANIEL, *Adv. Complex Syst.*, **11** (2008) 17.
- [16] PARRONDO JUAN M. R., HOROWITZ JORDAN M. and SAGAWA TAKAHIRO, *Nat. Phys.*, **11** (2015) 131.
- [17] COVER THOMAS M., *Elements of Information Theory* (John Wiley & Sons) 1999.
- [18] AMARI S. I., *Information Geometry and Its Applications*, Vol. **194** (Springer) 2016.
- [19] KUBO REP, *Rep. Prog. Phys.*, **29** (1966) 255.
- [20] KUBO RYOGO, *Science*, **233** (1986) 330.
- [21] MARINI BETTOLO MARCONI UMBERTO, PUGLISI ANDREA, RONDONI LAMBERTO and VULPIANI ANGELO, *Phys. Rep.*, **461** (2008) 111.
- [22] MAES CHRISTIAN, *Front. Phys.*, **8** (2020) 229.
- [23] DECHANT ANDREAS and SASA SHIN-ICHI, *Proc. Natl. Acad. Sci. U.S.A.*, **117** (2020) 6430.
- [24] RUNGE JAKOB, *Chaos*, **28** (2018) 075310.
- [25] SMIRNOV DMITRY A., *Phys. Rev. E*, **90** (2014) 062921.
- [26] SMIRNOV DMITRY A., *Phys. Rev. E*, **102** (2020) 062139.
- [27] ROLDÁN ÉDGAR, NERI IZAAK, DÖRPINGHAUS MEIK, MEYR HEINRICH and JÜLICHER FRANK, *Phys. Rev. Lett.*, **115** (2015) 250602.
- [28] ITO SOSUKE and DECHANT ANDREAS, *Phys. Rev. X*, **10** (2020) 021056.
- [29] WEI XUE-XIN and STOCKER ALAN A., *Neural Comput.*, **28** (2016) 305.
- [30] SAGAWA TAKAHIRO and UEDA MASAHIRO, *Phys. Rev. E*, **85** (2012) 021104.
- [31] ROSINBERG MARTIN LUC and HOROWITZ JORDAN M., *EPL*, **116** (2016) 10007.
- [32] RISKEN HANNES, *Fokker-Planck equation*, in *The Fokker-Planck Equation* (Springer) 1996, pp. 63–95.
- [33] BARNETT LIONEL, BARRETT ADAM B. and SETH ANIL K., *Phys. Rev. Lett.*, **103** (2009) 238701.
- [34] AUCONI ANDREA, GIANSAANTI ANDREA and KLIPP EDDA, *Entropy*, **21** (2019) 177.
- [35] BIALEK WILLIAM, CALLAN CURTIS G. and STRONG STEVEN P., *Phys. Rev. Lett.*, **77** (1996) 4693.
- [36] BIALEK WILLIAM, PALMER STEPHANIE E. and SCHWAB DAVID J., *What makes it possible to learn probability distributions in the natural world?*, arXiv preprint, arXiv:2008.12279 (2020).