Neural Estimation and Optimization of Directed Information Over Continuous Spaces

Dor Tsur, *Student Member, IEEE*, Ziv Aharoni, *Student Member, IEEE*, Ziv Goldfeld, *Member, IEEE*, and Haim Permuter, *Senior Member, IEEE*

***Abstract*— This work develops a new method for estimating and optimizing the directed information rate between two jointly stationary and ergodic stochastic processes. Building upon recent advances in machine learning, we propose a recurrent neural network (RNN)-based estimator which is optimized via gradient ascent over the RNN parameters. The estimator does not require prior knowledge of the underlying joint/marginal distributions and can be easily optimized over continuous input processes realized by a deep generative model. We prove consistency of the proposed estimation and optimization methods and combine them to obtain end-to-end performance guarantees. Applications for channel capacity estimation of continuous channels with memory are explored, and empirical results demonstrating the scalability and accuracy of our method are provided. When the channel is memoryless, we investigate the mapping learned by the optimized input generator.**

***Index Terms*— Channel capacity, directed information, neural estimation, recurrent neural networks.**

I. Introduction

DIRECTED information (DI), introduced by Massey [2], quantifies the amount of information one stochastic process causally conveys about another. It possesses structural properties that render it as the natural causal analog of mutual information (MI), and it emerges as the solution to various operational problems involving causality [3]. Applications of DI are abundant, from the capacity of communication channels with or without memory, which is generally given by maximized DI [4], [5], to causal hypothesis testing and portfolio theory [6], where DI intricately relates to optimal tests and investment strategies, respectively. DI has also seen a myriad applications in machine learning [7], [8], [9], [10], [11], neuroscience [12], [13], [14], and control [15], [16], to name a few. It is oftentimes of interest not only to evaluate DI but also to optimize it (e.g., to characterize capacity, to bound growth rates of optimal portfolios, to extract informative features, etc.). However, this optimization is challenging since analytic computation of DI requires knowledge of the underlying probability law, which is typically unavailable in practice. Furthermore, even when the probability law is given, tractable DI characterizations that lend well for optimization are rare [17], [18], as it is generally given by a multiletter expression. To address this, the goal of the paper is to develop a computable and provably accurate estimate of DI.

Manuscript received 28 March 2022; revised 14 December 2022; accepted 26 March 2023. Date of publication 5 April 2023; date of current version 14 July 2023. This work was supported in part by the German Research Foundation (DFG) via the German Israeli Project Cooperation (DIP), in part by the Israeli Science Foundation (ISF) under Grant 899/21, and in part by the Israeli Innovation Authority as part of the Worldwide Innovative Networking (WIN) Consortium. The work of Ziv Goldfeld was supported in part by the NSF CAREER Award under Grant CCF-2046018, in part by NSF under Grant DMS-2210368, and in part by the 2020 IBM Academic Award. An earlier version of this paper was presented in part at the International Symposium on Information Theory (ISIT) 2020 [DOI: 10.1109/ISIT44484.2020.9174109]. *(Corresponding author: Dor Tsur.)*

Dor Tsur, Ziv Aharoni, and Haim Permuter are with the Department of Electrical and Computer Engineering, Ben-Gurion University of the Negev, Beer-Sheva 84105, Israel (e-mail: dortz@post.bgu.ac.il).

Ziv Goldfeld is with the School of Electrical and Computer Engineering, Cornell University, Ithaca, NY 14850 USA.

Communicated by E. Gassiat, Associate Editor for Machine Learning and Statistics.

Digital Object Identifier 10.1109/TIT.2023.3264674

A. Estimation and Optimization of Directed Information

Existing estimators of DI operate under rather restrictive assumptions on the data, hence covering a small class of problems. DI estimation between discrete-valued processes using universal probability assignments and context tree weighting was studied in [19]. Their estimator is provably consistent, but requires that the depth of the context tree is greater than the assumed memory of the processes. An approach based on maximum likelihood estimation of the associated PMF was developed in [20]. However, both [19] and [20] are limited to the class of discrete-valued, stationary Markov processes of relatively small order. Continuous-valued processes, which are of central practical interest, were treated in [21], and [22] using k nearest neighbors (kNN) estimation techniques, but as the memory or dimension of the data increase, the performance of kNN-based techniques deteriorates, due to the curse of dimensionality [23].

Neural estimation is a modern technique for estimating divergences and information measures. Originally proposed in [24], the MI neural estimator (MINE) parametrizes the Donsker-Varadhan (DV) variational form [25] by a neural network (NN), and optimizes it over a parameter space. Several variations of the MINE were proposed in followup work, e.g., replacing the DV representation with other variational lower bounds [26], [27], or by incorporating auxiliary distributions [28]. Consistency of MINE in the infinite-width NN regime was established in [24], and non-asymptotic error bounds were later derived in [29] and [30]. The latter, in particular, showed that MINE is minimax optimal

under appropriate regularity assumptions on the distributions (see also [31] for formal limitations on MINE performance). For data with memory, [32] leveraged MINE for transfer entropy, while [33] constructed a conditional MI estimator and extended it to DI between 1st order Markov processes.

In many applications, it is of interest to optimize DI over the involved processes. A prominent example is channel capacity computation, which is also the main application considered herein. Tools from dynamic programming were used in [34] and [35] to estimate the feedback capacities of a class of binary finite state channels (FSCs). This approach was later generalized to large discrete alphabets using reinforcement learning [36]. Another approach towards maximizing information measures relies on the Blahut-Arimoto (BA) algorithm [37], [38], originally proposed for MI maximization between discrete random variables. Subsequently, the algorithm was extended to FSCs [39], to DI [40], and to MI between continuous random variables [41]. The main drawback of BA algorithms is that they require full knowledge of the involved densities or the availability of consistent estimates thereof. Moreover, the continuous BA algorithm is based on space quantization, and therefore its computational complexity grows exponentially with the variables dimension.

B. Contributions

Building on the computational potency of modern machine learning techniques, we develop herein a neural estimation and optimization framework for the DI rate between continuous-valued stochastic processes. Inspired by [24] and [28], we derive the DI neural estimator (DINE) by expressing DI in terms of certain Kullback-Leibler (KL) divergences (plus cross-entropy residuals) and invoking the DV representation to arrive at a variational form. To account for causal dependencies, we parametrize the DV feasible set with the set of recurrent neural networks (RNNs) and approximate expected values by sample means. This results in a parametrized empirical objective that lends well to gradient-based optimization. We prove that the DINE is consistent whenever the stochastic processes are stationary and ergodic. The proof is based in a generalized version of Birkhoff's ergodic theorem [42], martingale analysis, and the universal approximation property of RNNs [43].

Having the DINE, we consider optimization of the estimated DI rate over the input stochastic process. To that end, we simulate the input process by an RNN deep generative model, whose parameters can be tuned to increase the estimated DI rate. By jointly optimizing the DINE and the input generative model, we obtain an estimation-optimization scheme for estimating the capacity of continuous channels with memory. Consistency of the overall method is established using the functional representation lemma (FRL) [44], [45] and universal approximation arguments [43]. We provide an extensive empirical study of the proposed method, demonstrating its efficiency and accuracy in estimating the feedforward and feedback capacities of various channels with and without memory, encompassing the average and peak power constrained additive white Gaussian noise channels (AWGN) [46], [47], [48], moving-average (MA) AWGN [49] and MIMO auto-regressive (AR) AWGN channels [50]. Lastly, we discuss the structure of the learned optimal input distribution and furnish connections to probability integral transforms. We note that following the earlier conference version of this paper [1], several neural optimization techniques were proposed [51], [52], [53] and an empirical comparison was the focus of [54]. However, these methods are only applicable to memoryless channels.

C. Organization

The text is organized as follows. Section II provides preliminaries and technical background. Section III summarizes the main results of this paper. Section IV derives the DINE, provides theoretical guarantees, and discusses its implementation. The optimization procedure of DINE over continuous-valued input processes is the focus of Section V, where consistency of the overall method and implementation details are also given. Section VI provides empirical results for channel capacity estimation. Proofs are given in Section VIII, while Section VII provides concluding remarks and discusses future research directions.

II. Background and Preliminaries

A. Notation

Subsets of the d-dimensional Euclidean space are denoted by calligraphic letters, e.g., $\mathcal{X} \subseteq \mathbb{R}^d$. For any $n \in \mathbb{N}$, $\mathcal{X}^n$ is the n-fold Cartesian product of $\mathcal{X}$, while $x^n = (x_1, \ldots, x_n)$ denotes an element thereof. For $i, j \in \mathbb{Z}$ with $i \leq j$, we use the shorthand $x_i^j := (x_i, \ldots, x_j)$; the subscript is omitted when $i = 1$. We denote by $(\Omega, \mathcal{F}, \mathbb{P})$ the underlying probability space on which all random variables are defined, with $\mathbb{E}$ denoting expectation. The set of all Borel probability measures on $\mathcal{X} \subseteq \mathbb{R}^d$ is denoted by $\mathcal{P}(\mathcal{X})$. The subset of $\mathcal{P}(\mathcal{X})$ of Lebesgue absolutely continuous measures is denoted by $\mathcal{P}_{\mathsf{ac}}(\mathcal{X})$. The density of $P \in \mathcal{P}_{\mathsf{ac}}(\mathcal{X})$ is designated by its lowercase version p; n-fold product extensions of P and p are denoted by $P^{\otimes n}$ and $p^{\otimes n}$, respectively. Random variables are denoted by upper-case letters, e.g., X, using the same conventions as above for random vectors. Stochastic processes are denoted by blackboard bold letters, e.g., $\mathbb{X} := (X_i)_{i \in \mathbb{N}}$.

For $P, Q \in \mathcal{P}(\mathcal{X})$ such that $Q \ll P$, i.e., Q is absolutely continuous with respect to (w.r.t.) P, we denote the Radon-Nikodym derivative of P w.r.t. Q by $\frac{\mathrm{d}P}{\mathrm{d}Q}$. The KL divergence between P and Q is $\mathsf{D}_{\mathsf{KL}}(P\|Q) := \mathbb{E}_P\big[\log \frac{\mathrm{d}P}{\mathrm{d}Q}\big]$. If $Q \in \mathcal{P}_{\mathsf{ac}}(\mathcal{X})$ with probability density function (PDF) q, then the cross-entropy between P and Q is $\mathsf{h}_{\mathsf{CE}}(P, Q) := -\mathbb{E}_P[\log q]$. The MI between $(X, Y) \sim P_{XY} \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ is $\mathsf{I}(X;Y) := \mathsf{D}_{\mathsf{KL}}(P_{XY}\|P_X \otimes P_Y)$, where P_X and P_Y are the marginals of P_{XY}. The conditional KL divergence between two probability measures $P_{Y|X}$, $Q_{Y|X}$ with $X \sim P_X$ is given by $\mathsf{D}_{\mathsf{KL}}(P_{Y|X}\|Q_{Y|X}|P_X) := \mathbb{E}_{P_X}\big[\mathsf{D}_{\mathsf{KL}}(P_{Y|X}\|Q_{Y|X})\big]$. Consequently, for $(X, Y, Z) \sim P_{XYZ}$, we define the conditional MI as $I(X;Y|Z) := \mathsf{D}_{\mathsf{KL}}(P_{XY|Z}\|P_{X|Z} \otimes P_{Y|Z}|P_Z)$. The differential entropy of

$X \sim P$ is $\mathsf{h}(X) := \mathsf{h}_{\mathsf{CE}}(P,P)$ whenever $P \in \mathcal{P}_{\mathsf{ac}}(\mathcal{X})$. We denote the convolution between two probability measures μ and ν with $(\mu * \nu)(A) := \int\int \mathbb{1}_A(x+y)\,\mathrm{d}\mu(x)\,\mathrm{d}\nu(y)$ and $\mathbb{1}_A$ as the indicator of A For an open set $\mathcal{U} \subseteq \mathbb{R}^d$ and $k \in \mathbb{N}$, the class of functions such that all partial derivatives up to order k exist and are continuous is denoted by $\mathcal{C}^k(\mathcal{U})$, with $\mathcal{C}(\mathcal{U}) := \mathcal{C}^0(\mathcal{U})$ and we denote by $\partial^j_{x_i} f$ the jth order partial derivative of f w.r.t. x_i.

B. Directed Information and Channel Capacity

Originally proposed by Massey [2], DI quantifies the amount of information one sequence of random variables causally conveys about another.

Definition 1 (Directed Information): Let $(X^n, Y^n) \sim P_{X^nY^n} \in \mathcal{P}(\mathcal{X}^n \times \mathcal{Y}^n)$. The DI from X^n to Y^n is

$$\mathsf{I}(X^n \to Y^n) := \sum_{i=1}^{n} \mathsf{I}(X^i; Y_i | Y^{i-1}). \tag{1}$$

DI entails the concept of causal conditioning, i.e., conditioning only on present and past values of the sequences, which is seen through its decomposition using causal conditioned (CC) entropies [55]. For $(X^n, Y^n) \sim P_{X^nY^n} \in \mathcal{P}(\mathcal{X}^n \times \mathcal{Y}^n)$, the entropy of Y^n CC on X^n is given by

$$\mathsf{h}\left(Y^n \| X^n\right) := \mathbb{E}\left[-\log p_{Y^n\|X^n}\left(Y^n \| X^n\right)\right],$$

where $p_{Y^n\|X^n}\left(y^n \| x^n\right) := \prod_{i=1}^n p_{Y_i|Y^{i-1},X^i}\left(y_i | y^{i-1}, x^i\right)$ is the CC-PDF of Y^n given $X^n = x^n$. As the CC entropy can be expressed as $\mathsf{h}(Y^n\|X^n) := \sum_{i=1}^n \mathsf{h}(Y^i|X^i, Y^{i-1})$, we have the following representation for DI:

$$\mathsf{I}\left(X^n \to Y^n\right) = \mathsf{h}\left(Y^n\right) - \mathsf{h}\left(Y^n \| X^n\right). \tag{2}$$

This poses DI as the reduction in the uncertainty about Y^n as a result of causally observing (the elements of) X^n. Since DI (as well as MI) tends to grow with the number of observations, the appropriate figure of merit when considering stochastic processes is the DI *rate*.

Definition 2 (Directed Information Rate): Let $\mathbb{X}$ and $\mathbb{Y}$ be jointly stationary stochastic processes. The DI rate from $\mathbb{X}$ to $\mathbb{Y}$ is given by

$$\mathsf{I}(\mathbb{X} \to \mathbb{Y}) := \lim_{n\to\infty} \frac{1}{n} \mathsf{I}(X^n \to Y^n). \tag{3}$$

The limit exists whenever the processes are jointly stationary [56]. Due to the averaging, the DI rate captures prominent interactions, while the effect of transient phenomena decays to zero.

Remark 1 (Channel Capacity): We consider channels with and without a feedback link from the channel output back to the encoder. The feedforward capacity of a sequence of channels $\{P_{Y^n\|X^n}\}_{n\in\mathbb{N}}$ is [4][1]

$$C_{\mathsf{FF}} = \lim_{n\to\infty} \sup_{P_{X^n}} \frac{1}{n} \mathsf{I}(X^n; Y^n). \tag{4}$$

[1]This formula assumes the so-called information stability property (see [57]).

In the presence of feedback, the capacity becomes [58]

$$C_{\mathsf{FB}} = \lim_{n\to\infty} \sup_{P_{X^n\|Y^{n-1}}} \frac{1}{n} \mathsf{I}(X^n \to Y^n). \tag{5}$$

The achievability of (4) and (5) is discussed in [57] and [58], respectively. As shown in [2, Theorem 1], when feedback is not present, the optimization problem (5) (which amounts to optimizing over P_{X^n} rather than $P_{X^n\|Y^n}$) coincides with (4). Thus, DI provides a unified framework for the calculation of both feedforward and feedback capacities.

C. Neural Networks and Recurrent Neural Networks

The class of shallow NNs with fixed input and output dimensions is defined as follows [59].

Definition 3 (NN Function Class): For the ReLU activation function $\sigma_{\mathsf{R}}(x) = \max(x, 0)$ and $d_i, d_o \in \mathbb{N}$, define the class of neural networks with $k \in \mathbb{N}$ neurons as:

$$\mathcal{G}_k^{(d_i,d_o)} := \left\{ g : \mathbb{R}^{d_i} \to \mathbb{R}^{d_o} : g(x) = \sum_{j=1}^{k} \beta_j \sigma_{\mathsf{R}}(\mathrm{W}_j x - b_j), x \in \mathbb{R}^{d_i} \right\}, \tag{6}$$

where σ_{R} acts component-wise, $\beta_j \in \mathbb{R}, \mathrm{W}_j \in \mathbb{R}^{d_o \times d_i}$ and $b_j \in \mathbb{R}^{d_o}$ are the parameters of $g \in \mathcal{G}_k^{(d_i,d_o)}$. Then, the class of NNs with input and output dimensions (d_i, d_o) is given by

$$\mathcal{G}_{\mathsf{nn}}^{(d_i,d_o)} := \bigcup_{k\in\mathbb{N}} \mathcal{G}_k^{(d_i,d_o)}. \tag{7}$$

NNs form a universal approximation class under mild smoothness conditions [59]. However, feedforward networks such as those defined in (6) cannot capture temporal evolution, which is inherent to DI. Therefore, our neural estimator employs RNNs [60], which map an input sequence $(u_t)_{t=1}^T \subset \mathbb{R}^{d_i}$ to an output sequence $(x_t)_{t=1}^T \subset \mathbb{R}^{d_o}$ via a parametric nonlinear time-invariant transformation. The class of RNNs is defined as follows.

Definition 4 (RNN Function Class): Fix $k, d_i, d_o \in \mathbb{N}$. The class $\mathcal{G}_{\mathsf{rnn}}^{(d_i,d_o,k)}$ of RNNs with k neurons and input-output dimensions (d_i, d_o) is the set of discrete-time, nonlinear systems with the following structure:

$$\begin{aligned} s_t &= -\alpha s_{t-1} + \mathrm{A}\sigma(s_{t-1} + \mathrm{B}x_t), \\ y_t &= \mathrm{C}s_t, \end{aligned} \tag{8}$$

where $s_0 \in \mathbb{R}^k$ is the initial RNN state, $s_t \in \mathbb{R}^k$, $x_t \in \mathbb{R}^{d_i}$, and $y_t \in \mathbb{R}^{d_o}$ are, respectively, the state, input, and output (column) vectors, $\mathrm{A} \in \mathbb{R}^{k\times k}$, $\mathrm{B} \in \mathbb{R}^{k\times d_i}$, and $\mathrm{C} \in \mathbb{R}^{d_o\times k}$ are the associated weight matrices, $\alpha \in (-1, 1)$ is a fixed constant for controlling state decay, and $\sigma(x) = \frac{1}{1+e^{-x}}$ is the sigmoid function which acts on vectors component-wise. The class of RNNs (of all possible sizes) with dimensions (d_i, d_o) is defined as

$$\mathcal{G}_{\mathsf{rnn}}^{(d_i,d_o)} := \bigcup_{k\in\mathbb{N}} \mathcal{G}_{\mathsf{rnn}}^{(d_i,d_o,k)}. \tag{9}$$

Every element of $\mathcal{G}_{\mathsf{rnn}}^{(d_i,d_o)}$ is a function, such that for a given input sequence, its output is calculated sequentially according

to (8). An RNN is therefore a causal input-output discrete-time mapping of T elements of $\mathbb{R}^{d_i}$ to T elements of $\mathbb{R}^{d_o}$, for any $d_i, d_o, T \in \mathbb{N}$.

Note that both $\mathcal{G}_k^{(d_i,d_o)}$ and $\mathcal{G}_{\mathsf{rnn}}^{(d_i,d_o,k)}$ are parametric models whose (finitely many) parameters belong to some parameter space $\Theta \subset \mathbb{R}^d$, for an appropriate dimension d. When k is fixed, interchangeably denote functions from the above classes explicitly, as $g \in \mathcal{G}_k^{(d_i,d_o)}$, or in their corresponding parametrized form: g_θ where $\theta \in \Theta$.

D. Mutual Information Neural Estimation

The mutual information neural estimator (MINE) [24] is a NN-based estimator of the MI between two random variables. The technique relies on the DV variational representation of KL divergence [25, Theorem 3.2].

Theorem 1 (DV Representation): For any $P, Q \in \mathcal{P}(\mathcal{X})$, we have

$$\mathsf{D}_{\mathsf{KL}}(P\|Q) = \sup_{f:\mathcal{X}\to\mathbb{R}} \mathbb{E}_P[f] - \log\left(\mathbb{E}_Q[e^f]\right), \tag{10}$$

where the supremum is taken over all measurable functions f for which expectations are finite.

Given n pairwise independent and identically distributed (i.i.d.) samples $D_n := (X^n, Y^n)$ from $P_{XY} \in \mathcal{P}(\mathcal{X}\times\mathcal{Y})$, the MINE parametrizes f by a NN $g \in \mathcal{G}_{\mathsf{nn}} := \mathcal{G}_{\mathsf{nn}}^{(d_x+d_y,1)}$ and approximates expectations by sample means:

$$\widehat{\mathsf{I}}_{\mathsf{MI}}(D_n) := \sup_{g\in\mathcal{G}_{\mathsf{nn}}} \underbrace{\frac{1}{n}\sum_{i=1}^n g(X_i,Y_i) - \log\left(\frac{1}{n}\sum_{i=1}^n e^{g(X_i,\bar{Y}_i)}\right)}_{=\widehat{\mathsf{I}}_{\mathsf{MI}}(D_n,g)}, \tag{11}$$

where $(X_i, \bar{Y}_i) \sim P_X \otimes P_Y$. The functions over which we optimize the DV objective are termed DV potentials. We stress that only the correlated samples from D_n are given, so negative (i.e., independent) samples must be constructed from them, e.g., by random permutation [24]. In [24, Theorem 2] the strong consistency of MINE is proved, i.e., $\lim_{n\to\infty} \widehat{\mathsf{I}}_{\mathsf{MI}}(D_n) = \mathsf{I}(X;Y)$, $\mathbb{P}$-almost surely (a.s.).

Remark 2 (Non-Asymptotic Error Bound): Non-asymptotic error bounds for neural estimation of f-divergence were recently derived in [30]. Specifically, they established bounds on the effective (approximation plus empirical estimation) error of a neural estimator realized by a k-neuron shallow NN with bounded parameters and n data samples. Instantiating their result for the $\mathsf{D}_{\mathsf{KL}}(P\|Q)$ with $P = P_{XY}$ and $Q = P_X \otimes P_Y$ yields an $O\big(d^{1/2}k^{-1/2} + d^{3/2}(\log k)^7 n^{-1/2}\big)$ error bound for MI estimation, uniformly over a class of sufficiently regular d-dimensional distributions with bounded supports. Evidently, there is a fundamental tradeoff between the two sources of error: while good approximation needs the NN class to be rich and expressive, empirical estimation error bounds rely on controlling complexity.

Due to the consistency of the MINE, and since parameterization can only shrink the DV function class, it provably lower bounds the ground truth MI in the limit of large samples.

Lemma 1 (MINE Lower Bounds MI): For any $g \in \mathcal{G}_{\mathsf{nn}}$, we have

$$\mathsf{I}(X;Y) \geq \lim_{n\to\infty} \widehat{\mathsf{I}}_{\mathsf{MI}}(D_n, g), \qquad \mathbb{P}\text{-}a.s. \tag{12}$$

This property implies that the probability that MINE will overestimate MI is small. This property is central when the target MI is the underlying capacity of some communication channel, as we can state that the estimate provides a lower bound of it at worst. This property will be further discussed in the context of the proposed methods.

III. Main Results

This work develops a principled framework for neural estimation and optimization of information measures, which is then leveraged to estimate the feedforward and feedback capacities of general channels. To that end we propose the DINE, which generalizes the MINE for DI rate, and develop methods for optimizing MINE and DINE over continuous channel input distributions. While channel capacity estimation is the focus of this work, the proposed estimation and optimization techniques are applicable to any DI optimization scenario.

A. Directed Information Neural Estimation

We set up the DINE, state its consistency, and provide a pseudo-algorithm for its computation. We construct the DINE as the difference between two DV-based KL estimators. Given a sample $D_n = (X^n, Y^n) \sim P_{X^nY^n}$ and RNNs $g_y \in \mathcal{G}_{\mathsf{rnn}}^Y := \mathcal{G}_{\mathsf{rnn}}^{(d_y,1)}$ and $g_{xy} \in \mathcal{G}_{\mathsf{rnn}}^{XY} := \mathcal{G}_{\mathsf{rnn}}^{(d_y+d_x,1)}$, the DINE is given by

$$\widehat{\mathsf{I}}_{\mathsf{DI}}(D_n) := \sup_{g_{xy}\in\mathcal{G}_{\mathsf{rnn}}^{XY}} \widehat{\mathsf{D}}_{Y\|X}(D_n, g_{xy}) - \sup_{g_y\in\mathcal{G}_{\mathsf{rnn}}^{Y}} \widehat{\mathsf{D}}_Y(D_n, g_y),$$

where $\widehat{\mathsf{D}}_Y, \widehat{\mathsf{D}}_{Y\|X}$ are given by

$$\widehat{\mathsf{D}}_Y(D_n, g_y) := \frac{1}{n}\sum_{i=1}^n g_y\left(Y^i\right) - \log\left(\frac{1}{n}\sum_{i=1}^n e^{g_y\left(\widetilde{Y}_i, Y^{i-1}\right)}\right) \tag{13a}$$

$$\widehat{\mathsf{D}}_{Y\|X}(D_n, g_{xy}) := \frac{1}{n}\sum_{i=1}^n g_{xy}\left(Y^i, X^i\right) - \log\left(\frac{1}{n}\sum_{i=1}^n e^{g_{xy}\left(\widetilde{Y}_i, Y^{i-1}, X^i\right)}\right), \tag{13b}$$

and $\widetilde{Y}^n \overset{i.i.d.}{\sim} \mathsf{Unif}(\mathcal{Y})$. A full derivation of the estimator and further implementation details are provided in Section IV. As stated next, the DINE is a consistent estimator of the DI rate.

Theorem 2 (Consistency): Suppose $\mathbb{X}$ and $\mathbb{Y}$ are jointly stationary, ergodic, regular stochastic processes. Then the DINE is a strongly consistent estimator of $\mathsf{I}(\mathbb{X}\to\mathbb{Y})$, i.e., $\mathbb{P}$-*a.s.* for every $\epsilon > 0$, there exists $N \in \mathbb{N}$ such that for every $n > N$ we have

$$\left|\widehat{\mathsf{I}}_{\mathsf{DI}}(D_n) - \mathsf{I}(\mathbb{X}\to\mathbb{Y})\right| \leq \epsilon. \tag{14}$$

To compute the DINE in practice notice that $\mathcal{G}_{\mathsf{rnn}}^Y$ and $\mathcal{G}_{\mathsf{rnn}}^{XY}$ are parametric classes. We fix k and take their k-dimensional

counterparts whose (finitely many) parameters belong to some parameter space $\Theta \subset \mathbb{R}^d$, for an appropriate dimension d. We therefore denote the DINE RNNs with g_{θ_y} and $g_{\theta_{xy}}$, and optimize the DINE objective over their parameters (θ_y, θ_{xy}).

Algorithm 1 DINE

Input: Dataset D_n.

Output: $\widehat{\mathsf{I}}_{\mathsf{DI}}(D_n)$ DI rate estimate.

Initialize g_{θ_y}, $g_{\theta_{xy}}$ with parameters θ_y, θ_{xy}.

Step 1 – Parameter optimization:

repeat

Draw a batch B_m for $m < n$ & sample $P_{\tilde{Y}}$.

Compute $\widehat{\mathsf{D}}_{Y\|X}(B_m, g_{\theta_{xy}})$, $\widehat{\mathsf{D}}_{Y}(B_m, g_{\theta_y})$ using (13).

Update networks parameters:

$$\theta_{xy} \leftarrow \theta_{xy} + \nabla_{\theta_{xy}} \widehat{\mathsf{D}}_{Y\|X}(B_m, g_{\theta_{xy}})$$
$$\theta_{y} \leftarrow \theta_{y} + \nabla_{\theta_{y}} \widehat{\mathsf{D}}_{Y}(B_m, g_{\theta_{y}})$$

until convergence.

Step 2 – Evaluation: Evaluate over a sample D_n and subtract losses to obtain $\widehat{\mathsf{I}}_{\mathsf{DI}}(D_n)$ (20).

As delineated in Algorithm 1, the DI estimation algorithm consists of a training step and an evaluation step. At each iteration of the training step, a batch is drawn and set of reference samples is generated. The samples are then sequentially processed by the models $(g_{\theta_y}, g_{\theta_{xy}})$, whose outputs are fed to the KL estimates in (13). The estimates are then used to update the model parameters according to the steepest descent direction. When the training phase is done, an estimate of the DI rate is calculated from a long sequence (X^n, Y^n).

B. DINE Optimization Over Continuous Spaces

Given a sequence of transition kernels $\{P_{Y^n\|X^n}\}_{n\in\mathbb{N}}$ that models a communication channel, we propose a method for optimizing the DINE over continuous input distributions. Specifically, we employ an RNN generative model termed the *neural distribution transformer* (NDT), denoted $h_\phi \in \mathcal{G}_{\mathsf{rnn}}^{(d_x,d_x,k)}$, where $k, d_x \in \mathbb{N}$ and $\phi \in \Phi$ are its parameters. Let $U^n \sim P_U^{\otimes n}$ for some $P_U \in \mathcal{P}_{\mathsf{ac}}(\mathcal{U})$ and $\mathcal{U} \subset \mathbb{R}^{d_x}$. We define h_ϕ through the following recursive relation

$$h_\phi : (U_i, Z_{i-1}^\phi) \mapsto X_i^\phi, \qquad i = 1, \ldots, n,$$

where Z_i^ϕ is determined according to whether feedback is present or not. By sampling P_U and passing those samples through h_ϕ and the channel, we generate a dataset $D_n^\phi = (X^{\phi,n}, Y^{\phi,n})$. We optimize the DINE over ϕ such that D_n^ϕ corresponds to the distribution that maximizes the DINE.

As described by Algorithm 2, our scheme alternates between the optimization of the NDT and the DINE models. Each iteration begins with a choice of one of the models. Then, a set of channel input-output samples are calculated via the NDT mapping and channel models. These are then fed into the DINE and the corresponding DV-based loss is calculated, from which gradients are drawn for parameters update of the chosen model.

Algorithm 2 Continuous DINE Optimization

Input: Continuous channel, feedback indicator.

Output: $\widehat{\mathsf{I}}^\star_{\mathsf{DI}}(U^n)$, optimized NDT.

Initialize $g_{\theta_y}, g_{\theta_{xy}}$ and h_ϕ with parameters $\theta_y, \theta_{xy}, \phi$.

if feedback indicator **then**

Add feedback to NDT.

repeat

Draw noise U^m, $m < n$.

Compute B_m^ϕ using NDT and channel

if training DINE **then**

Perform DINE optimization according to step 1 in Algorithm 1.

else (Train NDT)

Compute $\widehat{\mathsf{I}}_{\mathsf{DI}}(B_m^\phi, g_{\theta_y}, g_{\theta_{xy}}, h_\phi)$ using (13).

Update NDT parameters:

$$\phi \leftarrow \phi + \nabla_\phi \widehat{\mathsf{I}}_{\mathsf{DI}}(B_m^\phi, g_{\theta_y}, g_{\theta_{xy}}, h_\phi)$$

until convergence.

Draw U^n and perform a Monte Carlo evaluation of $\widehat{\mathsf{I}}_{\mathsf{DI}}(D_n^\phi)$.

return $\widehat{\mathsf{I}}^\star_{\mathsf{DI}}(U^n)$, optimized NDT.

We prove the convergence of the joint estimation-optimization method. We assume that both the channel and input process adhere to a recursive nonlinear and stationary state space model. We further assume that the channel output and state mappings, given by f_{y} and f_{z}, meet some Lipschitz continuity criterion (this is summarized by Assumption 1, in Section V-B.1). We denote the class of such input processes by $\mathsf{X}_{\mathcal{S}}$ and denote the maximal DI rate over $\mathsf{X}_{\mathcal{S}}$ by $\underline{\mathsf{C}}_s$. We propose the following

Theorem 3 (Strong Consistency of the DINE-NDT Method): Let $\{P_{Y_i|Y^{i-1},X^i}\}_{i\in\mathbb{N}}$ be a continuous unifilar state channel, where $f_{\mathsf{y}}, f_{\mathsf{z}}$ satisfy Assumption 1. Then, $\mathbb{P}$-*a.s.* for every $\epsilon > 0$, there exist $N \in \mathbb{N}$ such that for every $n > N$ and $U^n \sim P_U^{\otimes n}$ we have

$$\left|\underline{\mathsf{C}}_s - \widehat{\mathsf{I}}^\star_{\mathsf{DI}}(U^n)\right| \leq \epsilon, \tag{15}$$

where $\widehat{\mathsf{I}}^\star_{\mathsf{DI}}(U^n) = \sup_{h_\phi \in \mathcal{G}_{\mathsf{rnn}}^{(d_x,d_x)}} \widehat{\mathsf{I}}_{\mathsf{DI}}(D_n^\phi, h_\phi)$.

For memoryless channels, where capacity is given by the maximized MI, we consider MINE optimization and identify the optimized NDT structure via multivariate generalization of the capacity achieving input cumulative distribution function (CDF), obtained by vectorizing the product of conditional CDFs of the entries of X (see Theorem 6). For the full details of the theoretical guarantees, see Section V-B.

As described in Algorithm 2, the joint DI estimation-maximization procedure involves alternating optimization between the DINE and NDT models. In Section VI, we demonstrate the end-to-end procedure by estimating the capacity of several channels, with and without memory, accounting for both feedforward and feedback capacities. We empirically demonstrate the accuracy of the algorithm by comparing it with known results/bounds and analyse the optimized NDT.

IV. Directed Information Neural Estimation

This section describes the DI estimation method introduced in Section III-A. We consider two jointly stationary and ergodic processes $\mathbb{X}$ and $\mathbb{Y}$, supported on $\mathcal{X} \subseteq \mathbb{R}^{d_x}$ and $\mathcal{Y} \subseteq \mathbb{R}^{d_y}$, respectively. Our goal is to devise a provably consistent neural estimator of the DI rate from $\mathbb{X}$ to $\mathbb{Y}$ based on a finite sample of these processes. The section is organized as follows. We begin by demonstrating the difficulty of generalizing the MINE framework to the DI estimation. We then derive the DINE, discuss theoretical guarantees, and illustrate its implementation.

A. Challenges in Generalizing MINE to Directed Information

Recall that the MINE (11) is derived by approximating the potentials in the DV variational formula with NNs, and estimating expectations by sample means. Generalizing to DI, we consider the conditional MI corresponding to the DI rate through $\lim_{n\to\infty} \mathsf{I}(X^n;Y_n|Y^{n-1}) = \mathsf{I}(\mathbb{X}\to\mathbb{Y})$ [19]. The corresponding KL term is given by

$$\mathsf{I}(X^n;Y_n|Y^{n-1}) = \mathsf{D}_{\mathsf{KL}}\big(\underbrace{P_{Y_n|Y^{n-1}X^n}}_{P_{\mathsf{pos}}} \big\| \underbrace{P_{Y_n|Y^{n-1}}}_{P_{\mathsf{neg}}} \big| P_{X^nY^{n-1}}\big), \tag{16}$$

where $\mathsf{D}_{\mathsf{KL}}(P_{Y|X}\|Q_{Y|X}|P_X)$ is the conditional KL divergence. Estimating the expectations in DV representation of (16) requires samples of both P_{pos} and P_{neg}, while only samples of $P_{X^nY^n}$ are available. Samples of P_{pos} are the sampled channel outputs. On the other hand, samples from P_{neg} require some manipulation of the data to break the relation between $\mathbb{X}$ and $\mathbb{Y}$, but maintain temporal inter dependencies. For i.i.d. data, random permutation of the samples is proposed [24], and for 1st order Markov processes the 1-nearest neighbors algorithm is utilized [61]. To the best of our knowledge, such a technique is unknown for unbounded memory, as previous methods either affect both dependencies between $\mathbb{X}$ and $\mathbb{Y}$ or the the temporal relations are restricted only to short memory. As a solution, we derive a DV-based estimator of the DI rate that solely relies on samples from P_{pos}, exploiting samples from an auxiliary distribution over $\mathcal{Y}$.

B. The Estimator

The DINE derivation relies on the following steps: First, we express DI as the difference between certain KL divergence terms. These are then represented via the DV variational formula (Theorem 1). Then, the DV potentials are parametrized using RNNs, and expected values are approximated by sample means. Recall that DI is given by

$$\mathsf{I}\left(X^n \to Y^n\right) = \mathsf{h}\left(Y^n\right) - \mathsf{h}\left(Y^n\|X^n\right). \tag{17}$$

For simplicity, assume $\mathcal{Y}$ is compact[2], and let $\widetilde{Y} \sim \mathsf{Unif}(\mathcal{Y}) =: P_{\widetilde{Y}}$ be independent of $\mathbb{X}$ and $\mathbb{Y}$. Using the uniform reference measure we expand each entropy term as

$$\mathsf{h}(Y^n) = \mathsf{h}_{\mathsf{CE}}\left(P_{Y^n}, P_{Y^{n-1}}\otimes P_{\widetilde{Y}}\right) - \mathsf{D}_{\mathsf{KL}}\left(P_{Y^n}\middle\|P_{Y^{n-1}}\otimes P_{\widetilde{Y}}\right) \tag{18a}$$

$$\mathsf{h}(Y^n\|X^n) = \mathsf{h}_{\mathsf{CE}}\left(P_{Y^n\|X^n}, P_{Y^{n-1}\|X^{n-1}}\otimes P_{\widetilde{Y}}\middle|P_{X^n\|Y^{n-1}}\right) - \mathsf{D}_{\mathsf{KL}}\left(P_{Y^n\|X^n}\middle\|P_{Y^{n-1}\|X^{n-1}}\otimes P_{\widetilde{Y}}\middle|P_{X^n\|Y^{n-1}}\right), \tag{18b}$$

where $\mathsf{h}_{\mathsf{CE}}(P_{Y|X},Q_{Y|X}|P_X)$ is the conditional cross-entropy. With some abuse of notation, let $\mathbb{X} := \{X_i\}_{i\in\mathbb{Z}}$ and $\mathbb{Y} := \{Y_i\}_{i\in\mathbb{Z}}$ be the two-sided extension of the considered processes (the underlying stationary and ergodic measure remains unchanged). Inserting (18a)-(18b) into (17) and using joint stationarity (which guarantees the existence of the following limit) we have

$$\mathsf{I}(\mathbb{X}\to\mathbb{Y}) = \mathsf{D}^{\infty}_{Y\|X} - \mathsf{D}^{\infty}_{Y} = \lim_{n\to\infty}\mathsf{D}^{n}_{Y\|X} - \lim_{n\to\infty}\mathsf{D}^{n}_{Y},$$

with

$$\mathsf{D}^n_Y := \mathsf{D}_{\mathsf{KL}}\left(P_{Y^0_{-(n-1)}}\middle\|P_{Y^{-1}_{-(n-1)}}\otimes P_{\widetilde{Y}}\right)$$

$$\mathsf{D}^n_{Y\|X} := \mathsf{D}_{\mathsf{KL}}\left(P_{Y^0_{-(n-1)}\|X^0_{-(n-1)}}\middle\|P_{Y^{-1}_{-(n-1)}\|X^{-1}_{-(n-1)}}\otimes P_{\widetilde{Y}}\middle|P_{X^0_{-(n-1)}\|Y^{-1}_{-(n-1)}}\right). \tag{19}$$

To arrive at a variational form we make use of the DV theorem. The optimal DV potentials for D^n_Y and $\mathsf{D}^n_{Y\|X}$ can be represented as dynamical systems that are given by the recursive relation $z_{t+1} = f(z_t,u_t)$ for inputs u_t and outputs z_t, respectively. The dynamical system formulation follows from a representation of the optimal potentials in terms of the corresponding likelihood ratios. As such, these potentials can be approximated to arbitrary precision by elements of the RNN function classes $\mathcal{G}^Y_{\mathsf{rnn}}$ and $\mathcal{G}^{XY}_{\mathsf{rnn}}$ [60]. The expectations in the DV formula are estimated with sample means (see Section VIII-A, where consistency of the DINE is proved, for details). The DINE objective is given by

$$\widehat{\mathsf{I}}_{\mathsf{DI}}(D_n,g_y,g_{xy}) := \widehat{\mathsf{D}}_{Y\|X}(D_n,g_{xy}) - \widehat{\mathsf{D}}_{Y}(D_n,g_y), \tag{20}$$

where

$$\widehat{\mathsf{D}}_Y(D_n,g_y) := \frac{1}{n}\sum_{i=1}^{n} g_y\left(Y^i\right) - \log\left(\frac{1}{n}\sum_{i=1}^{n} e^{g_y\left(\widetilde{Y}_i,Y^{i-1}\right)}\right), \tag{21a}$$

$$\widehat{\mathsf{D}}_{Y\|X}(D_n,g_{xy}) := \frac{1}{n}\sum_{i=1}^{n} g_{xy}\left(Y^i,X^i\right) - \log\left(\frac{1}{n}\sum_{i=1}^{n} e^{g_{xy}\left(\widetilde{Y}_i,Y^{i-1},X^i\right)}\right). \tag{21b}$$

Consequently, the DINE is given by the optimization of (20)

$$\widehat{\mathsf{I}}_{\mathsf{DI}}(D_n) := \sup_{g_{xy}\in\mathcal{G}^{XY}_{\mathsf{rnn}}}\widehat{\mathsf{D}}_{Y\|X}(D_n,g_{xy}) - \sup_{g_y\in\mathcal{G}^{Y}_{\mathsf{rnn}}}\widehat{\mathsf{D}}_{Y}(D_n,g_y) = \sup_{g_{xy}\in\mathcal{G}^{XY}_{\mathsf{rnn}}}\inf_{g_y\in\mathcal{G}^{Y}_{\mathsf{rnn}}}\widehat{\mathsf{I}}_{\mathsf{DI}}(D_n,g_y,g_{xy}). \tag{22}$$

[2]This is a technical assumption that arises due to the choice of a uniform reference measure. By changing $P_{\widetilde{Y}}$ to, e.g., Gaussian, this assumption is removed.

The optimization can be executed via gradient-ascent over the RNN parameters.

C. Theoretical Guarantees

We now provide formal guarantees for the DINE. To that end, we consider the class of jointly stationary, ergodic, and regular stochastic processes. A stationary process $\mathbb{X}$ is called *regular* [62], [63], [64] if and only if the logarithm of its power spectral density is absolutely integrable.[3] This condition is non-restrictive as it is met for stationary processes with finite variance whose power spectral density does not contain singularities. All the examples considered in this paper (cf. the experiments in Section VI) satisfy this regularity condition. The following theorem establishes consistency of the DINE.

Theorem 4 (Theorem 2, Restated): Suppose $\mathbb{X}$ and $\mathbb{Y}$ are jointly stationary, ergodic, regular stochastic processes. Then $\mathbb{P}$-*a.s.* for every $\epsilon > 0$, there exists $N \in \mathbb{N}$ such that for every $n > N$ we have

$$\left| \widehat{\mathsf{I}}_{\mathsf{DI}}(D_n) - \mathsf{I}(\mathbb{X} \to \mathbb{Y}) \right| \leq \epsilon. \tag{23}$$

The proof can be divided into three main steps. First, an *information-theoretic* step, in which we express the DI rate as the difference of KL divergence terms, and represent it with the DV formula (10). Second, an *estimation step*, that utilizes a generalization of Birkhoff's ergodic theorem [42], [65] to approximate the expectations of the DV representation by sample means. Third, an *approximation* step, in which we show that the sequence of optimal DV potentials possesses a certain sequential structure, and approximate it using RNNs, utilizing a universal approximation theorem for RNNs [60]. The proof is given in Section VIII-A. We stress that the choice of N depends on the underlying probability law $\mathbb{P}$. This dependence is implicitly encoded in the generalized AEP and Birkhoff's theorems (Theorems 7, 8 in Section VIII-A).

Remark 3 (Bound on the Underlying DI Rate): The DINE is constructed as a difference of two maximization problems. Therefore, while the DV representation induces a lower bound on each KL term for any choice of g_y and g_{xy}, the overall objective (20) does not bound the true DI-neither from above nor below. For a DINE variant that does bound $\mathsf{I}(\mathbb{X} \to \mathbb{Y})$, one would need a variational upper bound of KL divergences that can be optimized over RNNs. To the best of our knowledge, such a representation is not known.

D. Implementation

We describe the implementation details of the DINE. Fix k_y and k_{xy} with the corresponding RNN classes $\mathcal{G}_{\mathsf{rnn}}^{(d_y,1,k_y)}$ and $\mathcal{G}_{\mathsf{rnn}}^{(d_y,1,k_{xy})}$. The corresponding compact parameter subsets are denoted $\Theta_y \subseteq \mathbb{R}^{d_{\theta_y}}$ and $\Theta_{xy} \subseteq \mathbb{R}^{d_{\theta_{xy}}}$ with finite d_{θ_y} and $d_{\theta_{xy}}$. The RNNs over which we optimize comprise a *modified* long short-term memory (LSTM) layer and a fully connected (FC) network. The architecture for $\widehat{\mathsf{D}}_Y(D_n, g_{\theta_y})$ is depicted in Figure 1. We next present the modified LSTM cell, discuss the optimization procedure, and propose an adjustment for the DINE objective that accounts for possible estimation variance induced by the reference samples.

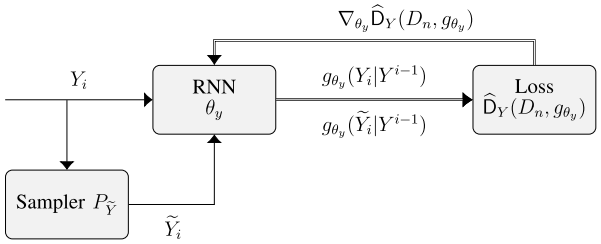

Fig. 1. The estimator architecture for the calculation of $\widehat{\mathsf{D}}_Y(D_n, g_{\theta_y})$.

1) Modified LSTM: Note that the RNN mappings in each KL estimate in (21) consists of the same mapping, each time differing on the ith input. Our goal is therefore to construct a unified mapping for samples of both the joint and reference distribution, while restricting memory to depend only on past samples from the joint distribution. To this end, we adjust the structure of the classic LSTM cell [66]. The modification is presented for $\widehat{\mathsf{D}}_Y$ and is straightforwardly adopted for $\widehat{\mathsf{D}}_{Y\|X}$. The classic LSTM is an RNN that recursively computes a hidden state s_i from its input y_i and the previous state s_{i-1}through a gating mechanism whose goal is to amplify relevant past information for computation of future states and outputs (see [66] for more background on LSTMs). We henceforth use the shorthand $s_i = f_{\mathsf{L}}(y_i, s_{i-1})$ for the relation between s_i and (y_i, s_{i-1}) defined by the LSTM. As DINE also employs the sequence $\widetilde{y}^n$ drawn from the reference distribution $P_{\widetilde{Y}}$, the modified LSTM collects hidden states for both y^n and $\widetilde{y}^n$. At time $i = 1, \ldots, n$, the cell takes a pair $(y_i, \widetilde{y}_i)$ as input, and outputs two hidden states $s_i = f_{\mathsf{L}}(y_i, s_{i-1})$ and $\widetilde{s}_i = f_{\mathsf{L}}(\widetilde{y}_i, s_{i-1})$, with only s_i passed on for calculating the next state. The state sequences are then processed by the FC network to obtain the elements of (21a). The states s_i and $\widetilde{s}_i$ calculate a summary of y^{i-1} and $\widetilde{y}^{i-1}$ through the LSTM cell recursive mapping. Therefore, we interpret the computation of $g_{\theta_y}(y^i)$ and $g_{\theta_y}(\widetilde{y}, y^{i-1})$ as conditioning on past inputs. With some abuse of notation, we use interchangeably the following conditional form for the DINE outputs.

$$\begin{aligned} g_{\theta_y}(y^i) &= g_{\theta_y}(y_i, s_{i-1}) = g_{\theta_y}(y_i|y^{i-1}) \\ g_{\theta_y}(\widetilde{y}_i, y^{i-1}) &= g_{\theta_y}(\widetilde{y}_i, s_{i-1}) = g_{\theta_y}(\widetilde{y}_i|y^{i-1}). \end{aligned} \tag{24}$$

The notation on the right-hand sides (RHSs) of (24) emphasizes that the input dimension is fixed for each time step. The calculation of the hidden states for $\widehat{\mathsf{D}}_{Y\|X}$ in (21b) is performed analogously, by replacing y_i and $\widetilde{y}_i$ with (x_i, y_i) and $(x_i, \widetilde{y}_i)$, respectively. The modified LSTM cell is shown in Figure 2.

2) Algorithm: The DINE algorithm computes $\widehat{\mathsf{I}}_{\mathsf{DI}}(D_n)$ by optimizing the parameters $\theta_y \in \Theta_y$ and $\theta_{xy} \in \Theta_{xy}$ of the RNNs g_{θ_y} and $g_{\theta_{xy}}$, respectively. We divide the dataset into batches of B sequences of length T, i.e., $B_m := (X^m, Y^m)$ with $m = BT < n$. For each batch, we provide samples of the

[3]Such processes are sometime also called 'purely non-deterministic'. Loosely speaking, the defining property of a regular (or purely non-deterministic) process $\mathbb{X}$ is that the amount of information contained in X^k for the prediction of X_t vanishes as $k \to -\infty$, i.e., remote past does not affect the prediction of X_t.

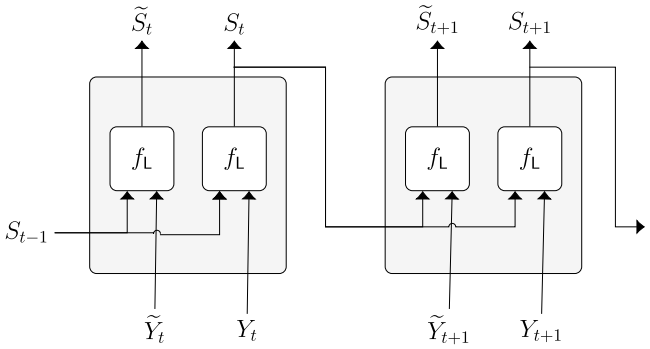

Fig. 2. The modified LSTM cell unrolled in the DINE architecture of $\widehat{\mathsf{D}}_Y$. Recursively, at each time t, (Y_t, S_{t-1}) and $(\widetilde{Y}_t, S_{t-1})$ are mapped to S_t and $\widetilde{S}_t$, respectively.

reference measure[4] and feed the sequences through the DINE architecture to obtain the DV potentials g_{θ_y} and $g_{\theta_{xy}}$. Those are then used to calculate the DINE objective (20), from which gradients are derived for the update of θ_y and θ_{xy}. We repeat the above steps until some convergence criteria is met. See Algorithm 1 in Section III for the full list of steps. The weights of the FC networks within each RNN are shared since we wish to produce the same function acting on different inputs.

3) Reference Samples: The exponential terms in (21a) and (21b) can potentially cause instability in the estimation process by biasing the estimate of the update gradients [24]. Existing methods to account for this problem include moving average filtering of the gradients [24] and clipping of the exponential terms [27]. Herein, we exploit the reference uniform measure. For each i, we collect K_U reference samples $\{\widetilde{Y}_{i,j}\}_{j=1}^{K_U}$. These are used to calculate the corresponding DV potentials by averaging over the reference samples,

$$\bar{g}_{\theta_y}(\widetilde{Y}_i|Y^{i-1}) := \frac{1}{K_U}\sum_{j=1}^{K_U} e^{g_y(\widetilde{Y}_{i,j}|Y^{i-1})},$$

$$\bar{g}_{\theta_{xy}}(\widetilde{Y}_i|Y^{i-1},X^i) := \frac{1}{K_U}\sum_{j=1}^{K_U} e^{g_{xy}(\widetilde{Y}_{i,j}|Y^{i-1},X^i)}.$$

We then use $\bar{g}_{\theta_y}$ and $\bar{g}_{\theta_{xy}}$ instead of the aforementioned exponential terms in (21a) and (21b). We observe empirically that the averaging reduces bias and numerical instability in the estimation process.

V. DINE Optimization Over Continuous Spaces

In this section we present our method for the optimization of the DINE over continuous input distributions. We utilize a generative model, whose objective is to construct a sample D_n that maximizes (22). In what follows, we derive the optimizer, discuss its theoretical properties, describe its implementation, and discuss the joint estimation-optimization procedure.

A. Optimizer Derivation

We consider the optimization $\sup_{\mathsf{P}_X} \mathsf{I}(\mathbb{X} \to \mathbb{Y})$, where $\mathsf{P}_X = \{P_{X_i|X^{i-1}}\}_{i\in\mathbb{N}}$ for feedforward channels and $\mathsf{P}_X = \{P_{X_i|X^{i-1}Y^{i-1}}\}_{i\in\mathbb{N}}$ for channels with feedback. To that end, we propose the NDT, an RNN-based generative model that maps an arbitrary i.i.d. sequence, $U^n \sim P_U^{\otimes n}$, to a sequence of channel inputs. The NDT is given by $h_\phi \in \mathcal{G}_{\mathsf{rnn}}^X = \mathcal{G}_{\mathsf{rnn}}^{(d_x,d_x,k)}$ with parameters $\phi \in \Phi$. Recall that h_ϕ recursively calculates the sequence of channel inputs $X^{\phi,n}$, where $X_0^\phi = 0$ and

$$X_i^\phi = h_\phi(U_i, Z_{i-1}^\phi), \quad i = 1\ldots, n. \tag{25}$$

The sequence $X^{\phi,n}$ is passed through the channel to obtain the corresponding outputs $Y^{\phi,n}$, to arrive at the dataset $D_n^\phi(U^n) := (X^{\phi,n}, Y^{\phi,n})$. For feedforward channels we take $Z_i^\phi = X_i^\phi$, while $Z_i^\phi = (X_i^\phi, Y_i^\phi)$ for channels with feedback. To simplify notation, we denote $D_n^\phi(U^n) = D_n^\phi$ and consider the same distribution P_U throughout. The overall optimization is given by

$$\begin{aligned}\widehat{\mathsf{I}}_{\mathsf{DI}}^\star(U^n) &:= \sup_{h_\phi\in\mathcal{G}_{\mathsf{rnn}}^X} \widehat{\mathsf{I}}_{\mathsf{DI}}(D_n^\phi, h_\phi) \\ &= \sup_{h_\phi\in\mathcal{G}_{\mathsf{rnn}}^X}\left(\sup_{g_{xy}\in\mathcal{G}_{\mathsf{rnn}}^{XY}} \inf_{g_y\in\mathcal{G}_{\mathsf{rnn}}^{Y}} \widehat{\mathsf{I}}_{\mathsf{DI}}(D_n^\phi, h_\phi, g_y, g_{xy})\right).\end{aligned} \tag{26}$$

The DINE objective (20) acts as a loss function for the optimization of h_ϕ, which is executed via gradient-based optimization over ϕ. When the channel is memoryless we focus on MI estimation and optimization, employing the MINE. Consequently, h_ϕ only takes U_i as input and the optimization is carried out over $\mathcal{G}_k^{(d_x,d_x)}$. We next inspect the theoretical properties of the combined estimation-optimization method.

B. Theoretical Guarantees

In this section we provide theoretical analysis of the performance and structure of the proposed method. We first account for the convergence of the joint optimization procedure. Then, restricting attention to MI optimization for memoryless channels, we characterize the optimized NDT structure.

1) Consistency: We show that under appropriate assumptions on the channel transition kernel and input distribution, the optimization in (26) converges to the maximal DI. We begin by describing the class of channel inputs that our result accounts for. We consider the class of stationary processes $\mathbb{X}$, for which there exist an auxiliary stationary process $\mathbb{S}$ over $\mathcal{S} \subseteq \mathbb{R}^{d'}$ and a function $f_{\mathsf{s}} \in \mathcal{C}(\mathcal{X}\times\mathcal{S})$ such that

$$S_i = f_{\mathsf{s}}(H_i, S_{i-1}), \quad i \in \mathbb{N},$$

and $H^{i-1} \leftrightarrow S_{i-1} \leftrightarrow X_i$ forms a Markov chain. We take $H_i = X_i$ for computing the feedforward capacity and $H_i = (X_i, Y_i)$ for the feedback capacity. We call such processes *recursive-state processes* (RSPs) and denote the class of RSPs by $\mathsf{X}_{\mathcal{S}}$. In Section VIII-B, where the consistency of the DINE-NDT method is proved, we show that $\mathsf{X}_{\mathcal{S}}$ can be represented as a special case of the general state-space model [67, Eqn. (3.1)-(3.2)] by constructing a functional reformulation of the aforementioned Markov relation. The structure allows f_{s}

[4]In practice, we sample uniformly from the smallest d-dimensional bounding hypercube of the samples Y^n.

to be a randomized function. To better understand the breadth of the class $\mathsf{X}_{\mathcal{S}}$, we make the following observation.

Lemma 2: The class of stationary Markov processes of finite order is a subset of $\mathsf{X}_{\mathcal{S}}$.

The proof is straightforward by choosing $S_i = [X_{i-(m-1)}, \ldots, X_i]$, for $i \geq m-1$, with Markov order m. When $i < m-1$, the ith to $(m-1)$th entries are zeros.

We next describe the considered class of channels. A unifilar state channel (USC) [68, Section 2] is a channel whose latent state Z_i evolves according to

$$Z_i = f_{\mathsf{z}}(Z_{i-1}, Y_i, X_i), \quad i \in \mathbb{N},$$

for some $f_{\mathsf{z}} \in \mathcal{C}^1(\mathcal{Z} \times \mathcal{X} \times \mathcal{Y})$, where $(X^{i-1}, Y^{i-1}) \leftrightarrow (X_i, Z_{i-1}, Y_i) \leftrightarrow Z_i$ forms a Markov chain. We consider USCs with continuous input and output spaces, whose outputs adhere to the functional relation

$$Y_i = f_{\mathsf{y}}(Z_i, X_i, K_i),$$

for some $f_{\mathsf{y}} \in \mathcal{C}^1(\mathcal{Z} \times \mathcal{X} \times \mathcal{Y})$ and an i.i.d. external process with $K_1 \sim P_K \in \mathcal{P}_{\mathsf{ac}}(\mathbb{R}^{d_K})$ for some $d_K \in \mathbb{N}$. This structure can be viewed as a variation of [69, Equation 7], in which the channel mapping also receives past outputs and the state is unifilar. To bound the effective estimation-optimization error, we impose the following Lipschitz condition on the functions $f_{\mathsf{z}}, f_{\mathsf{y}}$.

Assumption 1: f_{z} and f_{y} are Lipschitz continuous with Lipschitz constants M_y and M_z, respectively, such that $M_y(M_z+1) < 1$.

This assumption can be lifted if we do not permit any recursive relation in the channel structure (for more details, see Sec. VIII-B). In addition, we assume that the DINE RNNs, (g_y, g_{xy}), are Lipschitz with some finite Lipschitz constants M_1 and M_2. We have the following consistency claim.

Theorem 5 (Theorem 3, Restated): Let $\{P_{Y_i|Y^{i-1},X^i}\}_{i\in\mathbb{N}}$ be a continuous unifilar state channel, where $f_{\mathsf{y}}, f_{\mathsf{z}}$ satisfy Assumption 1. Then, $\mathbb{P}$-*a.s.* for every $\epsilon > 0$, there exist $N \in \mathbb{N}$ such that for every $n > N$ and $U^n \sim P_U^{\otimes n}$ we have

$$\left|\underline{\mathsf{C}}_s - \widehat{\mathsf{I}}^{\star}_{\mathsf{DI}}(U^n)\right| \leq \epsilon, \tag{27}$$

where $\widehat{\mathsf{I}}^{\star}_{\mathsf{DI}}(U^n) = \sup_{h_\phi \in \mathcal{G}_{\mathsf{rnn}}^{(d_x,d_x)}} \widehat{\mathsf{I}}_{\mathsf{DI}}(D_n^\phi, h_\phi)$.

The proof, which is given in Section VIII-B, is divided into two steps by dividing the consistency error into two pieces: (i) the error induced by replacing the optimal input distribution with the proxy coming from the NDT, and (ii) the error induced by estimating the ground-truth DI rate. The latter is bounded due to the DINE consistency, while the former uses the functional representation lemma to bound the error induced by using a dataset calculated by the NDT and control its propagation through the channel and DINE models. The full proof also generalizes the statement to the case of channels with feedback.

Remark 4 (Feasible Channels): In general, $\underline{\mathsf{C}}_s$ lower bounds the capacity of a given channel with memory, and the characterization of capacity-achieving input distributions of arbitrary stationary channels with continuous input and output spaces is currently an open problem. However, when the channel is Gaussian and the channel has a linear state-space model, the capacity achieving distribution can be reformulated as an RSP [50], [70].

2) Optimized NDT Structure: We now restrict attention to memoryless channels, thereby focusing on MI estimation and optimization. We employ MINE as the MI estimator and discuss the structure of the optimized NDT. To that end, we utilize the Rosenblatt transform [71], [72], also known as the Knöthe-Rosenblatt rearrangement. The Rosenblatt transform is a multivariate generalization of the CDF based on triangular transformations (cf. [73]). Consider a d-dimensional random vector $X := (X_1, \ldots, X_d) \sim P_X \in \mathcal{P}(\mathcal{X})$, where $\mathcal{X} \subseteq \mathbb{R}^d$, and define the associated vector-valued function $T_X : \mathcal{X} \to [0,1]^d$ by

$$\begin{aligned}
\big[T_X(x)\big]_1 &= \mathbb{P}(X_1 \leq x_1) \\
\big[T_X(x)\big]_i &= \mathbb{P}\Big(X_i \leq x_i \,\Big|\, \big[T_X(X)\big]_{i-1} = \big[T_X(x)\big]_{i-1}, \\
&\qquad \ldots, \big[T_X(X)\big]_1 = \big[T_X(x)\big]_1\Big), \\
&\qquad i = 2, \ldots d.
\end{aligned} \tag{28}$$

In words, for $x \in \mathcal{X}$, each entry $\big[T_X(x)\big]_i$ is given by the conditional distribution function of X_i at x_i given the values of the function in the preceding entries, i.e., $\big[T_X(x)\big]_1, \ldots, \big[T_X(x)\big]_{i-1}$. The mapping T_X^{-1} is the *inverse* Rosenblatt transform. We now have the following proposition.

Lemma 3: Let $X \sim P_X \in \mathcal{P}_{\mathsf{ac}}(\mathcal{X})$ with $\mathcal{X} \subseteq \mathbb{R}^{d_x}$, and consider the map $T_X : \mathcal{X} \to [0,1]^{d_x}$ defined above. Then:

(i) T_X is Borel measurable and the random variable $T_X(X)$ is uniformly distributed over $[0,1]^{d_x}$.

(ii) T_X is a bijection, and for $U \sim \mathsf{Unif}\big([0,1]^{d_x}\big)$, we have $T_X^{-1}(U) \overset{d}{=} X$.

The results of Lemma 3 can be attributed to several past works. Specifically, Part (i) was derived in [71], while Part (ii) follows from the results in [74] and [75]. We subsequently use this lemma to characterize the optimal NDT. The reader is referred to [71], [73], [76], [77], and [78] for further discussion and useful properties of the function T_X.

Lemma 3 provides a distribution-equivalent representation of continuous random variables as functions of uniformly distributed variables [5].

We leverage this fact to characterize to MINE-maximizing NDT. Let $\mathcal{P}_p(\mathbb{R}^d)$ be the class of Borel probability measures on $\mathbb{R}^d$ with finite p-th moment, i.e., $\int \|x\|^p \, \mathrm{d}\mu(x) < \infty$. For a given transition kernel $P_{Y|X}$, let C denote the capacity of the corresponding memoryless channel bound to a second moment input constraint. Denote the capacity-achieving distribution $P_{X^\star} := \operatorname{argmax}_{P_X \in \mathcal{P}_2(\mathcal{X})} \mathsf{I}(X;Y)$, let $X^\star \sim P_{X^\star}$, and consider its associated mapping $T_{X^\star}$. We quantify the distance between the NDT-induced probability distribution and $P_{X^\star}$ using the 2-Wasserstein distance. The p-Wasserstein distance between $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$ is given by

$$\mathsf{W}_p(\mu,\nu) := \left[\inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x-y\|^p \, \mathrm{d}\pi(x,y)\right]^{1/p},$$

[5] As a consequence of Lemma 3, we can construct a transformation between any two absolutely continuous random variables W and X provided they have the same dimension, by utilizing the composition $T_X^{-1} \circ T_W : \mathcal{W} \mapsto \mathcal{X}$ [79].

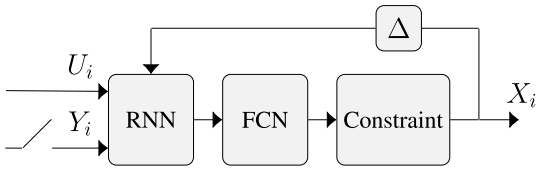

Fig. 3. The NDT. The noise and past channel output (if feedback is present) are fed into an RNN. The last layer imposes a constraint of our choice.

where $\Pi(\mu,\nu)$ is the set of all couplings of μ and ν. We propose the following theorem.

Theorem 6 (Optimal NDT): Fix $\epsilon > 0$ and let P_U be the uniform distribution over $\mathcal{U}$. Let $P_{Y|X}$ with a bounded and continuous PDF $p_{Y|X}$ such that it induces a finite second moment on the channel output for any second moment-bounded input distribution. Then, there exist $h_\phi \in \mathcal{G}_{\mathsf{nn}}^{(d_x,d_x)}$, such that

$$\mathsf{W}_2\left(h_{\phi\sharp}P_U, P_{X^\star}\right) \leq \epsilon, \tag{29}$$

where $h_{\phi\sharp}P_U$ is the pushforward measure of P_U by h_ϕ. Moreover, $\mathbb{P}$-*a.s.* for any $\epsilon > 0$ there exist $n_0 \in \mathbb{N}$ such that for any $n > n_0$ and $U^n \sim P_U^{\otimes n}$, we have

$$\left| \mathsf{C} - \widehat{\mathsf{I}}_{\mathsf{MI}}\left(D_n^{\phi_k}\right) \right| \leq \epsilon, \tag{30}$$

where, $D_n^\phi := \{(h_\phi(U_i), Y_i)\}_{i=1}^n$.

The proof is given in Section VIII-C. It utilizes Gaussian smoothing of probability distributions, which both helps us account for optimal input distributions that are not necessarily absolutely continuous, and induces additional structure and regularity into the framework. We first leverage the universal approximation of NNs and Gaussian smoothing to bound the W_2 error induced by the NN approximation of the optimal input distribution. Having that, we decompose the capacity estimation error into several terms, which are bounded using Wasserstein continuity of KL-divergence [80], weak continuity of differential entropy [81, Theorem 1], and the MINE consistency [24, Theorem 2]. Theorem 6 guarantees the existence of an NDT model that approximates the capacity achieving distribution (under the p-Wasserstein distance), which, in turn, yields a consistent MINE-based proxy of capacity. We therefore conjecture that the MINE-maximizing NDT is in fact an approximator of T_X^{-1}. We empirically validate this conjecture for the AWGN channel in the next section.

Remark 5 (Lower Bounding Channel Capacity): When the MINE is optimized, the NDT does not impede the DV-induced lower bound (Lemma 1). Consequently, for any $h_\phi \in \mathcal{G}_{\mathsf{nn}}^{(d_x,d_x)}$, the corresponding MINE output lower bounds the channel capacity. This property will serve us in Section VI-A.2 to propose a bound on the capacity of the peak-power constrained AWGN.

C. Implementation

The NDT is implemented using an LSTM stacked with 2 FC layers. Chanel input constraints, such as average or peak-power constraints, can be imposed on the NDT outputs, as long as these can be realized with a differential function of the ϕ.

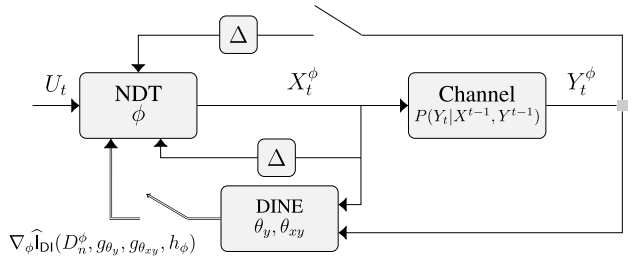

Fig. 4. The complete system for optimization over continuous spaces. On each step gradients are passed to a predetermined model, while the other one's parameters are fixed.

The NDT model is shown in Figure 3. The overall optimization over the NDT and DINE takes the form

$$\sup_{\phi\in\Phi,\theta_{xy}\in\Theta_{xy}} \inf_{\theta_y\in\Theta_y} \widehat{\mathsf{I}}_{\mathsf{DI}}(D_n^\phi, g_{\theta_y}, g_{\theta_{xy}}, h_\phi).$$

In every iteration, we draw a noise batch (U^m), from which $B_m^\phi = (X^{\phi,m}, Y^{\phi,m})B$ is computed. The batch B_m is processed by g_{θ_y} and $g_{\theta_{xy}}$, the loss $\widehat{\mathsf{I}}(B_m^\phi, g_{\theta_y}, g_{\theta_{xy}}, h_\phi)$ is calculated, and gradients are propagated to update the models weights. Figure 4 illustrates the complete architecture.

The training adheres to an alternating optimization procedure. Namely, we iterate between updating (θ_y, θ_{xy}) and ϕ, each time keeping the other parameters fixed. After the training is done, we perform a long Monte-Carlo (MC) evaluation to obtain an estimate of (20). The procedure is summarized in Algorithm 2 and its implementation is available on GitHub. This alternation between two models sharing a common loss is found in other fields, such as generative adversarial networks [82] and actor-critic algorithms [83]. We stress that the proposed optimization scheme can be applied to any NN-based estimator of information measures, inasmuch as it is differentiable w.r.t. the NDT outputs.

VI. Empirical Capacity Estimation Results

We demonstrate the performance of Algorithm 2 for continuous channel capacity estimation, considering both feedforward and feedback scenarios for several channel models. The numerical results are compared with the available theoretical solution/bounds to verify the effectiveness of the proposed method. The simulations are implemented in TensorFlow [84]. The DINE is implemented using a modified LSTM and two fully-connected layers with 50, 100 and 50 neurons, respectively. The NDT is implemented with an LSTM and two fully connected layers, each with 100 neurons, stacked with an output layer with d_x neurons.

We note that the term calculated by the DINE-NDT method differs from the general capacity expression in the following way. In (26), we take the supremum over the estimated DI *rate*, i.e., the limit is taken before the supremum. In contrast, the general capacity expression (4) considers the opposite order of limit and optimization. This order is known to be interchangeable for stationary Gaussian channels [85], and generally seems to have a minimal effect on the accuracy of the numerical results for the considered examples. We also stress that all methods with which we compare the DINE-NDT

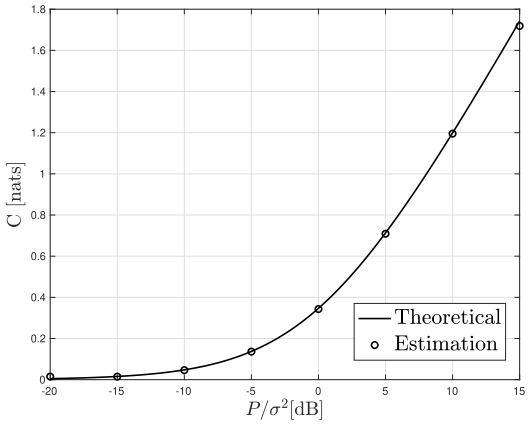

Fig. 5. AWGN with average power constraint.

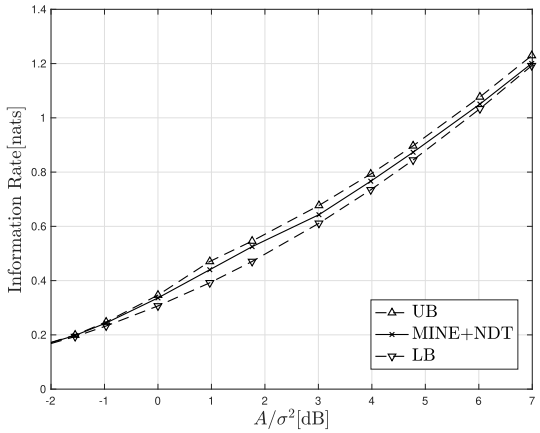

Fig. 6. AWGN with peak power constraint. Estimate compared with known capacity upper and lower bounds from [46] and [48], respectively.

method assume full knowledge of the channel model, which our approach does not require.

A. AWGN Channel

1) Power Constraint: We consider the AWGN channel

$$Y_i = X_i + Z_i, \quad i \in \mathbb{N}, \tag{31}$$

where $Z_i \sim \mathcal{N}\left(0, \sigma^2\right)$ are i.i.d. and X_i is the channel input sequence bound to the average power constraint $\mathbb{E}\left[X_i^2\right] \leq P$. The capacity of this channel is given by $\mathsf{C} = \frac{1}{2}\log\left(1 + \frac{P}{\sigma^2}\right)$ [56]. We set $\sigma^2 = 1$ and estimate the capacity via the optimized MINE for a range of P values. The numerical results are compared to the analytic solution in Figure 5, where a clear correspondence is seen.

2) Peak-Power Constraint: We consider the AWGN channel with a peak power constraint $|X| < A$, for some $A > 0$. The capacity of this channel is unknown, but upper and lower bounds on it are available in the literature [46], [48]. In Figure 6 we present a comparison of the capacity estimate obtained from our Algorithm 2 (with MINE instead of DINE) and the aforementioned bounds. Evidently, the estimate falls within the theoretical bounds. As MINE lower bounds the channel capacity for any choice of h_ϕ (cf., Remark 5), our estimate also provides new and tighter lower bounds on the capacity of this channel.

3) Optimized NDT Structure: Considering the average power constrained AWGN, we check two characteristics of the MINE-maximizing NDT. First, we empirically validate Theorem 6 by comparing the optimized NDT with $T_{X^\star}^{-1}$, where $X^\star \sim \mathcal{N}(0, P)$ is the capacity-achieving input. The correspondence is shown in Figure 7. Second, in Figure 8 we examine histograms to further verify that the optimized NDT maps the input samples U^n into samples of the capacity-achieving Gaussian distribution.

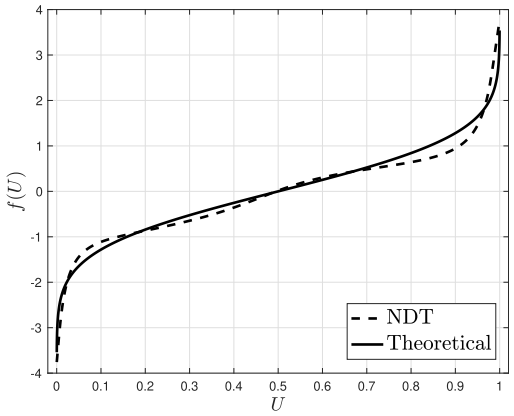

Fig. 7. Optimized NDT structure comparison with F_X^{-1} for AWGN with average power constraint.

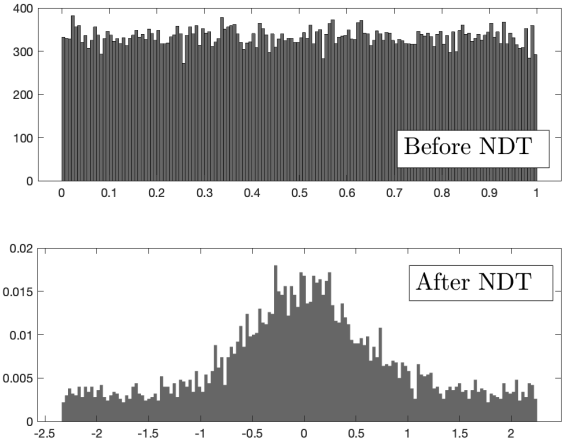

Fig. 8. NDT input vs. output histograms for AWGN with average power constraint.

B. Gaussian MA(1) Channel

We consider the MA-AWGN channel of order 1:

$$\begin{aligned} Z_i &= \alpha N_{i-1} + N_i \\ Y_i &= X_i + Z_i, \end{aligned} \tag{32}$$

where $N_i \sim \mathcal{N}(0,1)$ are i.i.d., X_i is the channel input sequence bound to the average power constraint $\mathbb{E}\left[X_i^2\right] \leq P$, and Y_i is the channel output. We consider both feedforward and feedback cases. The feedforward capacity can be calculated via the water-filling algorithm [56]. When feedback is present, we consider the capacity characterization from [49] as $-\log(x_0)$, where x_0 is a solution to a 4th order polynomial equation. In Figure 9, we compare our DINE-based capacity

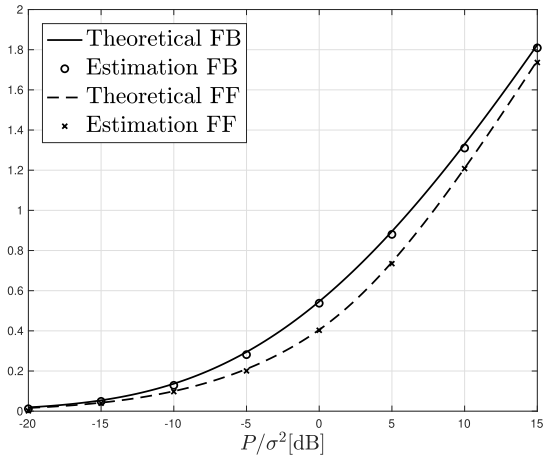

Fig. 9. MA(1)-AGN estimated capacity comparison with analytical solution.

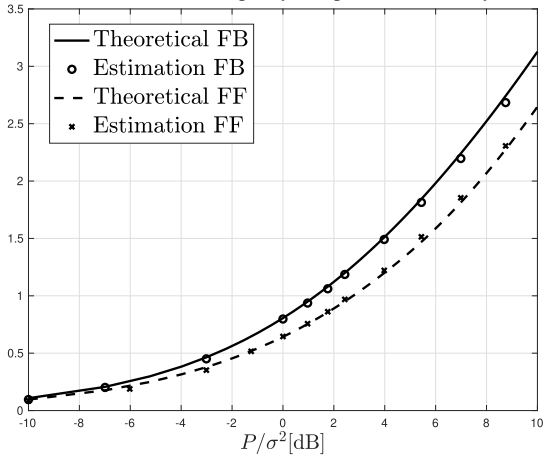

Fig. 10. MIMO AR(1)-AGN estimated capacity comparison with analytical solution.

estimator with the above solutions, again revealing clear correspondence.

C. MIMO Gaussian AR(1) Channel

The AR(1) Gaussian channel is given by

$$\begin{aligned} Z_i &= \alpha Z_{i-1} + N_i \\ Y_i &= X_i + Z_i, \end{aligned} \tag{33}$$

where $X_i \in \mathbb{R}^4$ and $N_i \sim \mathcal{N}(0, \mathrm{I}_4)$ where I_4 is the 4-dimensional identity matrix. We consider the power constraint $\mathrm{tr}(\mathrm{K}_{X_i}) \leq P$ for some $P \in \mathbb{R}_{\geq 0}$, where K_X is the covariance matrix of X. The feedforward capacity of (33) is obtained by the water filling algorithm, considering both the spatial and frequency domains. For the feedback capacity, the authors of [50] recently developed a method for calculating the capacity of a general class of MIMO Gaussian channels with memory through sequential convex optimization. This class subsumes the MIMO AR(1) channel as a special case. Figure 10 compares the performance of Algorithm 2 with the above methods. The convergence of the algorithm is shown in in Figure 11, presenting a long evaluation over 10^5 samples, taken every 20 training iterations. It is evident that our method converges in a relatively small number of iterations and the ground truth is attained in all considered cases.

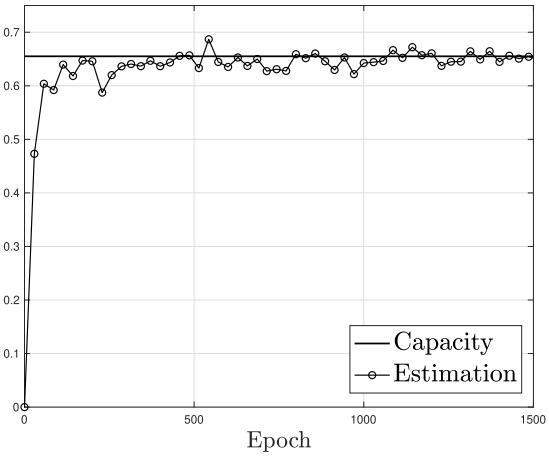

Fig. 11. Algorithm convergence for MIMO AR(1)-AGN with $P/\sigma^2 = 1$.

VII. Concluding Remarks and Future Work

This work proposed a new neural estimation-optimization framework of the DI rate between two jointly stationary and ergodic stochastic processes. Drawing upon recent neural estimation techniques and modifying the LSTM architecture, we developed the DINE, proved its consistency, and described its implementation. Then, we utilized an auxiliary deep generative model for the input process to obtain a provably consistent joint estimation-optimization scheme of DI rate. The method enables estimating channel capacity when the channel model is unknown (but can be sampled) or when the optimization objective is not tractable, accounting for both feedback and feedforward scenarios. We provided an empirical study that validated our theory and demonstrated the accuracy of the proposed framework for capacity estimation of various channel examples. The capacity estimates demonstrated significant correspondence with known theoretical solutions and/or bounds, and the learned input model was shown to approximate capacity-achieving input distributions.

Our method enables consistent estimation of channel capacity without the typically imposed model assumptions. However, the obtained estimate generally does not lower or upper bound the true capacity value. In future work, we plan to explore modified neural estimation techniques that would give rise to such theoretical bounds. Another appealing avenue is utilizing the learned NDT-based input distribution, or an appropriate adaptation thereof, to obtain explicit capacity-achieving coding schemes. We also plan to extend our method to multiuser channels with arbitrary input and output spaces, targeting a unified and scalable framework of channel capacity estimation. Moreover, we will look to apply the proposed scheme to other time-series domains, such as control, computer vision, speech recognition, and reinforcement learning.

VIII. Proofs

A. Proof of Theorem 2

With some abuse of notation, let $\{(X_i, Y_i)\}_{i\in\mathbb{Z}}$ be the two-sided extension of the considered processes, and $\mathbb{P}$ be

the underlying stationary ergodic measure over $\sigma(\mathbb{X},\mathbb{Y})$. An n-coordinate projection of $\mathbb{P}$ is denoted by $P_{X^nY^n} := \mathbb{P}\big|_{\sigma(X^n,Y^n)}$, where $\sigma(X^n,Y^n)$ is the σ-algebra generated by (X^n,Y^n). With this notation, $D_n = (X^n,Y^n) \sim P_{X^nY^n}$. Lastly, let $\widetilde{Y} \sim \mathsf{Unif}(\mathcal{Y})$ (recall that $\mathcal{Y} \subset \mathbb{R}^{d_y}$ is compact) be independent of $\{(X_i,Y_i)\}_{i\in\mathbb{Z}}$ and denote its distribution by $P_{\widetilde{Y}}$. We divide the proof into three steps: variational representation, estimation from samples, and functional approximation.

Step 1: Representation of DI rate. We first write the DI rate as the limit of certain KL divergence terms. To do so, we use to following lemma:

Lemma 4 (DI rate vs. $\mathsf{D_{KL}}$): Let

$$\mathsf{D}^{\infty}_{Y\|X} := \mathsf{D}_{\mathsf{KL}}\left(P_{Y^0_{-\infty}\|X^0_{-\infty}}\middle\|P_{Y^{-1}_{-\infty}\|X^{-1}_{-\infty}}\otimes P_{\widetilde{Y}}\middle|P_{X^0_{-\infty}\|Y^{-1}_{-\infty}}\right)$$

$$\mathsf{D}^{\infty}_{Y} := \mathsf{D}_{\mathsf{KL}}\left(P_{Y^0_{-\infty}}\middle\|P_{Y^{-1}_{-\infty}}\otimes P_{\widetilde{Y}}\right).$$

Then we have

$$\mathsf{I}(\mathbb{X}\to\mathbb{Y}) = \mathsf{D}^{\infty}_{Y\|X} - \mathsf{D}^{\infty}_{Y}. \tag{34}$$

Lemma 4 is proven in Appendix IX-A. The proof uses the stationarity of the considered processes and the monotone convergence theorem for the KL divergence (cf., e.g., [86, Corollary 3.2]). We henceforth focus on estimating D^{∞}_{Y} and $\mathsf{D}^{\infty}_{Y\|X}$. Using the DV representation (Theorem 1), we have

$$\mathsf{D}^{\infty}_{Y} = \sup_{f_y:\Omega_{\mathcal{Y}}\to\mathbb{R}} \mathbb{E}\left[f_y\big(Y^0_{-\infty}\big)\right] - \log\mathbb{E}\left[e^{f_y\left(Y^{-1}_{-\infty},\widetilde{Y}\right)}\right], \tag{35a}$$

where $\Omega_{\mathcal{Y}} = \mathcal{Y}^0_{-\infty}$. For $\mathsf{D}^{\infty}_{Y\|X}$, we use the KL divergence chain rule to write

$$\mathsf{D}^{\infty}_{Y\|X} = \mathsf{D}_{\mathsf{KL}}\Big(P_{X^0_{-\infty}\|Y^{-1}_{-\infty}}P_{Y^0_{-\infty}\|X^0_{-\infty}}\Big\|P_{X^0_{-\infty}\|Y^{-1}_{-\infty}}P_{Y^{-1}_{-\infty}\|X^{-1}_{-\infty}}\otimes P_{\widetilde{Y}}\Big)$$

and via the DV theorem obtain

$$\mathsf{D}^{\infty}_{Y\|X} = \sup_{f_{xy}:\Omega_{\mathcal{X}\times\mathcal{Y}}\to\mathbb{R}} \mathbb{E}\left[f_{xy}\big(X^0_{-\infty},Y^0_{-\infty}\big)\right] - \log\mathbb{E}\left[e^{f_2\left(X^0_{-\infty},Y^{-1}_{-\infty},\widetilde{Y}\right)}\right], \tag{35b}$$

where $\Omega_{\mathcal{X}\times\mathcal{Y}} = \mathcal{Y}^0_{-\infty}\times\mathcal{X}^0_{-\infty}$.

We now provide a full treatment of (35a). Afterwards, we refer back to (35b) and explain how its analysis reduces to that of (35a), without repeating the argument.

Step 2: Estimation. The supremum in (35a) is achieved by

$$f^{\star}_{y,\infty} := \log\left(\frac{\mathrm{d}P_{Y^0_{-\infty}}}{\mathrm{d}(P_{Y^{-1}_{-\infty}}\otimes P_{\widetilde{Y}})}\right) \overset{(a)}{=} \log p_{Y_0|Y^{-1}_{-\infty}} - \log p_{\widetilde{Y}}, \tag{36}$$

where (a) holds because $P_{Y^0_{-\infty}} \ll P_{Y^{-1}_{-\infty}}\otimes P_{\widetilde{Y}}$ and both measures have Lebesgue densities. Since $\widetilde{Y}$ is uniform, $p_{\widetilde{Y}}$ is a constant; denote $c_{\mathcal{Y}} := \log\big(p_{\widetilde{Y}}(y)\big)$, for any $y\in\mathcal{Y}$. We next show that the expectations in (35a) can be estimated with empirical means. Namely, for any $\epsilon>0$ and sufficiently large n, we have $\mathbb{P}$-*a.s.* that

$$\left|\mathbb{E}\Big[f^{\star}_{y,\infty}\big(Y^0_{-\infty}\big)\Big] - \frac{1}{n}\sum_{i=0}^{n-1} f^{\star}_{y,i}\left(Y^0_{-i}\right)\right| < \frac{\epsilon}{8} \tag{37a}$$

$$\left|\log\left(\mathbb{E}\Big[e^{f^{\star}_{y,\infty}\left(Y^{-1}_{-\infty},\widetilde{Y}\right)}\Big]\right) - \log\left(\frac{1}{n}\sum_{i=0}^{n-1} e^{f^{\star}_{y,i}\left(Y^{-1}_{-i},\widetilde{Y}\right)}\right)\right| < \frac{\epsilon}{8}, \tag{37b}$$

where $\{f^{\star}_{y,i}\}_{i\in\mathbb{N}}$ is the sequence of supremum achieving elements of $\left\{\mathsf{D}_{\mathsf{KL}}\left(P_{Y^0_{-i}}\middle\|P_{Y^{-1}_{-i}}\otimes P_{\widetilde{Y}}\right)\right\}_{i\in\mathbb{N}}$, with the $\mathbb{P}$-*a.s.* limit $\lim_{i\to\infty} f^{\star}_{y,i} = f^{\star}_{y,\infty}$. To simplify notation we denote the following empirical means over n samples as

$$\mathbb{E}_n[f^{\star}_y(Y^0_{-(n-1)})] := \frac{1}{n}\sum_{i=0}^{n-1} f^{\star}_{y,i}\left(Y^0_{-i}\right) \tag{38}$$

$$\mathbb{E}_n\left[e^{f^{\star}_y\left(\widetilde{Y},Y^{-1}_{-(n-1)}\right)}\right] := \frac{1}{n}\sum_{i=0}^{n-1} e^{f^{\star}_{y,i}\left(Y^{-1}_{-i},\widetilde{Y}\right)}, \tag{39}$$

and invoke the generalized form of the asymptotic equipartition (AEP) theorem [65], as stated next.

Theorem 7 (Generalized AEP): Suppose $\mathbb{M}$ is a v^{th} order Markov measure with a stationary transition kernel $\kappa(\mathrm{d}X_v|X_0^{v-1})$, and the finite-dimensional marginals of $\mathbb{M}$ are absolutely continuous w.r.t. the corresponding marginals of a stationary measure $\mathbb{P}$, i.e., if $\mathbb{P}$ is ergodic, $\mathbb{E}$ is the expectation w.r.t. $\mathbb{P}$ and $p_{X^0_{-(n-1)}} := \frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{M}}\Big|_{\sigma\left(X^0_{-(n-1)}\right)}$, then

$$\begin{aligned}&\frac{1}{n}\log\left(p_{X^0_{-(n-1)}}(X_0,\ldots,X_{-(n-1)})\right)\\ &= \frac{1}{n}\sum_{i=0}^{n-1}\log\left(p_{X_0|X^{-1}_{-i}}\left(X_0|X^{-1}_{-i}\right)\right)\\ &\xrightarrow[n\to\infty]{} \mathbb{E}\left[\log p_{X_0|X^{-1}_{-\infty}}(X_0|X^{-1}_{-\infty})\right], \qquad \mathbb{P}\text{-}a.s.\end{aligned} \tag{40}$$

By Theorem 7, we obtain

$$\lim_{n\to\infty}\mathbb{E}_n[f^{\star}_y(Y^0_{-(n-1)})] = \mathbb{E}\Big[f^{\star}_{y,\infty}\big(Y^0_{-\infty}\big)\Big], \qquad \mathbb{P}\text{-}a.s., \tag{41}$$

where $f^{\star}_{y,i} := \log p_{Y_0|Y^{-1}_{-i}} - c_{\mathcal{Y}}$.

Some additional work is needed to justify (37b). First, by [87, Proposition 2.6], we have that the sequence $\left(e^{f^{\star}_{y,n}+c_{\mathcal{Y}}}, \sigma\left(Y^{-1}_{-(n-1)},\widetilde{Y}\right)\right) = \left(p_{Y_0|Y^{-1}_{-(n-1)}}, \sigma\left(Y^{-1}_{-(n-1)},\widetilde{Y}\right)\right)$ is a positive supermartingale converging a.s. to $p_{Y_0|Y^{-1}_{-\infty}}$, which equals $e^{f^{\star}_{y,\infty}+c_{\mathcal{Y}}}$ with $f^{\star}_{y,\infty}$ given in (36). We now apply a generalization of Birkhoff's ergodic theorem (due to Breiman [42, Theorem 1]), as stated next.

Theorem 8 (The Generalized Birkhoff Theorem): Let T be a metrically transitive $1-1$ measure preserving transformation[6] of the probability space $(\Omega,\mathcal{F},\mathbb{P})$ onto itself. Let $g_0(\omega), g_y(\omega),\ldots$ be a sequence of measurable functions on Ω converging a.s. to the function $g(\omega)$ such that $\mathbb{E}[\sup_k|g_k|] \le$

[6]This translates into the condition $\mathbb{P}(A) = \mathbb{P}(T^{-1}(A))$ for any $A\in\mathcal{F}$. We consider the time shift transformation.

∞. Then,

$$\frac{1}{n}\sum_{k=1}^{n} g_k(T^k\omega) \xrightarrow{n\to\infty} \mathbb{E}[g], \qquad \mathbb{P}\text{-}a.s. \tag{42}$$

Applying Theorem 8 together with the continuous mapping theorem from [88, Corollary 2], we conclude that $\mathbb{P}$-*a.s.*

$$\lim_{n\to\infty} \log\left(\frac{1}{n}\sum_{i=0}^{n-1} e^{f^\star_{y,i}\left(Y_{-i}^{-1}\widetilde{Y}_0\right)}\right) = \log\left(\mathbb{E}\left[e^{f^\star_{y,\infty}\left(Y_{-\infty}^{-1},\widetilde{Y}\right)}\right]\right). \tag{43}$$

This, in turn, implies (37b) for a large enough n.

Step 3: Approximation. The last step is to approximate the functional space with the space of RNNs. namely, we define

$$\widehat{\mathsf{D}}_Y(D_n) := \sup_{g_y\in\mathcal{G}^Y_{\mathsf{rnn}}} \frac{1}{n}\sum_{i=0}^{n-1} g_y(Y_{-i}^0) - \log\left(\frac{1}{n}\sum_{i=0}^{n-1} e^{g_y(Y_{-i}^{-1},\widetilde{Y}_0)}\right), \tag{44}$$

and we want to show that for a given $\epsilon > 0$, we know that

$$\left|\widehat{\mathsf{D}}_Y(D_n) - \mathsf{D}_Y^{(\infty)}\right| \le \frac{\epsilon}{2}.$$

By Theorem 1, we have

$$\mathbb{E}\left[f^\star_{y,\infty}(Y_{-\infty}^0)\right] = \mathsf{D}_Y^{(\infty)}, \quad \mathbb{E}\left[f^\star_{y,\infty}(Y_{-\infty}^{-1},\widetilde{Y})\right] = 1.$$

We therefore bound the expression $\left|\widehat{\mathsf{D}}_Y(D_n) - \mathbb{E}\left[f^\star_y(Y_{-\infty}^0)\right]\right|$. First, by the identity $\log(x) \le x - 1$ for every $x \in \mathbb{R}_{\ge 0}$ we have

$$\begin{aligned}
&\left|\widehat{\mathsf{D}}_Y(D_n) - \mathbb{E}\left[f^\star_y(Y_{-\infty}^0)\right]\right| \\
&= \left|\sup_{g_y\in\mathcal{G}^Y_{\mathsf{rnn}}} \frac{1}{n}\sum_{i=0}^{n-1} g_y(Y_{-i}^0) - \log\left(\frac{1}{n}\sum_{i=0}^{n-1} e^{g_y(\widetilde{Y},Y_{-i}^{-1})}\right) - \mathbb{E}\left[f^\star_y(Y_{-\infty}^0)\right]\right| \\
&\le \left|\sup_{g_y\in\mathcal{G}^Y_{\mathsf{rnn}}} \frac{1}{n}\sum_{i=0}^{n-1} g_y(Y_{-i}^0) - \left(\frac{1}{n}\sum_{i=0}^{n-1} e^{g_y(\widetilde{Y},Y_{-i}^{-1})}\right) + 1 - \mathbb{E}\left[f^\star_y(Y_{-\infty}^0)\right]\right| \\
&\le \left|\sup_{g_y\in\mathcal{G}^Y_{\mathsf{rnn}}} \frac{1}{n}\sum_{i=0}^{n-1} g_y(Y_{-i}^0) - \left(\frac{1}{n}\sum_{i=0}^{n-1} e^{g_y(\widetilde{Y},Y_{-i}^{-1})}\right) + \mathbb{E}\left[e^{f^\star_{y,\infty}(\widetilde{Y},Y_{-\infty}^{-1})}\right] - \mathbb{E}\left[f^\star_y(Y_{-\infty}^0)\right]\right|.
\end{aligned} \tag{45}$$

Due to (37) and the a.s. convergence of $\{e^{f^\star_{y,n}}\}_{n\in\mathbb{N}}$, there exists an integer $N \in \mathbb{N}$ such that for every $n > N$

$$\begin{aligned}
\left|\mathbb{E}_n\left[f^\star_y(Y_{-(n-1)}^0)\right] - \mathbb{E}\left[f^\star_{y,\infty}(Y_{-\infty}^0)\right]\right| &\le \frac{\epsilon}{8}, \\
\left|\mathbb{E}_n\left[e^{f^\star_y(\widetilde{Y},Y_{-(n-1)}^{-1})}\right] - \mathbb{E}\left[e^{f^\star_{y,\infty}(\widetilde{Y},Y_{-\infty}^{-1})}\right]\right| &\le \frac{\epsilon}{8}
\end{aligned} \tag{46}$$

Plugging (46) into (45),we have

$$\begin{aligned}
&\left|\widehat{\mathsf{D}}_Y(D_n) - \mathsf{D}_Y^{\infty}\right| \\
&\le \left|\mathbb{E}_n\left[e^{f^\star_y(\widetilde{Y},Y_{-(n-1)}^{-1})}\right] - \mathbb{E}_n\left[f^\star_y(Y_{-(n-1)}^0)\right] - \sup_{g_y\in\mathcal{G}^Y_{\mathsf{rnn}}}\left\{\frac{1}{n}\sum_{i=0}^{n-1} g_y(Y_{-i}^0) - \left(\frac{1}{n}\sum_{i=0}^{n-1} e^{g_y(Y_{-i}^{-1},\widetilde{Y}_0)}\right)\right\}\right| + \frac{\epsilon}{4}.
\end{aligned}$$

By assumption, $\{f^\star_{y,i}\}_{i\in\mathbb{N}}$ is a sequence of functions converging a.s. to a function $f^\star_{y,\infty}$, uniformly bounded by some $M \in \mathbb{R}_{\ge 0}$. Since the exponent function is Lipschitz continuous with Lipschitz constant e^M on the interval $(-\infty, M]$, we obtain

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} e^{f^\star_{y,i}(\widetilde{Y},Y_{-i}^{-1})} - e^{g_y(\widetilde{Y},Y_{-i}^{-1})} \\
&\qquad \le e^M \frac{1}{n}\sum_{i=1}^{n}\left|f^\star_{y,i}(\widetilde{Y},Y_{-i}^{-1}) - g_y(\widetilde{Y},Y_{-i}^{-1})\right|.
\end{aligned}$$

We conclude this stage by applying the universal approximation theorem for RNNs [60]. To that end, we show that the sequence of supremum-achieving DV potentials are a dynamic system.

Definition 5 (Dynamic System): Let $d_i, d_o, T \in \mathbb{N}$, $\mathcal{Z} \subseteq \mathbb{R}^{d_o}$ and $\mathcal{U} \subseteq \mathbb{R}^{d_i}$ be open sets, $\mathcal{D}_z \subseteq \mathcal{Z}$ be a compact set and $f : \mathcal{Z}\times\mathcal{U} \mapsto \mathcal{Z}$ be a continuous vector-valued function. Then, the system $\mathsf{Z}^{(d_i,d_o)} := \{z_t\}_{t=1}^T$ defined by

$$z_{t+1} = f(z_t, u_t) \tag{47}$$

for $t \in \{1,\dots,T\}$ with some initial value $z_0 \in \mathcal{D}_z$ is a dynamic system.

We propose the following lemma, which provides an approximation of the process $\mathbb{Y}$ that adheres to the dynamic system structure

Lemma 5 (Dynamic System Representation): Let $\mathbb{Y}$ be a stationary and regular stochastic process. Then, for any $\epsilon > 0$, $T \in \mathbb{N}$ and $y^n \in \mathcal{Y}^n$ there exist a process $\widetilde{\mathbb{Y}}(\epsilon, n)$ over $\mathcal{Y}^\infty$ such that

$$\sup_{t=1,\dots,T} |\log p_{Y_t|Y^{t-1}}(y_t|y^{t-1}) - \log p_{\widetilde{Y}_t|\widetilde{Y}^{t-1}}(y_t|y^{t-1})| \le \epsilon$$

and the mapping $y^n \to (\log p_{\widetilde{Y}_t|\widetilde{Y}^{t-1}}(y_t|y^{t-1}))_{t=1}^n$ is a dynamical system.

The proof of Lemma 5 is in Appendix IX-B, and it follows a Wold-like decomposition of strictly stationary processes, which is then shown to adhere the desired structure. Let $\widetilde{\mathbb{Y}}$ be the approximating process. Thanks to the universal approximation theorem for RNNs [60, Theorem 2], we can approximate the corresponding dynamical system by elements of the class $\mathcal{G}^Y_{\mathsf{rnn}}$ to arbitrary precision.

Theorem 9 (Universal Approximation for RNNs): Let $\epsilon > 0$, $T \in \mathbb{N}$, $\mathcal{U} \subset \mathbb{R}^{d_x}$ be an open set and $\mathsf{Z}^{(d_i,d_o)}$ be a dynamic system as in Definition 5. There exist $k \in \mathbb{N}$ and a k-neuron RNN $g \in \mathcal{G}^{(d_i,d_o,k)}_{\mathsf{rnn}}$ (as in Definition 4), such that for any sequence of inputs $\{u_t\}_{t=1}^T \in \mathcal{U}^T$, we have

$$\max_{0\le t\le T} \|Z_t - g(u^t)\|_1 \le \epsilon, \tag{48}$$

where $g(u^t)$ denotes the output of the RNN g determined by the sequence u^t.

For given ϵ, M, and $T = n$, denote by $g_y^\star \in \mathcal{G}_{\mathsf{rnn}}^{(d_y,1,k)}$ and the RNN such that the approximation error is uniformly bounded by $e^{-M}\frac{\epsilon}{4}$ for all $t = 1 \dots n$. Finally, combining Theorem 9 and Lemma 5 we have

$$\begin{aligned}&\left|\widehat{\mathsf{D}}_Y(D_n) - \mathsf{D}_Y^{(\infty)}\right| \\ &\le (1+e^M)\frac{1}{n}\sum_{i=1}^{n}\left|f_{y,i}^\star(\widetilde{Y}, Y_{-i}^{-1}) - g_y^\star(\widetilde{Y}, Y_{-i}^{-1})\right| + \frac{\epsilon}{4} \\ &\le \frac{\epsilon}{2}. \end{aligned} \tag{49}$$

This concludes the proof of (35a). For (35b), note that

$$\begin{aligned} f_{xy,\infty}^\star &= \log\left(\frac{\mathrm{d}P_{X_{-\infty}^0\|Y_{-\infty}^{-1}} \otimes P_{Y_{-\infty}^0\|X_{-\infty}^0}}{\mathrm{d}P_{X_{-\infty}^0\|Y_{-\infty}^{-1}} \otimes P_{Y_{-\infty}^{-1}\|X_{-\infty}^{-1}} \otimes P_{\widetilde{Y}}}\right) \\ &= \log p_{Y_0|Y_{-\infty}^{-1}X_{-\infty}^0} - c_{\mathcal{Y}} \end{aligned}$$

achieves the supremum. Following similar arguments to those above, one may verify that

$$\left|\widehat{\mathsf{D}}_{Y\|X}(D_n) - \mathsf{D}_{Y\|X}^{\infty}\right| < \frac{\epsilon}{2}, \qquad \mathbb{P}\text{-}a.s., \tag{50}$$

where

$$\begin{aligned}\widehat{\mathsf{D}}_{Y\|X}(D_n) := \sup_{g_{xy}\in\mathcal{G}_{\mathsf{rnn}}^{XY}} &\frac{1}{n}\sum_{i=0}^{n-1} g_{xy}(Y_{-i}^0, X_{-i}^0) \\ &- \log\left(\frac{1}{n}\sum_{i=0}^{n-1} e^{g_{xy}(Y_{-i}^{-1}, X_{-i}^0, \widetilde{Y})}\right). \end{aligned} \tag{51}$$

Combining (49) and (50) concludes the proof. □

B. Proof of Theorem 5

Let $\epsilon > 0$ and $U^n \sim P_U^{\otimes n}$. Fix the USC $\{P_{Y_i|Y^{i-1}X^{i-1}}\}_{i\in\mathbb{Z}}$ as defined in Section V-B. Recall that X_s includes the class of stationary Markov processes of finite order and is therefore non-empty. Thus, there exist some $\mathbb{X}^\epsilon \in \mathsf{X}_s$ such that $|\mathsf{I}(\mathbb{X}^\epsilon \to \mathbb{Y}) - \underline{\mathsf{C}}_s| \le \epsilon/3$ by its definition as a supremum over a non-empty set. We denote a corresponding sample of $\mathbb{X}^\epsilon$ and the channel by $D_n^\epsilon = (X^{\epsilon,n}, Y^{\epsilon,n}) \sim \prod_{i=1}^n P_{X_i^\epsilon|X^{\epsilon,i-1}} P_{Y_i|X^iY^{i-1}}$. We have

$$\begin{aligned}\left|\underline{\mathsf{C}}_s - \widehat{\mathsf{I}}_{\mathsf{DI}}^\star(U^n)\right| &\le \frac{\epsilon}{3} + \left|\mathsf{I}(\mathbb{X}^\epsilon \to \mathbb{Y}) - \widehat{\mathsf{I}}_{\mathsf{DI}}(D_n^\epsilon)\right| \\ &\quad + \left|\widehat{\mathsf{I}}_{\mathsf{DI}}(D_n^\epsilon) - \widehat{\mathsf{I}}_{\mathsf{DI}}^\star(U^n)\right| \\ &\le \frac{2\epsilon}{3} + \left|\widehat{\mathsf{I}}_{\mathsf{DI}}(D_n^\epsilon) - \widehat{\mathsf{I}}_{\mathsf{DI}}^\star(U^n)\right| \qquad (52) \\ &= \frac{2\epsilon}{3} + \inf_{h_\phi\in\mathcal{G}_{\mathsf{rnn}}^X}\left|\widehat{\mathsf{I}}_{\mathsf{DI}}(D_n^\epsilon) - \widehat{\mathsf{I}}_{\mathsf{DI}}(D_n^\phi, h_\phi)\right|, \qquad (53) \end{aligned}$$

where (52) follows from Theorem 2 for a large enough $n \in \mathbb{N}$, and $\widehat{\mathsf{I}}_{\mathsf{DI}}(D_n^\epsilon)$ is given in (22). Therefore, our goal is to bound the remaining term in (53), which quantifies the DINE error induced by using the approximating dataset D_n^ϕ.

First, we show that the evolution of an RSP can be reformulated as an open dynamic system. Namely, an open dynamic system with inputs v^n, states s^n and outputs x^n taking values in $\mathcal{V} \subseteq \mathbb{R}^{d_v}$, $\mathcal{S} \subseteq \mathbb{R}^{d_s}$, $\mathcal{X} \subseteq \mathbb{R}^{d_x}$, respectively, is given by following set of equations [43, Eqn. 1].

$$s_{t+1} = f_1(s_t, v_t) \tag{54a}$$
$$x_t = f_{xy}(s_t), \tag{54b}$$

where f_1 is Borel measurable and $f_2 \in \mathsf{C}(\mathcal{S})$. Traditionally, an open dynamic system is a system that is driven by an external input sequence. Therefore, the system in Definition 4 is also an open dynamic system, by adding an external output signal that is defined via the identity mapping $Y_t = S_t$. Recall that the evolution of $\mathbb{X} \in \mathsf{X}_s$ is described by the relation

$$S_i = f_{\mathsf{s}}(X_i, S_{i-1}) \tag{55a}$$
$$P_{X_i|X^{i-1},S_{i-1}} = P_{X_i|S_{i-1}}. \tag{55b}$$

To show that (55) adheres to the relation presented in (54), we utilize the following lemma.

Lemma 6 (Functional representation of RSPs): For any $\mathbb{X} \in \mathsf{X}_{\mathcal{S}}$ with state process $\mathbb{S}$ and an i.i.d. process $\mathbb{W}$ with $W_1 \sim P_W \in \mathcal{P}_{\mathsf{ac}}(\mathcal{W})$ and $\mathcal{W} \subseteq \mathbb{R}^{d_x}$, there exists a function $f_{\mathsf{x}} : \mathcal{S} \times \mathcal{W} \to \mathcal{X}$ such that

$$X_i = f_{\mathsf{x}}(S_{i-1}, W_i), \qquad \forall i \in \mathbb{N}. \tag{56}$$

The proof is given in Appendix IX-C. It follows from the stationarity of $\mathbb{X}$, the FRL and Lemma 3. Lemma 6 provides us with f_{x} such that

$$S_i = f_{\mathsf{s}}(X_i^\epsilon, S_{i-1}), \quad X_i^\epsilon = f_{\mathsf{x}}(S_{i-1}, U_i).$$

As a final step towards the relation (54), denote $\widetilde{S}_i := (S_i, U_i)$ and $V_i := (U_i, X_i^\epsilon)$ and define $\widetilde{f}_{\mathsf{s}}$ such that the first d_s components of $\widetilde{S}_i$ are calculated from $f_{\mathsf{s}}(S_{i-1}, X_{i-1}^\epsilon)$ and the rest of its components comprise of replacing U_{i-1} with U_i. We therefore have the following open-dynamical system representation.

$$\begin{aligned}\widetilde{S}_i &= \widetilde{f}_{\mathsf{s}}(\widetilde{S}_{i-1}, V_i) \\ X_i^\epsilon &= f_{\mathsf{x}}(\widetilde{S}_i). \end{aligned} \tag{57}$$

Having an open-dynamical system representation of $\mathbb{X}^\epsilon$, we will approximate it with RNNs, due to the following Theorem [43, Theorem 2].

Theorem 10: Let $n \in \mathbb{N}$, $\epsilon > 0$, and let $u_t \in \mathbb{R}^{d_i}$, $s_t \in \mathbb{R}^{d_s}$ and $x_t \in \mathbb{R}^{d_o}$ be the inputs, states and outputs of an open dynamic system for $t = 1, \dots, n$. Then, there exists $k \in \mathbb{N}$ and $h_\phi \in \mathcal{G}_{\mathsf{rnn}}^{(d_i,d_o,k)}$ such that

$$\max_{t=1,\dots,n}\left\|h_\phi(u^i) - x_i\right\|_1 \le \epsilon. \tag{58}$$

Therefore, take $\epsilon' > 0$ and fix sample $u^n \in \mathcal{U}^n$ drawn according to $P_U^{\otimes n}$; there exists $k \in \mathbb{N}$ and $h_\phi \in \mathcal{G}_{\mathsf{rnn}}^X$ such that

$$\max_{t=1\dots n}\left\|x_i^\epsilon(u^i) - x_i^\phi(u^i)\right\|_1 \le \epsilon'. \tag{59}$$

Our next step is to bound $\left\|y_i^\epsilon - y_i^\phi\right\|_1$ in terms of $\left\|x_i^\epsilon - x_i^\phi\right\|_1$ for $i = 1, \dots, n$. To that end, consider the following lemma.

Lemma 7: Let $T \in \mathbb{N}$ and $\mathbb{Y}$ be the output of the USC described in Section V-B.1 with f_{y} and f_{z} satisfying Assumption 1 with Lipschitz constants M_y and M_z, respectively. Then, for any $n \in \mathbb{N}$, every pair of input sequences $(x^{1,n}, x^{2,n})$ such that $\max_{t=1,\dots,n} \|x_t^1 - x_t^2\|_1 \leq \eta$, we have

$$\max_{i=1,\dots,T} \|y_t^1 - y_t^2\|_1 \leq \frac{M_y(2 - M_z(M_y+1))}{1 - M_z(M_y+1)}\eta.$$

The proof of Lemma 7 is in Appendix IX-D. We further denote

$$\alpha(M_y, M_z) := \frac{M_y(2 - M_z(M_y+1))}{1 - M_z(M_y+1)}.$$

Finally, we have

$$\begin{aligned}
&\left| \widehat{\mathsf{I}}_{\mathsf{DI}}(D_n^{\epsilon}) - \widehat{\mathsf{I}}_{\mathsf{DI}}(D_n^{\phi}, h_{\phi}) \right| \\
&\leq \frac{1}{n}\sum_{i=1}^{n} \left| g_y(y_i^{\epsilon}|y^{\epsilon,i-1}) - g_y(y_i^{\phi}|y^{\phi,i-1}) \right| \\
&+ \frac{e^M}{n}\sum_{i=1}^{n} \left| g_y(\widetilde{y}|y^{\epsilon,i-1}) - g_y(\widetilde{y}|y^{\phi,i-1}) \right| \\
&+ \frac{1}{n}\sum_{i=1}^{n} \left| g_{xy}(y_i^{\epsilon}|y^{\epsilon,i-1}, x^{\epsilon,i}) - g_{xy}(y_i^{\phi}|y^{\phi,i-1}, x^{\phi,i}) \right| \\
&+ \frac{e^M}{n}\sum_{i=1}^{n} \left| g_{xy}(\widetilde{y}|y^{\epsilon,i-1}, x^{\epsilon,i}) - g_{xy}(\widetilde{y}|y^{\phi,i-1}, x^{\phi,i}) \right|. \quad (60)
\end{aligned}$$

By assumption, g_y and g_{xy} are Lipschitz continuous with Lipschitz constants M_1, M_2, respectively. Consequently, we have

$$\begin{aligned}
&\left| \widehat{\mathsf{I}}_{\mathsf{DI}}(D_n^{\epsilon}) - \widehat{\mathsf{I}}_{\mathsf{DI}}(D_n^{\phi}, h_{\phi}) \right| \\
&\leq \frac{(M_1+M_2)(1+e^M)}{n}\sum_{i=1}^{n} \left\| y_i^{\epsilon} - y_i^{\phi} \right\|_1 \\
&\qquad + \frac{M_2(1+e^M)}{n}\sum_{i=1}^{n} \left\| x_i^{\epsilon} - x_i^{\phi} \right\|_1 \\
&\leq \left((M_1+M_2)(1+e^M)\alpha(M_y, M_z) + M_2(1+e^M) \right) \epsilon'. \quad (61)
\end{aligned}$$

Take a large enough $k \in \mathbb{N}$ such that (61) is bounded by $\epsilon/3$. As the above steps hold for any realization of U^n and K^n, the inequality (61) holds $\mathbb{P}$-*a.s.* This concludes the proof. □

Remark 6 (Lipschitz Assumption): Lemma 7 calls for Assumption 1 due to the recursive nature of the proposed channel, i.e., Y_i and Z_i indirectly depend on their past values and the induced error accumulates over time. By restricting f_{z} to be a function of only X_i, the resulting Lipschitz constants M_z, M_y are no longer bound to $M_y(M_z+1) < 1$.

Remark 7 (Channels With Feedback): To account for the feedback scenario, we first consider a *conditional* version of $\mathsf{X}_{\mathcal{S}}$ that allows conditioning on past channel outputs. The state S_i is then taken as a function of (X_i, S_{i-1}, Y_{i-1}) and we require $P_{X_i|X^{i-1},Yi-1,S^{i-1}} = P_{X_i|S_i}$. Lemma 6 follows immediately, as the FRL holds even when conditioning on additional random variables. The rest of the proof follows by adding Y_i to the ith input of f_{s}.

C. Proof of Theorem 6

Let $U \sim P_U$ and $P_{Y|X}$ be a given transition kernel. Throughout this proof we employ the tools of Gaussian smoothing developed in [89] (see also [90], [91], [92], [93], [94]). To this end, we denote the isotropic d_x-dimensional Gaussian distribution with $\mathcal{N}_{\sigma} := \mathcal{N}(0, \sigma^2 \mathrm{I}_{d_x})$ with the corresponding PDF φ_{σ}. Let $P_{X^\star}$ be the MI maximizing input distribution for $P_{Y|X}$ and denote its corresponding smoothed distribution with $P_{X_\sigma^\star} := P_{X^\star} * \mathcal{N}_{\sigma}$. For any choice of $\sigma > 0$ we have $P_{X_\sigma^\star} \in \mathcal{P}_{\mathsf{ac}}(\mathcal{X})$, which implies the existence of the bijection $T_{X_\sigma^\star} \in \mathsf{C}^1(\mathcal{U}, \mathcal{X})$ due to Lemma 3. We utilize the universal approximation theorem for NNs with arbitrary finite output dimension [43, Corollary 1].

Lemma 8 (Universal Approximation of NNs): Let $\mathsf{C}(\mathcal{X}, \mathcal{Y})$ be the class continuous functions $f : \mathcal{X} \to \mathcal{Y}$ where $\mathcal{X} \subset \mathbb{R}^{\mathsf{d}_i}$ is compact and $\mathcal{Y} \subseteq \mathbb{R}^{\mathsf{d}_o}$. Then, the class of NNs $\mathcal{G}_{\mathsf{nn}}^{(\mathsf{d}_i, \mathsf{d}_o)}$ is dense in $\mathsf{C}(\mathcal{X}, \mathcal{Y})$, i.e., for every $f \in \mathsf{C}(\mathcal{X}, \mathcal{Y})$ and $\epsilon > 0$, there exist $g \in \mathcal{G}_{\mathsf{nn}}^{(\mathsf{d}_i, \mathsf{d}_o)}$ such that $\|f - g\|_{\infty} \leq \epsilon$.

By Lemma 8, we can construct a sequence of functions $\{h_{\phi,k}\}_{k\in\mathbb{N}} \subset \mathcal{G}_{\mathsf{nn}}^{(d_x, d_x)}$ such that $\|h_{\phi,k} - T_{X_\sigma^\star}^{-1}\|_{\infty} \to 0$. Setting $P_{X^{\phi_k}} := h_{\phi,k\sharp} P_U$, we therefore obtain $P_{X^{\phi_k}} \rightharpoonup P_{X_\sigma^\star}$, where $\rightharpoonup$ denotes weak convergence of probability measures.[7] As a consequence of the weak convergence, the compactness of $\mathcal{U}$, and the continuity of $h_{\phi,k}$, we have convergence of second moments, i.e., $\int_{\mathbb{R}^{d_x}} \|x\|^2 \,\mathrm{d}P_{X^{\phi_k}}(x) \to \int_{\mathbb{R}^{d_x}} \|x\|^2 \,\mathrm{d}P_{X_\sigma^\star}(x)$. As weak convergence plus convergence in 2-th moments is equivalent to convergence under the 2-Wasserstein distance, we obtain $\mathsf{W}_2(P_{X^{\phi_k}}, P_{X_\sigma^\star}) \to 0$ as $\sigma \to 0$[8] (cf., e.g., [95, Theorem 7.12]).

Given a non-increasing sequence $\sigma_i \searrow 0$, it is readily verified that $P_{X_{\sigma_i}^\star} \rightharpoonup P_{X^\star}$ and the second moments converge as well. Indeed, the former follows because weak convergence is equivalent to pointwise convergence of characteristic functions together with the fact that the characteristic function of $\mathcal{N}_{\sigma}$ never vanishes; the latter follows from a uniform integrability argument. We therefore have $\mathsf{W}_2(P_{X_{\sigma_i}^\star}, P_{X^\star}) \overset{i\to\infty}{\longrightarrow} 0$. To bound $\mathsf{W}_2(P_{X^\star}, P_{X^{\phi_k}})$ we perform two steps of approximation; first, we approximate $P_{X^\star}$ with $P_{X_{\sigma_i}^\star}$ which is then approximated with $P_{X^{\phi_k}}$. Take large enough $i, k \in \mathbb{N}$ such that the corresponding 2-Wasserstein metrics are bounded by $\epsilon/2$ and apply the triangle inequality to result with

$$\mathsf{W}_2(P_{X^\star}, P_{X^{\phi_k}}) \leq \mathsf{W}_2(P_{X^\star}, P_{X_{\sigma_i}^\star}) + \mathsf{W}_2(P_{X_{\sigma_i}^\star}, P_{X^{\phi_k}}) \leq \epsilon. \quad (62)$$

We stress that k is taken w.r.t. the chosen index of σ_i, but omit this in our notation for simplification.

All considered input-output pairs are distributed with the fixed transition kernel $P_{Y|X}$, therefore, they only differ by the input distribution. To bound the difference (30), we consider two intermediate steps of approximation. First, we consider the MI induced by the approximation of $X^\star$ by an element from the sequence of its Gaussian smoothed counterpart

[7] A sequence of measures $\{\mu_n\}_{n\in\mathbb{N}}$ converges weakly to a measure μ if $\int f \,\mathrm{d}\mu_n \to \int f \,\mathrm{d}\mu$ for any continuous and bounded function f.

[8] In general, we have convergence of any pth moment for any $p < \infty$, therefore, convergence of pth Wasserstein distance.

for some σ_i, denoted $X^\star_{\sigma_i} := X^\star + Z_{\sigma_i}$, where $Z_{\sigma_i} \sim \mathcal{N}(0, \sigma_i^2 \mathrm{I}_{d_x})$. Then, our task is to approximate the MI induced by $X^\star_{\sigma_i}$ with the MI induced by $X^{\phi_k} := h_{\phi,k}(U)$. To do so, we apply an additional intermediate step of an approximation of both elements with a smoothed version of X^{ϕ_k}, denoted $X^{\phi_k}_{\sigma_\ell} := h_{\phi,k}(U) + Z_{\sigma_\ell}$, where $Z_{\sigma_\ell} \sim \mathcal{N}(0, \sigma_\ell^2 \mathrm{I}_{d_x})$ for some σ_ℓ. The last step consists of approximating the MI induced by X^{ϕ_k} and its n-sample MINE approximation calculated from $D_n^{\phi_k} = \{(h_{\phi,k}(U_i), Y_i)\}_{i=1}^n$. By the triangle inequality, we have

$$\begin{aligned}\left|\mathsf{C} - \widehat{\mathsf{I}}_{\mathsf{MI}}\big(D_n^{\phi_k}\big)\right| &\le \left|\mathsf{I}(X^\star;Y^\star) - \mathsf{I}(X^\star_{\sigma_i};Y^\star_{\sigma_i})\right| \\ &+ \left|\mathsf{I}(X^\star_{\sigma_i};Y^\star_{\sigma_i}) - \mathsf{I}(X^{\phi_k}_{\sigma_\ell};Y^{\phi_k}_{\sigma_\ell})\right| \\ &+ \left|\mathsf{I}(X^{\phi_k}_{\sigma_\ell};Y^{\phi_k}_{\sigma_\ell}) - \mathsf{I}(X^{\phi_k};Y^{\phi_k})\right| \\ &+ \left|\mathsf{I}(X^{\phi_k};Y^{\phi_k}) - \widehat{\mathsf{I}}_{\mathsf{MI}}\big(D_n^{\phi_k}\big)\right|. \end{aligned} \tag{63}$$

To bound the first term in (63), we utilize the weak lower semicontinuity of MI [86, Section 3.5.2], i.e., $P_{X^\star_{\sigma_i},Y^\star_{\sigma_i}} \rightharpoonup P_{X^\star,Y^\star}$ implies

$$\mathsf{I}(X^\star;Y^\star) \le \liminf_{i\to\infty} \mathsf{I}(X^\star_{\sigma_i};Y^\star_{\sigma_i}). \tag{64}$$

With some abuse of notation, extract a subsequence $(X^\star_{\sigma_j}, Y^\star_{\sigma_j})_{j\in\mathbb{N}}$ that achieves the RHS of (64). Recall that $P_{X^\star_j,Y^\star_j} \rightharpoonup P_{X^\star,Y^\star}$. Along with the weak lower semicontinuity of MI and the fact that $X^\star$ achieves capacity for the fixed $P_{Y|X}$, there exist $j \in \mathbb{N}$ such that

$$|\mathsf{C} - \mathsf{I}(X^\star_{\sigma_j};Y^\star_{\sigma_j})| \le \frac{\epsilon}{3}. \tag{65}$$

To bound the second term in (63), we consider a non-increasing sequence $\sigma_\ell \searrow 0$ and denote $P_{X^{\phi_k}_{\sigma_\ell}} := P_{X^{\phi_k}} * \mathcal{N}_{\sigma_l}$, where $k_j \in \mathbb{N}$ is taken such that the bound (62) still holds. The second term in (63) can then be bounded as follows.

$$\begin{aligned}&\left|\mathsf{I}(X^\star_{\sigma_j};Y^\star_{\sigma_j}) - \mathsf{I}(X^{\phi_k}_{\sigma_\ell};Y^{\phi_k}_{\sigma_\ell})\right| \\ &= \left|\mathsf{D}_{\mathsf{KL}}(P_{X^\star_{\sigma_j}Y^\star_{\sigma_j}} \| P_{X^\star_{\sigma_j}} P_{Y^\star_{\sigma_j}}) - \mathsf{D}_{\mathsf{KL}}(P_{X^{\phi_k}_{\sigma_\ell}Y^{\phi_k}_{\sigma_\ell}} \| P_{X^{\phi_k}_{\sigma_\ell}} P_{Y^{\phi_k}_{\sigma_\ell}})\right| \\ &= \left|\mathbb{E}_{P_{X^\star_{\sigma_j}Y^\star_{\sigma_j}}}\left[\log \frac{p_{X^\star_{\sigma_j}Y^\star_{\sigma_j}}}{p_{X^{\phi_k}_{\sigma_\ell}Y^{\phi_k}_{\sigma_\ell}}}\right] + \mathbb{E}_{P_{X^{\phi_k}_{\sigma_\ell}Y^{\phi_k}_{\sigma_\ell}}}\left[\log \frac{p_{X^{\phi_k}_{\sigma_\ell}}p_{Y^{\phi_k}_{\sigma_\ell}}}{p_{X^\star_{\sigma_j}}p_{Y^\star_{\sigma_j}}}\right]\right| \\ &= \left|\mathsf{D}_{\mathsf{KL}}(P_{X^\star_{\sigma_j}} \| P_{X^{\phi_k}_{\sigma_\ell}}) + \mathbb{E}_{P_{X^{\phi_k}_{\sigma_\ell}Y^{\phi_k}_{\sigma_\ell}}}\left[\log \frac{p_{X^{\phi_k}_{\sigma_\ell}}p_{Y^{\phi_k}_{\sigma_\ell}}}{p_{X^\star_{\sigma_j}}p_{Y^\star_{\sigma_j}}}\right]\right| \end{aligned} \tag{66}$$

where (66) follows from the construction of both joint distributions with the same transition kernel $P_{Y|X}$. The second term in (66) can be represented as follows.

$$\begin{aligned}&\mathbb{E}_{P_{X^{\phi_k}_{\sigma_\ell}Y^{\phi_k}_{\sigma_\ell}}}\left[\log \frac{p_{X^{\phi_k}_{\sigma_\ell}}p_{Y^{\phi_k}_{\sigma_\ell}}}{p_{X^\star_{\sigma_j}}p_{Y^\star_{\sigma_j}}}\right] \\ &= \int_{\mathcal{X}}\int_{\mathcal{Y}} \log \frac{p_{X^{\phi_k}_{\sigma_\ell}}(x)p_{Y^{\phi_k}_{\sigma_\ell}}(y)}{p_{X^\star_{\sigma_j}}(x)p_{Y^\star_{\sigma_j}}(y)} p_{X^{\phi_k}_{\sigma_\ell}Y^{\phi_k}_{\sigma_\ell}}(x,y)\,\mathrm{d}x\,\mathrm{d}y \\ &= \int_{\mathcal{X}} \log \frac{p_{X^{\phi_k}_{\sigma_\ell}}(x)}{p_{X^\star_{\sigma_j}}(x)} p_{X^{\phi_k}_{\sigma_\ell}}(x)\,\mathrm{d}x + \int_{\mathcal{Y}} \log \frac{p_{Y^{\phi_k}_{\sigma_\ell}}(y)}{p_{Y^\star_{\sigma_j}}(y)} p_{Y^{\phi_k}_{\sigma_\ell}}(y)\,\mathrm{d}y \\ &= \mathsf{D}_{\mathsf{KL}}(P_{X^{\phi_k}_{\sigma_\ell}} \| P_{X^\star_{\sigma_j}}) + \mathsf{D}_{\mathsf{KL}}(P_{Y^{\phi_k}_{\sigma_\ell}} \| P_{Y^\star_{\sigma_j}}). \end{aligned} \tag{67}$$

Plug (67) into (66) and apply the data-processing inequality for KL divergences to obtain

$$\begin{aligned}&\left|\mathsf{I}(X^\star_{\sigma_j};Y^\star_{\sigma_j}) - \mathsf{I}(X^{\phi_k}_{\sigma_\ell};Y^{\phi_k}_{\sigma_\ell})\right| \\ &\le 2\left(\mathsf{D}_{\mathsf{KL}}(P_{X^\star_{\sigma_j}} \| P_{X^{\phi_k}_{\sigma_\ell}}) + \mathsf{D}_{\mathsf{KL}}(P_{X^{\phi_k}_{\sigma_\ell}} \| P_{X^\star_{\sigma_j}})\right). \end{aligned} \tag{68}$$

We note that both KL terms are well defined as both $P_{X^\star_{\sigma_j}}$ and $P_{X^{\phi_k}_{\sigma_\ell}}$ are defined and positive over the same space as Gaussian smoothed distributions. We will now upper bound the RHS of (68) with $\mathsf{W}_2(P_{X^\star_{\sigma_j}}, P_{X^{\phi_k}_{\sigma_\ell}})$, using the following result [80, Proposition 1], which was recently improved in [96, Lemma 1].

Theorem 11: Let U and V be random vectors with finite second moments. If both U and V are (c_1, c_2)-regular, then

$$\mathsf{D}_{\mathsf{KL}}(P_U\|P_V) + \mathsf{D}_{\mathsf{KL}}(P_V\|P_U) \le 2\Delta, \tag{69}$$

where P_U is (c_1, c_2)-regular if

$$\|\nabla \log p_U(u)\|_2 \le c_1\|u\|_2 + c_2, \tag{70}$$

and

$$\Delta := \left(\frac{c_1}{2}\left(\sqrt{\mathbb{E}\left[\|V\|_2^2\right]} + \sqrt{\mathbb{E}\left[\|U\|_2^2\right]}\right) + c_2\right)\mathsf{W}_2(P_U, P_V). \tag{71}$$

The (c_1, c_2) regularity of $P_{X^\star_{\sigma_j}}$ and $P_{X^{\phi_k}_{\sigma_\ell}}$ follows from the Gaussian smoothing of $P_{X^\star}$ and $P_{X^{\phi_k}}$ such that the regularity parameters depend on σ_j and σ_ℓ [80, Proposition 2]. Note that $P_{X^{\phi_k}} \in \mathcal{P}_2(\mathcal{X})$ follows from the compactness of $h_{\phi,k}(\mathcal{U})$. Consequently, $P_{X^\star_{\sigma_j}} \in \mathcal{P}_2(\mathcal{X})$ as $\mathbb{E}\big[\|X^\star_{\sigma_j}\|^2\big] = \mathbb{E}\big[\|X^\star\|^2\big] + \big[\|Z_{\sigma_j}\|^2\big]$, both having finite second moment. We can therefore bound (68) with $\mathsf{W}_2(P_{X^\star}, P_{X^{\phi_k}_{\sigma_\ell}})$, which by the triangle inequality, amounts to

$$\mathsf{W}_2(P_{X^\star_{\sigma_j}}, P_{X^{\phi_k}_{\sigma_\ell}}) \le \mathsf{W}_2(P_{X^\star_{\sigma_j}}, P_{X^{\phi_k}}) + \mathsf{W}_2(P_{X^{\phi_k}}, P_{X^{\phi_k}_{\sigma_\ell}}). \tag{72}$$

For given ϵ take k and ℓ large enough and utilize the weak continuity of $\mathsf{W}_2(P_{X^\star_{\sigma_j}}, P_{X^{\phi_k}_{\sigma_\ell}})$ to obtain an $\epsilon/6$ bound on (68).

We now describe the bound of the third term in the RHS of (63).

First, represent each MI as a combination of differential entropies to obtain the following bound

$$\begin{aligned}&\left|\mathsf{I}(X^{\phi_k}_{\sigma_\ell};Y^{\phi_k}_{\sigma_\ell}) - \mathsf{I}(X^{\phi_k};Y^{\phi_k})\right| \\ &\le \left|\mathsf{h}(X^{\phi_k}_{\sigma_l}) - \mathsf{h}(X^{\phi_k})\right| + \left|\mathsf{h}(Y^{\phi_k}_{\sigma_l}) - \mathsf{h}(Y^{\phi_k})\right| \\ &\qquad + \left|\mathsf{h}(X^{\phi_k}_{\sigma_l},Y^{\phi_k}_{\sigma_l}) - \mathsf{h}(X^{\phi_k},Y^{\phi_k})\right|. \end{aligned} \tag{73}$$

We will utilize the following Theorem [81, Theorem 1].

Theorem 12 (Convergence of differential entropies): Let $(X_i)_{i\in\mathbb{N}}$ be a sequence of continuous random variables with PDFs $(f_i)_{i\in\mathbb{N}}$ and X be a continuous random variable with PDF f such that $f_i \to f$ pointwise. If

$$\max\{\|f_i\|_\infty, \|f\|_\infty\} \le A_1 < \infty \tag{74a}$$

$$\max\left\{\int \|x\|^\kappa f_i(x)\,\mathrm{d}x, \int \|x\|^\kappa f(x)\,\mathrm{d}x\right\} \le A_2 < \infty, \tag{74b}$$

for some $\kappa > 1$ and for all $i \in \mathbb{N}$, then $h(X_i) \to h(X)$.

We will now show that the conditions of Theorem 12 hold in our case, focusing on $\kappa = 2$. Note that if such conditions hold for the input and output distributions, they hold for the joint distribution as well. To justify the pointwise convergence of PDFs we introduce the notion of asymptotic equicontinuity (a.e.c.). A function f is a.e.c. on $x \in \mathcal{X}$ if for every $\epsilon > 0$ there exist $\delta(x,\epsilon)$ and $n_0(x,\epsilon)$ such that whenever $\|x-y\|_1 < \delta(x,\epsilon)$, then $|f_n(x) - f_n(y)| < \epsilon$ for any $n > n_0$. We use following theorem [97, Theorem 1].

Theorem 13 (Pointwise Convergence of PDF Sequence): Let $(P_n)_{n\in\mathbb{N}} \subset \mathcal{P}(\mathcal{X})$ with PDFs $(p_n)_{n\in\mathbb{N}}$. The following statements are equivalent.

1) $(p_n)_{n\in\mathbb{N}}$ are a.e.c. on $\mathcal{X}$ and $P_n \rightharpoonup P$.
2) $p_n \to p$ pointwise, where p is the continuous PDF of P.

Recall that both $P_{X^{\phi_k}_{\sigma_\ell}}$ and $P_{Y^{\phi_k}_{\sigma_\ell}}$ weakly converge to $P_{X^{\phi_k}}$ and $P_{Y^{\phi_k}}$, respectively. The a.e.c. property of $p_{x^{\phi_k}_{\sigma_\ell}}$ follows from its structure is a convolution with a Gaussian density, as follows

$$\begin{aligned}
&\left|p_{X^{\phi_k}_{\sigma_\ell}}(x_1) - p_{X^{\phi_k}_{\sigma_\ell}}(x_2)\right| \\
&= \left| \int_{\mathbb{R}^{d_x}} p_{X^{\phi_k}}(x_1-u)\varphi_{\sigma_\ell}(u)\,\mathrm{d}u \right. \\
&\qquad\qquad \left. - \int_{\mathcal{X}} p_{X^{\phi_k}}(x_2-u)\varphi_{\sigma_\ell}(u)\,\mathrm{d}u \right| \\
&= \left| \int_{\mathbb{R}^{d_x}} \varphi_{\sigma_\ell}(u)\left(p_{X^{\phi_k}}(x_1-u) - p_{X^{\phi_k}}(x_2-u)\right)\,\mathrm{d}u \right| \\
&< \left| \int_{\mathbb{R}^{d_x}} \varphi_{\sigma_\ell}(u)\epsilon\,\mathrm{d}u \right| \qquad (75) \\
&= \epsilon, \qquad (76)
\end{aligned}$$

where (75) follows from the continuity of $p_{X^{\phi_k}_{\sigma_\ell}}$, taking appropriate $\delta > 0$. The a.e.c. property of $p_{Y^{\phi_k}_{\sigma_\ell}}$ follows from the same steps and the continuity of $p_{Y|X}$ on $\mathcal{Y}$. The boundedness of $p_{X^{\phi}_{k}}$ follows from the extreme value theorem, as it is a continuous function on $h_{\phi_k}(\mathcal{U})$. The PDF $p_{X^{\phi_k}_{\sigma_\ell}}$ is integrable due to Fubini's theorem. Consequently, the PDFs $p_{X^{\phi_k}_{\sigma_\ell}}$, $p_{Y^{\phi_k}}$ and $p_{Y^{\phi_k}_{\sigma_\ell}}$ are bounded as they are continuous integrable PDFs on $\mathbb{R}^{d_x}$. The second moment of X^{ϕ_k} is bounded by the compactness of $h_{\phi,k}(\mathcal{U})$ and the second moment bound of $X^{\phi_k}_{\sigma_\ell}$ follows from

$$\begin{aligned}
\mathbb{E}\left[\|X^{\phi_k}_{\sigma_\ell}\|_2^2\right] &= \mathbb{E}\left[\|X^{\phi_k}\|_2^2\right] + \mathbb{E}\left[\|Z_{\sigma_\ell}\|_2^2\right] \\
&= \mathbb{E}\left[\|X^{\phi_k}\|_2^2\right] + d_x\sigma_\ell^2 \\
&< \infty,
\end{aligned}$$

where $Z_{\sigma_\ell} \sim \mathcal{N}(0, \sigma_\ell \mathrm{I}_{d_x})$ is independent of X^{ϕ_k}. The second moment bound for Y^{ϕ_k} and $Y^{\phi_k}_{\sigma_\ell}$ follows from the assumption on $P_{Y|X}$. We can therefore apply Theorem 12 to bound the differences of differential entropies in (73). Take ℓ large enough such that both (73) and (68) are bounded by $\epsilon/6$.

Finally, the fourth term in (63) can be bounded by $\epsilon/3$ for large enough $n \in \mathbb{N}$ using the MINE consistency [24, Theorem 2], which concludes the proof. □

IX. Proofs of Lemmas

A. Proof of Lemma 4

Recall that

$$\begin{aligned}
D^n_{Y\|X} &:= \mathsf{D_{KL}}\left(P_{Y^n\|X^n}\middle\|P_{Y^{n-1}\|X^{n-1}} \otimes P_{\widetilde{Y}}\middle|P_{X^n\|Y^{n-1}}\right) \\
D^n_{Y} &:= \mathsf{D_{KL}}\left(P_{Y^n}\middle\|P_{Y^{n-1}} \otimes P_{\widetilde{Y}}\right).
\end{aligned}$$

We first show that $\mathsf{I}(\mathbb{X} \to \mathbb{Y}) = \lim_{n\to\infty}\left(D^n_{Y\|X} - D^n_Y\right)$. Recall that (see Section II-B)

$$\mathsf{I}(X^n \to Y^n) = \mathsf{h}(Y^n) - \mathsf{h}(Y^n\|X^n), \qquad (77)$$

and expand

$$\begin{aligned}
&\mathsf{h}(Y^n) \\
&= \mathsf{h_{CE}}\left(P_{Y^n}, P_{Y^{n-1}} \otimes P_{\widetilde{Y}}\right) - \mathsf{D_{KL}}\left(P_{Y^n}\middle\|P_{Y^{n-1}} \otimes P_{\widetilde{Y}}\right), \qquad (78a) \\
&\mathsf{h}(Y^n\|X^n) \\
&= \mathsf{h_{CE}}\left(P_{Y^n\|X^n}, P_{Y^{n-1}\|X^{n-1}} \otimes P_{\widetilde{Y}}\middle|P_{X^n\|Y^{n-1}}\right) \\
&\qquad - \mathsf{D_{KL}}\left(P_{Y^n\|X^n}\middle\|P_{Y^{n-1}\|X^{n-1}} \otimes P_{\widetilde{Y}}\middle|P_{X^n\|Y^{n-1}}\right). \qquad (78b)
\end{aligned}$$

Subtraction yields

$$\begin{aligned}
&\mathsf{I}(X^n \to Y^n) \\
&= \Big(\mathsf{h_{CE}}(P_{Y^n}, P_{Y^{n-1}} \otimes P_{\widetilde{Y}}) \\
&- \mathsf{h_{CE}}(P_{Y^n\|X^n}, P_{Y^{n-1}\|X^{n-1}} \otimes P_{\widetilde{Y}}|P_{X^0_{-(n-1)}\|Y^{-1}_{-(n-1)}})\Big) \\
&+ \Big(\mathsf{D_{KL}}(P_{Y^n\|X^n}\|P_{Y^{n-1}\|X^{n-1}} \otimes P_{\widetilde{Y}}|P_{X^0_{-(n-1)}\|Y^{-1}_{-(n-1)}}) \\
&\qquad - \mathsf{D_{KL}}(P_{Y^n}\|P_{Y^{n-1}} \otimes P_{\widetilde{Y}})\Big). \qquad (79)
\end{aligned}$$

Denote the residual cross-entropy terms by $\mathsf{h}_{\mathsf{CE},Y}$ and $\mathsf{h}_{\mathsf{CE},Y\|X}$, respectively. By stationarity and since $\widetilde{Y} \perp\!\!\!\perp \mathbb{X}$, we further obtain

$$\begin{aligned}
&\mathsf{h}_{\mathsf{CE},Y} - \mathsf{h}_{\mathsf{CE},Y\|X} \\
&= \mathbb{E}\left[-\log P_{Y^{-1}_{-(n-1)}} \otimes P_{\widetilde{Y}}(\widetilde{Y}, Y^{-1}_{-n})\right] \\
&- \mathbb{E}\left[-\log P_{Y^{-1}_{-(n-1)}\|X^{-1}_{-(n-1)}} \otimes P_{\widetilde{Y}}(\widetilde{Y}, Y^{-1}_{-n})\right] \\
&= \mathbb{E}\left[-\log P_{Y^{-1}_{-(n-1)}}(Y^{-1}_{-n})\right] \\
&- \mathbb{E}\left[-\log P_{Y^{-1}_{-(n-1)}\|X^{-1}_{-(n-1)}}(Y^{-1}_{-n})\right] \\
&\qquad + \mathbb{E}\left[-\log P_{\widetilde{Y}}(\widetilde{Y})\right] - \mathbb{E}\left[-\log P_{\widetilde{Y}}(\widetilde{Y})\right] \\
&= \mathsf{h}(Y^{-1}_{-(n-1)}) - \mathsf{h}(Y^{-1}_{-(n-1)}\|X^{-1}_{-(n-1)}) \\
&= \mathsf{I}(X^{n-1} \to Y^{n-1}).
\end{aligned}$$

Plugging this term into (79) implies

$$\begin{aligned}
D^n_{Y\|X} - D^n_Y &= \mathsf{I}(X^{n-1} \to Y^n) - \mathsf{I}(X^n \to Y^{n-1}) \\
&= \mathsf{I}(X^0_{-(n-1)}; Y_0|Y^{-1}_{-(n-1)}) \\
&= \mathsf{h}(Y_0|Y^{-1}_{-(n-1)}) - \mathsf{h}(Y_0|Y^{-1}_{-(n-1)}X^0_{-(n-1)}).
\end{aligned} \qquad (80)$$

We now use the following theorem, restated from [56, Theorem 4.2.1].

Theorem 14 (Entropy Rate of Stationary Processes): For a stationary process $\{Y_n\}_{n\in\mathbb{Z}}$, the following limits exist and are equal:

$$\lim_{n\to\infty}\frac{1}{n}\mathsf{h}(Y_{-(n-1)}^{0})=\lim_{n\to\infty}\mathsf{h}(Y_0|Y_{-(n-1)}^{1}). \tag{81}$$

Together with (80), the lemma implies

$$\begin{aligned}&\lim_{n\to\infty} D_{Y\|X}^{n}-D_{Y}^{n}\\&=\lim_{n\to\infty}\mathsf{h}(Y_0|Y_{-(n-1)}^{-1})-\mathsf{h}(Y_0|Y_{-(n-1)}^{-1}X_{-(n-1)}^{0})\\&=\lim_{n\to\infty}\frac{1}{n}\Big(\mathsf{h}\left(Y_{-(n-1)}^{0}\right)-\mathsf{h}\left(Y_{-(n-1)}^{0}\|X_{-(n-1)}^{0}\right)\Big)\\&=\mathsf{I}(\mathbb{X}\to\mathbb{Y}).\end{aligned}$$

Our last step is to identify the limiting KL divergence terms using the monotone convergence theorem (cf., e.g., [86, Corollary 3.2]).

Theorem 15 (D_{KL} Monotone Convergence): The following holds:

$$\begin{aligned}&D_{Y}^{n}\nearrow\mathsf{D}_{\mathsf{KL}}\left(P_{Y_{-\infty}^{0}}\middle\|P_{Y_{-\infty}^{-1}}\otimes P_{\tilde{Y}}\right)\\&D_{Y\|X}^{n}\nearrow\mathsf{D}_{\mathsf{KL}}\left(P_{Y_{-\infty}^{0}\|X_{-\infty}^{0}}\middle\|P_{Y_{-\infty}^{-1}\|X_{-\infty}^{-1}}\otimes P_{\tilde{Y}}\middle|P_{X_{-\infty}^{0}}\right).\end{aligned} \tag{82}$$

Recalling the definition of D_{Y}^{∞} and $D_{Y\|X}^{\infty}$, this concludes the proof. □

B. Proof of Lemma 5

Let $\mathbb{Y}$ be a strictly stationary process. We utilize the following representation theorem [63, Theorem 2].

Theorem 16 (Representation of Stationary Processes): Let $(Y_t)_{t\in\mathbb{Z}}$ be a strictly stationary completely non-deterministic process. Then, there is an i.i.d. process $(V_t)_{t\in\mathbb{Z}}$ such that for any $t\in\mathbb{Z}$, X_t is given by

$$X_t=\sum_{k=0}^{\infty}a_kV_{t-k}, \tag{83}$$

where the series converges in probability.

Theorem 16 is an adaptation of the Wold decomposition [62] to strictly stationary processes. Let $\mathbb{Y}_{T'}$ be the process obtained by (83) for a fixed horizon $T'<\infty$. By Theorem 16 we have $Y_{T',t}\to Y_t$ in probability for all $t\in\mathbb{Z}$, and therefore in distribution.

Our goal is to translate this convergence into pointwise convergence of the underlying PDFs. To do that, we utilize the notion of asymptotic equicontinuity (a.e.c.), that is introduced in Section VIII-C. Recall that a function sequence $(f_n)_{n\in\mathbb{N}}$ is a.e.c. on $x\in\mathcal{X}$ if for every $\epsilon>0$ there exist $\delta(x,\epsilon)$ and $n_0(x,\epsilon)$ such that whenever $\|x-y\|_1<\delta(x,\epsilon)$, then $|f_n(x)-f_n(y)|<\epsilon$ for any $n>n_0$. We will utilize the following theorem [97, Theorem 1].

Theorem 17 (Theorem 13, Restated): Let $(P_n)_{n\in\mathbb{N}}\subset\mathcal{P}(\mathcal{X})$ with PDFs $(p_n)_{n\in\mathbb{N}}$. The following statements are equivalent.

1) $(p_n)_{n\in\mathbb{N}}$ are a.e.c. on $\mathcal{X}$ and $P_n\rightharpoonup P$.
2) $p_n\to p$ pointwise, where p is the continuous PDF of P.

The a.e.c. property of $(p_{Y_{T',t}})_{T'\in\mathbb{N}}$ follows from the relation $Y_{T'+1,t}=Y_{T',t}+b_{T'+1}V_{T'+1}$. By continuity of $p_{Y_{T',t}}$,

$$\|y_1-y_2\|\le\delta(T',\epsilon)\implies|p_{Y_{T',t}}(y_1)-p_{Y_{T',t}}(y_2)|\le\epsilon.$$

Consequently, we have

$$\begin{aligned}&\left|p_{Y_{T'+1,t}}(y_1)-p_{Y_{T'+1,t}}(y_2)\right|\\&=\left|\int_{\mathcal{V}}p_{Y_{T',t}}(y_1-v)p_V(v)\,\mathrm{d}v-\int_{\mathcal{V}}p_{Y_{T',t}}(y_2-v)p_V(v)\,\mathrm{d}v\right|\\&\le\int_{\mathcal{V}}\left|p_{Y_{T',t}}(y_1-v)-p_{Y_{T',t}}(y_2-v)\right|p_V(v)\,\mathrm{d}v\\&\le\int_{\mathcal{V}}\epsilon p_V(v)\,\mathrm{d}v\\&\le\epsilon.\end{aligned}$$

The a.e.c. property then follows by induction on T'. Combining the pointwse convergence with the continuity of the logarithm function and Bayes' Theorem, we have that for any $\epsilon>0$, $n\in\mathbb{N}$, and sequence y^n, there exist $T_0\in\mathbb{N}$ such that for any $T'\ge T_0$ and $t\in\mathbb{Z}$

$$|\log p_{Y_0|Y_{-t}^{-1}}(y_t|y^{t-1})-\log p_{Y_{T',0}|Y_{T',-t}^{-1}}(y_t|y^{t-1})|\le\epsilon,$$

implying that

$$\sup_{t=1\dots n}|\log p_{Y_0|Y_{-t}^{-1}}(y_t|y^{t-1})-\log p_{Y_{T',0}|Y_{T',-t}^{-1}}(y_t|y^{t-1})|\le\epsilon.$$

This concludes the first part of the proof.

To avoid heavy notation, denote $l_t(y^t)=\log p_{Y_{T',0}|Y_{T',-t}^{-1}}(y_t|y^{t-1})$. To show that the mapping $y^n\mapsto(l_t(y^t))_{t=1}^n$ can be represented as a dynamic system (Definition 5), we will first show the process $(Y_{T',t})_{t\in\mathbb{Z}}$ is a hidden Markov model (HMM)

Definition 6 (Hidden Markov Model): A process $\mathbb{Y}$ is an HMM if there exist a Markov process $(S_t)_{t\in\mathbb{Z}}$ such that $Y_t=f(S_t)$ for some measurable function f.

Let $S_t:=V_{t-T'}^{t}$. For any $t\in\mathbb{Z}$, Y_t is a linear function of S_t, which leaves us with showing that $(S_t)_{t\in\mathbb{Z}}$ is a Markov process. Indeed, by the mutual Independence of $(V_t)_{t\in\mathbb{Z}}$, we have

$$\begin{aligned}P_{S_t|S^{t-1}}(S_t|S^{t-1})&=P_{V_{t-T'}^{t}|V^{t-1}}(V_t,\dots,V_{t-T'}|V^{t-1})\\&=P_{V_{t-T'}^{t}}(V_t,\dots,V_{t-T'}|V_{t-T'-1}^{t-1})\\&=P_{S_t|S^{t-1}}(S_t|S_{t-1}),\end{aligned}$$

therefore the process $(S_t)_{t\in\mathbb{Z}}$ is Markov,. Consequently $\mathbb{Y}_{T'}$ is an HMM as required. Finally, observe the conditional PDF $p_{Y_{T',t}|Y_{T'}^{t-1}}$:

$$\begin{aligned}&p_{Y_{T',t}|Y_{T'}^{t-1}}(y_t|y^{t-1})\\&=\int_{\mathcal{S}}p_{Y_{T',t}|S_t,Y_{T'}^{t-1}}(y_t|s_t,y^{t-1})p_{S_t|Y_{T'}^{t-1}}(s_t|y^{t-1})\,\mathrm{d}s_t &(84)\\&=\int_{\mathcal{S}}\delta(\omega_1)p_{S_t|Y_{T'}^{t-1}}(s_t|y^{t-1})\,\mathrm{d}s_t &(85)\\&=\int_{\mathcal{S}}\int_{\mathcal{S}}\delta(\omega_1)p_{S_{t-1}|Y_{T'}^{t-1}}(s_{t-1}|y^{t-1})\\&\qquad\times p_{S_t|S_{t-1},Y_{T'}^{t-1}}(s_t|s_{t-1}y^{t-1})\,\mathrm{d}s_t\,\mathrm{d}s_{t-1} &(86)\end{aligned}$$

$$= \int_{\mathcal{S}} \int_{\mathcal{S}} \delta(\omega_1) p_{S_{t-1}|Y_{T'}^{t-1}}(s_{t-1}|y^{t-1}) \times p_{S_t|S_{t-1}}(s_t|s_{t-1})\, \mathrm{d}s_t\, \mathrm{d}s_{t-1} \quad (87)$$

$$= \int_{\mathcal{X}} \int_{\mathcal{S}} \delta(\omega_1) p_{S_{t-1}|Y_{T',t-1}}(s_{t-1}|y_{t-1}) \times p_{S_t|S_{t-1}}(s_t|s_{t-1})\, \mathrm{d}s_t\, \mathrm{d}s_{t-1} \quad (88)$$

$$= \int_{\mathcal{S}} \int_{\mathcal{S}} \delta(\omega_1)\delta(\omega_2) p_{S_t|S_{t-1}}(s_t|s_{t-1})\, \mathrm{d}s_t\, \mathrm{d}s_{t-1}, \quad (89)$$

where (84), (86) and (87) follow from the chain rule, (85) and (89) follow from the functional relation between the elements of $(S_t)_{t\in\mathbb{Z}}$ and $(Y_{T',t})_{t\in\mathbb{Z}}$, with

$$\omega_1 := \{s \in \mathcal{S} | y_t = f(s)\}, \quad \omega_2 := \{s \in \mathcal{S} | y_{t-1} = f(s)\}, \quad (90)$$

and $\delta(\omega)$ being the delta functional of the event ω. As $\delta(\omega_1)$ and $\delta(\omega_2)$ determine (s_{t-1}, s_t) from y^{t-1} we see that $p_{Y_t|Y^{t-1}}$ and its logarithm $l_t(y^t)$ are, in fact, a function of the tuple (s_t, s_{t-1}, y_t). Consequently, set

$$z_t = (s_{t-1}, s_t, l_t(y^t)),$$

we have that z_t is determined from (z_{t-1}, y_t) via a time invariant function. This finishes the proof. □

C. Proof of Lemma 6

Let $\mathbb{X} \in \mathsf{X}_{\mathcal{S}}$ with corresponding stationary state process $\mathbb{S}$. By joint stationarity we have $P_{X_n|S_n} = P_{X|S}$ for any $n \in \mathbb{Z}$. To construct the desired relation we utilize the FRL [45, Theorem 1].

Theorem 18 (Functional representation lemma): For any pair of random variables $(X, Y) \sim P_{XY}$ (over a Polish space with a Borel probability measure) with $\mathsf{I}(X;Y) < \infty$, there exists a random variable Z independent of X such that Y can be expressed as a function $g(X, Z)$.

By Theorem 18 we know that there exist a random variable $V \sim P_V$ and a function f_{x} such that

$$X_n = f_{\mathsf{x}}(V_n, S_n). \quad (91)$$

As $P_{X_n|S_n}$ is independent of n, (91) holds for any n with the same choice of f_{x} and time-invariant distribution on V_n, i.e., define a sequence $\{V_n\}_{n\in\mathbb{Z}} \overset{i.i.d}{\sim} P_V$, we have

$$X_n = f_{\mathsf{x}}(V_n, S_n). \quad (92)$$

Let $U \sim \mathsf{Unif}[0,1]^{d_x}$ and T_V be as defined in V-B.2. By Lemma 3, $V = T_V^{-1}(U)$ for $U \sim \mathsf{Unif}([0,1]^{d_x})$. Take $W \sim P_W$ and let T_W be as in V-B.2. Lemma 3 shows that $T_W \sim \mathsf{Unif}[0,1]^d$. We therefore construct the composite function $\widetilde{f}_{\mathsf{v}} := T_V^{-1} \circ T_W : \mathcal{W} \mapsto \mathcal{V}$. By construction, $V = \widetilde{f}_{\mathsf{v}}(W)$. Plugging $\widetilde{f}_{\mathsf{v}}$ into (92), we have

$$X_i = f_{\mathsf{x}}(S_{n-1}, \widetilde{f}_{\mathsf{v}}(W_i)),$$

which completes the proof. □

D. Proof of Lemma 7

Let $\eta > 0$, fix $i \in \{1, 2, \ldots, n\}$ and let $x^{1,n}$, $x^{2,n}$ and k^n be realizations of $X^{1,n}$, $X^{2,n}$ and K^n, respectively. Let $y^{j,n}$ and $z^{j,n}$ be generated according to $x^{j,n}$ and k^n for $j = 1, 2$. Let $\Delta_{x,i}, \Delta_{z,i}$ and $\Delta_{y,i}$ be the L^1 distance of the channel inputs, states and outputs at the ith step, e.g., $\Delta_{x,i} = \|x_i^1 - x_i^2\|_1$. By the Lipschitz property of f_{y} and f_{z} and the triangle inequality, we have

$$\Delta_{y,i} \le M_y \left\|(x_i^1, z_i^1, k_i) - (x_i^2, z_i^2, k_i)\right\|_1 \le M_y\big(\Delta_{x,i} + \Delta_{z,i}\big) \quad (93a)$$

$$\Delta_{z,i} \le M_y \left\|(x_i^1, y_i^1, z_{i-1}^1) - (x_i^2, y_i^2, z_{i-1}^2)\right\|_1 \le M_z\big(\Delta_{x,i} + \Delta_{y,i} + \Delta_{z,i-1}\big). \quad (93b)$$

Combining (93a) and (93b), we obtain

$$\Delta_{z,i} \le (M_z + M_z M_y)\Delta_{x,i} + (M_z + M_z M_y)\Delta_{z,i-1}. \quad (94)$$

Recursively applying (94) yields

$$\Delta_{z,i} \le \sum_{j=0}^{i-1} (M_z + M_z M_y)^j \Delta_{x,i-j}. \quad (95)$$

Upper bound (95) with the infinite sum and assume $\max_{i=1,\ldots,n} \Delta_{x,i} \le \eta$. We have

$$\Delta_{z,i} \le \eta \sum_{j=0}^{\infty} (M_z + M_z M_y)^j = \eta \frac{1}{1 - M_z(M_y + 1)}, \quad (96)$$

where the sum converges due to Assumption 1. Plug (96) into (93a) to obtain

$$\Delta_{y,i} \le \eta \left(\frac{M_y(2 - M_z(M_y + 1))}{1 - M_z(M_y + 1)} \right), \quad (97)$$

which holds for any $i \le n$. The inequality (97) holds for any realization of $P_K^{\otimes n}$. □

ACKNOWLEDGMENT

The authors would like to thank the associate editor and anonymous reviewers for their constructive comments that helped improve the content organization of the paper.

REFERENCES

[1] Z. Aharoni, D. Tsur, Z. Goldfeld, and H. H. Permuter, "Capacity of continuous channels with memory via directed information neural estimator," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2020, pp. 2014–2019.

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

[3] M. Raginsky, "Directed information and Pearl's causal calculus," in *Proc. 49th Annu. Allerton Conf. Commun., Control, Comput. (Allerton)*, Sep. 2011, pp. 958–965.

[4] R. G. Gallager, *Information Theory and Reliable Communication*, vol. 2. New York, NY, USA: Springer, 1968.

[5] H. H. Permuter, T. Weissman, and A. J. Goldsmith, "Finite state channels with time-invariant deterministic feedback," *IEEE Trans. Inf. Theory*, vol. 55, no. 2, pp. 644–662, Feb. 2009.

[6] H. H. Permuter, Y.-H. Kim, and T. Weissman, "Interpretations of directed information in portfolio theory, data compression, and hypothesis testing," *IEEE Trans. Inf. Theory*, vol. 57, no. 6, pp. 3248–3259, Jun. 2011.

[7] R. Battiti, "Using mutual information for selecting features in supervised neural net learning," *IEEE Trans. Neural Netw.*, vol. 5, no. 4, pp. 537–550, Jul. 1994.

[8] A. J. Bell and T. J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Comput.*, vol. 7, no. 6, pp. 1129–1159, 1995.

[9] I. Higgins et al., "β-VAE: Learning basic visual concepts with a constrained variational framework," in *Proc. ICLR*, 2016, pp. 1–12.

[10] R. Shwartz-Ziv and N. Tishby, "Opening the black box of deep neural networks via information," 2017, *arXiv:1703.00810*.

[11] Z. Goldfeld et al., "Estimating information flow in deep neural networks," 2018, *arXiv:1810.05728*.

[12] A. G. Dimitrov, A. A. Lazar, and J. D. Victor, "Information theory in neuroscience," *J. Comput. Neurosci.*, vol. 30, no. 1, pp. 1–5, 2011.

[13] 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," *J. Comput. Neurosci.*, vol. 30, no. 1, pp. 17–44, 2011.

[14] M. Wibral, R. Vicente, and J. T. Lizier, *Directed Information Measures in Neuroscience*. Berlin, Germany: Springer, 2014.

[15] H. Touchette and S. Lloyd, "Information-theoretic limits of control," *Phys. Rev. Lett.*, vol. 84, no. 6, pp. 1156–1159, Feb. 2000.

[16] B. Grocholsky, "Information-theoretic control of multiple sensor platforms," School Aerosp., Mech. Mechtron., Univ. Sydney, 2002.

[17] H. Boche, R. F. Schaefer, and H. V. Poor, "Shannon meets turing: Non-computability and non-approximability of the finite state channel capacity," 2020, *arXiv:2008.13270*.

[18] A. Grigorescu, H. Boche, R. F. Schaefer, and H. V. Poor, "Capacity of finite state channels with feedback: Algorithmic and optimization theoretic properties," 2022, *arXiv:2201.11639*.

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

[20] C. J. Quinn, N. Kiyavash, and T. P. Coleman, "Directed information graphs," *IEEE Trans. Inf. Theory*, vol. 61, no. 12, pp. 6887–6909, Dec. 2015.

[21] Y. Murin, "k-NN estimation of directed information," 2017, *arXiv:1711.08516*.

[22] A. Rahimzamani, H. Asnani, P. Viswanath, and S. Kannan, "Estimators for multivariate information measures in general probability spaces," 2018, *arXiv:1810.11551*.

[23] R. B. Marimont and M. B. Shapiro, "Nearest neighbour searches and the curse of dimensionality," *IMA J. Appl. Math.*, vol. 24, no. 1, pp. 59–70, 1979.

[24] M. I. Belghazi et al., "Mutual information neural estimation," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 531–540.

[25] M. D. Donsker and S. R. S. Varadhan, "Asymptotic evaluation of certain Markov process expectations for large time. IV," *Commun. Pure Appl. Math.*, vol. 36, no. 2, pp. 183–212, Mar. 1983.

[26] B. Poole, S. Ozair, A. Van Den Oord, A. A. Alemi, and G. Tucker, "On variational lower bounds of mutual information," in *Proc. NeurIPS Workshop Bayesian Deep Learn.*, 2018, pp. 1–9.

[27] J. Song and S. Ermon, "Understanding the limitations of variational mutual information estimators," 2019, *arXiv:1910.06222*.

[28] C. Chan, A. Al-Bashabsheh, H. P. Huang, M. Lim, D. Sun Handason Tam, and C. Zhao, "Neural entropic estimation: A faster path to mutual information estimation," 2019, *arXiv:1905.12957*.

[29] Z. S. Zhang Sreekumar and Z. Goldfeld, "Non-asymptotic performance guarantees for neural estimation of f-divergences," in *Proc. Int. Conf. Artif. Intell. Statist. (AISTATS)*, vol. 130, Apr. 2021, pp. 3322–3330.

[30] S. Sreekumar and Z. Goldfeld, "Neural estimation of statistical divergences," *J. Mach. Learn. Res.*, vol. 23, no. 126, pp. 1–75, 2022.

[31] D. McAllester and K. Stratos, "Formal limitations on the measurement of mutual information," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2020, pp. 875–884.

[32] J. Zhang, O. Simeone, Z. Cvetkovic, E. Abela, and M. Richardson, "ITENE: Intrinsic transfer entropy neural estimator," 2019, *arXiv:1912.07277*.

[33] S. Molavipour, H. Ghourchian, G. Bassi, and M. Skoglund, "Neural estimator of information for time-series data with dependency," *Entropy*, vol. 23, no. 6, p. 641, May 2021.

[34] H. H. Permuter, P. Cuff, B. V. Roy, and T. Weissman, "Capacity of the trapdoor channel with feedback," *IEEE Trans. Inf. Theory*, vol. 54, no. 7, pp. 3150–3165, Jul. 2009.

[35] O. Elishco and H. Permuter, "Capacity and coding for the Ising channel with feedback," *IEEE Trans. Inf. Theory*, vol. 60, no. 9, pp. 5138–5149, Sep. 2014.

[36] Z. Aharoni, O. Sabag, and H. H. Permuter, "Feedback capacity of Ising channels with large alphabet via reinforcement learning," *IEEE Trans. Inf. Theory*, vol. 68, no. 9, pp. 5637–5656, Sep. 2022.

[37] R. E. Blahut, "Computation of channel capacity and rate-distortion functions," *IEEE Trans. Inf. Theory*, vol. IT-18, no. 4, pp. 460–473, Jul. 1972.

[38] S. Arimoto, "An algorithm for computing the capacity of arbitrary discrete memoryless channels," *IEEE Trans. Inf. Theory*, vol. IT-18, no. 1, pp. 14–20, Jan. 1972.

[39] P. O. Vontobel, A. Kavcic, D. M. Arnold, and H.-A. Loeliger, "A generalization of the Blahut–Arimoto algorithm to finite-state channels," *IEEE Trans. Inf. Theory*, vol. 54, no. 5, pp. 1887–1918, May 2008.

[40] I. Naiss and H. H. Permuter, "Extension of the Blahut–Arimoto algorithm for maximizing directed information," *IEEE Trans. Inf. Theory*, vol. 59, no. 1, pp. 204–222, Jan. 2013.

[41] J. Dauwels, "Numerical computation of the capacity of continuous memoryless channels," in *Proc. 26th Symp. Inf. Theory BENELUX*, 2005, pp. 221–228.

[42] L. Breiman, "The individual ergodic theorem of information theory," *Ann. Math. Statist.*, vol. 28, no. 3, pp. 809–811, 1957.

[43] A. M. Schäfer and H. G. Zimmermann, "Recurrent neural networks are universal approximators," in *Proc. Int. Conf. Artif. Neural Netw.* Cham, Switzerland: Springer, 2006, pp. 632–640.

[44] A. El Gamal and Y. H. Kim, *Network Information Theory*. Cambridge, U.K.: Cambridge Univ. Press, 2011.

[45] C. T. Li and A. El Gamal, "Strong functional representation lemma and applications to coding theorems," *IEEE Trans. Inf. Theory*, vol. 64, no. 11, pp. 6967–6978, Nov. 2018.

[46] L. H. Ozarow and A. D. Wyner, "On the capacity of the Gaussian channel with a finite number of input levels," *IEEE Trans. Inf. Theory*, vol. 36, no. 6, pp. 1426–1428, Nov. 1990.

[47] M. Raginsky, "On the information capacity of Gaussian channels under small peak power constraints," in *Proc. 46th Annu. Allerton Conf. Commun., Control, Comput.*, Sep. 2008, pp. 286–293.

[48] A. Thangaraj, G. Kramer, and G. Böcherer, "Capacity bounds for discrete-time, amplitude-constrained, additive white Gaussian noise channels," *IEEE Trans. Inf. Theory*, vol. 63, no. 7, pp. 4172–4182, Apr. 2017.

[49] S. Yang, A. Kavčić, and S. C. Tatikonda, "On the feedback capacity of power-constrained Gaussian noise channels with memory," *IEEE Trans. Inf. Theory*, vol. 53, no. 3, pp. 929–954, Mar. 2007.

[50] O. Sabag, V. Kostina, and B. Hassibi, "Feedback capacity of MIMO Gaussian channels," 2021, *arXiv:2106.01994*.

[51] F. Mirkarimi and N. Farsad, "Neural computation of capacity region of memoryless multiple access channels," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2021, pp. 2066–2071.

[52] N. A. Letizia and A. M. Tonello, "Capacity-driven autoencoders for communications," *IEEE Open J. Commun. Soc.*, vol. 2, pp. 1366–1378, 2021.

[53] N. A. Letizia and A. M. Tonello, "Discriminative mutual information estimators for channel capacity learning," 2021, *arXiv:2107.03084*.

[54] F. Mirkarimi, S. Rini, and N. Farsad, "A perspective on neural capacity estimation: Viability and reliability," 2022, *arXiv:2203.11793*.

[55] G. Kramer, *Directed Information for Channels With Feedback*, vol. 11. Princeton, NJ, USA: Citeseer, 1998.

[56] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. New York, NY, USA: Wiley, 2006.

[57] R. L. Dobrushin, "General formulation of Shannon's main theorem in information theory," *Amer. Math. Soc. Trans.*, vol. 33, no. 2, pp. 323–438, 1963.

[58] S. Tatikonda and S. Mitter, "The capacity of channels with feedback," *IEEE Trans. Inf. Theory*, vol. 55, no. 1, pp. 323–349, Jan. 2009.

[59] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural Netw.*, vol. 2, no. 5, pp. 359–366, Dec. 1989.

[60] L. Jin, M. M. Gupta, and P. N. Nikiforuk, "Universal approximation using dynamic recurrent neural networks: Discrete-time version," in *Proc. Int. Conf. Neural Netw. (ICNN)*, 1995, pp. 403–408.

[61] S. Molavipour, "Statistical inference of information in networks: Causality and directed information graphs," Ph.D. thesis, KTH Royal Inst. Technol., Stockholm, Sweden, 2021.

[62] H. Wold, "A study in the analysis of stationary time series," Ph.D. thesis, Almqvist & Wiksell, Stockholm, Sweden, 1938.

[63] K. Urbanik, "Prediction of strictly stationary sequences," in *Colloquium Mathematicum*, vol. 12. Warsaw, Poland: Instytut Matematyczny Polskiej Akademii Nauk, 1964, pp. 115–129.

[64] N. Wiener and P. Masani, "The prediction theory of multivariate stochastic processes," *Acta Math.*, vol. 98, no. 1, pp. 111–150, 1957.
[65] P. H. Algoet and T. M. Cover, "A sandwich proof of the Shannon–McMillan–Breiman theorem," *Ann. Probab.*, vol. 16, no. 2, pp. 899–909, Apr. 1988.
[66] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
[67] J. K. Møller, "Stochastic state space modelling of nonlinear systems-with application to marine ecosystems," Tech. Univ. Denmark, 2011.
[68] J. Ziv, "Universal decoding for finite-state channels," *IEEE Trans. Inf. Theory*, vol. IT-31, no. 4, pp. 453–460, Jul. 1985.
[69] Y.-H. Kim, "A coding theorem for a class of stationary channels with feedback," *IEEE Trans. Inf. Theory.*, vol. 54, no. 4, pp. 1488–1499, Apr. 2008.
[70] A. Reza Pedram and T. Tanaka, "Some results on the computation of feedback capacity of Gaussian channels with memory," in *Proc. 56th Annu. Allerton Conf. Commun., Control, Comput. (Allerton)*, Oct. 2018, pp. 919–926.
[71] M. Rosenblatt, "Remarks on a multivariate transformation," *Ann. Math. Statist.*, vol. 23, no. 3, pp. 470–472, 1952.
[72] H. Knöthe, "Contributions to the theory of convex bodies," *Michigan Math. J.*, vol. 4, no. 1, pp. 39–52, 1957.
[73] V. I. Bogachev, A. V. Kolesnikov, and K. V. Medvedev, "Triangular transformations of measures," *Sbornik, Math.*, vol. 196, no. 3, pp. 309–335, Apr. 2005.
[74] C. Huang, D. Krueger, A. Lacoste, and A. Courville, "Neural autoregressive flows," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 2078–2087.
[75] A. Spantini, D. Bigoni, and Y. Marzouk, "Inference via low-dimensional couplings," *J. Mach. Learn. Res.*, vol. 19, no. 1, pp. 2639–2709, 2018.
[76] A. Hyvärinen and P. Pajunen, "Nonlinear independent component analysis: Existence and uniqueness results," *Neural Netw.*, vol. 12, no. 3, pp. 429–439, Apr. 1999.
[77] F. Santambrogio, *Optimal Transport for Applied Mathematicians*, vol. 55, nos. 58–63. Basel, Switzerland: Birkäuser, 2015, p. 94.
[78] T. Cui and S. Dolgov, "Deep composition of tensor-trains using squared inverse rosenblatt transports," *Found. Comput. Math.*, vol. 22, no. 6, pp. 1–60, 2021.
[79] G. Papamakarios, E. Nalisnick, D. J. Rezende, S. Mohamed, and B. Lakshminarayanan, "Normalizing flows for probabilistic modeling and inference," *J. Mach. Learn. Res.*, vol. 22, no. 57, pp. 1–64, 2021.
[80] Y. Polyanskiy and Y. Wu, "Wasserstein continuity of entropy and outer bounds for interference channels," *IEEE Trans. Inf. Theory*, vol. 62, no. 7, pp. 3992–4002, Jul. 2016.
[81] M. Godavarti and A. Hero, "Convergence of differential entropies," *IEEE Trans. Inf. Theory*, vol. 50, no. 1, pp. 171–176, Jan. 2004.
[82] I. Goodfellow et al., "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014, pp. 1–14.
[83] V. R. Konda and J. N. Tsitsiklis, "Actor-critic algorithms," in *Proc. Adv. Neural Inf. Process. Syst.*, 2000, pp. 1008–1014.
[84] M. Abadi et al., "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015, *arXiv:1603.04467*.
[85] Y.-H. Kim, "Feedback capacity of stationary Gaussian channels," *IEEE Trans. Inf. Theory*, vol. 56, no. 1, pp. 57–85, Jan. 2010.
[86] Y. Polyanskiy and Y. Wu, "Lecture notes on information theory," *Lecture Notes ECE563*, vol. 6, p. 7, Feb. 2014.
[87] J. Neveu and T. P. Speed, *Discrete-Parameter Martingales*, vol. 10. Hoboken, NJ, USA: Wiley, 1975.
[88] H. B. Mann and A. Wald, "On stochastic limit and order relationships," *Ann. Math. Statist.*, vol. 14, no. 3, pp. 217–226, Sep. 1943.
[89] S. Nietert, Z. Goldfeld, and K. Kato, "Smooth p-Wasserstein distance: Structure, empirical approximation, and statistical applications," in *Proc. 38th Int. Conf. Mach. Learn.*, 2021, pp. 8172–8183.
[90] Z. Goldfeld, K. H. Greenewald, J. Niles-Weed, and Y. Polyanskiy, "Convergence of smoothed empirical measures with applications to entropy estimation," *IEEE Trans. Inf. Theory*, vol. 66, no. 7, pp. 1489–1501, Feb. 2020.
[91] Z. Goldfeld and K. Greenewald, "Gaussian-smoothed optimal transport: Metric structure and statistical efficiency," in *Proc. 23rd Int. Conf. Artif. Intell. Statist.*, 2020, pp. 3327–3337.
[92] Z. Goldfeld, K. Greenewald, and K. Kato, "Asymptotic guarantees for generative modeling based on the smooth Wasserstein distance," in *Proc. 34th Int. Conf. Neural Inf. Process. Syst.*, 2020, pp. 2527–2539.
[93] R. Sadhu, Z. Goldfeld, and K. Kato, "Limit distribution theory for the smooth 1-Wasserstein distance with applications," 2021, *arXiv:2107.13494*.
[94] Z. Goldfeld, K. Kato, S. Nietert, and G. Rioux, "Limit distribution theory for smooth p-Wasserstein distances," 2022, *arXiv:2203.00159*.
[95] C. Villani, *Topics in Optimal Transportation*. Washington, DC, USA: American Mathematical Soc., 2021.
[96] Z. Goldfeld, K. Greenewald, T. Nuradha, and G. Reeves, "k-sliced mutual information: A quantitative study of scalability with dimension," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2022, pp. 1–41.
[97] T. J. Sweeting, "On a converse to Scheffé's theorem," *Ann. Statist.*, vol. 14, no. 3, pp. 1252–1256, 1986.

Dor Tsur (Student Member, IEEE) received the B.Sc. (cum laude) and M.Sc. (summa cum laude) degrees in electrical and computer engineering from the Ben-Gurion University of the Negev, Israel, in 2020 and 2023, respectively, where he is currently pursuing the Ph.D. degree in electrical and computer engineering. His research interests include machine learning, information theory, and statistical signal processing.

Ziv Aharoni (Student Member, IEEE) received the B.Sc. and M.Sc. degrees in electrical and computer engineering from the Ben-Gurion University of the Negev, Israel, in 2017 and 2020, respectively, where he is currently pursuing the Ph.D. degree in electrical and computer engineering. His research interests include information theory and machine learning.

Ziv Goldfeld (Member, IEEE) received the B.Sc., M.Sc., and Ph.D. degrees in electrical and computer engineering from the Ben-Gurion University of the Negev, Israel, in 2012, 2014, and 2017, respectively. From 2017 to 2019, he was a Post-Doctoral Fellow with the Laboratory for Information and Decision Systems (LIDS), MIT. He is currently an Assistant Professor in electrical and computer engineering with Cornell University. He was a recipient of several awards, such as the 2020 NSF CRII Award, the 2020 IBM Academic Award, the Rothschild Postdoctoral Fellowship, and the Best Student Paper Award in the IEEE 28th Convention of Electrical and Electronics Engineers in Israel.

Haim Permuter (Senior Member, IEEE) received the B.Sc. and M.Sc. degrees (summa cum laude) in electrical and computer engineering from the Ben-Gurion University of the Negev, Israel, in 1997 and 2003, respectively, and the Ph.D. degree in electrical engineering from Stanford University, CA, USA, in 2008. From 1997 to 2004, he was an Officer with a research and development unit of the Israeli Defense Forces. Since 2009, he has been with the Department of Electrical and Computer Engineering, Ben-Gurion University of the Negev, where he is currently a Professor and the Luck-Hille Chair of Electrical Engineering and also the Head of Communication, Cyber, and Information Track in his department. He was a recipient of several awards, among them are the Fullbright Fellowship, the Stanford Graduate Fellowship (SGF), Allon Fellowship, and the U.S. Israel Binational Science Foundation Bergmann Memorial Award. He served on the Editorial Board of the IEEE TRANSACTIONS ON INFORMATION THEORY from 2013 to 2016.