

Directed Information on Abstract Spaces: Properties and Variational Equalities

Charalambos D. Charalambous and Photios A. Stavrou

Abstract

Directed information or its variants are utilized extensively in the characterization of the capacity of channels with memory and feedback, nonanticipative lossy data compression, and their generalizations to networks.

In this paper, we derive several functional and topological properties of directed information for general abstract alphabets (complete separable metric spaces) using the topology of weak convergence of probability measures. These include convexity of the set of consistent distributions, which uniquely define causally conditioned distributions, convexity and concavity of directed information with respect to the sets of consistent distributions, weak compactness of these sets of distributions, their joint distributions and their marginals. Furthermore, we show lower semicontinuity of directed information, and under certain conditions we also establish continuity of directed information. Finally, we derive variational equalities for directed information, including sequential versions. These may be viewed as the analogue of the variational equalities of mutual information (utilized in Blahut-Arimoto algorithm).

In summary, we extend the basic functional and topological properties of mutual information to directed information. These properties are discussed in the context of extremum problems of directed information.

Index Terms

Directed information, weak convergence, convexity, concavity, lower semicontinuity, continuity, variational equalities.

I. INTRODUCTION

Directed information quantifies the directivity of information defined by a causal sequence of feedback and feedforward channel conditional distributions [4], [5]. Specifically, given two sequences of Random Variables (RV's) $X^n \triangleq \{X_0, X_1, \dots, X_n\} \in \mathcal{X}_{0,n} \triangleq \times_{i=0}^n \mathcal{X}_i$, $Y^n \triangleq \{Y_0, Y_1, \dots, Y_n\} \in \mathcal{Y}_{0,n} \triangleq \times_{i=0}^n \mathcal{Y}_i$, where \mathcal{X}_i and \mathcal{Y}_i are the input and output alphabets of a channel, respectively, and $\mathcal{B}(\mathcal{X}_i)$, $\mathcal{B}(\mathcal{Y}_i)$, the corresponding measurable spaces, directed information from X^n to Y^n is often defined via conditional mutual information [5], [6] as follows.

$$I(X^n \rightarrow Y^n) \triangleq \sum_{i=0}^n I(X^i; Y_i | Y^{i-1}) \quad (\text{I.1})$$

$$= \sum_{i=0}^n \int_{\mathcal{X}_{0,i} \times \mathcal{Y}_{0,i}} \log \left(\frac{dP_{Y_i | Y^{i-1}, X^i}(\cdot | y^{i-1}, x^i)}{dP_{Y_i | Y^{i-1}}(\cdot | y^{i-1})}(y_i) \right) P_{X^i, Y^i}(dx^i, dy^i) \quad (\text{I.2})$$

$$\equiv \mathbb{I}_{X^n \rightarrow Y^n}(P_{X_i | X^{i-1}, Y^{i-1}}, P_{Y_i | Y^{i-1}, X^i} : i = 0, 1, \dots, n) \quad (\text{I.3})$$

where notion (I.3) indicates that directed information $I(X^n \rightarrow Y^n)$ is a functional of two collections of causally conditioned distributions, $\{P_{Y_i | Y^{i-1}, X^i} : i = 0, \dots, n\}$, and $\{P_{X_i | X^{i-1}, Y^{i-1}} : i = 0, 1, \dots, n\}$, called feedforward distribution, and feedback distribution, respectively, which uniquely define

This work was financially supported by a medium size University of Cyprus grant entitled DIMITRIS. Parts of the material in this paper were presented at the IEEE International Symposium on Information Theory, Boston MA, July 1–6 2012 [1], at the IEEE International Symposium on Information Theory, Istanbul, Turkey, July 7–12 2013 [2], and in book series “Lecture Notes in Control and Information Sciences” [3].

The authors are with the Department of Electrical and Computer Engineering (ECE), University of Cyprus, 75 Kallipoleos Avenue, P.O. Box 20537, Nicosia, 1678, Cyprus, Email: {chadcha, stavrou.fotios}@ucy.ac.cy

the joint distribution $\{P_{X^i, Y^i} : i = 0, 1, \dots, n\}$ and the conditional distribution $\{P_{Y_i|Y^{i-1}} : i = 0, 1, \dots, n\}$ of the RV's $\{(X^i, Y^i) : i = 0, 1, \dots, n\}$.

By Bayes' rule, for any $A_j \in \mathcal{B}(\mathcal{X}_j), B_j \in \mathcal{B}(\mathcal{Y}_j), j = 0, 1, \dots, i$, the joint distribution decomposes into

$$P_{X^i, Y^i}(A_0, B_0, \dots, A_i, B_i) = \int_{A_0} P_{X_0}(dx_0) \int_{B_0} P_{Y_0|X_0, Y^{-1}}(dy_0|x_0, y^{-1}) \dots \\ \dots \int_{A_i} P_{X_i|X^{i-1}, Y^{i-1}}(dx_i|x^{i-1}, y^{i-1}) \int_{B_i} P_{Y_i|Y^{i-1}, X^i}(dy_i|y^{i-1}, x^i), \quad i = 0, 1, \dots, n. \quad (\text{I.4})$$

Formally, we represent (I.4) by $P_{X^i, Y^i}(dx^i, dy^i) = \otimes_{j=0}^i (P_{X_j|X^{j-1}, Y^{j-1}} \otimes P_{Y_j|Y^{j-1}, X^j})$, and we call it an $(n+1)$ -fold compound probability distribution.

If the distributions $\{P_{X_i|X^{i-1}, Y^{i-1}}, P_{Y_i|Y^{i-1}, X^i} : i = 0, \dots, n\}$ are defined with respect to the probability density functions of continuous valued RV's $\{(X_i, Y_i) : i = 0, 1, \dots, n\}$, denoted by, $\{f_{X_i|X^{i-1}, Y^{i-1}}, f_{Y_i|Y^{i-1}, X^i} : i = 0, \dots, n\}$, then (I.1) reduces to

$$I(X^n \rightarrow Y^n) = \sum_{i=0}^n \int_{\mathcal{X}_{0,i} \times \mathcal{Y}_{0,i}} \log \left(\frac{f_{Y_i|Y^{i-1}, X^i}(y_i|y^{i-1}, x^i)}{f_{Y_i|Y^{i-1}}(y_i|y^{i-1})} \right) f_{X^i, Y^i}(x^i, y^i) dx^i dy^i.$$

If the distributions $\{P_{X_i|X^{i-1}, Y^{i-1}}, P_{Y_i|Y^{i-1}, X^i} : i = 0, \dots, n\}$ are defined with respect to the probability mass functions of countable or finite alphabet valued RV's $\{(X_i, Y_i) : i = 0, \dots, n\}$, denoted by, $\{p_{X_i|X^{i-1}, Y^{i-1}}, p_{Y_i|Y^{i-1}, X^i} : i = 0, \dots, n\}$, then (I.1) reduces to

$$I(X^n \rightarrow Y^n) = \sum_{i=0}^n \sum_{(x^i, y^i) \in \mathcal{X}_{0,i} \times \mathcal{Y}_{0,i}} \log \left(\frac{p_{Y_i|Y^{i-1}, X^i}(y_i|y^{i-1}, x^i)}{p_{Y_i|Y^{i-1}}(y_i|y^{i-1})} \right) p_{X^i, Y^i}(x^i, y^i).$$

In information theory, directed information (I.1) or its variants are used to characterize capacity of channels with memory and feedback [7]–[14], lossy data compression of sequential codes [7], [15], lossy data compression with feedforward information at the decoder [16], and capacity of networks, such as, the two-way channel, the multiple access channel [6], [17], etc. Some of the above references derive coding theorems for an anthology of problems of information theory, under any one of the assumptions: (a) stationary ergodic processes $\{(X_i, Y_i) : i = 0, 1, \dots\}$, (b) Dobrushin's stability of the information density $\sum_{i=0}^n \log \left(\frac{dP_{Y_i|Y^{i-1}, X^i}}{dP_{Y_i|Y^{i-1}}} \right)$, (c) Verdú and Han's information spectrum methods [18]. Moreover, directed information is also utilized in a variety of problems subject to causality constraints, such as, gambling, portfolio theory, data compression and hypothesis testing [19], in biology as an alternative to Granger's measure of causality [20]–[22], and in relating Bayesian filtering theory to sequential and nonanticipative RDF [23], [24].

Directed information is initially introduced by Marko [4] by decomposing Shannon's self-mutual information into two directional parts, and then taking expectation. Although, directed information is defined via a sequence of conditional mutual informations (i.e., (I.1)), for general abstract alphabets (i.e., continuous) or distributions which are not necessarily continuous (i.e., induced by mixture of continuous and finite alphabet RVs) its functional and topological properties are not well understood [6].

Further, for such alphabet spaces or distributions, specific functional properties of mutual information expressed as a functional $I(X^n; Y^n) \equiv \mathbb{I}_{X^n, Y^n}(P_{X^n}, P_{Y^n|X^n})$, of the two distributions $\{P_{X^n}, P_{Y^n|X^n}\}$, such as, convexity, concavity, and topological properties such as lower semicontinuity (with respect to the topology of weak convergence of probability measures), at first glance, do not translate into analogous properties for directed information. The reason is that directed information $I(X^n \rightarrow Y^n) \equiv \mathbb{I}_{X^n \rightarrow Y^n}(P_{X_i|X^{i-1}, Y^{i-1}}, P_{Y_i|Y^{i-1}, X^i} : i = 0, 1, \dots, n)$ is a functional of two sequences of distributions $\{P_{X_i|X^{i-1}, Y^{i-1}}, P_{Y_i|Y^{i-1}, X^i} : i = 0, 1, \dots, n\}$, and the joint and marginal distributions are induced from these sequences of distributions. Such properties are important in extremum problems of directed information.

Similarly, it is not obvious whether the well-known variational equalities of mutual information, which involve a single maximization or minimization of appropriate functionals over appropriate convex sets, have counter parts, for directed information, which involve nested maximization and minimization operations of appropriate functionals over appropriate convex sets, giving rise to sequential variational equalities. Such sequential variational equalities, are important to develop computationally efficient sequential algorithms to compute capacity of channels with memory and feedback, similar to the Blahut-Arimoto algorithm [25], of memoryless channels.

These properties together with compactness of subsets of the sets of the conditional distributions $\{P_{X_i|X^{i-1}, Y^{i-1}} : i = 0, 1, \dots, n\}$ and $\{P_{Y_i|Y^{i-1}, X^i} : i = 0, 1, \dots, n\}$, are fundamental to analyze extremum problems of directed information related to channel capacity, sequential and nonanticipative RDF, their generalizations to networks, etc, for countable and abstract alphabets.

Recently, in [26] it is demonstrated via several examples that Shannon information measures, such as, entropy, relative entropy, mutual information, and conditional mutual information, when defined on countable alphabets, are discontinuous with respect to strong topologies (i.e., induced by total variational distance metrics on the space of probability distributions). Since directed information in (I.1) involves a sequence of conditional mutual informations, the observations in [26] also apply to directed information. The lack of continuity is attributed to the fact that mutual information and directed information are defined from relative entropy, and relative entropy is lower semicontinuous with respect to distributions [27]. For such abstract alphabets problems, it was recognized many years ago (see [28], [29]) that the analysis of capacity formulae based on single letter mutual information formulae requires tools from the topology of weak convergence of probability measures (or equivalently the weak* topology), in order to identify global and local analytical properties of channel input distributions which maximize mutual information.

The main objective of this paper is to derive functional properties, topological properties, and sequential variational equalities, for directed information, when the distributions are defined on abstract alphabets, and to provide appropriate conditions for these to hold. The methodology and the main results are summarized below.

- R1) Introduce an equivalent directed information definition expressed via information divergence $\mathbb{D}(\cdot||\cdot)$, as a functional of two consistent families of conditional distributions $\mathbf{P}(\cdot|\mathbf{y})$ on $\mathcal{X}^{\mathbb{N}_0} \triangleq \times_{i=0}^{\infty} \mathcal{X}_i$ parametrized by $\mathbf{y} = (y_0, y_1, \dots) \in \mathcal{Y}^{\mathbb{N}_0} \triangleq \times_{i=0}^{\infty} \mathcal{Y}_i$, and $\mathbf{Q}(\cdot|\mathbf{x})$ on $\mathcal{Y}^{\mathbb{N}_0}$ parametrized by $\mathbf{x} \in \mathcal{X}^{\mathbb{N}_0}$, which uniquely define $\{P_{X_i|X^{i-1}, Y^{i-1}} : i \in \mathbb{N}_0\}$ and $\{P_{Y_i|Y^{i-1}, X^i} : i \in \mathbb{N}_0\}$, respectively, and vice-versa, and their $(n+1)$ -fold compound probability distributions $\overleftarrow{P}_{0,n}(dx^n|y^{n-1}) \triangleq \otimes_{i=0}^n P_{X_i|X^{i-1}, Y^{i-1}}(dx_i|x^{i-1}, y^{i-1})$, $\overrightarrow{Q}_{0,n}(dy^n|x^n) \triangleq \otimes_{i=0}^n P_{Y_i|Y^{i-1}, X^i}(dy_i|y^{i-1}, x^i)$.
- R2) Show convexity of the consistent families of the conditional distributions $\mathbf{P}(\cdot|\mathbf{y})$ for $\mathbf{y} \in \mathcal{Y}^{\mathbb{N}_0}$, $\mathbf{Q}(\cdot|\mathbf{x})$ for $\mathbf{x} \in \mathcal{X}^{\mathbb{N}_0}$.
- R3) Show convexity and concavity of directed information as a functional with respect to the consistent families of conditional distributions $\mathbf{Q}(\cdot|\mathbf{x})$ for $\mathbf{x} \in \mathcal{X}^{\mathbb{N}_0}$, and $\mathbf{P}(\cdot|\mathbf{y})$ for $\mathbf{y} \in \mathcal{Y}^{\mathbb{N}_0}$, respectively.
- R4) Show under certain conditions, weak compactness of the consistent families of conditional distributions $\mathbf{P}(\cdot|\mathbf{x})$ for $\mathbf{x} \in \mathcal{X}^{\mathbb{N}_0}$, and $\mathbf{Q}(\cdot|\mathbf{y})$ for $\mathbf{y} \in \mathcal{Y}^{\mathbb{N}_0}$, and of their marginals and joint distribution.
- R5) Show lower semicontinuity of directed information as a functional of the consistent families of the conditional distributions $\mathbf{P}(\cdot|\mathbf{y})$ for $\mathbf{y} \in \mathcal{Y}^{\mathbb{N}_0}$, and $\mathbf{Q}(\cdot|\mathbf{x})$ for $\mathbf{x} \in \mathcal{X}^{\mathbb{N}_0}$, and under certain conditions, continuity of directed information as a functional of the family $\mathbf{P}(\cdot|\mathbf{y})$ for $\mathbf{y} \in \mathcal{Y}^{\mathbb{N}_0}$.
- R6) Express directed information in terms of variational equalities involving sequential minimization and sequential maximization operations over conditional distributions.
- R7) Illustrate that R1)–R6) extend naturally to three sequences of RV's $X^n \in \mathcal{X}_{0,n}$, $Y^n \in \mathcal{Y}_{0,n}$, $Z^n \in \mathcal{Z}_{0,n}$, or more, which cover directed information measures for networks, and possible problems with side information.
- R8) Discuss applications of R1)–R6).

The above functional and topological properties are shown by invoking the topology of weak convergence

of probability measures on Polish spaces and Prohorov's theorems [30], [31]. Some of the results described above are obtained by utilizing analogies between communication channels with memory and feedback, and stochastic optimal control problems in which the control element and the controlled element are the sequences of conditional distributions, $\{P_{X_i|X^{i-1}, Y^{i-1}} : i = 0, 1, \dots\}$ and $\{P_{Y_i|Y^{i-1}, X^i} : i = 0, 1, \dots\}$, respectively, [32], [33].

Items R1)-R7) extend various functional and topological properties of mutual information $I(X^n; Y^n) \equiv \mathbb{I}_{X^n; Y^n}(P_{X^n}, P_{Y^n|X^n})$ as a functional of $\{P_{X^n}, P_{Y^n|X^n}\}$ to directed information.

From the practical point of view, there are many potential applications of R1)-R7). Below, we briefly discuss some of them.

The concavity and convexity properties are important in deriving tight bounds in applications of converse coding theorems, in identifying properties of extremum problems involving feedback capacity [6], [34] and sequential and nonanticipative lossy data compression via the nonanticipative RDF [35], in relating Bayesian filtering theory and nonanticipative RDF [23], in network communication applications [36], [37], etc. The semicontinuity and continuity of directed information, and the compactness of the consistent families of distributions $\mathbf{P}(\cdot|\mathbf{y})$ for $\mathbf{y} \in \mathcal{Y}^{\mathbb{N}_0}$, and $\mathbf{Q}(\cdot|\mathbf{x})$ for $\mathbf{x} \in \mathcal{X}^{\mathbb{N}_0}$, are crucial, when addressing questions of existence of extremum solutions to problems involving feedback capacity, sequential and nonanticipative lossy data compression, computations of extremum solutions and their properties, and in extending existing coding theorems to abstract alphabets [38]. For example, the converse part of coding theorem for feedback capacity presupposes existence of optimal channel input distribution maximizing directed information, and existence of its per unit time limit. The variational equalities are important in generalizing Blahut-Arimoto computation schemes of single letter mutual information expressions [39] to sequential Blahut-Arimoto schemes, involving extremum problems of directed information, such as, in problems of evaluating feedback capacity (see [40]).

Throughout the paper, we illustrate applications of the results to the following extremum problems.

Capacity of channels with memory and feedback. Consider the extremum problem of channel capacity with memory and feedback. Under the assumption of stationary ergodic processes $\{(X_i, Y_i) : i = 0, 1, \dots\}$ or Dobrushin's directed information stability and transmission cost stability, the operational definition of capacity is given by the following extremum problem [11].

$$C^{fb}(P) \triangleq \liminf_{n \rightarrow \infty} \sup_{\{P_{X_i|X^{i-1}, Y^{i-1}} : i=0,1,\dots,n\} \in \mathcal{P}_{0,n}(P)} \frac{1}{n+1} I(X^n \rightarrow Y^n), \quad (\text{I.5})$$

where $\mathcal{P}_{0,n}(P)$ is the transmission cost constraint set defined by

$$\mathcal{P}_{0,n}(P) \triangleq \left\{ P_{X_i|X^{i-1}, Y^{i-1}}, i = 0, 1, \dots, n : \frac{1}{n+1} \mathbb{E}\{c_{0,n}(x^n, y^{n-1})\} \leq P \right\}, P \geq 0 \quad (\text{I.6})$$

and $c_{0,n} : \mathcal{X}_{0,n} \times \mathcal{Y}_{0,n-1} \mapsto [0, \infty)$, $c_{0,n}(x^n, y^{n-1}) \triangleq \sum_{i=0}^n g_i(x^i, y^{i-1})$ is a measurable function denoting the cost of transmitting symbols over the channel.

The task of showing existence of a sequence of probability distributions $\{P_{X_i|X^{i-1}, Y^{i-1}} : i = 0, 1, \dots, n\} \in \mathcal{P}_{0,n}(P)$ which achieves the supremum in (I.5) for continuous or countable alphabet spaces is not easy. The main difficulty arises from the fact that $I(X^n \rightarrow Y^n)$ is a functional of the two sequences of distributions $\{P_{X_i|X^{i-1}, Y^{i-1}}, P_{Y_i|Y^{i-1}, X^i} : i = 0, 1, \dots, n\}$, unlike mutual information $I(X^n; Y^n) \equiv \mathbb{I}_{X^n; Y^n}(P_{X^n}, P_{Y^n|X^n})$, which inherits most of its properties from those of relative entropy between the joint distribution P_{Y^n, X^n} and the product of its marginals $P_{X^n} \times P_{Y^n}$. However, we show by utilizing some of the results described under R1)–R6), existence of such conditional distribution and identify several properties of the optimal conditional channel input distribution.

Generalized Information Nonanticipative or Sequential RDF. Consider the extremum problem of general information nonanticipative RDF, or sequential RDF [7], which is a variant of classical RDF [41],

defined by [23], [42]

$$R^{na}(D) \triangleq \limsup_{n \rightarrow \infty} \inf_{\{P_{Y_i|Y^{i-1}, X^i}, i=0,1,\dots,n\} \in \mathcal{Q}_{0,n}(D)} \frac{1}{n+1} I(X^n \rightarrow Y^n), \quad (\text{I.7})$$

where $\mathcal{Q}_{0,n}(D)$ is the fidelity constraint set defined by

$$\mathcal{Q}_{0,n}(D) \triangleq \left\{ Q_{Y_i|Y^{i-1}, X^i}, i = 0, 1, \dots, n : \frac{1}{n+1} \mathbb{E}\{d_{0,n}(x^n, y^n)\} \leq D \right\}, \quad D \geq 0 \quad (\text{I.8})$$

and $d_{0,n} : \mathcal{X}_{0,n} \times \mathcal{Y}_{0,n} \mapsto [0, \infty]$, $d_{0,n}(x^n, y^n) \triangleq \sum_{i=0}^n \rho_i(x^i, y^i)$ is a measurable function denoting the distortion function of reconstructing x_i by y_i , $i = 0, 1, \dots, n$. Note that if $P_{X_i|X^{i-1}, Y^{i-1}} = P_{X_i|X^{i-1}}$, a.a. (x^{i-1}, y^{i-1}) , $i = 0, 1, \dots, n$, then it can be shown that (I.7), (I.8) are degraded to Gorbunov and Pinsker's nonanticipatory ϵ -entropy [43].

For both extremum problems (I.5), (I.7), we illustrate applications of R1)–R6) in showing existence of solutions, identifying properties of optimal solutions, and in constructing sequential versions of Blahut Arimoto Algorithm (BAA) [39].

The rest of the paper is structured as follows. Section II introduces two equivalent definitions of nonanticipative channels on abstract spaces (R1)). Section III derives the functional and topological properties of directed information (R2)–R5)). Section IV derives sequential variational equalities of directed information (R6)).

II. EQUIVALENT NONANTICIPATIVE CHANNELS ON ABSTRACT SPACES

In this section, our aim is to establish two equivalent definitions of the sequence of conditional distributions or basic processes, which define any probabilistic channel with nonanticipative (causal) feedback, that relate causally the input-output behavior of the channel. This formulation is utilized extensively to establish the results stated under R1)–R7). The first definition of conditional distributions is the usual one found in many papers, e.g., [6], [7], [10]–[13], for finite alphabets spaces. The aforementioned definition is described via a family of multi-fold compound conditional distributions (see Fig. II.1, (a)). The second definition is described via a family of conditional distributions defined on product alphabets, which satisfy a certain consistency condition (see Fig. II.1, (b)). The second definition is often utilized in

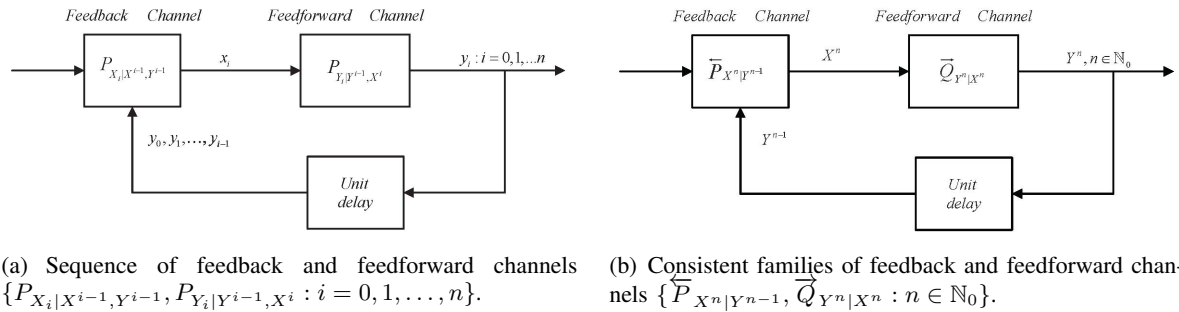


Fig. II.1. Equivalent Representations of Feedback/Feedforward Channels.

the stochastic control literature, in which there is a control process and a controlled process [32], [33]. Indeed, the analogy is that $\{X_i : i = 0, 1, \dots\}$ is the control process, $\{Y_i : i = 0, 1, \dots\}$ is the controlled process, $\{P_{X_i|X^{i-1}, Y^{i-1}} : i = 0, 1, \dots\}$ is the control element, and $\{P_{Y_i|Y^{i-1}, X^i} : i = 0, 1, \dots\}$ is the controlled element. The second definition is more convenient, because the directed information density $i(X^n \rightarrow Y^n) \triangleq \log \left(\bigotimes_{i=0}^n \frac{dP_{Y_i|Y^{i-1}, X^i}}{dP_{Y_i|Y^{i-1}}} \right) = \sum_{i=0}^n \log \left(\frac{dP_{Y_i|Y^{i-1}, X^i}}{dP_{Y_i|Y^{i-1}}} \right)$ corresponding to $I(X^n \rightarrow Y^n)$, can be equivalently expressed in terms of two consistent families of conditional distributions, namely, $Q(\cdot|x)$ on

$\mathcal{Y}^{\mathbb{N}_0}$ given $\mathbf{x} = (x_0, x_1, \dots) \in \mathcal{X}^{\mathbb{N}_0}$, and $\mathbf{P}(\cdot|\mathbf{y})$ on $\mathcal{X}^{\mathbb{N}_0}$ given $\mathbf{y} = (y_0, y_1, \dots) \in \mathcal{Y}^{\mathbb{N}_0}$, which uniquely define $\{P_{Y_i|Y^{i-1}, X^i} : i = 0, 1, \dots\}$ and $\{P_{X_i|X^{i-1}, Y^{i-1}} : i = 0, 1, \dots\}$, respectively, and vice-versa, such that $i(X^n \rightarrow Y^n) = \log \left(\frac{d\mathbf{Q}(\cdot|\mathbf{x}^n)}{d\nu^{\mathbf{P} \otimes \mathbf{Q}}(\cdot)}(y^n) \right) - a.s.$, where $\nu^{\mathbf{P} \otimes \mathbf{Q}}(\cdot)$ is the marginal distribution on $\times_{i=0}^n \mathcal{Y}_i$ obtained from $\mathbf{P}(\cdot|\mathbf{y})$ and $\mathbf{Q}(\cdot|\mathbf{x})$. Once the conditions on the abstract spaces $\{(\mathcal{Y}_i, \mathcal{X}_i) : i = 0, 1, \dots\}$ are identified, and the consistency conditions are introduced, then it can be shown that $i(X^n \rightarrow Y^n)$ has another version given by $i(X^n \rightarrow Y^n) = \log \left(\frac{d(\mathbf{P}(\cdot|\cdot) \otimes \mathbf{Q}(\cdot|\cdot))}{d(\mathbf{P}(\cdot|\cdot) \otimes \nu^{\mathbf{P} \otimes \mathbf{Q}}(\cdot))}(x^n, y^n) \right) - a.s.$, where \otimes denotes the compound probability distribution defined by $\mathbf{P}(\cdot|\cdot)$ and $\mathbf{Q}(\cdot|\cdot)$, and similarly for the rest of the measures. Consequently, directed information can be expressed in terms of Kullback-Leibler distance $\mathbb{D}(\mathbf{P} \otimes \mathbf{Q} || \mathbf{P} \otimes \nu^{\mathbf{P} \otimes \mathbf{Q}})$ ¹.

Notations and Preliminaries.

Denote the set of nonnegative integers by $\mathbb{N}_0 \triangleq \{0, 1, 2, \dots\}$, and the restriction of \mathbb{N}_0 to positive integers by $\mathbb{N}_1 \triangleq \{1, 2, \dots\}$, and to a finite set by $\mathbb{N}_0^n \triangleq \{0, 1, 2, \dots, n\}$. Introduce two sequences of spaces $\{(\mathcal{X}_n, \mathcal{B}(\mathcal{X}_n)) : n \in \mathbb{N}_0\}$ and $\{(\mathcal{Y}_n, \mathcal{B}(\mathcal{Y}_n)) : n \in \mathbb{N}_0\}$, called basic measurable spaces, where $\mathcal{X}_n, \mathcal{Y}_n, n \in \mathbb{N}_0$ are topological spaces, and $\mathcal{B}(\mathcal{X}_n)$ and $\mathcal{B}(\mathcal{Y}_n)$ are Borel σ -algebras of subsets of \mathcal{X}_n and \mathcal{Y}_n , respectively. The set of probability measures on any measurable space $(\mathcal{Z}, \mathcal{B}(\mathcal{Z}))$ is denoted by $\mathcal{M}_1(\mathcal{Z})$.

For each $n \in \mathbb{N}_0$ define the product spaces

$$(\mathcal{X}_{0,n}, \mathcal{B}(\mathcal{X}_{0,n})) \triangleq (\times_{i=0}^n \mathcal{X}_i, \otimes_{i=0}^n \mathcal{B}(\mathcal{X}_i)), (\mathcal{Y}_{0,n}, \mathcal{B}(\mathcal{Y}_{0,n})) \triangleq (\times_{i=0}^n \mathcal{Y}_i, \otimes_{i=0}^n \mathcal{B}(\mathcal{Y}_i)).$$

For each $n \in \mathbb{N}_0$, let \mathcal{X}_n and \mathcal{Y}_n be the spaces of all possible outcomes. Given the data up to and including the n th time, specifically, $(x_i, y_i) \in \mathcal{X}_i \times \mathcal{Y}_i$, $i = 0, 1, \dots, n$, the probability distributions at time $(n+1)$ are $p_{n+1}(A_{n+1}|x_0, \dots, x_n, y_0, \dots, y_n)$ and $q_{n+1}(B_{n+1}|y_0, \dots, y_n, x_0, \dots, x_{n+1})$, $A_{n+1} \in \mathcal{B}(\mathcal{X}_{n+1})$, $B_{n+1} \in \mathcal{B}(\mathcal{Y}_{n+1})$. Hence, each possible outcome of the experiment is a sequence $\omega = (x_0, y_0, x_1, y_1, \dots)$ with $x_n \in \mathcal{X}_n, y_n \in \mathcal{Y}_n$ for each $n \in \mathbb{N}_0$ (here, no time ordering is required).

Consequently, define the sample space Ω and the algebra \mathcal{F} of all experiments by

$$(\Omega, \mathcal{F}) \triangleq \left(\times_{n \in \mathbb{N}_0} (\mathcal{X}_n \times \mathcal{Y}_n), \otimes_{n \in \mathbb{N}_0} (\mathcal{B}(\mathcal{X}_n) \otimes \mathcal{B}(\mathcal{Y}_n)) \right).$$

Associated with the basic measurable spaces there are two basic sequences of Random Variables (RV's) $\{X_n : n \in \mathbb{N}_0\}$ and $\{Y_n : n \in \mathbb{N}_0\}$, such that for each $n \in \mathbb{N}_0$, they take values $X_n \in \mathcal{X}_n$ and $Y_n \in \mathcal{Y}_n$. These are introduced as follows.

Let $X_0, Y_0, X_1, Y_1, \dots$ be the coordinate RV's. For each $n \in \mathbb{N}_0$

$$X_n(\omega) = x_n, Y_n(\omega) = y_n \text{ if } \omega = (x_0, y_0, x_1, y_1, \dots).$$

Clearly, $X_n : (\Omega, \mathcal{F}) \mapsto (\mathcal{X}_n, \mathcal{B}(\mathcal{X}_n))$, $Y_n : (\Omega, \mathcal{F}) \mapsto (\mathcal{Y}_n, \mathcal{B}(\mathcal{Y}_n))$, and for each outcome $\omega \in \Omega$ of the experiment, $X_n(\omega), Y_n(\omega)$ are the results of the n th time. Similarly, $X^n \triangleq \{X_0, \dots, X_n\}$ and $Y^n \triangleq \{Y_0, \dots, Y_n\}$ denote the result of the trials up to and including the n th time; they are RV taking values in $(\mathcal{X}_{0,n}, \mathcal{B}(\mathcal{X}_{0,n}))$ and $(\mathcal{Y}_{0,n}, \mathcal{B}(\mathcal{Y}_{0,n}))$, respectively. The objective is to construct a measure \mathbb{P} on (Ω, \mathcal{F}) consistent with the data (e.g., measurable spaces and conditional distributions).

For every $n \in \mathbb{N}_0$, define the σ -algebras generated by $\{X_0, X_1, \dots, X_n\}$ and $\{Y_0, Y_1, \dots, Y_n\}$ by

$$\mathcal{F}(X^n) \triangleq \sigma\{X_0, X_1, \dots, X_n\}, \mathcal{F}(Y^n) \triangleq \sigma\{Y_0, Y_1, \dots, Y_n\}.$$

Then every event $H \in \mathcal{F}(X^n)$ has the form

$$H = \left\{ (X_0, X_1, \dots, X_n) \in A \right\} = A \times \mathcal{X}_{n+1} \times \mathcal{X}_{n+2} \dots, A \in \mathcal{B}(\mathcal{X}_{0,n})$$

¹In the rest of the paper we write ν instead of $\nu^{\mathbf{P} \otimes \mathbf{Q}}$ omitting its explicit dependence on $\mathbf{P}(\cdot|\mathbf{y})$ and $\mathbf{Q}(\cdot|\mathbf{x})$.

and H is called a cylinder set with base $A \in \mathcal{B}(\mathcal{X}_{0,n})$. Similarly, for an event $J \in \mathcal{F}(Y^n)$

$$J = \{(Y_0, Y_1, \dots, Y_n) \in B\} = B \times \mathcal{Y}_{n+1} \times \mathcal{Y}_{n+2} \dots, \quad B \in \mathcal{B}(\mathcal{Y}_{0,n})$$

and J is a cylinder set with base $B \in \mathcal{B}(\mathcal{Y}_{0,n})$.

Points in the Cartesian countable product spaces $\mathcal{X}^{\mathbb{N}_0} \triangleq \times_{n \in \mathbb{N}_0} \mathcal{X}_n$, $\mathcal{Y}^{\mathbb{N}_0} \triangleq \times_{n \in \mathbb{N}_0} \mathcal{Y}_n$ are denoted by $\mathbf{x} \triangleq \{x_0, x_1, \dots\} \in \mathcal{X}^{\mathbb{N}_0}$, $\mathbf{y} \triangleq \{y_0, y_1, \dots\} \in \mathcal{Y}^{\mathbb{N}_0}$, respectively. Similarly, for $n \in \mathbb{N}_0$, points in $\mathcal{X}_{0,n} \triangleq \times_{i=0}^n \mathcal{X}_i$, $\mathcal{Y}_{0,n} \triangleq \times_{i=0}^n \mathcal{Y}_i$ are denoted by $x^n \triangleq \{x_0, x_1, \dots, x_n\} \in \mathcal{X}_{0,n}$, $y^n \triangleq \{y_0, y_1, \dots, y_n\} \in \mathcal{Y}_{0,n}$, respectively. Let $\mathcal{B}(\mathcal{X}^{\mathbb{N}_0})$ and $\mathcal{B}(\mathcal{Y}^{\mathbb{N}_0})$ denote the σ -algebras in $\mathcal{X}^{\mathbb{N}_0}$ and $\mathcal{Y}^{\mathbb{N}_0}$, respectively, generated by cylinder sets (e.g., $\mathcal{B}(\mathcal{X}^{\mathbb{N}_0})$ is the smallest Borel σ -algebra containing all cylinder sets $\{\mathbf{x} = (x_0, x_1, \dots) \in \mathcal{X}^{\mathbb{N}_0} : x_0 \in A_0, x_1 \in A_1, \dots, x_n \in A_n\}, A_i \in \mathcal{B}(\mathcal{X}_i), i \in \mathbb{N}_0$). The Borel σ -algebra $\mathcal{B}(\mathcal{X}^{\mathbb{N}_0})$ is denoted by $\otimes_{i \in \mathbb{N}_0} \mathcal{B}(\mathcal{X}_i)$. Hence, $\mathcal{B}(\mathcal{X}_{0,n})$ and $\mathcal{B}(\mathcal{Y}_{0,n})$ denote the σ -algebras of cylinder sets in $\mathcal{X}^{\mathbb{N}_0}$ and $\mathcal{Y}^{\mathbb{N}_0}$, respectively, with bases over $A_i \in \mathcal{B}(\mathcal{X}_i)$, $i \in \mathbb{N}_0^n$, and $B_i \in \mathcal{B}(\mathcal{Y}_i)$, $i \in \mathbb{N}_0^n$, respectively.

Backward or Feedback Channel.

Suppose for each $n \in \mathbb{N}_0$, the conditional distribution of the RV $X_n \in \mathcal{X}_n$ is determined provided the values of the basic processes $X^{n-1} = x^{n-1} \in \mathcal{X}_{0,n-1}$ and $Y^{n-1} = y^{n-1} \in \mathcal{Y}_{0,n-1}$ are known, and let $\{p_n(dx_n|x^{n-1}, y^{n-1}) : n \in \mathbb{N}_0\}$ denote the collection of these distributions. At $n = 0$, the distribution is $p_0(dx_0|x^{-1}, y^{-1})$, where (x^{-1}, y^{-1}) are either fixed, or $p_0(dx_0|x^{-1}, y^{-1}) = p(dx_0)$, depending on the convention used. Without loss of generality, we assume $p_0(dx_0|x^{-1}, y^{-1}) \triangleq p_0(x_0)$ (i.e., $\sigma\{X^{-1}, Y^{-1}\} = \{\emptyset, \Omega\}$). For each $n \in \mathbb{N}_0$, the functions $p_n(\cdot|\cdot, \cdot) : \mathcal{X}_n \times \mathcal{X}_{0,n-1} \times \mathcal{Y}_{0,n-1} \mapsto [0, 1]$ are candidates of distributions of the sequence of RV's $\{X_n : n \in \mathbb{N}_0\}$ on $\{(\mathcal{X}_n, \mathcal{B}(\mathcal{X}_n)) : n \in \mathbb{N}_0\}$ if and only if the following conditions hold.

- i) For every $n \in \mathbb{N}_0$, and $x^{n-1} \in \mathcal{X}_{0,n-1}$, $y^{n-1} \in \mathcal{Y}_{0,n-1}$, $p_n(\cdot|x^{n-1}, y^{n-1})$ is a probability measure on $\mathcal{B}(\mathcal{X}_n)$;
- ii) For every $n \in \mathbb{N}_0$, and $A_n \in \mathcal{B}(\mathcal{X}_n)$, $p_n(A_n|\cdot, \cdot)$ is an $\otimes_{i=0}^{n-1} (\mathcal{B}(\mathcal{X}_i) \otimes \mathcal{B}(\mathcal{Y}_i))$ -measurable function of $x^{n-1} \in \mathcal{X}_{0,n-1}$, $y^{n-1} \in \mathcal{Y}_{0,n-1}$.

For every $n \in \mathbb{N}_0$, the set of all functions that satisfy **i)**, **ii)**, are called *stochastic kernels* on \mathcal{X}_n given $\mathcal{X}_{0,n-1} \times \mathcal{Y}_{0,n-1}$, and these are denoted by

$$\mathcal{Q}(\mathcal{X}_n|\mathcal{X}_{0,n-1} \times \mathcal{Y}_{0,n-1}) \triangleq \{p_n(\cdot|x^{n-1}, y^{n-1}) \in \mathcal{M}_1(\mathcal{X}_n) : x^{n-1} \in \mathcal{X}_{0,n-1}, y^{n-1} \in \mathcal{Y}_{0,n-1} \text{ and ii) holds}\}.$$

Given the collection of functions $\{p_n(\cdot|\cdot, \cdot) : n \in \mathbb{N}_0\}$ satisfying conditions **i)**, **ii)**, one can construct a family of measures on the product space $(\mathcal{X}^{\mathbb{N}_0}, \mathcal{B}(\mathcal{X}^{\mathbb{N}_0})) \triangleq (\times_{i \in \mathbb{N}_0} \mathcal{X}_i, \otimes_{i \in \mathbb{N}_0} \mathcal{B}(\mathcal{X}_i))$ as follows.

Let $C \in \mathcal{B}(\mathcal{X}_{0,n})$ be a cylinder set of the form

$$C \triangleq \{\mathbf{x} \in \mathcal{X}^{\mathbb{N}_0} : x_0 \in C_0, x_1 \in C_1, \dots, x_n \in C_n\}, \quad C_i \in \mathcal{B}(\mathcal{X}_i), \quad i \in \mathbb{N}_0^n, \quad C_{0,n} = \times_{i=0}^n C_i.$$

Define a family of measures $\mathbf{P}(\cdot|\mathbf{y})$ parametrized by $\mathbf{y} \in \mathcal{Y}^{\mathbb{N}_0}$ on $\mathcal{B}(\mathcal{X}^{\mathbb{N}_0})$ by

$$\mathbf{P}(C|\mathbf{y}) \triangleq \int_{C_0} p_0(dx_0) \int_{C_1} p_1(dx_1|x_0, y_0) \dots \int_{C_n} p_n(dx_n|x^{n-1}, y^{n-1}) \quad (\text{II.1})$$

$$\equiv \overleftarrow{P}_{0,n}(C_{0,n}|y^{n-1}). \quad (\text{II.2})$$

The notation $\overleftarrow{P}_{0,n}(\cdot|y^{n-1})$ is used to denote the causal conditioning dependence of the measure $\mathbf{P}(\cdot|\mathbf{y})$ defined on cylinder sets $C \in \mathcal{B}(\mathcal{X}_{0,n})$, for any $n \in \mathbb{N}_0$. The right hand side (RHS) of (II.1) uniquely defines a measure on $(\mathcal{X}^{\mathbb{N}_0}, \mathcal{B}(\mathcal{X}^{\mathbb{N}_0}))$. Moreover, for each $n \in \mathbb{N}_0$ the family of measures $\mathbf{P}(\cdot|\mathbf{y})$ parametrized by $\mathbf{y} \in \mathcal{Y}^{\mathbb{N}_0}$, satisfies the following property (inherited from condition **ii)**): for $E \in \mathcal{B}(\mathcal{X}^{\mathbb{N}_0})$, $\mathbf{P}(E|\cdot)$ is $\mathcal{B}(\mathcal{Y}^{\mathbb{N}_0})$ -measurable, and for $E \in \mathcal{B}(\mathcal{X}_{0,n})$, $\mathbf{P}(E|\cdot)$ is $\mathcal{B}(\mathcal{Y}_{0,n-1})$ -measurable.

Thus, if conditions **i)** and **ii)** hold then for each $\mathbf{y} \in \mathcal{Y}^{\mathbb{N}_0}$, the RHS of (II.1) defines a consistent family

of finite-dimensional distribution, and hence there exists a unique measure on $(\mathcal{X}^{\mathbb{N}_0}, \mathcal{B}(\mathcal{X}^{\mathbb{N}_0}))$, for which $p_n(dx_n|x^{n-1}, y^{n-1})$ is obtained. This leads to the first definition of a feedback channel, as a family of functions $\{p_n(\cdot|\cdot, \cdot) \in \mathcal{Q}(\mathcal{X}_n|\mathcal{X}_{0,n-1} \times \mathcal{Y}_{0,n-1}) : n \in \mathbb{N}_0\}$, i.e., satisfying conditions **i)** and **ii)**. This definition is used extensively by many authors [6], [7], [10]–[13], when the alphabet spaces have finite cardinality.

An alternative, equivalent definition of a feedback channel is established as follows. Consider a family of measures $\mathbf{P}(\cdot|\mathbf{y})$ on $(\mathcal{X}^{\mathbb{N}_0}, \mathcal{B}(\mathcal{X}^{\mathbb{N}_0}))$ parametrized by $\mathbf{y} \in \mathcal{Y}^{\mathbb{N}_0}$ satisfying the following consistency condition.

C1: If $E \in \mathcal{B}(\mathcal{X}_{0,n})$ then $\mathbf{P}(E_{0,n}|\cdot)$ is $\mathcal{B}(\mathcal{Y}_{0,n-1})$ –measurable function of $\mathbf{y} \in \mathcal{Y}^{\mathbb{N}_0}$.

Clearly, if conditions **i)** and **ii)** are satisfied, then the family of measures $\mathbf{P}(\cdot|\mathbf{y})$ defined via the RHS of (II.1) satisfies consistency condition **C1**. The question we address next is whether for any family of measures $\mathbf{P}(\cdot|\mathbf{y})$ on $(\mathcal{X}^{\mathbb{N}_0}, \mathcal{B}(\mathcal{X}^{\mathbb{N}_0}))$ parametrized by $\mathbf{y} \in \mathcal{Y}^{\mathbb{N}_0}$, satisfying consistency condition **C1**, one can construct a collection of functions $\{p_n(\cdot|\cdot, \cdot) \in \mathcal{Q}(\mathcal{X}_n|\mathcal{X}_{0,n-1} \times \mathcal{Y}_{0,n-1}) : n \in \mathbb{N}_0\}$, i.e., satisfying conditions **i)** and **ii)**, which are connected to $\mathbf{P}(\cdot|\mathbf{y})$ via relation (II.1). To illustrate this point, let $A^{(n)} = \{\mathbf{x} \in \mathcal{X}^{\mathbb{N}_0} : x_n \in A\}$, $A \in \mathcal{B}(\mathcal{X}_n)$, and let $\mathbf{P}(A^{(n)}|\mathcal{B}(\mathcal{X}_{0,n-1})|\mathbf{y})$ denote the conditional probability of $A^{(n)}$ with respect to $\mathcal{B}(\mathcal{X}_{0,n-1})$ calculated on the probability space $(\mathcal{X}^{\mathbb{N}_0}, \mathcal{B}(\mathcal{X}^{\mathbb{N}_0}), \mathbf{P}(\cdot|\mathbf{y}))$. Then

$$\mathbf{P}(A^{(n)}|\mathcal{B}(\mathcal{X}_{0,n-1})|\mathbf{y}) = p_n(A|x^{n-1}, y^{n-1}), \quad A^{(n)} \in \mathcal{B}(\mathcal{X}_{0,n}), \quad (\text{II.3})$$

for $\mathbf{P}(\cdot|\mathbf{y})$ –almost all $\mathbf{x} \in \mathcal{X}^{\mathbb{N}_0}$. Clearly, the function on the RHS of (II.3), $p_n(A|x^{n-1}, y^{n-1})$ is $\mathcal{B}(\mathcal{X}_{0,n-1})$ –measurable for a fixed $A \in \mathcal{B}(\mathcal{X}_n)$ and $y^{n-1} \in \mathcal{Y}_{0,n-1}$, but it cannot be claimed that $p_n(\cdot|x^{n-1}, y^{n-1})$ is a probability measure on \mathcal{X}_n . However, under the general assumption that $\{(\mathcal{X}_n, \mathcal{B}(\mathcal{X}_n)) : n \in \mathbb{N}_0\}$ are complete separable metric spaces (Polish spaces), with $\mathcal{B}(\mathcal{X}_n)$ the σ –algebra of Borel sets, it is shown in [32], that the RHS of (II.3) represents a version of conditional probability (*a.s.*), i.e., condition **i)** holds as well. Therefore, to establish the second equivalent definition of a family of measures defined by (II.1) with elements $\{p_n(\cdot|\cdot, \cdot) \in \mathcal{Q}(\mathcal{X}_n|\mathcal{X}_{0,n-1} \times \mathcal{Y}_{0,n-1}) : n \in \mathbb{N}_0\}$, we introduce the following condition on the alphabet spaces.

iii) $\{\mathcal{X}_n : n \in \mathbb{N}_0\}$ are complete separable metric spaces and $\{\mathcal{B}(\mathcal{X}_n) : n \in \mathbb{N}_0\}$ are the σ –algebras of Borel sets.

By [32], if condition **iii)** holds, then for any family of measures $\mathbf{P}(\cdot|\mathbf{y})$ parametrized by $\mathbf{y} \in \mathcal{Y}^{\mathbb{N}_0}$ satisfying **C1** one can construct a collection of versions of conditional distributions $\{p_n(dx_n|x^{n-1}, y^{n-1}) : n \in \mathbb{N}_0\}$ satisfying conditions **i)** and **ii)** which are connected with $\mathbf{P}(\cdot|\mathbf{y})$ via relation (II.1), and hence the following conclusion.

When $\{\mathcal{X}_n : n \in \mathbb{N}_0\}$ are Polish Spaces with $\{\mathcal{B}(\mathcal{X}_n) : n \in \mathbb{N}_0\}$ the σ –algebra of Borel sets, there are two equivalent definitions of a feedback channel. The first definition is the usual one given by a collection of functions $\{p_n(\cdot|\cdot, \cdot) \in \mathcal{Q}(\mathcal{X}_n|\mathcal{X}_{0,n-1} \times \mathcal{Y}_{0,n-1}) : n \in \mathbb{N}_0\}$, i.e., satisfying conditions **i)** and **ii)**. The second definition is given by a family of measures $\mathbf{P}(\cdot|\mathbf{y})$ on $(\mathcal{X}^{\mathbb{N}_0}, \mathcal{B}(\mathcal{X}^{\mathbb{N}_0}))$ depending parametrically on $\mathbf{y} \in \mathcal{Y}^{\mathbb{N}_0}$ and satisfying the consistency condition **C1**.

The second equivalent definition of a feedback channel, together with an analogous equivalent definition for the forward channel will be used throughout the paper.

Feedforward Channel.

The above methodology is repeated to obtain two equivalent definitions for the forward channel as well. Suppose for each $n \in \mathbb{N}_0$, the conditional distribution of the RV $Y_n \in \mathcal{Y}_n$ is determined provided the values of the basic processes $Y^{n-1} \in \mathcal{Y}_{0,n-1}$ and $X^n = x^n \in \mathcal{X}_{0,n}$ are known, and let $\{q_n(dy_n|y^{n-1}, x^n) : n \in \mathbb{N}_0\}$ denotes this collection of distributions. At $n = 0$, $q_0(dy_0|y_{-1}, x_0)$, where y_{-1} is either fixed or its distribution is fixed (depending on the convection used). Without loss of generality, we assume $q_0(dy_0|y_{-1}, x_0) \triangleq q_0(dy_0|x_0)$. The functions $\{q_n(\cdot|\cdot, \cdot) : n \in \mathbb{N}_0\}$ satisfy the following conditions.

iv) For every $n \in \mathbb{N}_0$, and $y^{n-1} \in \mathcal{Y}_{0,n-1}$, $x^n \in \mathcal{X}_{0,n}$, $q_n(\cdot|y^{n-1}, x^n)$ is a probability measure $\mathcal{B}(\mathcal{Y}_n)$;

v) For every $n \in \mathbb{N}_0$, and $B_n \in \mathcal{B}(\mathcal{Y}_n)$, $q_n(B_n|\cdot, \cdot)$ is an $\otimes_{i=0}^{n-1}(\mathcal{B}(\mathcal{Y}_i) \otimes \mathcal{B}(\mathcal{X}_i)) \otimes \mathcal{B}(\mathcal{X}_n)$ -measurable function of $x^n \in \mathcal{X}_{0,n}$, $y^{n-1} \in \mathcal{Y}_{0,n-1}$.

For every $n \in \mathbb{N}_0$, the set of all functions that satisfy **iv)**, **v)**, are called stochastic kernels on \mathcal{Y}_n given $\mathcal{Y}_{0,n-1} \times \mathcal{X}_{0,n}$, and these are denoted by

$$\mathcal{Q}(\mathcal{Y}_n|\mathcal{Y}_{0,n-1} \times \mathcal{X}_{0,n}) = \{q_n(\cdot|y^{n-1}, x^n) \in \mathcal{M}_1(\mathcal{Y}_n) : y^{n-1} \in \mathcal{Y}_{0,n-1}, x^n \in \mathcal{X}_{0,n} \text{ and } \mathbf{v}) \text{ holds}\}.$$

Similarly as before, using the collection of functions $\{q_n(\cdot|\cdot, \cdot) \in \mathcal{Q}(\mathcal{Y}_n|\mathcal{Y}_{0,n-1} \times \mathcal{X}_{0,n}) : n \in \mathbb{N}_0\}$ one can construct a family of measures $\mathbf{Q}(\cdot|\mathbf{x})$ on $(\mathcal{Y}^{\mathbb{N}_0}, \mathcal{B}(\mathcal{Y}^{\mathbb{N}_0}))$ which depend parametrically on $\mathbf{x} \in \mathcal{X}^{\mathbb{N}_0}$, as follows.

Consider a cylinder set $D \in \mathcal{B}(\mathcal{Y}_{0,n})$ of the form

$$D \triangleq \left\{ \mathbf{y} \in \mathcal{Y}^{\mathbb{N}_0} : y_0 \in D_0, y_1 \in D_1, \dots, y_n \in D_n \right\}, \quad D_i \in \mathcal{B}(\mathcal{Y}_i), \quad n \in \mathbb{N}_0, \quad D_{0,n} = \times_{i=0}^n D_i.$$

Define a family of measures on $\mathcal{B}(\mathcal{Y}^{\mathbb{N}_0})$ parametrized by $\mathbf{x} \in \mathcal{X}^{\mathbb{N}_0}$ by

$$\mathbf{Q}(D|\mathbf{x}) \triangleq \int_{D_0} q_0(dy_0|x_0) \int_{D_1} q_1(dy_1|y_0, x^1) \dots \int_{D_n} q_n(dy_n|y^{n-1}, x^n) \quad (\text{II.4})$$

$$\equiv \vec{Q}_{0,n}(D_{0,n}|\mathbf{x}). \quad (\text{II.5})$$

Since, for each $\mathbf{x} \in \mathcal{X}^{\mathbb{N}_0}$ the RHS of (II.4) defines a consistent family of finite dimensional distribution, then there exist a unique measure on $(\mathcal{Y}^{\mathbb{N}_0}, \mathcal{B}(\mathcal{Y}^{\mathbb{N}_0}))$ from which the family of distributions $\{q_n(dy_n|y^{n-1}, x^n) : n \in \mathbb{N}_0\}$ satisfying **iv)**, **v)** can be obtained. Moreover, the family of measures $\mathbf{Q}(\cdot|\mathbf{x})$ parametrized by $\mathbf{x} \in \mathcal{X}^{\mathbb{N}_0}$ satisfies the following consistency condition.

C2: If $F \in \mathcal{B}(\mathcal{Y}_{0,n})$, then $\mathbf{Q}(F|\cdot)$ is a $\mathcal{B}(\mathcal{X}_{0,n})$ -measurable function of $\mathbf{x} \in \mathcal{X}^{\mathbb{N}_0}$.

By [32], to obtain another equivalent definition for the forward channel introduce the following condition on the output alphabet.

vi) $\{\mathcal{Y}_n : n \in \mathbb{N}_0\}$ are Polish Spaces and $\{\mathcal{B}(\mathcal{Y}_n) : n \in \mathbb{N}_0\}$ are the σ -algebra of Borel sets.

If condition **vi)** holds, then for any family of measures $\mathbf{Q}(\cdot|\mathbf{x})$ on $(\mathcal{Y}^{\mathbb{N}_0}, \mathcal{B}(\mathcal{Y}^{\mathbb{N}_0}))$ parametrized by $\mathbf{x} \in \mathcal{X}^{\mathbb{N}_0}$ satisfying consistency condition **C2**, one can construct a collection of functions $\{q_n(\cdot|\cdot, \cdot) \in \mathcal{Q}(\mathcal{Y}_n|\mathcal{Y}_{0,n-1} \times \mathcal{X}_{0,n}) : n \in \mathbb{N}_0\}$, i.e., satisfying conditions **iv)** and **v)**, which are connected with $\mathbf{Q}(\cdot|\mathbf{x})$ via relation (II.4). Therefore, we arrive at two equivalent definitions for the forward channel as well.

We conclude this section by constructing the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, as stated earlier, and the sequence of RV's $\{(X_n, Y_n) : n \in \mathbb{N}_0\}$ defined on it. Given the basic measures $\mathbf{P}(\cdot|\mathbf{y})$ on $\mathcal{X}^{\mathbb{N}_0}$ satisfying consistency condition **C1** and $\mathbf{Q}(\cdot|\mathbf{x})$ on $\mathcal{Y}^{\mathbb{N}_0}$ satisfying consistency condition **C2**, one can construct a sequence of RV's $\{X_n, Y_n : n \in \mathbb{N}_0\}$ or conditional distributions as follows.

Suppose **iii)**, **iv)** hold. Let $A^{(n)} = \{\mathbf{x} : x_n \in A\}$, $A \in \mathcal{B}(\mathcal{X}_n)$ and $B^{(n)} = \{\mathbf{y} : y_n \in B\}$, $B \in \mathcal{B}(\mathcal{Y}_n)$. In addition, let $\mathbf{P}(A^{(n)}|\mathcal{B}(\mathcal{X}_{0,n-1})|\mathbf{y})$ denote the conditional probability of $A^{(n)}$ with respect to $\mathcal{B}(\mathcal{X}_{0,n-1})$ calculated on the probability space $(\mathcal{X}^{\mathbb{N}_0}, \mathcal{B}(\mathcal{X}^{\mathbb{N}_0}), \mathbf{P}(\cdot|\mathbf{y}))$, and $\mathbf{Q}(B^{(n)}|\mathcal{B}(\mathcal{Y}_{0,n-1})|\mathbf{x})$ denote the conditional probability of $B^{(n)}$ with respect to $\mathcal{B}(\mathcal{Y}_{0,n-1})$ calculated on the probability space $(\mathcal{Y}^{\mathbb{N}_0}, \mathcal{B}(\mathcal{Y}^{\mathbb{N}_0}), \mathbf{Q}(\cdot|\mathbf{x}))$. Then for each $n \in \mathbb{N}_0$, by conditioning it follows that

$$\begin{aligned} \mathbb{P}\{X_n \in A | X^{n-1} = x^{n-1}, Y^{n-1} = y^{n-1}\} &= \mathbf{P}(\{\mathbf{x} : x_n \in A\} | \mathcal{B}(\mathcal{X}_{0,n-1}) | \mathbf{y}), \quad A \in \mathcal{B}(\mathcal{X}_n) \\ &= p_n(A | x^{n-1}, y^{n-1}) \end{aligned} \quad (\text{II.6})$$

$$\begin{aligned} \mathbb{P}\{Y_n \in B | Y^{n-1} = y^{n-1}, X^n = x^n\} &= \mathbf{Q}(\{\mathbf{y} : y_n \in B\} | \mathcal{B}(\mathcal{Y}_{0,n-1}) | \mathbf{x}), \quad B \in \mathcal{B}(\mathcal{Y}_n) \\ &= q_n(B | y^{n-1}, x^n) \end{aligned} \quad (\text{II.7})$$

for almost all $\mathbf{x} \in \mathcal{X}^{\mathbb{N}_0}$ in measure $\mathbf{P}(\cdot|\mathbf{y})$, and for almost all $\mathbf{y} \in \mathcal{Y}^{\mathbb{N}_0}$ in measure $\mathbf{Q}(\cdot|\mathbf{x})$. Note that for each $n \in \mathbb{N}_0$, $p_n(\cdot; \cdot, \cdot) \in \mathcal{Q}(\mathcal{X}_n|\mathcal{X}_{0,n-1}, \mathcal{Y}_{0,n-1})$ and $q_n(\cdot|\cdot, \cdot) \in \mathcal{Q}(\mathcal{Y}_n|\mathcal{Y}_{0,n-1}, \mathcal{X}_n)$ are stochastic kernels

determined from $\mathbf{P}(\cdot|\cdot)$ and $\mathbf{Q}(\cdot|\cdot)$, respectively, (e.g., they are related via (II.1) and (II.4), respectively). Consequently, the finite dimensional distributions of the sequence of RV's $\{(X_n, Y_n) : n \in \mathbb{N}_0\}$ is defined by

$$\mathbb{P}\{X_0 \in A_0, Y_0 \in B_0, \dots, X_n \in A_n, Y_n \in B_n\} = \int_{A_0} p_0(dx_0) \int_{B_0} q_0(dy_0|x_0) \dots \int_{A_n} p_n(dx_n|x^{n-1}, y^{n-1}) \int_{B_n} q_n(dy_n|y^{n-1}, x^n). \quad (\text{II.8})$$

Hence, given the two Polish spaces $\mathcal{X}^{\mathbb{N}_0}$ and $\mathcal{Y}^{\mathbb{N}_0}$, for any $\mathbf{P}(\cdot|\cdot)$ and $\mathbf{Q}(\cdot|\cdot)$ satisfying the consistency conditions **C1**, **C2**, respectively, there exist a probability space and a sequence of RV's $\{(X_n, Y_n) : n \in \mathbb{N}_0\}$ defined on it, whose joint probability distribution is uniquely defined by (II.8), via $\mathbf{P}(\cdot|\cdot)$ and $\mathbf{Q}(\cdot|\cdot)$.

The following remark summarizes the previous discussion on the two equivalent definitions of forward and feedback channels.

Remark II.1.

Suppose $\{\mathcal{X}_n : n \in \mathbb{N}_0\}$, $\{\mathcal{Y}_n : n \in \mathbb{N}_0\}$, are complete, separable metric spaces (Polish spaces) and $\{\mathcal{B}(\mathcal{X}_n) : n \in \mathbb{N}_0\}$, $\{\mathcal{B}(\mathcal{Y}_n) : n \in \mathbb{N}_0\}$ are respectively, the σ -algebras of Borel sets.

Then

- 1) *The collection of stochastic kernels $\{p_n(\cdot|\cdot, \cdot) \in \mathcal{Q}(\mathcal{X}_n|\mathcal{X}_{0,n-1} \times \mathcal{Y}_{0,n-1}) : n \in \mathbb{N}_0\}$ uniquely define a family of probability measures on $(\mathcal{X}^{\mathbb{N}_0}, \mathcal{B}(\mathcal{X}^{\mathbb{N}_0}))$ parametrized by $\mathbf{y} \in \mathcal{Y}^{\mathbb{N}_0}$ via (II.1).*
- 2) *For any family of probability measures $\mathbf{P}(\cdot|\mathbf{y})$ on $(\mathcal{X}^{\mathbb{N}_0}, \mathcal{B}(\mathcal{X}^{\mathbb{N}_0}))$ parametrized by $\mathbf{y} \in \mathcal{Y}^{\mathbb{N}_0}$, satisfying consistency condition **C1** there exists a collection of stochastic kernels $\{p_n(\cdot|\cdot, \cdot) \in \mathcal{Q}(\mathcal{X}_n|\mathcal{X}_{0,n-1} \times \mathcal{Y}_{0,n-1}) : n \in \mathbb{N}_0\}$ connected to $\mathbf{P}(\cdot|\cdot)$ via (II.1).*
- 3) *The collection of stochastic kernels $\{q_n(\cdot|\cdot, \cdot) \in \mathcal{Q}(\mathcal{Y}_n|\mathcal{Y}_{0,n-1} \times \mathcal{X}_{0,n}) : n \in \mathbb{N}_0\}$ uniquely define a family of probability measures on $(\mathcal{Y}^{\mathbb{N}_0}, \mathcal{B}(\mathcal{Y}^{\mathbb{N}_0}))$ parametrized by $\mathbf{x} \in \mathcal{X}^{\mathbb{N}_0}$ via (II.4).*
- 4) *For any family of probability measures $\mathbf{Q}(\cdot|\mathbf{x})$ on $(\mathcal{Y}^{\mathbb{N}_0}, \mathcal{B}(\mathcal{Y}^{\mathbb{N}_0}))$ parametrized by $\mathbf{x} \in \mathcal{X}^{\mathbb{N}_0}$ satisfying consistency condition **C2** there exists a collection of stochastic kernels $\{q_n(\cdot|\cdot, \cdot) \in \mathcal{Q}(\mathcal{Y}_n|\mathcal{Y}_{0,n-1} \times \mathcal{X}_{0,n}) : n \in \mathbb{N}_0\}$ connected to $\mathbf{Q}(\cdot|\cdot)$ via (II.4).*

The point to be made here is that directed information as defined by (I.1)-(I.3) can be expressed via the equivalent definitions of Remark II.1, 2) and 4) rather than 1) and 3). We use this equivalent definition of directed information, to derive the functional and topological properties of directed information on general abstract spaces. Throughout the rest of the paper it is assumed that the conditions of Remark II.1 are satisfied, i.e., all spaces are Polish spaces.

III. PROPERTIES OF DIRECTED INFORMATION

In this section, we define the feedforward information $I(X^n \rightarrow Y^n)$ on abstract spaces (Polish spaces), via the Kullback-Leibler distance (or relative entropy), using the basic family of measures $\mathbf{P}(\cdot|\mathbf{y})$ on $(\mathcal{X}^{\mathbb{N}_0}, \mathcal{B}(\mathcal{X}^{\mathbb{N}_0}))$, and $\mathbf{Q}(\cdot|\mathbf{x})$ on $(\mathcal{Y}^{\mathbb{N}_0}, \mathcal{B}(\mathcal{Y}^{\mathbb{N}_0}))$, which satisfy consistency condition **C1** and **C2**, respectively. Once this is established, then following Pinsker [44], it will become obvious that directed information permits a representation as a supremum of relative entropy between two distributions, where the supremum is taken over all measurable partitions on a given σ -algebra of subsets of a set \mathcal{Z} . Further, in a subsequent subsection, we use the definition of directed information in terms of $\mathbf{P}(\cdot|\mathbf{y})$ and $\mathbf{Q}(\cdot|\mathbf{x})$, to derive several of its properties, such as, convexity, concavity, lower semicontinuity, with respect to these two families of measures.

To present the precise expression for the directed information, we first introduce the measures of interest constructed from the basic consistent families of conditional distributions. Introduce the following notation.

The set of stochastic kernels by

$$\begin{aligned}\mathcal{Q}^{\mathbf{C1}}(\mathcal{X}^{\mathbb{N}_0}|\mathcal{Y}^{\mathbb{N}_0}) &\triangleq \left\{ \mathbf{P}(\cdot|\mathbf{y}) \in \mathcal{M}_1(\mathcal{X}^{\mathbb{N}_0}) : \mathbf{y} \in \mathcal{Y}^{\mathbb{N}_0} \text{ and consistency condition } \mathbf{C1} \text{ holds} \right\} \\ &\equiv \left\{ \mathbf{P}(\cdot|\cdot) \in \mathcal{Q}(\mathcal{X}^{\mathbb{N}_0}|\mathcal{Y}^{\mathbb{N}_0}) : \text{consistency condition } \mathbf{C1} \text{ holds} \right\}.\end{aligned}\quad (\text{III.1})$$

Note that for each $\mathbf{y} \in \mathcal{Y}^{\mathbb{N}_0}$, elements of this set are probability distributions on $\mathcal{X}^{\mathbb{N}_0}$ denoted by

$$\mathcal{M}_1^{\mathbf{C1}}(\mathcal{X}^{\mathbb{N}_0}) \triangleq \left\{ \mathbf{P}(\cdot|\mathbf{y}) \in \mathcal{M}_1(\mathcal{X}^{\mathbb{N}_0}) : \text{consistency condition } \mathbf{C1} \text{ holds} \right\} \quad (\text{III.2})$$

Similarly,

$$\begin{aligned}\mathcal{Q}^{\mathbf{C2}}(\mathcal{Y}^{\mathbb{N}_0}|\mathcal{X}^{\mathbb{N}_0}) &\triangleq \left\{ \mathbf{Q}(\cdot|\mathbf{x}) \in \mathcal{M}_1(\mathcal{Y}^{\mathbb{N}_0}) : \mathbf{x} \in \mathcal{X}^{\mathbb{N}_0} \text{ and consistency condition } \mathbf{C2} \text{ holds} \right\} \\ &\equiv \left\{ \mathbf{Q}(\cdot|\cdot) \in \mathcal{Q}(\mathcal{Y}^{\mathbb{N}_0}|\mathcal{X}^{\mathbb{N}_0}) : \text{consistency condition } \mathbf{C2} \text{ holds} \right\}.\end{aligned}\quad (\text{III.3})$$

and for each $\mathbf{x} \in \mathcal{X}^{\mathbb{N}_0}$, elements of this set are probability distributions on $\mathcal{Y}^{\mathbb{N}_0}$, denoted by

$$\mathcal{M}_1^{\mathbf{C2}}(\mathcal{Y}^{\mathbb{N}_0}) \triangleq \left\{ \mathbf{Q}(\cdot|\mathbf{x}) \in \mathcal{M}_1(\mathcal{Y}^{\mathbb{N}_0}) : \text{consistency condition } \mathbf{C2} \text{ holds} \right\} \quad (\text{III.4})$$

The projection of $\mathcal{M}_1^{\mathbf{C1}}(\mathcal{X}^{\mathbb{N}_0})$, $\mathcal{M}_1^{\mathbf{C2}}(\mathcal{Y}^{\mathbb{N}_0})$, $\mathcal{Q}^{\mathbf{C1}}(\mathcal{X}^{\mathbb{N}_0}|\mathcal{Y}^{\mathbb{N}_0})$, and $\mathcal{Q}^{\mathbf{C2}}(\mathcal{Y}^{\mathbb{N}_0}|\mathcal{X}^{\mathbb{N}_0})$ to finite number of coordinates is denoted by $\mathcal{M}_1^{\mathbf{C1}}(\mathcal{X}_{0,n})$, $\mathcal{M}_1^{\mathbf{C2}}(\mathcal{Y}_{0,n})$, $\mathcal{Q}^{\mathbf{C1}}(\mathcal{X}_{0,n}|\mathcal{Y}_{0,n-1})$, and $\mathcal{Q}^{\mathbf{C2}}(\mathcal{Y}_{0,n}|\mathcal{X}_{0,n})$, respectively. Since the spaces are complete separable metric spaces then $\mathbf{P}(\cdot|\mathbf{y}) \in \mathcal{M}_1(\mathcal{X}^{\mathbb{N}_0})$, for fixed $\mathbf{y} \in \mathcal{Y}^{\mathbb{N}_0}$, and $\mathbf{Q}(\cdot|\mathbf{x}) \in \mathcal{M}_1(\mathcal{Y}^{\mathbb{N}_0})$, for fixed $\mathbf{x} \in \mathcal{X}^{\mathbb{N}_0}$, are regular conditional probability distributions [30].

Next, we define the distributions of interest. Given any $\mathbf{P}(\cdot|\cdot) \in \mathcal{Q}^{\mathbf{C1}}(\mathcal{X}^{\mathbb{N}_0}|\mathcal{Y}^{\mathbb{N}_0})$ and $\mathbf{Q}(\cdot|\cdot) \in \mathcal{Q}^{\mathbf{C2}}(\mathcal{Y}^{\mathbb{N}_0}|\mathcal{X}^{\mathbb{N}_0})$, by utilizing the construction of Section II, we can define uniquely $\{p_n(\cdot|\cdot, \cdot) : n \in \mathbb{N}_0\}$ and $\{q_n(\cdot|\cdot, \cdot) : n \in \mathbb{N}_0\}$, (see (II.6), (II.7)) and the following distributions.

P1: The joint distribution on $\mathcal{X}^{\mathbb{N}_0} \times \mathcal{Y}^{\mathbb{N}_0}$ of the basic sequence $\{X_n, Y_n : n \in \mathbb{N}_0\}$ constructed from $\mathbf{P}(\cdot|\cdot) \in \mathcal{Q}^{\mathbf{C1}}(\mathcal{X}^{\mathbb{N}_0}|\mathcal{Y}^{\mathbb{N}_0})$ and $\mathbf{Q}(\cdot|\cdot) \in \mathcal{Q}^{\mathbf{C2}}(\mathcal{Y}^{\mathbb{N}_0}|\mathcal{X}^{\mathbb{N}_0})$, defined uniquely for $A_i \in \mathcal{B}(\mathcal{X}_i)$, $B_i \in \mathcal{B}(\mathcal{Y}_i)$, $\forall i \in \mathbb{N}_0^n$, by

$$\begin{aligned}(\overleftarrow{P}_{0,n} \otimes \overrightarrow{Q}_{0,n})(\times_{i=0}^n (A_i \times B_i)) &\triangleq \mathbb{P}\{X_0 \in A_0, Y_0 \in B_0, \dots, X_n \in A_n, Y_n \in B_n\} \\ &= \int_{A_0} p_0(dx_0) \int_{B_0} q_0(dy_0|x_0) \dots \int_{A_n} p_n(dx_n|x^{n-1}, y^{n-1}) \int_{B_n} q_n(dy_n|y^{n-1}, x^n).\end{aligned}\quad (\text{III.5})$$

Formally, the $(n+1)$ fold compound joint distribution defined by (III.5) is written as $(\overleftarrow{P}_{0,n} \otimes \overrightarrow{Q}_{0,n})(dx^n, dy^n)$ or $\overleftarrow{P}_{0,n}(dx^n|y^{n-1}) \otimes \overrightarrow{Q}_{0,n}(dy^n|x^n)$.

P2: The marginal distributions on $\mathcal{X}^{\mathbb{N}_0}$ of the sequence $\{X_n : n \in \mathbb{N}_0\}$ constructed from $\mathbf{P}(\cdot|\cdot) \in \mathcal{Q}^{\mathbf{C1}}(\mathcal{X}^{\mathbb{N}_0}|\mathcal{Y}^{\mathbb{N}_0})$ and $\mathbf{Q}(\cdot|\cdot) \in \mathcal{Q}^{\mathbf{C2}}(\mathcal{Y}^{\mathbb{N}_0}|\mathcal{X}^{\mathbb{N}_0})$, defined uniquely by²

$$\mu_{0,n}(\times_{i=0}^n A_i) \triangleq \mathbb{P}\{X_0 \in A_0, Y_0 \in \mathcal{Y}_0, \dots, X_n \in A_n, Y_n \in \mathcal{Y}_n\}, \quad A_i \in \mathcal{B}(\mathcal{X}_i), \quad \forall i \in \mathbb{N}_0^n \quad (\text{III.6})$$

$$\begin{aligned}&= (\overleftarrow{P}_{0,n} \otimes \overrightarrow{Q}_{0,n})(\times_{i=0}^n (A_i \times \mathcal{Y}_i)) \\ &= \int_{A_0} p_0(dx_0) \int_{\mathcal{Y}_0} q_0(dy_0|x_0) \dots \int_{A_n} p_n(dx_n|x^{n-1}, y^{n-1}) \int_{\mathcal{Y}_n} q_n(dy_n|y^{n-1}, x^n).\end{aligned}\quad (\text{III.7})$$

²Actually $\mu \equiv \mu^{\mathbf{P} \otimes \mathbf{Q}}$ but we omit the superscript throughout the paper.

Formally, (III.7) is written as $\mu_{0,n}(dx^n) = (\overleftarrow{P}_{0,n} \otimes \overrightarrow{Q}_{0,n})(dx^n, \mathcal{Y}_{0,n})$, and by Bayes' rule $\mu_{0,n}(dx^n) = \otimes_{i=0}^n \mu_i(dx_i|x^{i-1})$.

P3: The marginal distributions on $\mathcal{Y}^{\mathbb{N}_0}$ of the sequence $\{Y_n : n \in \mathbb{N}_0\}$ constructed from $\mathbf{P}(\cdot|\cdot) \in \mathcal{Q}^{\mathbf{C1}}(\mathcal{X}^{\mathbb{N}_0}|\mathcal{Y}^{\mathbb{N}_0})$ and $\mathbf{Q}(\cdot|\cdot) \in \mathcal{Q}^{\mathbf{C2}}(\mathcal{Y}^{\mathbb{N}_0}|\mathcal{X}^{\mathbb{N}_0})$, defined uniquely by³

$$\nu_{0,n}(\times_{i=0}^n B_i) \triangleq \mathbb{P}\left\{X_0 \in \mathcal{X}_0, Y_0 \in B_0, \dots, X_n \in \mathcal{X}_n, Y_n \in B_n\right\}, \quad B_i \in \mathcal{B}(\mathcal{Y}_i), \quad \forall i \in \mathbb{N}_0^n \quad (\text{III.8})$$

$$\begin{aligned} &= (\overleftarrow{P}_{0,n} \otimes \overrightarrow{Q}_{0,n})(\times_{i=0}^n (\mathcal{X}_i \times B_i)) \\ &= \int_{\mathcal{X}_0} p_0(dx_0) \int_{B_0} q_0(dy_0|x_0) \dots \int_{\mathcal{X}_n} p_n(dx_n|x^{n-1}, y^{n-1}) \int_{B_n} q_n(dy_n|y^{n-1}, x^n). \end{aligned} \quad (\text{III.9})$$

Formally, (III.9) is written as $\nu_{0,n}(dy^n) = (\overleftarrow{P}_{0,n} \otimes \overrightarrow{Q}_{0,n})(\mathcal{X}_{0,n}, dy^n)$, and by Bayes' rule $\nu_{0,n}(dy^n) = \otimes_{i=0}^n \nu_i(dy_i|y^{i-1})$.

P4: The distribution $\overrightarrow{\Pi}_{0,n} : \mathcal{B}(\mathcal{X}_{0,n}) \otimes \mathcal{B}(\mathcal{Y}_{0,n}) \mapsto [0, 1]$ constructed from $\overleftarrow{P}_{0,n}(\cdot|\cdot) \in \mathcal{Q}^{\mathbf{C1}}(\mathcal{X}_{0,n}|\mathcal{Y}_{0,n-1})$ and $\nu_{0,n}(dy^n) = (\overleftarrow{P}_{0,n} \otimes \overrightarrow{Q}_{0,n})(\mathcal{X}_{0,n}, dy^n) \in \mathcal{M}_1(\mathcal{Y}_{0,n})$ of (III.8), defined uniquely by

$$\begin{aligned} \overrightarrow{\Pi}_{0,n}(\times_{i=0}^n (A_i \times B_i)) &\triangleq (\overleftarrow{P}_{0,n} \otimes \nu_{0,n})(\times_{i=0}^n (A_i \times B_i)), \quad A_i \in \mathcal{B}(\mathcal{X}_i), \quad B_i \in \mathcal{B}(\mathcal{Y}_i), \quad \forall i \in \mathbb{N}_0^n \\ &= \int_{A_0} p_0(dx_0) \int_{B_0} \nu_0(dy_0) \int_{A_1} p_1(dx_1|x_0, y_0) \int_{B_1} \nu_1(dy_1|y_0) \dots \\ &\dots \int_{A_n} p_n(dx_n|x^{n-1}, y^{n-1}) \int_{B_n} \nu_n(dy_n|y^{n-1}). \end{aligned} \quad (\text{III.10})$$

Formally, (III.10) is written as $\overrightarrow{\Pi}_{0,n}(dx^n, dy^n) = \overleftarrow{P}_{0,n}(dx^n|y^{n-1}) \otimes \nu_{0,n}(dy^n) \in \mathcal{M}_1(\mathcal{X}_{0,n} \times \mathcal{Y}_{0,n})$.

P5: The distribution $\overleftarrow{\Pi}_{0,n} : \mathcal{B}(\mathcal{Y}_{0,n}) \otimes \mathcal{B}(\mathcal{X}_{0,n}) \mapsto [0, 1]$ constructed from $\overrightarrow{Q}_{0,n}(\cdot|\cdot) \in \mathcal{Q}^{\mathbf{C2}}(\mathcal{Y}_{0,n}|\mathcal{X}_{0,n})$ and $\mu_{0,n}(dx^n) = (\overleftarrow{P}_{0,n} \otimes \overrightarrow{Q}_{0,n})(dx^n, \mathcal{Y}_{0,n}) \in \mathcal{M}_1(\mathcal{X}_{0,n})$ of (III.7), defined uniquely by

$$\begin{aligned} \overleftarrow{\Pi}_{0,n}(\times_{i=0}^n (A_i \times B_i)) &\triangleq (\mu_{0,n} \otimes \overrightarrow{Q}_{0,n})(\times_{i=0}^n (A_i \times B_i)), \quad A_i \in \mathcal{B}(\mathcal{X}_i), \quad B_i \in \mathcal{B}(\mathcal{Y}_i), \quad \forall i \in \mathbb{N}_0^n \\ &= \int_{A_0} \mu_0(dx_0) \int_{B_0} q_0(dy_0|x_0) \int_{A_1} \mu_1(dx_1|x_0) \int_{B_1} q_1(dy_1|y_0, x_0) \dots \\ &\dots \int_{A_n} \mu_n(dx_n|x^{n-1}) \int_{B_n} q_n(dy_n|y^{n-1}, x^n). \end{aligned} \quad (\text{III.11})$$

Formally, (III.11) is written as $\overleftarrow{\Pi}_{0,n}(dx^n, dy^n) = \mu_{0,n}(dx^n) \otimes \overrightarrow{Q}_{0,n}(dy^n|x^n) \in \mathcal{M}_1(\mathcal{X}_{0,n} \times \mathcal{Y}_{0,n})$.

From the above definitions, for each $n \in \mathbb{N}_0$, an alternative way to construct the conditional distributions of Y_n given $Y^{n-1} = y^{n-1}$, $\nu_n(\cdot|y^{n-1}) \in \mathcal{M}_1(\mathcal{Y}_n)$, and X_n given $X^{n-1} = x^{n-1}$, $\mu_n(\cdot|x^{n-1}) \in \mathcal{M}_1(\mathcal{X}_n)$ is as follows. Let $A^{(n)} = \{\mathbf{x} : x_n \in A\}$, $A \in \mathcal{B}(\mathcal{X}_n)$, $B^{(n)} = \{\mathbf{y} : y_n \in B\}$, $B \in \mathcal{B}(\mathcal{Y}_n)$, and let $\overrightarrow{\Pi}_{0,n}(A^{(n)}, B^{(n)}|\mathcal{B}(\mathcal{X}_{0,n-1}) \otimes \mathcal{B}(\mathcal{Y}_{0,n-1}))$ denote the joint conditional probability of $A^{(n)} \times B^{(n)}$ with respect to $\mathcal{B}(\mathcal{X}_{0,n-1}) \otimes \mathcal{B}(\mathcal{Y}_{0,n-1})$ calculated on the probability space $(\mathcal{X}^{\mathbb{N}_0} \otimes \mathcal{Y}^{\mathbb{N}_0}, \mathcal{B}(\mathcal{X}^{\mathbb{N}_0}) \otimes \mathcal{B}(\mathcal{Y}^{\mathbb{N}_0}), \overrightarrow{\Pi}_{0,n}(\cdot))$. Then, for $A \in \mathcal{B}(\mathcal{X}_n)$, $B \in \mathcal{B}(\mathcal{Y}_n)$ we obtain

$$\overrightarrow{\Pi}_{0,n}(A^{(n)}, B^{(n)}|\mathcal{B}(\mathcal{X}_{0,n-1}) \otimes \mathcal{B}(\mathcal{Y}_{0,n-1})) = p_n(A|x^{n-1}, y^{n-1}) \times \nu_n(B|y^{n-1}). \quad (\text{III.12})$$

Hence, $\nu_n(\cdot|y^{n-1}) \in \mathcal{M}_1(\mathcal{Y}_n)$ is given by $\nu_n(dy_n|y^{n-1}) = \int_{\mathcal{X}_n} \overrightarrow{\Pi}_{0,n}(dx_n, dy_n|x^{n-1}, y^{n-1})$, from which $\nu_{0,n}(dy^n) \in \mathcal{M}_1(\mathcal{Y}_{0,n})$ is also obtained. Similarly, let $\overleftarrow{\Pi}_{0,n}(A^{(n)}, B^{(n)}|\mathcal{B}(\mathcal{Y}_{0,n-1}) \otimes \mathcal{B}(\mathcal{X}_{0,n-1}))$ denote

³Similarly, $\nu \equiv \nu^{\mathbf{P} \otimes \mathbf{Q}}$.

the joint conditional probability of $A^{(n)} \times B^{(n)}$ with respect to $\mathcal{B}(\mathcal{Y}_{0,n-1}) \otimes \mathcal{B}(\mathcal{X}_{0,n-1})$ calculated on the probability space $(\mathcal{Y}^{\mathbb{N}_0} \times \mathcal{X}^{\mathbb{N}_0}, \mathcal{B}(\mathcal{Y}^{\mathbb{N}_0}) \otimes \mathcal{B}(\mathcal{X}^{\mathbb{N}_0}), \bar{\Pi}_{0,n}(\cdot))$. Then, for $B \in \mathcal{B}(\mathcal{Y}_n)$ we have

$$\bar{\Pi}_{0,n}(A^{(n)}, B^{(n)} | \mathcal{B}(\mathcal{X}_{0,n-1}) \otimes \mathcal{B}(\mathcal{Y}_{0,n-1})) = \int_{A_n} q_n(B | y^{n-1}, x^n) \otimes \mu_n(dx_n | x^{n-1}) \quad (\text{III.13})$$

from which $\mu_n(\cdot | x^{n-1}) \in \mathcal{M}_1(\mathcal{X}_n)$ and $\mu_{0,n}(dx^n) \in \mathcal{M}_1(\mathcal{X}_{0,n})$ are obtained. Similarly, from (III.10) and (III.12) we can obtain any of the individual conditional distributions $p_n(\cdot | x^{n-1}, y^{n-1}) \in \mathcal{M}_1(\mathcal{X}_n)$ and $q_n(\cdot | y^{n-1}, x^n) \in \mathcal{M}_1(\mathcal{Y}_n)$ appearing in their RHS by proper conditional expectations.

Using the first definition of basic processes, that is, given a collection of stochastic kernels $\{p_n(\cdot | \cdot, \cdot) \in \mathcal{Q}(\mathcal{X}_n | \mathcal{X}_{0,n-1} \times \mathcal{Y}_{0,n-1}) : n \in \mathbb{N}_0\}$ and $\{q_n(\cdot | \cdot, \cdot) \in \mathcal{Q}(\mathcal{Y}_n | \mathcal{Y}_{0,n-1} \times \mathcal{X}_{0,n}) : n \in \mathbb{N}_0\}$, the joint distribution, as well as the conditional distributions are defined via **P1 – P5**. Consequently, it is well-known that directed information is defined via relative entropy as follows [11]

$$\begin{aligned} I(X^n \rightarrow Y^n) &\triangleq \sum_{i=0}^n I(X^i; Y_i | Y^{i-1}) \\ &= \sum_{i=0}^n \int_{\mathcal{Y}_{0,i-1}} \int_{\mathcal{X}_{0,i} \times \mathcal{Y}_i} \log \left(\frac{dP_{0,i}(\cdot, \cdot | y^{i-1})}{d(P_{0,i}(\cdot | y^{i-1}) \times \nu_i(\cdot | y^{i-1}))}(x^i, y_i) \right) P_{0,i}(dx^i, dy_i | y^{i-1}) P_{0,i-1}(dy^{i-1}) \quad (\text{III.14}) \end{aligned}$$

$$\begin{aligned} &= \sum_{i=0}^n \int_{\mathcal{X}_{0,i} \times \mathcal{Y}_{0,i-1}} \mathbb{D}(q_i(\cdot | y^{i-1}, x^i) || \nu_i(\cdot | y^{i-1})) p_i(dx_i | x^{i-1}, y^{i-1}) \\ &\quad \otimes_{j=0}^{i-1} (q_j(dy_j | y^{j-1}, x^j) \otimes p_j(dx_j | x^{j-1}, y^{j-1})) \quad (\text{III.15}) \end{aligned}$$

$$\equiv \mathbb{I}_{X^n \rightarrow Y^n}(p_i(\cdot | \cdot, \cdot), q_i(\cdot | \cdot, \cdot) : i = 0, 1, \dots, n). \quad (\text{III.16})$$

The RHS in (III.14) follows from the definition of conditional mutual information. In (III.16), we use the notation $\mathbb{I}_{X^n \rightarrow Y^n}(p_i(\cdot | \cdot, \cdot), q_i(\cdot | \cdot, \cdot) : i = 0, 1, \dots, n)$ to indicate that $I(X^n \rightarrow Y^n)$ is a functional of $\{p_i(\cdot | \cdot, \cdot) \in \mathcal{Q}(\mathcal{X}_i | \mathcal{X}_{0,i-1} \times \mathcal{Y}_{0,i-1}), q_i(\cdot | \cdot, \cdot) \in \mathcal{Q}(\mathcal{Y}_i | \mathcal{Y}_{0,i-1} \times \mathcal{X}_{0,i}) : i = 0, 1, \dots, n\}$.

A. Directed Information Functional of Consistent Conditional Distributions

Now we consider the second definition of basic process introduced in Section II. Given any $\mathbf{P}(\cdot | \cdot) \in \mathcal{Q}^{\mathbf{C1}}(\mathcal{X}^{\mathbb{N}_0} | \mathcal{Y}^{\mathbb{N}_0})$ and $\mathbf{Q}(\cdot | \cdot) \in \mathcal{Q}^{\mathbf{C2}}(\mathcal{X}^{\mathbb{N}_0} | \mathcal{Y}^{\mathbb{N}_0})$ the distributions under **P1 – P5** are constructed. Next, we define directed information via relative entropy as often done for mutual information [28]. By Lemma A.9, $\bar{P}_{0,n} \otimes \bar{Q}_{0,n} << \bar{P}_{0,n} \otimes \nu_{0,n}$ if and only if $\bar{Q}_{0,n}(\cdot | x^n) << \nu_{0,n}(\cdot)$ for $\bar{P}_{0,n}$ -almost all $x^n \in \mathcal{X}_{0,n}$. Utilizing the Radon-Nikodym derivative (RND) $\frac{d(\bar{P}_{0,n} \otimes \bar{Q}_{0,n})}{d(\bar{P}_{0,n} \otimes \nu_{0,n})}(x^n, y^n)$, define the relative entropy of $\bar{P}_{0,n} \otimes \bar{Q}_{0,n}$ with respect to $\bar{\Pi}_{0,n}$ as follows.

$$\begin{aligned} \mathbb{I}_{X^n \rightarrow Y^n}(\bar{P}_{0,n}, \bar{Q}_{0,n}) &\triangleq \mathbb{D}(\bar{P}_{0,n} \otimes \bar{Q}_{0,n} || \bar{\Pi}_{0,n}) \\ &= \int_{\mathcal{X}_{0,n} \times \mathcal{Y}_{0,n}} \log \left(\frac{d(\bar{P}_{0,n} \otimes \bar{Q}_{0,n})}{d(\bar{P}_{0,n} \otimes \nu_{0,n})}(x^n, y^n) \right) (\bar{P}_{0,n} \otimes \bar{Q}_{0,n})(dx^n, dy^n) \quad (\text{III.17}) \end{aligned}$$

$$\begin{aligned} &= \int_{\mathcal{X}_{0,n} \times \mathcal{Y}_{0,n}} \log \left(\frac{d\bar{Q}_{0,n}(\cdot | x^n)}{d\nu_{0,n}(\cdot)}(y^n) \right) (\bar{P}_{0,n} \otimes \bar{Q}_{0,n})(dx^n, dy^n) \quad (\text{III.18}) \end{aligned}$$

$$\equiv \mathbb{I}_{X^n \rightarrow Y^n}(\bar{P}_{0,n}, \bar{Q}_{0,n}) \quad (\text{III.19})$$

Note that (III.18) is obtained by utilizing the fact that if $\bar{P}_{0,n} \otimes \bar{Q}_{0,n} << \bar{P}_{0,n} \otimes \nu_{0,n}$ then the RND $\frac{d(\bar{P}_{0,n} \otimes \bar{Q}_{0,n})}{d(\bar{P}_{0,n} \otimes \nu_{0,n})}(x^n, y^n)$ represents a version of $\frac{d\bar{Q}_{0,n}(\cdot | x^n)}{d\nu_{0,n}(\cdot)}(y^n)$, $\bar{P}_{0,n}$ -a.s for all $x^n \in \mathcal{X}_{0,n}$. On the other hand,

using Lemma A.9, $\vec{Q}_{0,n}(\cdot|x^n) \ll \nu_{0,n}(\cdot)$, $\overleftarrow{P}_{0,n}$ -almost $x^n \in \mathcal{X}_{0,n}$, and by Radon-Nikodym theorem, there exists a version of the RND $\bar{\xi}_{0,n}(x^n, y^n) \triangleq \frac{d\vec{Q}_{0,n}(\cdot|x^n)}{d\nu_{0,n}(\cdot)}(y^n)$ which is a non-negative measurable function of $(x^n, y^n) \in \mathcal{X}_{0,n} \times \mathcal{Y}_{0,n}$. Hence another version of $\bar{\xi}_{0,n}(\cdot, \cdot)$ is $\bar{\xi}_{0,n}(x^n, y^n) = \frac{d(\overleftarrow{P}_{0,n} \otimes \vec{Q}_{0,n})}{d(\overleftarrow{P}_{0,n} \otimes \nu_{0,n})}(x^n, y^n)$. We use notation $\mathbb{I}_{X^n \rightarrow Y^n}(\overleftarrow{P}_{0,n}, \vec{Q}_{0,n})$ given in (III.19) to illustrate that $\mathbb{D}(\overleftarrow{P}_{0,n} \otimes \vec{Q}_{0,n} || \vec{\Pi}_{0,n})$ is a functional of $\{\overleftarrow{P}_{0,n}(\cdot|\cdot), \vec{Q}_{0,n}(\cdot|\cdot)\} \in \mathcal{Q}^{C1}(\mathcal{X}_{0,n}|\mathcal{Y}_{0,n-1}) \times \mathcal{Q}^{C2}(\mathcal{Y}_{0,n}|\mathcal{X}_{0,n})$.

In the next Remark we summarize the equivalent definitions of directed information based on the two equivalent definitions of channels, that is, the one based on (III.15), (III.16), and the one based on (III.17), (III.18).

Remark III.1.

Let $\mathbf{P}(\cdot|\cdot) \in \mathcal{Q}^{C1}(\mathcal{X}^{\mathbb{N}_0}|\mathcal{Y}^{\mathbb{N}_0})$ and $\mathbf{Q}(\cdot|\cdot) \in \mathcal{Q}^{C2}(\mathcal{Y}^{\mathbb{N}_0}|\mathcal{X}^{\mathbb{N}_0})$. By repeated application of Lemma A.9, and the chain rule of relative entropy [30, Theorem B.2.1., p. 326], directed information admits the following equivalent definitions.

$$I(X^n \rightarrow Y^n) \triangleq \sum_{i=0}^n I(X^i; Y_i | Y^{i-1}) = \mathbb{D}(\overleftarrow{P}_{0,n} \otimes \vec{Q}_{0,n} || \vec{\Pi}_{0,n}) \quad (\text{III.20})$$

$$= \int_{\mathcal{X}_{0,n} \times \mathcal{Y}_{0,n}} \log \left(\frac{d\vec{Q}_{0,n}(\cdot|x^n)}{d\nu_{0,n}(\cdot)}(y^n) \right) (\overleftarrow{P}_{0,n} \otimes \vec{Q}_{0,n})(dx^n, dy^n) \equiv \mathbb{I}_{X^n \rightarrow Y^n}(\overleftarrow{P}_{0,n}, \vec{Q}_{0,n}). \quad (\text{III.21})$$

Clearly, (III.21) is valid even when $(\overleftarrow{P}_{0,n} \otimes \vec{Q}_{0,n})(dx^n, dy^n)$ is singular with respect to $(\overleftarrow{P}_{0,n} \otimes \nu_{0,n})(dx^n, dy^n)$, in which case its value is $+\infty$. The point to be made here is that we will show the convexity, concavity, lower semicontinuity properties of directed information using the definition $I(X^n \rightarrow Y^n) = \mathbb{D}(\overleftarrow{P}_{0,n} \otimes \vec{Q}_{0,n} || \vec{\Pi}_{0,n}) \equiv \mathbb{I}_{X^n \rightarrow Y^n}(\overleftarrow{P}_{0,n}, \vec{Q}_{0,n})$, as a functional of $\overleftarrow{P}_{0,n}(\cdot|y^{n-1}) \in \mathcal{M}_1^{C1}(\mathcal{X}_{0,n})$ and $\vec{Q}_{0,n}(\cdot|x^n) \in \mathcal{M}_1^{C2}(\mathcal{Y}_{0,n})$. We will also use the directed information definition $\mathbb{D}(\overleftarrow{P}_{0,n} \otimes \vec{Q}_{0,n} || \vec{\Pi}_{0,n})$, as a functional of $\{\overleftarrow{P}_{0,n}, \vec{Q}_{0,n}\}$ to show lower semicontinuity, convexity and concavity properties. Then we will use these functional and topological properties to demonstrate how to establish existence of optimal solutions to the two extremum problems defined by (I.5) and (I.8), respectively.

B. Convexity and Concavity of Directed Information

First, we show that the set of conditional distributions $\mathbf{P}(\cdot|\mathbf{y}) \in \mathcal{M}_1^{C1}(\mathcal{X}^{\mathbb{N}_0})$ and $\mathbf{Q}(\cdot|\mathbf{x}) \in \mathcal{M}_1^{C2}(\mathcal{Y}^{\mathbb{N}_0})$, i.e., satisfying consistency conditions C1 and C2, are convex, and then we show convexity of directed information with respect to $\mathbf{Q}(\cdot|\mathbf{x})$ and concavity with respect to $\mathbf{P}(\cdot|\mathbf{y})$.

Recall that the set of all distributions $\mathbf{P}(\cdot|\mathbf{y}) \in \mathcal{M}_1(\mathcal{X}^{\mathbb{N}_0})$ and $\mathbf{Q}(\cdot|\mathbf{x}) \in \mathcal{M}_1(\mathcal{Y}^{\mathbb{N}_0})$ (i.e., without imposing consistency conditions C1 and C2) are convex, that is, given $\{\mathbf{P}^1(\cdot|\mathbf{y}), \mathbf{P}^2(\cdot|\mathbf{y})\} \in \mathcal{M}_1(\mathcal{X}^{\mathbb{N}_0}) \times \mathcal{M}_1(\mathcal{X}^{\mathbb{N}_0})$, and $\lambda \in (0, 1)$, there exists a probability measure \tilde{P} on $(\mathcal{X}^{\mathbb{N}_0} \times \mathcal{Y}^{\mathbb{N}_0}, \mathcal{B}(\mathcal{X}^{\mathbb{N}_0}) \otimes \mathcal{B}(\mathcal{Y}^{\mathbb{N}_0}))$ whose regular distribution $\tilde{P}(\cdot|\mathbf{y})$ satisfies $\tilde{P}(\cdot|\mathbf{y}) = \lambda \mathbf{P}^1(\cdot|\mathbf{y}) + (1 - \lambda) \mathbf{P}^2(\cdot|\mathbf{y}) \in \mathcal{M}_1(\mathcal{X}^{\mathbb{N}_0})$.

Next, we show convexity of the sets $\mathcal{M}_1^{C1}(\mathcal{X}^{\mathbb{N}_0})$ and $\mathcal{M}_1^{C2}(\mathcal{Y}^{\mathbb{N}_0})$.

Theorem III.2. (Convexity of sets $\mathcal{M}_1^{C1}(\mathcal{X}^{\mathbb{N}_0})$, $\mathcal{M}_1^{C2}(\mathcal{Y}^{\mathbb{N}_0})$)

Let $\{\mathcal{X}_n : n \in \mathbb{N}_0\}$, $\{\mathcal{Y}_n : n \in \mathbb{N}_0\}$ be Polish spaces with $\mathcal{B}(\mathcal{X}_n)$, $\mathcal{B}(\mathcal{Y}_n)$, respectively, the σ -algebras of Borel sets. Then the sets of distributions $\mathbf{P}(\cdot|\mathbf{y}) \in \mathcal{M}_1^{C1}(\mathcal{X}^{\mathbb{N}_0})$ and $\mathbf{Q}(\cdot|\mathbf{x}) \in \mathcal{M}_1^{C2}(\mathcal{Y}^{\mathbb{N}_0})$ are convex, and similarly, their projection to finite number of coordinates, that is, $\overleftarrow{P}_{0,n}(\cdot|y^{n-1}) \in \mathcal{M}_1^{C1}(\mathcal{X}_{0,n})$ and $\vec{Q}_{0,n}(\cdot|x^n) \in \mathcal{M}_1^{C2}(\mathcal{Y}_{0,n})$, are also convex.

Proof: Since the methodology is similar for both sets, only the derivation for $\mathcal{M}_1^{C1}(\mathcal{X}^{\mathbb{N}_0})$ is given. By definition, the set of distributions $\mathcal{M}_1^{C1}(\mathcal{X}^{\mathbb{N}_0})$ is convex if for a given $\{\mathbf{P}^1(\cdot|\mathbf{y}), \mathbf{P}^2(\cdot|\mathbf{y})\} \in \mathcal{M}_1^{C1}(\mathcal{X}^{\mathbb{N}_0}) \times$

$\mathcal{M}_1^{\mathbf{C1}}(\mathcal{X}^{\mathbb{N}_0})$, and a given $\lambda \in (0, 1)$, there exists a probability measure \tilde{P} on $(\mathcal{X}^{\mathbb{N}_0} \times \mathcal{Y}^{\mathbb{N}_0}, \mathcal{B}(\mathcal{X}^{\mathbb{N}_0}) \otimes \mathcal{B}(\mathcal{Y}^{\mathbb{N}_0}))$, whose regular conditional measure $\tilde{P}(\cdot|\mathbf{y})$ is a convex combination $\tilde{P}(\cdot|\mathbf{y}) = \lambda \mathbf{P}^1(\cdot|\mathbf{y}) + (1 - \lambda) \mathbf{P}^2(\cdot|\mathbf{y})$, *a.e.* $\mathbf{y} \in \mathcal{Y}^{\mathbb{N}_0}$, and consistency condition **C1** holds, i.e., $\lambda \mathbf{P}^1(\cdot|\mathbf{y}) + (1 - \lambda) \mathbf{P}^2(\cdot|\mathbf{y}) \in \mathcal{M}_1^{\mathbf{C1}}(\mathcal{X}^{\mathbb{N}_0})$. By [45], the set of distributions $\mathcal{M}_1(\mathcal{X}^{\mathbb{N}_0})$ is convex, and since $\{\mathbf{P}^1(\cdot|\mathbf{y}), \mathbf{P}^2(\cdot|\mathbf{y})\} \in \mathcal{M}_1(\mathcal{X}^{\mathbb{N}_0}) \times \mathcal{M}_1(\mathcal{X}^{\mathbb{N}_0})$, then there is a probability measure \tilde{P} on $\mathcal{M}_1(\mathcal{X}^{\mathbb{N}_0} \times \mathcal{Y}^{\mathbb{N}_0}, \mathcal{B}(\mathcal{X}^{\mathbb{N}_0}) \otimes \mathcal{B}(\mathcal{Y}^{\mathbb{N}_0}))$, whose regular distribution $\tilde{P}(\cdot|\mathbf{y})$, $\mathbf{y} \in \mathcal{Y}^{\mathbb{N}_0}$, satisfies

$$\tilde{P}(\cdot|\mathbf{y}) = \lambda \mathbf{P}^1(\cdot|\mathbf{y}) + (1 - \lambda) \mathbf{P}^2(\cdot|\mathbf{y}) \in \mathcal{M}_1(\mathcal{X}^{\mathbb{N}_0}), \quad \forall \lambda \in (0, 1).$$

Moreover, if $\mathbf{P}^1(\cdot|\mathbf{y})$, and $\mathbf{P}^2(\cdot|\mathbf{y})$ satisfy consistency condition **C1**, then their convex combination also satisfies consistency condition **C1**, and consequently $\lambda \mathbf{P}^1(\cdot|\mathbf{y}) + (1 - \lambda) \mathbf{P}^2(\cdot|\mathbf{y}) \in \mathcal{M}_1^{\mathbf{C1}}(\mathcal{X}^{\mathbb{N}_0})$, i.e., the consistency condition **C1** holds. The derivation for $\mathbf{Q}(\cdot|\mathbf{x}) \in \mathcal{M}_1^{\mathbf{C2}}(\mathcal{Y}^{\mathbb{N}_0})$ is similar. The derivation for the projection to finite number of coordinates is done as follows. Let $A^{(n)} = \{\mathbf{x} : x_n \in A\}$, $A \in \mathcal{B}(\mathcal{X}_n)$, and let $\mathbf{P}(A^{(n)}|\mathcal{B}(\mathcal{X}_{0,n-1})|\mathbf{y})$ denote the conditional probability of $A^{(n)}$ with respect to $\mathcal{B}(\mathcal{X}_{0,n-1})$ calculated on the probability space $(\mathcal{X}^{\mathbb{N}}, \mathcal{B}(\mathcal{X}^{\mathbb{N}}), \mathbf{P}(\cdot|\mathbf{y}))$. From the definition of regular conditional probability measures, it follows that

$$\begin{aligned} \tilde{P}(A^{(n)}|\mathcal{B}(\mathcal{X}_{0,n-1})|\mathbf{y}) &= \lambda \mathbf{P}^1(A^{(n)}|\mathcal{B}(\mathcal{X}_{0,n-1})|\mathbf{y}) + (1 - \lambda) \mathbf{P}^2(A^{(n)}|\mathcal{B}(\mathcal{X}_{0,n-1})|\mathbf{y}) - a.s. \\ &= \lambda p_n^1(A|x^{n-1}, y^{n-1}) + (1 - \lambda) p_n^2(A|x^{n-1}, y^{n-1}) - a.s. \end{aligned}$$

where $p_n^1(\cdot|x^{n-1}, y^{n-1})$, $p_n^2(\cdot|x^{n-1}, y^{n-1})$ are regular conditional distributions. Since convex combination of regular conditional distributions is also a regular conditional distribution, by Remark II.1 the set $\tilde{P}_{0,n}(\cdot|y^{n-1}) \in \mathcal{M}_1^{\mathbf{C1}}(\mathcal{X}_{0,n})$ is convex, and the derivation is complete. \square

Since $\mathcal{M}_1^{\mathbf{C1}}(\mathcal{X}_{0,n})$ and $\mathcal{M}_1^{\mathbf{C2}}(\mathcal{Y}_{0,n})$ are convex, then we proceed further to show that directed information $\mathbb{I}_{X^n \rightarrow Y^n}(\tilde{P}_{0,n}, \vec{Q}_{0,n})$, as a functional of $\tilde{P}_{0,n}(\cdot|y^{n-1}) \in \mathcal{M}_1^{\mathbf{C1}}(\mathcal{X}_{0,n})$, for a fixed $\vec{Q}_{0,n}(\cdot|x^n) \in \mathcal{M}_1^{\mathbf{C2}}(\mathcal{Y}_{0,n})$, is concave, and as a functional of $\vec{Q}_{0,n}(\cdot|x^n) \in \mathcal{M}_1^{\mathbf{C2}}(\mathcal{Y}_{0,n})$, for a fixed $\tilde{P}_{0,n}(\cdot|y^{n-1}) \in \mathcal{M}_1^{\mathbf{C1}}(\mathcal{X}_{0,n})$, is convex. These results are shown in the next theorem.

Theorem III.3. (*Convexity of conditional distributions*)

Let $\{\mathcal{X}_n : n \in \mathbb{N}_0\}$, $\{\mathcal{Y}_n : n \in \mathbb{N}_0\}$ be Polish spaces with $\mathcal{B}(\mathcal{X}_n)$, $\mathcal{B}(\mathcal{Y}_n)$, respectively, the σ -algebras of Borel sets. Consider the directed information functional $I(X^n \rightarrow Y^n) = \mathbb{I}_{X^n \rightarrow Y^n}(\tilde{P}_{0,n}, \vec{Q}_{0,n})$, $\mathbb{I}_{X^n \rightarrow Y^n} : \mathcal{M}_1^{\mathbf{C1}}(\mathcal{X}_{0,n}) \times \mathcal{M}_1^{\mathbf{C2}}(\mathcal{Y}_{0,n}) \mapsto [0, \infty]$ defined by (III.21).

Then the following hold.

- 1) $\mathbb{I}_{X^n \rightarrow Y^n}(\tilde{P}_{0,n}, \vec{Q}_{0,n})$ is a convex functional of $\vec{Q}_{0,n}(\cdot|x^n) \in \mathcal{M}_1^{\mathbf{C2}}(\mathcal{Y}_{0,n})$ for a fixed $\tilde{P}_{0,n}(\cdot|y^{n-1}) \in \mathcal{M}_1^{\mathbf{C1}}(\mathcal{X}_{0,n})$.
- 2) $\mathbb{I}_{X^n \rightarrow Y^n}(\tilde{P}_{0,n}, \vec{Q}_{0,n})$ is a concave functional of $\tilde{P}_{0,n}(\cdot|y^{n-1}) \in \mathcal{M}_1^{\mathbf{C1}}(\mathcal{X}_{0,n})$ for a fixed $\vec{Q}_{0,n}(\cdot|x^n) \in \mathcal{M}_1^{\mathbf{C2}}(\mathcal{Y}_{0,n})$.
- 3) $\mathbb{I}_{X^n \rightarrow Y^n}(\tilde{P}_{0,n}, \cdot)$ is a strictly convex functional on the set $\{\vec{Q}_{0,n}(\cdot|x^n) \in \mathcal{M}_1^{\mathbf{C2}}(\mathcal{Y}_{0,n}) : \mathbb{I}_{X^n \rightarrow Y^n}(\tilde{P}_{0,n}, \vec{Q}_{0,n}) < \infty\}$ for a fixed $\tilde{P}_{0,n}(\cdot|y^{n-1}) \in \mathcal{M}_1^{\mathbf{C1}}(\mathcal{X}_{0,n})$.

Proof: By Theorem III.2, the sets $\mathcal{M}_1^{\mathbf{C1}}(\mathcal{X}_{0,n})$ and $\mathcal{M}_1^{\mathbf{C2}}(\mathcal{Y}_{0,n})$ are convex. Therefore, to show parts 1), 2), 3) we utilize the consistency of the two families of conditional distributions and we apply the log-sum formulae, and the existence of certain Radon-Nikodym Derivatives (RNDs). The complete derivation is given in Appendix B. \square

Theorem III.3 is analogous to mutual information $I(X^n; Y^n) \equiv \mathbb{I}_{X^n, Y^n}(P_{X^n}, P_{Y^n|X^n})$, expressed as a functional of input distribution $P_{X^n}(\cdot) \in \mathcal{M}_1(\mathcal{X}_{0,n})$ and the channel $P_{Y^n|X^n}(\cdot|x^n) \in \mathcal{M}_1(\mathcal{Y}_{0,n})$, which is known to be a convex (respectively concave) functional of $P_{Y^n|X^n}(\cdot|x^n) \in \mathcal{M}_1(\mathcal{Y}_{0,n})$ (respectively $P_{X^n}(\cdot) \in \mathcal{M}_1(\mathcal{X}_{0,n})$), for a fixed $P_{X^n}(\cdot) \in \mathcal{M}_1(\mathcal{X}_{0,n})$ (respectively $P_{Y^n|X^n}(\cdot|x^n) \in \mathcal{M}_1(\mathcal{Y}_{0,n})$). It is important to point out that if one considers the alternative definition of directed information (III.14), (III.16), as a

functional of the sequence of input channel distributions, $I(X^n \rightarrow Y^n) \equiv \mathbb{I}_{X^n \rightarrow Y^n}(p_i(\cdot|\cdot, \cdot), q_i(\cdot|\cdot, \cdot) : i = 0, 1, \dots, n)$, then it is not clear to us whether it is possible to establish convexity and concavity with respect to q_i and p_i .

For finite alphabet spaces, the convexity of the set of causally conditioned probability mass functions $P(x^n|y^{n-1}) \triangleq \prod_{i=0}^n p(x_i|x^{i-1}, y^{i-1})$ and $Q(y^n|x^n) \triangleq \prod_{i=0}^n q(y_i|y^{i-1}, x^i)$ is shown in [46, Lemma 1], under the assumption that for each $n \in \mathbb{N}_0$, the ratios $\frac{P(x^n|y^{n-1})}{P(x^{n-1}|y^{n-1})}$ and $\frac{Q(y^n|x^n)}{Q(y^{n-1}|x^{n-1})}$ exist, and they are given by $p(x_n|x^{n-1}, y^{n-1})$ and $q(y_n|y^{n-1}, x^n)$, respectively. The derivation in [46] is based on showing that the set of all causally conditioned distributions $P(x^n|y^{n-1})$ is a polyhedron. The method described in [46] does not apply to conditional distributions defined on continuous alphabets. Theorem III.2 and Theorem III.3, hold for general conditional distributions defined on abstract alphabet spaces, and they do not require existence of probability density functions (corresponding to the causally conditioned distributions for each $n \in \mathbb{N}_0$), hence they compliment the work in [46].

C. Weak Convergence and Compactness of Conditional Distributions

In this section we give general sufficient conditions for weak compactness of the set of probability distributions $\overleftarrow{P}_{0,n}(\cdot|y^{n-1}) \in \mathcal{M}_1^{\mathcal{C}^1}(\mathcal{X}_{0,n})$ and $\overrightarrow{Q}_{0,n}(\cdot|x^n) \in \mathcal{M}_1^{\mathcal{C}^2}(\mathcal{Y}_{0,n})$, and compactness of the set of joint and marginal measures with respect to the topology of weak convergence of probability measures. These conditions are sufficient to show lower semicontinuity of $\mathbb{I}_{X^n \rightarrow Y^n}(\overleftarrow{P}_{0,n}, \overrightarrow{Q}_{0,n})$ for fixed $\overleftarrow{P}_{0,n}(\cdot|y^{n-1}) \in \mathcal{M}_1^{\mathcal{C}^1}(\mathcal{X}_{0,n})$ (respectively $\overrightarrow{Q}_{0,n}(\cdot|x^n) \in \mathcal{M}_1^{\mathcal{C}^2}(\mathcal{Y}_{0,n})$) with respect to $\overrightarrow{Q}_{0,n}(\cdot|x^n) \in \mathcal{M}_1^{\mathcal{C}^2}(\mathcal{Y}_{0,n})$ (respectively $\overleftarrow{P}_{0,n}(\cdot|y^{n-1}) \in \mathcal{M}_1^{\mathcal{C}^1}(\mathcal{X}_{0,n})$). The lower semicontinuity of directed information is the analogue of the lower semicontinuity of mutual information, extensively utilized in information theory and statistics (see [28], [47]).

Before we state the main theorem, we introduce the following notation. Let $BC(\mathcal{X})$ denote the set of bounded, continuous real-valued function f defined on a metric space (\mathcal{X}, d) endowed with the supremum norm $\|f\| = \sup_{x \in \mathcal{X}} |f(x)|$. A sequence of probability measures $\{P_\alpha : \alpha = 1, 2, \dots\} \subset \mathcal{M}_1(\mathcal{X})$ is said to *converge weakly* to a probability measure $P \in \mathcal{M}_1(\mathcal{X})$ if [31]

$$\lim_{\alpha \rightarrow \infty} \int_{\mathcal{X}} f(x) dP_\alpha(x) = \int_{\mathcal{X}} f(x) dP(x), \quad \forall f \in BC(\mathcal{X}).$$

Weak convergence of $\{P_\alpha : \alpha = 1, 2, \dots\}$ to P is denoted by $P_\alpha \xrightarrow{w} P$. A family of probability measures $M \subset \mathcal{M}_1(\mathcal{X})$ is called *relatively compact or weakly compact* if every sequence in M contains a weakly convergent subsequence that converges to $\mathcal{M}_1(\mathcal{X})$ but not necessarily to M . Appendix A summarizes well-known theorems of weak convergence, compactness, tightness, and Prohorov's theorem, which we invoke to derive the results of this section.

Throughout sequences of points in $\mathcal{X}^{\mathbb{N}_0}$ and $\mathcal{Y}^{\mathbb{N}_0}$ are denoted by $\mathbf{x}^{(\alpha)} \triangleq \{x_0^{(\alpha)}, x_1^{(\alpha)}, \dots\} \in \mathcal{X}^{\mathbb{N}_0}$, $\mathbf{y}^{(\alpha)} \triangleq \{y_0^{(\alpha)}, y_1^{(\alpha)}, \dots\} \in \mathcal{Y}^{\mathbb{N}_0}$, $\alpha = 1, 2, \dots$. Moreover, a sequence of points $\mathbf{x}^{(\alpha)} \in \mathcal{X}^{\mathbb{N}_0}$, $\alpha = 1, 2, \dots$ is said to converge to $\mathbf{x}^{(o)} \in \mathcal{X}^{\mathbb{N}_0}$ as $\alpha \rightarrow \infty$, if $\lim_{\alpha \rightarrow \infty} x_n^{(\alpha)} = x_n^{(o)}$ for every $n \in \mathbb{N}_0$. Sequences of such points in $\mathcal{X}_{0,n} \triangleq \times_{i=0}^n \mathcal{X}_i$ and $\mathcal{Y}_{0,n} \triangleq \times_{i=0}^n \mathcal{Y}_i$ are denoted by $x^{n,(\alpha)} \triangleq \{x_0^{(\alpha)}, x_1^{(\alpha)}, \dots, x_n^{(\alpha)}\}$ and $y^{n,(\alpha)} \triangleq \{y_0^{(\alpha)}, y_1^{(\alpha)}, \dots, y_n^{(\alpha)}\}$, $\alpha = 1, 2, \dots$.

The next remark, is introduced to illustrate that in applications of weak convergence of probability distributions, weak continuity of probability distributions is natural, when analyzing conditional distributions with discontinuities, such as, distributions induced by mixture of discrete and continuous RVs.

Remark III.4. (Weak continuity vs. Strong continuity)

Let $q(\cdot|x) \in \mathcal{Q}(\mathcal{Y}|\mathcal{X})$ be a conditional distribution, and suppose there is a distribution $\mu(dx) \in \mathcal{M}_1(\mathcal{X})$ such that for every $x \in \mathcal{X}$, $q(\cdot|x)$ has a density $\bar{q}(\cdot|x)$ with respect to $\mu(\cdot)$, i.e.,

$$q(B|x) = \int_B \bar{q}(y|x) \mu(dx), \quad \forall B \in \mathcal{B}(\mathcal{Y}), \quad \forall x \in \mathcal{X}.$$

For example, if $\mathcal{X} \in \mathbb{R}$ then $\mu(dx) = dx$ is the Lebesgue measure on \mathbb{R} . If $\bar{q}(y|\cdot)$ is continuous on \mathcal{X} for every $y \in \mathcal{Y}$, then $q(\cdot|\cdot) \in \mathcal{Q}(\mathcal{Y}|\mathcal{X})$ is strongly continuous (i.e., $q(B|\cdot)$ is continuous on \mathcal{X} for every $B \in \mathcal{B}(\mathcal{Y})$). Strong continuity of channel models is rather restrictive, because it rules out conditional distributions which have discontinuities, such as, additive noise channels, in which noise is a mixture of a continuous RV (i.e., Gaussian distributed RV) and a finite alphabet valued RV.

Consider a channel model with feedback described by the nonlinear recursive equation

$$Y_n = h_n(Y^{n-1}, X_n, V_n), \quad Y^{-1} = y^{-1}, \quad n = 0, 1, \dots$$

where $\{h_n : \mathcal{Y}_{0,n-1} \times \mathcal{X}_n \times \mathcal{V}_n \mapsto \mathcal{Y}_n : n = 0, 1, \dots\}$, is a sequence of measurable functions and $\{V_n : n = 0, 1, \dots\}$ is a sequence of $\{\mathcal{V}_n : n = 0, 1, \dots\}$ -valued RV's, representing the channel noise. Suppose the following condition holds.

$$P_{V_n|Y^{n-1}, X^n, Y^{n-1}}(dv_n|v^{n-1}, x^n, y^{n-1}) = P_{V_n}(v_n), \quad n = 0, 1, \dots$$

Then the channel distribution induced by the above model is

$$\begin{aligned} q_n(B|y^{n-1}, x^n) &= \mathbb{P}\{Y_n \in B | Y^{n-1} = y^{n-1}, X^n = x^n\}, \quad B \in \mathcal{B}(\mathcal{Y}) \\ &= \mathbb{P}\{h_n(Y^{n-1}, X_n, V_n) \in B | Y^{n-1} = y^{n-1}, X^n = x^n\}, \\ &= \mathbf{P}(\{v_n \in V_n : h_n(y^{n-1}, x_n, v_n) \in B\}) = \int_{\mathcal{V}_n} I_B(h_n(y^{n-1}, x_n, v_n)) P_{V_n}(dv_n) \\ &\equiv q_n(B|y^{n-1}, x_n) \end{aligned}$$

where $I_B(\cdot)$ is the indicator function. If for each n , the function $h_n(\cdot, \cdot, v_n)$ is continuous on $\mathcal{Y}_{0,n-1} \times \mathcal{X}_n$ for every $v_n \in \mathcal{V}_n$, $n = 0, 1, \dots$, then by bounded convergence theorem $\{q_n(\cdot|\cdot, \cdot) \in \mathcal{Q}(\mathcal{Y}_n|\mathcal{Y}_{0,n-1} \times \mathcal{X}_n) : n = 0, 1, \dots\}$ is weakly continuous (see Definition A.4), i.e., for each sequence $\{(y^{n-1,(\alpha)}, x_n^{(\alpha)}) : \alpha = 1, \dots\} \subset \mathcal{Y}_{0,n-1} \times \mathcal{X}_n$ such that $(y^{n-1,(\alpha)}, x_n^{(\alpha)}) \rightarrow (y^{n-1,(o)}, x_n^{(o)})$, then $\lim_{\alpha \rightarrow \infty} \int_{\mathcal{Y}_n} g(y_n) q_n(dy_n|y^{n-1,(\alpha)}, x_n^{(\alpha)}) = \int_{\mathcal{Y}_n} g(y_n) q_n(dy_n|y^{n-1,(o)}, x_n^{(o)})$, for all bounded continuous functions $g(\cdot) \in BC(\mathcal{Y}_n)$. Hence, no requirement is imposed on the distribution of $\{P_{V_n}(\cdot) \in \mathcal{M}_1(\mathcal{V}_n) : n = 0, 1, \dots\}$.

On the other hand, consider the special case of an additive channel, of the form

$$Y_n = \bar{h}_n(Y^{n-1}, X_n) + V_n, \quad n = 0, 1, \dots$$

where $P_{V_n}(dv_n)$ is assumed to have a density, $\bar{p}(v_n)$, i.e., $P_{V_n}(dv_n) = \bar{p}(v_n)dv_n$, $n = 0, 1, \dots$. Then $\{q_n(\cdot|\cdot, \cdot) \in \mathcal{Q}(\mathcal{Y}_n|\mathcal{Y}_{0,n-1} \times \mathcal{X}_n) : n = 0, 1, \dots\}$ is strongly continuous if for each n , $\bar{h}(\cdot, \cdot)$ is continuous on $\mathcal{Y}_{0,n-1} \times \mathcal{X}_n$ and $\bar{p}(\cdot)$ is continuous on \mathcal{V}_n , for $n = 0, 1, \dots$.

Clearly, when proving properties of mutual information or directed information, weak continuity is more general (less restrictive), compared to strong continuity, which by definition rules out many interesting application examples.

Next, we state the main theorem which is also used to show lower semicontinuity of directed information. The theorem consists of two parts depending on whether, A) $\mathcal{Y}_{0,n}$ is compact and $p_n(dx_n|\cdot, \cdot)$ as a function of $(x^{n-1}, y^{n-1}) \in \mathcal{X}_{0,n-1} \times \mathcal{Y}_{0,n-1}$ is weakly continuous, and B) $\mathcal{X}_{0,n}$ is compact and $q_n(dy_n|\cdot, \cdot)$ as a function of $(y^{n-1}, x^n) \in \mathcal{Y}_{0,n-1} \times \mathcal{X}_{0,n}$ is weakly continuous. In applications of information theory either one of them or both may be required, depending on the context of the application considered.

Theorem III.5.

Part A. For each $n \in \mathbb{N}_0$, let $\mathcal{Y}_{0,n}$ be a compact Polish space, $\mathcal{X}_{0,n}$ a Polish space, and assume the collection of conditional distributions $\{p_n(\cdot|\cdot, \cdot) \in \mathcal{Q}(\mathcal{X}_n|\mathcal{X}_{0,n-1} \times \mathcal{Y}_{0,n-1}) : n \in \mathbb{N}_0\}$ satisfy the following condition.

CA: For all $g(\cdot) \in BC(\mathcal{X}_{0,n})$, the function

$$(x^{n-1}, y^{n-1}) \in \mathcal{X}_{0,n-1} \times \mathcal{Y}_{0,n-1} \mapsto \int_{\mathcal{X}_n} g(x) p_n(dx|x^{n-1}, y^{n-1}) \in \mathbb{R} \quad (\text{III.22})$$

is continuous jointly in the variables $(x^{n-1}, y^{n-1}) \in \mathcal{X}_{0,n-1} \times \mathcal{Y}_{0,n-1}$.

Then the following hold.

A1) Let $\bar{P}_{0,n}(\cdot|y^{n-1}) \in \mathcal{M}_1^{\mathbf{C}^1}(\mathcal{X}_{0,n})$ and consider a sequence of forward channels $\{\bar{Q}_{0,n}^\alpha(\cdot|x^n) : \alpha = 1, 2, \dots\} \subset \mathcal{M}_1^{\mathbf{C}^2}(\mathcal{Y}_{0,n})$. Then the sequence of joint measures $\{(\bar{P}_{0,n} \otimes \bar{Q}_{0,n}^\alpha) : \alpha = 1, 2, \dots\}$ converges weakly to a joint measure $P^o(dx^n, dy^n)$, that is,

$$(\bar{P}_{0,n} \otimes \bar{Q}_{0,n}^\alpha)(dx^n, dy^n) \xrightarrow{w} P^o(dx^n, dy^n) = (\bar{P}_{0,n} \otimes \bar{Q}_{0,n}^o)(dx^n, dy^n) \in \mathcal{M}_1(\mathcal{X}_{0,n} \times \mathcal{Y}_{0,n}) \quad (\text{III.23})$$

where the joint measure $P^o(dx^n, dy^n)$ corresponds to the same backward channel $\bar{P}_{0,n}(\cdot|y^{n-1}) \in \mathcal{M}_1^{\mathbf{C}^1}(\mathcal{X}_{0,n})$ and a forward channel $\bar{Q}_{0,n}^o(\cdot|x^n) \in \mathcal{M}_1(\mathcal{Y}_{0,n})$ (i.e., not necessarily in $\mathcal{M}_1^{\mathbf{C}^2}(\mathcal{Y}_{0,n})$). Equivalently, $\{(\bar{P}_{0,n} \otimes \bar{Q}_{0,n}^\alpha) : \alpha = 1, 2, \dots\}$ is relatively or weakly compact.

Moreover, the corresponding sequence of marginal measures $\{\nu_{0,n}^\alpha(\cdot) \in \mathcal{M}_1(\mathcal{Y}_{0,n}) : \alpha = 1, 2, \dots\}$ on $\mathcal{Y}_{0,n}$ and $\{\mu_{0,n}^\alpha(\cdot) \in \mathcal{M}_1(\mathcal{X}_{0,n}) : \alpha = 1, 2, \dots\}$ on $\mathcal{X}_{0,n}$, converges weakly, that is,

$$\nu_{0,n}^\alpha(dy^n) \xrightarrow{w} \nu_{0,n}^o(dy^n) \text{ and } \mu_{0,n}^\alpha(dx^n) \xrightarrow{w} \mu_{0,n}^o(dx^n)$$

where $\nu_{0,n}^o(\cdot) \in \mathcal{M}_1(\mathcal{Y}_{0,n})$ and $\mu_{0,n}^o(\cdot) \in \mathcal{M}_1(\mathcal{X}_{0,n})$ are the marginals of the joint measure in (III.23).

A2) The set of measures $\bar{P}_{0,n}(\cdot|y^{n-1}) \in \mathcal{M}_1^{\mathbf{C}^1}(\mathcal{X}_{0,n})$ is uniformly tight.

A3) The set of measures $\bar{Q}_{0,n}^\alpha(\cdot|x^n) \in \mathcal{M}_1^{\mathbf{C}^2}(\mathcal{Y}_{0,n})$ is relatively compact.

A4) Let $\bar{P}_{0,n}(\cdot|y^{n-1}) \in \mathcal{M}_1^{\mathbf{C}^1}(\mathcal{X}_{0,n})$, $\{\bar{Q}_{0,n}^\alpha(\cdot|x^n) : \alpha = 1, 2, \dots\} \subset \mathcal{M}_1^{\mathbf{C}^2}(\mathcal{Y}_{0,n})$, where $\{\nu_{0,n}^\alpha(\cdot) \in \mathcal{M}_1(\mathcal{Y}_{0,n}) : \alpha = 1, 2, \dots\}$ are the marginals of $\{(\bar{P}_{0,n} \otimes \bar{Q}_{0,n}^\alpha)(dx^n, dy^n) \in \mathcal{M}_1(\mathcal{X}_{0,n} \times \mathcal{Y}_{0,n}) : \alpha = 1, 2, \dots\}$. Then

$$\bar{\Pi}_{0,n}^\alpha(dx^n, dy^n) \equiv \bar{P}_{0,n}(dx^n|dy^{n-1}) \otimes \nu_{0,n}^\alpha(dy^n) \xrightarrow{w} \bar{P}_{0,n}(dx^n|dy^{n-1}) \otimes \nu_{0,n}^o(dy^n) \equiv \bar{\Pi}_{0,n}^o(dx^n, dy^n)$$

where $\nu_{0,n}^o(\cdot) \in \mathcal{M}_1(\mathcal{Y}_{0,n})$ is the weak limit of the marginal in (III.23).

Part B. For each $n \in \mathbb{N}_0$, let $\mathcal{X}_{0,n}$ be a compact Polish space, $\mathcal{Y}_{0,n}$ a Polish space, and assume the collection of conditional distributions $\{q_n(\cdot|\cdot, \cdot) \in \mathcal{Q}(\mathcal{Y}_n|\mathcal{Y}_{0,n-1} \times \mathcal{X}_{0,n}) : n \in \mathbb{N}_0\}$ satisfy the following condition.

CB: For all $h(\cdot) \in BC(\mathcal{Y}_{0,n})$, the function

$$(x^n, y^{n-1}) \in \mathcal{X}_{0,n} \times \mathcal{Y}_{0,n-1} \mapsto \int_{\mathcal{Y}_n} h(y) q_n(dy|y^{n-1}, x^n) \in \mathbb{R} \quad (\text{III.24})$$

is continuous jointly in the variables $(x^n, y^{n-1}) \in \mathcal{X}_{0,n} \times \mathcal{Y}_{0,n-1}$.

Then the following hold.

B1) Let $\bar{Q}_{0,n}(\cdot|x^n) \in \mathcal{M}_1^{\mathbf{C}^2}(\mathcal{Y}_{0,n})$ and consider a sequence of backward channels $\{\bar{P}_{0,n}^\alpha(\cdot|y^{n-1}) : \alpha = 1, 2, \dots\} \subset \mathcal{M}_1^{\mathbf{C}^1}(\mathcal{X}_{0,n})$. Then, the joint measures $\{(\bar{P}_{0,n}^\alpha \otimes \bar{Q}_{0,n}) : \alpha = 1, 2, \dots\}$ converges weakly to a joint measure $P^o(dx^n, dy^n)$, that is,

$$(\bar{P}_{0,n}^\alpha \otimes \bar{Q}_{0,n})(dx^n, dy^n) \xrightarrow{w} P^o(dx^n, dy^n) = (\bar{P}_{0,n}^o \otimes \bar{Q}_{0,n})(dx^n, dy^n) \in \mathcal{M}_1(\mathcal{X}_{0,n} \times \mathcal{Y}_{0,n}) \quad (\text{III.25})$$

where the joint measure $P^o(dx^n, dy^n)$ corresponds to the same forward channel $\bar{Q}_{0,n}(\cdot|x^n) \in \mathcal{M}_1^{\mathbf{C}^2}(\mathcal{Y}_{0,n})$ and a backward channel $\bar{P}_{0,n}^o(\cdot|y^{n-1}) \in \mathcal{M}_1(\mathcal{X}_{0,n})$ (i.e., not necessarily in $\mathcal{M}_1^{\mathbf{C}^1}(\mathcal{X}_{0,n})$). Equivalently, $\{(\bar{P}_{0,n}^\alpha \otimes \bar{Q}_{0,n}) : \alpha = 1, 2, \dots\}$ is relatively or weakly compact.

Moreover, the corresponding sequence of marginal measures $\{\nu_{0,n}^\alpha(\cdot) \in \mathcal{M}_1(\mathcal{Y}_{0,n}) : \alpha = 1, 2, \dots\}$ on $\mathcal{Y}_{0,n}$ and $\{\mu_{0,n}^\alpha(\cdot) \in \mathcal{M}_1(\mathcal{X}_{0,n}) : \alpha = 1, 2, \dots\}$ on $\mathcal{X}_{0,n}$, converges weakly, that is,

$$\nu_{0,n}^\alpha(dy^n) \xrightarrow{w} \nu_{0,n}^o(dy^n) \text{ and } \mu_{0,n}^\alpha(dx^n) \xrightarrow{w} \mu_{0,n}^o(dx^n)$$

where $\nu_{0,n}^o(\cdot) \in \mathcal{M}_1(\mathcal{Y}_{0,n})$ and $\mu_{0,n}^o(\cdot) \in \mathcal{M}_1(\mathcal{X}_{0,n})$ are the marginals of (III.25).

B2) The set of measures $\vec{Q}_{0,n}(\cdot|x^n) \in \mathcal{M}_1^{\mathbf{C}^2}(\mathcal{Y}_{0,n})$ is uniformly tight.

B3) The set of measures $\vec{P}_{0,n}(\cdot|y^{n-1}) \in \mathcal{M}_1^{\mathbf{C}^1}(\mathcal{X}_{0,n})$ is relatively compact.

B4) Let $\vec{Q}_{0,n}(\cdot|x^n) \in \mathcal{M}_1^{\mathbf{C}^2}(\mathcal{Y}_{0,n})$, $\{\vec{P}_{0,n}^\alpha(\cdot|y^{n-1}) : \alpha = 1, 2, \dots\} \subset \mathcal{M}_1^{\mathbf{C}^1}(\mathcal{X}_{0,n})$, where $\{\mu_{0,n}^\alpha(\cdot) \in \mathcal{M}_1(\mathcal{X}_{0,n}) : \alpha = 1, 2, \dots\}$ are the marginals of $\{(\vec{P}_{0,n}^\alpha \otimes \vec{Q}_{0,n})(dx^n, dy^n) \in \mathcal{M}_1(\mathcal{X}_{0,n} \times \mathcal{Y}_{0,n}) : \alpha = 1, 2, \dots\}$. Then

$$\overleftarrow{\Pi}^\alpha(dx^n, dy^n) \equiv \vec{Q}_{0,n}(dy^n|dx^n) \otimes \mu_{0,n}^\alpha(dx^n) \xrightarrow{w} \vec{Q}_{0,n}(dy^n|dx^n) \otimes \mu_{0,n}^o(dx^n) \equiv \overleftarrow{\Pi}^o(dx^n, dy^n)$$

where $\mu_{0,n}^o(\cdot) \in \mathcal{M}_1(\mathcal{X}_{0,n})$ is the weak limit of the marginal in (III.25).

Proof: See Appendix C. □

Note that additional conditions are required to show that the limiting joint distribution (III.23) (respectively, (III.25)) corresponds to a $\vec{Q}^o(\cdot|x^n) \in \mathcal{M}_1^{\mathbf{C}^2}(\mathcal{Y}_{0,n})$ (respectively, $\vec{P}^o(\cdot|y^{n-1}) \in \mathcal{M}_1^{\mathbf{C}^1}(\mathcal{X}_{0,n})$). Conditions for this to hold are given in Section III-D.

Below, we illustrate analogies and differences between Theorem III.5 and currently known results regarding mutual information found in [28], [47]. To this end, consider Part B., B1). If we use mutual information [28, Lemma 2], then the sequence of joint measures is defined by $P_{X^n, Y^n}^\alpha(dx^n, dy^n) \triangleq P_{Y^n|X^n}(dy^n|x^n) \otimes P_{X^n}^\alpha(dx^n)$, and showing weak convergence of this family is much simpler compared to the sequence of joint distributions $(\vec{P}_{0,n}^\alpha \otimes \vec{Q}_{0,n})(dx^n, dy^n)$, because $P_{X^n}(dx^n)$ is not conditioned on $y^n \in \mathcal{Y}_{0,n}$. Clearly, if the mapping $x^n \rightarrow P_{Y^n|X^n}(\cdot|x^n)$ is weakly continuous (i.e., special case of III.24), and $P_{X^n}^\alpha(dx^n)$ converges weakly to $P_{X^n}^o(dx^n)$, then $P_{X^n, Y^n}^\alpha(dx^n, dy^n)$ converges weakly to $P_{Y^n|X^n}(dy^n|x^n) \otimes P_{X^n}^o(dx^n) = P_{X^n, Y^n}^o(dx^n, dy^n)$, and so does its marginal on $\mathcal{Y}_{0,n}$. On the other hand, if we use directed information, then the joint measure $P_{X^n, Y^n}(dx^n, dy^n) \triangleq \otimes_{i=0}^n P_{Y_i|Y^{i-1}, X^i}(dy_i|y^{i-1}, x^i) \otimes P_{X_i|X^{i-1}, Y^{i-1}}(dx_i|x^{i-1}, y^{i-1})$ involves an $(n+1)$ -fold compound probability distribution defined by (I.4), and $P_{X_i|X^{i-1}, Y^{i-1}}(\cdot|\cdot, \cdot)$ is a function of $y^{n-1} \in \mathcal{Y}_{0,n-1}$, hence a significant level of additional complexity incurs, compared to mutual information. Nevertheless, condition CB is the natural generalization to causally conditioned $(n+1)$ -fold compound probability distributions of the weak continuity of the mapping $x^n \rightarrow P_{Y^n|X^n}(\cdot|x^n)$, assumed for the mutual information by Csiszár in [28].

Theorem III.5 is important for several extremum problems involving directed information. Such applications are discussed in the next section.

D. Applications of Theorem III.5

In this section, we discuss applications of Theorem III.5 to the extremum problems of feedback capacity and nonanticipative RDF, defined by (I.5) and (I.7), respectively.

Existence of optimal channel input distribution for channels with memory and feedback. Consider extremum problems of capacity of channels with memory and feedback defined by (I.5), without any transmission cost constraint. The aim is to show existence of a channel input conditional distribution $\vec{P}(\cdot|y^{n-1}) \in \mathcal{M}_1^{\mathbf{C}^1}(\mathcal{X}_{0,n})$, $y^{n-1} \in \mathcal{Y}_{0,n-1}$, which achieves the supremum of directed information. To show that such a conditional distribution exists, it is sufficient to show compactness of the set of channel input conditional distributions (i.e., this set is closed and uniformly tight) and upper semicontinuity (or continuity) of $\mathbb{I}_{X^n \rightarrow Y^n}(\vec{P}_{0,n}, \vec{Q}_{0,n})$, with respect to $\vec{P}(\cdot|y^{n-1}) \in \mathcal{M}_1^{\mathbf{C}^1}(\mathcal{X}_{0,n})$ for a fixed channel $\vec{Q}_{0,n}(\cdot|x^n) \in \mathcal{M}_1^{\mathbf{C}^2}(\mathcal{Y}_{0,n})$. Since Theorem III.13, Part A. A2) uniform tightness of $\vec{P}(\cdot|y^{n-1}) \in \mathcal{M}_1^{\mathbf{C}^1}(\mathcal{X}_{0,n})$, it remains to show this set is closed. This is shown in the next lemma, by introducing additional assumptions.

Lemma III.6. (Compactness of $\vec{P}(\cdot|y^{n-1}) \in \mathcal{M}_1^{\mathbf{C}^1}(\mathcal{X}_{0,n})$)

Suppose the conditions of Theorem III.5, Part A. hold, and for each compact subset $K_{0,i-1} \subset \mathcal{X}_{0,i-1}$, and

each $h_i(\cdot) \in BC(\mathcal{X}_i)$,

$$\lim_{\alpha \rightarrow \infty} \sup_{x^{i-1} \in K_{0,i-1}} \left| \int_{\mathcal{X}_i} h_i(x) p_i^\alpha(dx|x^{i-1}, y^{i-1}) - \int_{\mathcal{X}_i} h_i(x) p_i(dx|x^{i-1}, y^{i-1}) \right| = 0, \quad i = 0, 1, \dots, n \quad (\text{III.26})$$

Then,

$$\overleftarrow{P}_{0,n}^\alpha(dx^n|y^{n-1}) \xrightarrow{w} \overleftarrow{P}_{0,n}^o(dx^n|y^{n-1}) \text{ for each } y^{n-1} \in \mathcal{Y}_{0,n-1} \quad (\text{III.27})$$

i.e., the set $\overleftarrow{P}(\cdot|y^{n-1}) \in \mathcal{M}_1^{\mathbf{C1}}(\mathcal{X}_{0,n})$ is closed with respect to the topology of weak convergence, and moreover, it is also compact (i.e., closed and tight).

Proof: See Appendix D. □

Remark III.7. (Compactness of channel input distributions with transmission cost)

In the presence of power constraints $\overleftarrow{P}(\cdot|y^{n-1}) \in \mathcal{P}_{0,n}(P) \subset \mathcal{Q}^{\mathbf{C1}}(\mathcal{X}_{0,n}|\mathcal{Y}_{0,n-1})$, by Prohorov's theorem (Appendix A, Theorem A.3), to show compactness of $\mathcal{P}_{0,n}(P)$, it is sufficient to show that this set is closed and uniformly tight. By invoking Lemma III.6, it suffices to show $\mathcal{P}_{0,n}(P)$ is a closed subset of the weakly compact set $\mathcal{M}^{\mathbf{C1}}(\mathcal{X}_{0,n})$ (as a closed subset of a weakly compact set is weakly compact).

Existence of optimal reproduction distribution of nonanticipative RDF. Consider a special case of extremum problems of nonanticipative RDF defined by (I.7), with distortion constraint defined by (I.8), when the source distribution is causally independent of past reproduction symbols, that is, $p_i(dx_i|x^{i-1}, y^{i-1}) = \mu_i(dx_i|x^{i-1})$, -a.a. (x^{i-1}, y^{i-1}) , $i = 0, 1, \dots, n$. Then, the finite time version of (I.7) is given by

$$R_{0,n}^{na}(D) = \inf_{\overrightarrow{Q}_{0,n}(dy^n|x^n) \in \mathcal{Q}_{0,n}(D)} \int_{\mathcal{X}_{0,n} \times \mathcal{Y}_{0,n}} \log \left(\frac{d\overrightarrow{Q}_{0,n}(\cdot|x^n)}{d\nu_{0,n}(\cdot)}(y^n) \right) \overrightarrow{Q}_{0,n}(dy^n|x^n) \otimes \mu_{0,n}(dx^n) \quad (\text{III.28})$$

$$\equiv \inf_{\overrightarrow{Q}_{0,n}(dy^n|x^n) \in \mathcal{Q}_{0,n}(D)} \mathbb{I}_{X^n \rightarrow Y^n}(\mu_{0,n}, \overrightarrow{Q}_{0,n}) \quad (\text{III.29})$$

where $\mu_{0,n}(dx^n) = \otimes_{i=0}^n \mu_i(dx_i|x^{i-1})$, $\nu_{0,n}(dy^n) = \int_{\mathcal{X}_{0,n}} \overrightarrow{Q}_{0,n}(dy^n|x^n) \otimes \mu_{0,n}(dx^n)$, and the fidelity constraint is defined by

$$\mathcal{Q}_{0,n}(D) \triangleq \left\{ \overrightarrow{Q}_{0,n}(dy^n|x^n) \in \mathcal{M}_1^{\mathbf{C2}}(\mathcal{Y}_{0,n}) : \frac{1}{n+1} \int_{\mathcal{X}_{0,n} \times \mathcal{Y}_{0,n}} d_{0,n}(x^n, y^n) \overrightarrow{Q}_{0,n}(dy^n|x^n) \otimes \mu_{0,n}(dx^n) \leq D \right\}, \quad D \geq 0 \quad (\text{III.30})$$

and $d_{0,n} : \mathcal{X}_{0,n} \times \mathcal{Y}_{0,n} \mapsto [0, \infty]$, $d_{0,n}(x^n, y^n) \triangleq \sum_{i=0}^n \rho_i(x^i, y^i)$ is a measurable function denoting the distortion function of reconstructing x_i by y_i , $i = 0, 1, \dots, n$.

The information nonanticipative RDF defined by (III.28), (III.30), is an equivalent notion to the nonanticipative epsilon entropy investigated by Gorbunov and Pinsker [43] (see Charalambous et al. in [23] for relations to filtering theory).

The aim is to show existence of a conditional distribution $\overrightarrow{Q}_{0,n}(\cdot|x^n) \in \mathcal{M}_1^{\mathbf{C2}}(\mathcal{Y}_{0,n})$, which achieves infimum in (III.28). Since $\mathcal{Q}_{0,n}(D) \subset \mathcal{M}_1^{\mathbf{C2}}(\mathcal{Y}_{0,n})$, to show such a conditional distribution exists, it is sufficient to show compactness of $\mathcal{M}_1^{\mathbf{C2}}(\mathcal{Y}_{0,n})$ (closed and uniformly tight), the set $\mathcal{Q}_{0,n}(D)$ is a closed subset of $\mathcal{M}_1^{\mathbf{C2}}(\mathcal{Y}_{0,n})$, and $\mathbb{I}_{X^n \rightarrow Y^n}(\overleftarrow{P}_{0,n}, \overrightarrow{Q}_{0,n})$ is lower semicontinuous with respect to $\overrightarrow{Q}(\cdot|x^n) \in \mathcal{M}_1^{\mathbf{C2}}(\mathcal{Y}_{0,n})$, for a fixed $\mu_{0,n}(dx^n) \in \mathcal{M}_1(\mathcal{X}_{0,n})$. This can be done by invoking a combination of the assumptions of Theorem III.5 Part A. or Part B., depending on whether $\mathcal{Y}_{0,n}$ is compact and $\mathcal{X}_{0,n}$ is arbitrary or $\mathcal{X}_{0,n}$ is compact and $\mathcal{Y}_{0,n}$ is arbitrary, respectively. Since in general, $\mathcal{Y}_{0,n} \subseteq \mathcal{X}_{0,n}$, it is more appropriate to assume $\mathcal{Y}_{0,n}$ is compact.

Lemma III.8. (Compactness of $\vec{Q}(\cdot|x^n) \in \mathcal{Q}_{0,n}(D)$)

(1) Suppose $\mathcal{X}_{0,n}$ are Polish spaces, and $\mathcal{Y}_{0,n}$ is compact, the sequence $\{q_n(\cdot|\cdot, \cdot) \in \mathcal{Q}(\mathcal{Y}_n|\mathcal{Y}_{0,n-1} \times \mathcal{X}_{0,n}) : n \in \mathbb{N}_0\}$ is weakly continuous, i.e., it satisfies (III.24), and for each compact subset $\Phi_{0,i-1} \subset \mathcal{Y}_{0,i-1}$, and each $h_i(\cdot) \in BC(\mathcal{Y}_i)$,

$$\lim_{\alpha \rightarrow \infty} \sup_{y^{i-1} \in \Phi_{0,i-1}} \left| \int_{\mathcal{Y}_i} h_i(x) q_i^\alpha(dy|y^{i-1}, x^i) - \int_{\mathcal{Y}_i} h_i(y) q_i(dy|y^{i-1}, x^i) \right| = 0, \quad \forall x^i \in \mathcal{X}_{0,i}, \quad i = 0, 1, \dots, n. \quad (\text{III.31})$$

Then,

$$\vec{Q}_{0,n}^\alpha(dy^n|x^n) \xrightarrow{w} \vec{Q}_{0,n}^\circ(dy^n|x^n) \text{ for each } x^n \in \mathcal{X}_{0,n}$$

i.e., the set $\mathcal{M}_1^{\mathcal{C}^2}(\mathcal{Y}_{0,n})$ is closed with respect to the topology of weak convergence. Moreover, $\mathcal{M}_1^{\mathcal{C}^2}(\mathcal{Y}_{0,n})$ is compact (closed and tight).

(2) In addition, suppose the distortion function $d_{0,n}(x^n, \cdot) : \mathcal{X}_{0,n} \times \mathcal{Y}_{0,n} \mapsto [0, \infty]$ is Borel measurable relative to $\mathcal{B}(\mathcal{X}_{0,n}) \otimes \mathcal{B}(\mathcal{Y}_{0,n})$ and continuous on $y^n \in \mathcal{Y}_{0,n}$.

Then, the fidelity set $\mathcal{Q}_{0,n}(D)$ is compact (it is a closed subset of the compact set $\mathcal{M}_1^{\mathcal{C}^2}(\mathcal{Y}_{0,n})$).

Proof: See Appendix E. □

Theorem III.5 gives the flexibility of choosing either $\mathcal{X}_{0,n}$ or $\mathcal{Y}_{0,n}$ to be compact; it has several applications in other extremum problems of directed information. In the following remark, we discuss such applications.

Remark III.9. (Additional Applications)

- (1) Consider extremum problems of capacity for a class of channels with memory and feedback, such as, arbitrary varying channels [28]. Such problems are defined by the max-min operations of directed information, where the minimizer is over the class of channels [48]. To investigate such capacity problems one has to establish coding theorems, and showing compactness over the class of channel conditional distributions, in addition to channel input distributions is very helpful. Theorem III.5, Part B., B3) gives conditions of weak compactness of channels $\vec{Q}_{0,n}(\cdot|x^n) \in \mathcal{M}_1^{\mathcal{C}^2}(\mathcal{Y}_{0,n})$.
- (2) Consider extremum problems of sequential or nonanticipative lossy data compression for a class of sources. Then such problems are defined by mini-max operations of directed information, where the maximizer is over the class of source distributions [49]. To investigate such data compression problems, one has to establish coding theorems, and to show compactness over the class of source distributions, in addition to the reproduction distributions, Theorem III.5, Part A., A3) is crucial.

E. Lower Semicontinuity of Directed Information

We are now ready to utilize the results of Theorem III.5, to show lower semicontinuity of directed information $I(X^n \rightarrow Y^n) \equiv \mathbb{I}_{X^n \rightarrow Y^n}(\vec{P}_{0,n}, \vec{Q}_{0,n})$. This may be viewed as a generalization of lower semicontinuity of mutual information $I(X^n; Y^n) \equiv \mathbb{I}_{X^n, Y^n}(P_{X^n}, Q_{Y^n|X^n})$, with respect to P_{X^n} for fixed $Q_{Y^n|X^n}$, and with respect to $Q_{Y^n|X^n}$ for fixed P_{X^n} .

Theorem III.10. (Lower semicontinuity)

1) Suppose the conditions in Theorem III.5, Part A., hold.

For fixed $\vec{P}_{0,n}(\cdot|y^{n-1}) \in \mathcal{M}_1^{\mathcal{C}^1}(\mathcal{X}_{0,n})$, if the family $\mathcal{M}_1^{\mathcal{C}^2}(\mathcal{Y}_{0,n})$ is closed (i.e., $\{\vec{Q}_{0,n}^\alpha(\cdot|x^n) : \alpha = 1, 2, \dots\} \in \mathcal{M}_1^{\mathcal{C}^2}(\mathcal{Y}_{0,n})$ converges weakly to $\vec{Q}_{0,n}^\circ(\cdot|x^n) \in \mathcal{M}_1^{\mathcal{C}^2}(\mathcal{Y}_{0,n})$) then

$$\mathbb{I}_{X^n \rightarrow Y^n}(\vec{P}_{0,n}, \vec{Q}_{0,n}^\circ) \leq \liminf_{\alpha \rightarrow \infty} \mathbb{I}_{X^n \rightarrow Y^n}(\vec{P}_{0,n}, \vec{Q}_{0,n}^\alpha)$$

i.e., $\mathbb{I}_{X^n \rightarrow Y^n}(\cdot, \vec{Q}_{0,n})$ is lower semicontinuous on $\vec{Q}_{0,n}(\cdot|x^n) \in \mathcal{M}_1^{\mathcal{C}^2}(\mathcal{Y}_{0,n})$.

2) Suppose the conditions in Theorem III.5, Part B., hold.

For fixed $\vec{Q}_{0,n}(\cdot|x^n) \in \mathcal{M}_1^{\mathbf{C}^2}(\mathcal{Y}_{0,n})$, if the family $\mathcal{M}_1^{\mathbf{C}^1}(\mathcal{X}_{0,n})$ is closed (i.e., $\{\bar{P}_{0,n}^\alpha(\cdot|y^{n-1}) : \alpha = 1, 2, \dots\} \in \mathcal{M}_1^{\mathbf{C}^1}(\mathcal{X}_{0,n})$ converges weakly to $\bar{P}_{0,n}^o(\cdot|y^{n-1}) \in \mathcal{M}_1^{\mathbf{C}^1}(\mathcal{X}_{0,n})$) then

$$\mathbb{I}_{X^n \rightarrow Y^n}(\bar{P}_{0,n}^o, \vec{Q}_{0,n}) \leq \liminf_{\alpha \rightarrow \infty} \mathbb{I}_{X^n \rightarrow Y^n}(\bar{P}_{0,n}^\alpha, \vec{Q}_{0,n})$$

i.e., $\mathbb{I}_{X^n \rightarrow Y^n}(\bar{P}_{0,n}, \cdot)$ is lower semicontinuous on $\bar{P}_{0,n}(\cdot|y^{n-1}) \in \mathcal{M}_1^{\mathbf{C}^1}(\mathcal{X}_{0,n})$.

Proof: See Appendix F. □

Recall that conditions for the sets $\mathcal{M}_1^{\mathbf{C}^1}(\mathcal{X}_{0,n})$, $\mathcal{M}_1^{\mathbf{C}^2}(\mathcal{Y}_{0,n})$ to be closed are given in Lemma III.6 and Lemma III.8, respectively.

Comparing Theorem III.10, 1), with the lower semicontinuity of mutual information $I(X^n; Y^n) \equiv \mathbb{I}_{X^n; Y^n}(P_{X^n}, Q_{Y^n|X^n})$, it is clear that directed information requires additional assumptions for its derivation (e.g., those given in Theorem III.5).

Theorem III.5 together with Theorem III.10 are important to establish existence of the optimal reproduction distribution for the nonanticipative rate distortion functions defined by (I.7) [23], [42] (by utilizing Weierstrass' Theorem) and in general extremum problems of directed information involving minimization over $\vec{Q}_{0,n}(\cdot|x^n)$ in some subset of $\mathcal{M}_1^{\mathbf{C}^2}(\mathcal{Y}_{0,n})$. This is formally stated in the next theorem.

Theorem III.11. (Existence of information nonanticipative RDF)

Under the conditions of Lemma III.8 and Theorem III.10, the infimum over $\vec{Q}_{0,n}(\cdot|x^n) \in \mathcal{Q}_{0,n}(D)$ in $R_{0,n}^{na}(D)$, defined by (III.28), is achieved by some $\vec{Q}_{0,n}^*(\cdot|x^n) \in \mathcal{Q}_{0,n}(D)$.

F. Continuity of Directed Information

Many problems in information theory involve extremum problems defined as maximizations of directed information, with respect to the feedback channels $\{p_i(dx_i|x^{i-1}, y^{i-1}) \in \mathcal{M}_1(\mathcal{X}_i) : i = 0, 1, \dots, n\}$, such as, extremum problems of feedback capacity of channels with memory with transmission cost constraint defined by (I.5). For such problems it is desirable to have upper semicontinuity of directed information with respect to $\bar{P}_{0,n}(\cdot|y^{n-1}) \in \mathcal{M}_1^{\mathbf{C}^1}(\mathcal{X}_{0,n})$. Since by Theorem III.10, directed information is lower semicontinuous with respect to $\bar{P}_{0,n}(\cdot|y^{n-1}) \in \mathcal{M}_1^{\mathbf{C}^1}(\mathcal{X}_{0,n})$, to investigate extremum problems involving feedback capacity (maximization problems), it is sufficient to show continuity of the functional $\mathbb{I}_{X^n \rightarrow Y^n}(\bar{P}_{0,n}, \vec{Q}_{0,n})$ with respect to $\bar{P}_{0,n}(\cdot|y^{n-1}) \in \mathcal{M}_1^{\mathbf{C}^1}(\mathcal{X}_{0,n})$ for a fixed $\vec{Q}_{0,n}(\cdot|x^n) \in \mathcal{M}_1^{\mathbf{C}^2}(\mathcal{Y}_{0,n})$. Continuity of mutual information based on single letter expression is shown in [28, Lemma 7], and under weaker conditions in [29, Theorem 3.2]. Here, we show continuity of directed information by following the procedure in [29], generalized to the directed information functional $\mathbb{I}_{X^n \rightarrow Y^n}(\bar{P}_{0,n}, \vec{Q}_{0,n})$. First, we shall need the following Lemma.

Lemma III.12.

For a given $\bar{P}_{0,n}(\cdot|\cdot) \in \mathcal{Q}^{\mathbf{C}^1}(\mathcal{X}_{0,n}|\mathcal{Y}_{0,n-1})$ and $\vec{Q}_{0,n}(\cdot|\cdot) \in \mathcal{Q}^{\mathbf{C}^2}(\mathcal{Y}_{0,n}|\mathcal{X}_{0,n})$ define

$$|\mathbb{I}_{X^n \rightarrow Y^n}|(\bar{P}_{0,n}, \vec{Q}_{0,n}) \triangleq \int_{\mathcal{X}_{0,n} \times \mathcal{Y}_{0,n}} \left| \log \left(\frac{d(\bar{P}_{0,n} \otimes \vec{Q}_{0,n})}{d(\bar{P}_{0,n} \otimes \nu_{0,n})} \right) \right| d(\bar{P}_{0,n} \otimes \vec{Q}_{0,n}).$$

Then the following inequalities hold.

$$\mathbb{I}_{X^n \rightarrow Y^n}(\bar{P}_{0,n}, \vec{Q}_{0,n}) \leq |\mathbb{I}_{X^n \rightarrow Y^n}|(\bar{P}_{0,n}, \vec{Q}_{0,n}) \leq \mathbb{I}_{X^n \rightarrow Y^n}(\bar{P}_{0,n}, \vec{Q}_{0,n}) + \frac{2}{e \ln 2}. \quad (\text{III.32})$$

Proof: Recall directed information defined in Remark III.1. Then

$$\begin{aligned} \mathbb{I}_{X^n \rightarrow Y^n}(\overleftarrow{P}_{0,n}, \overrightarrow{Q}_{0,n}) &= \int_{\mathcal{X}_{0,n} \times \mathcal{Y}_{0,n}} \log \left(\frac{d(\overleftarrow{P}_{0,n} \otimes \overrightarrow{Q}_{0,n})}{d(\overleftarrow{P}_{0,n} \otimes \nu_{0,n})} \right) d(\overleftarrow{P}_{0,n} \otimes \overrightarrow{Q}_{0,n}) \\ &= \int_{\mathcal{X}_{0,n} \times \mathcal{Y}_{0,n}} \log \left(\frac{d(\overleftarrow{P}_{0,n} \otimes \overrightarrow{Q}_{0,n})}{d(\overleftarrow{P}_{0,n} \otimes \nu_{0,n})} \right) \left(\frac{d(\overleftarrow{P}_{0,n} \otimes \overrightarrow{Q}_{0,n})}{d(\overleftarrow{P}_{0,n} \otimes \nu_{0,n})} \right) d(\overleftarrow{P}_{0,n} \otimes \nu_{0,n}). \end{aligned} \quad (\text{III.33})$$

The first inequality in (III.32) is obvious. To show the second inequality in (III.32), recall the inequality [44, Section 2.3, p. 13] $-\frac{1}{e \ln 2} \leq x \log_2 x$, $x \in [0, \infty)$ ($0 \log 0$ is assumed to be 0). Then,

$$|x \log_2 x| \leq x \log_2 x + \frac{2}{e \ln 2}. \quad (\text{III.34})$$

Using (III.34) in (III.33), with $x = \left(\frac{d(\overleftarrow{P}_{0,n} \otimes \overrightarrow{Q}_{0,n})}{d(\overleftarrow{P}_{0,n} \otimes \nu_{0,n})} \right)$, establishes the second inequality in (III.32). \square

Now, we are ready to state the Theorem, which establishes continuity with respect to weak convergence of $\mathbb{I}_{X^n \rightarrow Y^n}(\overleftarrow{P}_{0,n}, \overrightarrow{Q}_{0,n})$ for a fixed $\overrightarrow{Q}_{0,n}(\cdot|x^n) \in \mathcal{M}_1^{\text{C}^2}(\mathcal{Y}_{0,n})$, as a functional of $\overleftarrow{P}_{0,n}(\cdot|y^{n-1}) \in \mathcal{M}_1^{\text{C}^1}(\mathcal{X}_{0,n})$.

Theorem III.13. (Continuity)

Consider a forward channel $\overrightarrow{Q}_{0,n}(\cdot|x^n) \in \mathcal{M}_1^{\text{C}^2}(\mathcal{Y}_{0,n})$, and a closed family of feedback channels $\overleftarrow{P}_{0,n}(\cdot|y^{n-1}) \in \mathcal{M}_1^{\text{C}^1, \text{cl}}(\mathcal{X}_{0,n}) \subseteq \mathcal{M}_1^{\text{C}^1}(\mathcal{X}_{0,n})$. Suppose the following conditions hold.

- A) There exists a measure $\bar{\nu}_{0,n}(dy^n)$ on $\mathcal{Y}_{0,n}$ such that $\overrightarrow{Q}_{0,n}(\cdot|x^n) \ll \bar{\nu}_{0,n}(dy^n)$ with RND or density $\xi_{\bar{\nu}_{0,n}}(x^n, y^n) \triangleq \frac{d\overrightarrow{Q}_{0,n}(\cdot|x^n)}{d\bar{\nu}_{0,n}(\cdot)}(y^n)$.
 - B) The RND $\xi_{\bar{\nu}_{0,n}}(x^n, y^n)$ is continuous on $\mathcal{X}_{0,n} \times \mathcal{Y}_{0,n}$, and $\xi_{\bar{\nu}_{0,n}}(x^n, y^n) \log \xi_{\bar{\nu}_{0,n}}(x^n, y^n)$ is uniformly integrable over $\left\{ (\bar{\nu}_{0,n} \otimes \overleftarrow{P}_{0,n})(dx^n, dy^n) : \overleftarrow{P}_{0,n}(\cdot|y^{n-1}) \in \mathcal{M}_1^{\text{C}^1, \text{cl}}(\mathcal{X}_{0,n}) \right\}$.
 - C) For a fixed $y^n \in \mathcal{Y}_{0,n}$, the RND $\xi_{\bar{\nu}_{0,n}}(x^n, y^n)$ is uniformly integrable over $\mathcal{M}_1^{\text{C}^1, \text{cl}}(\mathcal{X}_{0,n})$.
- Then, $\mathbb{I}_{X^n \rightarrow Y^n}(\overleftarrow{P}_{0,n}, \overrightarrow{Q}_{0,n})$ as a functional of $\overleftarrow{P}_{0,n}(\cdot|y^{n-1}) \in \mathcal{M}_1^{\text{C}^1, \text{cl}}(\mathcal{X}_{0,n})$ is bounded and weakly continuous over $\mathcal{M}_1^{\text{C}^1, \text{cl}}(\mathcal{X}_{0,n})$, for fixed $\overrightarrow{Q}_{0,n}(\cdot|x^n) \in \mathcal{M}_1^{\text{C}^2}(\mathcal{Y}_{0,n})$.

Proof: The derivation is shown in Appendix G. \square

Note that Theorem III.5 gives conditions for weak compactness of $\overleftarrow{P}_{0,n}(\cdot|y^{n-1}) \in \mathcal{M}_1^{\text{C}^1}(\mathcal{X}_{0,n})$, and Lemma III.6 gives conditions for compactness of $\overleftarrow{P}_{0,n}(\cdot|y^{n-1}) \in \mathcal{M}_1^{\text{C}^1}(\mathcal{X}_{0,n})$. In addition, Theorem III.13 gives conditions of weak continuity of $\mathbb{I}_{X^n \rightarrow Y^n}(\overleftarrow{P}_{0,n}, \overrightarrow{Q}_{0,n})$ with respect to $\overleftarrow{P}_{0,n}(\cdot|y^{n-1}) \in \mathcal{M}_1^{\text{C}^1}(\mathcal{X}_{0,n})$, for fixed $\overrightarrow{Q}_{0,n}(\cdot|x^n) \in \mathcal{M}_1^{\text{C}^2}(\mathcal{Y}_{0,n})$. Hence, sufficient conditions are identified to address existence of solution to the extremum problem of feedback capacity. This is stated in the next theorem.

Theorem III.14. (Existence of information feedback capacity without transmission cost constraint)

Under the conditions of Lemma III.6 and Theorem III.13, the supremum over $\overleftarrow{P}_{0,n}(\cdot|y^{n-1}) \in \mathcal{M}_1^{\text{C}^1}(\mathcal{X}_{0,n})$ in the extremum problem of information feedback capacity

$$C_{0,n}^{\text{fb}} \triangleq \sup_{\{P_{X_i|X^{i-1}, Y^{i-1}}: i=0,1,\dots,n\} \in \mathcal{M}_1^{\text{C}^1}(\mathcal{X}_{0,n})} \frac{1}{n+1} I(X^n \rightarrow Y^n) \quad (\text{III.35})$$

is achieved by some $\overleftarrow{P}_{0,n}^*(\cdot|y^{n-1}) \in \mathcal{M}_1^{\text{C}^1}(\mathcal{X}_{0,n})$.

G. Extension of Directed Information to Arbitrary Number of Sequences of RV's

In this section, we demonstrate how the previous results are easily generalized to three, or more, sequences of RV's. These extensions have implications in communication networks, and in communication with side information at either the transmitter or the receiver [36], [37].

To facilitate the demonstration, first consider the following case.

Case 1: The sequence of RV's $X^n \in \mathcal{X}_{0,n}$ is defined by $X^n = (X^{1,n}, X^{2,n}) \in \mathcal{X}_{0,n}^1 \times \mathcal{X}_{0,n}^2 \equiv \mathcal{X}_{0,n}$, where $X^{1,n} = \{X_i^1 : i = 0, 1, \dots, n\}$ and $X^{2,n} = \{X_i^2 : i = 0, 1, \dots, n\}$.

Then, the two sequences of conditional distributions are $\{p_i(dx_i^1, dx_i^2 | x^{1,i-1}, x^{2,i-1}, y^{i-1}) : i = 0, 1, \dots, n\}$ and $\{q_i(dy_i^1 | y^{1,i-1}, x^{1,i}, x^{2,i}) : i = 0, 1, \dots, n\}$, respectively. Consequently, the constructions of consistent families of conditional distributions, and the results obtained so far, extend naturally to directed information $\mathbb{I}_{(X^{1,n}, X^{2,n}) \rightarrow Y^n}(\overleftarrow{P}_{0,n}, \overrightarrow{Q}_{0,n})$, where $\overleftarrow{P}_{0,n}(dx^{1,n}, dx^{2,n} | y^{n-1}) = \otimes_{i=0}^n p_i(dx_i^1, dx_i^2 | x^{1,i-1}, x^{2,i-1}, y^{i-1})$, and $\overrightarrow{Q}_{0,n}(dy^n | x^{1,n}, x^{2,n}) = \otimes_{i=0}^n q_i(dy_i | y^{1,i-1}, x^{1,i}, x^{2,i})$.

Next, we consider the following case.

Case 2: The sequence of RV's $Y^n \in \mathcal{Y}_{0,n}$ is defined by $Y^n \triangleq (Y^{1,n}, Y^{2,n}) \in \mathcal{Y}_{0,n}^1 \times \mathcal{Y}_{0,n}^2 \equiv \mathcal{Y}_{0,n}$, where $Y^{1,n} = \{Y_i^1 : i = 0, 1, \dots, n\}$ and $Y^{2,n} = \{Y_i^2 : i = 0, 1, \dots, n\}$.

Then, the two sequences of conditional distributions are $\{p_i(dx_i | x^{i-1}, y^{1,i-1}, y^{2,i-1}) : i = 0, 1, \dots, n\}$ and $\{q_i(dy_i^1, dy_i^2 | y^{1,i-1}, y^{2,i-1}, x^i) : i = 0, 1, \dots, n\}$, respectively. Consequently, the constructions of consistent families of conditional distributions, and the results obtained so far, extend naturally to directed information $\mathbb{I}_{X^n \rightarrow (Y^{1,n}, Y^{2,n})}(\overleftarrow{P}_{0,n}, \overrightarrow{Q}_{0,n})$, where $\overleftarrow{P}_{0,n}(dx^n | y^{1,n-1}, y^{2,n-1}) = \otimes_{i=0}^n p_i(dx_i | x^{i-1}, y^{1,i-1}, y^{2,i-1})$, and $\overrightarrow{Q}_{0,n}(dy^{1,n}, dy^{2,n} | x^n) = \otimes_{i=0}^n q_i(dy_i^1, dy_i^2 | y^{1,i-1}, y^{2,i-1}, x^i)$.

Clearly, **Case 1** and **Case 2** can be generalized to an arbitrary number of sequences of RV's.

IV. SEQUENTIAL VARIATIONAL EQUALITIES OF DIRECTED INFORMATION

In this section we derive variational equalities including their sequential versions for directed information. Moreover, we illustrate an application of these variational equalities in feedback capacity computation, by developing the main ingredient of a sequential algorithm using dynamic programming.

The variational equalities of directed information may be viewed as generalizations of the well-known variational equalities of mutual information $I(X^n; Y^n) \equiv \mathbb{I}_{X^n; Y^n}(P_{X^n}, P_{Y^n|X^n})$, expressed as minimizations or maximizations of relative entropy functionals, as follows [25].

Min: Given a channel distribution $P_{Y^n|X^n}(dy^n | x^n)$, a source distribution P_{X^n} , and any arbitrary distribution $V_{Y^n}(dy^n)$ on $\mathcal{Y}_{0,n}$ then

$$\mathbb{I}_{X^n; Y^n}(P_{X^n}, P_{Y^n|X^n}) = \inf_{V_{Y^n}(dy^n) \in \mathcal{M}_1(\mathcal{Y}_{0,n})} \int_{\mathcal{X}_{0,n} \times \mathcal{Y}_{0,n}} \log \left(\frac{dP_{Y^n|X^n}(\cdot | x^n)}{dV_{Y^n}(\cdot)}(y^n) \right) P_{Y^n|X^n}(dy^n | x^n) \otimes P_{X^n}(dx^n) \quad (\text{IV.1})$$

and the infimum is achieved at $V_{Y^n}(dy^n) \equiv P_{Y^n}(dy^n)$ given by

$$P_{Y^n}(dy^n) = \int_{\mathcal{X}_{0,n}} P_{Y^n|X^n}(dy^n | x^n) \otimes P_{X^n}(dx^n). \quad (\text{IV.2})$$

Max: Given a channel distribution $P_{Y^n|X^n}(dy^n | x^n)$, a source distribution $P_{X^n}(dx^n)$, and any arbitrary conditional distribution $V_{X^n|Y^n}(dx^n | y^n)$ on $\mathcal{X}_{0,n}$ parametrized by $y^n \in \mathcal{Y}_{0,n}$ then

$$\mathbb{I}_{X^n; Y^n}(P_{X^n}, P_{Y^n|X^n}) = \sup_{\substack{V_{X^n|Y^n}(dx^n | y^n) \\ \in \mathcal{M}_1(\mathcal{X}_{0,n})}} \int_{\mathcal{X}_{0,n} \times \mathcal{Y}_{0,n}} \log \left(\frac{dV_{X^n|Y^n}(\cdot | y^n)}{dP_{X^n}(\cdot)}(x^n) \right) P_{Y^n|X^n}(dy^n | x^n) \otimes P_{X^n}(dx^n) \quad (\text{IV.3})$$

and the supremum is achieved at $V_{X^n|Y^n}(dx^n | y^n) \equiv P_{X^n|Y^n}(dx^n | y^n)$ given by

$$P_{X^n|Y^n}(dx^n | y^n) = \frac{P_{Y^n|X^n}(dy^n | x^n) \otimes P_{X^n}(dx^n)}{\int_{\mathcal{X}_{0,n}} P_{Y^n|X^n}(dy^n | x^n) \otimes P_{X^n}(dx^n)}. \quad (\text{IV.4})$$

That is, in (IV.1) and (IV.3) the optimal distribution is generated by the joint distribution induced by $\{P_{Y^n|X^n}, P_{X^n}\}$. Both variational equalities are used in the Blahut-Arimoto algorithm (BAA) [25], [39] to derive iterative computational schemes for channel capacity of memoryless channels, via max-max operations, and for RDF of memoryless sources via mini-min operations.

Recently, a version of (IV.3) is applied in [50, eq. (9)] to develop a BAA for capacity of channels with memory and feedback, defined on finite alphabet spaces. Specifically, the authors in [50] consider causally conditioned probability mass functions, $P(x^n|y^{n-1}) \triangleq \prod_{i=0}^n p(x_i|x^{i-1}, y^{i-1})$, $Q(y^n|x^n) \triangleq \prod_{i=0}^n q(y_i|y^{i-1}, x^i)$, where $P(y^n) = \prod_{i=0}^n p(y_i|y^{i-1})$ is generated by $P(x^n, y^n) \triangleq P(x^n|y^{n-1}) \otimes Q(y^n|x^n) = \prod_{i=0}^n p(x_i|x^{i-1}, y^{i-1}) \otimes q(y_i|y^{i-1}, x^i)$, and utilize the identity $P(x^n|y^n) = \prod_{i=0}^n p(x_i|x^{i-1}, y^n)$, to rewrite $\frac{Q(y^n|x^n)}{P(y^n)} = \frac{P(x^n|y^n)}{P(x^n|y^{n-1})}$, and to express (IV.3) as follows.

$$I(X^n \rightarrow Y^n) = \sup_{P(x^n|y^n)} \sum_{(x^n, y^n) \in \mathcal{X}_{0,n} \times \mathcal{Y}_{0,n}} \log \left(\frac{P(x^n|y^n)}{P(x^n|y^{n-1})} \right) P(x^n|y^{n-1}) \otimes Q(y^n|x^n). \quad (\text{IV.5})$$

Based on (IV.5), the authors in [50] developed an algorithm, which computes the causally conditioned product $P^*(x^n|y^{n-1})$ that maximizes (IV.5), similar to the BAA [25], [39], over the product space $\mathcal{X}_{0,n} = \times_{i=0}^n \mathcal{X}_i$. The variational equalities introduced in this paper and the envisioned applications compliment [50], in the sense that our emphasis is on generalizing classical variational equalities, by developing sequential variational equalities, which can be used to develop sequential computational algorithms.

A. Variational Equalities of Directed Information

In this section, our emphasis is to develop variational equalities of directed information, and equivalent sequential variational equalities.

The variational equalities of directed information are based on two families of distributions, similar to $\mathbf{P}(\cdot|\cdot) \in \mathcal{Q}^{\mathbf{C1}}(\mathcal{X}^{\mathbb{N}_0}|\mathcal{Y}^{\mathbb{N}_0})$ and $\mathbf{Q}(\cdot|\cdot) \in \mathcal{Q}^{\mathbf{C2}}(\mathcal{Y}^{\mathbb{N}_0}|\mathcal{X}^{\mathbb{N}_0})$, which are introduced below.

Let $P_{0,n}(dx^n, dy^n) = \overleftarrow{P}_{0,n}(dx^n|y^{n-1}) \otimes \overrightarrow{Q}_{0,n}(dy^n|x^n)$ be the given distribution constructed from the basic feedback channel $\mathbf{P}(\cdot|y) \in \mathcal{M}_1^{\mathbf{C1}}(\mathcal{X}^{\mathbb{N}_0})$ and forward channel $\mathbf{Q}(\cdot|x) \in \mathcal{M}_1^{\mathbf{C2}}(\mathcal{Y}^{\mathbb{N}_0})$ (by projection onto finite number of coordinates).

Let $\mathbf{S}(\cdot|x)$ be any probability measure on $(\mathcal{Y}^{\mathbb{N}_0}, \mathcal{B}(\mathcal{Y}^{\mathbb{N}_0}))$ depending parametrically on $\mathbf{x} \in \mathcal{X}^{\mathbb{N}_0}$ satisfying the following consistency condition.

C3: If $F \in \mathcal{B}(\mathcal{Y}_{0,n})$, then $\mathbf{S}(F|x)$ is a $\mathcal{B}(\mathcal{X}_{0,n-1})$ -measurable.

For fixed $\mathbf{x} \in \mathcal{X}^{\mathbb{N}_0}$, the set of measures on $(\mathcal{Y}^{\mathbb{N}_0}, \mathcal{B}(\mathcal{Y}^{\mathbb{N}_0}))$ satisfying consistency condition **C3** is denoted by $\mathcal{M}_1^{\mathbf{C3}}(\mathcal{Y}^{\mathbb{N}_0})$ and the corresponding family by $\mathcal{Q}^{\mathbf{C3}}(\mathcal{Y}^{\mathbb{N}_0}|\mathcal{X}^{\mathbb{N}_0})$. By Remark II.1, for any family of probability measures $\mathbf{S}(\cdot|x)$ on $(\mathcal{Y}^{\mathbb{N}_0}, \mathcal{B}(\mathcal{Y}^{\mathbb{N}_0}))$ parametrized by $\mathbf{x} \in \mathcal{X}^{\mathbb{N}_0}$, satisfying consistency condition **C3**, there exists a collection of stochastic kernels $\{s_n(\cdot|\cdot, \cdot) \in \mathcal{Q}(\mathcal{Y}_n|\mathcal{Y}_{0,n-1} \times \mathcal{X}_{0,n-1}) : n \in \mathbb{N}_0\}$ connected to $\mathbf{S}(\cdot|x)$ as follows.

$$\mathbf{S}(D|x) = \int_{D_0} s_0(dy_0) \int_{D_1} s_1(dy_1|y_0, x_0) \dots \int_{D_n} s_n(dy_n|y^{n-1}, x^{n-1}) \equiv \overleftarrow{S}_{0,n}(\times_{i=0}^n D_i|x^{n-1}) \quad (\text{IV.6})$$

where

$$D \triangleq \{\mathbf{y} \in \mathcal{Y}^{\mathbb{N}_0} : y_0 \in D_0, y_1 \in D_1, \dots, y_n \in D_n\}, \quad D_i \in \mathcal{B}(\mathcal{Y}_i), \quad \forall i \in \mathbb{N}_0^n.$$

Note that $\overleftarrow{S}_{0,n}(\cdot|x^{n-1}) \in \mathcal{M}_1^{\mathbf{C3}}(\mathcal{Y}_{0,n})$ is conditioned on $x^{n-1} \in \mathcal{X}_{0,n-1}$, unlike $\overrightarrow{Q}_{0,n}(\cdot|x^n) \in \mathcal{M}_1^{\mathbf{C2}}(\mathcal{Y}_{0,n})$, which is conditioned on $x^n \in \mathcal{X}_{0,n}$.

Let $\mathbf{R}(\cdot|y)$ be any family of probability measures on $(\mathcal{X}^{\mathbb{N}_0}, \mathcal{B}(\mathcal{X}^{\mathbb{N}_0}))$ depending parametrically on $\mathbf{y} \in \mathcal{Y}^{\mathbb{N}_0}$

satisfying the following consistency condition.

C4: If $E \in \mathcal{B}(\mathcal{X}_{0,n})$, then $\mathbf{R}(E|\mathbf{y})$ is a $\mathcal{B}(\mathcal{Y}_{0,n})$ -measurable.

For fixed $\mathbf{y} \in \mathcal{Y}^{\mathbb{N}_0}$, the set of measures on $(\mathcal{Y}^{\mathbb{N}_0}, \mathcal{B}(\mathcal{Y}^{\mathbb{N}_0}))$ satisfying consistency condition **C4** is denoted by $\mathcal{M}_1^{\mathbf{C4}}(\mathcal{X}^{\mathbb{N}_0})$ and the corresponding family by $\mathcal{Q}^{\mathbf{C4}}(\mathcal{X}^{\mathbb{N}_0}|\mathcal{Y}^{\mathbb{N}_0})$. Similarly as before, by Remark **II.1**, for any family of measures $\mathbf{R}(\cdot|\mathbf{y})$ on $(\mathcal{X}^{\mathbb{N}_0}, \mathcal{B}(\mathcal{X}^{\mathbb{N}_0}))$ parametrized by $\mathbf{y} \in \mathcal{Y}^{\mathbb{N}_0}$ satisfying consistency condition **C4**, there exists a collection of stochastic kernels $\{r_n(\cdot|\cdot, \cdot) \in \mathcal{Q}(\mathcal{X}_n|\mathcal{X}_{0,n-1} \times \mathcal{Y}_{0,n}) : n \in \mathbb{N}_0\}$ connected to $\mathbf{R}(\cdot|\mathbf{y})$ as follows.

$$\mathbf{R}(E|\mathbf{y}) = \int_{E_0} r_0(dx_0|y_0) \int_{E_1} r_1(dx_1|x_0, y^1) \dots \int_{E_n} r_n(dx_n|x^{n-1}, y^n) \equiv \vec{R}_{0,n}(\times_{i=0}^n E_i|y^n) \quad (\text{IV.7})$$

where

$$E \triangleq \{\mathbf{x} \in \mathcal{X}^{\mathbb{N}_0} : x_0 \in E_0, x_1 \in E_1, \dots, x_n \in E_n\}, \quad E_i \in \mathcal{B}(\mathcal{X}_i), \quad \forall i \in \mathbb{N}_0^n.$$

The joint distribution on $(\mathcal{X}^{\mathbb{N}_0} \times \mathcal{Y}^{\mathbb{N}_0}, \otimes_{n \in \mathbb{N}_0} \mathcal{B}(\mathcal{X}_n) \otimes \mathcal{B}(\mathcal{Y}_n))$ constructed from $\mathbf{S}(\cdot|\cdot) \in \mathcal{Q}^{\mathbf{C3}}(\mathcal{Y}^{\mathbb{N}_0}|\mathcal{X}^{\mathbb{N}_0})$ and $\mathbf{R}(\cdot|\cdot) \in \mathcal{Q}^{\mathbf{C4}}(\mathcal{X}^{\mathbb{N}_0}|\mathcal{Y}^{\mathbb{N}_0})$, is defined uniquely for $D_i \in \mathcal{B}(\mathcal{Y}_i)$, $E_i \in \mathcal{B}(\mathcal{X}_i)$, $\forall i \in \mathbb{N}_0^n$, by

$$\begin{aligned} (\overleftarrow{S}_{0,n} \otimes \vec{R}_{0,n})(\times_{i=0}^n (D_i \times E_i)) &= \int_{D_0} s_0(dy_0) \int_{E_0} r_0(dx_0|y_0) \dots \\ &\dots \int_{D_n} s_n(dy_n|y^{n-1}, x^{n-1}) \int_{E_n} r_n(dx_n|x^{n-1}, y^n). \end{aligned} \quad (\text{IV.8})$$

Formally, the $(n+1)$ fold compound joint distribution defined by (IV.8) is written as $(\overleftarrow{S}_{0,n} \otimes \vec{R}_{0,n})(dx^n, dy^n)$. Note the difference between the stochastic kernels $\{p_i(dx_i|x^{i-1}, y^{i-1}) : i \in \mathbb{N}_0\}$, $\{q_i(dy_i|y^{i-1}, x^i) : i \in \mathbb{N}_0\}$, which define $\overleftarrow{P}_{0,n}(dx^n|y^{n-1})$, $\overrightarrow{Q}_{0,n}(dy^n|x^n)$, respectively, as well as the joint measure $(\overleftarrow{P}_{0,n} \otimes \overrightarrow{Q}_{0,n})(dx^n, dy^n)$, and the stochastic kernels $\{r_i(dx_i|x^{i-1}, y^i) : i \in \mathbb{N}_0^n\}$, $\{s_i(dy_i|y^{i-1}, y^{i-1}) : i \in \mathbb{N}_0^n\}$ which define $\vec{R}_{0,n}(dx^n|y^n)$, $\overleftarrow{S}_{0,n}(dy^n|x^{n-1})$, respectively, and the joint measure $(\overleftarrow{S} \otimes \vec{R})(dx^n, dy^n)$.

The following theorem gives two variational equalities of directed information, including their sequential versions, which are analogous to (IV.1), (IV.3).

Theorem IV.1. (*Variational equalities*)

Let $\{\mathcal{X}_n : n \in \mathbb{N}_0\}$ and $\{\mathcal{Y}_n : n \in \mathbb{N}_0\}$ be Polish spaces. Let $\mathbf{P}(\cdot|\cdot) \in \mathcal{Q}^{\mathbf{C1}}(\mathcal{X}^{\mathbb{N}_0}|\mathcal{Y}^{\mathbb{N}_0})$ and $\mathbf{Q}(\cdot|\cdot) \in \mathcal{Q}^{\mathbf{C2}}(\mathcal{Y}^{\mathbb{N}_0}|\mathcal{X}^{\mathbb{N}_0})$, and for any $n \in \mathbb{N}_0$, construct from them the joint distribution $P_{0,n}(dx^n, dy^n) = (\overleftarrow{P}_{0,n} \otimes \overrightarrow{Q}_{0,n})(dx^n, dy^n)$, and the distributions $\nu_{0,n}(dy^n) = P_{0,n}(\mathcal{X}_{0,n}, dy^n) = \otimes_{i=0}^n \nu_i(dy_i|y^{i-1})$, $\{\nu_i(dy_i|y^{i-1}) \in \mathcal{M}_1(\mathcal{Y}_i) : i = 0, 1, \dots, n\}$, $\vec{\Pi}(dx^n, dy^n) = \overleftarrow{P}_{0,n}(dx^n|y^{n-1}) \otimes \nu_{0,n}(dy^n)$, (defined by (III.5), (III.9), (III.10)).

Then the following variational equalities hold.

Part A. (i) For any arbitrary distribution $V_{0,n}(dy^n) \in \mathcal{M}_1(\mathcal{Y}_{0,n})$ we have

$$\begin{aligned} I(X^n \rightarrow Y^n) &= \mathbb{I}_{X^n \rightarrow Y^n}(\overleftarrow{P}_{0,n}, \overrightarrow{Q}_{0,n}) \triangleq \mathbb{D}(P_{0,n} || \vec{\Pi}_{0,n}) \\ &= \inf_{V_{0,n}(dy^n) \in \mathcal{M}_1(\mathcal{Y}_{0,n})} \mathbb{D}(\overleftarrow{P}_{0,n} \otimes \overrightarrow{Q}_{0,n} || \overleftarrow{P}_{0,n} \otimes V_{0,n}) \end{aligned} \quad (\text{IV.9})$$

$$= \inf_{V_{0,n}(dy^n) \in \mathcal{M}_1(\mathcal{Y}_{0,n})} \left\{ \int_{\mathcal{X}_{0,n} \times \mathcal{Y}_{0,n}} \log \left(\frac{d\overrightarrow{Q}_{0,n}(\cdot|x^n)}{dV_{0,n}(\cdot)}(y^n) \right) (\overleftarrow{P}_{0,n} \otimes \overrightarrow{Q}_{0,n})(dx^n, dy^n) \right\} \quad (\text{IV.10})$$

and the infimum is achieved at $V_{0,n}(dy^n) \equiv \nu_{0,n}(dy^n) \in \mathcal{M}_1(\mathcal{Y}_{0,n})$ given by

$$\nu_{0,n}(dy^n) = \int_{\mathcal{X}_{0,n}} (\overleftarrow{P}_{0,n} \otimes \overrightarrow{Q}_{0,n})(dx^n, dy^n). \quad (\text{IV.11})$$

(ii) For any arbitrary conditional distribution $V_i(dy_i|y^{i-1}) \in \mathcal{M}_1(\mathcal{Y}_i)$, $i = 0, 1, \dots, n$, we have

$$\begin{aligned} I(X^n \rightarrow Y^n) &\equiv \mathbb{I}_{X^n \rightarrow Y^n}(p_i, q_i : i = 0, 1, \dots, n) \\ &= \inf_{\{V_i(dy_i|y^{i-1}) \in \mathcal{M}_1(\mathcal{Y}_i) : i=0,1,\dots,n\}} \sum_{i=0}^n \int_{\mathcal{X}_{0,i} \times \mathcal{Y}_{0,i-1}} \log \left(\frac{dq_i(\cdot|y^{i-1}, x^i)}{dV_i(\cdot|y^{i-1})}(y_i) \right) \\ &\quad p_i(dx_i|x^{i-1}, y^{i-1}) \otimes (\overleftarrow{P}_{0,i-1} \otimes \overrightarrow{Q}_{0,n-1})(dy^{i-1}, dx^{i-1}) \end{aligned} \quad (\text{IV.12})$$

and the infimum is achieved at $V_i(dy_i|y^{i-1}) = \nu_i(dy_i|y^{i-1})$ given by

$$\nu_i(dy_i|y^{i-1}) = \int_{\mathcal{X}_{0,i}} q_i(dy_i|y^{i-1}, x^i) \otimes p_i(dx_i|x^{i-1}, y^{i-1}) \otimes (\overleftarrow{P}_{0,i-1} \otimes \overrightarrow{Q}_{0,i-1})(dx^{i-1}, dy^{i-1}), \quad i = 0, 1, \dots, n. \quad (\text{IV.13})$$

Part B. (i) For any $\mathbf{S}(\cdot|\cdot) \in \mathcal{Q}^{\mathbf{C3}}(\mathcal{Y}^{\mathbb{N}_0}|\mathcal{X}^{\mathbb{N}_0})$ and $\mathbf{R}(\cdot|\cdot) \in \mathcal{Q}^{\mathbf{C4}}(\mathcal{X}^{\mathbb{N}_0}|\mathcal{Y}^{\mathbb{N}_0})$ then

$$\begin{aligned} \mathbb{I}_{X^n \rightarrow Y^n}(\overleftarrow{P}_{0,n}, \overrightarrow{Q}_{0,n}) &= \mathbb{D}(P_{0,n} || \overrightarrow{\Pi}_{0,n}) \\ &= \sup_{\substack{(\overleftarrow{S}_{0,n} \otimes \overrightarrow{R}_{0,n})(dx^n, dy^n) \in \mathcal{M}_1(\mathcal{X}_{0,n} \times \mathcal{Y}_{0,n}) : \\ \overleftarrow{S}_{0,n}(dy^n|x^{n-1}) \in \mathcal{M}_1^{\mathbf{C3}}(\mathcal{Y}_{0,n}), \overrightarrow{R}_{0,n}(dx^n|y^n) \in \mathcal{M}_1^{\mathbf{C4}}(\mathcal{X}_{0,n})}} \int_{\mathcal{X}_{0,n} \times \mathcal{Y}_{0,n}} \log \left(\frac{d(\overleftarrow{S}_{0,n} \otimes \overrightarrow{R}_{0,n})}{d\overrightarrow{\Pi}_{0,n}}(x^n, y^n) \right) \\ &\quad (\overleftarrow{P}_{0,n} \otimes \overrightarrow{Q}_{0,n})(dx^n, dy^n) \end{aligned} \quad (\text{IV.14})$$

and the supremum is achieved at $(\overleftarrow{S}_{0,n} \otimes \overrightarrow{R}_{0,n})(dx^n, dy^n) = (\overleftarrow{P}_{0,n} \otimes \overrightarrow{Q}_{0,n})(dx^n, dy^n)$, given by the RND

$$\Lambda_{0,n}(x^n, y^n) \triangleq \frac{d(\overleftarrow{P}_{0,n} \otimes \overrightarrow{Q}_{0,n})}{d(\overleftarrow{S}_{0,n} \otimes \overrightarrow{R}_{0,n})}(x^n, y^n) = 1 - a.s., \quad n \in \mathbb{N}_0. \quad (\text{IV.15})$$

Equivalently,

$$\lambda_i(x^i, y^i) \triangleq \frac{dp_i(\cdot|x^{i-1}, y^{i-1})}{dr_i(\cdot|x^{i-1}, y^i)}(x_i) \cdot \frac{dq_i(\cdot|y^{i-1}, x^i)}{ds_i(\cdot|y^{i-1}, x^{i-1})}(y_i) = 1 - a.s., \quad i = 0, 1, \dots, n. \quad (\text{IV.16})$$

Moreover, if $q_i(\cdot|y^{i-1}, x^i) \ll s_i(\cdot|y^{i-1}, x^{i-1})$ -a.a. (y^{i-1}, x^i) and $p_i(\cdot|x^{i-1}, y^{i-1}) \ll r_i(\cdot|x^{i-1}, y^i)$ -a.a. (x^{i-1}, y^i) , $i = 0, 1, \dots, n$, then

$$\Pi_{i=0}^n \frac{dq_i(\cdot|y^{i-1}, x^i)}{ds_i(\cdot|y^{i-1}, x^{i-1})}(y_i) = \Pi_{i=0}^n \left(\frac{dp_i(\cdot|x^{i-1}, y^{i-1})}{dr_i(\cdot|x^{i-1}, y^i)}(x_i) \right)^{-1} - a.s., \quad n \in \mathbb{N}_0 \quad (\text{IV.17})$$

or equivalently,

$$\frac{dq_i(\cdot|y^{i-1}, x^i)}{ds_i(\cdot|y^{i-1}, x^{i-1})}(y_i) = \left(\frac{dp_i(\cdot|x^{i-1}, y^{i-1})}{dr_i(\cdot|x^{i-1}, y^i)}(x_i) \right)^{-1} - a.s., \quad i = 0, 1, \dots, n. \quad (\text{IV.18})$$

(ii) For any arbitrary collection of stochastic kernels $\{r_i(\cdot|\cdot, \cdot) \in \mathcal{Q}(\mathcal{X}_i|\mathcal{X}_{0,i-1} \times \mathcal{Y}_{0,i-1}), i = 0, 1, \dots, n\}$, and $\{s_i(\cdot|\cdot, \cdot) \in \mathcal{Q}(\mathcal{Y}_i|\mathcal{Y}_{0,i-1} \times \mathcal{X}_{0,i-1}), i = 0, 1, \dots, n\}$, define

$$\begin{aligned} \mathbb{I}(p_i, q_i, s_i, r_i : i = 0, 1, \dots, n) &\triangleq \sum_{i=0}^n \int_{\mathcal{X}_{0,i} \times \mathcal{Y}_{0,i}} \log \left(\frac{dr_i(\cdot|x^{i-1}, y^i)}{dp_i(\cdot|x^{i-1}, y^{i-1})}(x_i) \cdot \frac{ds_i(\cdot|y^{i-1}, x^{i-1})}{d\nu_i(\cdot|y^{i-1})}(y_i) \right) \\ &\quad \otimes_{k=0}^i (p_k(dx_k|x^{k-1}, y^{k-1}) \otimes q_k(dy_k|y^{k-1}, x^k)). \end{aligned}$$

Then

$$I(X^n \rightarrow Y^n) \equiv \mathbb{I}_{X^n \rightarrow Y^n}(p_i(\cdot|\cdot, \cdot), q_i(\cdot|\cdot, \cdot) : i = 0, 1, \dots, n) \\ = \sup \mathbb{I}(p_i, q_i, s_i, r_i : i = 0, 1, \dots, n) \quad (\text{IV.19}) \\ \left\{ s_i(dy_i|y^{i-1}, x^{i-1}) \otimes r_i(dx_i|x^{i-1}, y^i) \in \mathcal{M}(\mathcal{X}_i \times \mathcal{Y}_i) \right\}, i=0,1,\dots,n \\ \left\{ s_i(dy_i|y^{i-1}, x^{i-1}) \in \mathcal{M}_1(\mathcal{Y}_i), r_i(dx_i|x^{i-1}, y^i) \in \mathcal{M}_1(\mathcal{X}_i) \right\}$$

and the supremum is achieved when (IV.16) or (IV.18) hold.

Proof: Part A. (i) From Theorem III.1, then

$$\mathbb{D}(\overleftarrow{P}_{0,n} \otimes \overrightarrow{Q}_{0,n} || \overleftarrow{P}_{0,n} \otimes V_{0,n}) = \int_{\mathcal{X}_{0,n} \times \mathcal{Y}_{0,n}} \log \left(\frac{d\overrightarrow{Q}_{0,n}(\cdot|x^n)}{dV_{0,n}(\cdot)}(y^n) \right) (\overleftarrow{P}_{0,n} \otimes \overrightarrow{Q}_{0,n})(dx^n, dy^n) \quad (\text{IV.20})$$

$$= \int_{\mathcal{X}_{0,n} \times \mathcal{Y}_{0,n}} \log \left(\frac{d\overrightarrow{Q}_{0,n}(\cdot|x^n)}{d\nu_{0,n}(\cdot)}(y^n) \right) (\overleftarrow{P}_{0,n} \otimes \overrightarrow{Q}_{0,n})(dx^n, dy^n) + \mathbb{D}(\nu_{0,n} || V_{0,n}) \quad (\text{IV.21})$$

$$\geq \mathbb{D}(\overleftarrow{P}_{0,n} \otimes \overrightarrow{Q}_{0,n} || \overleftarrow{P}_{0,n} \otimes \nu_{0,n}). \quad (\text{IV.22})$$

Moreover, equality holds in (IV.22) when $V_{0,n} = \nu_{0,n}$ given by (IV.11). Hence, $\mathbb{D}(P_{0,n} || \overrightarrow{\Pi}_{0,n})$ in (III.20) can be expressed via variational equality (IV.10).

(ii) The derivation of (IV.12) is similar to (IV.9), (IV.10), but it is done with respect to each component $V_i(dy_i|y^{i-1}) \in \mathcal{M}_1(\mathcal{Y}_i)$, starting at $i = n$ and moving sequentially backward to $i = 0$.

Part B. (i) Consider the difference between $I(X^n \rightarrow Y^n) = \mathbb{D}(\overleftarrow{P}_{0,n} \otimes \overrightarrow{Q}_{0,n} || \overrightarrow{\Pi}_{0,n})$ given by (III.20) and the LHS of (IV.14) (without the supremum). Then

$$\mathbb{I}_{X^n \rightarrow Y^n}(\overleftarrow{P}_{0,n}, \overrightarrow{Q}_{0,n}) - \int_{\mathcal{X}_{0,n} \times \mathcal{Y}_{0,n}} \log \left(\frac{d(\overleftarrow{S}_{0,n} \otimes \overrightarrow{R}_{0,n})}{d(\overleftarrow{P}_{0,n} \otimes \nu_{0,n})}(x^n, y^n) \right) (\overleftarrow{P}_{0,n} \otimes \overrightarrow{Q}_{0,n})(dx^n, dy^n) \\ = \int_{\mathcal{X}_{0,n} \times \mathcal{Y}_{0,n}} \log \left(\frac{d(\overleftarrow{P}_{0,n} \otimes \overrightarrow{Q}_{0,n})}{d(\overleftarrow{S}_{0,n} \otimes \overrightarrow{R}_{0,n})}(x^n, y^n) \right) (\overleftarrow{P}_{0,n} \otimes \overrightarrow{Q}_{0,n})(dx^n, dy^n) \\ \geq \int_{\mathcal{X}_{0,n} \times \mathcal{Y}_{0,n}} \left(1 - \frac{d(\overleftarrow{S}_{0,n} \otimes \overrightarrow{R}_{0,n})}{d(\overleftarrow{P}_{0,n} \otimes \overrightarrow{Q}_{0,n})}(x^n, y^n) \right) (\overleftarrow{P}_{0,n} \otimes \overrightarrow{Q}_{0,n})(dx^n, dy^n) = 0 \quad (\text{IV.23})$$

where (IV.23) follows from the inequality $\log x \geq 1 - \frac{1}{x}$, $x > 0$, which holds with equality if and only if $x = 1$. Furthermore, equality holds in (IV.23), when the RND $\Lambda_{0,n}(x^n, y^n) \triangleq \frac{d(\overleftarrow{P}_{0,n} \otimes \overrightarrow{Q}_{0,n})}{d(\overleftarrow{S}_{0,n} \otimes \overrightarrow{R}_{0,n})}(x^n, y^n) = 1$, $\overleftarrow{S}_{0,n} \otimes \overrightarrow{R}_{0,n} - a.s.$ in (x^n, y^n) . Since $(\overleftarrow{P}_{0,n} \otimes \overrightarrow{Q}_{0,n})(\mathcal{X}_{0,n} \times \mathcal{Y}_{0,n}) = (\overleftarrow{S}_{0,n} \otimes \overrightarrow{R}_{0,n})(\mathcal{X}_{0,n} \times \mathcal{Y}_{0,n}) = 1$, this condition is equivalent to $\overleftarrow{P}_{0,n} \otimes \overrightarrow{Q}_{0,n} = \overleftarrow{S}_{0,n} \otimes \overrightarrow{R}_{0,n}$. By conditioning (IV.15) on $\mathcal{B}(\mathcal{X}_{0,n-1}) \otimes \mathcal{B}(\mathcal{Y}_{0,n-1})$ one obtains (IV.16). Furthermore, (IV.17) is obtained from (IV.15), while (IV.18) is obtained by conditioning.

(ii) The derivation of (IV.19) is similar to (IV.14) but it is done with respect to each component $s_i \otimes r_i$, starting at $i = n$ and moving backward sequentially to $i = 0$. \square

Note that Theorem IV.1, Part A. (ii), Part B. (ii) are sequential versions of Part A. (i), Part B. (i), respectively.

Next, we discuss the relation between the variational equality of directed information (IV.14) and the variational equality of mutual information (IV.3). Clearly, (IV.3) is also equivalent to

$$\sup_{V_{X^n|Y^n} \otimes P_{Y^n}} \int_{\mathcal{X}_{0,n} \times \mathcal{Y}_{0,n}} \log \left(\frac{d(V_{X^n|Y^n}(\cdot|y^n) \otimes P_{Y^n}(\cdot))}{d(P_{X^n}(\cdot) \times P_{Y^n}(\cdot))}(x^n, y^n) \right) P_{Y^n|X^n}(dy^n|x^n) \otimes P_{X^n}(dx^n) \quad (\text{IV.24})$$

since the RND in (IV.24) is another version of the one in (IV.3). Hence, (IV.14) is the analogue of (IV.24). Further, to obtain the analogue of the maximizing measure in (IV.3), given by (IV.4), suppose $q_i(\cdot|y^{i-1}, x^i) \ll s_i(\cdot|y^{i-1}, x^{i-1}) - a.a. (x^i, y^{i-1})$, $i = 0, 1, \dots, n$, and $\{s_i(\cdot|y^{i-1}, x^{i-1}) : i = 0, 1, \dots, n\}$ is fixed, and generated by $\vec{P}_{0,n}(\cdot|y^{n-1}) \in \mathcal{M}^{C1}(\mathcal{X}_{0,n})$ and $\vec{Q}_{0,n}(\cdot|x^n) \in \mathcal{M}^{C2}(\mathcal{Y}_{0,n})$. Then from (IV.16) we obtain

$$r_i(dx_i|x^{i-1}, y^i) = \left(\frac{dq_i(\cdot|y^{i-1}, x^i)}{ds_i(\cdot|y^{i-1}, x^{i-1})}(y_i) \right) p_i(dx_i|x^{i-1}, y^{i-1}), \quad i = 0, 1, \dots, n \quad (\text{IV.25})$$

$$= \frac{q_i(dy_i|y^{i-1}, x^i)}{\int_{\mathcal{X}_i} q_i(dy_i|y^{i-1}, x^i) \otimes p_i(dx_i|x^{i-1}, y^{i-1})} p_i(dx_i|x^{i-1}, y^{i-1}), \quad i = 0, 1, \dots, n. \quad (\text{IV.26})$$

Obviously, for a fixed $\{s_i(\cdot|y^{i-1}, x^{i-1}) : i = 0, 1, \dots, n\}$, (IV.25), (IV.26) are the sequential versions of maximizing distribution satisfying (IV.15), given by

$$\vec{R}_{0,n}(dx^n|y^n) = \otimes_{i=0}^n \frac{q_i(dy_i|y^{i-1}, x^i)}{\int_{\mathcal{X}_i} q_i(dy_i|y^{i-1}, x^i) \otimes p_i(dx_i|x^{i-1}, y^{i-1})} p_i(dx_i|x^{i-1}, y^{i-1}), \quad n \in \mathbb{N}_0. \quad (\text{IV.27})$$

Clearly, (IV.27) is the analogue of the maximizing distribution $P_{X^n|Y^n}$ in (IV.3).

Note that the optimization in (IV.14) can be done by keeping $\vec{S}_{0,n}(\cdot|x^{n-1})$ fixed, generated by $\mathbf{P}(\cdot|\cdot) \in \mathcal{Q}^{C1}(\mathcal{X}^{\mathbb{N}_0}|\mathcal{Y}^{\mathbb{N}_0})$ and $\mathbf{Q}(\cdot|\cdot) \in \mathcal{Q}^{C2}(\mathcal{Y}^{\mathbb{N}_0}|\mathcal{X}^{\mathbb{N}_0})$, and maximizing only over $\vec{R}_{0,n}(\cdot|y^n) \in \mathcal{M}_1(\mathcal{X}_{0,n})$ as demonstrated above.

For extremum problems of directed information, such as, the channel capacity with memory with and without feedback, it is desirable to invoke a sequential version of variational equalities, in order to derive sequential algorithms. This point is illustrated in the next section.

B. Applications of Sequential Variational Equalities to Feedback Capacity Computations

Consider the extremum problem of feedback capacity given by (I.5), without transmission cost constraint. Expressed in terms of channel distributions $\{q_i(dy_i|y^{i-1}, x^i) \in \mathcal{M}_1(\mathcal{Y}_i) : i = 0, 1, \dots, n\}$ and the channel input distributions $\{p_i(dx_i|x^{i-1}, y^{i-1}) \in \mathcal{M}_1(\mathcal{X}_i) : i = 0, 1, \dots, n\}$, then $C^{fb} \triangleq \liminf_{n \rightarrow \infty} \frac{1}{n+1} C_{0,n}^{fb}$ where

$$C_{0,n}^{fb} \triangleq \sup_{\{p_i(dx_i|x^{i-1}, y^{i-1}) \in \mathcal{M}_1(\mathcal{X}_i) : i=0,1,\dots,n\}} \sum_{i=0}^n I(X^i; Y_i|Y^{i-1}). \quad (\text{IV.28})$$

Given a specific channel, Theorem IV.1, Part B. (ii) can be used to develop a sequential alternating double maximization algorithm over appropriate sets of distributions, which computes C^{fb} via (IV.28) (i.e., $\frac{C_{0,n}^{fb}}{n+1}$), for large enough n , starting at n and moving sequentially in time to $n-1, n-2, \dots, 0$. This is illustrated next, by considering a simple example.

Unit Memory Channel. Consider a channel defined by $\{q_i(dy_i|y_{i-1}, x_i) \in \mathcal{M}_1(\mathcal{Y}_i) : i = 0, 1, \dots, n\}$, called Unit Memory Channel Output (UMCO). Then, (IV.28) reduces to

$$C_{0,n}^{fb,UMCO} \triangleq \sup_{\{p_i(dx_i|x^{i-1}, y^{i-1}) \in \mathcal{M}_1(\mathcal{X}_i) : i=0,1,\dots,n\}} \sum_{i=0}^n \mathbb{E} \left\{ \log \left(\frac{dq_i(\cdot|Y_{i-1}, X_i)}{d\nu_i(\cdot|Y^{i-1})}(Y_i) \right) \right\}. \quad (\text{IV.29})$$

It is conjectured by Chen and Berger [8] (see also [51], [52]) that the optimal channel input distribution in (IV.29) satisfies the conditional independence $p_i(dx_i|x^{i-1}, y^{i-1}) = \pi_i(dx_i|y_{i-1}) - a.a. (x^{i-1}, y^{i-1}) \in \mathcal{X}_{0,n-1} \times \mathcal{Y}_{0,n-1}$, which then implies the corresponding joint process $\{(X_i, Y_i) : i = 0, 1, \dots, n\}$ is first

order Markov, the output process $\{Y_i : i = 0, 1, \dots, n\}$ is first order Markov, and consequently, (IV.29) reduces to the following expression⁴.

$$C_{0,n}^{fb,UMCO} \triangleq \sup_{\{\pi_i(dx_i|y_{i-1}) \in \mathcal{M}_1(\mathcal{X}_i) : i=0,1,\dots,n\}} \sum_{i=0}^n \int_{\mathcal{Y}_{i-1,i} \times \mathcal{X}_i} \log \left(\frac{dq_i(\cdot|y_{i-1}, x_i)}{d\nu_i^\pi(\cdot|y_{i-1})}(y_i) \right) q_i(dy_i|y_{i-1}, x_i) \otimes \pi_i(dx_i|y_{i-1}) \otimes \nu_i^\pi(dy_{i-1}) \quad (\text{IV.30})$$

$$= \sup_{\{\pi_i(dx_i|y_{i-1}) \in \mathcal{M}_1(\mathcal{X}_i) : i=0,1,\dots,n\}} \sum_{i=0}^n I(X_i; Y_i|Y_{i-1}) \quad (\text{IV.31})$$

where

$$\nu_i^\pi(\cdot|y_{i-1}) = \int_{\mathcal{X}_i} q_i(dy_i|y_{i-1}, x_i) \otimes \pi_i(dx_i|y_{i-1}), \quad i = 0, 1, \dots, n. \quad (\text{IV.32})$$

The conjecture by Chen and Berger [8] (i.e., (IV.30)-(IV.32)) is recently shown in [53], by invoking the variational equality (IV.12) in extremum problems of feedback capacity, to identify information structures of the optimal channel input distribution for general channels with finite memory.

By Theorem IV.1, **Part B. (ii)**, for a fixed $\{\pi_i(dx_i|y_{i-1}) \in \mathcal{M}_1(\mathcal{X}_i) : i = 0, 1, \dots, n\}$, the expression inside the maximization in (IV.30) or (IV.31) is expressed as

$$\sum_{i=0}^n I(X_i; Y_i|Y_{i-1}) = \sup_{\{r_i(dx_i|y_{i-1}, y_i) \in \mathcal{M}_1(\mathcal{X}_i) : i=0,1,\dots,n\}} \sum_{i=0}^n \int_{\mathcal{Y}_{i-1,i} \times \mathcal{X}_i} \log \left(\frac{dr_i(\cdot|y_{i-1}, y_i)}{d\pi_i(\cdot|y_{i-1})}(x_i) \right) q_i(dy_i|y_{i-1}, x_i) \otimes \pi_i(dx_i|y_{i-1}) \otimes \nu_i^\pi(dy_{i-1}) \quad (\text{IV.33})$$

where the supremum in (IV.33) is achieved at

$$r_i^\pi(dx_i|y_{i-1}, y_i) = \left(\frac{dq_i(\cdot|y_{i-1}, x_i)}{d\nu_i^\pi(\cdot|y_{i-1})}(y_i) \right) \pi_i(dx_i|y_{i-1}), \quad i = 0, 1, \dots, n. \quad (\text{IV.34})$$

Next, we convert $C_{0,n}^{fb,UMCO}$ into a sequential alternating maximization problem over appropriate sets of distributions, by using dynamic programming.

Let $C_t : \mathcal{Y}_{t-1} \mapsto [0, \infty)$ represent the maximum expected total pay-off in (IV.30) on the future time horizon $\{t, t+1, \dots, n\}$, given $Y_{t-1} = y_{t-1}$ at time $t-1$, defined by

$$C_t(y_{t-1}) = \sup_{\{\pi_i(dx_i|y_{i-1}) \in \mathcal{M}_1(\mathcal{X}_i) : i=t,t+1,\dots,n\}} \mathbb{E}^\pi \left\{ \sum_{i=t}^n \log \left(\frac{dq_i(\cdot|y_{i-1}, x_i)}{d\nu_i^\pi(\cdot|y_{i-1})}(y_i) \right) q_i(dy_i|y_{i-1}, x_i) \otimes \pi_i(dx_i|y_{i-1}) \middle| Y_{t-1} = y_{t-1} \right\}. \quad (\text{IV.35})$$

By standard arguments (see [33]), and in view of the Markov property of $\{Y_i : i = 0, 1, \dots, n\}$, it follows that (IV.35) satisfies the following dynamic programming recursions.

$$C_n(y_{n-1}) = \sup_{\pi_n(dx_n|y_{n-1}) \in \mathcal{M}_1(\mathcal{X}_n)} \int_{\mathcal{X}_n \times \mathcal{Y}_n} \log \left(\frac{dq_n(\cdot|y_{n-1}, x_n)}{d\nu_n^\pi(\cdot|y_{n-1})}(y_n) \right) q_n(dy_n|y_{n-1}, x_n) \otimes \pi_n(dx_n|y_{n-1}) \quad (\text{IV.36})$$

$$C_t(y_{t-1}) = \sup_{\pi_t(dx_t|y_{t-1}) \in \mathcal{M}_1(\mathcal{X}_t)} \left\{ \int_{\mathcal{X}_t \times \mathcal{Y}_t} \log \left(\frac{dq_t(\cdot|y_{t-1}, x_t)}{d\nu_t^\pi(\cdot|y_{t-1})}(y_t) \right) q_t(dy_t|y_{t-1}, x_t) \otimes \pi_t(dx_t|y_{t-1}) \right. \\ \left. + \int_{\mathcal{X}_t \times \mathcal{Y}_t} C_{t+1}(y_t) q_t(dy_t|y_{t-1}, x_t) \otimes \pi_t(dx_t|y_{t-1}) \right\}, \quad t = 0, 1, \dots, n-1. \quad (\text{IV.37})$$

⁴superscript π on various distributions indicates their dependence on $\{\pi_i(dx_i|y_{i-1}) : i = 0, 1, \dots, n\}$.

It is well-known that the computation of the optimal channel input distribution in (IV.36), (IV.37) suffers from the so-called, curse of dimensionality (i.e., it is often computationally prohibitive, even for finite alphabet spaces). However, by applying Theorem IV.1, Part B. (ii), to the dynamic programming recursions (IV.36), (IV.37), we can show that these can be converted to equivalent alternating maximizations over convex sets. Consequently, (IV.30) can be expressed via sequential alternating maximizations, of concave functionals over convex sets, as stated in the next theorem.

Theorem IV.2. (Sequential double maximization of feedback capacity of UMCO)

Consider the UMCO defined by $\{q_i(dy_i|y_{i-1}, x_i) \in \mathcal{M}_1(\mathcal{Y}_i) : i = 0, 1, \dots, n\}$, and $C_{0,n}^{fb,UMCO}$ defined by (IV.30), for a fixed $\text{Prob}\{Y_{-1} \in dy_{-1}\} \triangleq \nu_{-1}(dy_{-1})$.

Part A. The dynamic programming recursions (IV.36), (IV.37) are equivalent to the following sequential double maximization dynamic programming recursions.

$$C_n(y_{n-1}) = \sup_{\pi_n(dx_n|y_{n-1}) \in \mathcal{M}_1(\mathcal{X}_n)} \sup_{r_n(dx_n|y_{n-1}, y_n) \in \mathcal{M}_1(\mathcal{X}_n)} \left\{ \int_{\mathcal{X}_n \times \mathcal{Y}_n} \log \left(\frac{dr_n(\cdot|y_{n-1}, y_n)}{d\pi_n(\cdot|y_{n-1})}(x_n) \right) q_n(dy_n|y_{n-1}, x_n) \otimes \pi_n(dx_n|y_{n-1}) \right\} \quad (\text{IV.38})$$

$$C_t(y_{t-1}) = \sup_{\pi_t(dx_t|y_{t-1}) \in \mathcal{M}_1(\mathcal{X}_t)} \sup_{r_t(dx_t|y_{t-1}, y_t) \in \mathcal{M}_1(\mathcal{X}_t)} \left\{ \int_{\mathcal{X}_t \times \mathcal{Y}_t} \log \left(\frac{dr_t(\cdot|y_{t-1}, y_t)}{d\pi_t(\cdot|y_{t-1})}(x_t) \right) q_t(dy_t|y_{t-1}, x_t) \otimes \pi_t(dx_t|y_{t-1}) + \int_{\mathcal{X}_t \times \mathcal{Y}_t} C_{t+1}(y_t) q_t(dy_t|y_{t-1}, x_t) \otimes \pi_t(dx_t|y_{t-1}) \right\}, \quad t = 0, 1, \dots, n-1. \quad (\text{IV.39})$$

and $C_{0,n}^{fb,UMCO}$ is given by

$$C_{0,n}^{fb,UMCO} = \int_{\mathcal{Y}_{-1}} C_0(y_{-1}) \nu_{-1}(dy_{-1}).$$

Moreover, the following hold.

Maximizations in (IV.38).

(i) For a fixed $\pi_n(dx_n|y_{n-1}) \in \mathcal{M}_1(\mathcal{X}_n)$, the maximum in (IV.38) over $r_n(dx_n|y_{n-1}, y_n) \in \mathcal{M}_1(\mathcal{X}_n)$ occurs at $r_n(\cdot|\cdot, \cdot) = r_n^{*,\pi}(\cdot|\cdot, \cdot)$ given by

$$r_n^{*,\pi}(dx_n|y_{n-1}, y_n) = \left(\frac{dq_n(\cdot|y_{n-1}, x_n)}{d\nu_n^\pi(\cdot|y_{n-1})}(y_n) \right) \pi_n(dx_n|y_{n-1}). \quad (\text{IV.40})$$

(ii) For a fixed $r_n(dx_n|y_{n-1}, y_n) \in \mathcal{M}_1(\mathcal{X}_n)$, the maximum in (IV.38) over $\pi_n(dx_n|y_{n-1}) \in \mathcal{M}_1(\mathcal{X}_n)$ occurs at $\pi_n(\cdot|\cdot) = \pi_n^{*,r}(\cdot|\cdot)$ ⁵ given by

$$\pi_n^{*,r}(dx_n|y_{n-1}) = \frac{\exp \left\{ \int_{\mathcal{Y}_n} \log \left(\frac{dr_n(\cdot|y_{n-1}, y_n)}{d\pi_n(\cdot|y_{n-1})}(x_n) \right) q_n(dy_n|y_{n-1}, x_n) \right\} \pi_n(dx_n|y_{n-1})}{\int_{\mathcal{X}_n} \exp \left\{ \int_{\mathcal{Y}_n} \log \left(\frac{r_n^\pi(\cdot|y_{n-1}, y_n)}{\pi_n(\cdot|y_{n-1})}(x_n) \right) q_n(dy_n|y_{n-1}, x_n) \right\} \pi_n(dx_n|y_{n-1})} \quad (\text{IV.41})$$

Moreover, when (IV.41) is evaluated at $r_n(\cdot|\cdot, \cdot) = r_n^{*,\pi}(\cdot|\cdot, \cdot)$ given by (IV.40) then

$$\pi_n^{*,r^*}(dx_n|y_{n-1}) = \frac{\exp \left\{ \int_{\mathcal{Y}_n} \log \left(\frac{dq_n(\cdot|y_{n-1}, x_n)}{d\nu_n^\pi(\cdot|y_{n-1})}(y_n) \right) q_n(dy_n|y_{n-1}, x_n) \right\} \pi_n(dx_n|y_{n-1})}{\int_{\mathcal{X}_n} \exp \left\{ \int_{\mathcal{Y}_n} \log \left(\frac{dq_n(\cdot|y_{n-1}, x_n)}{d\nu_n^\pi(\cdot|y_{n-1})}(y_n) \right) q_n(dy_n|y_{n-1}, x_n) \right\} \pi_n(dx_n|y_{n-1})}. \quad (\text{IV.42})$$

Maximizations in (IV.39).

(iii) For a fixed $\pi_t(dx_t|y_{t-1}) \in \mathcal{M}_1(\mathcal{X}_t)$, the maximum in (IV.39) over $r_t(dx_t|y_{t-1}, y_t) \in \mathcal{M}_1(\mathcal{X}_t)$ occurs at $r_t(\cdot|\cdot, \cdot) = r_t^{*,\pi}(\cdot|\cdot, \cdot)$ given by

$$r_t^{*,\pi}(dx_t|y_{t-1}, y_t) = \left(\frac{dq_t(\cdot|y_{t-1}, x_t)}{d\nu_t^\pi(\cdot|y_{t-1})}(y_t) \right) \pi_t(dx_t|y_{t-1}), \quad t = n-1, n-2, \dots, 0. \quad (\text{IV.43})$$

⁵superscript r indicates the dependence on the distribution $\{r_i(dx_i|y_{i-1}, y_i) : i = 0, 1, \dots, n\}$.

(iv) For a fixed $r_t(dx_t|y_{n-1}, y_t) \in \mathcal{M}_1(\mathcal{X}_t)$, the maximum in (IV.39) over $\pi_t(dx_t|y_{t-1}) \in \mathcal{M}_1(\mathcal{X}_t)$, occurs at $\pi_t(\cdot|\cdot) = \pi_t^{*,r}(\cdot|\cdot)$, $t = n-1, n-2, \dots, 0$, given by

$$\pi_t^{*,r}(dx_t|y_{t-1}) = \frac{\exp \left\{ \int_{\mathcal{Y}_t} \left\{ \log \left(\frac{dr_t(\cdot|y_{t-1}, y_t)}{d\pi_t(\cdot|y_{t-1})}(x_t) \right) + C_{t+1}(y_t) \right\} q_t(dy_t|y_{t-1}, x_t) \right\} \pi_t(dx_t|y_{t-1})}{\int_{\mathcal{X}_t} \exp \left\{ \int_{\mathcal{Y}_t} \left\{ \log \left(\frac{dr_t(\cdot|y_{t-1}, y_t)}{d\pi_t(\cdot|y_{t-1})}(x_t) \right) + C_{t+1}(y_t) \right\} q_t(dy_t|y_{t-1}, x_t) \right\} \pi_t(dx_t|y_{t-1})}. \quad (\text{IV.44})$$

Moreover, when (IV.44) is evaluated at $r_t(\cdot|\cdot, \cdot) = r_t^{*,\pi}(\cdot|\cdot, \cdot)$, $t = n-1, n-2, \dots, 0$, given by (IV.43) then

$$\pi_t^{*,r^*}(dx_t|y_{t-1}) = \frac{\exp \left\{ \int_{\mathcal{Y}_t} \left\{ \log \left(\frac{dq_t(\cdot|y_{t-1}, x_t)}{d\nu_t^{\pi}(\cdot|y_{t-1})}(y_t) \right) + C_{t+1}(y_t) \right\} q_t(dy_t|y_{t-1}, x_t) \right\} \pi_t(dx_t|y_{t-1})}{\int_{\mathcal{X}_t} \exp \left\{ \int_{\mathcal{Y}_t} \left\{ \log \left(\frac{dq_t(\cdot|y_{t-1}, x_t)}{d\nu_t^{\pi}(\cdot|y_{t-1})}(y_t) \right) + C_{t+1}(y_t) \right\} q_t(dy_t|y_{t-1}, x_t) \right\} \pi_t(dx_t|y_{t-1})}. \quad (\text{IV.45})$$

Part B. The extremum problem $C_{0,n}^{fb,UMCO}$ defined by (IV.30) is equivalent to the following sequential double maximization problem.

$$C_{0,n}^{fb,UMCO} = \sup_{\pi_0(dx_0|y_{-1}) \in \mathcal{M}_1(\mathcal{X}_0)} \sup_{r_0^{\pi}(dx_0|y_{-1}, y_0) \in \mathcal{M}_1(\mathcal{X}_0)} \dots \sup_{\pi_n(dx_n|y_{n-1}) \in \mathcal{M}_1(\mathcal{X}_n)} \sup_{r_n^{\pi}(dx_n|y_{n-1}, y_n) \in \mathcal{M}_1(\mathcal{X}_n)} \left\{ \sum_{i=0}^n \int_{\mathcal{Y}_{i-1,i} \times \mathcal{X}_i} \log \left(\frac{dr_i^{\pi}(\cdot|y_{i-1}, y_i)}{d\pi_i(\cdot|y_{i-1})}(x_i) \right) q_i(dy_i|y_{i-1}, x_i) \otimes \pi_i(dx_i|y_{i-1}) \otimes \nu_i^{\pi}(dy_{i-1}) \right\} \quad (\text{IV.46})$$

and statements (i)-(iv) hold.

Proof: **Part A.** (i) (IV.38) and (IV.40) follow directly from (IV.36). (ii) (IV.41) is obtained as follows. Fix $r_n(dx_n|y_{n-1}, y_n) \in \mathcal{M}_1(\mathcal{X}_n)$, calculate the Gâteaux differential inside the maximization in (IV.38) at $\pi_n^{*,r}(dx_n|y_{n-1})$ in the direction $\pi_n^r(dx_n|y_{n-1}) - \pi_n^{*,r}(dx_n|y_{n-1})$, i.e., $\pi_n^{\epsilon,r}(dx_n|y_{n-1}) \triangleq \pi_n^{*,r}(dx_n|y_{n-1}) - \epsilon \{ \pi_n^r(dx_n|y_{n-1}) - \pi_n^{*,r}(dx_n|y_{n-1}) \}$, $\epsilon \in [0, 1]$, by incorporating the constraint $\int_{\mathcal{X}_n} \pi_n^r(dx_n|y_{n-1}) = 1$ via a Lagrange multiplier $\lambda_n(y_{n-1})$. The Gâteaux differential gives (IV.41). Then substitute (IV.40) into (IV.41) to obtain (IV.42). (iii) For fixed $\pi_t(dx_t|y_{t-1}) \in \mathcal{M}_1(\mathcal{X}_t)$, the second RHS term in (IV.37) is a function of the channel distribution, hence (IV.39) and (IV.43) follow directly as in (i). (iv) To show (IV.44), (IV.45), compute the Gâteaux differential as in (ii), by tracking the additional second RHS term in (IV.39).

Part B. Since $\nu_{-1}(dy_{-1}) \in \mathcal{M}_1(\mathcal{Y}_{-1})$ is fixed, then (IV.46) follows directly from **Part A.**, and the definition of $C_t(y_{t-1})$ evaluated at $t = 0$. \square

Theorem IV.2, specifically (IV.42), (IV.45), are the equations, which should be used to derive a sequential algorithm to compute numerically the optimal channel input distribution.

Below, we discuss applications of Theorem IV.2, and identify generalizations, and directions for future research.

Remark IV.3. (Sequential algorithms for feedback capacity)

- (1) For the UMCO, Theorem IV.2 provides all necessary ingredients to derive a sequential algorithm at each time step, $t = n, n-1, \dots, 0$, similar to the BAA. It remains to show at each time step, $t = n, n-1, \dots, 0$, that (IV.42), (IV.44) have fixed points corresponding to the optimal channel input distribution, and to derive upper and lower bounds on $C_t(y_{t-1})$, $t = n, n-1, \dots, 0$, to stop the iterations at each time step of the algorithm. For finite alphabet spaces $\{(\mathcal{X}_i, \mathcal{Y}_i) : i = 0, 1, \dots, n\}$, these additional steps can be carried out using Theorem IV.2 and the procedure in [39].
- (2) For the UMCO, if the alphabet spaces $\mathcal{X}_i \equiv \mathcal{X}$, $\mathcal{Y}_i \equiv \mathcal{Y}$, $i = 0, 1, \dots$, and the joint process $\{(X_i, Y_i) : i = 0, 1, \dots\}$ is stationary ergodic or directed information stable, then the per unit time limiting version of dynamic programming recursive equations (IV.36), (IV.37) can be derived [54], and these involve only a single stage maximization over $\pi(dx_i|y_{i-1}) \in \mathcal{M}_1(\mathcal{X})$, $\forall i$. Hence, a theorem similar to Theorem IV.2 can be derived.

- (3) For general channels, it is possible to derive the analogue of Theorem IV.2, provided the set of optimal channel input distributions, which maximize $\sum_{i=0}^n I(X^i; Y_i | Y^{i-1})$ is identified, as in the case of UMC0 (see [53]).

V. CONCLUSION

In this paper we derive functional and topological properties of directed information, for abstract alphabet spaces (i.e., complete separable metric spaces). These include, convexity of the set of consistent family of distributions, which uniquely define causally conditioned compound distributions, convexity and concavity of directed information with respect to consistent family of distributions, and a general theorem on weak compactness of causally conditioned distributions, their joint distributions, and marginals, which are utilized to define directed information. Further, we use this main theorems to show lower semicontinuity of directed information as a functional of two causally conditioned distributions, and under additional conditions continuity of directed information. In addition, we derive sequential variational equalities for directed information. Throughout the paper, we discuss application examples in the context of extremum problems of directed information, such as, in feedback capacity, nonanticipative RDF, and in developing sequential computational algorithms, similar to the Blahut-Arimoto algorithm [39].

APPENDIX A BACKGROUND MATERIAL

In this section, we introduce some of the basic analytical concepts which are used throughout the paper.

Weak Convergence and Compactness.

The main notions discussed are weak convergence of probability measures, the relation to convergence with respect to Prohorov metric, tightness of a family of probability measures and relative compactness [31].

Let (\mathcal{X}, d) be a metric space, $\mathcal{B}(\mathcal{X})$ the σ -algebra of Borel subsets of \mathcal{X} , and $\mathcal{M}_1(\mathcal{X})$ the family of probability measures on \mathcal{X} . Let $BC(\mathcal{X})$ denote the set of bounded, continuous real-valued function f on (\mathcal{X}, d) , endowed with the supremum norm $\|f\| = \sup_{x \in \mathcal{X}} |f(x)|$. A sequence of probability measures $\{P_n : n = 1, 2, \dots\} \subset \mathcal{M}_1(\mathcal{X})$ is said to *converge weakly* to a probability measure $P \in \mathcal{M}_1(\mathcal{X})$ if

$$\lim_{n \rightarrow \infty} \int_{\mathcal{X}} f(x) dP_n(x) = \int_{\mathcal{X}} f(x) dP(x), \quad \forall f \in BC(\mathcal{X}).$$

Weak convergence of $\{P_n : n = 1, 2, \dots\}$ to P is denoted by $P_n \xrightarrow{w} P$. The space of probability measures $\mathcal{M}_1(\mathcal{X})$ is metrizable with respect to the Prohorov metric (see [30]).

A crucial result for the characterization of compact subsets of $\mathcal{M}_1(\mathcal{X})$ is the next theorem due to Prohorov, which relates compactness and tightness of a family of measures.

Definition A.1. (*Tightness and Relative Compactness*) [30, p. 308]

Let $M \subset \mathcal{M}_1(\mathcal{X})$ be a family of probability measures on a metric space (\mathcal{X}, d) . M is said to be

- 1) *tight or uniformly tight* if for every $\epsilon > 0$ there exists a compact set $K^{(\epsilon)} \subset \mathcal{X}$ such that $\inf_{P \in M} P(K^{(\epsilon)}) \geq 1 - \epsilon$;
- 2) *relatively compact or weakly compact* if every sequence in M contains a weakly convergent subsequence, that is, for every sequence $\{P_n : n = 1, 2, \dots\}$ in M there is a subsequence $\{P_{n_i} : i \in \{1, 2, \dots\}\}$ and a $P \in \mathcal{M}_1(\mathcal{X})$ such that $P_{n_i} \xrightarrow{w} P$. Here, the limit P is not required to belong to M , but all is required is to belong to $\mathcal{M}_1(\mathcal{X})$.

Prohorov states that for (\mathcal{X}, d) a metric space and \mathcal{X} compact, then any sequence $\{P_n : n = 1, 2, \dots\}$ of probability measures on \mathcal{X} possess a convergent subsequence. The following theorem due to Prohorov, relates weak compactness and tightness of a family of probability measures.

Theorem A.2. (*Prohorov's Theorem*) [30, Theorem A.3.15, p. 309]

Let $M \subset \mathcal{M}_1(\mathcal{X})$ be a family of probability measures on a metric space (\mathcal{X}, d) .

- 1) If M is tight, then it is relative compact.
- 2) Suppose \mathcal{X} is separable and complete. If M is relatively compact, then it is tight.

Thus, a family of probability measures $M \subset \mathcal{M}_1(\mathcal{X})$ on a complete separable metric space (\mathcal{X}, d) is weakly compact or relatively compact with respect to weak convergence if and only if it is tight. Moreover, if $P_n \xrightarrow{w} P$, then the family $\{P_n : n = 1, 2, \dots\}$ is tight.

Finally, we give another version due to Prohorov for a family of measures $M \subset \mathcal{M}_1(\mathcal{X})$ to be compact.

Theorem A.3. (*Corollary of Prohorov's Theorem*)

Let (\mathcal{X}, d) be a separable metric and $M \subset \mathcal{M}_1(\mathcal{X})$ a set of measures. The following hold.

- (a) If M is closed and tight, then M is compact.
- (b) Suppose \mathcal{X} is complete. If M is compact then M is closed and tight.

In what follows, we give the definition of weak continuity of conditional distributions, which is often associated with proving results using weak convergence of probability distributions, and we distinguish it from strong continuity.

Definition A.4. (*Strong and weak continuity*)

Let (\mathcal{X}, d) , (\mathcal{Y}, d') be metric spaces, $Q(\cdot|\cdot) \in \mathcal{Q}(\mathcal{Y}|\mathcal{X})$ a conditional distribution, and define by $BM(\mathcal{Y})$ the set of bounded measurable functions on \mathcal{Y} . Then $Q(\cdot|\cdot) \in \mathcal{Q}(\mathcal{Y}|\mathcal{X})$ is said to be

- 1) strongly continuous if the function mapping

$$x \mapsto \int_{\mathcal{Y}} f(y) Q(dy|x) \in BC(\mathcal{Y})$$

whenever $f(\cdot) \in BM(\mathcal{Y})$,

- 2) weakly continuous if the function mapping

$$x \mapsto \int_{\mathcal{Y}} f(y) Q(dy|x) \in BC(\mathcal{Y})$$

whenever $f(\cdot) \in BC(\mathcal{Y})$.

It can be shown that strong continuity is equivalent to $Q(B|\cdot)$ is continuous on \mathcal{Y} for every set $B \in \mathcal{B}(\mathcal{Y})$ (i.e., its conditional distribution is continuous), and this is much stronger than weak continuity of $Q(\cdot|\cdot) \in \mathcal{Q}(\mathcal{Y}|\mathcal{X})$.

Uniform Integrability.

In this paper we shall also need stronger sufficient conditions to verify convergence of a sequence of integrals using the concept of uniform integrability. We state this next.

Definition A.5. (*Uniform Integrability of RV's*) [45, Definition 4, p. 188]

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. A sequence of RV's $\{X_n : n \in \mathbb{N}_1\}$, $\mathbb{N}_1 \triangleq \{1, 2, \dots\}$, is said to be uniformly \mathbb{P} -integrable if

$$\lim_{c \rightarrow \infty} \sup_{n \in \mathbb{N}_1} \int_{\{\omega : |X_n(\omega)| \geq c\}} |X_n(\omega)| d\mathbb{P}(\omega) = 0.$$

Note that if $\{X_n : n \in \mathbb{N}_1\}$ satisfy $|X_n| \leq Y$ and $\mathbb{E}\{Y\} < \infty$, then the sequence $\{X_n : n \in \mathbb{N}_1\}$ is uniformly integrable.

The following theorem gives some properties for a family of uniformly integrable RV's.

Theorem A.6. (*Uniform Integrability of RV's*) [45, Theorem 4, pp. 188-189]

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $\{X_n : n \in \mathbb{N}_1\}$ a uniformly \mathbb{P} -integrable family of RV's. Then

- (a) $\mathbb{E} \liminf_n X_n \leq \liminf_n \mathbb{E} X_n \leq \limsup_n \mathbb{E} X_n \leq \mathbb{E} \limsup_n X_n$.

(b) If $X_n \xrightarrow{a.s.} X$, then $\mathbb{E}|X| < \infty$, $\lim_{n \rightarrow \infty} \mathbb{E}|X_n| = \mathbb{E}|X|$ and $\lim_{n \rightarrow \infty} \mathbb{E}\{|X_n - X|\} = 0$.

The next definition of uniform integrability is with respect to a family of probability measures for a fixed integrand.

Definition A.7. (*Uniform Integrability for a family of probability measures*)

Let $M \subset \mathcal{M}_1(\mathcal{X})$ be a family of probability measures on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$. A measurable function f on \mathcal{X} is said to be uniformly integrable over M if

$$\lim_{c \rightarrow \infty} \sup_{P \in M} \int_{\{x \in \mathcal{X} : |f(x)| > c\}} |f(x)| dP(x) = 0.$$

A sufficient condition for the convergence of a sequence of integrals of a function with respect to a weakly convergent sequence of measures is the following.

Theorem A.8. [29, Appendix, Theorem A.2, p. 3084]

Let $M \subset \mathcal{M}_1(\mathcal{X})$ be a closed family of probability measures on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$, and let $\{P_n : n \in \mathbb{N}_1\} \subset M$ be a weakly convergent sequence in M . If f is a continuous function on \mathcal{X} and uniformly integrable over $\{P_n : n \in \mathbb{N}_1\}$ then $\lim_{n \rightarrow \infty} \int f(x) dP_n(x) = \int f(x) dP(x)$.

Absolute Continuity of Probability Measures.

Let (Ω, \mathcal{F}) be a measurable space. Given two probability measures P, Q on (Ω, \mathcal{F}) , Q is said to be *absolutely continuous* with respect to P (denoted $P \ll Q$) if for every $A \in \mathcal{F}$ such that $P(A) = 0$ then $Q(A) = 0$. If $Q \ll P$, by Radon-Nikodym Derivative theorem, there exists a P -integrable and \mathcal{F} -measurable function f such that for every $A \in \mathcal{F}$, $Q(A) = \int_A f(\omega) dP(\omega)$. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and \mathcal{G} be a sub- σ -algebra of \mathcal{F} . A *regular conditional probability distribution* $P(\cdot | \mathcal{G})$ on (Ω, \mathcal{F}) exist, when \mathcal{G} is generated by a countable partition of Ω . Moreover, if (Ω, d) is a metric space which is complete and separable (Polish space), and \mathcal{F} is a Borel σ -algebra, then for any probability measure P on (Ω, \mathcal{F}) and any sub- σ -algebra $\mathcal{G} \subseteq \mathcal{F}$, a regular conditional probability measure of P given \mathcal{G} always exists.

The next lemma summarizes certain relationships between the absolute continuity of probability measures.

Lemma A.9. (*Absolute Continuity of Probability Measures*) [55, Lemma 4.4.7, pp. 149-150]

- a) Suppose $Q_{\mathcal{G}} \ll P_{\mathcal{G}}$. If $Q(\cdot | \mathcal{G})(\omega) \ll P(\cdot | \mathcal{G})(\omega)$, $Q_{\mathcal{G}} - a.s.$, then $Q \ll P$.
- b) Conversely, if $Q \ll P$, then $Q(\cdot | \mathcal{G})(\omega) \ll P(\cdot | \mathcal{G})(\omega)$, $P(\cdot | \mathcal{G})(\omega) - a.s.$

If $Y : (\Omega, \mathcal{F}) \mapsto (\mathcal{Y}, \mathcal{A})$ is a RV on (Ω, \mathcal{F}) into a measurable space $(\mathcal{Y}, \mathcal{A})$ and \mathcal{Y} is a Polish space, then a regular conditional distribution for Y given the sub- σ -algebra \mathcal{G} of \mathcal{F} denoted by $P(dy | \mathcal{G})(\omega)$, always exists. Additionally, if $X : (\Omega, \mathcal{F}) \mapsto (\mathcal{X}, \mathcal{B})$ is a RV on (Ω, \mathcal{F}) into a measurable space $(\mathcal{X}, \mathcal{B})$, and \mathcal{G} is the sub- σ -algebra of \mathcal{F} generated by X , then $P(dy | X)(\omega)$ is called the *regular conditional distribution* of Y given X . One can go one step further to define an equivalent definition of a regular conditional distribution for Y given $X = x$ as a quantity $P(dy | X = x)$ called stochastic kernel.

APPENDIX B

PROOF OF THEOREM III.3

1) Fix $\overleftarrow{P}_{0,n}(\cdot | y^{n-1}) \in \mathcal{M}_1^{C1}(\mathcal{X}_{0,n})$ and let $\overrightarrow{Q}_{0,n}^1(\cdot | x^n), \overrightarrow{Q}_{0,n}^2(\cdot | x^n) \in \mathcal{M}_1^{C2}(\mathcal{Y}_{0,n})$. Then, the joint distributions corresponding to $\overrightarrow{Q}_{0,n}^1(\cdot | x^n), \overrightarrow{Q}_{0,n}^2(\cdot | x^n)$ are

$$(\overleftarrow{P}_{0,n} \otimes \overrightarrow{Q}_{0,n}^1)(dx^n, dy^n) \text{ and } (\overleftarrow{P}_{0,n} \otimes \overrightarrow{Q}_{0,n}^2)(dx^n, dy^n),$$

and the marginals are

$$\nu_{0,n}^1(dy^n) = (\overleftarrow{P}_{0,n} \otimes \overrightarrow{Q}_{0,n}^1)(\mathcal{X}_{0,n}, dy^n), \quad \nu_{0,n}^2(dy^n) = (\overleftarrow{P}_{0,n} \otimes \overrightarrow{Q}_{0,n}^2)(\mathcal{X}_{0,n}, dy^n).$$

Since the set $\mathcal{M}_1^{\mathcal{C}2}(\mathcal{Y}_{0,n})$ is convex, given $\lambda \in (0, 1)$ there exists a probability measure $\tilde{\mathbf{P}}$ on $(\mathcal{X}^{\mathbb{N}_0} \times \mathcal{Y}^{\mathbb{N}_0}, \mathcal{B}(\mathcal{X}^{\mathbb{N}_0}) \otimes \mathcal{B}(\mathcal{Y}^{\mathbb{N}_0}))$ whose regular conditional measure $\mathbf{Q}(\cdot|\mathbf{x}) \in \mathcal{M}_1(\mathcal{Y}^{\mathbb{N}_0})$ satisfies

$$\vec{\mathcal{Q}}_{0,n}(\cdot|x^n) = \lambda \vec{\mathcal{Q}}_{0,n}^1(\cdot|x^n) + (1 - \lambda) \vec{\mathcal{Q}}_{0,n}^2(\cdot|x^n), \quad \tilde{\mathbf{P}}|_{\mathcal{B}(\mathcal{X}_{0,n})} - a.e. \ x^n$$

and C1 holds. Define

$$\nu_{0,n}(dy^n) = \lambda \nu_{0,n}^1(dy^n) + (1 - \lambda) \nu_{0,n}^2(dy^n).$$

Introduce the RNDs $\Lambda_{0,n}^i(x^n, y^n) = \frac{d\vec{\mathcal{Q}}_{0,n}^i(\cdot|x^n)}{d\nu_{0,n}^i(\cdot)}(y^n)$, $\Psi_{0,n}^i(x^n, y^n) = \frac{d\vec{\mathcal{Q}}_{0,n}^i(\cdot|x^n)}{d\nu_{0,n}(\cdot)}(y^n)$, $K_{0,n}^i(y^n) = \frac{d\nu_{0,n}^i(\cdot)}{d\nu_{0,n}(\cdot)}(y^n)$ and $\Lambda_{0,n}(x^n, y^n) = \frac{d\vec{\mathcal{Q}}_{0,n}(\cdot|x^n)}{d\nu_{0,n}(\cdot)}(y^n)$, $i = 1, 2$. Then,

$$\begin{aligned} \lambda \Psi_{0,n}^1(x^n, y^n) + (1 - \lambda) \Psi_{0,n}^2(x^n, y^n) &= \lambda \frac{d\vec{\mathcal{Q}}_{0,n}^1(\cdot|x^n)}{d\nu_{0,n}(\cdot)}(y^n) + (1 - \lambda) \frac{d\vec{\mathcal{Q}}_{0,n}^2(\cdot|x^n)}{d\nu_{0,n}(\cdot)}(y^n) \\ &= \frac{d(\lambda \vec{\mathcal{Q}}_{0,n}^1(\cdot|x^n) + (1 - \lambda) \vec{\mathcal{Q}}_{0,n}^2(\cdot|x^n))}{d(\lambda \nu_{0,n}^1(\cdot) + (1 - \lambda) \nu_{0,n}^2(\cdot))}(y^n) = \Lambda_{0,n}(x^n, y^n) \end{aligned}$$

and

$$\lambda K_{0,n}^1(y^n) + (1 - \lambda) K_{0,n}^2(y^n) = \lambda \frac{d\nu_{0,n}^1(\cdot)}{d\nu_{0,n}(\cdot)}(y^n) + (1 - \lambda) \frac{d\nu_{0,n}^2(\cdot)}{d\nu_{0,n}(\cdot)}(y^n) = \frac{d(\lambda \nu_{0,n}^1(\cdot) + (1 - \lambda) \nu_{0,n}^2(\cdot))}{d(\lambda \nu_{0,n}^1(\cdot) + (1 - \lambda) \nu_{0,n}^2(\cdot))}(y^n) = 1.$$

Applying the log-sum formula [56, Theorem 2.7.1, p. 31] yields

$$\begin{aligned} &\lambda \Psi_{0,n}^1(x^n, y^n) \log \Lambda_{0,n}^1(x^n, y^n) + (1 - \lambda) \Psi_{0,n}^2(x^n, y^n) \log \Lambda_{0,n}^2(x^n, y^n) \\ &= \lambda \Psi_{0,n}^1(x^n, y^n) \log \left(\frac{\frac{d\vec{\mathcal{Q}}_{0,n}^1(\cdot|x^n)}{d\nu_{0,n}(\cdot)}(y^n)}{\frac{d\nu_{0,n}^1(\cdot)}{d\nu_{0,n}(\cdot)}(y^n)} \right) + (1 - \lambda) \Psi_{0,n}^2(x^n, y^n) \log \left(\frac{\frac{d\vec{\mathcal{Q}}_{0,n}^2(\cdot|x^n)}{d\nu_{0,n}(\cdot)}(y^n)}{\frac{d\nu_{0,n}^2(\cdot)}{d\nu_{0,n}(\cdot)}(y^n)} \right) \\ &= \lambda \frac{d\vec{\mathcal{Q}}_{0,n}^1(\cdot|x^n)}{d\nu_{0,n}(\cdot)}(y^n) \log \left(\frac{\lambda \frac{d\vec{\mathcal{Q}}_{0,n}^1(\cdot|x^n)}{d\nu_{0,n}(\cdot)}(y^n)}{\lambda \frac{d\nu_{0,n}^1(\cdot)}{d\nu_{0,n}(\cdot)}(y^n)} \right) + (1 - \lambda) \frac{d\vec{\mathcal{Q}}_{0,n}^2(\cdot|x^n)}{d\nu_{0,n}(\cdot)}(y^n) \log \left(\frac{(1 - \lambda) \frac{d\vec{\mathcal{Q}}_{0,n}^2(\cdot|x^n)}{d\nu_{0,n}(\cdot)}(y^n)}{(1 - \lambda) \frac{d\nu_{0,n}^2(\cdot)}{d\nu_{0,n}(\cdot)}(y^n)} \right) \\ &\geq \left(\lambda \frac{d\vec{\mathcal{Q}}_{0,n}^1(\cdot|x^n)}{d\nu_{0,n}(\cdot)}(y^n) + (1 - \lambda) \frac{d\vec{\mathcal{Q}}_{0,n}^2(\cdot|x^n)}{d\nu_{0,n}(\cdot)}(y^n) \right) \log \left(\frac{\lambda \frac{d\vec{\mathcal{Q}}_{0,n}^1(\cdot|x^n)}{d\nu_{0,n}(\cdot)}(y^n) + (1 - \lambda) \frac{d\vec{\mathcal{Q}}_{0,n}^2(\cdot|x^n)}{d\nu_{0,n}(\cdot)}(y^n)}{\lambda \frac{d\nu_{0,n}^1(\cdot)}{d\nu_{0,n}(\cdot)}(y^n) + (1 - \lambda) \frac{d\nu_{0,n}^2(\cdot)}{d\nu_{0,n}(\cdot)}(y^n)} \right) \\ &= \frac{d\vec{\mathcal{Q}}_{0,n}(\cdot|x^n)}{d\nu_{0,n}(\cdot)}(y^n) \log \frac{d\vec{\mathcal{Q}}_{0,n}(\cdot|x^n)}{d\nu_{0,n}(\cdot)}(y^n). \end{aligned}$$

Integrating the above with respect to $\nu_{0,n}(dy^n) \otimes \overleftarrow{P}_{0,n}(dx^n|y^{n-1})$ yields:

$$\begin{aligned} &\int_{\mathcal{X}_{0,n} \times \mathcal{Y}_{0,n}} \log \left(\frac{d\vec{\mathcal{Q}}_{0,n}(\cdot|x^n)}{d\nu_{0,n}(\cdot)}(y^n) \right) \frac{d\vec{\mathcal{Q}}_{0,n}(\cdot|x^n)}{d\nu_{0,n}(\cdot)}(y^n) (\nu_{0,n}(dy^n) \otimes \overleftarrow{P}_{0,n}(dx^n|y^{n-1})) \\ &= \int_{\mathcal{X}_{0,n} \times \mathcal{Y}_{0,n}} \log \left(\frac{d\vec{\mathcal{Q}}_{0,n}(\cdot|x^n)}{d\nu_{0,n}(\cdot)}(y^n) \right) (\vec{\mathcal{Q}}_{0,n} \otimes \overleftarrow{P}_{0,n})(dx^n, dy^n) \\ &\leq \lambda \int_{\mathcal{X}_{0,n} \times \mathcal{Y}_{0,n}} \log \left(\frac{d\vec{\mathcal{Q}}_{0,n}^1(\cdot|x^n)}{d\nu_{0,n}^1(\cdot)}(y^n) \right) (\vec{\mathcal{Q}}_{0,n}^1 \otimes \overleftarrow{P}_{0,n})(dx^n, dy^n) \\ &\quad + (1 - \lambda) \int_{\mathcal{X}_{0,n} \times \mathcal{Y}_{0,n}} \log \left(\frac{d\vec{\mathcal{Q}}_{0,n}^2(\cdot|x^n)}{d\nu_{0,n}^2(\cdot)}(y^n) \right) (\vec{\mathcal{Q}}_{0,n}^2 \otimes \overleftarrow{P}_{0,n})(dx^n, dy^n). \end{aligned}$$

Hence,

$$\mathbb{I}_{X^n \rightarrow Y^n}(\overleftarrow{P}_{0,n}, \lambda \overrightarrow{Q}_{0,n}^1 + (1-\lambda) \overrightarrow{Q}_{0,n}^2) \leq \lambda \mathbb{I}_{X^n \rightarrow Y^n}(\overleftarrow{P}_{0,n}, \overrightarrow{Q}_{0,n}^1) + (1-\lambda) \mathbb{I}_{X^n \rightarrow Y^n}(\overleftarrow{P}_{0,n}, \overrightarrow{Q}_{0,n}^2).$$

This completes the derivation of 1).

2) Fix $\overrightarrow{Q}_{0,n}(\cdot|x^n) \in \mathcal{M}_1^{\mathbf{C2}}(\mathcal{Y}_{0,n})$ and let $\overleftarrow{P}_{0,n}^1(\cdot|y^{n-1}), \overleftarrow{P}_{0,n}^2(\cdot|y^{n-1}) \in \mathcal{M}_1^{\mathbf{C1}}(\mathcal{X}_{0,n})$. Then, the joint distributions corresponding to $\overleftarrow{P}_{0,n}^1(\cdot|y^{n-1}), \overleftarrow{P}_{0,n}^2(\cdot|y^{n-1})$ are

$$(\overleftarrow{P}_{0,n}^1 \otimes \overrightarrow{Q}_{0,n})(dx^n, dy^n) \text{ and } (\overleftarrow{P}_{0,n}^2 \otimes \overrightarrow{Q}_{0,n})(dx^n, dy^n).$$

The marginals corresponding to $\overleftarrow{P}_{0,n}^1(\cdot|y^{n-1}), \overleftarrow{P}_{0,n}^2(\cdot|y^{n-1})$ are

$$\nu_{0,n}^1(dy^n) = (\overleftarrow{P}_{0,n}^1 \otimes \overrightarrow{Q}_{0,n})(\mathcal{X}_{0,n}, dy^n), \quad \nu_{0,n}^2(dy^n) = (\overleftarrow{P}_{0,n}^2 \otimes \overrightarrow{Q}_{0,n})(\mathcal{X}_{0,n}, dy^n).$$

Since the set $\mathcal{M}_1^{\mathbf{C1}}(\mathcal{X}_{0,n})$ is convex, given $\lambda \in (0, 1)$ there exists a probability measure $\tilde{\mathbf{P}}$ on $(\mathcal{X}^{\mathbb{N}_0} \times \mathcal{Y}^{\mathbb{N}_0}, \mathcal{B}(\mathcal{X}^{\mathbb{N}_0}) \otimes \mathcal{B}(\mathcal{Y}^{\mathbb{N}_0}))$ whose regular conditional measure $\mathbf{P}(\cdot|y) \in \mathcal{M}_1(\mathcal{X}^{\mathbb{N}_0})$ satisfies

$$\overleftarrow{P}_{0,n}(\cdot|y^{n-1}) = \lambda \overleftarrow{P}_{0,n}^1(\cdot|y^{n-1}) + (1-\lambda) \overleftarrow{P}_{0,n}^2(\cdot|y^{n-1}), \quad \tilde{\mathbf{P}}|_{\mathcal{B}(\mathcal{Y}_{0,n-1})} - a.e. \ y^{n-1}$$

and **C2** holds. Then, corresponding to $\overleftarrow{P}_{0,n}(\cdot|y^{n-1})$ and $\overrightarrow{Q}_{0,n}(\cdot|x^n)$ we have

$$\begin{aligned} \nu_{0,n}(dy^n) &= \int_{\mathcal{X}_{0,n}} (\lambda \overleftarrow{P}_{0,n}^1(dx^n|y^{n-1}) + (1-\lambda) \overleftarrow{P}_{0,n}^2(dx^n|y^{n-1})) \otimes \overrightarrow{Q}_{0,n}(dy^n|x^n) \\ &= \lambda (\overleftarrow{P}_{0,n}^1 \otimes \overrightarrow{Q}_{0,n})(\mathcal{X}_{0,n}, dy^n) + (1-\lambda) (\overleftarrow{P}_{0,n}^2 \otimes \overrightarrow{Q}_{0,n})(\mathcal{X}_{0,n}, dy^n) = \lambda \nu_{0,n}^1(dy^n) + (1-\lambda) \nu_{0,n}^2(dy^n). \end{aligned}$$

Pick any measure $U_{0,n}(dy^n) \in \mathcal{M}_1(\mathcal{Y}_{0,n})$ with $\mathbb{D}(\nu_{0,n}||U_{0,n}) < \infty$, e.g., such that $\nu_{0,n}(\cdot) \ll U_{0,n}(\cdot)$. Since $\overrightarrow{Q}_{0,n}(\cdot|x^n) \ll \nu_{0,n}(\cdot)$, for almost all $x^n \in \mathcal{X}_{0,n}$, and $\nu_{0,n}(\cdot) \ll U_{0,n}(\cdot)$, then $\overrightarrow{Q}_{0,n}(\cdot|x^n) \ll U_{0,n}(\cdot)$, for almost all $x^n \in \mathcal{X}_{0,n}$. Consider

$$\begin{aligned} \mathbb{I}_{X^n \rightarrow Y^n}(\overleftarrow{P}_{0,n}, \overrightarrow{Q}_{0,n}) &= \int_{\mathcal{X}_{0,n} \times \mathcal{Y}_{0,n}} \log \left(\frac{d\overrightarrow{Q}_{0,n}(\cdot|x^n)}{d\nu_{0,n}(\cdot)}(y^n) \right) (\overrightarrow{Q}_{0,n} \otimes \overleftarrow{P}_{0,n})(dx^n, dy^n) \\ &= \int_{\mathcal{X}_{0,n} \times \mathcal{Y}_{0,n}} \log \left(\frac{d(\overrightarrow{Q}_{0,n}(\cdot|x^n) \times U_{0,n}(\cdot))}{d(\nu_{0,n}(\cdot) \times U_{0,n}(\cdot))}(y^n) \right) (\overrightarrow{Q}_{0,n} \otimes \overleftarrow{P}_{0,n})(dx^n, dy^n) \\ &= \int_{\mathcal{X}_{0,n} \times \mathcal{Y}_{0,n}} \log \left(\frac{d\overrightarrow{Q}_{0,n}(\cdot|x^n)}{dU_{0,n}(\cdot)}(y^n) \right) (\overrightarrow{Q}_{0,n} \otimes \overleftarrow{P}_{0,n})(dx^n, dy^n) \\ &\quad - \int_{\mathcal{X}_{0,n} \times \mathcal{Y}_{0,n}} \log \left(\frac{d\nu_{0,n}(\cdot)}{dU_{0,n}(\cdot)}(y^n) \right) (\overrightarrow{Q}_{0,n} \otimes \overleftarrow{P}_{0,n})(dx^n, dy^n) \\ &= \int_{\mathcal{X}_{0,n} \times \mathcal{Y}_{0,n}} \log \left(\frac{d\overrightarrow{Q}_{0,n}(\cdot|x^n)}{dU_{0,n}(dy^n)}(y^n) \right) (\overrightarrow{Q}_{0,n} \otimes \overleftarrow{P}_{0,n})(dx^n, dy^n) \\ &\quad - \int_{\mathcal{Y}_{0,n}} \log \left(\frac{d\nu_{0,n}(\cdot)}{dU_{0,n}(\cdot)}(y^n) \right) \left(\int_{\mathcal{X}_{0,n}} (\overrightarrow{Q}_{0,n} \otimes \overleftarrow{P}_{0,n})(dx^n, dy^n) \right) \\ &= \int_{\mathcal{X}_{0,n} \times \mathcal{Y}_{0,n}} \log \left(\frac{d\overrightarrow{Q}_{0,n}(\cdot|x^n)}{dU_{0,n}(\cdot)}(y^n) \right) (\overrightarrow{Q}_{0,n} \otimes \overleftarrow{P}_{0,n})(dx^n, dy^n) - \int_{\mathcal{Y}_{0,n}} \log \left(\frac{d\nu_{0,n}(\cdot)}{dU_{0,n}(\cdot)}(y^n) \right) \nu_{0,n}(dy^n). \end{aligned}$$

Hence,

$$\begin{aligned} \mathbb{I}_{X^n \rightarrow Y^n}(\lambda \overleftarrow{P}_{0,n}^1 + (1-\lambda) \overleftarrow{P}_{0,n}^2, \overrightarrow{Q}_{0,n}) &= \int_{\mathcal{X}_{0,n} \times \mathcal{Y}_{0,n}} \log \left(\frac{d\overrightarrow{Q}_{0,n}(\cdot|x^n)}{dU_{0,n}(\cdot)}(y^n) \right) \\ &\times \overrightarrow{Q}_{0,n}(dy^n|x^n) \otimes (\lambda \overleftarrow{P}_{0,n}^1(dx^n|y^{n-1}) + (1-\lambda) \overleftarrow{P}_{0,n}^2(dx^n|y^{n-1})) - \int_{\mathcal{Y}_{0,n}} \log \left(\frac{d\nu_{0,n}(\cdot)}{dU_{0,n}(\cdot)}(y^n) \right) \nu_{0,n}(dy^n). \end{aligned}$$

Moreover, relative entropy is convex in both arguments (e.g., $\mathbb{D}(\cdot||U_{0,n})$ is convex for fixed $U_{0,n}$), hence

$$\begin{aligned} \mathbb{I}_{X^n \rightarrow Y^n}(\lambda \overleftarrow{P}_{0,n}^1 + (1-\lambda) \overleftarrow{P}_{0,n}^2, \overrightarrow{Q}_{0,n}) &\geq \lambda \int_{\mathcal{X}_{0,n} \times \mathcal{Y}_{0,n}} \log \left(\frac{d\overrightarrow{Q}_{0,n}(\cdot|x^n)}{dU_{0,n}(\cdot)}(y^n) \right) (\overrightarrow{Q}_{0,n} \otimes \overleftarrow{P}_{0,n}^1)(dx^n, dy^n) \\ &- \lambda \int_{\mathcal{Y}_{0,n}} \log \left(\frac{d\nu_{0,n}^1(\cdot)}{dU_{0,n}(\cdot)}(y^n) \right) \nu_{0,n}^1(dy^n) \\ &+ (1-\lambda) \int_{\mathcal{X}_{0,n} \times \mathcal{Y}_{0,n}} \log \left(\frac{d\overrightarrow{Q}_{0,n}(\cdot|x^n)}{dU_{0,n}(\cdot)}(y^n) \right) (\overrightarrow{Q}_{0,n} \otimes \overleftarrow{P}_{0,n}^2)(dx^n, dy^n) \\ &- (1-\lambda) \int_{\mathcal{Y}_{0,n}} \log \left(\frac{d\nu_{0,n}^2(\cdot)}{dU_{0,n}(\cdot)}(y^n) \right) \nu_{0,n}^2(dy^n). \end{aligned}$$

Finally, since $\nu_{0,n}^1(\cdot) \ll U_{0,n}(\cdot)$ and $\nu_{0,n}^2(\cdot) \ll U_{0,n}(\cdot)$ by substituting the following versions

$\frac{d(\overrightarrow{Q}_{0,n}(\cdot|x^n) \times \nu_{0,n}^i(\cdot))}{d(U_{0,n}(\cdot) \times \nu_{0,n}^i(\cdot))}(y^n)$, $i = 1, 2$, of the RND for $\frac{d\overrightarrow{Q}_{0,n}(\cdot|x^n)}{dU_{0,n}(\cdot)}(y^n)$ in the first and third RHS expression in the preceding equations yields

$$\mathbb{I}_{X^n \rightarrow Y^n}(\lambda \overleftarrow{P}_{0,n}^1 + (1-\lambda) \overleftarrow{P}_{0,n}^2, \overrightarrow{Q}_{0,n}) \geq \lambda \mathbb{I}_{X^n \rightarrow Y^n}(\overleftarrow{P}_{0,n}^1, \overrightarrow{Q}_{0,n}) + (1-\lambda) \mathbb{I}_{X^n \rightarrow Y^n}(\overleftarrow{P}_{0,n}^2, \overrightarrow{Q}_{0,n}).$$

This completes the derivation of 2).

3) Here, it will be shown that for $\overrightarrow{Q}_{0,n}^1(\cdot|x^n), \overrightarrow{Q}_{0,n}^2(\cdot|x^n) \in \mathcal{M}_1^{\mathbf{C}2}(\mathcal{Y}_{0,n})$ such that $\overrightarrow{Q}_{0,n}^1(\cdot|x^n) \neq \overrightarrow{Q}_{0,n}^2(\cdot|x^n)$, and $\lambda \in (0, 1)$, then $\mathbb{I}_{X^n \rightarrow Y^n}(\overleftarrow{P}_{0,n}, \lambda \overrightarrow{Q}_{0,n}^1 + (1-\lambda) \overrightarrow{Q}_{0,n}^2) < \lambda \mathbb{I}_{X^n \rightarrow Y^n}(\overleftarrow{P}_{0,n}, \overrightarrow{Q}_{0,n}^1) + (1-\lambda) \mathbb{I}_{X^n \rightarrow Y^n}(\overleftarrow{P}_{0,n}, \overrightarrow{Q}_{0,n}^2)$, for a fixed $\overleftarrow{P}_{0,n}(\cdot|y^{n-1}) \in \mathcal{M}_1^{\mathbf{C}1}(\mathcal{X}_{0,n})$.

It is already known that $\mathbb{I}_{X^n \rightarrow Y^n}(\overleftarrow{P}_{0,n}, \overrightarrow{Q}_{0,n})$ is a convex functional on $\overrightarrow{Q}_{0,n}(\cdot|x^n) \in \mathcal{M}_1^{\mathbf{C}2}(\mathcal{Y}_{0,n})$ for a fixed $\overleftarrow{P}_{0,n}(\cdot|y^{n-1}) \in \mathcal{M}_1^{\mathbf{C}1}(\mathcal{X}_{0,n})$. All is required to show in order to have strict convexity is that $\mathbb{I}_{X^n \rightarrow Y^n}(\overleftarrow{P}_{0,n}, \overrightarrow{Q}_{0,n}) < \infty$. This can be easily obtained from part 1) since $\overleftarrow{P}_{0,n} \otimes \overrightarrow{Q}_{0,n} \ll \overleftarrow{P}_{0,n} \otimes \nu_{0,n}$ if and only if $\overrightarrow{Q}_{0,n}(\cdot|x^n) \ll \nu_{0,n}(\cdot)$, for $\mu_{0,n}$ -almost all $x^n \in \mathcal{X}_{0,n}$. Hence, from the strict convexity of the function $s \log s$, $s \in [0, \infty)$, and the expression of directed information as a functional of $\{\overleftarrow{P}_{0,n}(\cdot|y^{n-1}), \overrightarrow{Q}_{0,n}(\cdot|x^n)\} \in \mathcal{M}_1^{\mathbf{C}1}(\mathcal{X}_{0,n}) \times \mathcal{M}_1^{\mathbf{C}2}(\mathcal{Y}_{0,n})$, with $\overrightarrow{Q}_{0,n}(\cdot|x^n) = \lambda \overrightarrow{Q}_{0,n}^1(\cdot|x^n) + (1-\lambda) \overrightarrow{Q}_{0,n}^2(\cdot|x^n)$,

$\lambda) \vec{Q}_{0,n}^2(\cdot|x^n)$ it follows that

$$\begin{aligned}
\mathbb{I}_{X^n \rightarrow Y^n}(\overleftarrow{P}_{0,n}, \overrightarrow{Q}_{0,n}) &= \int_{\mathcal{X}_{0,n} \times \mathcal{Y}_{0,n}} \log \left(\frac{d(\overleftarrow{P}_{0,n} \otimes \overrightarrow{Q}_{0,n})(\cdot, \cdot)}{\overrightarrow{\Pi}(\cdot, \cdot)}(x^n, y^n) \right) (\overrightarrow{Q}_{0,n} \otimes \overleftarrow{P}_{0,n})(dx^n, dy^n) \\
&= \int_{\mathcal{X}_{0,n} \times \mathcal{Y}_{0,n}} \log \left(\frac{d\overrightarrow{Q}_{0,n}(\cdot|x^n)}{d\nu_{0,n}(\cdot)}(y^n) \right) (\overrightarrow{Q}_{0,n} \otimes \overleftarrow{P}_{0,n})(dx^n, dy^n) \\
&\leq \int_{\mathcal{X}_{0,n} \times \mathcal{Y}_{0,n}} \lambda \log \left(\frac{d\overrightarrow{Q}_{0,n}^1(\cdot|x^n)}{d\nu_{0,n}(\cdot)}(y^n) \right) (\overrightarrow{Q}_{0,n}^1 \otimes \overleftarrow{P}_{0,n})(dx^n, dy^n) \\
&\quad + \int_{\mathcal{X}_{0,n} \times \mathcal{Y}_{0,n}} (1 - \lambda) \log \left(\frac{d\overrightarrow{Q}_{0,n}^2(\cdot|x^n)}{d\nu_{0,n}(\cdot)}(y^n) \right) (\overrightarrow{Q}_{0,n}^2 \otimes \overleftarrow{P}_{0,n})(dx^n, dy^n) \\
&= \lambda \mathbb{I}_{X^n \rightarrow Y^n}(\overleftarrow{P}_{0,n}, \overrightarrow{Q}_{0,n}^1) + (1 - \lambda) \mathbb{I}_{X^n \rightarrow Y^n}(\overleftarrow{P}_{0,n}, \overrightarrow{Q}_{0,n}^2) < \infty.
\end{aligned}$$

This completes the derivation of 3).

APPENDIX C PROOF OF THEOREM III.5

Part A. Let $\overrightarrow{Q}_{0,n}^\alpha(\cdot|\cdot) \in \mathcal{Q}^{\mathbf{C}^2}(\mathcal{Y}_{0,n}|\mathcal{X}_{0,n})$, $\alpha = 1, 2, \dots$, be a sequence of forward channels and $(X^{n,(\alpha)}, Y^{n,(\alpha)})$, $\alpha = 1, 2, \dots$ a sequence of the basic joint process corresponding to the backward channel $\overleftarrow{P}_{0,n}(\cdot|\cdot) \in \mathcal{Q}^{\mathbf{C}^1}(\mathcal{X}_{0,n}|\mathcal{Y}_{0,n-1})$ and the sequence of forward channels $\overrightarrow{Q}_{0,n}^\alpha(\cdot|\cdot) \in \mathcal{Q}^{\mathbf{C}^2}(\mathcal{Y}_{0,n}|\mathcal{X}_{0,n})$, $\alpha = 1, 2, \dots$. The important steps for the derivation of A1) are outlined in [32] for stochastic control problems with randomized controls. Since we shall use A1) and parts of its derivation to show A2)–A4), we give the details of the derivation.

A1) First, it is shown that the joint distribution of the basic joint process $\{(X_i^{(\alpha)}, Y_i^{(\alpha)}) : i \in \mathbb{N}_0\}$ converges as $\alpha \rightarrow \infty$ to the joint distribution of a joint process $\{(X_i^{(o)}, Y_i^{(o)}) : i \in \mathbb{N}_0\}$ and secondly, that this limiting joint process $\{(X_i^{(o)}, Y_i^{(o)}) : i \in \mathbb{N}_0\}$ is also a basic joint process corresponding to the backward channel $\overleftarrow{P}_{0,n}(\cdot|\cdot) \in \mathcal{Q}^{\mathbf{C}^1}(\mathcal{X}_{0,n}|\mathcal{Y}_{0,n-1})$, that is, $(\overleftarrow{P}_{0,n} \otimes \overrightarrow{Q}_{0,n}^\alpha)(dx^n, dy^n) \xrightarrow{w} (\overleftarrow{P}_{0,n} \otimes \overrightarrow{Q}_{0,n}^o)(dx^n, dy^n) \in \mathcal{M}_1(\mathcal{X}_{0,n} \times \mathcal{Y}_{0,n})$ and that $(\overleftarrow{P}_{0,n} \otimes \overrightarrow{Q}_{0,n}^o)(dx^n, dy^n)$ has backward channel $\overleftarrow{P}_{0,n}(\cdot|\cdot) \in \mathcal{Q}^{\mathbf{C}^1}(\mathcal{X}_{0,n}|\mathcal{Y}_{0,n-1})$, but $\overrightarrow{Q}_{0,n}^o(\cdot|x^n) \in \mathcal{M}_1(\mathcal{Y}_{0,n})$ is not necessarily an element of $\mathcal{M}_1^{\mathbf{C}^2}(\mathcal{Y}_{0,n})$.

For any $g(\cdot) \in BC(\mathcal{X}_n)$, by condition **CA**, the function

$$f : \mathcal{X}_{0,n-1} \times \mathcal{Y}_{0,n-1} \mapsto \mathbb{R}, \quad f(x^{n-1}, y^{n-1}) \triangleq \int_{\mathcal{X}_n} g(x) p_n(dx_n | x^{n-1}, y^{n-1})$$

is continuous, and hence for any compact sets $K_i \in \mathcal{X}_i$, $i = 0, 1, \dots, n-1$, and by the compactness of $\mathcal{Y}_{0,n-1}$, the image of $f(\cdot, \cdot)$ under $K_{0,n-1} \times \mathcal{Y}_{0,n-1} \triangleq K_0 \times K_1 \times \dots \times K_{n-1} \times \mathcal{Y}_{0,n-1}$, $f(K_{0,n-1} \times \mathcal{Y}_{0,n-1}) = \mathcal{R} \subset \mathbb{R}$, and \mathcal{R} is compact (since the image of any real-valued continuous function on a compact set is compact). Thus, by condition **CA** and the compactness of $\{\mathcal{Y}_i : i \in \mathbb{N}_0^n\}$, for any compact sets $K_0 \in \mathcal{X}_0, K_1 \in \mathcal{X}_1, \dots, K_{n-1} \in \mathcal{X}_{n-1}$ the family of distributions $\{p_n(\cdot|x^{n-1}, y^{n-1}) : x_0 \in K_0, x_1 \in K_1, \dots, x_{n-1} \in K_{n-1}, y^{n-1} \in \mathcal{Y}_{0,n-1}\}$ is compact. Indeed, given any sequence $\{x_0^{(\alpha)}, \dots, x_{n-1}^{(\alpha)}, y_0^{(\alpha)}, \dots, y_{n-1}^{(\alpha)}\}$, by selecting a subsequence α_i such that the subsequence $\{x_0^{(\alpha_i)}, \dots, x_{n-1}^{(\alpha_i)}, y_0^{(\alpha_i)}, \dots, y_{n-1}^{(\alpha_i)}\}$ converges to $\{x_0^{(o)}, \dots, x_{n-1}^{(o)}, y_0^{(o)}, \dots, y_{n-1}^{(o)}\}$, a weakly convergent subsequence of measures $p_n(\cdot|x_0^{(\alpha_i)}, \dots, x_{n-1}^{(\alpha_i)}, y_0^{(\alpha_i)}, \dots, y_{n-1}^{(\alpha_i)})$ is obtained. Utilizing Prohorov's theorem (see Theorem A.2), we verify that for any sequence of compact sets $K_0 \subset \mathcal{X}_0, K_1 \subset \mathcal{X}_1, \dots, K_{n-1} \subset \mathcal{X}_{n-1}$, and $\epsilon_1 > 0$ a compact set $K_n \subset \mathcal{X}_n$ can be constructed such that $p_n(K_n | x^{n-1}, y^{n-1}) \geq 1 - \epsilon_1$, for any $y^{n-1} \in \mathcal{Y}_{0,n-1}$. To this end, pick $\epsilon_1 > 0$ and construct the compact sets as follows. Choose compact set $K_0 \subset \mathcal{X}_0$ such that $p_0(K_0) \geq 1 - \frac{\epsilon_1}{2}$, compact set $K_1 \subset \mathcal{X}_1$ such that $p_1(K_1 | x_0, y_0) \geq 1 - \frac{\epsilon_1}{2^2}$, for any $x_0 \in K_0, y_0 \in \mathcal{Y}_0$,

compact set $K_2 \subset \mathcal{X}_2$ such that $p_2(K_2|x_0, x_1, y_0, y_1) \geq 1 - \frac{\epsilon_1}{2^3}$, for any $x_0 \in K_0, x_1 \in K_1, y_0 \in \mathcal{Y}_0, y_1 \in \mathcal{Y}_1$, and compact set K_n such that

$$p_n(K_n|x^{n-1}, y^{n-1}) \geq 1 - \frac{\epsilon_1}{2^{n+1}}. \quad (\text{C.1})$$

Utilizing (C.1) then

$$\begin{aligned} \mathbb{P}\{X_0^{(\alpha)} \in K_0, \dots, X_n^{(\alpha)} \in K_n\} &= \mathbb{P}\{X_0^{(\alpha)} \in K_0, \dots, X_n^{(\alpha)} \in K_n, Y_0^{(\alpha)} \in \mathcal{Y}_0, \dots, Y_{n-1}^{(\alpha)} \in \mathcal{Y}_{n-1}\} \\ &= \int_{\times_{i=0}^n K_i} \int_{\mathcal{Y}_{0,n-1}} \mathbb{P}\{X_n^{(\alpha)} \in K_n | X_0^{(\alpha)} = x_0, \dots, X_{n-1}^{(\alpha)} = x_{n-1}, Y_0^{(\alpha)} = y_0, \dots, Y_{n-1}^{(\alpha)} = y_{n-1}\} \\ &\quad \mathbb{P}\{X_0^{(\alpha)} \in dx_0, \dots, X_{n-1}^{(\alpha)} \in dx_{n-1}, Y_0^{(\alpha)} \in dy_0, \dots, Y_{n-1}^{(\alpha)} \in dy_{n-1}\} \\ &\geq \left(1 - \frac{\epsilon_1}{2^{n+1}}\right) \int_{\times_{i=0}^{n-1} K_i} \mathbb{P}\{X_0^{(\alpha)} \in dx_0, \dots, X_{n-1}^{(\alpha)} \in dx_{n-1}\} \\ &= \left(1 - \frac{\epsilon_1}{2^{n+1}}\right) \mathbb{P}\{X_0^{(\alpha)} \in K_0, \dots, X_{n-1}^{(\alpha)} \in K_{n-1}\} \\ &\geq \left(1 - \frac{\epsilon_1}{2^{n+1}}\right) \left(1 - \frac{\epsilon_1}{2^n}\right) \mathbb{P}\{X_0^{(\alpha)} \in K_0, \dots, X_{n-2}^{(\alpha)} \in K_{n-2}\} \\ &= \left(1 - \frac{\epsilon_1}{2^{n+1}} - \frac{\epsilon_1}{2^n} + \frac{\epsilon_1^2}{2^{2n+1}}\right) \mathbb{P}\{X_0^{(\alpha)} \in K_0, \dots, X_{n-2}^{(\alpha)} \in K_{n-2}\} \\ &\geq \left(1 - \frac{\epsilon_1}{2^{n+1}} - \frac{\epsilon_1}{2^n}\right) \mathbb{P}\{X_0^{(\alpha)} \in K_0, \dots, X_{n-2}^{(\alpha)} \in K_{n-2}\} \\ &\geq \left(1 - \frac{\epsilon_1}{2^{n+1}} - \frac{\epsilon_1}{2^n}\right) \left(1 - \frac{\epsilon_1}{2^{n-1}}\right) \mathbb{P}\{X_0^{(\alpha)} \in K_0, \dots, X_{n-3}^{(\alpha)} \in K_{n-3}\} \\ &= \left(1 - \frac{\epsilon_1}{2^{n+1}} - \frac{\epsilon_1}{2^n} - \frac{\epsilon_1}{2^{n-1}} + \frac{\epsilon_1^2}{2^{2n}} + \frac{\epsilon_1^2}{2^{2n-1}}\right) \mathbb{P}\{X_0^{(\alpha)} \in K_0, \dots, X_{n-3}^{(\alpha)} \in K_{n-3}\} \\ &\geq \left(1 - \frac{\epsilon_1}{2^{n+1}} - \frac{\epsilon_1}{2^n} - \frac{\epsilon_1}{2^{n-1}}\right) \mathbb{P}\{X_0^{(\alpha)} \in K_0, \dots, X_{n-3}^{(\alpha)} \in K_{n-3}\}. \end{aligned} \quad (\text{C.2})$$

Iterating the RHS of (C.2) we obtain

$$\begin{aligned} \mathbb{P}\{X_0^{(\alpha)} \in K_0, \dots, X_n^{(\alpha)} \in K_n\} &\geq 1 - \frac{\epsilon_1}{2^{n+1}} - \frac{\epsilon_1}{2^n} - \frac{\epsilon_1}{2^{n-1}} - \dots - \frac{\epsilon_1}{2^1} = 1 - \epsilon_1 \sum_{i=1}^n \frac{1}{2^{i+1}} \\ &\geq 1 - \epsilon_1, \quad \text{for all } \alpha = 1, 2, \dots, \text{ and any } n \in \mathbb{N}_0. \end{aligned} \quad (\text{C.3})$$

By (C.3), the family of marginal distributions of the joint process $\{(X_i^{(\alpha)}, Y_i^{(\alpha)}) : i \in \mathbb{N}_0\}$, $\alpha = 1, 2, \dots$ on $\mathcal{X}_{0,n}$ is uniformly tight, and by Prohorov's theorem [57] it has a weakly convergent subsequence. On the other hand, since $\{\mathcal{Y}_i : i \in \mathbb{N}_0^n\}$ are compact metric spaces, the family of marginal distributions of the joint sequence $\{(X_i^{(\alpha)}, Y_i^{(\alpha)}) : i \in \mathbb{N}_0\}$ on $\mathcal{Y}_{0,n}$ is uniformly tight. Utilizing the uniform tightness of the marginal distribution of the joint process $\{(X_i^{(\alpha)}, Y_i^{(\alpha)}) : i \in \mathbb{N}_0\}$, then the family of joint distributions of the joint process $\{(X_i^{(\alpha)}, Y_i^{(\alpha)}) : i \in \mathbb{N}_0\}$ is uniformly tight. By Prohorov's theorem [57], the sequence of joint distribution of the joint process $\{(X_i^{(\alpha)}, Y_i^{(\alpha)}) : i \in \mathbb{N}_0\}$ possess a weakly convergent subsequence to a joint process $\{(X_i^{(o)}, Y_i^{(o)}) : i \in \mathbb{N}_0\}$. A restatement of Prohorov's theorem states that, if \mathcal{Z} is a

separable metric space then every uniformly tight sequence of measures $\{\gamma^\alpha : \alpha = 1, 2, \dots\}$ on \mathcal{Z} has a subsubsequence which is weakly convergent. Moreover, by [57], if each subsequence $\{\gamma^{\alpha_i} : i = 1, 2, \dots\}$ of $\{\gamma^\alpha : \alpha = 1, 2, \dots\}$ contains a further subsequence $\{\gamma^{\alpha_{im}} : m = 1, 2, \dots\}$ such that $\gamma^{\alpha_{im}} \xrightarrow{w} \gamma^o$ as $m \rightarrow \infty$, then $\gamma^\alpha \xrightarrow{w} \gamma^o$ as $\alpha \rightarrow \infty$. Utilizing these facts, then the joint distribution of the joint process $\{(X_i^{(\alpha)}, Y_i^{(\alpha)}) : i \in \mathbb{N}_0\}$ converges weakly to a joint process $\{(X_i^{(o)}, Y_i^{(o)}) : i \in \mathbb{N}_0\}$. Next, we show that the limiting joint process $\{(X_i^{(o)}, Y_i^{(o)}) : i \in \mathbb{N}_0\}$ is a basic joint process with the same backward channel $\bar{P}(\cdot|\cdot) \in \mathcal{Q}^{\mathbf{C1}}(\mathcal{X}_{0,n}|\mathcal{Y}_{0,n-1})$. For any $n \in \mathbb{N}_0$, consider bounded and continuous real-valued functions $g_n(\cdot) \in BC(\mathcal{X}_n)$ and $\Psi_{0,n-1}(\cdot, \cdot) \in BC(\mathcal{X}_{0,n-1} \times \mathcal{Y}_{0,n-1})$. By the weak convergence of the joint measures corresponding to $\{(X_i^{(\alpha)}, Y_i^{(\alpha)}) : i \in \mathbb{N}_0\}$ to the joint measures corresponding to $\{(X_i^{(o)}, Y_i^{(o)}) : i \in \mathbb{N}_0\}$ denoted by $(\bar{P}_{0,n} \otimes \bar{Q}_{0,n}^\alpha)(dx^n, dy^n) \xrightarrow{w} P_{0,n}^o(dx^n, dy^n)$, the continuity of $g_n(\cdot)$ and the continuity of the function mapping $(x^{n-1}, y^{n-1}) \in \mathcal{X}_{0,n-1} \times \mathcal{Y}_{0,n-1} \mapsto \int_{\mathcal{X}_n} g_n(x) p_n(dx|x^{n-1}, y^{n-1}) \in \mathbb{R}$, given $\epsilon > 0$ there exists $N \in \mathbb{N}_0$ such that for all $\alpha \geq N$

$$\left| \int_{\mathcal{X}_{0,n-1} \times \mathcal{Y}_{0,n-1}} \left(\int_{\mathcal{X}_n} g_n(x) p_n(dx|x^{n-1}, y^{n-1}) \right) \Psi_{0,n-1}(x^{n-1}, y^{n-1}) P_{0,n-1}^o(dx^{n-1}, dy^{n-1}) \right. \\ \left. - \int_{\mathcal{X}_{0,n-1} \times \mathcal{Y}_{0,n-1}} \left(\int_{\mathcal{X}_n} g_n(x) p_n(dx|x^{n-1}, y^{n-1}) \right) \Psi_{0,n-1}(x^{n-1}, y^{n-1}) P_{0,n-1}^\alpha(dx^{n-1}, dy^{n-1}) \right| \leq \epsilon.$$

Since $\epsilon > 0$ is arbitrary, then

$$\lim_{\alpha \rightarrow \infty} \mathbb{E} \left\{ g_n(X_n^{(\alpha)}) \Psi(X_0^{(\alpha)}, \dots, X_{n-1}^{(\alpha)}, Y_0^{(\alpha)}, \dots, Y_{n-1}^{(\alpha)}) \right\} = \mathbb{E} \left\{ g_n(X_n^{(o)}) \Psi(X_0^{(o)}, \dots, X_{n-1}^{(o)}, Y_0^{(o)}, \dots, Y_{n-1}^{(o)}) \right\}. \quad (\text{C.4})$$

Moreover, for all $\alpha = 1, 2, \dots$, then

$$\begin{aligned} & \mathbb{E} \left\{ g_n(X_n^{(\alpha)}) \Psi(X_0^{(\alpha)}, \dots, X_{n-1}^{(\alpha)}, Y_0^{(\alpha)}, \dots, Y_{n-1}^{(\alpha)}) \right\} \\ &= \mathbb{E} \left\{ \Psi(X_0^{(\alpha)}, \dots, X_{n-1}^{(\alpha)}, Y_0^{(\alpha)}, \dots, Y_{n-1}^{(\alpha)}) \mathbb{E} \left\{ g_n(X_n^{(\alpha)}) | X_0^{(\alpha)}, \dots, X_{n-1}^{(\alpha)}, Y_0^{(\alpha)}, \dots, Y_{n-1}^{(\alpha)} \right\} \right\} \\ &= \mathbb{E} \left\{ \left(\int_{\mathcal{X}_n} g_n(x) p_n(dx | X_0^{(\alpha)}, \dots, X_{n-1}^{(\alpha)}, Y_0^{(\alpha)}, \dots, Y_{n-1}^{(\alpha)}) \right) \Psi(X_0^{(\alpha)}, \dots, X_{n-1}^{(\alpha)}, Y_0^{(\alpha)}, \dots, Y_{n-1}^{(\alpha)}) \right\}. \end{aligned}$$

Hence, (C.4) is equivalent to

$$\begin{aligned} & \lim_{\alpha \rightarrow \infty} \mathbb{E} \left\{ \int_{\mathcal{X}_n} g_n(x) p_n(dx | X_0^{(\alpha)}, \dots, X_{n-1}^{(\alpha)}, Y_0^{(\alpha)}, \dots, Y_{n-1}^{(\alpha)}) \Psi(X_0^{(\alpha)}, \dots, X_{n-1}^{(\alpha)}, Y_0^{(\alpha)}, \dots, Y_{n-1}^{(\alpha)}) \right\} \\ &= \mathbb{E} \left\{ \int_{\mathcal{X}_n} g_n(x) p_n(dx | X_0^{(o)}, \dots, X_{n-1}^{(o)}, Y_0^{(o)}, \dots, Y_{n-1}^{(o)}) \Psi(X_0^{(o)}, \dots, X_{n-1}^{(o)}, Y_0^{(o)}, \dots, Y_{n-1}^{(o)}) \right\}. \end{aligned}$$

From the previous equality, the following identity is obtained.

$$\mathbb{E} \left\{ g_n(X_n^{(o)}) | X_0^{(o)}, \dots, X_{n-1}^{(o)}, Y_0^{(o)}, \dots, Y_{n-1}^{(o)} \right\} = \int_{\mathcal{X}_n} g_n(x) p_n(dx | X_0^{(o)}, \dots, X_{n-1}^{(o)}, Y_0^{(o)}, \dots, Y_{n-1}^{(o)}) - a.s. \quad (\text{C.5})$$

Since for any indicator function I_E , $E \in \mathcal{B}(\mathcal{X}_n)$ there exists a sequence $\{g_{n,j} : j = 1, 2, \dots\} \subset BC(\mathcal{X}_n)$ which is nondecreasing such that $g_{n,j} \uparrow I_E$, by utilizing such a sequence in (C.5), and by invoking Lebesgue's monotone convergence theorem then

$$\mathbb{P} \left\{ X_n^{(o)} \in E | X_0^{(o)}, \dots, X_{n-1}^{(o)}, Y_0^{(o)}, \dots, Y_{n-1}^{(o)} \right\} = p_n(E | X_0^{(o)}, \dots, X_{n-1}^{(o)}, Y_0^{(o)}, \dots, Y_{n-1}^{(o)}).$$

This shows that the limiting joint process $\{(X_i^{(o)}, Y_i^{(o)}) : i \in \mathbb{N}_0\}$ is a basic process corresponding to the backward channel $\overleftarrow{P}_{0,n}(\cdot|y^{n-1}) \in \mathcal{M}_1^{\mathbf{C1}}(\mathcal{X}_{0,n})$ and a forward channel $\overrightarrow{Q}_{0,n}(\cdot|x^n) \in \mathcal{M}_1(\mathcal{Y}_{0,n})$. Moreover, the marginal distributions of the basic joint process $\{(X_i^{(\alpha)}, Y_i^{(\alpha)}) : i \in \mathbb{N}_0\}$ converge to the marginal distributions of the basic joint process $\{(X_i^{(o)}, Y_i^{(o)}) : i \in \mathbb{N}_0\}$ corresponding to the backward channel $\overleftarrow{P}_{0,n}(\cdot|y^{n-1}) \in \mathcal{M}_1^{\mathbf{C1}}(\mathcal{X}_{0,n})$ and a forward channel $\overrightarrow{Q}_{0,n}(\cdot|x^n) \in \mathcal{M}_1(\mathcal{Y}_{0,n})$. This completes the derivation of A1).

A2) By consistency condition **C1**, any $\overleftarrow{P}_{0,n}(\cdot|\cdot) \in \mathcal{Q}^{\mathbf{C1}}(\mathcal{X}_{0,n}|\mathcal{Y}_{0,n-1})$ uniquely defines a family $\{p_i(\cdot|\cdot, \cdot) \in \mathcal{Q}(\mathcal{X}_i|\mathcal{X}_{0,i-1} \times \mathcal{Y}_{0,i-1}), i \in \mathbb{N}_0^n\}$ via (II.1). Hence, (II.1) can be used to relate tightness of $p_i(\cdot|x^{i-1}, y^{i-1}) \in \mathcal{M}_1(\mathcal{X}_i)$, $(x^{i-1}, y^{i-1}) \in \mathcal{X}_{0,i-1} \times \mathcal{Y}_{0,i-1}$, $i \in \mathbb{N}_0^n$, to tightness of $\overleftarrow{P}_{0,n}(\cdot|y^{n-1}) \in \mathcal{M}_1^{\mathbf{C1}}(\mathcal{X}_{0,n})$, $y^{n-1} \in \mathcal{Y}_{0,n-1}$.

By recalling the derivation A1), condition (C.1), for $K_{0,n} = \times_{i=0}^n K_i$, $K_i \in \mathcal{B}(\mathcal{X}_i)$ compact sets, $i \in \mathbb{N}_0^n$, then

$$\begin{aligned} \mathbf{P}(K_{0,n}|\mathbf{y}) &\triangleq \int_{K_0} p_0(dx_0) \int_{K_1} p_1(dx_1|x^0, y^0) \dots \int_{K_n} p_n(dx_n|x^{n-1}, y^{n-1}) \\ &\geq \left(1 - \frac{\epsilon_1}{2^{n+1}}\right) \int_{K_0} p_0(dx_0) \int_{K_1} p_1(dx_1|x^0, y^0) \dots \int_{K_{n-1}} p_{n-1}(dx_{n-1}|x^{n-2}, y^{n-2}) \\ &\geq \left(1 - \frac{\epsilon_1}{2^{n+1}}\right) \left(1 - \frac{\epsilon_1}{2^n}\right) \int_{K_0} p_0(dx_0) \int_{K_1} p_1(dx_1|x^0, y^0) \dots \int_{K_{n-2}} p_{n-2}(dx_{n-2}|x^{n-3}, y^{n-3}) \\ &= \left(1 - \frac{\epsilon_1}{2^{n+1}} - \frac{\epsilon_1}{2^n} + \frac{\epsilon_1^2}{2^{2n+1}}\right) \int_{K_0} p_0(dx_0) \int_{K_1} p_1(dx_1|x^0, y^0) \dots \int_{K_{n-2}} p_{n-2}(dx_{n-2}|x^{n-3}, y^{n-3}) \\ &\geq \left(1 - \frac{\epsilon_1}{2^{n+1}} - \frac{\epsilon_1}{2^n}\right) \int_{K_0} p_0(dx_0) \int_{K_1} p_1(dx_1|x^0, y^0) \dots \int_{K_{n-2}} p_{n-2}(dx_{n-2}|x^{n-3}, y^{n-3}). \end{aligned}$$

By repeating the above procedure the following bound is obtained.

$$\begin{aligned} \mathbf{P}(K_{0,n}|\mathbf{y}) &\geq 1 - \frac{\epsilon_1}{2^{n+1}} - \frac{\epsilon_1}{2^n} - \frac{\epsilon_1}{2^{n-1}} - \dots - \frac{\epsilon_1}{2^1} = 1 - \epsilon_1 \sum_{i=1}^n \frac{1}{2^{i+1}} \\ &\geq 1 - \epsilon_1, \quad \text{for any } n \in \mathbb{N}_0 \text{ and for every } \mathbf{y} \in \mathcal{Y}^{\mathbb{N}_0}. \end{aligned}$$

Since $\{K_i : i = 0, 1, \dots, n\}$ are compact, from the last inequality it follows that the family of measures $\overleftarrow{P}_{0,n}(\cdot|y^{n-1}) \in \mathcal{M}_1^{\mathbf{C1}}(\mathcal{X}_{0,n})$, $y^{n-1} \in \mathcal{Y}_{0,n-1}$ is uniformly tight. This completes the derivation of A2).

A3) Weak compactness of the family of measures $\overrightarrow{Q}_{0,n}(\cdot|x^n) \in \mathcal{M}_1^{\mathbf{C2}}(\mathcal{Y}_{0,n})$ for fixed $x^n \in \mathcal{X}_{0,n}$ follows from the fact that $\mathcal{Y}_{0,n}$ is a compact Polish space.

A4) Utilizing the weak convergence $\nu_{0,n}^\alpha \xrightarrow{w} \nu_{0,n}^o$ (shown in A2)), we shall show weak convergence of the convolution of measures $\overrightarrow{\Pi}_{0,n}^\alpha(dx^n, dy^n) \equiv \overleftarrow{P}_{0,n}(dx^n|y^{n-1}) \otimes \nu_{0,n}^\alpha(dy^n) \xrightarrow{w} \overleftarrow{P}_{0,n}(dx^n|y^{n-1}) \otimes \nu_{0,n}^o(dy^n) \equiv \overrightarrow{\Pi}_{0,n}^o(dx^n, dy^n)$, when $\overleftarrow{P}_{0,n}(\cdot|y^{n-1}) \in \mathcal{M}_1^{\mathbf{C1}}(\mathcal{X}_{0,n})$ is fixed. We show weak convergence by considering integrals with respect to $g_{0,n}(x^n)h_{0,n}(y^n)$, where $g_{0,n}(\cdot) \in BC(\mathcal{X}_{0,n})$ and $h_{0,n}(\cdot) \in BC(\mathcal{Y}_{0,n})$. Let $\epsilon > 0$ be given. Condition CA implies that the function mapping

$$y^{n-1} \in \mathcal{Y}_{0,n-1} \mapsto \int_{\mathcal{X}_{0,n}} g(x^n) \overleftarrow{P}_{0,n}(dx^n|y^{n-1}) \in \mathbb{R} \quad (\text{C.6})$$

is continuous. Hence, by the weak convergence $\nu_{0,n}^\alpha \xrightarrow{w} \nu_{0,n}^o$ and the continuity of the function mapping (C.6) then there exists $N \in \mathbb{N}_0$ such that for all $\alpha \geq N$

$$\left| \int_{\mathcal{Y}_{0,n}} \left(\int_{\mathcal{X}_{0,n}} g(x^n) \overleftarrow{P}_{0,n}(dx^n|y^{n-1}) \right) h(y^n) \nu_{0,n}^o(dy^n) - \int_{\mathcal{Y}_{0,n}} \left(\int_{\mathcal{X}_{0,n}} g(x^n) \overleftarrow{P}_{0,n}(dx^n|y^{n-1}) \right) h(y^n) \nu_{0,n}^\alpha(dy^n) \right| \leq \epsilon.$$

Since $\epsilon > 0$ is arbitrary, then the derivation of A5) is complete.

Part B. The methodology is similar to that of **Part A.**, hence it is omitted.

APPENDIX D PROOF OF LEMMA III.6

By Theorem III.5, **Part A.**, A2), the family of measures $\overleftarrow{P}_{0,n}(\cdot|y^{n-1}) \in \mathcal{M}_1^{\text{C1}}(\mathcal{X}_{0,n})$, $y^{n-1} \in \mathcal{Y}_{0,n-1}$ are tight, and by Appendix C, (C.1), $\{p_i(\cdot|x^{i-1}, y^{i-1}) \in \mathcal{M}_1^{\text{C1}}(\mathcal{X}_i) : i = 0, 1, \dots, n\}$ are tight. Since $p_i(\cdot|x^{i-1}, y^{i-1})$ are probability measures on $\mathcal{M}_1^{\text{C1}}(\mathcal{X}_i)$, $i = 0, 1, \dots, n$, for any sequence $\overleftarrow{P}_{0,n}^\alpha(\cdot|y^{n-1}) \in \mathcal{M}_1^{\text{C1}}(\mathcal{X}_{0,n})$, $\alpha = 1, 2, \dots$, there is a collection $\{p_i^\alpha(\cdot|x^{i-1}, y^{i-1}) : i = 0, 1, \dots, n\}$, $\alpha = 1, 2, \dots$, such that

$$p_i^\alpha(\cdot|x^{i-1}, y^{i-1}) \xrightarrow{w} p_i^o(\cdot|x^{i-1}, y^{i-1}), \quad i = 0, 1, \dots, n.$$

Hence, to show closedness of $\overleftarrow{P}_{0,n}(\cdot|y^{n-1}) \in \mathcal{M}_1^{\text{C1}}(\mathcal{X}_{0,n})$, $y^{n-1} \in \mathcal{Y}_{0,n-1}$ it suffices to show that

$$\otimes_{i=0}^n p_i^\alpha(\cdot|x^{i-1}, y^{i-1}) \xrightarrow{w} \otimes_{i=0}^n p_i^o(\cdot|x^{i-1}, y^{i-1})$$

whenever $p_i^\alpha(\cdot|x^{i-1}, y^{i-1}) \xrightarrow{w} p_i^o(\cdot|x^{i-1}, y^{i-1})$, for each (x^{i-1}, y^{i-1}) , $i = 0, 1, \dots, n$. This will be shown by induction.

Consider $n = 0$. For any $h_0(\cdot) \in BC(\mathcal{X}_0)$, by definition of weak convergence we have

$$\lim_{\alpha \rightarrow \infty} \int_{\mathcal{X}_0} h_0(x) p_0^\alpha(dx_0) = \int_{\mathcal{X}_0} h_0(x) p_0^o(dx_0).$$

Consider $n = 1$. For any $h_0(\cdot) \in BC(\mathcal{X}_0)$, $h_1(\cdot) \in BC(\mathcal{X}_1)$, we need to show $\forall \epsilon > 0$, there exists an $N \in \mathbb{N}_+ \triangleq \{1, 2, \dots\}$ such that for $\alpha > N$

$$\left| \int_{\mathcal{X}_0} h_0(x_0) p_0^\alpha(dx_0) \int_{\mathcal{X}_1} h_1(x_1) p_1^\alpha(dx_1|x_0, y_0) - \int_{\mathcal{X}_0} h_0(x_0) p_0^o(dx_0) \int_{\mathcal{X}_1} h_1(x_1) p_1^o(dx_1|x_0, y_0) \right| \leq \epsilon. \quad (\text{D.1})$$

From the left hand side (LHS) of (D.1), by adding and subtracting terms, we have the following upper bound.

$$\begin{aligned} A_{0,1} &\triangleq \left| \int_{\mathcal{X}_0 \times \mathcal{X}_1} h_0(x_0) h_1(x_1) p_1^\alpha(dx_1|x_0, y_0) p_0^\alpha(dx_0) - \int_{\mathcal{X}_0 \times \mathcal{X}_1} h_0(x_0) h_1(x_1) p_1^o(dx_1|x_0, y_0) p_0^o(dx_0) \right| \\ &\leq \underbrace{\left| \int_{\mathcal{X}_0 \times \mathcal{X}_1} h_0(x_0) h_1(x_1) p_1^o(dx_1|x_0, y_0) p_0^\alpha(dx_0) - \int_{\mathcal{X}_0 \times \mathcal{X}_1} h_0(x_0) h_1(x_1) p_1^o(dx_1|x_0, y_0) p_0^o(dx_0) \right|}_{\text{Term-1}} \\ &\quad + \underbrace{\left| \int_{\mathcal{X}_0 \times \mathcal{X}_1} h_0(x_0) h_1(x_1) p_1^\alpha(dx_1|x_0, y_0) p_0^\alpha(dx_0) - \int_{\mathcal{X}_0 \times \mathcal{X}_1} h_0(x_0) h_1(x_1) p_1^\alpha(dx_1|x_0, y_0) p_0^o(dx_0) \right|}_{\text{Term-2}}. \quad (\text{D.2}) \end{aligned}$$

Term-1: Let $\epsilon_0 > 0$ be given, and consider *Term-1*. By the continuity of the function mapping $(x_0, y_0) \in \mathcal{X}_0 \times \mathcal{Y}_0 \mapsto \int_{\mathcal{X}_1} h(x_1) p_1(dx_1|x_0, y_0)$ and the weak convergence $p_1^\alpha(\cdot|x_0, y_0) \xrightarrow{w} p_1^o(\cdot|x_0, y_0)$, for each $(x_0, y_0) \in \mathcal{X}_0 \times \mathcal{Y}_0$, then there exists an $N_1 \in \mathbb{N}_+$ such that for all $\alpha \geq N_1$

$$\left| \int_{\mathcal{X}_0} h_0(x_0) \left(\int_{\mathcal{X}_1} h_1(x_1) p_1^o(dx_1|x_0, y_0) \right) (p_0^\alpha(dx_0) - p_0^o(dx_0)) \right| \leq \epsilon_0. \quad (\text{D.3})$$

Term-2: Consider *Term-2*. By the weak convergence, $p_0^\alpha(dx_0) \xrightarrow{w} p_0^o(dx_0)$, $p_1^\alpha(dx_1|x_0, y_0) \xrightarrow{w} p_1^o(dx_1|x_0, y_0)$, for each $(x_0, y_0) \in \mathcal{X}_0 \times \mathcal{Y}_0$. According to Prohorov's theorem there exist compact subset $K_0 \subset \mathcal{X}_0$ such that $p_0^\alpha(K_0^c) \leq \epsilon_1$, $\alpha = 1, 2, \dots$, and compact subset $K_1 \subset \mathcal{X}_1$ such that $p_1^\alpha(K_1^c|x_0, y_0) \leq \epsilon_2$, $\alpha = 1, 2, \dots$, for each $(x_0, y_0) \in \mathcal{X}_0 \times \mathcal{Y}_0$.

Hence, *Term-2* is written as follows.

$$\begin{aligned} & \left| \int_{K_0 \cup K_0^c} h_0(x_0) \left(\int_{\mathcal{X}_1} h_1(x_1) p_1^\alpha(dx_1|x_0, y_0) \right) p_0^\alpha(dx_0) - \int_{K_0 \cup K_0^c} h_0(x_0) \left(\int_{\mathcal{X}_1} h_1(x_1) p_1^o(dx_1|x_0, y_0) \right) p_0^\alpha(dx_0) \right| \\ &= \left| \int_{K_0^c} h_0(x_0) \left(\int_{\mathcal{X}_1} h_1(x_1) p_1^\alpha(dx_1|x_0, y_0) \right) p_0^\alpha(dx_0) - \int_{K_0^c} h_0(x_0) \left(\int_{\mathcal{X}_1} h_1(x_1) p_1^o(dx_1|x_0, y_0) \right) p_0^\alpha(dx_0) \right| \\ &+ \left| \int_{K_0} h_0(x_0) \left(\int_{\mathcal{X}_1} h_1(x_1) p_1^\alpha(dx_1|x_0, y_0) \right) p_0^\alpha(dx_0) - \int_{K_0} h_0(x_0) \left(\int_{\mathcal{X}_1} h_1(x_1) p_1^o(dx_1|x_0, y_0) \right) p_0^\alpha(dx_0) \right| \\ &\leq \int_{K_0^c} \|h_0(\cdot)\|_\infty \|h_1(\cdot)\|_\infty p_0^\alpha(dx_0) + \int_{K_0^c} \|h_0(\cdot)\|_\infty \|h_1(\cdot)\|_\infty p_0^\alpha(dx_0) \\ &\quad + \left| \int_{K_0} h_0(x_0) \left(\int_{\mathcal{X}_1} h_1(x_1) p_1^\alpha(dx_1|x_0, y_0) - \int_{\mathcal{X}_1} h_1(x_1) p_1^o(dx_1|x_0, y_0) \right) p_0^\alpha(dx_0) \right| \\ &\leq 2. \|h_0(\cdot)\|_\infty \|h_1(\cdot)\|_\infty p_0^\alpha(K_0^c) \\ &\quad + \left| \int_{K_0} h_0(x_0) \left(\int_{\mathcal{X}_1} h_1(x_1) p_1^\alpha(dx_1|x_0, y_0) - \int_{\mathcal{X}_1} h_1(x_1) p_1^o(dx_1|x_0, y_0) \right) p_0^\alpha(dx_0) \right| \\ &\leq 2. \|h_0(\cdot)\|_\infty \|h_1(\cdot)\|_\infty \epsilon_1 + \|h_0(\cdot)\|_\infty \sup_{x_0 \in K_0} \left| \int_{\mathcal{X}_1} h_1(x_1) p_1^\alpha(dx_1|x_0, y_0) - \int_{\mathcal{X}_1} h_1(x_1) p_1^o(dx_1|x_0, y_0) \right|. \end{aligned} \quad (\text{D.5})$$

By utilizing condition (III.26), $\forall \epsilon_2 > 0$ there exists $N_2 \in \mathbb{N}_1$ such that for all $\alpha \geq N_2$

$$\sup_{x_0 \in K_0} \left| \int_{\mathcal{X}_1} h_1(x_1) p_1^\alpha(dx_1|x_0, y_0) - \int_{\mathcal{X}_1} h_1(x_1) p_1^o(dx_1|x_0, y_0) \right| < \epsilon_2, \quad \forall y_0 \in \mathcal{Y}_0 \quad (\text{D.6})$$

Hence, by (D.3), (D.5), (D.6), there exists an $N \in \mathbb{N}_1$ large enough such that for all $\alpha \geq N_2$, expression (D.2) is further bounded by

$$A_{0,1} \leq \epsilon_0 + 2. \|h_0(\cdot)\|_\infty \|h_1(\cdot)\|_\infty \epsilon_1 + \|h_0(\cdot)\|_\infty \epsilon.$$

Since $\epsilon_0, \epsilon_1, \epsilon_2 > 0$ are arbitrary, the claim holds for $n = 1$, as well.

Suppose that for $n = k$, and for each $h_i(\cdot) \in BC(\mathcal{X}_i)$, $i = 0, 1, \dots, k$, and $\forall \epsilon > 0$, there exists $N^k \in \mathbb{N}_1$ such that for each $\alpha \geq N^k$

$$\left| \int_{\mathcal{X}_{0,k}} \bigotimes_{i=0}^k h_i(x_i) p_i^\alpha(dx_i|x^{i-1}, y^{i-1}) - \int_{\mathcal{X}_{0,k}} \bigotimes_{i=0}^k h_i(x_i) p_i^o(dx_i|x^{i-1}, y^{i-1}) \right| \leq \epsilon. \quad (\text{D.7})$$

We need to show that (D.7) holds for $n = k + 1$, i.e.,

$$\otimes_{i=0}^{k+1} p_i^\alpha(\cdot | x^{i-1}, y^{i-1}) \xrightarrow{w} \otimes_{i=0}^{k+1} p_i^o(\cdot | x^{i-1}, y^{i-1})$$

whenever $p_i^\alpha(\cdot | x^{i-1}, y^{i-1}) \xrightarrow{w} p_i^o(\cdot | x^{i-1}, y^{i-1})$, $i = 0, 1, \dots, k+1$, and provided that $\otimes_{i=0}^k p_i^\alpha(\cdot | x^{i-1}, y^{i-1}) \xrightarrow{w} \otimes_{i=0}^k p_i^o(\cdot | x^{i-1}, y^{i-1})$. The derivation is similar to showing (D.1), hence it is omitted.

This shows (III.27), hence the set $\bar{P}_{0,n}(\cdot | y^{n-1}) \in \mathcal{M}_1^{\mathbf{C1}}(\mathcal{X}_{0,n}), y^{n-1} \in \mathcal{Y}_{0,n-1}$ is closed. By Theorem III.5, Part A. A2), this set is also tight, hence by Prohorov's theorem (Appendix A, Theorem A.3) it is compact. This completes the derivation. \square

APPENDIX E PROOF OF LEMMA III.8

(1) Since every probability measure on a compact metric space is weakly compact, then the set $\bar{Q}_{0,n}(\cdot | x^n) \in \mathcal{M}_1^{\mathbf{C2}}(\mathcal{Y}_{0,n}), x^n \in \mathcal{X}_{0,n}$ is weakly compact. This means that any sequence $\{\bar{Q}_{0,n}^\alpha(\cdot | x^n) : \alpha = 1, 2, \dots\}$, possesses a weakly convergent subsequence $\bar{Q}_{0,n}^{\alpha_i}(dy^n | x^n) \xrightarrow{w} \bar{Q}_{0,n}^o(dy^n | x^n)$, for each $x^n \in \mathcal{X}_{0,n}$, and hence tight (by Prohorov's theorem, see Appendix A, Theorem A.2), but $\bar{Q}_{0,n}^o(dy^n | x^n)$ may not be an element of $\mathcal{M}_1^{\mathbf{C2}}(\mathcal{Y}_{0,n})$ (i.e., it may fail to satisfy consistency condition C2). By Prohorov's theorem, to show compactness of $\bar{Q}_{0,n}(\cdot | x^n) \in \mathcal{M}_1^{\mathbf{C2}}(\mathcal{Y}_{0,n}), x^n \in \mathcal{X}_{0,n}$, we need to show $\bar{Q}_{0,n}^o(\cdot | x^n) = \bar{Q}_{0,n}^o(\cdot | x^n) \triangleq \otimes_{i=0}^n q_i^o(dy_i | y^{i-1}, x^i)$, whenever $q_i^\alpha(dy_i | y^{i-1}, x^i) \xrightarrow{w} q_i^o(dy_i | y^{i-1}, x^i)$, $i = 0, 1, \dots, n$ (since \mathcal{Y}_i , $i = 0, 1, \dots, n$ are compact Polish spaces). The method is precisely the same as in Lemma III.6, hence it is omitted. Therefore, the set $\bar{Q}_{0,n}(\cdot | x^n) \in \mathcal{M}_1^{\mathbf{C2}}(\mathcal{Y}_{0,n}), x^n \in \mathcal{X}_{0,n}$ is closed, and since it is also tight, it is compact.

(2) Next, we discuss how the fidelity set $\mathcal{Q}_{0,n}(D)$ is a closed subset of the compact set $\mathcal{M}_1^{\mathbf{C2}}(\mathcal{Y}_{0,n})$, hence compact itself, that is, for each sequence $\{\bar{Q}_{0,n}^\alpha(\cdot | x^n) : \alpha = 1, 2, \dots\} \in \mathcal{Q}_{0,n}(D)$ there is a subsequence such that $\bar{Q}_{0,n}^\alpha(\cdot | x^n) \xrightarrow{w} \bar{Q}_{0,n}^o(\cdot | x^n) \in \mathcal{Q}_{0,n}(D)$. We outline the derivation. Let $\{\bar{Q}_{0,n}^\alpha(\cdot | x^n) : \alpha = 1, 2, \dots\} \in \mathcal{Q}_{0,n}(D) \subset \mathcal{M}_1^{\mathbf{C2}}(\mathcal{Y}_{0,n})$. Since $\mathcal{M}_1^{\mathbf{C2}}(\mathcal{Y}_{0,n})$ is closed and uniformly tight, and hence compact, there exists a subsequence $\{\bar{Q}_{0,n}^{\alpha_i}(\cdot | x^n) : i = 1, 2, \dots\} \in \mathcal{M}_1^{\mathbf{C2}}(\mathcal{Y}_{0,n})$ and a measure $\bar{Q}_{0,n}^o(\cdot | x^n) \in \mathcal{M}_1^{\mathbf{C2}}(\mathcal{Y}_{0,n})$ such that $\bar{Q}_{0,n}^{\alpha_i}(\cdot | x^n) \xrightarrow{w} \bar{Q}_{0,n}^o(\cdot | x^n)$ for each $x^n \in \mathcal{X}_{0,n}$. Recall that $d_{0,n} : \mathcal{X}_{0,n} \times \mathcal{Y}_{0,n} \mapsto [0, \infty]$ is a Borel measurable, non-negative, and continuous function on $y^n \in \mathcal{Y}_{0,n}$. Consider the sequence $\{d_{0,n}^{(k)} \triangleq d_{0,n} \wedge k : k \in \mathbb{N}_0\}$, $\mathbb{N}_1 \triangleq \{1, 2, \dots\}$, which is bounded, and continuous function in the second argument $y^n \in \mathcal{Y}_{0,n}$. By Lebesgue's monotone convergence theorem and Fatou's lemma it can be shown that $\mathcal{Q}_{0,n}(D)$ is closed with respect to the topology of weak convergence. Since a closed subset of a compact set is compact, then $\mathcal{Q}_{0,n}(D)$ is compact. This completes the derivation. \square

APPENDIX F PROOF OF THEOREM III.10

1) We need to show that for any sequence $\{\bar{Q}_{0,n}^\alpha(\cdot | x^n) \in \mathcal{M}_1^{\mathbf{C2}}(\mathcal{Y}_{0,n}) : \alpha = 1, 2, \dots\}$, such that $\bar{Q}_{0,n}^\alpha(\cdot | x^n) \xrightarrow{w} \bar{Q}_{0,n}^o(\cdot | x^n)$ for each $x^n \in \mathcal{X}_{0,n}$ then

$$\mathbb{I}_{X^n \rightarrow Y^n}(\bar{P}_{0,n}, \bar{Q}_{0,n}^o) \leq \liminf_{\alpha \rightarrow \infty} \mathbb{I}_{X^n \rightarrow Y^n}(\bar{P}_{0,n}, \bar{Q}_{0,n}^\alpha).$$

Define the sequence of joint distribution $P_{0,n}^\alpha(dx^n, dy^n) \triangleq (\bar{P}_{0,n} \otimes \bar{Q}_{0,n}^\alpha)(dx^n, dy^n)$, $\alpha = 1, 2, \dots$. Weak convergence $P_{0,n}^\alpha(dx^n, dy^n) \xrightarrow{w} (\bar{P}_{0,n} \otimes \bar{Q}_{0,n}^o)(dx^n, dy^n) \equiv P_{0,n}^o(dx^n, dy^n)$ is shown by considering integrals with respect to a test function $\phi_{0,n}(\cdot, \cdot) \in BC(\mathcal{X}_{0,n} \times \mathcal{Y}_{0,n})$ via

$$\int_{\mathcal{X}_{0,n} \times \mathcal{Y}_{0,n}} \phi_{0,n}(x^n, y^n) P_{0,n}^\alpha(dx^n, dy^n) = \int_{\mathcal{X}_{0,n} \times \mathcal{Y}_{0,n}} \phi_{0,n}(x^n, y^n) (\bar{P}_{0,n} \otimes \bar{Q}_{0,n}^\alpha)(dx^n, dy^n).$$

By Theorem III.5, Part A., A1), $P_{0,n}^\alpha(dx^n, dy^n) \xrightarrow{w} P_{0,n}^o(dx^n, dy^n)$. Similarly, consider $\vec{\Pi}_{0,n}^\alpha \triangleq \overleftarrow{P}_{0,n} \otimes \nu_{0,n}^\alpha$, $\alpha = 1, 2, \dots$, where $\{\nu_{0,n}^\alpha : \alpha = 1, 2, \dots\}$ are the marginals of $\{P_{0,n}^\alpha : \alpha = 1, 2, \dots\}$. Then by Theorem III.5, Part A., A4) we have

$$\vec{\Pi}_{0,n}^\alpha = \overleftarrow{P}_{0,n} \otimes \nu_{0,n}^\alpha \xrightarrow{w} \vec{\Pi}_{0,n}^o = \overleftarrow{P}_{0,n} \otimes \nu_{0,n}^o.$$

Recall the definition of directed information via relative entropy given by

$$\mathbb{D}(P_{0,n} || \vec{\Pi}_{0,n}) = \mathbb{D}(\overleftarrow{P}_{0,n} \otimes \vec{Q}_{0,n} || \overleftarrow{P}_{0,n} \otimes \nu_{0,n}) = \mathbb{I}_{X^n \rightarrow Y^n}(\overleftarrow{P}_{0,n}, \vec{Q}_{0,n}). \quad (\text{F.1})$$

It is well known that relative entropy is lower semicontinuous, hence

$$\mathbb{D}(P_{0,n}^o || \vec{\Pi}_{0,n}^o) = \mathbb{D}(\overleftarrow{P}_{0,n} \otimes \vec{Q}_{0,n}^o || \vec{\Pi}_{0,n}^o) \leq \liminf_{\alpha \rightarrow \infty} \mathbb{D}(P_{0,n}^\alpha || \vec{\Pi}_{0,n}^\alpha). \quad (\text{F.2})$$

By (F.1) it follows that (F.2) is also equivalent to

$$\mathbb{I}_{X^n \rightarrow Y^n}(\overleftarrow{P}_{0,n}, \vec{Q}_{0,n}^o) \leq \liminf_{\alpha \rightarrow \infty} \mathbb{I}_{X^n \rightarrow Y^n}(\overleftarrow{P}_{0,n}, \vec{Q}_{0,n}^\alpha)$$

Hence, directed information is lower semicontinuous as a functional of $\vec{Q}_{0,n}(\cdot|x^n) \in \mathcal{M}_1^{\text{C2}}(\mathcal{Y}_{0,n})$ for a fixed $\overleftarrow{P}_{0,n}(\cdot|y^{n-1}) \in \mathcal{M}_1^{\text{C1}}(\mathcal{X}_{0,n})$. This completes the derivation of 1).

2) The derivation is similar to 1). \square

APPENDIX G

PROOF OF THEOREM III.13

To show continuity of $\mathbb{I}_{X^n \rightarrow Y^n}(\cdot, \vec{Q}_{0,n})$ we need to show that for every sequence $\{\overleftarrow{P}_{0,n}^\alpha(\cdot|y^{n-1}) : \alpha = 1, 2, \dots\}$ such that $\overleftarrow{P}_{0,n}^\alpha \xrightarrow{w} \overleftarrow{P}_{0,n}^o$, we have

$$\mathbb{I}_{X^n \rightarrow Y^n}(\overleftarrow{P}_{0,n}^\alpha, \vec{Q}_{0,n}) \longrightarrow \mathbb{I}_{X^n \rightarrow Y^n}(\overleftarrow{P}_{0,n}^o, \vec{Q}_{0,n}).$$

The derivation is based on the procedure utilized in [29] to show continuity for single letter mutual information. First, decompose directed information into two terms as follows.

$$\begin{aligned} \mathbb{I}_{X^n \rightarrow Y^n}(\overleftarrow{P}_{0,n}, \vec{Q}_{0,n}) &= \int_{\mathcal{X}_{0,n} \times \mathcal{Y}_{0,n}} \log \left(\frac{d\vec{Q}_{0,n}(\cdot|x^n)}{d\nu_{0,n}(\cdot)}(y^n) \right) (\overleftarrow{P}_{0,n} \otimes \vec{Q}_{0,n})(dx^n, dy^n) \\ &= \int_{\mathcal{X}_{0,n} \times \mathcal{Y}_{0,n}} \log \left(\frac{d\vec{Q}_{0,n}(\cdot|x^n)}{d\nu_{0,n}(\cdot)}(y^n) \right) (\overleftarrow{P}_{0,n} \otimes \vec{Q}_{0,n})(dx^n, dy^n) \\ &\quad - \int_{\mathcal{Y}_{0,n}} \log \left(\frac{d\vec{Q}_{0,n}(\cdot|x^n)}{d\nu_{0,n}(\cdot)}(y^n) \right) \nu_{0,n}(dy^n) \\ &= \int_{\mathcal{X}_{0,n} \times \mathcal{Y}_{0,n}} \left(\xi_{\bar{\nu}_{0,n}}(x^n, y^n) \log \xi_{\bar{\nu}_{0,n}}(x^n, y^n) \right) \overleftarrow{P}_{0,n}(dx^n|y^{n-1}) \otimes \bar{\nu}_{0,n}(dy^n) \\ &\quad - \int_{\mathcal{Y}_{0,n}} \left(\xi_{\bar{\nu}_{0,n}, \overleftarrow{P}_{0,n}}(y^n) \log \xi_{\bar{\nu}_{0,n}, \overleftarrow{P}_{0,n}}(y^n) \right) \bar{\nu}_{0,n}(dy^n), \end{aligned} \quad (\text{G.1})$$

where $\xi_{\bar{\nu}_{0,n}, \overleftarrow{P}_{0,n}}(y^n) \triangleq \frac{d\nu_{0,n}(\cdot)}{d\bar{\nu}_{0,n}(\cdot)}(y^n)$ emphasizes the fact that this RND depends on $\overleftarrow{P}_{0,n}(\cdot|y^{n-1})$ via $\bar{\nu}(\cdot)$. For now, assume that both terms in on the RHS of the above formula are finite; the validity of this assumption will be established at the end. Thus, we only need to show that both terms are bounded and continuous in the weak sense over $\mathcal{M}_1^{\text{C1,cl}}(\mathcal{X}_{0,n})$.

Continuity of the first term. Since $\overleftarrow{P}_{0,n}^\alpha(\cdot|y^{n-1}) \xrightarrow{w} \overleftarrow{P}_{0,n}^o(\cdot|y^{n-1})$, by [30, Theorem A.5.8, p. 320], utilizing Lebesgue's dominated convergence theorem, we have $\overleftarrow{P}_{0,n}^\alpha \otimes \bar{\nu}_{0,n} \xrightarrow{w} \overleftarrow{P}_{0,n}^o \otimes \bar{\nu}_{0,n}$. Since $\xi_{\bar{\nu}_{0,n}}(x^n, y^n)$ is continuous, then so is $\xi_{\bar{\nu}_{0,n}}(x^n, y^n) \log \xi_{\bar{\nu}_{0,n}}(x^n, y^n)$. By hypothesis, $\xi_{\bar{\nu}_{0,n}}(x^n, y^n) \log \xi_{\bar{\nu}_{0,n}}(x^n, y^n)$ is

uniformly integrable over $\{\bar{\nu}_{0,n} \otimes \bar{P}_{0,n} : \bar{P}_{0,n}(\cdot|y^{n-1}) \in \mathcal{M}_1^{\text{C1},cl}(\mathcal{X}_{0,n})\}$. Therefore, using Theorem A.8, Appendix A, we conclude that

$$\begin{aligned} \lim_{\alpha \rightarrow \infty} \int_{\mathcal{X}_{0,n} \times \mathcal{Y}_{0,n}} \xi_{\bar{\nu}_{0,n}}(x^n, y^n) \log \xi_{\bar{\nu}_{0,n}}(x^n, y^n) \bar{P}_{0,n}^\alpha(dx^n|y^{n-1}) \otimes \bar{\nu}_{0,n}(dy^n) \\ = \int_{\mathcal{X}_{0,n} \times \mathcal{Y}_{0,n}} \xi_{\bar{\nu}_{0,n}}(x^n, y^n) \log \xi_{\bar{\nu}_{0,n}}(x^n, y^n) P_{0,n}^o(dx^n|y^{n-1}) \otimes \bar{\nu}_{0,n}(dy^n). \end{aligned} \quad (\text{G.2})$$

This proves the continuity of the first term. The finiteness of the first term is obtained from uniform integrability as follows. For a given $\epsilon > 0$ and sufficiently large $c > 0$

$$\begin{aligned} & \sup_{\bar{P}_{0,n}(\cdot|y^{n-1}) \in \mathcal{M}_1^{\text{C1},cl}(\mathcal{X}_{0,n})} \left\{ \int_{\mathcal{X}_{0,n} \times \mathcal{Y}_{0,n}} |\xi_{\bar{\nu}_{0,n}}(x^n, y^n) \log \xi_{\bar{\nu}_{0,n}}(x^n, y^n)| I_{\{|\xi_{\bar{\nu}_{0,n}}(x^n, y^n) \log \xi_{\bar{\nu}_{0,n}}(x^n, y^n)| \geq c\}} \right. \\ & \quad \times \bar{P}_{0,n}^\alpha(dx^n|y^{n-1}) \otimes \bar{\nu}_{0,n}(dy^n) \\ & + \int_{\mathcal{X}_{0,n} \times \mathcal{Y}_{0,n}} |\xi_{\bar{\nu}_{0,n}}(x^n, y^n) \log \xi_{\bar{\nu}_{0,n}}(x^n, y^n)| I_{\{|\xi_{\bar{\nu}_{0,n}}(x^n, y^n) \log \xi_{\bar{\nu}_{0,n}}(x^n, y^n)| < c\}} \bar{P}_{0,n}^\alpha(dx^n|y^{n-1}) \otimes \bar{\nu}_{0,n}(dy^n) \Big\} \\ & \leq \sup_{\bar{P}_{0,n}(\cdot|y^{n-1}) \in \mathcal{M}_1^{\text{C1},cl}(\mathcal{X}_{0,n})} \left\{ \int_{\{|\xi_{\bar{\nu}_{0,n}}(x^n, y^n) \log \xi_{\bar{\nu}_{0,n}}(x^n, y^n)| \geq c\}} |\xi_{\bar{\nu}_{0,n}}(x^n, y^n) \log \xi_{\bar{\nu}_{0,n}}(x^n, y^n)| \right. \\ & \quad \times \bar{P}_{0,n}^\alpha(dx^n|y^{n-1}) \otimes \bar{\nu}_{0,n}(dy^n) \Big\} \\ & + \sup_{\bar{P}_{0,n}(\cdot|y^{n-1}) \in \mathcal{M}_1^{\text{C1},cl}(\mathcal{X}_{0,n})} \left\{ \int_{\{|\xi_{\bar{\nu}_{0,n}}(x^n, y^n) \log \xi_{\bar{\nu}_{0,n}}(x^n, y^n)| < c\}} |\xi_{\bar{\nu}_{0,n}}(x^n, y^n) \log \xi_{\bar{\nu}_{0,n}}(x^n, y^n)| \right. \\ & \quad \times \bar{P}_{0,n}^\alpha(dx^n|y^{n-1}) \otimes \bar{\nu}_{0,n}(dy^n) \Big\} \leq \epsilon + c. \end{aligned}$$

Continuity of the second term. For a fixed $y^n \in \mathcal{Y}_{0,n}$, since $\xi_{\bar{\nu}_{0,n}}(x^n, y^n)$ is uniformly integrable over $\mathcal{M}_1^{\text{C1},cl}(\mathcal{X}_{0,n})$, by Theorem A.6, Appendix A, we obtain that $\bar{P}_{0,n}^\alpha \xrightarrow{w} \bar{P}_{0,n}^o$, implies pointwise convergence of $\xi_{\bar{\nu}_{0,n}, \bar{P}_{0,n}^\alpha}(y^n) \rightarrow \xi_{\bar{\nu}_{0,n}, \bar{P}_{0,n}^o}(y^n)$. By continuity of the logarithm, we obtain the pointwise convergence of $\xi_{\bar{\nu}_{0,n}, \bar{P}_{0,n}^\alpha}(y^n) \log \xi_{\bar{\nu}_{0,n}, \bar{P}_{0,n}^\alpha}(y^n) \rightarrow \xi_{\bar{\nu}_{0,n}, \bar{P}_{0,n}^o}(y^n) \log \xi_{\bar{\nu}_{0,n}, \bar{P}_{0,n}^o}(y^n)$. It only remains to show convergence under the integral with respect to $\bar{\nu}_{0,n}$. By (III.34), then $\forall \alpha$

$$\begin{aligned} \left| \xi_{\bar{\nu}_{0,n}, \bar{P}_{0,n}^\alpha}(y^n) \log \xi_{\bar{\nu}_{0,n}, \bar{P}_{0,n}^\alpha}(y^n) \right| & \leq \frac{2}{e \ln 2} + \xi_{\bar{\nu}_{0,n}, \bar{P}_{0,n}^\alpha}(y^n) \log \xi_{\bar{\nu}_{0,n}, \bar{P}_{0,n}^\alpha}(y^n) \\ & = \frac{2}{e \ln 2} \int_{\mathcal{X}_{0,n}} \xi_{\bar{\nu}_{0,n}, \bar{P}_{0,n}^\alpha}(y^n) \log \xi_{\bar{\nu}_{0,n}, \bar{P}_{0,n}^\alpha}(y^n) \bar{P}_{0,n}^\alpha(dx^n|y^{n-1}) \\ & \leq \frac{2}{e \ln 2} + \int_{\mathcal{X}_{0,n}} \left(\xi_{\bar{\nu}_{0,n}}(x^n, y^n) \log \xi_{\bar{\nu}_{0,n}}(x^n, y^n) \right) \bar{P}_{0,n}^\alpha(dx^n|y^{n-1}). \end{aligned} \quad (\text{G.3})$$

where (G.3) follows from (G.1) and the nonnegativity of $\mathbb{I}_{X^n \rightarrow Y^n}(\bar{P}_{0,n}, \bar{Q}_{0,n})$. By (G.2), the integration of the RHS over $\bar{\nu}_{0,n}$ converges. Thus, by the generalized Lebesgue's dominated convergence theorem [58, p. 59], we conclude that

$$\int_{\mathcal{Y}_{0,n}} \xi_{\bar{\nu}_{0,n}, \bar{P}_{0,n}^\alpha}(y^n) \log \xi_{\bar{\nu}_{0,n}, \bar{P}_{0,n}^\alpha}(y^n) \bar{\nu}_{0,n}(dy^n) \xrightarrow{\alpha \rightarrow \infty} \int_{\mathcal{Y}_{0,n}} \xi_{\bar{\nu}_{0,n}, \bar{P}_{0,n}^o}(y^n) \log \xi_{\bar{\nu}_{0,n}, \bar{P}_{0,n}^o}(y^n) \bar{\nu}_{0,n}(dy^n).$$

This implies the continuity of the second term. Furthermore, its finiteness follows as before. Since both terms are finite and continuous we deduce continuity of the directed information $\mathbb{I}_{X^n \rightarrow Y^n}(\cdot, \bar{Q}_{0,n})$ with respect to $\bar{P}_{0,n}(\cdot|y^{n-1})$, for fixed $\bar{Q}_{0,n}(\cdot|x^n)$. This completes the derivation.

REFERENCES

- [1] C. D. Charalambous and P. A. Stavrou, "Directed information on abstract spaces: properties and extremum problems," in *IEEE International Symposium on Information Theory (ISIT)*, July 1-6 2012, pp. 518–522.
- [2] P. A. Stavrou and C. D. Charalambous, "Variational equalities of directed information and applications," in *IEEE International Symposium on Information Theory Proceedings (ISIT)*, July 2013, pp. 2577–2581.
- [3] C. D. Charalambous, P. A. Stavrou, and C. K. Kourtellaris, "Directed information on abstract spaces: Properties and extremum problems," in *Coordination Control of Distributed Systems*, ser. Lecture Notes in Control and Information Sciences, J. H. van Schuppen and T. Villa, Eds. Springer International Publishing, 2015, vol. 456, pp. 307–315.
- [4] H. Marko, "The bidirectional communication theory—A generalization of information theory," *IEEE Transactions on Communications*, vol. 21, no. 12, pp. 1345–1351, Dec. 1973.
- [5] J. L. Massey, "Causality, feedback and directed information," in *International Symposium on Information Theory and its Applications (ISITA '90)*, Nov. 27-30 1990, pp. 303–305.
- [6] G. Kramer, "Directed information for channels with feedback," Ph.D. dissertation, Swiss Federal Institute of Technology (ETH), December 1998.
- [7] S. C. Tatikonda, "Control over communication constraints," Ph.D. dissertation, Mass. Inst. of Tech. (M.I.T.), Cambridge, MA, 2000.
- [8] J. Chen and T. Berger, "The capacity of finite-state Markov channels with feedback," *IEEE Transactions on Information Theory*, vol. 51, no. 3, pp. 780–798, Mar. 2005.
- [9] S. Yang, A. Kavcic, and S. Tatikonda, "Feedback capacity of finite-state machine channels," *IEEE Transactions on Information Theory*, vol. 51, no. 3, pp. 799–810, Mar. 2005.
- [10] H. Permuter, P. Cuff, B. Van Roy, and T. Weissman, "Capacity of the trapdoor channel with feedback," *IEEE Transactions on Information Theory*, vol. 54, no. 7, pp. 3150–3165, July 2008.
- [11] S. Tatikonda and S. Mitter, "The capacity of channels with feedback," *IEEE Transactions on Information Theory*, vol. 55, no. 1, pp. 323–349, Jan. 2009.
- [12] H. H. Permuter, T. Weissman, and J. Chen, "Capacity region of the finite-state multiple-access channel with and without feedback," *IEEE Transactions on Information Theory*, vol. 55, no. 6, pp. 2455–2477, June 2009.
- [13] H. H. Permuter, T. Weissman, and A. J. Goldsmith, "Finite state channels with time-invariant deterministic feedback," *IEEE Transactions on Information Theory*, vol. 55, no. 2, pp. 644–662, Feb. 2009.
- [14] B. Shrader and H. Permuter, "Feedback capacity of the compound channel," *IEEE Transactions on Information Theory*, vol. 55, no. 8, pp. 3629–3644, Aug. 2009.
- [15] N. Ma and P. Ishwar, "On delayed sequential coding of correlated sources," *IEEE Transactions on Information Theory*, vol. 57, no. 6, pp. 3763–3782, 2011.
- [16] R. Venkataramanan and S. S. Pradhan, "Source coding with feed-forward: Rate-distortion theorems and error exponents for a general source," *IEEE Transactions on Information Theory*, vol. 53, no. 6, pp. 2154–2179, 2007.
- [17] G. Kramer, "Capacity results for the discrete memoryless network," *IEEE Transactions on Information Theory*, vol. 49, no. 1, pp. 4–21, Jan. 2003.
- [18] T. S. Han and S. Verdú, "Approximation theory of output statistics," *IEEE Transactions on Information Theory*, vol. 39, no. 3, pp. 752–772, May 1993.
- [19] H. H. Permuter, Y.-H. Kim, and T. Weissman, "Interpretations of directed information in portfolio theory, data compression, and hypothesis testing," *IEEE Transactions on Information Theory*, vol. 57, no. 6, pp. 3248–3259, 2011.
- [20] V. Solo, "On causality and mutual information," in *47th IEEE Conference on Decision and Control (CDC '08)*, Dec. 2008, pp. 4939–4944.
- [21] P. O. Amblard and O. J. J. Michel, "On directed information theory and granger causality graphs," *Journal of Computational Neuroscience*, vol. 30, no. 1, pp. 7–16, Feb. 2011.
- [22] 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, Feb. 2011.
- [23] C. D. Charalambous, P. A. Stavrou, and N. U. Ahmed, "Nonanticipative rate distortion function and relations to filtering theory," *IEEE Transactions on Automatic Control*, vol. 59, no. 4, pp. 937–952, April 2014.
- [24] C. D. Charalambous and P. A. Stavrou, "Optimization of directed information and relations to filtering theory," in *European Control Conference (ECC)*, June 2014, pp. 1385–1390.
- [25] R. E. Blahut, *Principles and Practice of Information Theory*, ser. in Electrical and Computer Engineering. Reading, MA: Addison-Wesley Publishing Company, 1987.
- [26] S.-W. Ho and R. Yeung, "On the discontinuity of the shannon information measures," *IEEE Transactions on Information Theory*, vol. 55, no. 12, pp. 5362–5374, Dec 2009.
- [27] S. Ihara, *Information theory - for Continuous Systems*. World Scientific, 1993.
- [28] I. Csiszár, "Arbitrarily varying channels with general alphabets and states," *IEEE Transactions on Information Theory*, vol. 38, no. 6, pp. 1725–1742, Nov. 1992.
- [29] M. Fozunbal, S. McLaughlin, and R. Schafer, "Capacity analysis for continuous-alphabet channels with side information, part I: A general framework," *IEEE Transactions on Information Theory*, vol. 51, no. 9, pp. 3075–3085, Sep. 2005.
- [30] P. Dupuis and R. S. Ellis, *A Weak Convergence Approach to the Theory of Large Deviations*. John Wiley & Sons, Inc., New York, 1997.
- [31] P. Billingsley, *Convergence of Probability Measures*, 2nd ed., ser. Wiley Series in Probability and Statistics: Probability and Statistics. John Wiley & Sons Inc., 1999.
- [32] I. I. Gihman and A. V. Skorohod, *Controlled Stochastic Processes*. Springer-Verlag, New York Inc., 1979, translated by Samuel Kotz.
- [33] D. P. Bertsekas and S. E. Shreve, *Stochastic Optimal Control: The Discrete-Time Case*, ser. Optimization and Neural Computation Series. Athena Scientific, 2007.

- [34] S. Tatikonda and S. Mitter, "The capacity of channels with feedback," *IEEE Transactions on Information Theory*, vol. 55, no. 1, pp. 323–349, Jan. 2009.
- [35] P. A. Stavrou, C. D. Charalambous, and C. K. Kourtellis, "Optimal nonstationary reproduction distribution for nonanticipative rdf on abstract alphabets," *CoRR*, vol. abs/1301.6522, 2013. [Online]. Available: <http://arxiv.org/abs/1301.6522>
- [36] G. Kramer, "Topics in multi-user information theory," *Foundations and Trends in Communications and Information Theory*, vol. 4, no. 4-5, pp. 265–444, Apr. 2007.
- [37] E. A. Gamal and H. Y. Kim, *Network Information Theory*. Cambridge University Press, December 2011.
- [38] T. Berger, "Rate distortion theory for sources with abstract alphabets and memory," *Information and Control*, vol. 13, no. 3, pp. 254–273, Sep. 1968.
- [39] R. Blahut, "Computation of channel capacity and rate-distortion functions," *IEEE Transactions on Information Theory*, vol. 18, no. 4, pp. 460–473, July 1972.
- [40] P. A. Stavrou, C. D. Charalambous, and I. Tzortzis, "Algorithms and dynamic programming for feedback capacity computation of channels with memory," in *preparation for submission to IEEE Transactions on Information Theory*, 2015.
- [41] T. Berger, *Rate Distortion Theory: A Mathematical Basis for Data Compression*. Englewood Cliffs, NJ: Prentice-Hall, 1971.
- [42] P. A. Stavrou, C. K. Kourtellis, and C. D. Charalambous, "Information nonanticipative rate distortion function and its applications," *submitted to IEEE Transactions on Information Theory*, 2015. [Online]. Available: <http://arxiv.org/abs/1405.1593>
- [43] A. K. Gorbunov and M. S. Pinsker, "Nonanticipatory and prognostic epsilon entropies and message generation rates," *Problems of Information Transmission*, vol. 9, no. 3, pp. 184–191, July-Sept. 1973.
- [44] M. Pinsker, *Information and Information Stability of Random Variables and Processes*. Holden-Day Inc, San Francisco, 1964, translated by Amiel Feinstein.
- [45] A. N. Shiryaev, *Probability*, 2nd ed., ser. Graduate Texts in Mathematics. Springer-Verlag, Berlin, Heidelberg, New York, 1996.
- [46] H. Permuter, H. Asnani, and T. Weissman, "Capacity of a post channel with and without feedback," *IEEE Transactions on Information Theory*, vol. 60, no. 10, pp. 6041–6057, Oct 2014.
- [47] I. Csiszár, "On an extremum problem of information theory," *Studia Scientiarum Mathematicarum Hungarica*, vol. 9, pp. 57–71, 1974.
- [48] A. Lapidoth and P. Narayan, "Reliable communication under channel uncertainty," *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2148–2177, Oct. 1998.
- [49] D. J. Sakrison, "The rate distortion function for a class of sources," *Information and Control*, vol. 15, no. 2, pp. 165–195, Aug. 1969.
- [50] I. Naiss and H. H. Permuter, "Extension of the blahut-arimoto algorithm for maximizing directed information," *IEEE Transactions on Information Theory*, vol. 59, no. 1, pp. 204–222, Jan. 2013.
- [51] C. K. Kourtellis and C. D. Charalambous, "Capacity of binary state symmetric channel with and without feedback and transmission cost," in *IEEE Information Theory Workshop (ITW)*, April 2015, pp. 1–5.
- [52] C. K. Kourtellis, C. D. Charalambous, and J. J. Boutros, "Nonanticipative transmission for sources and channels with memory," in *IEEE International Symposium on Information Theory (ISIT)*, June 2015, pp. 521–525.
- [53] C. K. Kourtellis and C. D. Charalambous, "Information structures of capacity achieving distributions for feedback channels with memory and transmission cost: stochastic optimal control & variational equalities-part I," *IEEE Transactions on Information Theory* (submitted), 2015. [Online]. Available: <http://arxiv.org/pdf/1512.04514>
- [54] O. Hernández-Lerma and J.-B. Lasserre, *Discrete-time Markov control processes : basic optimality criteria*, ser. Applications of mathematics. New York: Springer, 1996, vol. 30.
- [55] J. D. Deuschel and D. W. Stroock, *Large Deviations*. San Diego: Academic Press, Inc., 1989.
- [56] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. John Wiley & Sons, Inc., Hoboken, New Jersey, 2006.
- [57] K. R. Parthasarathy, *Probability Measures on Metric Spaces*. Academic Press, New York, 1967.
- [58] G. B. Folland, *Real Analysis: Modern Techniques and Their Applications*, 2nd ed. John Wiley & Sons Inc., April 1999.