This paper was presented as part of the Workshop on Network Science for Communication Networks (NetSciCom)

A Generalized Prediction Framework for Granger Causality

Christopher J. Quinn
Department of Electrical and Computer Engineering
University of Illinois
Urbana, Illinois 61801
Email: quinn7@illinois.edu

Todd P. Coleman
Department of Electrical and Computer Engineering
University of Illinois
Urbana, Illinois 61801
Email: colemant@illinois.edu

Negar Kiyavash
Department of Industrial and Enterprise Systems Engineering
University of Illinois
Urbana, Illinois 61801
Email: kiyavash@illinois.edu

***Abstract*—In his 1969 paper, Granger proposed a statistical definition of causality between stochastic processes. It is based on whether causal side information helps in a sequential prediction task. However, his formulation was limited to linear predictors. We describe a generalized framework, where predictions are beliefs and compare the best predictor with side information to the best predictor without side information. The difference in the prediction performance, i.e., regret of such predictors, is used as a measure of causal influence of the side information. Specifically when log loss is used to quantify each predictor's loss and an expectation over the outcomes is used to quantify the regret, we show that the directed information, an information theoretic quantity, quantifies Granger causality. We also explore a more pessimistic setup perhaps better suited for adversarial settings where minimax criterion is used to quantify the regret.**

I. Introduction

In his 1969 paper [1], Granger proposed a framework for identifying statistically causal relationships between stochastic processes, based on sequential prediction. It has been widely adopted in a number of research fields, including economics, biology, and social sciences [2], [3]. His framework is [1]:

> "We say that X_t is causing Y_t if we are better able to predict Y_t using all available information than if the information apart from X_t had been used."

This was motivated by earlier work by Wiener [1]. Granger formulated this framework using linear regression models of stochastic processes [1]. While this version of the framework has been widely adopted in econometrics and other disciplines [3], there have been attempts to extend it to nonlinear processes. The directed transfer function, for example, extends the framework into the spectral domain [3]. However, all known formulations of Granger's principle are designed for specific classes of processes.

Granger's principle is based on how much causal side information helps in a sequential prediction task. There is a large body of research on sequential prediction [4], [5]. Some researchers have focused on predicting stochastic processes, often with modeling assumptions, but there has been increasing focus on sequential prediction of general sequences, a problem known as "on-line" prediction. In this setting, the outcome sequence could be generated stochastically, deterministically, or even generated sequentially by an adversary [4]. Much of the work in this field has focused on comparing performance of a predictor to the best "expert" from a group of experts. Before the predictor makes a decision, he learns what each of the experts predict.

Although there has been significant advances in the field of sequential prediction, there has been little work characterizing how much side information (knowledge of the X_t process) helps. The works that examine problems with side information, such as [4], [6], [7], compare a predictor with side information to a group of experts with the same side information. Also, most works assume both the predictor and experts will use the side information in the same manner [4].

In this paper, inspired by Granger's philosophy, we develop a generalized framework for measuring causal influences characterized by how much side information helps in sequential prediction. We focus on the setting where experts assign probabilities to outcomes. (This setting reveals the experts' certainties on all outcomes, not just one.) The goodness of prediction is measured with log loss. The comparison of the performance, i.e., regret of the best predictor with side information to the best predictor without side information is used as our causality metric. When the comparison of regret is done taking an expectation over all possible outcomes, we show that directed information, an information theoretic quantity, captures Granger viewpoint on definition of causality. We first consider a two process problem and then generalize it. Moreover, we explore Grangers principle in the minimax setting, where the performance of the predictor with side information is compared to the predictor without it for the worst-case outcome.

II. Notation

A. Sequential Prediction

We first introduce notation for sequential prediction.

- There are two, competing decision makers (or predictors) f and q from classes of predictors $\mathcal{F}$ and $\mathcal{Q}$, respectively, who sequentially predict an outcome sequence $y_1, y_2, \ldots$ composed of elements from an outcome space $\mathcal{Y}$. For simplicity we consider discrete $\mathcal{Y}$.
- At time i, the decision makers make predictions f_i and q_i respectively in a decision space $\mathcal{D}$ for the next outcome

978-1-4244-9920-5/11/$26.00 ©2011 IEEE

y_i. To make this decision, both have access to the past outcomes $y^{i-1} = (y_1, \ldots, y_{i-1})$ and q_i additionally has access to causal side information $x^i = (x_1, \ldots, x_i)$. (The causal property is that at time i, the side information for the future is not revealed.)

- The "goodness" of the predictions is measured by a nonnegative loss function $l : \mathcal{D} \times \mathcal{Y} \to \mathbb{R}_+$. The decision makers incur losses $l(f_i, y_i)$ and $l(q_i, y_i)$ respectively.
- Denote the cumulative losses for f and q as

$$L_n(f, y^n) \triangleq \frac{1}{n}\sum_{i=1}^{n} l(f_i, y_i) \qquad \text{and}$$
$$L_n(q, y^n, x^n) \triangleq \frac{1}{n}\sum_{i=1}^{n} l(q_i, y_i).$$

- We are interested in characterizing the regret:

$$R_n(f, q, y^n, x^n) \triangleq L_n(f, y^n) - L_n(q, y^n, x^n) \qquad (1)$$

between the "best" decision makers $f \in \mathcal{F}$ and $q \in \mathcal{Q}$. f_i is a function of y^{i-1} and q_i is a function of x^{i-1} and y^{i-1}. With appropriate $\mathcal{F}$ and $\mathcal{Q}$, the regret quantifies how much the side information helps on average over time.

B. Information Theory

Now we introduce some information theoretic notation.

- Let X^n, Y^n, and Z^n be three, discrete stochastic processes with joint distribution $P_{X^n,Y^n,Z^n}(x^n, y^n, z^n)$.
- We denote the space of all possible distributions on these processes as $\mathcal{P}(\mathcal{X}^n \times \mathcal{Y}^n \times \mathcal{Z}^n)$.
- The entropy of Y^n is [8]:

$$H(Y^n) \triangleq \mathrm{E}_{P_{Y^n}}\left[-\log P_{Y^n}(Y^n)\right].$$

- The causally conditional entropy of Y^n causally conditioned on X^n is (Ch 3 in [9]) :

$$\begin{aligned} H(Y^n\|X^n) &\triangleq \mathrm{E}_{P_{Y^n,X^n}}\left[-\log P_{Y^n\|X^n}(Y^n\|X^n)\right] \\ &= \sum_{i=1}^{n} \mathrm{E}_{P_{Y^i,X^i}}\left[-\log P_{Y_i|Y^{i-1},X^i}(Y_i|Y^{i-1}, X^i)\right] \end{aligned}$$

- For two distributions P_{Y^n} and Q_{Y^n} on $\mathcal{Y}^n$, the Kullback-Leibler divergence between P_{Y^n} and Q_{Y^n} is [8]

$$D(P_{Y^n}\|Q_{Y^n}) \triangleq \mathrm{E}_{P_{Y^n}}\left[\log \frac{P_{Y^n}(Y^n)}{Q_{Y^n}(Y^n)}\right].$$

$D(P_{Y^n}||Q_{Y^n}) \geq 0$ with equality iff $P_{Y^n} \equiv Q_{Y^n}$.

- For two distributions P_{Y^n,X^n} and Q_{Y^n,X^n} on $\mathcal{Y}^n \times \mathcal{X}^n$, the conditional Kullback-Leibler divergence between $P_{Y^n|X^n}$ and $Q_{Y^n|X^n}$ is [8]:

$$\begin{aligned} &D\big(P_{Y^n|X^n}\|Q_{Y^n|X^n}\big|P_{X^n}\big) \\ &\quad \triangleq \mathrm{E}_{P_{X^n}}\mathrm{E}_{P_{Y^n|X^n}}\left[\log \frac{P_{Y^n|X^n}(Y^n|X^n)}{Q_{Y^n|X^n}(Y^n|X^n)}\Big|X^n\right]. \end{aligned}$$

- The directed information from X^n to Y^n is

$$\begin{aligned} &\mathrm{I}(X^n \to Y^n) \\ &\quad \triangleq \sum_{i=1}^{n} \mathrm{E}_{P_{X^n,Y^n}}\left[\log \frac{P_{Y_i|Y^{i-1},X^i}(Y_i|Y^{i-1}, X^i)}{P_{Y_i|Y^{i-1}}(Y_i|Y^{i-1})}\right] \\ &\quad = H(Y^n) - H(Y^n\|X^n). \end{aligned}$$

This was formally introduced by Massey [10]. Massey's definition was motivated by Marko's work [11]. Related work was independently done by Rissanen [12].

- The causally conditioned directed information from X^n to Y^n causally conditioned on Z^n is [9]

$$\begin{aligned} &\mathrm{I}(X^n \to Y^n\|Z^n) \\ &\quad \triangleq \sum_{i=1}^{n} \mathrm{E}_{P_{X^n,Y^n,Z^n}}\left[\log \frac{P_{Y_i|Y^{i-1},X^i,Z^i}(Y_i|Y^{i-1},X^i,Z^i)}{P_{Y_i|Y^{i-1},Z^i}(Y_i|Y^{i-1},Z^i)}\right] \\ &\quad = H(Y^n\|Z^n) - H(Y^n\|X^n, Z^n). \end{aligned}$$

III. Setup

The overall goal is to characterize how much side information helps in sequential prediction. For this, we will develop a general framework in which the best possible predictor $f \in \mathcal{F}$ without side information competes with the best possible predictor $q \in \mathcal{Q}$ with causal side information. For this we consider decision spaces and loss functions which are meaningful and not restrictive.

First consider the decision space. A general type of prediction problem involves the decision makers sequentially predicting probabilities, or "beliefs," of the next symbol [4], [5]. Predicting belief functions, for which the decision maker must assign a confidence to each possible outcome, is much more informative than just seeing the single outcome the decision maker thought was most likely. At time i, the decision makers will each predict a probability vector, assigning a probability to each of the possible outcomes

$$f_i = \{\ f_i(y)\ \}_{y\in\mathcal{Y}} \quad \text{and} \quad q_i = \{\ q_i(y)\ \}_{y\in\mathcal{Y}}.$$

The assignments are nonnegative and normalized. The decision space is the set of all probability measures on $\mathcal{Y}$: $\mathcal{D} = \mathcal{P}(\mathcal{Y})$. The decision makers will choose their decisions f_i and q_i using different information. f_i will be a distribution of y_i conditioned on the past outcomes y^{i-1} and q_i a distribution also conditioned on side information x^i.

Any set of sequential predictions on the outcome sequence $\{f_i(y_i|y^{i-1})\}_{i=1}^n$ has a corresponding joint $f(y^n) = \prod_{i=1}^n f_i(y^i|y^{i-1})$. Likewise, any joint distribution on the outcome sequence $f(y^n)$, through marginalizations, can be used to form sequential predictions:

$$f_i(y_i|y^{i-1}) = \frac{f(y^i)}{f(y^{i-1})}.$$

Thus, the class of sequential predictors $\mathcal{F}$ could be any subset of probability distributions on the whole outcome sequence $\mathcal{F} \subseteq \mathcal{P}(\mathcal{Y}^n)$.

However, this is not quite the case with the class $\mathcal{Q}$ which has access to side information. Any set of sequential

predictions on the outcome sequence with causal knowledge of the side information $\{q_i(y_i|y^{i-1},x^i)\}_{i=1}^n$ can be combined to form a *causally conditioned* distribution (Ch. 3 in [9]) on the y^n sequence:

$$q(y^n\|x^n)=\prod_{i=1}^n q_i(y^i|y^{i-1},x^i). \tag{2}$$

Any causally conditioned distribution $q(y^n\|x^n)$ can be equivalently deconstructed to form sequential predictions

$$q_i(y_i|y^{i-1},x^i)=\frac{q(y^i\|x^i)}{q(y^{i-1}\|x^{i-1})}.$$

Thus, the class of sequential predictors $\mathcal{Q}$ could be any subset of causally conditioned distributions $\mathcal{Q}\subseteq\mathcal{P}(\mathcal{Y}^n\|\mathcal{X}^n)$. However, note that the outcome sequence need not have a causal dependence on the side information, so not all distributions in $\mathcal{P}(\mathcal{Y}^n\times\mathcal{X}^n)$ have corresponding causally conditioned marginal distributions in $\mathcal{P}(\mathcal{Y}^n\|\mathcal{X}^n)$. This limitation is due to the q_i's only having causal access to the side information.

For prediction problems where the predictions are probability assignments, a widely used loss function is the "log-loss," also called "self-information loss" [4], which for probability assignment $p=\{p(y):y\in\mathcal{Y}\}$ and outcome $y\in\mathcal{Y}$,

$$l(p,y)=-\log p(y).$$

This has meaningful interpretations in areas such as data compression, gambling, and portfolio theory [4], [5]. In sequential data compression, if a stochastic sequence Z^n is sequentially generated from a distribution $P_Z(z)$, then the "ideal" codelength of a symbol z is $-\log P_Z(z)$ [8]. This code, known as the Shannon code, achieves the minimum expected total codelength for any uniquely decodable code [8]. Log loss is commonly used to characterize the growth rate in wealth in sequential gambling and in portfolio theory [4], [8]. Log loss also has the property that it can break up products of terms (such as products of conditional probabilities) into a summation of those terms. We will now investigate the expected regret between the best decision makers (in expectation), where the predictions are probability measures.

In characterizing how much causal side information helps in sequential prediction, we consider the regret between the best predictor with side information to the best predictor without. "Best" can be defined in a number of ways. The best $f\in\mathcal{F}$ could be specified as the one that minimizes its loss $\arg\min_{f\in\mathcal{F}} L_n(f,y^n)$. In this case, there is a different "best" f for each outcome sequence y^n. Alternatively, if the outcome sequence is stochastic, best could be the $f\in\mathcal{F}$ which minimizes its expected loss $\arg\min_{f\in\mathcal{F}}\mathrm{E}_{P_{Y^n}}L_n(f,Y^n)$. In this case, there is a single f. Likewise, best could be the f whose *worst-case* loss is minimal, $\arg\min_{f\in\mathcal{F}}\max_{y^n\in\mathcal{Y}^n} L_n(f,y^n)$.

For these, the choice of f only depends on the class $\mathcal{F}$. Alternatively, it could also depend the other class $\mathcal{Q}$. For instance, best could be the f which has the least worst-case regret when compared to the $q\in\mathcal{Q}$ which minimizes the loss $L_n(q,y^n,x^n)$:

$$\arg\min_{f\in\mathcal{F}}\max_{y^n\in\mathcal{Y}^n,x^n\in\mathcal{X}^n}\max_{q\in\mathcal{Q}} L_n(f,y^n)-\inf_{f\in\mathcal{F}} L_n(q,y^n,x^n).$$

There are a variety of settings that could be considered. We will focus on three. The first will be the setting where best for both $\mathcal{F}$ and $\mathcal{Q}$ will be the f and q respectively which minimize the expected loss with respect to their classes. The second is a minimax-type setting. The best $q\in\mathcal{Q}$ will be the one that for any particular outcome and side information sequence, has smallest loss. The best $f\in\mathcal{F}$ will be the one which has least regret compared to the best q for the worst outcome and side information sequences. The third is similar to the second, except instead of worst-case side information, it will be in expectation over side information.

IV. Best Experts in Expectation

A. Two processes

Consider an outcome sequence Y^n and side information sequence X^n which are both stochastic and generated according to the distribution P_{X^n,Y^n}. Our goal is to characterize the expected regret between the best predictor without side information and the best predictor with side information. "Best" is in terms of having the minimal expected cumulative loss. The best $f\in\mathcal{F}$ for some $\mathcal{F}\subseteq\mathcal{P}(\mathcal{Y}^n)$ is

$$f^*=\arg\min_{f\in\mathcal{F}}\mathrm{E}_{P_{Y^n}}\left[L_n(f,Y^n)\right].$$

Likewise, the best $q\in\mathcal{Q}$ for some $\mathcal{Q}\subseteq\mathcal{P}(\mathcal{Y}^n\|\mathcal{X}^n)$is

$$q^*=\arg\min_{q\in Q}\mathrm{E}_{P_{X^n,Y^n}}\left[L_n(q,X^n,Y^n)\right].$$

We now consider the value of the expected cumulative regret. It turns out to be the directed information plus divergences which act as correction terms.

Lemma IV.1. *The expected cumulative regret between the best predictors, f^* and q^* is*

$$\begin{aligned}&\mathrm{E}_{P_{X^n,Y^n}}\left[R_n(f^*,q^*,X^n,Y^n)\right]\\&\quad=\frac{1}{n}\mathrm{I}\left(X^n\to Y^n\right)+\left[\sum_{i=1}^n D\big(P_{Y_i|Y^{i-1}}\|f_i^*|P_{Y_i}\big)\right.\\&\qquad\left.-\sum_{i=1}^n D\big(P_{Y_i|Y^{i-1},X^{i-1}}\|q_i^*\big|P_{Y^{i-1},X^{i-1}}\big)\right]\end{aligned} \tag{3}$$

Proof: By linearity of expectation,

$$\begin{aligned}&\mathrm{E}_{P_{X^n,Y^n}}\left[R_n(f^*,q^*,X^n,Y^n)\right]\\&\quad=\mathrm{E}_{P_{Y^n}}\left[L_n(f^*,Y^n)\right]-\mathrm{E}_{P_{X^n,Y^n}}\left[L_n(q^*,X^n,Y^n)\right]\end{aligned} \tag{4}$$

$$=\sum_{i=1}^n\mathrm{E}_{P_{Y^i}}\left[l(f_i^*,Y_i)\right]-\sum_{i=1}^n\mathrm{E}_{P_{X^i,Y^i}}\left[l(q_i^*,Y_i)\right] \tag{5}$$

where f_i^* is allowed to depend on Y^{i-1} and q_i^* is allowed to depend on both Y^{i-1} and X^i. Consider the sum on the left

in (5). Note that

$$\sum_{i=1}^{n} \mathrm{E}_{P_{Y^i}} \left[l(f^*, Y_i) \right] \tag{6}$$

$$= \sum_{i=1}^{n} \mathrm{E}_{P_{Y^{i-1}}} \mathrm{E}_{P_{Y_i|Y^{i-1}}} \left[-\log f_i^*(Y_i) \middle| Y^{i-1} \right] \tag{7}$$

$$= \sum_{i=1}^{n} \mathrm{E}_{P_{Y^{i-1}}} \mathrm{E}_{P_{Y_i|Y^{i-1}}} \left[\log \frac{P_{Y_i|Y^{i-1}}(Y_i|Y^{i-1})}{f_i^*(Y_i)} \middle| Y^{i-1} \right] + H(Y_i|Y^{i-1}) \tag{8}$$

$$= H(Y^n) + \sum_{i=1}^{n} D\big(P_{Y_i|Y^{i-1}} \| f_i^* \big| P_{Y^{i-1}}\big) \tag{9}$$

Now consider the sum on the right in (5).

$$\sum_{i=1}^{n} \mathrm{E}_{P_{Y^i,X^i}} \left[l(q^*, Y_i) \right]$$

$$= \sum_{i=1}^{n} \mathrm{E}_{P_{Y^{i-1},X^i}} \mathrm{E}_{P_{Y_i|Y^{i-1},X^i}} \left[-\log q_i^*(Y_i) \middle| Y^{i-1}, X^i \right]$$

$$= \sum_{i=1}^{n} \mathrm{E}_{P_{Y^{i-1},X^i}} \mathrm{E}_{P_{Y_i|Y^{i-1},X^i}} \left[\log \frac{P_{Y_i|Y^{i-1},X^i}(Y_i|Y^{i-1},X^i)}{q_i^*(Y_i)} \middle| Y^{i-1}, X^i \right] + H(Y_i|Y^{i-1}, X^i) \tag{10}$$

$$= H(Y^n\|X^n) + \sum_{i=1}^{n} D\big(P_{Y_i|Y^{i-1},X^i} \| q_i^* \big| P_{Y^{i-1},X^i}\big) . \tag{11}$$

Combining (9) and (11) gives (13). ∎

In Lemma IV.1, we considered arbitrary classes $\mathcal{F} \subseteq \mathcal{P}(\mathcal{Y}^n)$ and $\mathcal{Q} \subseteq \mathcal{P}(\mathcal{Y}^n\|\mathcal{X}^n)$. Now we consider the specific case where f can be any joint distribution on the outcome sequence and q any causally conditioned distribution on the outome sequence: $\mathcal{F} = \mathcal{P}(\mathcal{Y}^n)$ and $\mathcal{Q} = \mathcal{P}(\mathcal{Y}^n\|\mathcal{X}^n)$. In this case, we find that the divergence correction terms vanish, and the expected regret is precisely the directed information.

Theorem IV.2. *If $\mathcal{F} = \mathcal{P}(\mathcal{Y}^n)$ and $\mathcal{Q} = \mathcal{P}(\mathcal{Y}^n \parallel \mathcal{X}^n)$, then $f_i^* \equiv P_{Y_i|Y^{i-1}}$, $q_i^* \equiv P_{Y_i|Y^{i-1},X^i}$, and the expected cumulative regret has value*

$$\mathrm{E}_{P_{X^n,Y^n}} \left[R_n(f^*, q^*, X^n, Y^n) \right] = \frac{1}{n} \mathrm{I}\left(X^n \to Y^n\right). \tag{12}$$

Proof: The expected cumulative loss of f^* is

$$\min_{f \in \mathcal{F}} \mathrm{E}_{P_{Y^n}} \left[L_n(f, Y^n) \right] = \min_{f \in \mathcal{F}} H(Y^n) + \sum_{i=1}^{n} D\big(P_{Y_i|Y^{i-1}} \| f_i \big| P_{Y^{i-1}}\big) .$$

Consider for each i that $f_i \equiv P_{Y_i|Y^{i-1}}$. Then

$$D\big(P_{Y_i|Y^{i-1}} \| f_i \big| P_{Y^{i-1}}\big) = 0.$$

By the nonnegativity of KL divergence, and since $H(Y^n)$ does not depend on f, $f_i^* \equiv P_{Y_i|Y^{i-1}}$ is the minimizer and the expected cumulative loss of f^* is $H(Y^n)$.

The expected cumulative loss of q^* is

$$\min_{q \in \mathcal{Q}} \mathrm{E}_{P_{Y^n,X^n}} \left[L_n(q, X^n, Y^n) \right] = \min_{q \in \mathcal{Q}} H(Y^n\|X^n) + \sum_{i=1}^{n} D\big(P_{Y_i|Y^{i-1},X^i} \| q_i \big| P_{Y^{i-1},X^i}\big) .$$

Consider for each i that $q_i \equiv P_{Y_i|Y^{i-1},X^i}$. Then

$$D\big(P_{Y_i|Y^{i-1},X^i} \| q_i \big| P_{Y^{i-1},X^i}\big) = 0.$$

By the nonnegativity of KL divergence, and since $H(Y^n\|X^n)$ does not depend on q, $q_i^* \equiv P_{Y_i|Y^{i-1},X^i}$ is the minimizer and the expected cumulative loss of q^* is $H(Y^n\|X^n)$. ∎

Since

$$\prod_{i=1}^{n} f_i^*(y_i) = \prod_{i=1}^{n} P_{Y_i|Y^{i-1}}(y_i|y^{i-1}) = P_{Y^n}(y^n),$$

we can write the predictor f^* over the whole outcome sequence as the joint $f^*(y^n) = P_{Y^n}(y^n)$. Similarly, since

$$\prod_{i=1}^{n} q_i^*(y_i) = \prod_{i=1}^{n} P_{Y_i|Y^{i-1},x^i}(y_i|y^{i-1}, x^i) = P_{Y^n\|X^n}(y^n\|x^n),$$

we can write the predictor q^* over the whole outcome with causal access to the side information as the causally conditional distribution $q^*(y^n\|x^n) = P_{Y^n\|X^n}(y^n\|x^n)$.

In Theorem IV.2, since $\mathcal{F}$ and $\mathcal{Q}$ were both as large as possible, this suggests that the directed information characterizes how much causal knowledge of the side information helps, and thus captures Granger's principle when sequentially predicting one sequence with sequential access to another.

B. More than two processes

We now examine a generalization where there are more than two processes. Now, both decision makers have causal access to an additional side information sequence $z^n \in \mathcal{Z}^n$. At time i, the predictions are $f_i(y_i|y^{i-1}, z^i)$ and $q_i(y_i|y^{i-1}, x^i, z^i)$. Let P_{X^n,Y^n,Z^n} denote the joint distribution. This scenario is a fuller generalization of Granger's statement, as "all other knowledge" besides the past Y_i's can be represented in the Z^n process.

"Best" is still in terms of having the minimal expected cumulative loss. The best $f \in \mathcal{F}$ for some $\mathcal{F} \subseteq \mathcal{P}(\mathcal{Y}^n\|\mathcal{Z}^n)$ is

$$f^* = \arg\min_{f \in \mathcal{F}} \mathrm{E}_{P_{Y^n}} \left[L_n(f, Y^n, Z^n) \right].$$

Likewise, the best $q \in \mathcal{Q}$ for some $\mathcal{Q} \subseteq \mathcal{P}(\mathcal{Y}^n\|\mathcal{X}^n, \mathcal{Z}^n)$ is

$$q^* = \arg\min_{q \in Q} \mathrm{E}_{P_{X^n,Y^n,Z^n}} \left[L_n(q, X^n, Y^n, Z^n) \right].$$

We now consider the value of the expected cumulative regret. It turns out to be the causally conditioned directed information plus divergences which act as correction terms.

Lemma IV.3. *The expected cumulative regret between the best predictors, f^* and q^* is*

$$\begin{aligned}\mathrm{E}_{P_{X^n,Y^n,Z^n}}&\left[R_n(f^*,q^*,X^n,Y^n,Z^n)\right]\\ &=\frac{1}{n}\mathrm{I}\left(X^n\to Y^n\|Z^n\right)+\left[\sum_{i=1}^{n}D\big(P_{Y_i|Y^{i-1},Z^i}\|f_i^*|P_{Y_i,Z^i}\big)\right.\\ &\qquad\left.-\sum_{i=1}^{n}D\big(P_{Y_i|Y^{i-1},X^i,Z^i}\|q_i^*\big|P_{Y^{i-1},X^i,Z^i}\big)\right] \qquad (13)\end{aligned}$$

The proof is similar to the proof of Lemma IV.2.

We now consider the specific case that the classes of predictors is as large as possible: $\mathcal{F}=\mathcal{P}(\mathcal{Y}^n\|\mathcal{Z}^n)$ and $\mathcal{Q}=\mathcal{P}(\mathcal{Y}^n\|\mathcal{X}^n,\mathcal{Z}^n)$.

Corollary IV.4. *If $\mathcal{F}=\mathcal{P}(\mathcal{Y}^n\|\mathcal{Z}^n)$ and $\mathcal{Q}=\mathcal{P}(\mathcal{Y}^n\|\mathcal{X}^n,\mathcal{Z}^n)$, then $f_i^*\equiv P_{Y_i|Y^{i-1},Z^i}$, $q_i^*\equiv P_{Y_i|Y^{i-1},X^i,Z^i}$, and the expected cumulative regret has value*

$$\begin{aligned}\mathrm{E}_{P_{X^n,Y^n,Z^n}}&\left[R_n(f^*,q^*,X^n,Y^n,Z^n)\right]\\ &=\frac{1}{n}\mathrm{I}\left(X^n\to Y^n\|Z^n\right) \qquad (14)\end{aligned}$$

This result suggests that the causally conditioned directed information characterizes Granger causality in the setting where average regret is considered.

V. Minimax case

Previously we considered the expected regret between the best predictors, where best meant minimizing the expected loss with respect to the class. We now consider an alternative setting, where best for the class $\mathcal{Q}$ is the one that minimizes loss for a paticular outcome and side information sequence: $\arg\inf_{q\in Q} L_n(q,y^n,x^n)$. The best for the class $\mathcal{F}$ is the one whose worst case regret with respect to the best $q\in\mathcal{Q}$ is least. If the worst case side information is considered, this corresponds to

$$\arg\inf_{f\in\mathcal{F}}\max_{y^n\in\mathcal{Y}^n}\left[L_n(f,y^n)-\inf_{q\in\mathcal{Q}}L_n(q,y^n,x^n)\right]. \qquad (15)$$

The previous setting, where expectations were considered, is meaningful when the outcome and side information sequences are stochastic. Considering the worst case performance is meaningful in adversarial cases. For example, if the outcome sequence is decided by an adversary, who can observe the predictor's decisions at each step, the adversary would decide on the outcome that resulted in the predictor having largest loss.

This will be similar to the traditional minimax sequential prediction problem [4], where a predictor p competes against a class of predictors $\mathcal{F}$:

$$\inf_{p}\max_{y^n\in\mathcal{Y}^n}\left[L_n(p,y^n)-\inf_{f\in\mathcal{F}}L_n(f,y^n)\right].$$

Note that in this setting, though, p and the predictors in class $\mathcal{F}$ have the same information. In (15), however, the predictors in $\mathcal{Q}$ have the additional knowledge of the side information. Thus, in this setting, the regret value could convey some (causal) relationship between the side information sequence and the outcome sequence.

The difference in loss between a predictor $f\in\mathcal{F}$ which does not have access to side information and the "best" predictor (one with smallest loss) $\arg\inf_{q\in Q} L_n(q,y^n,x^n)$ with access to side information is the *regret*:

$$R(f,y^n,x^n)\triangleq L_n(f,y^n)-\inf_{q\in Q}L_n(q,y^n,x^n). \qquad (16)$$

Note that

$$\begin{aligned}-\inf_{q\in Q}L(q,y^n,x^n)&=-\inf_{q\in Q}-\log q(y^n\|x^n)\\ &=\log\sup_{q\in Q}q(y^n\|x^n)=\log q_{ML}(y^n\|x^n)\end{aligned}$$

where $q_{ML}(y^n\|x^n)$ denotes the maximum likelihood of y^n causally conditioned on x^n for the class $\mathcal{Q}$. Thus

$$R(p,y^n,x^n)=-\log p(y^n)+\log q_{ML}(y^n\|x^n). \qquad (17)$$

Note that we will not consider the case that $\mathcal{Q}=\mathcal{P}(\mathcal{Y}^n\|\mathcal{X}^n)$ because for any outcome sequence $y^n\in\mathcal{Y}^n$, there is a distribution that assigns probability one to that sequence and probability zero to all others [5]. In this case, $q_{ML}(y^n)=1$ uniformly.

We now introduce a lemma which will be used in later proofs. It characterizes the "normalized maximum likelihood" predictor as the minimax optimal predictor [4].

Lemma V.1. *The $f\in\mathcal{F}=\mathcal{P}(\mathcal{Y}^n)$ that minimizes*

$$\inf_{f\in\mathcal{F}}\max_{y^n\in\mathcal{Y}^n}\left[-\log f(y^n)+\log g(y^n)\right] \qquad (18)$$

for some function $g:\mathcal{Y}^n\to\mathbb{R}^+$ is

$$f^*(y^n)=\frac{g(y^n)}{\sum_{z^n\in\mathcal{Y}^n}g(z^n)}. \qquad (19)$$

Proof: The proof is to first show f^* achieves uniform regret over all sequences. We refer to $[-\log f(y^n)+\log g(y^n)]$ as the regret. The second step is to show that any other distribution f' does worse than f^* *for some outcome* y^n, and thus the worst case regret of f' is larger than that of f^*. Let $\phi=\sum_{z^n\in\mathcal{Y}^n}g(y^n)$ denote the normalization constant.

That f^* achieves uniform regret over all sequences follows from:

$$\begin{aligned}R(f^*,y^n)&=-\log f^*(y^n)+\log g(y^n) &(20)\\ &=-\log\frac{g(y^n)}{\phi}+\log g(y^n) &(21)\\ &=\log\phi &(22)\end{aligned}$$

Now consider any other predictor f'. Since both f' and f^* are normalized, there is some outcome sequence $z^n\in\mathcal{Y}^n$ for which $f'(z^n)<f^*(z^n)$, which implies that $-\log f'(z^n)>-\log f^*(z^n)$. Thus we have that

$$\begin{aligned}R(f',z^n)&=-\log f'(z^n)+\log g(z^n)\\ &>-\log f^*(z^n)+\log g(z^n)=R(f^*,z^n)\end{aligned}$$

Thus, $\max_{y^n\in\mathcal{Y}^n}R(f',y^n)>\max_{y^n\in\mathcal{Y}^n}R(f^*,y^n)$. ∎

We now consider the best possible worst case performance a predictor $f \in \mathcal{F}$ without side information could do against a family $\mathcal{Q}$ of predictors with side information. We will consider the specific setting where $\mathcal{F} = \mathcal{P}(\mathcal{Y}^n)$. Here the environment is considered adversarial, such that it will give the worst possible outcome sequence and the worst possible side information sequence, where worst means larger regret for the predictors in $\mathcal{F}$.

Lemma V.2. *The $f \in \mathcal{F} = \mathcal{P}(\mathcal{Y}^n)$ that minimizes*

$$\inf_{f\in\mathcal{F}} \max_{y^n\in\mathcal{Y}^n} \max_{x^n\in\mathcal{X}^n} \sup_{q\in Q} R_n(f,q,y^n,x^n) \tag{23}$$

is

$$f^*(y^n) = \frac{\max_{x^n\in\mathcal{X}^n} q_{ML}(y^n \parallel x^n)}{\sum_{z^n\in\mathcal{Y}^n} \max_{x^n\in\mathcal{X}^n} q_{ML}(z^n \parallel x^n)}. \tag{24}$$

The value of (23) *is*

$$\frac{1}{n}\log \sum_{z^n\in\mathcal{Y}^n} \max_{x^n\in\mathcal{X}^n} q_{ML}(z^n \parallel x^n).$$

Proof:

$$\max_{x^n\in\mathcal{X}^n} \sup_{q\in Q} R_n(f,q,y^n,x^n) \tag{25}$$
$$= \max_{x^n\in\mathcal{X}^n} [-\log p(y^n) + \log q_{ML}(y^n \parallel x^n)] \tag{26}$$
$$= \left[-\log p(y^n) + \log \max_{x^n\in\mathcal{X}^n} q_{ML}(y^n \parallel x^n)\right] \tag{27}$$
$$= [-\log p(y^n) + \log g(y^n)] \tag{28}$$

with $g(y^n) = \max_{x^n\in\mathcal{X}^n} q_{ML}(y^n \parallel x^n)$ and (27) uses that $\log$ is one to one increasing so $\max\log$ same as $\log\max$. By Lemma V.1, Lemma V.2 holds. ∎

In the previous setting, we considered the regret with the worst case side information (the worst case for the predictor f without it). That is meaningful when the environment is adversarial and gives both the worst possible outcome and side information sequence. An alternative setting is where the environment gives the worst possible outcome sequence, but the side information is stochastic, and conditionally (but not necessarily causally) dependent on the outcome sequence with a distribution $P_{X^n|Y^n}$. Here, the average regret over possible side information sequences is considered.

Lemma V.3. *The $f \in \mathcal{F} = \mathcal{P}(\mathcal{Y}^n)$ that minimizes*

$$\inf_{f\in\mathcal{F}} \max_{y^n\in\mathcal{Y}^n} \mathrm{E}_{P_{X^n|Y^n=y^n}} \sup_{q\in Q} R_n(f,q,y^n,X^n) \tag{29}$$

is

$$f^*(y^n) = \frac{\prod_{x^n\in\mathcal{X}^n} q_{ML}(y^n \parallel x^n)^{P_{X^n|Y^n}(x^n|y^n)}}{\sum_{z^n\in\mathcal{Y}^n} \prod_{x^n\in\mathcal{X}^n} q_{ML}(z^n \parallel x^n)^{P_{X^n|Y^n}(x^n|y^n)}}. \tag{30}$$

The value of (29) *is*

$$\log \sum_{z^n\in\mathcal{Y}^n} \prod_{x^n\in\mathcal{X}^n} q_{ML}(z^n \parallel x^n)^{P_{X^n|Y^n}(x^n,y^n)}. \tag{31}$$

Proof:

$$\mathrm{E}_{P_{X^n|Y^n=y^n}} \sup_{q\in Q} R_n(f,q,y^n,X^n)$$
$$= \mathrm{E}_{P_{X^n|Y^n=y^n}} [-\log p(y^n) + \log q_{ML}(y^n \parallel X^n)] \tag{32}$$
$$= \left[-\log p(y^n) + \mathrm{E}_{P_{X^n|Y^n=y^n}} \log q_{ML}(y^n \parallel X^n)\right] \tag{33}$$

Recall the property of logarithms for positive constants a,b,c,d: $a\log b + c\log d = \log b^a + \log d^c = \log b^a d^c$. Using this,

$$\mathrm{E}_{P_{X^n|Y^n=y^n}} \log q_{ML}(y^n \parallel X^n)$$
$$= \sum_{x^n\in\mathcal{X}^n} P_{X^n|Y^n}(x^n|y^n) \log q_{ML}(y^n \parallel x^n) \tag{34}$$
$$= \log \prod_{x^n\in\mathcal{X}^n} q_{ML}(y^n \parallel x^n)^{P_{X^n|Y^n}(x^n|y^n)} \tag{35}$$
$$= \log g(y^n) \tag{36}$$

for

$$g(y^n) = \prod_{x^n\in\mathcal{X}^n} q_{ML}(y^n \parallel x^n)^{P_{X^n|Y^n}(x^n|y^n)}.$$

By Lemma V.1, Lemma V.3 holds. ∎

The values of the minimax regrets in Lemma V.2 and Lemma V.3 can be interpreted as potentially offering a characterization of how much, from a sequential prediction perspective, the side information causally influences the outcome sequence, and thus a form of Granger causality in adversarial settings.

REFERENCES

[1] C. Granger, "Investigating causal relations by econometric models and cross-spectral methods," *Econometrica*, vol. 37, no. 3, pp. 424–438, 1969.

[2] J. Geweke, R. Meese, and W. Dent, "Comparing alternative tests of causality in temporal systems : Analytic results and experimental evidence," *Journal of Econometrics*, vol. 21, no. 2, pp. 161 – 194, 1983.

[3] M. Kamiński, M. Ding, W. Truccolo, and S. Bressler, "Evaluating causal relations in neural systems: Granger causality, directed transfer function and statistical assessment of significance," *Biological Cybernetics*, vol. 85, no. 2, pp. 145–157, 2001.

[4] N. Cesa-Bianchi and G. Lugosi, *Prediction, learning, and games*. Cambridge University Press, 2006.

[5] N. Merhav and M. Feder, "Universal prediction," *Information Theory, IEEE Transactions on*, vol. 44, no. 6, pp. 2124–2147, 2002.

[6] S. Kozat and A. Singer, "Min-max optimal universal prediction with side information," in *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, vol. 5. IEEE, 2004.

[7] Q. Xie and A. Barron, "Asymptotic minimax regret for data compression, gambling, and prediction," *Information Theory, IEEE Transactions on*, vol. 46, no. 2, pp. 431–445, 2002.

[8] T. Cover and J. Thomas, *Elements of information theory*. Wiley-Interscience, 2006.

[9] G. Kramer, "Directed information for channels with feedback," Ph.D. dissertation, University of Manitoba, Canada, 1998.

[10] J. Massey, "Causality, feedback and directed information," in *Proc. 1990 Intl. Symp. on Info. Th. and its Applications*. Citeseer, 1990, pp. 27–30.

[11] H. Marko, "The bidirectional communication theory–a generalization of information theory," *Communications, IEEE Transactions on*, vol. 21, no. 12, pp. 1345–1351, Dec 1973.

[12] J. Rissanen and M. Wax, "Measures of mutual and causal dependence between two time series (Corresp.)," *IEEE Transactions on Information Theory*, vol. 33, no. 4, pp. 598–601, 1987.