Empirical Coordination over Markov Channel with Independent Source

Mengyuan Zhao*, Maël Le Treust†, Tobias J. Oechtering*

*KTH Royal Institute of Technology, 100 44 Stockholm, Sweden, Email: {mzhao, oech}@kth.se
†CNRS, University of Rennes, Inria, IRISA UMR 6074, F-35000 Rennes, France, Email: mael.le-treust@cnrs.fr

Abstract—We study joint source–channel coding over Markov channels through the empirical coordination framework. More specifically, we aim at determining the empirical distributions of source and channel symbols that can be induced by a coding scheme. We consider strictly causal encoders that generate channel inputs, without access to the past channel states, henceforth driving the current Markov state evolution. Our main result is the single-letter inner and outer bounds of the set of achievable joint distributions, coordinating all the symbols in the network. To establish the inner bound, we introduce a new notion of typicality, the input-driven Markov typicality, and develop its fundamental properties. Contrary to the classical block-Markov coding schemes that rely on blockwise independence for discrete memoryless channels, our analysis directly exploits the Markov channel structure and improves beyond the independence-based arguments.

Index Terms—joint source-channel coding, Markov channel, coordination coding, typicality.

I. INTRODUCTION

Markov chains, as canonical models of stochastic recursion, are often represented as being driven by exogenous independent randomness [1]. This perspective casts the evolution as a state update rule acted on by fresh external randomness at each step, unifying formulations across random dynamical systems and stochastic algorithms. When combined with structural conditions such as monotonicity or contractivity, it leads to tractable analyses of stationarity and convergence [2], [3]. From a control perspective, such exogenous inputs correspond to the classical open-loop operation [4]–[6].

In information theory literature, Markov channels, sometimes known as finite state channels (FSC), have been studied extensively through the lens of channel capacity and control mechanisms [7]–[11]. When feedback is available, the encoder observes the channel state and can adapt its inputs accordingly, leading to a rich body of results on FSCs with feedback [12]–[17]. These capacity characterizations, however, are complicated and typically involve multi-letter expressions. Recently, under the assumption of unichain and ergodicity, a single-letter capacity representation was obtained for Markov channels with feedback [18].

In this work, we study the Markov evolution structure from a joint source–channel coding (JSCC) perspective [19], [20]. Specifically, the source sequence is generated independently at
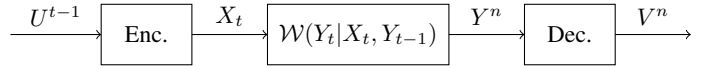
Fig. 1: Joint source-channel coding with strictly causal encoder and noncausal decoder over Markov channel

each time, and the channel input is generated by strictly causal encoders based solely on the past source sequence, and does not adapt to the latent channel state history. The channel input, together with the previous state, updates the current channel state. The decoder operates noncausally, observing the entire state sequence to produce its action sequence.

Empirical coordination coding framework, introduced in [21]–[23], characterizes cooperative behavior among agents by requiring the empirical distribution of symbol sequences to converge to a prescribed target distribution. In this framework, encoders and decoders select their actions based on local observations not only to ensure reliable communication, but also to induce a desired joint behavior. Coordination results for JSCC with (strictly) causal decision-makers over discrete memoryless channel (DMC) were established in [24], [25]. Extensions to point-to-point settings with channel state or feedback were developed in [26], [27], and recently have been further applied to distributed decision-making [28]–[30] etc..

To the best of our knowledge, this work is the first to study JSCC over Markov channels through the lens of coordination coding. Under our Markov channel model assumption, we derive single-letter inner and outer bounds that characterize the set of achievable joint distributions that can be arbitrarily well approximated by suitable coding schemes. To establish the inner bound, we introduce a new notion of typicality, the input-driven Markov typicality, and develop its fundamental properties, such as asymptotic equipartition property (AEP) and a packing lemma, etc.. In contrast to the standard random-coding arguments for DMC that rely on block independence in the block-Markov coding constructions, such as in [27], [31], this notion of typicality enables us to exploit the Markov channel structure directly, allowing the analysis to go beyond independence-based arguments while still recovering the results. The outer bound shares the same information constraint expression as the inner bound, but is defined over a generally larger set of distributions.

This paper is structured as follows: Our problem is formulated in Section II. Section III presents our main results,

which comprise the inner and outer bounds for the achievable distribution region. In the Appendices, we introduce the input-driven Markov typicality and prove the main results.

II. System Model

We study the problem depicted in Fig. 1. Capital letters, like U, denote random variables, calligraphic fonts, like $\mathcal{U}$, denote alphabets, and lowercase letters, like $u \in \mathcal{U}$, denote the realizations of random variables. We assume all the alphabets considered here $\mathcal{U}, \mathcal{X}, \mathcal{Y}, \mathcal{V}$ are finite. The ℓ_1 distance between two probability distributions $\mathcal{Q}$ and $\mathcal{P}$ is denoted by $\|\mathcal{Q} - \mathcal{P}\|_1 = \sum_u |\mathcal{Q}(u) - \mathcal{P}(u)|$. The probability is denoted by $\mathbb{P}(\cdot)$.

We assume the source is drawn i.i.d. $\sim \mathcal{P}_U$, and the channel is stationary Markov with one step memory of the latent state, with conditional distribution $\mathcal{W}_{Y|X,Y'}$. Moreover, the statistics of the source and channel are known by both agents.

Definition 1. *A code with strictly causal encoder and non-causal decoder is a tuple of (deterministic) functions $c = (\{f^{(t)}_{X_t|U^{t-1}}\}_{t=1}^n, g_{V^n|Y^n})$ defined by*

$$f^{(t)}_{X_t|U^{t-1}} : \mathcal{U}^{t-1} \longrightarrow \mathcal{X}, \quad g_{V^n|Y^n} : \mathcal{Y}^n \longrightarrow \mathcal{V}^n, \quad (1)$$

where $U_0 \triangleq \varnothing$. We denote by $\mathcal{C}(n)$ the set of codes with strictly causal encoder and non-causal decoder.

We assume the initial channel state Y_0 is arbitrary. Then, a code $c \in \mathcal{C}(n)$ induces a joint distribution on sequences (U^n, X^n, Y^n, V^n) given by

$$\mathcal{P}_U^{\otimes n} \prod_{t=1}^n \left[f^{(t)}_{X_t|U^{t-1}} \cdot \mathcal{W}_{Y_t|X_t,Y_{t-1}} \right] g_{V^n|Y^n}, \quad (2)$$

from which we have the following:

$$U_t \perp\!\!\!\perp (X_t, Y_t), \quad (3)$$
$$X_t \;\multimap\!\!\!- U^{t-1} -\!\!\!\multimap\; Y^{t-1}, \quad (4)$$
$$Y_t \;\multimap\!\!\!- (X_t, Y_{t-1}) -\!\!\!\multimap\; (U^{t-1}, X^{t-1}, Y^{t-2}), \quad (5)$$

where (3) comes from the independent source and strictly causal encoding, (4) is our strictly causal encoder, and (5) is the Markov channel.

We denote the empirical distribution $Q^n \in \Delta(\mathcal{U} \times \mathcal{X} \times \mathcal{Y} \times \mathcal{Y} \times \mathcal{V})$ of sequences (u^n, x^n, y^n, v^n) by

$$Q^n(u, x, y', y, v) \quad (6)$$
$$= \frac{1}{n} \sum_{t=1}^n \mathbf{1}\{u_t = u, x_t = x, y_{t-1} = y', y_t = y, v_t = v\},$$
$$\forall (u, x, y', y, v) \in \mathcal{U} \times \mathcal{X} \times \mathcal{Y} \times \mathcal{Y} \times \mathcal{V}.$$

Note here, we take into account the adjacent pairs (y_{t-1}, y_t) in the Markov sequence y^n.

Definition 2. *We define a joint distribution $\mathcal{P}_{U,X,Y',Y,V} \in \Delta(\mathcal{U} \times \mathcal{X} \times \mathcal{Y} \times \mathcal{Y} \times \mathcal{V})$ to be achievable, if for all $\varepsilon \geq 0$, there exists $N_0 \in \mathbb{N}$, such that for all $n \geq N_0$, we can find*

a sequence of codes $c \in \mathcal{C}(n)$, such that the joint empirical distribution given by (6) satisfying that

$$\mathbb{P}(\||Q^n - \mathcal{P}_{U,X,Y',Y,V}\||_1 \geq \varepsilon) \leq \varepsilon. \quad (7)$$

We impose the following assumption:

Assumption A. *For every stationary channel input distribution $\mathcal{P}_X$ on $\mathcal{X}$, the induced Markov chain $Y^n = \{Y_1, \ldots, Y_n\}$ admits a unique recurrent class $\mathcal{R}_Y \subseteq \mathcal{Y}$ that is irreducible and aperiodic.*

The above assumption is standard, see [18], [32], [33], [34, Sec. 4.1.2]. This assumption guarantees that the long-run behavior of the Markov chain Y^n does not depend on its starting point and is ergodic within $\mathcal{R}_Y$. Moreover, it ensures the following result:

Lemma 1. *The Markov chain Y^n admits a unique equilibrium distribution $\pi_Y \in \Delta(\mathcal{R}_Y)$, which is uniquely determined by the stationary input distribution $\mathcal{P}_X$ and the channel transition kernel $\mathcal{W}_{Y|X,Y'}$ by*

$$\pi_Y(y) = \sum_{x,y'} \pi_Y(y') \mathcal{P}_X(x) \mathcal{W}_{Y|X,Y'}(y|x, y'). \quad (8)$$

Next, we present our main result: the single-letter characterization of the achievable distribution region, defined in Definition 2.

III. Main Results

Our main result provides necessary and sufficient conditions for a distribution that is achievable.

Theorem 1 (Main Result). *A joint probability distribution $\mathcal{P}_{U,X,Y',Y,V}$ is achievable, if there exists an auxiliary random variable W, satisfying*

$$\mathcal{P}_U \mathcal{P}_X \mathcal{P}_{W|U,X} \pi_{Y'} \mathcal{W}_{Y|X,Y'} \mathcal{P}_{V|Y,X,W} \quad (9)$$

with marginal $\mathcal{P}_{U,X,Y',Y,V}$ such that we have

$$I(X; Y|Y') - I(U; W|X) \geq 0, \quad (10)$$

where $\pi_{Y'}$ is the unique equilibrium distribution (8). Conversely, $\mathcal{P}_{U,X,Y',Y,V}$ is achievable only if there exists an auxiliary random variable W satisfying (10) and

$$\mathcal{P}_U \mathcal{P}_X \mathcal{P}_{Y'|X} \mathcal{W}_{Y|X,Y'} \mathcal{P}_{W|U,X,Y,Y'} \mathcal{P}_{V|Y,X,W} \quad (11)$$

with marginal $\mathcal{P}_{U,X,Y',Y,V}$.

The first part of Theorem 1 is an inner bound of the achievable distribution region, whereas the second part is the outer bound.

When the Markov channel reduces to a DMC (e.g. take $Y' = \varnothing$), our inner bound specializes to the coordination coding result for joint source–channel coding with strictly causal encoders established in [24, Theorem 3].

The auxiliary random variable W plays a role analogous to that in coordination coding for DMC and channels with state [24], [27], [31]: it enables the encoder to convey compressed source information to the decoder in order to create the desired

coordination. Although the inner and outer bounds share the same information constraint (10), the outer bound is generally looser, as it is defined over a larger class of joint distributions.

Remark 1 (Source-Channel Separation). *In the case when the random variables of the source (U,V) are independent of those of the channel (X, Y', Y), i.e.,*

$$\mathcal{P}_{U,X,Y',Y,V} = \mathcal{P}_{U,V} \cdot \mathcal{P}_{X,Y',Y},$$

the information constraint (10) *reduces to a form analogous to Shannon's source–channel separation theorem* [19]. *In particular,*

$$0 \leq I(X;Y|Y') - I(U;W|X)$$
$$\overset{(a)}{\leq} I(X;Y|Y') - I(U;V|X)$$
$$= I(X;Y|Y') - I(U;V)$$

where the inequality (a) comes from the Markov chain $V -\!\!\!\circ\!\!\!- (X, Y, W) -\!\!\!\circ\!\!\!- (U, Y')$ at the noncausal decoder and the data processing inequality. Inequality (a) becomes equality if and only if we take $W = U$. The resulting expression consists the channel capacity expression for Markov channel (but with feedback) derived in [18], minus the source rate required at the decoder.

To establish the inner bound, we introduce a new notion of typicality – the input-driven Markov typicality – for sequences (X^n, Y^n) in Appendix A. The proof of the inner bound is a random-coding achievability argument, given in the Appendix B. Finally, the outer bound is proved in Appendix C.

APPENDIX A
INPUT-DRIVEN MARKOV TYPICALITY

In this section, we introduce the input-driven Markov typicality for two sequences (X^n, Y^n), as an extension to the strong Markov typicality for a single Markov sequence analyzed in [34]–[36]. We consider $X^n \sim \mathcal{P}_X^{\otimes n}$ i.i.d. generated, and Y^n a Markov chain with starting state Y_0 arbitrary, $Y_t \sim \mathcal{W}_{Y|X,Y'}(\cdot|X_t, Y_{t-1})$ conditioning on X_t for every $t = 1, ..., n$. In other words,

$$(X^n, Y^n) \sim \mathcal{P}_X^{\otimes n} \cdot \prod_{t=1}^{n} \mathcal{W}(Y_t|X_t, Y_{t-1}). \tag{12}$$

Now, for simplicity, we denote the target product distribution by $\mathcal{Q}_{Y'XY} = \pi_{Y'} \mathcal{P}_X \mathcal{W}_{Y|X,Y'}$.

Definition 3 (Input-driven Markov Typicality). *For sequences (x^n, y^n), define the empirical distribution of joint symbols over adjacent indices of triplets:*

$$Q_{Y'XY}^n(i,x,j) = \frac{1}{n} \sum_{t=1}^{n} \mathbf{1}\{(y_{t-1}, x_t, y_t) = (i,x,j)\},$$
$$x \in \mathcal{X}, (i,j) \in \mathcal{Y}^2.$$

We define the set of input-driven Markov typicality sequences by

$$\mathcal{T}_\varepsilon^{(n)}(\mathcal{Q}_{Y'XY}) = \{(x^n, y^n) : \|Q_{Y'XY}^n - \mathcal{Q}_{Y'XY}\|_1 \leq \varepsilon\}.$$

For a fixed $x^n \in \mathcal{X}^n$, the set of conditionally Markov typical sequences $\mathcal{T}_\varepsilon^{(n)}(\mathcal{Q}_{Y'XY}|x^n)$ is defined by

$$\mathcal{T}_\varepsilon^{(n)}(\mathcal{Q}_{Y'XY}|x^n) = \{y^n \in \mathcal{Y}^n :$$
$$\|Q_{Y'XY}^n - \pi_{Y'}Q_X^n \mathcal{W}_{Y|X,Y'}\|_1 \leq \varepsilon\}.$$

where

$$Q_X^n(x) = \frac{1}{n} \sum_{t=1}^{n} \mathbf{1}\{x_t = x\} = \sum_{i,j \in \mathcal{Y}} Q_{Y'XY}^n(i,x,j), \forall x \in \mathcal{X}.$$

We have the following theorems for the input-driven Markov typicality:

Theorem 2 (Ergodicity). *Let (X^n, Y^n) generated according to* (12), *then*

$$\forall \delta > 0, \lim_{n \to \infty} \mathbb{P}((X^n, Y^n) \in \mathcal{T}_\delta^{(n)}(\mathcal{Q}_{Y'XY})) = 1 \tag{13}$$

This theorem shows that no matter how the initial symbol of is generated, the Markov type of X^n, Y^n always converges to their joint stationary equilibrium distribution $\mathcal{Q}_{Y'XY}$.

We denote $H(X)$ by the entropy for random variable $X \sim \mathcal{P}_X$ and $H(Y|X, Y')$ for jointly discrete random variables $(Y', X, Y) \sim \mathcal{Q}_{Y'XY}$ which quantifies the remaining randomness in Y given the independent input X and the past symbol Y' in the Markov chain. We have

Theorem 3 (AEP and Cardinality Bound). *For any $\delta > 0$, there exists $\varepsilon_0 > 0, N_0 \in \mathbb{N}$, such that $\forall \varepsilon < \varepsilon_0, n > N_0$, and $\forall (x^n, y^n) \in \mathcal{T}_\varepsilon^{(n)}(\mathcal{Q}_{Y'XY})$, if a random sequence (X^n, Y^n) is generated according to* (12), *then it satisfies that*

1) *AEP:* $2^{-n(H(X,Y|Y')+\delta)} < \mathbb{P}((X^n, Y^n) = (x^n, y^n))$
$$< 2^{-n(H(X,Y|Y')-\delta)},$$

2) *Cardinality bound:* $\left|\mathcal{T}_\varepsilon^{(n)}(\mathcal{Q}_{Y'XY})\right| < 2^{n(H(X,Y|Y')+\delta)}.$

More properties of the input-driven typicality and the proofs for the above theorems are given in the supplementary material.

APPENDIX B
ACHIEVABILITY PROOF OF THEOREM 1

The coding argument relies on the block-Markov coding scheme [31], extended to the coordination framework [27].

We consider a probability distribution $\mathcal{P} = \mathcal{P}_{U,X,Y',Y,V}$ decomposed as (9) that satisfies (10) and (8). There exists a $\delta > 0$ and rate $R > 0$ such that

$$R \geq I(U;W|X) + \delta, \tag{14}$$
$$R \leq I(X;Y|Y') - \delta. \tag{15}$$

We consider a block-Markov random code $c \in \mathcal{C}(nB)$ over $B \in \mathbb{N}$ blocks of length $n \in \mathbb{N}$.

Random codebook: We generate $|\mathcal{M}| = 2^{nR}$ sequences $X^n(m)$, where each drawn from the i.i.d. distribution $\mathcal{P}_X^{\otimes n}$ with index $m \in \mathcal{M}$. For each $m \in \mathcal{M}$, we generate the same number $|\mathcal{M}| = 2^{nR}$ of sequences $W^n(m, \hat{m})$ with index $\hat{m} \in \mathcal{M}$, drawn from the conditional distribution $\mathcal{P}_{W|X}^{\otimes n}$ conditioning on the sequence $X^n(m)$.

Encoding function: Let m_b denote the message generated during block $b \in [1 : B]$. During the first block, without loss of generality, the encoder takes $m_1 = 1$ and returns $X^n(m_1)$. At the beginning of block $b \in \{2, ..., B\}$, the encoder observes the sequence of source U_{b-1}^n of the previous block $b - 1$. It recalls the index $m_{b-1} \in \mathcal{M}$ of the sequence $X^n(m_{b-1})$ used for block $b - 1$. It finds an index m_b such that sequences

$$(U_{b-1}^n, X^n(m_{b-1}), W^n(m_{b-1}, m_b)) \in \mathcal{T}_\varepsilon^{(n)}(\mathcal{P}_{U,X,W})$$

are jointly typical. The encoder sends the sequence $X^n(m_b)$ corresponding to the current block b. We denote by $W_{b-1}^n \triangleq W^n(m_{b-1}, m_b), X_b^n \triangleq X^n(m_b)$.

Channel Transmission: During each block $b \in \{1, ..., B\}$, at time instant $t = 1$, the channel initializes by generating $Y_{b,1} \sim \mathcal{W}(\cdot|X_{b,1}, Y_{b-1,n})$ conditioning on the current input $X_{b,1}$ and the last channel output $Y_{b-1,n}$ from the last block $b-1$. Then, the channel generates the current sequence $Y_{b,2}^n \sim \prod_{t=2}^n \mathcal{W}(Y_{b,t}|X_{b,t}, Y_{b,t-1})$ depending on X_b^n corresponding to block b. Note that in this scheme, because the channel is Markov, the transmission for each block b is *not* independent.

Decoding function: The decoder first returns $\tilde{m}_1 = 1$. During block $b \in [2 : B]$, the decoder recalls the past sequence $Y_{1,b-1}^n$ and the index $\tilde{m}_{b-1}$ that corresponds to the sequence $\tilde{X}_{b-1}^n = X^n(\tilde{m}_{b-1})$. It observes the channel output $Y_{1,b}^n$ and finds the unique index $\tilde{m}_b$ such that

$$(Y_b^n, X^n(\tilde{m}_b)) \in \mathcal{T}_\varepsilon^{(n)}(\mathcal{P}_{X,Y',Y}),$$
$$(Y_{b-1}^n, X^n(\tilde{m}_{b-1}), W^n(\tilde{m}_{b-1}, \tilde{m}_b)) \in \mathcal{T}_\varepsilon^{(n)}(\mathcal{P}_{X,W,Y',Y})$$

are input-driven Markov typical. We denote by $\tilde{X}_b^n = X^n(\tilde{m}_b)$ and $\tilde{W}_{b-1}^n = W^n(\tilde{m}_{b-1}, \tilde{m}_b)$ as our choice. The decoder non-causally generates $V_b^n \sim \mathcal{P}_{V|Y,X,W}^{\otimes n}$ depending on sequences $(Y_b^n, \tilde{X}_b^n, \tilde{W}_b^n)$ for $b \in [1 : B - 1]$. As for the last block, the decoder simply outputs an all zero sequence, i.e., $V_B^n = \mathbf{0}$. Usually, sequences are *not* jointly typical in the last block.

Error Analysis: Next, we show that given the above coding scheme, the generated sequences $(U_b^n, X_b^n, W_b^n, Y_b^n, V_b^n)$ for each block $b \in \{1, ..., B - 1\}$ are jointly (Markov) typical with respect to the joint distribution given in (9) with high probability. For every $\varepsilon > 0$, there exists $N_0 \in \mathbb{N}$, such that the probability of error events for all $n \geq N_0$:

(a) $\mathbb{E}_c \left[\mathbb{P} \left(\forall m \in \mathcal{M}, (U_{b-1}^n, X^n(m_{b-1}), W^n(m_{b-1}, m)) \right. \right.$
$$\left. \left. \notin \mathcal{T}_\varepsilon^{(n)}(\mathcal{P}_{UXW})) \right] \leq \frac{\varepsilon}{3},$$

(b) $\mathbb{E}_c \left[\mathbb{P} \left((X_b^n, Y_b^n) \notin \mathcal{T}_\varepsilon^{(n)}(\mathcal{P}_{Y'XY}) \right) \right] \leq \frac{\varepsilon}{3}$,

(c) $\mathbb{E}_c \left\{ \mathbb{P} \left(\exists \tilde{m} \neq m, \text{s.t. } \{(Y_b^n, X^n(\tilde{m})) \in \mathcal{T}_\varepsilon^{(n)}(\mathcal{P}_{Y'XY})\} \cap \right. \right.$
$$\left. \left. \{(Y_{b-1}^n, X^n(m_{b-1}), W^n(m_{b-1}, \tilde{m})) \in \mathcal{T}_\varepsilon^{(n)}(\mathcal{P}_{WY'XY})\} \right) \right\} \leq \frac{\varepsilon}{3}.$$

Here, (a) is guaranteed by covering lemma for classical i.i.d. sequences [37] and the condition (14). Item (b) is guaranteed by the sequential generation of (X^n, Y^n) and Theorem 2 for input-driven Markov typical. Item (c) is satisfied by a joint packing lemma shown in Appendix B-A below such that

both (dependent) events hold when the rate constraint (15) is satisfied.

We denote by $\tilde{Q}^n \in \Delta(\mathcal{U} \times \mathcal{X} \times \mathcal{W} \times \mathcal{R}_Y \times \mathcal{R}_Y \times \mathcal{V})$ the empirical distribution of symbols (U, X, W, Y', Y, V) over blocks $b = 1, ..., B - 1$, where (Y', Y) are the adjacent symbols of the Markov chain. We now show that $\tilde{Q}^n$ is close to the empirical distribution Q^n over all B blocks for B sufficiently large. We denote by Q_B the empirical distribution of all symbols over the last block. Now,

$$\|Q^n - \tilde{Q}^n\|_1 = \|\frac{1}{B} \left((B-1)\tilde{Q}^n + Q_B \right) - \tilde{Q}^n\|_1$$
$$= \frac{1}{B}\|Q_B - \tilde{Q}^n\|_1$$
$$\leq \frac{2}{B} \cdot |\mathcal{U} \times \mathcal{X} \times \mathcal{W} \times \mathcal{R}_Y \times \mathcal{R}_Y \times \mathcal{V}| \leq \varepsilon$$

when $B \geq \frac{2}{\varepsilon} |\mathcal{U} \times \mathcal{X} \times \mathcal{W} \times \mathcal{R}_Y \times \mathcal{R}_Y \times \mathcal{V}|$. Then, the expected error probability

$$\mathbb{E}_c[\mathbb{P}_e(c)] = \mathbb{E}_c \left[\mathbb{P} \left(\|Q^n - \mathcal{P}\|_1 \geq 2\varepsilon \right) \right]$$
$$\leq \mathbb{E}_c \left[\mathbb{P} \left(\|Q^n - \tilde{Q}^n\|_1 + \|\tilde{Q}^n - \mathcal{P}\|_1 \geq 2\varepsilon \right) \right]$$
$$\leq \mathbb{E}_c \left[\mathbb{P} \left(\|\tilde{Q}^n - \mathcal{P}\|_1 \geq \varepsilon \right) \right] \leq \varepsilon.$$

This implies the existence of a code $c^* \in \mathcal{C}(nB)$ with an error probability below ε for all $n \geq N_0 B$.

A. Joint Packing Lemma

We denote the following events for block b by

$$A_b(\tilde{m}) \triangleq \{(Y_b^n, X^n(\tilde{m})) \in \mathcal{T}_\varepsilon^{(n)}(\mathcal{P}_{Y'XY})\},$$
$$B_b(\tilde{m}) \triangleq \{(Y_b^n, X^n(m_b), W^n(m_b, \tilde{m})) \in \mathcal{T}_\varepsilon^{(n)}(\mathcal{P}_{WY'XY})\}.$$

Note that, compared to the joint packing lemma for the DMC scenario, e.g., in [27], here, since the channel has memory, $A_b(\tilde{m})$ and $B_{b-1}(\tilde{m})$ are not independent anymore, due to the channel state $Y_{b-1,n}$. However, by conditioning on $Y_{b-1,n} = y'$, applying the Markov property, we can have conditional independence and therefore show a similar result. $\square$

Lemma 2 (Joint Packing Lemma). *For each block $b \in [2 : B]$ and for each previous message m_{b-1}, if $R \leq I(X; Y|Y') - \delta$, then, $\forall \varepsilon > 0$, $\exists N_0 \in \mathbb{N}$, such that for all $n \geq N$,*

$$\mathbb{E}_c \{\mathbb{P} (\exists \tilde{m} \neq m, \text{s.t. } A_b(\tilde{m}) \cap B_{b-1}(\tilde{m}))\} \leq \varepsilon$$

Proof. Consider $\varepsilon > 0$ that satisfies $4\varepsilon < \delta$. We have

$$R - I(Y; X|Y') + 3\varepsilon \leq -\delta + 3\varepsilon < -\varepsilon.$$

$$\mathbb{E}_c \{\mathbb{P} (\exists \tilde{m} \neq m, \text{s.t. } A_b(\tilde{m}) \cap B_{b-1}(\tilde{m}))\}$$
$$\leq \sum_{\tilde{m} \neq m} \mathbb{E}_c \{\mathbb{P} (A_b(\tilde{m}) \cap B_{b-1}(\tilde{m}))\}$$
$$= \sum_{\tilde{m} \neq m} \mathbb{E}_c \left\{ \sum_y \mathbb{P}(Y_{b-1,n} = y) \right.$$
$$\left. \cdot \mathbb{P} \left(A_b(\tilde{m}) \cap B_{b-1}(\tilde{m}) \middle| Y_{b-1,n} = y \right) \right\}$$

$$= \sum_{\tilde{m} \neq m} \mathbb{E}_c \left\{ \sum_y \mathbb{P}(Y_{b-1,n} = y) \cdot \mathbb{P}\left(B_{b-1}(\tilde{m}) \mid Y_{b-1,n} = y\right) \right.$$
$$\left. \cdot \mathbb{P}\left(A_b(\tilde{m}) \mid B_{b-1}(\tilde{m}), Y_{b-1,n} = y\right) \right\}$$

$$\stackrel{(a)}{=} \sum_{\tilde{m} \neq m} \mathbb{E}_c \left\{ \sum_y \mathbb{P}(Y_{b-1,n} = y) \cdot \mathbb{P}\left(B_{b-1}(\tilde{m}) \middle| Y_{b-1,n} = y\right) \right.$$
$$\left. \cdot \mathbb{P}\left(A_b(\tilde{m}) \middle| Y_{b-1,n} = y\right) \right\}$$

$$= \sum_{\tilde{m} \neq m} \mathbb{E}_c \left\{ \sum_y \mathbb{P}(Y_{b-1,n} = y) \cdot \mathbb{P}\left(B_{b-1}(\tilde{m}) \middle| Y_{b-1,n} = y\right) \right.$$
$$\left. \cdot \sum_{(y^n, x^n) \in \mathcal{T}_\varepsilon^{(n)}(\mathcal{P})} \mathbb{P}\left(Y_b^n = y^n X^n(\tilde{m})) = x^n \middle| Y_{b-1,n} = y\right) \right\}$$

$$\stackrel{(b)}{=} \sum_{\tilde{m} \neq m} \mathbb{E}_c \left\{ \sum_y \mathbb{P}(Y_{b-1,n} = y) \cdot \mathbb{P}\left(B_{b-1}(\tilde{m}) \middle| Y_{b-1,n} = y\right) \right.$$
$$\left. \cdot \sum_{(y^n, x^n) \in \mathcal{T}_\varepsilon^{(n)}(\mathcal{P})} \mathbb{P}\left(Y_b^n = y^n \middle| Y_{b-1,n} = y\right) \cdot \mathbb{P}\left(X^n(\tilde{m})) = x^n\right) \right\}$$

$$\stackrel{(c)}{\leq} \sum_{\tilde{m} \neq m} \mathbb{E}_c \left\{ \sum_y \mathbb{P}(Y_{b-1,n} = y) \cdot \mathbb{P}\left(B_{b-1}(\tilde{m}) \middle| Y_{b-1,n} = y\right) \right.$$
$$\left. \cdot \sum_{(y^n, x^n) \in \mathcal{T}_\varepsilon^{(n)}(\mathcal{P})} 2^{-n(H(Y|Y')-\epsilon)} \times 2^{-n(H(X)-\epsilon)} \right\}$$

$$\stackrel{(d)}{\leq} 2^{n(H(Y,X|Y')-H(Y|Y')-H(X)+3\epsilon)} \cdot \sum_{\tilde{m} \neq m} \mathbb{E}_c \left\{ \mathbb{P}(B_{b-1}(\tilde{m})) \right\}$$

$$\stackrel{(e)}{\leq} 2^{nR} \times 2^{n(-I(X;Y|Y')+3\varepsilon)}$$
$$\leq 2^{-n\epsilon},$$

where,
(a) Markov property of the channel output generation: the dependence of the current block's sequence on the last block only through its last symbol $Y_{b-1,n}$;
(b) the independence of the random variable Y_b^n given $Y_{b-1,n}$ with $X^n(\tilde{m})$ where $\tilde{m} \neq m$;
(c) AEP of the strong Markov typicality for Markov sequence Y^n, which leverages ergodicity properties that the effect of the boundary initial state can be averaged out, see [34, Proposition 4.1.1], and also $X \perp\!\!\!\perp Y'$;
(d) Cardinality bounds for the input-driven Markov typical sequence, i.e., Lemma 3;
(e) $\mathbb{E}_c \left\{ \mathbb{P}(B_{b-1}(\tilde{m})) \right\} \leq 1$ and the choice of R. $\square$

APPENDIX C
CONVERSE PROOF OF THEOREM 1

For $n \in \mathbb{N}$, we consider a code $c \in \mathcal{C}(n)$ that induces a joint distribution of form (2).

$$0 = I(U^n; Y^n) - I(U^n; Y^n)$$
$$\stackrel{(a)}{=} \sum_t I(U^n; Y_t | Y^{t-1}) - \sum_i I(U_t; Y^n | U^{t-1})$$
$$\stackrel{(b)}{=} \sum_t I(U^{t-1}; Y_t | Y^{t-1}) - \sum_i I(U_t; Y^n | U^{t-1})$$
$$\leq \sum_t I(U^{t-1}; Y_t | Y^{t-1}) - \sum_t I(U_t; Y^{t-1}, Y_{t+1}^n | U^{t-1})$$
$$\stackrel{(c)}{=} \sum_t I(X_t, U^{t-1}; Y_t | Y^{t-1}) - \sum_t I(U_t; Y^{t-1}, Y_{t+1}^n | U^{t-1}, X_t)$$
$$\stackrel{(d)}{=} \sum_t I(X_t, U^{t-1}; Y_t | Y^{t-1}) - \sum_t I(U_t; U^{t-1}, Y^{t-1}, Y_{t+1}^n | X_t)$$
$$\stackrel{(e)}{\leq} \sum_t I(X_t; Y_t | Y_{t-1}) - \sum_t I(U_t; W_t | X_t)$$
$$\stackrel{(f)}{=} n \cdot (I(X_T; Y_T | Y_{T-1}, T) - I(U_T; W_T | X_T, T))$$
$$= n \cdot (H(Y_T | Y_{T-1}, T) - H(Y_T | X_T, Y_{T-1}, T)$$
$$- I(U_T; W_T, T | X_T))$$
$$\stackrel{(g)}{\leq} n \cdot (H(Y_T | Y_{T-1}) - H(Y_T | X_T, Y_{T-1}) - I(U_T; W_T, T | X_T))$$
$$= n \cdot (I(X_T; Y_T | Y_{T-1}) - I(U_T; W_T, T | X_T))$$
$$\stackrel{(h)}{=} n \cdot (I(X; Y | Y') - I(U; W | X)),$$

where,
(a) chain rule of mutual information;
(b) because the source is generated i.i.d., we have $U_t^n \perp\!\!\!\perp (U^{t-1}, Y^{t-1}, Y_t)$, hence $I(U_t^n; Y_t | Y^{t-1}, U^{t-1}) = 0$;
(c) deterministic strictly causal encoding function, i.e., $X_t = f^{(t)}(U^{t-1})$;
(d) i.i.d. source and strictly causal encoding, we hence have $I(U_t; U^{t-1} | X_t) = 0$;
(e) because of conditioning reduces entropy, and the Markov chain $Y_t \, \text{\small$\multimap$} \, (X_t, Y_{t-1}) \, \text{\small$\multimap$} \, (U^{t-1}, Y^{t-2})$. Also, in the second term, we identify $W_t = (U^{t-1}, Y^{t-1}, Y_{t+1}^n)$;
(f) the introduction of the uniform random variable T over $\{1, ..., n\}$ and the introduction of the corresponding mean random variables U_T, X_T, W_T, Y_T, V_T and Y_{T-1} represents the last symbol of the current state Y_T;
(g) conditioning reduces entropy, and stationarity of the Markov channel;
(h) the assignment of $U = U_T, X = X_T, W = (W_T, T), Y' = Y_{T-1}, Y = Y_T, V = V_T$.

Lastly, we show that, random variables (U, X, W, Y', Y, V) defined above satisfy the following Markov chains

- $U \perp\!\!\!\perp (X, Y', Y)$ comes from the strictly causal encoding,
- $V \, \text{\small$\multimap$} \, (X, Y, W) \, \text{\small$\multimap$} \, (U, Y')$ comes from the noncausal decoding, and the fact that Y^n is included in (W_t, Y_t) for all $t \in \{1, ..., n\}$, hence is included in $(W_T, T, Y_T) = (W, Y)$. $\square$

References

[1] P. Diaconis and D. Freedman, "Iterated random functions," *SIAM review*, vol. 41, no. 1, pp. 45–76, 1999.

[2] N. Hermer, D. R. Luke, and A. Sturm, "Rates of convergence for chains of expansive markov operators," *Transactions of Mathematics and Its Applications*, vol. 7, no. 1, p. tnad001, 2023.

[3] A. Gupta, R. Jain, and P. Glynn, "Probabilistic contraction analysis of iterated random operators," *IEEE Transactions on Automatic Control*, vol. 69, no. 9, pp. 5947–5962, 2024.

[4] K. J. Åström, *Introduction to stochastic control theory*. Courier Corporation, 2012.

[5] D. Bertsekas, *Dynamic programming and optimal control: Volume I*, vol. 4. Athena scientific, 2012.

[6] T. Tanaka, A. Z. W. Cheng, and C. Langbort, "A dynamic pivot mechanism with application to real time pricing in power systems," in *2012 American Control Conference (ACC)*, pp. 3705–3711, IEEE, 2012.

[7] D. Blackwell, L. Breiman, and A. J. Thomasian, "Proof of shannon's transmission theorem for finite-state indecomposable channels," *The Annals of Mathematical Statistics*, pp. 1209–1220, 1958.

[8] R. G. Gallager, *Information theory and reliable communication*, vol. 588. Springer, 1968.

[9] R. Gray, M. Dunham, and R. Gobbi, "Ergodicity of markov channels," *IEEE transactions on information theory*, vol. 33, no. 5, pp. 656–664, 1987.

[10] S. Verdú *et al.*, "A general formula for channel capacity," *IEEE Transactions on Information Theory*, vol. 40, no. 4, pp. 1147–1157, 1994.

[11] A. J. Goldsmith and P. P. Varaiya, "Capacity, mutual information, and coding for finite-state markov channels," *IEEE transactions on Information Theory*, vol. 42, no. 3, pp. 868–886, 2002.

[12] J. Massey *et al.*, "Causality, feedback and directed information," in *Proc. Int. Symp. Inf. Theory Applic.(ISITA-90)*, vol. 2, p. 1, 1990.

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

[14] J. Chen and T. Berger, "The capacity of finite-state markov channels with feedback," *IEEE Transactions on Information Theory*, vol. 51, no. 3, pp. 780–798, 2005.

[15] S. Tatikonda and S. Mitter, "The capacity of channels with feedback," *IEEE Transactions on Information Theory*, vol. 55, no. 1, pp. 323–349, 2008.

[16] H. H. Permuter, H. Asnani, and T. Weissman, "Capacity of a post channel with and without feedback," *IEEE Transactions on Information Theory*, vol. 60, no. 10, pp. 6041–6057, 2014.

[17] A. Grigorescu, H. Boche, R. F. Schaefer, and H. V. Poor, "Capacity of finite state channels with feedback: Algorithmic and optimization theoretic properties," *IEEE Transactions on Information Theory*, vol. 70, no. 8, pp. 5413–5426, 2024.

[18] H. Wu, G. Chen, and D. Gündüz, "Actions speak louder than words: Rate-reward trade-off in markov decision processes," *arXiv preprint arXiv:2502.03335*, 2025.

[19] C. E. Shannon, "A mathematical theory of communication," *The Bell system technical journal*, vol. 27, no. 3, pp. 379–423, 1948.

[20] D. Gündüz, M. A. Wigger, T.-Y. Tung, P. Zhang, and Y. Xiao, "Joint source–channel coding: Fundamentals and recent progress in practical designs," *Proceedings of the IEEE*, 2024.

[21] P. Cuff and L. Zhao, "Coordination using implicit communication," in *2011 IEEE Information Theory Workshop*, pp. 467–471, IEEE, 2011.

[22] M. Raginsky, "Empirical processes, typical sequences, and coordinated actions in standard borel spaces," *IEEE Transactions on Information Theory*, vol. 59, no. 3, pp. 1288–1301, 2012.

[23] P. W. Cuff, H. H. Permuter, and T. M. Cover, "Coordination capacity," *IEEE Transactions on Information Theory*, vol. 56, no. 9, pp. 4181–4206, 2010.

[24] P. Cuff and C. Schieler, "Hybrid codes needed for coordination over the point-to-point channel," in *2011 49th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 235–239, IEEE, 2011.

[25] M. Le Treust, "Empirical coordination with channel feedback and strictly causal or causal encoding," in *2015 IEEE International Symposium on Information Theory (ISIT)*, 2015.

[26] M. Le Treust, "Joint empirical coordination of source and channel," *IEEE Transactions on Information Theory*, vol. 63, no. 8, pp. 5087–5114, 2017.

[27] M. Le Treust, "Empirical coordination with two-sided state information and correlated source and state," in *2015 IEEE International Symposium on Information Theory (ISIT)*, pp. 466–470, IEEE, 2015.

[28] M. Le Treust and T. J. Oechtering, "Power-estimation trade-off of vector-valued Witsenhausen counterexample with causal decoder," *IEEE Transactions on Information Theory*, vol. 70, no. 3, pp. 1588–1609, 2024.

[29] M. Zhao, M. Le Treust, and T. J. Oechtering, "Coordination coding with causal encoder for vector-valued Witsenhausen counterexample," in *2024 IEEE International Symposium on Information Theory (ISIT)*, pp. 3255–3260, IEEE, 2024.

[30] M. Zhao, M. Le Treust, and T. J. Oechtering, "Causal vector-valued Witsenhausen counterexamples with feedback," in *2024 IEEE Information Theory Workshop (ITW)*, pp. 687–692, IEEE, 2024.

[31] C. Choudhuri, Y.-H. Kim, and U. Mitra, "Causal state communication," *IEEE Transactions on Information Theory*, vol. 59, no. 6, pp. 3709–3719, 2013.

[32] J. R. Norris, *Markov chains*. No. 2, Cambridge university press, 1998.

[33] L. Breuer and D. Baum, *An introduction to queueing theory and matrix-analytic methods*. Springer, 2005.

[34] S. Huang, *Linear Coding, Applications and Supremus Typicality*. PhD thesis, KTH Royal Institute of Technology, 2015.

[35] I. Csiszár, "The method of types [information theory]," *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2505–2523, 2002.

[36] L. Davisson, G. Longo, and A. Sgarro, "The error exponent for the noiseless encoding of finite ergodic markov sources," *IEEE Transactions on Information Theory*, vol. 27, no. 4, pp. 431–438, 2003.

[37] T. M. Cover and J. A. Thomas, *Elements of information theory*. John Wiley & Sons, 1999.

Under Assumption A, the induced transition matrix $T = [T_{i,j}]_{i,j \in \mathcal{Y}}$ of the Markov chain Y^n does not depend on time $t \in [1:n]$, where

$$
\begin{aligned}
T_{i,j} &= \mathbb{P}(Y_k = j | Y_{k-1} = i) \\
&= \sum_{x \in \mathcal{X}} \mathcal{P}_X(x) \mathcal{W}(Y_k = j | X_k = x, Y_{k-1} = i).
\end{aligned}
$$

We can prove the following property of the input-driven Markov typicality:

Proposition 1. *For $n \in \mathbb{N}$ and $\varepsilon > 0$, the typical sequences satisfy the following properties*
1) If $(x^n, y^n) \in \mathcal{T}_\varepsilon^{(n)}(\mathcal{Q}_{Y'XY})$, then $x^n \in \mathcal{T}_\varepsilon^{(n)}(\mathcal{P}_X)$, $y^n \in \mathcal{T}_\varepsilon^{(n)}(\pi \cdot T)$, $y^n \in \mathcal{T}_{2\varepsilon}^{(n)}(\mathcal{Q}_{Y'XY}|x^n)$ conditional typical, where T is the marginalized transition matrix of Markov sequence Y^n given in (16).
2) If $x^n \in \mathcal{T}_\varepsilon^{(n)}(\mathcal{P}_X)$, $y^n \in \mathcal{T}_{2\varepsilon}^{(n)}(\mathcal{Q}_{Y'XY}|x^n)$, then $(x^n, y^n) \in \mathcal{T}_{2\varepsilon}^{(n)}(\mathcal{Q}_{Y'XY})$

Proof. 1) Proof of item 1: If $(x^n, y^n) \in \mathcal{T}_\varepsilon^{(n)}(\mathcal{Q}_{Y'XY})$, then

$$
\begin{aligned}
\varepsilon &\geq \|Q_{Y'XY}^n - \mathcal{Q}_{Y'XY}\|_1 \\
&= \sum_{x \in \mathcal{X}, i,j \in \mathcal{Y}} |Q_{Y'XY}^n(i,x,j) - \pi_{Y'}(i)\mathcal{P}_X(x)\mathcal{W}(j|x,i)| \\
&\geq \sum_x \left| \sum_{i,j} (Q_{Y'XY}^n(i,x,j) - \pi_{Y'}(i)\mathcal{P}_X(x)\mathcal{W}(j|x,i)) \right| \\
&= \sum_x |Q_X^n(x) - \mathcal{P}_X(x)| \\
&= \|Q_X^n - \mathcal{P}_X\|_1
\end{aligned}
$$

Similarly,

$$
\begin{aligned}
\varepsilon &\geq \|Q_{Y'XY}^n - \mathcal{Q}_{Y'XY}\|_1 \\
&= \sum_{x \in \mathcal{X}, i,j \in \mathcal{Y}} |Q_{Y'XY}^n(i,x,j) - \pi_{Y'}(i)\mathcal{P}_X(x)\mathcal{W}(j|x,i)| \\
&\geq \sum_{i,j} \left| \sum_x (Q_{Y'XY}^n(i,x,j) - \pi_{Y'}(i)\mathcal{P}_X(x)\mathcal{W}(j|x,i)) \right| \\
&= \sum_{i,j} |Q_Y^n(i,j) - \pi_{Y'}(i)T(j|i)| \\
&= \|Q_Y^n - \pi \cdot T\|_1
\end{aligned}
$$

Moreover, we also have

$$
\begin{aligned}
&\|Q_{Y'XY}^n - \pi_{Y'}Q_X^n\mathcal{W}\|_1 \\
&= \|Q_{Y'XY}^n - \mathcal{Q}_{Y'XY} + \mathcal{Q}_{Y'XY} - \pi_{Y'}Q_X^n\mathcal{W}\|_1 \\
&\leq \|Q_{Y'XY}^n - \mathcal{Q}_{Y'XY}\|_1 + \|\mathcal{Q}_{Y'XY} - \pi_{Y'}Q_X^n\mathcal{W}\|_1 \\
&\leq \|Q_{XY}^n - \mathcal{Q}_{Y'XY}\|_1 + \sum_{x,i,j} |\pi_{Y'}(i)\mathcal{P}_X(x)\mathcal{W}(j|x,i) - \pi_{Y'}(i)Q_X^n(x)\mathcal{W}(j|x,i)| \\
&\leq \|Q_{XY}^n - \mathcal{Q}_{Y'XY}\|_1 + \sum_{x,i,j} |\mathcal{P}_X(x) - Q_X^n(x)| \cdot (\pi_{Y'}(i)\mathcal{W}(j|x,i)) \\
&= \|Q_{XY}^n - \mathcal{Q}_{Y'XY}\|_1 + \sum_x |\mathcal{P}_X(x) - Q_X^n(x)| \cdot \sum_{i,j} (\pi_{Y'}(i)\mathcal{W}(j|x,i)) \\
&= \|Q_{XY}^n - \mathcal{Q}_{Y'XY}\|_1 + \|\mathcal{P}_X - Q_X^n\|_1 \\
&\leq 2 \cdot \varepsilon.
\end{aligned}
$$

2) Proof of item 2:

$$
\|Q_{Y'XY}^n - \pi_{Y'}\mathcal{P}_X\mathcal{W}\|_1
$$

$$= \|Q^n_{Y'XY} - \pi_{Y'}Q^n_X\mathcal{W} + \pi_{Y'}Q^n_X\mathcal{W} - \pi_{Y'}\mathcal{P}_X\mathcal{W}\|_1$$
$$\leq \|Q^n_{Y'XY} - \pi_{Y'}Q^n_X\mathcal{W}\|_1 + \|\pi_{Y'}Q^n_X\mathcal{W} - \pi_{Y'}\mathcal{P}_X\mathcal{W}\|_1$$
$$\leq 2\varepsilon.$$

$\square$

APPENDIX E
PROOF FOR THEOREM 2

Define the stochastic process for $t \geq 1$:

$$S_t = (Y_{t-1}, X_t, Y_t) \tag{16}$$

It is obviously a Markov chain.

We show first, that $\{S_t\}$ is a time-homogeneous Markov chain: Given the current state $S_t = (Y_{t-1} = i, X_t = x, Y_t = j)$, then for the next time

- $X_{t+1} \sim \mathcal{P}_X$ independent of the past,
- $Y_{t+1} \sim \mathcal{W}(\cdot|X_{t+1}, Y_t = j)$

Therefore, the transition probability

$$\mathbb{P}(S_{t+1} = (j, x', k)|S_t = (i, x, j)) = \mathcal{P}_X(x')\mathcal{W}(k|x', j) \tag{17}$$

does not depend on time. Hence S is homogeneous. Moreover, since $\{Y_k\}$ has only one irreducible aperiodic recurrent set $\mathcal{R}_Y$, so does $\{S_k\}$, and it is simply $\mathcal{R}_S = \mathcal{R}_Y \times \mathcal{X} \times \mathcal{R}_Y$. Therefore, S has a unique equilibrium distribution on $\mathcal{R}_S$. Next, we prove the equilibrium distribution of the process S is simply $\mathcal{Q}_{Y'XY} = \pi_{Y'}\mathcal{P}_X\mathcal{W}_{Y|X,Y'}$, i.e., we want to have

$$\sum_{i,x} \mathcal{Q}_{Y'XY}(i, x, j)\underbrace{\mathcal{P}_X(x')\mathcal{W}(k|x', j)}_{\text{transition matrix of } S} = \mathcal{Q}_{Y'XY}(j, x', k) \tag{18}$$

Now, because

$$\sum_{i,x} \mathcal{Q}_{Y'XY}(i, x, j)\mathcal{P}_X(x')\mathcal{W}(k|x', j)$$

$$= \left(\sum_{i,x} \pi(i)\mathcal{P}_X(x)\mathcal{W}(j|x, i)\right) \mathcal{P}_X(x')\mathcal{W}(k|x', j)$$

$$= \left(\sum_i \pi(i) \sum_x [\mathcal{P}_X(x)\mathcal{W}(j|x, i)]\right) \mathcal{P}_X(x')\mathcal{W}(k|x', j)$$

$$= \pi(j)\mathcal{P}_X(x')\mathcal{W}(k|x', j)$$

$$= \mathcal{Q}_{Y'XY}(j, x', k).$$

$\mathcal{Q}_{Y'XY} = \pi_{Y'}\mathcal{P}_X\mathcal{W}_{Y|X,Y'}$ is the invariant distribution of S. Since $\{S_t\}$ is a Markov chain with finite states and a unique irreducible aperiodic recurrent class, it is then ergodic. By the ergodic theorem, for each state $(i, x, j) \in \mathcal{R}_S$, we have

$$\frac{1}{n}\sum_{t=1}^n \mathbf{1}\{S_t = (i, x, j)\} \xrightarrow{a.s.} \mathcal{Q}_{Y'XY}(i, x, j) \tag{19}$$

where the left hand side above is exactly the joint empirical $Q^n_{Y'XY}(i, x, j)$. Therefore

$$\|Q^n_{Y'XY} - \mathcal{Q}_{Y'XY}\|_1 \xrightarrow{a.s.} 0. \tag{20}$$

Since almost sure convergence implies converge in probability, we have

$$\forall \delta > 0, \lim_{n\to\infty} \mathbb{P}((X^n, Y^n) \in \mathcal{T}_\delta^{(n)}(\mathcal{Q}_{Y'XY})) = 1. \tag{21}$$

$\square$

Proof of 1)

Let $\mathbb{P}((X_1, Y_1) = (x_1, y_1)) = c$. Then,

$$\mathbb{P}((X^n, Y^n) = (x^n, y^n))$$

$$= \mathbb{P}((X_1, Y_1) = (x_1, y_1)) \prod_{i=2}^n \mathcal{P}_X(x_i) \mathcal{W}(y_i | x_i, y_{i-1})$$

$$= c \prod_{x \in \mathcal{X}} \mathcal{P}_X(x)^{N(x|x^n)} \prod_{x \in \mathcal{X}, i, j \in \mathcal{Y}} \mathcal{W}(j|x, i)^{N(i,x,j|x^n, y^n)}$$

where $N(x|x^n) = \sum_{t=1}^n \mathbf{1}\{x_t = x\}$, $N(i, x, j|x^n, y^n) = \sum_{t=1}^n \mathbf{1}\{(y_{t-1}, x_t, y_t) = (i, x, j)\}$.

Now, take $\log_2$ on both sides and multiply by $-\frac{1}{n}$. Then we have

$$-\frac{1}{n} \log_2 \mathbb{P}((x^n, y^n))$$

$$= -\frac{1}{n} \log_2 c - \sum_x Q_X^n(x) \log_2 \mathcal{P}_X(x) - \sum_{x,i,j} Q_{Y'XY}^n(i, x, j) \log_2 \mathcal{W}(j|x, i)$$

$$= -\frac{1}{n} \log_2 c - \sum_x (Q_X^n(x) - \mathcal{P}_X(x)) \log_2 \mathcal{P}_X(x) - \sum_x \mathcal{P}_X(x) \log_2 \mathcal{P}_X(x) \tag{22}$$

$$- \sum_{x,i,j} (Q_{Y'XY}^n(i, x, j) - \pi_{Y'}(i) \mathcal{P}_X(x) \mathcal{W}(j|x, i)) \log_2 \mathcal{W}(j|x, i) - \sum_{x,i,j} \pi_{Y'}(i) \mathcal{P}_X(x) \mathcal{W}(j|x, i) \log_2 \mathcal{W}(j|x, i)$$

Because $\mathcal{X}, \mathcal{Y}$ are finite, then

$$L_X := \max_{x: \mathcal{P}_X(x) > 0} |\log_2 \mathcal{P}_X(x)|, \quad L_W := \max_{(i,x,j): \mathcal{W}(j|x,i) > 0} |\log_2 \mathcal{W}(j|x, i)| \tag{23}$$

are both finite numbers. Now, since $(x^n, y^n) \in \mathcal{T}_\varepsilon^{(n)}(\mathcal{Q}_{Y'XY})$ input-driven Markov typical, we also have $\|Q_X^n - \mathcal{P}_X\|_1 \leq \varepsilon$ provided by Proposition 1. Therefore, we obtain that

$$-\varepsilon \cdot L_X \leq \sum_x (Q_X^n(x) - \mathcal{P}_X(x)) \log_2 \mathcal{P}_X(x) \leq \varepsilon \cdot L_X$$

and

$$-\varepsilon \cdot L_W \leq \sum_{x,i,j} (Q_{Y'XY}^n(i, x, j) - \pi_{Y'}(i) \mathcal{P}_X(x) \mathcal{W}(j|x, i)) \log_2 \mathcal{W}(j|x, i) \leq \varepsilon \cdot L_W.$$

Now, for any fixed $\delta > 0$, choose N_0 big enough such that

$$-\delta/3 \leq -\frac{1}{n} \log_2 c \leq \delta/3, \quad \forall n > N_0,$$

and ε_0 small enough such that

$$\varepsilon_0 (L_X + L_W) \leq \delta/3.$$

Then, $\forall \varepsilon < \varepsilon_0$ and $n > N_0$, from (22) we obtain that every $(x^n, y^n) \in \mathcal{T}_\varepsilon^{(n)}(\mathcal{Q}_{Y'XY})$ satisfies

$$-\frac{1}{n} \log_2 \mathbb{P}((x^n, y^n)) < \delta + H(X) + H(Y|X, Y') = \delta + H(X, Y|Y')$$

$$-\frac{1}{n} \log_2 \mathbb{P}((x^n, y^n)) > -\delta + H(X) + H(Y|X, Y') = -\delta + H(X, Y|Y')$$

Hence, for every $(x^n, y^n) \in \mathcal{T}_\varepsilon^{(n)}(\mathcal{Q}_{Y'XY})$,

$$2^{-n[H(X,Y|Y') + \delta]} < \mathbb{P}((x^n, y^n)) < 2^{-n[H(X,Y|Y') - \delta]}$$

Item 2) follows from the lower bound in item 1). $\qquad \square$