Non-interactive Remote Coordination

Yassine Hamdi
Department of Electrical
and Electronic Engineering
Imperial College London
United Kingdom
y.hamdi@imperial.ac.uk

Xueyan Niu
Theory Lab, 2012 Labs
Huawei Technologies Co. Ltd.
Hong Kong
China
niuxueyan3@huawei.com

Bo Bai
Theory Lab, 2012 Labs
Huawei Technologies Co. Ltd.
Hong Kong
China
baibo8@huawei.com

Deniz Gündüz
Department of Electrical
and Electronic Engineering
Imperial College London
United Kingdom
d.gunduz@imperial.ac.uk

Abstract

In multi-agent systems coordination is essential to enable autonomous agents to carry out joint tasks. In stochastic environments, this is usually achieved by the agents' actions to approximate a certain desired joint distribution. In this work, we consider coordination among two remote agents, controlled by two controllers through separate rate-limited communication channels. The controllers cannot communicate with each other, but they have access to correlated observations, which can provide a certain level of coordination. Our goal is to explore the effect of limited communication links on the coordination capabilities of the agents. The studied problem can be considered as distributed compression for coordination, with implications in multi-agent reinforcement learning and game theory.

1 Introduction

Recent advances in edge computing have witnessed the proliferation of agent-based systems such as the Internet of Things, federated learning, and autonomous agents. Achieving high-level coordination under communication constraints is essential for agents to accomplish real-world tasks from policy making [9], to distributed learning [3], and generative modeling [12]. For example, automated cars at high-speed need to make real-time decisions; drones at high altitudes need to operate jointly with limited communications. On the one hand, communications among agents/edge devices are not always possible due to practical constraints such as wireless environments, computational resources, latency, and privacy. On the other hand, coordination is needed to support robust actions, such as rapid eye-hand coordination for competitive games [6]. Therefore, understanding the fundamental limits of coordination under rate-limited communication is essential. Moreover, limited communication allows decision-making with low-latency, which is crucial for applications such as industrial automation.

In this paper, we consider the non-interactive remote coordination problem which is illustrated in Figure 1. Here, two non-interacting controllers, called Encoder A and Encoder B, observe two correlated source sequences, U^n and V^n, respectively. The goal of the encoders is to help two decoders, Decoder A and Decoder B, to generate coordinated actions, X^n and Y^n, respectively, by communicating with them over rate-limited channels. Input U^n is compressed by Encoder A into nR_A bits, which Decoder A uses to generate its output X^n. Independently, Encoder B compresses V^n into nR_B bits, used by Decoder B to output Y^n. Moreover, each encoder-decoder pair has access to a common

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

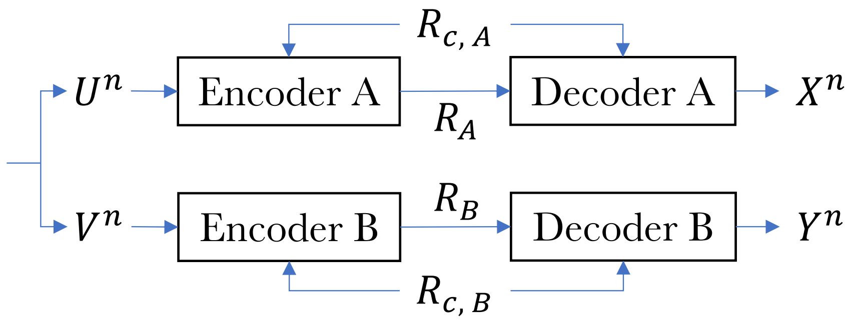

Figure 1: The system model. Encoder A (resp. B) and Decoder A (resp. B) have access to a common randomness at a limited rate $R_{c,A}$ (resp. $R_{c,B}$), and the goal of Encoders A and B is to help Decoders A and B to generate outputs from a certain desired joint distribution, through communicating over rate-limited channels.

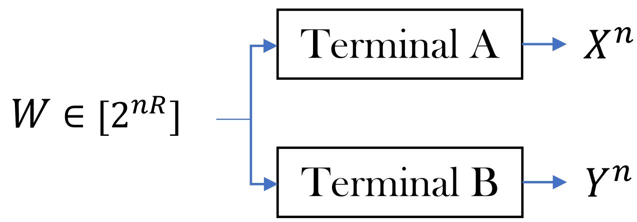

Figure 2: Wyner's setup [11].

randomness at a limited rate ($R_{c,A}$ and $R_{c,B}$, respectively), in the form of uniform random variables (J_A and J_B). The goal is for the generated outputs X^n and Y^n to approximately follow a given target joint distribution $q_{X,Y}^{\otimes n}$, as formalized with a negligible total variation distance, in the limit of large n. This formulation encompasses various machine learning problems. These include the generation of coordinated policies among distributed agents in multi-agent reinforcement learning problems, or neural based distributed compression of correlated images with the goal of realism, where the realism will require the reconstructed images to follow the joint distribution of the original images.

The problem of generating a target joint distribution, in a distributed manner, from some common randomness, has long been of interest [11, 8, 7]. In particular, several notions of common information [11, 7] playing a fundamental role in such problems have been uncovered. Wyner [11] considered the distributed generation task depicted in Figure 2. Two terminals have access to a common random variable W, and independently generate two correlated outputs. Wyner [11] studied the case where the target distribution is a product distribution $q_{X,Y}^{\otimes n}$, for some integer n. He characterized the necessary and sufficient amount of entropy $H(W)$, of the common randomness, for $q_{X,Y}$ to be *asymptotically simulable*, i.e. for there to exist generators achieving an arbitrarily small mismatch between $q_{X,Y}^{\otimes n}$ and the joint distribution of X^n and Y^n, for large enough n. The aforementioned entropy threshold exhibited by Wyner [11] is known as the Wyner common information between of X and Y, which is a function of $q_{X,Y}$. Furthermore, in multi-modal learning, data from different modalities are embedded into a shared representation space using methods such as contrastive learning. [10] revisits the Platonic representation hypothesis, which suggests that data from different modalities all converge to a "true" distribution through training. This encoding process resembles the Gács- Körner (GK) common information [7], where Encoder A and Encoder B each compute a deterministic function of their input, to agree on a common random variable. Berg et. al. [2] went further, by considering two generators having distinct but correlated inputs U^n and V^n, instead of a common random input W, as depicted in Figure 3. They designed non-trivial coding schemes by combining results on the Wyner common information and the GK common information. The problem studied by Berg et. al. [2] is the special case of our problem, in which $R_A = R_B = \infty, R_{c,A} = R_{c,B} = 0$.

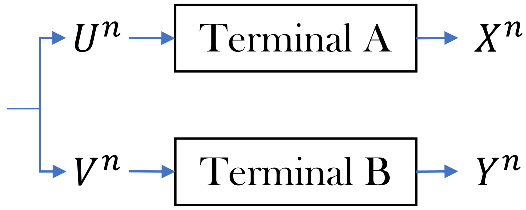

Figure 3: Setup of Berg et. al. [2]. The inputs U^n and V^n follow a joint distribution of the form $p_{U,V}^{\otimes n}$.

In this paper, for the remote scenario in Fig. 1, we identify sufficient conditions on the rate tuple $(R_A, R_{c,A}, R_B, R_{c,B})$ and the distribution $q_{U,V}$ of inputs, in order for a given target distribution $q_{X,Y}$ to be asymptotically simulable. In particular, these are sufficient conditions to allow for *enough* correlation to carry over from the sources to the remote outputs. These correspond to a scheme we propose, which relies on a special channel simulation protocol, enabling a generalization of the scheme in [2].

In Section 2, we provide a formal description of the problem. In Section 3, we present our main result, then compare to two simpler kinds of schemes, for certain families of distributions $q_{U,V}$ and $q_{X,Y}$. We show that our scheme leverages the correlation between the sources more efficiently. As a byproduct, this analysis exhibits the role that common randomness can play for rate efficiency. The main components of our derivation are provided in Section 4, and the remainder is deferred to the Appendix.

2 Problem formulation

2.1 Notation

Calligraphic letters such as $\mathcal{X}$ denote sets, except in $p_{\mathcal{J}}^{\mathcal{U}}$, which denotes the uniform distribution over alphabet $\mathcal{J}$. Random variables are denoted using upper case letters such as X, and their realizations using lower case letters such as x. For a distribution P, the expression P_X denotes the marginal of variable X, while $P(x)$ denotes the probability of the event $X{=}x$. Similarly, $P_{X|Y=y}$ denotes a distribution over $\mathcal{X}$, and $P_{X|Y=y}(x)$ a real number. We denote by $[a]$ the set $\{1, ..., \lfloor a \rfloor\}$, and by x^n the finite sequence $(x_1, ..., x_n)$. The total variation distance (TVD) between distributions p and q on a space $(\mathcal{X}, \mathcal{B})$ is defined by

$$\|p - q\|_{TV} := \sup_{B \in \mathcal{B}} |p(B) - q(B)|.$$

2.2 Definitions

Let $\mathcal{U}, \mathcal{V}, \mathcal{X}, \mathcal{Y}$ be four finite alphabets, $q_{U,V}$ a distribution on $\mathcal{U} \times \mathcal{V}$, and $q_{X,Y}$ a distribution on $\mathcal{X} \times \mathcal{Y}$. As shown in Figure 1, we consider two encoder-decoder pairs, with no communication across pairs, and with each pair having access to a rate-limited uniformly distributed common randomness. Each encoder can send a rate-limited message to the corresponding decoder. Private randomness is unconstrained.

Definition 2.1. *Given a couple $(R, R_c) \in \mathbb{R}^2_{\geq 0}$, an input alphabet $\mathcal{Z}$, and an output alphabet $\mathcal{O}$, an (n, R, R_c) code for $(\mathcal{Z}, \mathcal{O})$ is an encoder-decoder pair $(F^{(n)}, G^{(n)})$, consisting of a privately randomized mapping $F^{(n)}_{M|Z^n,J}$ from $\mathcal{Z}^n \times [2^{\lfloor nR_c \rfloor}]$ to $[2^{\lfloor nR \rfloor}]$, and a privately randomized mapping $G^{(n)}_{O^n|J,M}$ from $[2^{\lfloor nR_c \rfloor}] \times [2^{\lfloor nR \rfloor}]$ to $\mathcal{O}^n$. The distribution induced by the latter code is given by*

$$P^{(n)}_{Z^n,J,M,O^n} := p_Z^{\otimes n} \cdot p^{\mathcal{U}}_{[2^{\lfloor nR_c \rfloor}]} \cdot F^{(n)}_{M|Z^n,J} \cdot G^{(n)}_{O^n|J,M}.$$

Given a quadruplet $(R_A, R_{c,A}, R_B, R_{c,B}) \in \mathbb{R}^4_{\geq 0}$, an $(n, R_A, R_{c,A}, R_B, R_{c,B})$ code is a quadruplet $(F^{(A,n)}, G^{(A,n)}, F^{(B,n)}, G^{(B,n)})$ such that $(F^{(A,n)}, G^{(A,n)})$ (resp. $(F^{(B,n)}, G^{(B,n)})$) is an $(n, R_A, R_{c,A})$ (resp. $(n, R_B, R_{c,B})$) code for $(\mathcal{U}, \mathcal{X})$ (resp. $(\mathcal{V}, \mathcal{Y})$). The distribution induced by the *latter code is given by*

$$P^{(n)}_{U^n,V^n,J_A,J_B,M_A,M_B,X^n,Y^n} := p^{\otimes n}_{U,V} \cdot p^{\mathcal{U}}_{[2^{\lfloor nR_{c,A}\rfloor}]} \cdot p^{\mathcal{U}}_{[2^{\lfloor nR_{c,B}\rfloor}]} \cdot F^{(A,n)}_{M_A|U^n,J_A} \cdot F^{(B,n)}_{M_B|V^n,J_B} \cdot G^{(A,n)}_{X^n|J_A,M_A} \cdot G^{(B,n)}_{Y^n|J_B,M_B}.$$

Definition 2.2. *For a given target distribution $q_{X,Y}$, a tuple $(R_A, R_{c,A}, R_B, R_{c,B}) \in \mathbb{R}^4_{\geq 0}$ is said to be achievable if there exists a sequence of $(n, R_A, R_{c,A}, R_B, R_{c,B})$ codes with induced distributions $\{P^{(n)}\}_n$ such that*

$$\|P^{(n)}_{X^n,Y^n} - q^{\otimes n}_{X,Y}\|_{TV} \underset{n\to\infty}{\longrightarrow} 0. \quad (1)$$

3 Main results

3.1 Achievable region

Theorem 3.1. *All tuples in the interior of the following region $\mathcal{S}$ are achievable.*

$$\left\{ \begin{array}{rcl} (R_A, R_{c,A}, R_B, R_{c,B}) & : & \exists\, p_{U,V,K,W,L_A,L_B,X,Y} \in \mathcal{D}, \\ R_A & \geq & I_p(U,W;L_A) \\ R_A + R_{c,A} & \geq & I_p(U,W,X;L_A) \\ R_B & \geq & I_p(V,W;L_B) \\ R_A + R_{c,B} & \geq & I_p(V,W,Y;L_B) \end{array} \right\},$$

with $\mathcal{D}$ defined as

$$\left\{ \begin{array}{c} p_{U,V,K,W,L_A,L_B,X,Y} : p_{X,Y} \equiv q_{X,Y},\ p_{U,V} \equiv q_{U,V}, \\ W - K - (U,V),\ H(K|U) = H(K|V) = 0 \\ X - L_A - (U,W) - (V,W) - L_B - Y \\ H_p(K) \geq I_p(X,Y;K,W) \end{array} \right\}.$$

The corresponding scheme is depicted in Figure 4, and is closely related to the one in [2]. Encoder A first extracts the GK common part K^n of U^n and V^n, then deterministically maps K^n to a string $w^n(K^n)$. Instead of sampling X^n from $\prod_{t=1}^n p_{X|U=u_t,W=w_t(K^n)}$ locally, the channel is simulated in a distributed manner, between Encoder A and Decoder A. The two other terminals proceed similarly, using the same map $w^n(\cdot)$, and simulate channel $\prod_{t=1}^n p_{Y|V=v_t,W=w_t(K^n)}$.

3.2 Corollaries and comparison to other schemes

The merits of using channel simulation, and leveraging common randomness, can be seen when comparing to certain other classes of schemes.

3.2.1 Schemes based on scalar relations between U^n and X^n and between V^n and Y^n

Consider distributions $q_{U,V}$ and $q_{X,Y}$ such that there exists a coupling p of the latter, satisfying $X - U - V - Y$. This is a special case of the relation $X - (U,W) - (V,W) - V$ of $\mathcal{D}$. Then, it is natural to consider schemes such that the output X^n of Decoder A approximately satisfies the following scalar relation to the input U^n of Encoder A:

$$X^n \sim \prod_{t=1}^n p_{X|U=u_t}.$$

That is, channel simulation schemes. We claim that our scheme performs at least as well as any such scheme. Of course, sampling such a X^n locally at Encoder A would incur a communication cost of $H(X)$, and sending U^n would incur a cost of $H(U)$. Our scheme specializes to the case $X - U - V - Y$ as follows. Consider the subset $\mathcal{D}^{(a)}$ of $\mathcal{D}$ defined as

$$\{p \in \mathcal{D} \mid W \text{ is independent of } (X,Y,U,V,L_A,L_B)\}. \quad (2)$$

Note that every $p \in \mathcal{D}^{(a)}$ satisfies the Markov chain relation $X - L_A - U - V - L_B - Y$. The set $\mathcal{D}^{(a)}$ corresponds to a subset of the achievable region $\mathcal{S}$.

Corollary 3.2. *All tuples in the interior of the following region $\mathcal{S}^{(a)}$ are achievable.*

$$\left\{\begin{array}{rcl} (R_A, R_{c,A}, R_B, R_{c,B}) & : & \exists\, p_{U,V,K,W,L_A,L_B,X,Y} \in \mathcal{D}^{(a)}, \\ R_A & \geq & I_p(U; L_A) \\ R_A + R_{c,A} & \geq & I_p(U, X; L_A) \\ R_B & \geq & I_p(V; L_B) \\ R_A + R_{c,B} & \geq & I_p(V, Y; L_B) \end{array}\right\}$$

The channel simulated in our scheme is $\prod p_{X|U,W} \equiv \prod p_{X|U}$. Note that the inequalities involving R_A and $R_{c,A}$ are exactly the ones in the standard channel simulation result [4, Theorem II.1] for channel $p_{X|U}$: our scheme reduces to a pure channel simulation scheme.

3.2.2 Schemes with a vector encoding of U^n and V^n into a common W^n

Consider distributions $q_{U,V}$ and $q_{X,Y}$ such that there exists a coupling $p_{U,V,X,Y,W}$ satisfying relation $X - W - U - V - W - Y$, and hence $X - W - Y$. This is a special case of the relation $X - (U, W) - (V, W) - V$ of $\mathcal{D}$. Consider schemes involving a vector operation $U^n \mapsto W^n$, and the sampling of X^n approximately from $\prod_{t=1}^n p_{X|W=w_t}$. Assume that the vector operation is not randomized, so as not to impede Encoder B and Decoder B from using the same realization of W^n, as Enocder A and Decoder A. Assume that the set of possible realizations of W^n is limited to some relatively small codebook. Then, it may seem efficient to compute W^n at Encoder A, and send the corresponding index to Decoder A, which samples X^n using memoryless channel $\prod p_{X|W}$. The converse to the soft covering lemma [4, Appendix 4] implies that, for a given channel $p_{X|W}$, in order to satisfy (1), the codebook of w^n strings must have a rate of at least $\min I_p(X, Y; W)$. The minimum is over all p_W such that $\sum_w p_W(w) p_{X|W=w} p_{Y|W=w} \equiv q_{X,Y}$. When restricted to distributions $p \in \mathcal{D}$ satisfying Markov chain relation $X - W - Y$, our scheme specializes to a distributed channel simulation scheme for $\prod p_{X|W}$ - and a similar scheme for $\prod p_{Y|W}$ -, which may be more efficient, as discussed in the remainder of this section. Consider the subset $\mathcal{D}^{(d)}$ of $\mathcal{D}$ defined as

$$\begin{aligned} \{p \in \mathcal{D} \mid\ & X - L_A - W - L_B - Y \\ & \text{and } (X, L_A, L_B, Y) - W - K - (U, V)\}. \end{aligned} \tag{3}$$

The set $\mathcal{D}^{(d)}$ corresponds to a subset of the achievable region $\mathcal{S}$.

Corollary 3.3. *All tuples in the interior of the following region $\mathcal{S}^{(d)}$ are achievable.*

$$\left\{\begin{array}{rcl} (R_A, R_{c,A}, R_B, R_{c,B}) & : & \exists\, p_{U,V,K,W,L_A,L_B,X,Y} \in \mathcal{D}^{(d)}, \\ R_A & \geq & I_p(W; L_A) \\ R_A + R_{c,A} & \geq & I_p(W, X; L_A) \\ R_B & \geq & I_p(W; L_B) \\ R_A + R_{c,B} & \geq & I_p(W, Y; L_B) \end{array}\right\}$$

The channel simulated in our scheme is $\prod p_{X|U,W} \equiv \prod p_{X|W}$. Note that the inequalities involving R_A and $R_{c,A}$ are exactly the ones in the standard channel simulation result for channel $p_{X|W}$. If $R_{c,A}$ is large enough, then a rate $R_A = I(W; X)$ is sufficient -which is less or equal to $I(W; L_A)$ by the data processing inequality. This can be achieved by applying our scheme to a modified version of p, obtained by replacing L_A with $\tilde{L}_A = X$. Hence, when common randomness is available at a sufficient rate in each encoder-decoder pair, the rate required for our channel simulation scheme is lower ($I(W; X) \leq I(X, Y; W)$) than the rate required for the aforementioned scheme, in which the same string W^n is conveyed to both decoders.

4 Main components of the proof of Theorem 3.1

The problem studied in [2] is the special case where $R_A = R_B = \infty$ and $R_{c,A} = R_{c,B} = 0$, where the notation $R_A = R_B = \infty$ refers to any rates for which zero-error compression is possible, i.e. $R_A > \log(|\mathcal{U}|), R_B > \log(|\mathcal{V}|)$. Then, Encoder A and Decoder A behave as a single terminal, which we shall call Terminal A, and Terminal B is defined similarly. Using the notation in the present paper, we can reformulate [2, Theorem 2] as follows.

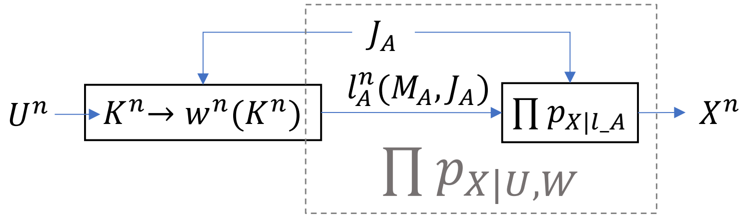

Figure 4: Coding scheme.

Theorem 4.1. *Consider any $p \in \mathcal{D}$. Then, tuple $(\infty, 0, \infty, 0)$ is achieved by the following scheme.*

- *Terminal A passes its input U^n through channel $\prod p_{K|U}$ (which is a deterministic mapping), resulting in K^n.*
- *It then passes K^n through a well-chosen deterministic mapping $w^n(\cdot)$.*
- *Terminal B proceeds in the same manner, using the same mapping $w^n(\cdot)$, resulting in the same realizations of K^n and $w^n(K^n)$ as obtained at Terminal A.*
- *Terminal A samples X^n from $\prod_{t=1}^n p_{X|U=U_t,W=w_t(K^n)}$.*
- *Terminal B samples Y^n from $\prod_{t=1}^n p_{Y|V=V_t,W=w_t(K^n)}$.*

In order to prove Theorem 3.1, we use the same scheme, but replace the last two steps with a channel simulation protocol: given U^n and $w^n(K^n)$ at Encoder A, the latter sends a message at rate R_A allowing the decoder to produce X^n distributed approximately according to $\prod_{t=1}^n p_{X|U=U_t,W=w_t(K^n)}$. Encoder B and Decoder B proceed similarly. One can find a channel simulation protocol of adequate communication and common randomness rates, provided that $(U^n, w^n(K^n))$ is a typical sequence, as formalized in the following result. Regarding notation, we denote the empirical distribution, which we also call *type*, of a string $z^n \in \mathcal{Z}^n$ as $t_{\mathcal{Z}}(z^n)$, and the set of all types as

$$\mathcal{T}^{(n)}(\mathcal{Z}) := \left\{ \{t_z\}_{z\in\mathcal{Z}} \;\middle|\; \exists z^n \in \mathcal{Z}^n, t_{\mathcal{Z}}(z^n) = \{t_z\}_{z\in\mathcal{Z}} \right\}.$$

We use similar notation for joint types. We use the notion of typical sets defined in [5], and the notation $\mathcal{T}_\varepsilon^{(n)}(p_Z)$, instead of the similar notation in [5].

Proposition 4.2. *Let $\mathcal{Z}, \mathcal{L}, \mathcal{O}$ be three finite alphabets, and $p_{Z,L,O}$ be a distribution on their product, satisfying Markov chain $Z - L - O$. For any $\varepsilon > 0$, there exists a real $\delta > 0$ and an integer $N \geq 1$ such that, for any integer $n \geq N$, there exists a $(n, I_p(Z;L) + \varepsilon, I_p(Z,O;L) - I_p(Z;L) + \varepsilon)$ code for $(\mathcal{Z}, \mathcal{O})$, such that for any $z^n \in \mathcal{T}_\delta^{(n)}(p_Z)$,*

$$\left\| P^{(n)}_{O^n|Z^n=z^n} - \prod_{t=1}^n p_{O|Z=z_t} \right\|_{TV} \leq \varepsilon. \tag{4}$$

Proposition 4.2 is a generalization of the standard channel simulation result in [4], which only pertains to the problem of identifying the rates for which it is possible to achieve a vanishing average of (4) over $Z^n \sim p_Z^{\otimes n}$. Proposition 4.2 is a corollary of the proofs from [1], where the problem of interest is identifying the rates for which it is possible to achieve a vanishing average of (4) under all possible distributions for Z^n, with a single encoder-decoder pair. The rate region appearing in Proposition 4.2 is optimal for the corresponding channel simulation task. Due to space limitations, we defer the remainder of the proof to the Appendix.

5 Conclusions

We have considered a non-interactive remote coordination problem, where two non-interacting controllers observe correlated sources, and can each communicate with a decoder, with the aim of helping the two decoders to generate coordinated actions. We have uncovered sufficient conditions on the compression rates, common randomness rates, and the distribution of inputs, in order for a given target output distribution to be asymptotically simulable. In particular, these are sufficient conditions to allow for *enough* correlation to carry over from the sources to the remote outputs. These correspond to a scheme we propose, which relies on a special channel simulation protocol, to generalize prior work. We compared to two simpler kinds of schemes, and showed that our scheme leverages the correlation between the sources more efficiently. Thereby, we exhibited the role that common randomness between a controller and its decoder can play for rate efficiency. While we have focused on the theoretical aspects of non-interactive remote coordination in this paper, we believe the tools developed here can find applications in various machine learning problems. These include the generation of coordinated policies among distributed agents in multi-agent reinforcement learning problems, or neural based distributed compression of correlated images with the goal of realism, where the realism will require the reconstructed images to follow the joint distribution of the original images. It is relevant to mention that, in rare cases, the latter kind of applications, which may prioritize realism over fidelity to the original images, may lead to negative societal impacts, such as false identity recognition.

Acknowledgments and Disclosure of Funding

This research was supported by the European Union's Horizon 2020 program through Marie Skłodowska Curie Innovative Training Network Greenedge (No. 953775), and by the United Kingdom Engineering and Physical Sciences Research Council (EPSRC) for the projects AIR (ERC Consolidator Grant, EP/X030806/1) and INFORMED-AI (EP/Y028732/1).

References

[1] C. H. Bennett, I. Devetak, A. W. Harrow, P. W. Shor, and A. Winter. The quantum reverse shannon theorem and resource tradeoffs for simulating quantum channels. *IEEE Transactions on Information Theory*, 60(5):2926–2959, 2014.

[2] T. Berg, O. Shayevitz, Y. H. Kim, and L. Wang. Distributed source simulation with no communication. *IEEE Transactions on Information Theory*, 67(2):886–901, 2021.

[3] Mingzhe Chen, Deniz Gündüz, Kaibin Huang, Walid Saad, Mehdi Bennis, Aneta Vulgarakis Feljan, and H Vincent Poor. Distributed learning in wireless networks: Recent progress and future challenges. *IEEE Journal on Selected Areas in Communications*, 39(12):3579–3605, 2021.

[4] P. Cuff. Distributed Channel Synthesis. *IEEE Transactions on Information Theory*, 59(11), 2013.

[5] Abbas El Gamal and Young-Han Kim. *Network Information Theory*. Cambridge University Press, Cambridge (England), 2011.

[6] D'Ambrosio et. al. Achieving Human Level Competitive Robot Table Tennis. *arXiv preprint arXiv:2408.03906*, 2024.

[7] Peter Gács, Janos Korner, et al. Common information is far less than mutual information. *Problems of Control and Information Theory*, 2:149–162, 1973.

[8] T.S. Han and S. Verdu. Approximation theory of output statistics. *IEEE Transactions on Information Theory*, 39(3):752–772, 1993.

[9] Ruiquan Huang, Yuan Cheng, Jing Yang, Vincent Tan, and Yingbin Liang. Provable benefits of multi-task RL under non-markovian decision making processes. In *The Twelfth International Conference on Learning Representations*, 2024.

[10] Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis. *arXiv preprint arXiv:2405.07987*, 2024.

[11] A. Wyner. The common information of two dependent random variables. *IEEE Transactions on Information Theory*, 21(2):163–179, 1975.

[12] Selim F Yilmaz, Xueyan Niu, Bo Bai, Wei Han, Lei Deng, and Deniz Gündüz. High perceptual quality wireless image delivery with denoising diffusion models. In *IEEE INFOCOM 2024-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pages 1–5. IEEE, 2024.

A Continuation of the proof of Theorem 3.1

After stating basic lemmas on the TVD in Section A.1, we prove a technical lemma in Section A.2. Then, we prove Proposition 4.2 in Section A.3, and Theorem 3.1 in Section A.4.

A.1 Lemmas regarding the total variation distance

Lemma A.1. *[4, Lemma V.1] Let Π and Γ be two distributions on an alphabet $\mathcal{W} \times \mathcal{L}$. Then*

$$\|\Pi_W - \Gamma_W\|_{TV} \leq \|\Pi_{W,L} - \Gamma_{W,L}\|_{TV}.$$

Lemma A.2. *[4, Lemma V.2] Let Π and Γ be two distributions on an alphabet $\mathcal{W} \times \mathcal{L}$. Then when using the same channel $\Pi_{L|W}$ we have*

$$\|\Pi_W \Pi_{L|W} - \Gamma_W \Pi_{L|W}\|_{TV} = \|\Pi_W - \Gamma_W\|_{TV}.$$

A.2 On joint typicality in Theorem 4.1

Lemma A.3. *In the scheme which yields Theorem 4.1, for any $\delta \in (0,1)$, the deterministic mappings $\{w^n(\cdot)\}_{n\geq 1}$ can be chosen such that the probability of the event*

$$(U^n, V^n, K^n, w^n(K^n)) \notin \mathcal{T}_\delta^{(n)}(p_{U,V,K,W}) \tag{5}$$

vanishes as n goes to infinity.

Proof: The proof of Theorem 4.1, i.e. that of [2, Theorem 2], consists in showing the following. Let $P^{(n,w^n(\cdot))}$ denote the distribution induced by the scheme described in Theorem 4.1, with a choice of mapping $w^n(\cdot)$. Then,

$$\mathbb{E}_{\mathcal{C}^{(n)}}\left[\|P_{X^n,Y^n}^{(n,\mathcal{C}^{(n)})} - q_{X,Y}^{\otimes n}\|_{TV}\right] \underset{n\to\infty}{\longrightarrow} 0. \tag{6}$$

where $\mathcal{C}^{(n)}$ is a family $\{W_{k^n}^n\}_{k^n\in\mathcal{K}^n}$ of mutually independent variables, with the variable of index k^n being distributed according to $\prod_{t=1}^n p_{W|K=k_t}$. The latter proof is concluded by choosing any realization if $\mathcal{C}^{(n)}$ achieving a TVD below average. See the end of [2, Section III-C] and [2, Equation 45]. Therefore, to guarantee the existence of a realization of $\mathcal{C}^{(n)}$ which also satisfies (5), it only remains to show that the probability that $(U^n, V^n, K^n, W_{K^n}^n)$ is not δ-typical vanishes as n goes to infinity. Fix $\delta \in (0,1)$, and $\delta' < \delta$. Then, we have

$$\begin{aligned}
&\mathbb{P}_{\mathcal{C}^{(n)},U^n,V^n,K^n}\Big((U^n, V^n, K^n, W_{K^n}^n) \notin \mathcal{T}_\delta^{(n)}(p)\Big)\\
&\leq p_{U,V,K}^{\otimes n}\Big((U^n, V^n, K^n) \notin \mathcal{T}_{\delta'}^{(n)}(p)\Big) + \sup_{(u^n,v^n,k^n)\in\mathcal{T}_{\delta'}^{(n)}}\\
&\Big(\prod_{t=1}^n p_{W|U=u_t,V=v_t,K=k_t}\Big)\Big((u^n, v^n, k^n, W_{k^n}^n) \notin \mathcal{T}_\delta^{(n)}(p)\Big).
\end{aligned}$$

The first probability in the right hand side vanishes as n goes to infinity, by definition of the typical set. We show that the second probability is upper bounded uniformly over $\mathcal{T}_{\delta'}^{(n)}$, by a term which vanishes as n goes to infinity. Fix $(u_0, v_0, k_0, w_0) \in \mathcal{U} \times \mathcal{V} \times \mathcal{K} \times \mathcal{W}$, such that $p_{Z,L,O}(u_0, v_0, k_0, w_0) > 0$. The probability according to $\prod_{t=1}^n p_{W|U=u_t,V=v_t,K=k_t}$ of the following event

$$\begin{aligned}
\Big|&\#\{i \in [n] | (u_i, v_i, k_i, W_i) = (u_0, v_0, k_0, w_0)\}/n\\
&-p_{Z,L,O}(z_0, l_0, o_0)\Big| \leq \delta p_{Z,L,O}(z_0, l_0, o_0)
\end{aligned} \tag{7}$$

is upper bounded, in [5, Appendix 2A], by a term of the form $\phi(a_{u^n,v^n,k^n})$. Map $\phi : \mathbb{N}_{\geq 0} \to [0,\infty)$ is a deterministic function which vanishes at infinity, and a_{u^n,v^n,k^n} is the number of occurrences of (u_0, v_0, k_0) in (u^n, v^n, k^n). We know that typicality entails $a_{u^n,v^n,k^n} \geq (1-\delta')np_{U,V,K}(u_0, v_0, k_0)$, for every (u^n, v^n, k^n) in $\mathcal{T}_{\delta'}^{(n)}(p_{U,V,K})$ [5], and we have assumed $p_{Z,L,O}(u_0, v_0, k_0) > 0$. Therefore, the probability of the event in (7) tends to 0 as n goes to infinity, uniformly over all (u^n, v^n, k^n) in $\mathcal{T}_{\delta'}^{(n)}(p_{U,V,K})$. A union bound over all $(u_0, v_0, k_0, w_0) \in \mathcal{U} \times \mathcal{V} \times \mathcal{K} \times \mathcal{W}$ concludes the proof. ■

A.3 Proof of Proposition 4.2

We use the following result, from the main proof in [1] - precisely the end of Section III.C therein. It pertains to the existence of a good channel simulation scheme tailored to a given joint type. To prove Proposition 4.2, we then use this only for typical types.

Proposition A.4. *For every $\varepsilon>0$, there exists a positive integer N_ε, such that for all $n\geq N_\varepsilon$, there exists a family $\{(F^{(n,\tau)}, G^{(n,\tau)})\}_{\tau\in\mathcal{T}^{(n)}(\mathcal{Z}\times\mathcal{L}\times\mathcal{O})}$ of codes, with that of index τ being a $(n, I_\tau(Z;L)+\varepsilon, I_\tau(Z,O;L)-I_\tau(Z;L)+\varepsilon)$ code for $(\mathcal{Z},\mathcal{O})$, such that the following holds. Consider the following code for $(\mathcal{Z},\mathcal{O})$, where common randomness is assumed to be available at rate R_c satisfying*

$$\forall\tau\in\mathcal{T}^{(n)}(\mathcal{Z}\times\mathcal{L}\times\mathcal{O}),\ R_c\geq I_\tau(Z,O;L)-I_\tau(Z;L)+\varepsilon := R_{c,\tau}.$$

- *From its input z^n, the encoder samples $(\tilde{L}^n,\tilde{O}^n)\sim\prod_{t=1}^n p_{L,O|Z=z_t}$, and sends the joint type $\tau := t_{\mathcal{Z}\times\mathcal{L}\times\mathcal{O}}(z^n,\tilde{L}^n,\tilde{O}^n)$. There shall be no further use of $\tilde{L}^n,\tilde{O}^n$.*
- *Using $F^{(n,\tau)}$ and the integer corresponding to the first $\lfloor nR_{c,\tau}\rfloor$ bits of the common randomness' binary representation, it then encodes z^n and sends the resulting message.*
- *The decoder (which has received τ) uses $G^{(n,\tau)}$, the received message, and the integer corresponding to the first $\lfloor nR_{c,\tau}\rfloor$ bits of the common randomness' binary representation, to output a string O^n.*

Then, for any $n\geq N_\varepsilon$, and every z^n satisfying $p_Z^{\otimes n}(z^n)>0$, the above scheme results in a conditional distribution $P^{(n)}_{O^n|Z^n=z^n}$ such that

$$\left\|P^{(n)}_{O^n|Z^n=z^n}-\prod_{t=1}^n p_{O|Z=z_t}\right\|_{TV}\leq\varepsilon. \tag{8}$$

We proceed to proving Proposition 4.2. Let $p_{Z,L,O}$ be as in the latter. Fix some $\varepsilon>0$, and let $N_\varepsilon, R_c, \{P^{(n)}\}_{n\geq 1}$ be the corresponding objects from Proposition A.4. We use the scheme of the latter whenever $(z^n,\tilde{L}^n,\tilde{O}^n)\in\mathcal{T}_\delta^{(n)}(p_{Z,L,O})$, for some small enough $\delta>0$ to be specified later. Otherwise, the encoder sends an error message, and the decoder produces a default output. The remainder of this section is dedicated to analyzing this scheme. From Proposition A.4, the definition of $\mathcal{T}_\delta^{(n)}(p_{Z,L,O})$ [5], and the continuity of mutual information, there exists a choice of δ such that for any $n\geq N_\varepsilon$, our scheme defines a $(n, I_p(Z;L)+2\varepsilon, I_p(Z,O;L)-I_p(Z;L)+2\varepsilon)$ code for $(\mathcal{Z},\mathcal{O})$. Let $\tilde{P}^{(n)}_{Z^n,J,M,O^n,\tilde{L}^n,\tilde{O}^n}$ denote the distribution induced by this code, and let $\tilde{R}_c := I_p(Z,O;L)-I_p(Z;L)+2\varepsilon$. For any $j\in[2^{nR_c}]$, let $j_{\lfloor n\tilde{R}_c\rfloor}$ denote the sequence of first $\lfloor n\tilde{R}_c\rfloor$ bits of j. For any $n\geq N_\varepsilon$, any $\tilde{j}\in[2^{n\tilde{R}_c}]$, and any $(z^n,\tilde{l}^n,\tilde{o}^n)\in\mathcal{T}_\delta^{(n)}(p_{Z,L,O})$, we have

$$\begin{aligned}&\tilde{P}^{(n)}_{M,O^n|Z^n=z^n,\tilde{L}^n=\tilde{l}^n,\tilde{O}^n=\tilde{o}^n,J=\tilde{j}}\\ &\equiv P^{(n)}_{M,O^n|Z^n=z^n,\tilde{L}^n=\tilde{l}^n,\tilde{O}^n=\tilde{o}^n,J_{\lfloor n\tilde{R}_c\rfloor}=\tilde{j}},\end{aligned} \tag{9}$$

$$\begin{aligned}&\text{and}\ \ \tilde{P}^{(n)}(Z^n=z^n,\tilde{L}^n=\tilde{l}^n,\tilde{O}^n=\tilde{o}^n,J=\tilde{j})\\ &=P^{(n)}(Z^n=z^n,\tilde{L}^n=\tilde{l}^n,\tilde{O}^n=\tilde{o}^n,J_{\lfloor n\tilde{R}_c\rfloor}=\tilde{j}).\end{aligned} \tag{10}$$

Indeed, under both $\tilde{P}^{(n)}$ and $P^{(n)}$, the common randomness J is independent from $(Z^n,\tilde{L}^n,\tilde{O}^n)$. Thus, conditioned on a realization of the latter, for any $k\leq R_c$, the distribution of the first k bits of J is independent of R_c. From (9), (10), Proposition A.4, and the triangle inequality for the TVD, for any $n\geq N_\varepsilon$ and any $z^n\in\mathcal{Z}^n$ such that $p_Z^{\otimes n}(z^n)>0$, we have

$$\begin{aligned}&\left\|\tilde{P}^{(n)}_{O^n|Z^n=z^n}-\textstyle\prod_{t=1}^n p_{O|Z=z_t}\right\|_{TV}\leq\\ &\left\|P^{(n)}_{O^n|Z^n=z^n}-\textstyle\prod_{t=1}^n p_{O|Z=z_t}\right\|_{TV}+\left\|\tilde{P}^{(n)}_{O^n|Z^n=z^n}-P^{(n)}_{O^n|Z^n=z^n}\right\|_{TV}\\ &\leq\varepsilon+\left(\textstyle\prod_{t=1}^n p^{\otimes n}_{L,O|Z=z_t}\right)\left((z^n,\tilde{L}^n,\tilde{O}^n)\notin\mathcal{T}_\delta^{(n)}(p_{Z,L,O})\right).\end{aligned} \tag{11}$$

For any given $\delta>0$, the probability in the right hand side of (11) is upper bounded uniformly over all z^n in $\mathcal{T}_{\delta/2}^{(n)}(p_Z)$, by a term which vanishes as n goes to infinity. This is shown, albeit with different notation, in the proof of Lemma A.3. Since this analysis is valid for any ε, then Proposition 4.2 is proved.

A.4 Proof of Theorem 3.1

A.4.1 Scheme and relation to Theorem 4.1

Fix $(R_A, R_{c,A}, R_B, R_{c,B})$ in the interior of $\mathcal{S}$. Then, there exists some $\varepsilon^* > 0$ such that for any $\varepsilon < \varepsilon^*$, we have $(R_A - \varepsilon, R_{c,A} - \varepsilon, R_B - \varepsilon, R_{c,B} - \varepsilon) \in \mathcal{S}$. Consider $p \in \mathcal{D}$ corresponding to the latter tuple. Since p satisfies $(U, W) - L_A - X$, Proposition 4.2 applies. Let $\delta_A, N_A, \{(F^{(A,n,1)}, G^{(A,n,1)})\}_{n \geq 1}$ be the objects from the latter, corresponding to ε. Note that for any $n \geq N_A$, $(F^{(A,n,1)}, G^{(A,n,1)})$ is a $(n, R_A, R_{c,A})$ code for $(\mathcal{U} \times \mathcal{W}, \mathcal{X})$. Define $\delta_B, N_B, \{(F^{(B,n,1)}, G^{(B,n,1)})\}_{n \geq 1}$ similarly. We apply Theorem 4.1, with the minor improvement of Lemma A.3, for $\delta = \min(\delta_A, \delta_B)$. Let $P^{(n,2)}_{U^n,V^n,K^n,W^n,X^n,Y^n}$ denote the distribution induced by the corresponding scheme. Then, there exists $N_1 \in \mathbb{N}$ such that for any $n \geq N_1$, we have

$$\|P^{(n,2)}_{X^n,Y^n} - q^{\otimes n}_{X,Y}\|_{TV} \leq \varepsilon. \tag{12}$$

Let $N_2 := \max(N_1, N_A, N_B)$. For any $n \geq N_2$, consider the following $(n, R_A, R_{c,A}, R_B, R_{c,B})$ code, with induced distribution denoted $P^{(n,1)}$.

- Using its input U^n, Encoder A samples (K^n, W^n) using $P^{(n,2)}_{K^n,W^n|U^n}$.
- If $(U^n, W^n) \in \mathcal{T}^{(n)}_{\delta_A}(p_{U,W})$, then Encoder A chooses a message using $F^{(A,n,1)}$, then Decoder A uses the message and $G^{(A,n,1)}$ to produce an output string X^n.
- Otherwise, Encoder A sends an error message, and Decoder A outputs a default string.
- Encoder B and Decoder B follow similar steps.

From (4), for any $n \geq N_2$, for all $(u^n, w^n) \in \mathcal{T}^{(n)}_{\delta_A}(p_{U,W})$, we have

$$\left\| P^{(n,1)}_{X^n|U^n=u^n,W^n=w^n} - \prod_{t=1}^n p_{X|U=u_t,W=w_t} \right\|_{TV} \leq \varepsilon,$$

i.e.

$$\left\| P^{(n,1)}_{X^n|U^n=u^n,W^n=w^n} - P^{(n,2)}_{X^n|U^n=u^n,W^n=w^n} \right\|_{TV} \leq \varepsilon. \tag{13}$$

Then, since $P^{(n,1)}_{U^n,W^n} \equiv P^{(n,2)}_{U^n,W^n}$, we have

$$\begin{aligned} &\left\| P^{(n,1)}_{U^n,X^n} - P^{(n,2)}_{U^n,X^n} \right\|_{TV} \\ &\leq \left\| P^{(n,1)}_{U^n,W^n,X^n} - P^{(n,2)}_{U^n,W^n,X^n} \right\|_{TV} &(14)\\ &\leq \varepsilon + P^{(n,2)}_{U^n,W^n}\big((U^n, W^n) \notin \mathcal{T}^{(n)}_{\delta_A}(p_{U,W})\big) &(15)\\ &\leq 2\varepsilon, &(16)\end{aligned}$$

where (14) follows from Lemma A.1; (15) follows form (13) and the triangle inequality for the TVD; and (16) holds for large enough n, and follows from (5) and the fact that $\delta \leq \delta_A$. The same inequality holds for variables (V^n, Y^n), for large enough n.

A.4.2 Proof conclusion

Hence, since $P^{(n,1)}$ and $P^{(n,2)}$ satisfy Markov chain $X^n - U^n - V^n - Y^n$, then we have

$$\begin{aligned} &\left\| P^{(n,1)}_{U^n,V^n,X^n,Y^n} - P^{(n,2)}_{U^n,V^n,X^n,Y^n} \right\|_{TV} \\ &= \left\| P^{(n,1)}_{U^n,X^n} \cdot P^{(n,1)}_{V^n,Y^n|U^n} - P^{(n,2)}_{V^n,Y^n} \cdot P^{(n,2)}_{U^n,X^n|V^n} \right\|_{TV} \\ &= \Big\| P^{(n,1)}_{U^n,X^n} \cdot P^{(n,1)}_{V^n,Y^n|U^n} - P^{(n,2)}_{U^n,X^n} \cdot P^{(n,1)}_{V^n,Y^n|U^n} \\ &+ P^{(n,1)}_{V^n,Y^n} \cdot P^{(n,2)}_{U^n,X^n|V^n} - P^{(n,2)}_{V^n,Y^n} \cdot P^{(n,2)}_{U^n,X^n|V^n} \Big\|_{TV} \\ &\leq 2\varepsilon + 2\varepsilon, &(17)\end{aligned}$$

where (17) follows from the triangle inequality for the TVD, Lemma A.2, and (16). This yields, for large enough n,

$$\begin{aligned}
&\|P^{(n,1)}_{X^n,Y^n} - q^{\otimes n}_{X,Y}\|_{TV} \\
&\leq \big\|P^{(n,1)}_{X^n,Y^n} - P^{(n,2)}_{X^n,Y^n}\big\|_{TV} + \|P^{(n,2)}_{X^n,Y^n} - q^{\otimes n}_{X,Y}\|_{TV} \qquad (18)\\
&\leq \big\|P^{(n,1)}_{U^n,V^n,X^n,Y^n} - P^{(n,2)}_{U^n,V^n,X^n,Y^n}\big\|_{TV} + \varepsilon \leq 5\varepsilon, \qquad (19)
\end{aligned}$$

where (18) follows from the triangle inequality for the TVD; and (19) follows from Lemma A.1, (12), and (17). Since this construction and analysis are valid for any $\varepsilon \in (0, \varepsilon^*)$, then $(R_A, R_{c,A}, R_B, R_{c,B})$ is achievable, which concludes the proof of Theorem 3.1.