# COMMUNICATION IN NETWORKS FOR COORDINATING BEHAVIOR

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF ELECTRICAL ENGINEERING
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Paul W. Cuff
July 2009

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Thomas M. Cover) Principal Adviser

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Tsachy Weissman)

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Abbas El Gamal)

Approved for the University Committee on Graduate Studies.

# Preface

The theory of information, as charted by Claude Shannon, has influenced many fields and introduced a new way of thinking about information. Beyond the immediate applications to digital communication, the ideas introduced by information theory have found their way into the study of biology and DNA, computation and complexity, and machine learning. Indeed, the field provides a concrete way of dealing with the otherwise nebulous substance of information. Although many of the basic questions have been answered, even by Shannon himself, many fundamental questions dealing with multiple users and multiple sources of information have remained unanswered for decades. We are still lacking in our understanding of how to best structure and correlate codebooks for communicating in network settings. Answers to the building blocks of network information theory provide insight into how information can be reinforced in a complex setting, ultimately giving us principles for better technology and greater understanding.

Beyond the intriguing array of classical open problems related to communicating in networks, the field itself is open to broader interpretation. By considering new purposes for communication, besides transporting information from one location to another, we find many new problems of interest.

In this work, we develop elements of a theory of coordination in networks. With rate-limited communication between the nodes of a network, we ask for the set of all possible joint distributions $p(x_1, ..., x_m)$ of actions at the nodes. Several networks are solved, including arbitrarily large cascade networks. Distributed coordination can be the solution to many problems such as distributed games, distributed control, and establishing bounds on the physical influence of one part of a system on another.

# Acknowledgement

Larisa and I made a decision in 2004 to move out to California, live in cramped graduate student housing, raise a small family while we're at it, and enjoy a Stanford education. First and foremost I thank her for giving me the opportunity.

It was clear from the beginning of graduate school that I would enjoy going deeper in familiar areas of science, math, and engineering. But it came as a surprise that my depth would ultimately be developed in a field completely unknown to me at the start. The door to an entirely new way of thinking was hiding behind a few well-taught lectures and fascinating conversations. If anyone wonders how I lost focus during my time at Stanford, I have Tom Cover to blame, for leaking the best-kept secrets of applied math and engineering. He taught me not only information theory but to continually look for the big picture and let curiosity guide. He made my grad school experience.

A number of other mentors will have a lasting influence on me, including Tsachy Weissman, Abbas El Gamal, Balaji Prabhakar, and Bernard Widrow. We've had fun adventures and collaborations together.

For all the times I needed a friend, colleague, baby-sitter, or brain, Haim Permuter was there from day one. We started graduate school together, and not long after I introduced him to information theory he was trying to get me to commit as well. Each member of Tom Cover's and Tsachy Weissman's research groups became a close friend. Young-Han Kim in particular served not only as a friend but also as an adviser for career and research decisions.

# Contents

# Chapter 1

# Coordination of Actions

Communication is required to establish cooperative behavior. In a network of nodes where relevant information is known at only some nodes in the network, finding the minimum communication requirements to coordinate actions can be posed as a network source coding problem. This breaks somewhat from traditional source coding. Rather than focusing on sending data from one point to another with a fidelity constraint, we can consider the communication needed to establish coordination summarized by a joint probability distribution of behavior among all nodes in the network.

A large variety of research addresses the challenge of collecting or moving information in networks. Network coding [1] seeks to efficiently move independent flows of information over shared communication links. On the other hand, distributed average consensus [2] involves collecting related information. Sensors in a network collectively compute the average of their measurements in a distributed fashion. The network topology and dynamics determine how many rounds of communication among neighbors are needed to converge to the average and how good the estimate will be at each node [3]. Similarly, in the gossiping Dons problem [4], each node starts with a unique piece of gossip, and one wishes to know how many exchanges of gossip are required to make everything known to everyone. Computing functions in a network is considered in [5], [6], and [7].

Our work has several distinctions from the network communication examples mentioned. First, we keep the purpose for communication very general, which means

sometimes we get away with saying very little about the information in the network while still achieving the desired coordination. We are concerned with the joint distribution of actions taken at the various nodes in the network, and the "information" that enters the network is nothing more than actions that are selected randomly by nature and assigned to certain nodes. Secondly, we consider quantization and rates of communication in the network, as opposed to only counting the number of exchanges. We find that we can gain efficiency by using vector quantization specifically tailored to the network topology.

Figure 1.1 shows an example of a network with rate-limited communication links. In general, each node in the network performs an action where some of these actions are selected randomly by nature. In this network, the source set $\mathcal{S}$ indicates which actions are chosen by nature: Actions $X_1$, $X_2$, and $X_3$ are assigned randomly according to the joint distribution $p_0(x_1, x_2, x_3)$. Then, using the communication and common randomness that is available to all nodes, the actions $Y_1$, $Y_2$, and $Y_3$ outside of $\mathcal{S}$ are produced. We ask, which conditional distributions $p(y_1, y_2, y_3|x_1, x_2, x_3)$ are compatible with the network constraints.

A variety of applications are encompassed in this framework. This could be used to model sensors in a sensor network, sharing information in the standard sense, while also cooperating in their transmission of data. Similarly, a wireless ad hoc network can improve performance by cooperating among nodes to allow beam-forming and interference alignment. On the other hand, some settings do not involve moving information in the usual sense. The nodes in the network might comprise a distributed control system, where the behavior at each node must be related to the behavior at other nodes and the information coming into the system. Also, with computing technology continuing to move in the direction of parallel processing, even across large networks, a network of computers must coherently perform computations while distributing the work load across the participating machines. Alternatively, the nodes might each be agents taking actions in a multiplayer game.

Network communication can be revisited from the viewpoint of coordinated actions. Rate distortion theory becomes a special case. More generally, we ask how we can build dependence among the nodes. What is it good for? How do we use it?

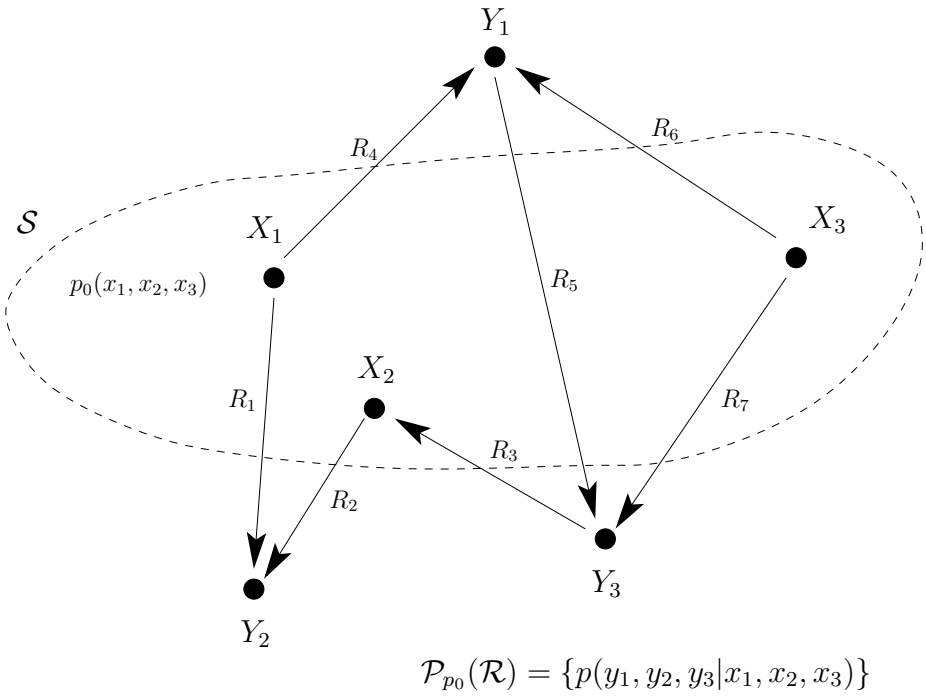


Figure 1.1: *Coordination capacity.* This network represents the general framework we consider. The nodes in this network have rate-limited links of communication between them. Each node performs an action. The actions $X_1$, $X_2$, and $X_3$ in the source set $\mathcal{S}$ are chosen randomly by nature according to $p_0(x_1, x_2, x_3)$, while the actions $Y_1$, $Y_2$, and $Y_3$ are produced based on the communication and common randomness in the network. What joint distributions $p_0(x_1, x_2, x_3)p(y_1, y_2, y_3|x_1, x_2, x_3)$ can be achieved?

A desired joint distribution of actions in the network is said to be achievable for *empirical coordination* if under the communication constraints in the network the empirical joint distribution of the actions at all of the nodes, over multiple instances, can be made statistically indistinguishable (as measured by total variation) from the desired distribution. In situations where average behavior over time is the concern, the empirical joint distribution captures the network's performance. On the other hand, a desired joint distribution is achievable for *strong coordination* if the actions can be generated randomly to match the desired distribution with negligible total variation.

Before developing the mathematical formulation, consider the first surprising observation.

*No communication:* Suppose we have three nodes choosing actions and no communication is allowed between the nodes (Fig. 1.2). We assume that common randomness is available to all the nodes. What is the set of joint distributions $p(x, y, z)$ that can be achieved at these isolated nodes? The answer turns out to be any joint distribution whatsoever. The nodes can agree ahead of time on how they will behave in the presence of common randomness (for example, a time stamp used as a seed for a random number generator). Any triple of random variables can be created as functions of common randomness.

This would seem to be the end of the problem, but the problem changes dramatically when one of the nodes is specified by nature to take on a certain value, as will be the case in each of the scenarios following.

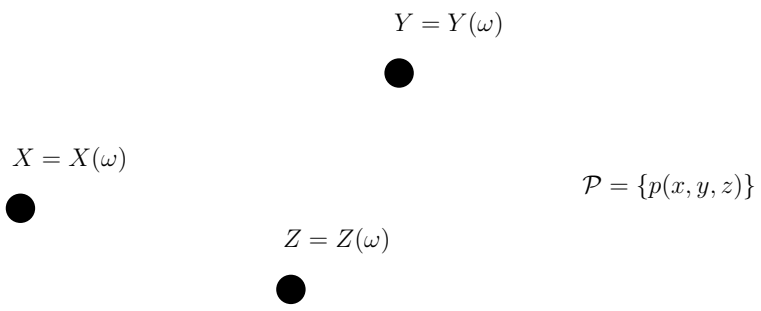


Figure 1.2: *No communication.* Any distribution $p(x, y, z)$ can be achieved without communication between nodes. Define three random variables $X(\cdot)$, $Y(\cdot)$, and $Z(\cdot)$ with the appropriate joint distribution, on the standard probability space $(\Omega, \mathcal{B}, \mathcal{P})$, and let the actions at the nodes be $X(\omega)$, $Y(\omega)$, and $Z(\omega)$, where $\omega \in \Omega$ is the common randomness.

An eclectic collection of work, ranging from game theory to quantum information theory, has a number of close relationships to our approach and results. For example, Anantharam and Borkar [8] let two agents generate actions for a multiplayer game based on correlated observations and common randomness and ask what kind of correlated actions are achievable. From a quantum mechanics perspective, Barnum et. al. [9] consider quantum coding of mixed quantum states. Kramer and Savari [10] look at communication for the purpose of "communicating probability distributions"

in the sense that they care about reconstructing a sequence with the proper empirical distribution of the sources rather than the sources themselves. Weissman and Ordentlich [11] make statements about the empirical distributions of sub-blocks of source and reconstruction symbols in a rate-constrained setting. And Han and Verdú [12] consider generating a random process via use of a memoryless channel, while Bennett et. al. [13] propose a "reverse Shannon theorem" stating the amount of noise-free communication necessary to synthesize a memoryless channel.

In this work, we consider coordination of actions in two and three node networks. These serve as building blocks for understanding larger networks. Some of the actions at the nodes are given by nature, and some are constructed by the node itself.

Chapter 2 deals with empirical coordination. For some network settings we characterize the entire solution, but for others we give partial results including bounds and solutions to special cases. The complete results include a variant of the multiterminal source coding problem (Section 2.2.4). Among the partial results, a consistent trend in coordination strategies becomes apparent, and the golden ratio makes a surprise appearance. Also, rate-distortion regions are shown to be projections of the coordination capacity region.

In Chapter 3 we consider strong coordination. We characterize the communication requirements in two fundamental settings and discuss the role of common randomness. If common randomness is available to all nodes in the network, then empirical coordination and strong coordination seem to require equivalent communication resources, consistent with the implications of the "reverse Shannon theorem" [13]. Furthermore, we can quantify the amount of common randomness needed, treating common randomness itself as a scarce resource.

We can use these building blocks to compare the efficiency of network topologies in larger networks. An example of this idea can be found in Chapter 4 along with some natural inquiries about coordination across noisy channels.

# Chapter 2

# Empirical Coordination

## 2.1 Introduction

In this chapter we address questions of this nature: If three different tasks are to be performed in a shared effort between three people, but one of them is randomly assigned his responsibility, how much must he tell the others about his assignment?

We consider coordination in a variety of two and three node networks. The basic meaning of empirical coordination according to a desired distribution is the same for each network—we use the network communication to construct a sequence of actions that have an empirical joint distribution closely matching the desired distribution. What's different from one problem to the next is the set of nodes whose actions are selected randomly by nature and the communication limitations imposed by the network topology.

Here we define the problem in the context of the cascade network of Section 2.2.3 shown in Figure 2.1. These definitions have obvious generalizations to other networks.

In the cascade network of Figure 2.1, node $X$ has a sequence of actions $X_1, X_2, ...$ specified randomly by nature. Note that a node is allowed to see all of its actions before it summarizes them for the next node. Communication is used to give Node $Y$ and Node $Z$ enough information to choose sequences of actions that are empirically correlated with $X_1, X_2, ...$ according to a desired joint distribution $p_0(x)p(y, z|x)$. The communication travels in a cascade, first from Node $X$ to Node $Y$ at rate $R_1$ bits per

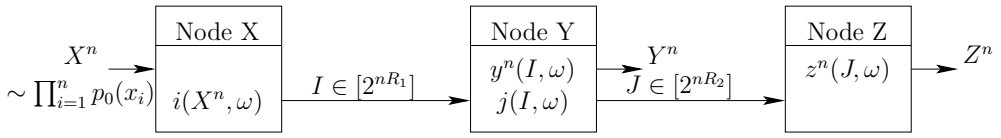


Figure 2.1: *Cascade network.* Node X is assigned actions $X^n$ chosen by nature according to $p(x^n) = \prod_{i=1}^n p_0(x_i)$. A message $I$ in the set $\{1, ..., 2^{nR_1}\}$ is constructed based on $X^n$ and the common randomness $\omega$ and sent to Node Y, which constructs both an action sequence $Y^n$ and a message $J$ in the set $\{1, ..., 2^{nR_2}\}$. Finally, Node Z produces actions $Z^n$ based on the message $J$ and the common randomness $\omega$. This is summarized in Figure 2.2.

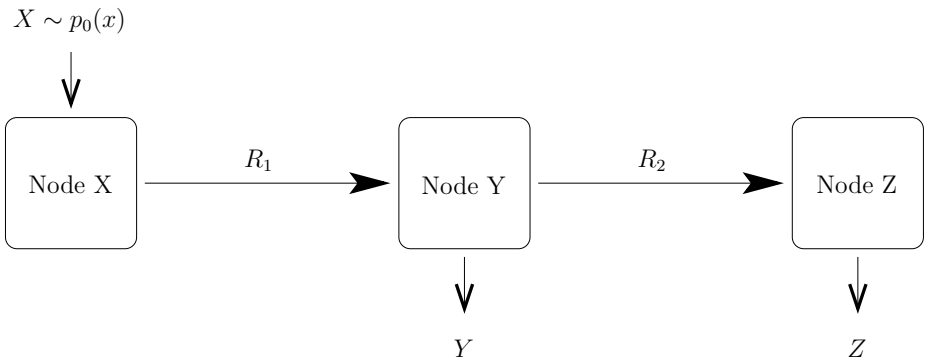


Figure 2.2: Shorthand notation for the cascade network of Figure 2.1.

action, and then from Node $Y$ to Node $Z$ at rate $R_2$ bits per action.

Specifically, a $(2^{nR_1}, 2^{nR_2}, n)$ *coordination code* is used as a protocol to coordinate the actions in the network for a block of $n$ time periods. The coordination code and the distribution of the random actions $X^n$ induce a joint distribution on the actions in the network. If the joint type of the actions in the network can be made arbitrarily close to a desired distribution $p_0(x)p(y, z|x)$ with high probability, according to the distribution induced by a $(2^{nR_1}, 2^{nR_2}, n))$ coordination code, then $p_0(x)p(y, z|x)$ is achievable with the rate pair $(R_1, R_2)$.

**Definition 1** (Coordination code)**.** *A $(2^{nR_1}, 2^{nR_2}, n)$ coordination code for the cascade network of Figure 2.1 consists of four functions—an encoding function*

$$i \;:\; \mathcal{X}^n \times \Omega \longrightarrow \{1, ..., 2^{nR_1}\},$$

*a recoding function*

$$j \;:\; \{1,...,2^{nR_1}\} \times \Omega \longrightarrow \{1,...,2^{nR_2}\},$$

*and two decoding functions*

$$\begin{aligned} y^n &: \{1,...,2^{nR_1}\} \times \Omega \longrightarrow \mathcal{Y}^n, \\ z^n &: \{1,...,2^{nR_2}\} \times \Omega \longrightarrow \mathcal{Z}^n. \end{aligned}$$

**Definition 2** (Induced distribution)**.** *The induced distribution $p(x^n, y^n, z^n)$ is the resulting joint distribution of the actions in the network when a $(2^{nR_1}, 2^{R_2}, n)$ coordination code is used. The actions $X^n$ are chosen by nature i.i.d. according to $p_0(x)$ and independent of the common randomness $\omega$. Thus, the joint distribution of $X^n$ and $\omega$ is*

$$p(x^n, \omega) \;=\; p(\omega) \prod_{i=1}^{n} p_0(x_i).$$

*The actions $Y^n$ and $Z^n$ are functions of $X^n$ and $\omega$ given by implementing the coordination code as*

$$\begin{aligned} Y^n &= y^n(i(X^n, \omega), \omega), \\ Z^n &= z^n(j(i(X^n, \omega), \omega), \omega). \end{aligned}$$

**Definition 3** (Joint type)**.** *The joint type $P_{x^n,y^n,z^n}$ of a tuple of sequences $(x^n, y^n, z^n)$ is the empirical probability mass function, given by*

$$P_{x^n,y^n,z^n}(x, y, z) \;\triangleq\; \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}((x_i, y_i, z_i) = (x, y, z)),$$

*for all $(x, y, z) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$, where $\mathbf{1}$ is the indicator function.*

**Definition 4** (Total variation)**.** *The total variation between two probability mass functions is half the $L_1$ distance between them, given by*

$$\|p(x,y,z) - q(x,y,z)\|_{TV} \triangleq \frac{1}{2}\sum_{x,y,z} |p(x,y,z) - q(x,y,z)|.$$

**Definition 5** (Achievability)**.** *A desired distribution $p_0(x)p(y,z|x)$ is achievable with the rate pair $(R_1, R_2)$ if there exists a sequence of $(2^{nR_1}, 2^{nR_2}, n)$ coordination codes and a choice of $p(\omega)$ such that under the induced distribution the total variation between the joint type of the actions in the network and the desired distribution goes to zero in probability. That is,*

$$\|P_{X^n,Y^n,Z^n}(x,y,z) - p_0(x)p(y,z|x)\|_{TV} \longrightarrow 0 \text{ in probability.}$$

**Definition 6** (Coordination capacity region)**.** *The coordination capacity region $\mathcal{C}_{p_0}$ for the source distribution $p_0(x)$ is the closure of the set of rate-coordination tuples $(R_1, R_2, p(y,z|x))$ that are achievable:*

$$\mathcal{C}_{p_0} \triangleq \mathbf{Cl}\left\{ \begin{array}{l} (R_1, R_2, p(y,z|x)) : \\ p_0(x)p(y,z|x) \text{ is achievable at rates } (R_1, R_2) \end{array} \right\}.$$

**Definition 7** (Rate-coordination region)**.** *The rate-coordination region $\mathcal{R}_{p_0}$ is a slice of the coordination capacity region corresponding to a fixed distribution $p(y,z|x)$:*

$$\mathcal{R}_{p_0}(p(y,z|x)) \triangleq \{(R_1, R_2) : (R_1, R_2, p(y,z|x)) \in \mathcal{C}_{p_0}\}.$$

**Definition 8** (Coordination-rate region)**.** *The coordination-rate region $\mathcal{P}_{p_0}$ is a slice of the coordination capacity region corresponding to a tuple of rates $(R_1, R_2)$:*

$$\mathcal{P}_{p_0}(R_1, R_2) \triangleq \{p(y,z|x) : (R_1, R_2, p(y,z|x)) \in \mathcal{C}_{p_0}\}.$$

### 2.1.1 Preliminary observations

**Lemma 1** (Convexity of coordination)**.** *$\mathcal{C}_{p_0}$, $\mathcal{R}_{p_0}$, and $\mathcal{P}_{p_0}$ are all convex sets.*

*Proof.* The coordination capacity region $\mathcal{C}_{p_0}$ is convex because time-sharing can be used to achieve any point on the chord between two achievable rate-coordination pairs. Simply combine two sequences of coordination codes that achieve the two points in the coordination capacity region by using one code and then the other in a proportionate manner to achieve any point on the chord. The definition of joint type in Definition 3 involves an average over time. Thus if one sequence is concatenated with another sequence, the resulting joint type is a weighted average of the joint types of the two composing sequences. Rates of communication also combine according to the same weighted average. The rate of the resulting concatenated code is the weighted average of the two rates.

The rate-coordination region $\mathcal{R}_{p_0}$ is the intersection of the coordination capacity region $\mathcal{C}_{p_0}$ with a hyperplane, which are both convex sets. Likewise for the coordination-rate region $\mathcal{P}_{p_0}$. Therefore, $\mathcal{R}_{p_0}$ and $\mathcal{P}_{p_0}$ are both convex. □

Although common randomness is available as a resource, the following theorem shows that it doesn't play a necessary role in achieving empirical coordination. Therefore, we will not bother to include common randomness in the construction of coordination codes. However, in Chapter 3 we show that common randomness is a valuable resource for achieving strong coordination, yet to be precisely defined.

**Theorem 2** (Common randomness doesn't help)**.** *Any desired distribution $p_0(x)p(y,z|x)$ that is achievable for empirical coordination with the rate pair $(R_1, R_2)$ can be achieved with $\Omega = \emptyset$.*

*Proof.* Suppose that $p_0(x)p(y,z|x)$ is achievable for empirical coordination with the rate pair $(R_1, R_2)$. Then there exists a sequence of $(2^{nR_1}, 2^{nR_2}, n)$ coordination codes for which the expected total variation between the joint type and $p(x,y,z)$ goes to zero with respect to the induced distribution. This follows from the bounded convergence theorem since total variation is bounded by one. By iterated expectation,

$$\mathbf{E}\left[\mathbf{E}\left[\left\|P_{X^n,Y^n,Z^n} - p_0(x)(y,z|x)\right\|_{TV}|\omega\right]\right] \;=\; \mathbf{E}\left\|P_{X^n,Y^n,Z^n} - p_0(x)(y,z|x)\right\|_{TV}.$$

Therefore, there exists a value $\omega^*$ such that

$$\mathbf{E}\left[\left\|P_{X^n,Y^n,Z^n} - p_0(x)(y,z|x)\right\|_{TV} |\omega^*\right] \leq \mathbf{E}\left\|P_{X^n,Y^n,Z^n} - p_0(x)(y,z|x)\right\|_{TV}.$$

Define a new coordination code that doesn't depend on $\omega$ and at the same time doesn't increase the expected total variation:

$$\begin{aligned} i^*(x^n) &= i(x^n,\omega^*), \\ j^*(i) &= j(i,\omega^*), \\ y^{n*}(i) &= Y^n(i,\omega^*), \\ z^{n*}(j) &= Z^n(j,\omega^*). \end{aligned}$$

This can be done for each $(2^{nR_1}, 2^{nR_2}, n)$ coordination code for $n = 1, 2, ....$ □

### 2.1.2 Generalize

We will investigate empirical coordination in a variety of networks. In each case, we explicitly specify the structure and implementation of the coordination codes, similar to Definitions 1 and 2, while all other definitions carry over in a straightforward manner.

We use a *shorthand* notation in order to illustrate each network setting with a simple and consistent figure. Figure 2.2 shows the shorthand notation for the cascade network of Figure 2.1. The random actions that are specified by nature are shown with arrows pointing down toward the node (represented by a block). Actions constructed by the nodes themselves are shown coming out of the node with an arrow downward. And arrows indicating communication from one node to another are labeled with the rate limits for the communication along those links.

We fully characterize the coordination capacity regions $\mathcal{C}_{p_0}$ for empirical coordination in four network settings: a network of two nodes (Section 2.2.1); an isolated node network (Section 2.2.2); a cascade network (Section 2.2.3); and a degraded source network (Section 2.2.4). Additionally, we give bounds on the coordination capacity region in two more network settings: a broadcast network (Section 2.2.5);

and a cascade multiterminal network (Section 2.2.5). Proofs are left to Section 2.4.

A communication technique that we find useful in several settings is to use a portion of the communication to send identical messages to all nodes in the network. The common message serves to correlate the codebooks used on different communication links and can result in reduced rates in the network.

## 2.2 Network results

### 2.2.1 Two nodes

In the simplest network setting shown in Figure 2.3, we consider two nodes, X and Y. The action $X$ is specified by nature according to $p_0(x)$, and a message is sent at rate $R$ to node Y.

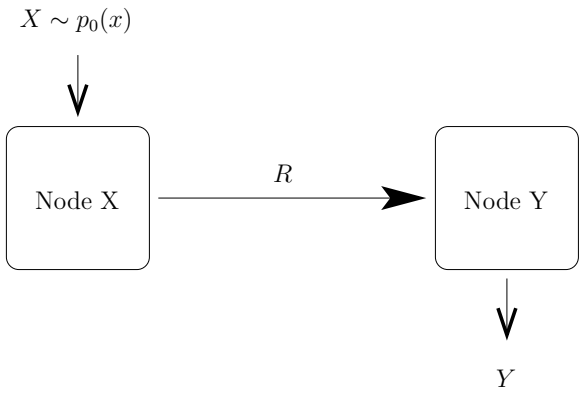


Figure 2.3: *Two nodes.* The action $X$ is chosen by nature according to $p_0(x)$. A message is sent to node Y at rate $R$. The coordination capacity region $\mathcal{C}_{p_0}$ is the set of rate-coordination pairs where the rate is greater than the mutual information between $X$ and $Y$.

The $(2^{nR}, n)$ coordination codes consist of an encoding function

$$i \;:\; \mathcal{X}^n \longrightarrow \{1, ..., 2^{nR}\},$$

and a decoding function

$$y^n \;:\; \{1, ..., 2^{nR}\} \longrightarrow \mathcal{Y}^n.$$

The actions $X^n$ are chosen by nature i.i.d. according to $p_0(x)$, and the actions $Y^n$ are functions of $X^n$ given by implementing the coordination code as

$$Y^n \;=\; y^n(i(X^n)).$$

**Theorem 3** (Coordination capacity region)**.** *The coordination capacity region $\mathcal{C}_{p_0}$ for empirical coordination in the two-node network of Figure 2.3 is the set of rate-coordination pairs where the rate is greater than the mutual information between $X$ and $Y$. Thus,*

$$\mathcal{C}_{p_0} \;=\; \Big\{(R, p(y|x)) \;:\; R \geq I(X;Y)\Big\}.$$

*Discussion:* The coordination capacity region in this setting yields the rate-distortion result of Shannon [14].

**Example 1** (Task assignment)**.** *Suppose there are $k$ tasks numbered 1 through $k$. One task is dealt randomly to node X, and node Y needs to choose one of the remaining tasks. This coordinated behavior can be summarized by a distribution $\hat{p}$. The action $X$ is given by nature according to $\hat{p}_0(x)$, the uniform distribution on the set $\{1, ..., k\}$. The desired conditional distribution of the action $Y$ is $\hat{p}(y|x)$, the uniform distribution on the set of tasks different from $x$. Therefore, the joint distribution $\hat{p}_0(x)\hat{p}(y|x)$ is the uniform distribution on pairs of differing tasks from the set $\{1, ..., k\}$. Figure 2.4 illustrates a valid outcome for $k$ larger than 5.*

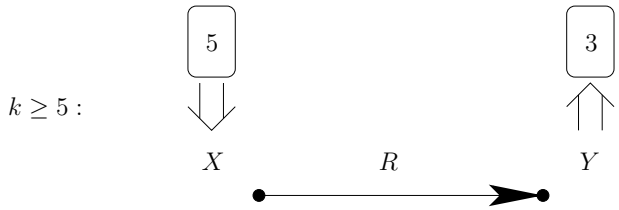


Figure 2.4: *Task assignment in the two-node network.* A task from a set of tasks numbered $1, ..., k$ is to be assigned uniquely to each of the nodes X and Y in the two-node network setting. The task assignment for X is given randomly by nature. The communication rate $R \geq \log(k/k-1)$ is necessary and sufficient to allow Y to select a different task from X.

*By applying Theorem 3, we find that the rate-coordination region* $\mathcal{R}_{\hat{p}_0}(\hat{p}(y|x))$ *is given by*

$$\mathcal{R}_{\hat{p}_0}(\hat{p}(y|x)) \;=\; \left\{R \;:\; R \geq \log\left(\frac{k}{k-1}\right)\right\}.$$

### 2.2.2 Isolated node

Now we derive the coordination capacity region for the isolated-node network of Figure 2.5. Node X has an action chosen by nature according to $p_0(x)$, and a message is sent at rate $R$ from node X to node Y from which node Y produces its action. Node Z also produces an action but receives no communication. What is the set of all achievable coordination distributions $p(y, z|x)$?

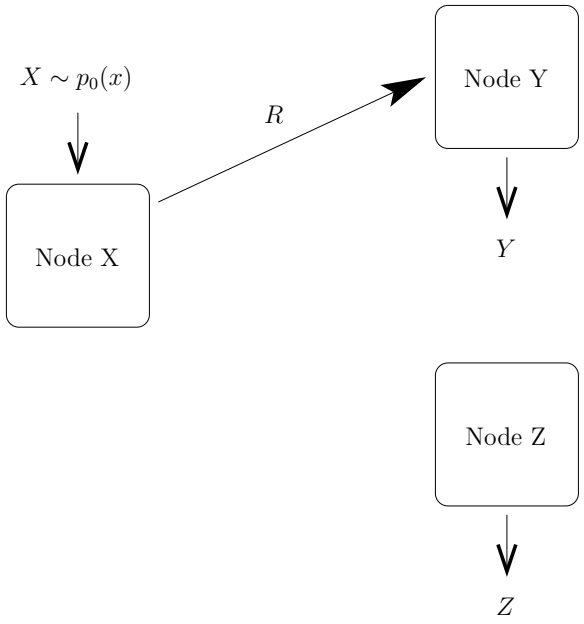


Figure 2.5: *Isolated node.* The action $X$ is chosen by nature according to $p_0(x)$, and a message is sent at rate $R$ from node X to node Y. Node Z receives no communication. The coordination capacity region $\mathcal{C}_{p_0}$ is the set of rate-coordination pairs where $p(x, y, z) = p_0(x)p(z)p(y|x, z)$ and the rate $R$ is greater than the conditional mutual information between $X$ and $Y$ given $Z$.

We formalize this problem as follows. The $(2^{nR}, n)$ coordination codes consist of

an encoding function

$$i \;:\; \mathcal{X}^n \longrightarrow \{1, ..., 2^{nR}\},$$

a decoding function

$$y^n \;:\; \{1, ..., 2^{nR}\} \longrightarrow \mathcal{Y}^n,$$

and a deterministic sequence

$$z^n \;\in\; \mathcal{Z}^n.$$

The actions $X^n$ are chosen by nature i.i.d. according to $p_0(x)$, and the actions $Y^n$ are functions of $X^n$ given by implementing the coordination code as

$$\begin{aligned} Y^n &= y^n(i(X^n)), \\ Z^n &= z^n. \end{aligned}$$

It is tempting to believe that the coordination-rate region $\mathcal{P}_{p_0}(R)$ is the set of all distributions $p_0(x)p(y|x)p(z)$ such that $R \geq I(X;Y)$. However, $Z$ need not be independent of $Y$, even though there is no communication to node Z. The entire region is given in the following theorem.

**Theorem 4** (Coordination capacity region)**.** *The coordination capacity region $\mathcal{C}_{p_0}$ for empirical coordination in the isolated-node network of Figure 2.5 is the set of rate-coordination pairs where $Z$ is independent of $X$ and the rate $R$ is greater than the conditional mutual information between $X$ and $Y$ given $Z$. Thus,*

$$\mathcal{C}_{p_0} \;=\; \Big\{ (R, p(z)p(y|x,z)) \;:\; R \geq I(X;Y|Z) \Big\}.$$

*Discussion:* How can $Y$ and $Z$ have a dependence when there is no communication between them? This dependence is possible because neither $Y$ nor $Z$ is chosen randomly by nature. In an extreme case, we could let node Y ignore the incoming

message from node X and let the actions at node Y and node Z be equal, $Y = Z$. Thus we can immediately see that with no communication the coordination region consists of all distributions of the form $p_0(x)p(y,z)$.

It is interesting to note that there is a tension between the correlation of $X$ and $Y$ and the correlation of $Y$ and $Z$. For instance, if the communication is used to make perfect correlation between $X$ and $Y$ then any potential correlation between $Y$ and $Z$ is forfeited.

Within the results for the more general cascade network in the sequel (Section 2.2.3) we will find that Theorem 4 is an immediate consequence of Theorem 5 by letting $R_2 = 0$.

**Example 2** (Jointly Gaussian)**.** *Jointly Gaussian distributions illustrate the tradeoff between the correlation of $X$ and $Y$ and the correlation of $Y$ and $Z$ in the isolated-node network. Consider the portion of the coordination-rate region $\mathcal{P}_{p_0}(R)$ that consists of jointly Gaussian distributions. If $X$ is distributed according to $N(0,\sigma_X^2)$, what set of covariance matrices can be achieved at rate $R$?*

*So far we have discussed coordination for distribution functions with finite alphabets. Extending to infinite alphabet distributions, achievability means that any finite quantization of the joint distribution is achievable.*

*Using Theorem 4, we bound the correlations as follows:*

$$\begin{aligned}
R &\geq I(X;Y|Z) \\
&= I(X;Y,Z) \\
&= \frac{1}{2}\log\frac{|K_x||K_{yz}|}{|K_{XYZ}|} \\
&\overset{(a)}{=} \frac{1}{2}\log\frac{\sigma_x^2(\sigma_y^2\sigma_z^2-\sigma_{yz}^2)}{\sigma_x^2\sigma_y^2\sigma_z^2-\sigma_x^2\sigma_{yz}^2-\sigma_z^2\sigma_{xy}^2} \\
&\overset{(b)}{=} \frac{1}{2}\log\frac{1-\left(\frac{\sigma_{yz}}{\sigma_y\sigma_z}\right)^2}{1-\left(\frac{\sigma_{yz}}{\sigma_y\sigma_z}\right)^2-\left(\frac{\sigma_{xy}}{\sigma_x\sigma_y}\right)^2} \\
&= \frac{1}{2}\log\frac{1-\rho_{yz}^2}{1-\rho_{yz}^2-\rho_{xy}^2},
\end{aligned} \tag{2.1}$$

*where $\rho_{xy}$ and $\rho_{yz}$ are correlation coefficients. Equality (a) holds because $\sigma_{xz} = 0$ due to the independence between $X$ and $Z$. Obtain equality (b) by dividing the numerator and denominator of the argument of the* log *by $\sigma_x^2\sigma_y^2\sigma_z^2$.*

*Unfolding (2.1) yields a linear tradeoff between the $\rho_{xy}^2$ and $\rho_{yz}^2$, given by*

$$(1 - 2^{-2R})^{-1}\rho_{xy}^2 + \rho_{yz}^2 \quad \leq \quad 1.$$

*Thus any correlation coefficients $\rho_{xy}$ and $\rho_{yz}$ are achievable at rate $R$ if they satisfy the above constraint.*

### 2.2.3 Cascade

We now give the coordination capacity region for the cascade of communication in Figure 2.6. In this setting, the action at node X is chosen by nature. A message at rate $R_1$ is sent from node X to node Y, and subsequently a message at rate $R_2$ is sent from node Y to node Z based on the message received from node X. Nodes Y and Z produce actions based on the messages they receive.

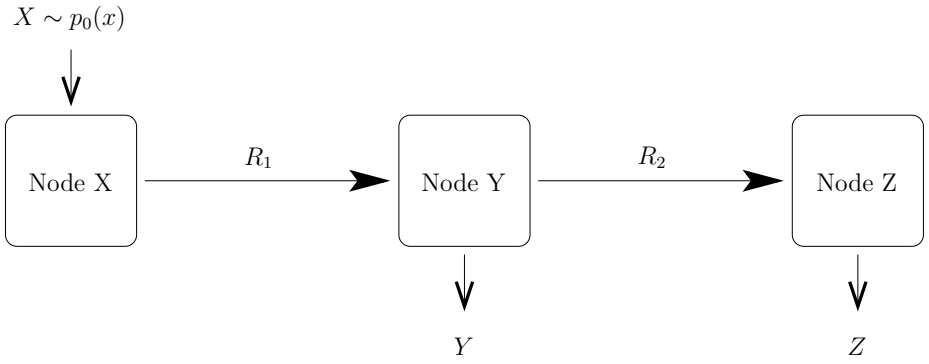


Figure 2.6: *Cascade.* The action $X$ is chosen by nature according to $p_0(x)$. A message is sent from node X to node Y at rate $R_1$. Node Y produces an action $Y$ and a message to send to node Z based on the message received from node X. Node Z then produces an action $Z$ based on the message received from node Y. The coordination capacity region $\mathcal{C}_{p_0}$ is the set of rate-coordination triples where the rate $R_1$ is greater than the mutual information between $X$ and $(Y, Z)$, and the rate $R_2$ is greater than the mutual information between $X$ and $Z$.

The formal statement is as follows. The $(2^{nR_1}, 2^{nR_2}, n)$ coordination codes consist

of four functions—an encoding function

$$i \;:\; \mathcal{X}^n \longrightarrow \{1, ..., 2^{nR_1}\},$$

a recoding function

$$j \;:\; \{1, ..., 2^{nR_1}\} \longrightarrow \{1, ..., 2^{nR_2}\},$$

and two decoding functions

$$\begin{aligned} y^n &: \{1, ..., 2^{nR_1}\} \longrightarrow \mathcal{Y}^n, \\ z^n &: \{1, ..., 2^{nR_2}\} \longrightarrow \mathcal{Z}^n. \end{aligned}$$

The actions $X^n$ are chosen by nature i.i.d. according to $p_0(x)$, and the actions $Y^n$ and $Z^n$ are functions of $X^n$ given by implementing the coordination code as

$$\begin{aligned} Y^n &= y^n(i(X^n)), \\ Z^n &= z^n(j(i(X^n))). \end{aligned}$$

This network was considered by Yamamoto [15] in the context of rate-distortion theory. The same optimal encoding scheme from his work achieves the coordination capacity region as well.

**Theorem 5** (Coordination capacity region)**.** *The coordination capacity region $\mathcal{C}_{p_0}$ for empirical coordination in the cascade network of Figure 2.6 is the set of rate-coordination triples where the rate $R_1$ is greater than the mutual information between $X$ and $(Y, Z)$, and the rate $R_2$ is greater than the mutual information between $X$ and $Z$. Thus,*

$$\mathcal{C}_{p_0} = \left\{ (R_1, R_2, p(y, z|x)) \;:\; \begin{array}{l} R_1 \geq I(X; Y, Z), \\ R_2 \geq I(X; Z). \end{array} \right\}.$$

*Discussion:* The coordination capacity region $\mathcal{C}_{p_0}$ meets the cut-set bound. The trick to achieving this bound is to first specify $Z$ and then specify $Y$ conditioned on $Z$.

**Example 3** (Task assignment)**.** *Consider a task assignment setting where three tasks are to be assigned without duplication to the three nodes X, Y, and Z, and the assignment for node X is chosen uniformly at random by nature. A distribution capturing this coordination behavior is the uniform distribution over the six permutations of task assignments. Let $\hat{p}_0(x)$ be the uniform distribution on the set $\{1, 2, 3\}$, and let $\hat{p}(y, z|x)$ give equal probability to both of the assignments to Y and Z that produce different tasks at the three nodes. Figure 2.7 illustrates a valid outcome of the task assignments.*

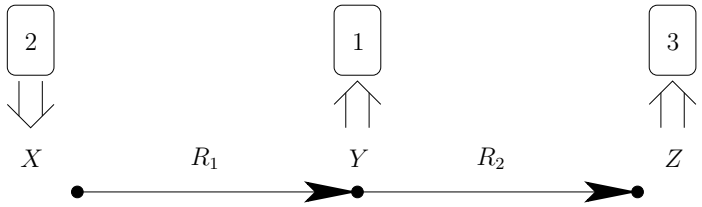


Figure 2.7: *Task assignment in the cascade network.* Three tasks, numbered 1, 2, and 3, are distributed among three nodes X, Y, and Z in the cascade network setting. The task assignment for X is given randomly by nature. The rates $R_1 \geq \log 3$ and $R_2 \geq \log 3 - \log 2$ are required to allow Y and Z to choose different tasks from X and from each other.

*According to Theorem 5, the rate-coordination region $\mathcal{R}_{\hat{p}_0}(\hat{p}(y, z|x))$ is given by*

$$\mathcal{R}_{\hat{p}_0}(\hat{p}(y,z|x)) \;=\; \left\{ (R_1, R_2) \;:\; \begin{array}{l} R_1 \geq \log 3, \\ R_2 \geq \log 3 - \log 2. \end{array} \right\}.$$

### 2.2.4 Degraded source

Here we present the coordination capacity region for the degraded-source network shown in Figure 2.8. Nodes X and Y each have an action specified by nature, and $Y$ is a function of $X$. That is, $p_0(x, y) = p_0(x)\mathbf{1}(y = f_0(x))$, where $\mathbf{1}(\cdot)$ is the indicator function. Node X sends a message to node Y at rate $R_1$ and a message to node Z at

rate $R_2$. Node Y, upon receiving the message from node X, sends a message at rate $R_3$ to node Z. Node Z produces an action based on the two messages it receives.

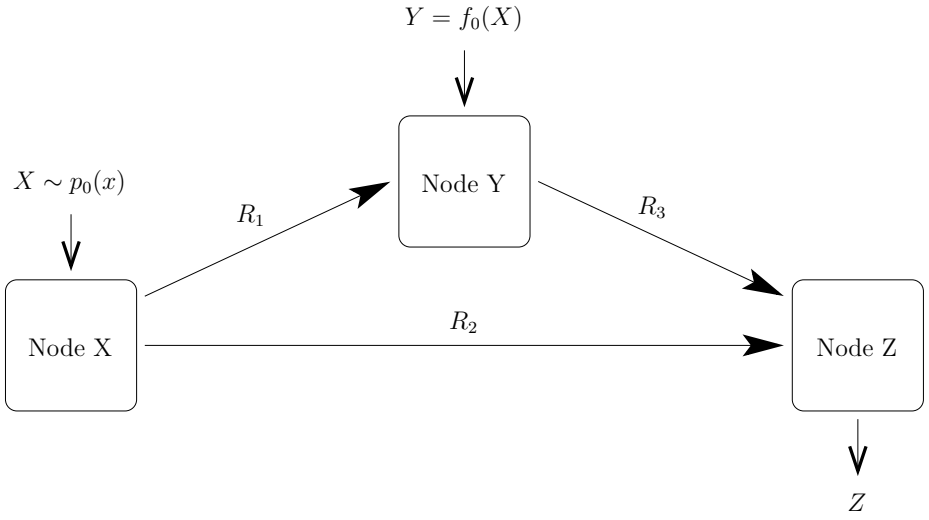


Figure 2.8: *Degraded source:* The action $X$ is specified by nature according to $p_0(x)$, and the action $Y$ is a function $f_0$ of $X$. A message is sent from node X to node Y at rate $R_1$, after which node Y constructs a message for node Z at rate $R_3$ based on the incoming message from node X and the action $Y$. Node X also sends a message directly to node Z at rate $R_2$. The coordination capacity region $\mathcal{C}_{p_0}$ is given in Theorem 6.

The $(2^{nR_1}, 2^{nR_2}, 2^{nR_3}, n)$ coordination codes for Figure 2.8 consist of four functions—two encoding functions

$$\begin{aligned} i &: \mathcal{X}^n \longrightarrow \{1, ..., 2^{nR_1}\}, \\ j &: \mathcal{X}^n \longrightarrow \{1, ..., 2^{nR_2}\}, \end{aligned}$$

a recoding function

$$k : \{1, ..., 2^{nR_1}\} \times \mathcal{Y}^n \longrightarrow \{1, ..., 2^{nR_3}\},$$

and a decoding function

$$z^n : \{1, ..., 2^{nR_2}\} \times \{1, ..., 2^{nR_3}\} \longrightarrow \mathcal{Y}^n.$$

The actions $X^n$ and $Y^n$ are chosen by nature i.i.d. according to $p_0(x, y)$, having the property that $Y_i = f_0(X_i)$ for all $i$, and the actions $Z^n$ are a function of $X^n$ and $Y^n$ given by implementing the coordination code as

$$Y^n = y^n(j(X^n), k(i(X^n), Y^n)).$$

Others have investigated source coding networks in the rate-distortion context where two sources are encoded at separate nodes to be reconstructed at a third node. Kaspi and Berger [16] consider a variety of cases where the encoders share some information. Also, Barros and Servetto [17] articulate the compress and bin strategy for more general bi-directional exchanges of information among the encoders. While falling under the same general compression strategy, the degraded source network is a special case where optimality can be established, yielding a characterization of the coordination capacity region.

**Theorem 6** (Coordination capacity region)**.** *The coordination capacity region $\mathcal{C}_{p_0}$ for empirical coordination in the degraded-source network of Figure 2.8 is given by*

$$\mathcal{C}_{p_0} = \left\{ (R_1, R_2, R_3, p(z|x, y)) \; : \; \begin{array}{l} \exists p(u|x, y, z) \text{ such that} \\ |\mathcal{U}| \leq |\mathcal{X}||\mathcal{Z}| + 2, \\ R_1 \geq I(X; U|Y), \\ R_2 \geq I(X; Z|U), \\ R_3 \geq I(X; U). \end{array} \right\}.$$

### 2.2.5 Broadcast

We now give bounds on the coordination capacity region for the broadcast network of Figure 2.9. In this setting, node X has an action specified by nature according to $p_0(x)$ and sends one message to node Y at rate $R_1$ and a separate message to node Z at rate $R_2$. Nodes Y and Z each produce an action based on the message they receive.

Node X serves as the controller for the network. Nature assigns an action to node X, which then tells node Y and node Z which actions to take.

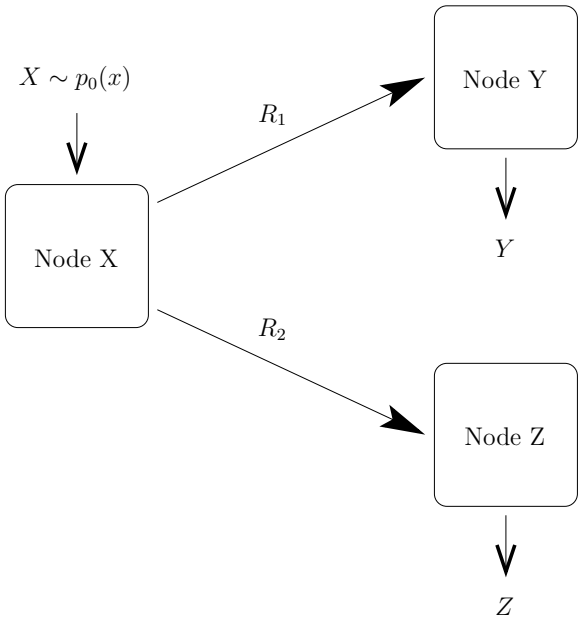


Figure 2.9: *Broadcast.* The action $X$ is chosen by nature according to $p_0(x)$. A message is sent from node X to node Y at rate $R_1$, and a separate message is sent from node X to node Z at rate $R_2$. Nodes Y and Z produce actions based on the messages they receive. Bounds on the coordination capacity region $\mathcal{C}_{p_0}$ are given in Theorem 7.

The $(2^{nR_1}, 2^{nR_2}, n)$ coordination codes consist of two encoding functions

$$\begin{aligned} i &: \mathcal{X}^n \longrightarrow \{1, ..., 2^{nR_1}\}, \\ j &: \mathcal{X}^n \longrightarrow \{1, ..., 2^{nR_2}\}, \end{aligned}$$

and two decoding functions

$$\begin{aligned} y^n &: \{1, ..., 2^{nR_1}\} \longrightarrow \mathcal{Y}^n. \\ z^n &: \{1, ..., 2^{nR_2}\} \longrightarrow \mathcal{Z}^n. \end{aligned}$$

The actions $X^n$ are chosen by nature i.i.d. according to $p_0(x)$, and the actions $Y^n$

and $Z^n$ are functions of $X^n$ given by implementing the coordination code as

$$\begin{aligned} Y^n &= y^n(i(X^n)). \\ Z^n &= z^n(j(X^n)). \end{aligned}$$

From a rate-distortion point of view, the broadcast network is not a likely candidate for consideration. The problem separates into two non-interfering rate-distortion problems, and the relationship between the sequences $Y^n$ and $Z^n$ is ignored. However, the problem of multiple descriptions, where the combination of two messages $I$ and $J$ are used to make a third estimate of the source $X$, demands consideration of the relationship between the two messages. In fact, the communication scheme for the multiple descriptions problem presented by Zhang and Berger [18] coincides with our inner bound for the coordination capacity region in the broadcast network.

The set of rate-coordination tuples $\mathcal{C}_{p_0,in}$ is an inner bound on the coordination capacity region, given by

$$\mathcal{C}_{p_0,in} \triangleq \left\{ (R_1, R_2, p(y,z|x)) : \begin{array}{l} \exists p(u|x,y,z) \text{ such that} \\ R_1 \geq I(X;U,Y), \\ R_2 \geq I(X;U,Z), \\ R_1 + R_2 \geq I(X;U,Y) + I(X;U,Z) + I(Y;Z|X,U). \end{array} \right\}.$$

The set of rate-coordination tuples $\mathcal{C}_{p_0,out}$ is an outer bound on the coordination capacity region, given by

$$\mathcal{C}_{p_0,out} \triangleq \left\{ (R_1, R_2, p(y,z|x)) : \begin{array}{l} R_1 \geq I(X;Y), \\ R_2 \geq I(X;Z), \\ R_1 + R_2 \geq I(X;Y,Z). \end{array} \right\}.$$

Also, define $\mathcal{R}_{p_0,in}(p(y,z|x))$ and $\mathcal{R}_{p_0,out}(p(y,z|x))$ to be the sets of rate pairs in $\mathcal{C}_{p_0,in}$ and $\mathcal{C}_{p_0,out}$ corresponding to the desired distribution $p(y,z|x)$.

**Theorem 7** (Coordination capacity region bounds)**.** *The coordination capacity region $\mathcal{C}_{p_0}$ for empirical coordination in the broadcast network of Figure 2.9 is bounded by*

$$\mathcal{C}_{p_0,in} \subset \mathcal{C}_{p_0} \subset \mathcal{C}_{p_0,out}.$$

*Discussion:* The regions $\mathcal{C}_{p_0,in}$ and $\mathcal{C}_{p_0,out}$ are convex. A time-sharing random variable can be lumped into the auxiliary random variable $U$ in the definition of $\mathcal{C}_{p_0,in}$ to show convexity.

The inner bound $\mathcal{C}_{p_0,in}$ is achieved by first sending a common message, represented by $U$, to both receivers and then private messages to each. The common message effectively correlates the two codebooks to reduce the required rates for specifying the actions $Y^n$ and $Z^n$. The sum rate takes a penalty of $I(Y;Z|X,U)$ in order to assure that $Y$ and $Z$ are coordinated with each other as well as with $X$.

The outer bound $\mathcal{C}_{p_0,out}$ is a consequence of applying the two-node result of Theorem 3 in three different ways, once for each receiver, and once for the pair of receivers with full cooperation.

For many distributions, the bounds in Theorem 7 are tight and the rate-coordination region $\mathcal{R}_{p_0} = \mathcal{R}_{p_0,in} = \mathcal{R}_{p_0,out}$. This is true for all distributions where $X$, $Y$, and $Z$ form a Markov chain in any order. It is also true for distributions where $Y$ and $Z$ are independent or where $X$ is independent pairwise with both $Y$ and $Z$. For each of these cases, Table 2.1 shows the choice of auxiliary random variable $U$ in the definition of $\mathcal{R}_{p_0,in}$ that yields $\mathcal{R}_{p_0,in} = \mathcal{R}_{p_0,out}$. In case 5, the region $\mathcal{R}_{p_0,in}$ is optimized by time-sharing between $U = Y$ and $U = Z$.

Table 2.1: Known capacity region (cases where $\mathcal{R}_{p_0,in} = \mathcal{R}_{p_0,out}$).

| | Condition | Auxiliary |
|---|---|---|
| Case 1: | $Y - X - Z$ | $U = \emptyset$ |
| Case 2: | $X - Y - Z$ | $U = Z$ |
| Case 3: | $X - Z - Y$ | $U = Y$ |
| Case 4: | $Y \perp Z$ | $U = \emptyset$ |
| Case 5: | $X \perp Y$ and $X \perp Z$ | $U = Y, U = Z$ |

Notice that if $R_2 = 0$ in the broadcast network we find ourselves in the isolated

node setting of Section 2.2.2. Consider a particular distribution $p_0(x)p(z)p(y|x,z)$ that could be achieved in the isolated node network. In the setting of the broadcast network, it might seem that the message from node X to node Z is useless for achieving $p_0(x)p(z)p(y|x,z)$, since $X$ and $Z$ are independent. However, this is not the case. For some desired distributions $p_0(x)p(z)p(y|x,z)$, a positive rate $R_2$ in the broadcast network actually helps reduce the required rate $R_1$.

To highlight a specific case where a message to node Z is useful even though $Z$ is independent of $X$ in the desired distribution, consider the following. Let $\overline{p}_0(x)\overline{p}(z)\overline{p}(y|x,z)$ be the uniform distribution over all combinations of binary $x$, $y$, and $z$ with even parity. The variables $X$, $Y$, and $Z$ are each Bernoulli-half and pairwise independent, and $X \oplus Y \oplus Z = 0$, where $\oplus$ is addition modulo two. This distribution satisfies both case 4 and case 5 from Table 2.1, so we know that $\mathcal{R}_{\overline{p}_0} = \mathcal{R}_{\overline{p}_0,out}$. Therefore, the rate-coordination region $\mathcal{R}_{\overline{p}_0}(\overline{p}(y,z|x))$ is characterized by a single inequality,

$$\mathcal{R}_{\overline{p}_0}(\overline{p}(y,z|x)) \;\; = \;\; \{(R_1,R_2) \in \mathbb{R}^{2+} \;:\; R_1 + R_2 \geq 1 \text{ bit}\}.$$

The minimum rate $R_1$ needed when no message is sent from node X to node Z is 1 bit, while the required rate in general is $1 - R_2$ bits.

**Example 4** (Task assignment)**.** *Consider a task assignment setting similar to Example 3, where three tasks are to be assigned without duplication to the three nodes X, Y, and Z, and the assignment for node X is chosen uniformly at random by nature. A distribution capturing this coordination behavior is the uniform distribution over the six permutations of task assignments. Let $\hat{p}_0(x)$ be the uniform distribution on the set $\{0,1,2\}$, and let $\hat{p}(y,z|x)$ give equal probability to both of the assignments to Y and Z that produce different tasks at the three nodes. Figure 2.10 illustrates a valid outcome of the task assignments.*

*We can explore the achievable rate region $\mathcal{R}_{\hat{p}_0}(\hat{p}(y,z|x))$ by using the bounds in Theorem 7. In this process, we find rates as low as $\log 3 - \log \phi$ to be sufficient on each link, where $\phi = \frac{\sqrt{5}+1}{2}$ is the golden ratio.*

*First consider the points in the inner bound $\mathcal{R}_{\hat{p}_0,in}(\hat{p}(y,z|x))$ that are achieved without the use of the auxiliary variable $U$. This consists of a pentagonal region of rate*

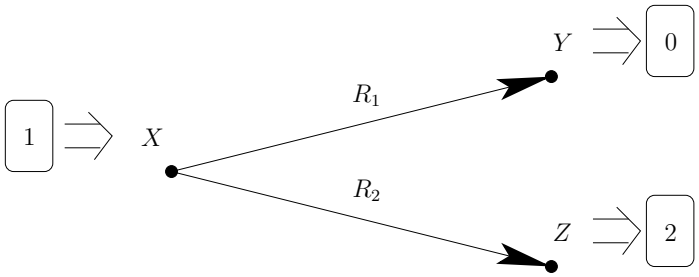


Figure 2.10: *Task assignment in the broadcast network.* Three tasks, numbered 0, 1, and 2, are distributed among three nodes X, Y, and Z in the broadcast network setting. The task assignment for X is given randomly by nature. What rates $R_1$ and $R_2$ are necessary to allow Y and Z to choose different tasks from X and each other?

*pairs. The extreme point* $A = (\log(3/2), \log 3)$*, shown in Figure 2.11, corresponds to the a simple communication approach. First node X coordinates with node Y. Theorem 3 for the two-node network declares the minimum rate needed to be* $R_1 = \log(3/2)$*. After action Y has been established, node X specifies action Z in it's entire detail using the rate* $R_2 = \log 3$*. A complementary scheme achieves the extreme point B in Figure 2.11. The sum rate achieved by these points is* $R_1 + R_2 = 2(\log_2 3 - 1/2)$ *bits.*

*We can explore more of the inner bound* $\mathcal{R}_{\hat{p}_0,in}(\hat{p}(y,z|x))$ *by adding the element of time-sharing. That is, use an auxiliary variable* $U$ *that is independent of* $X$*. As long as we can assign tasks in the network so that* $X$*,* $Y$*, and* $Z$ *are each unique, then there will be a method of using time-sharing that will achieve the desired uniform distribution over unique task assignments* $\hat{p}$*. For example, devise six task assignment schemes from the one successful scheme by mapping the tasks onto the six different permutations of* $\{0,1,2\}$*. By time-sharing equally among these six schemes, we achieve the desired distribution.*

*With the idea of time-sharing in mind, we achieve a better sum rate by restricting the domain of* $Y$ *to* $\{0,1\}$ *and* $Z$ *to* $\{0,2\}$ *and letting them be functions of* $X$ *in the*

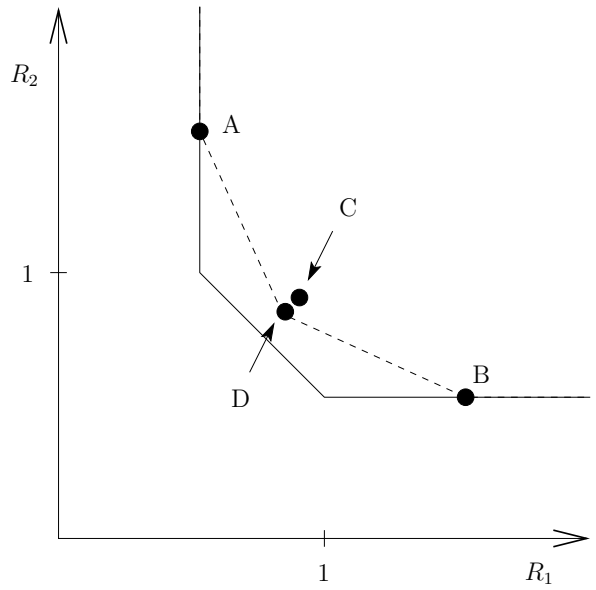


Figure 2.11: *Rate region bounds for task assignment.* Points $A$, $B$, $C$, and $D$ are achievable rates for the task assignment problem in the broadcast network. The solid line indicates the outer bound $\mathcal{R}_{\hat{p}_0,out}(\hat{p}(y,z|x))$, and the dashed line indicates a subset of the inner bound $\mathcal{R}_{\hat{p}_0,in}(\hat{p}(y,z|x))$. Points $A$ and $B$ are achieved by letting $U = \emptyset$. Point $C$ uses $U$ as time-sharing, independent of $X$. Point $D$ uses $U$ to describe $X$ partially to each of the nodes Y and Z.

*following way:*

$$Y = \begin{cases} 1, & X \neq 1, \\ 0, & X = 1, \end{cases} \tag{2.2}$$

$$Z = \begin{cases} 2, & X \neq 2, \\ 0, & X = 2. \end{cases} \tag{2.3}$$

*We can say that $Y$ takes on a default value of 1, and $Z$ takes on a default value of 2. Node $X$ just tells nodes $Y$ and $Z$ when they need to get out of the way, in which case they switch to task 0. To achieve this we only need $R_1 \geq H(Y) = \log_3 -2/3$ bits and $R_2 \geq H(Z) = \log_2 3 - 2/3$ bits, represented by point $C$ in Figure 2.11.*

*Finally, we achieve an even smaller sum rate in the inner bound $\mathcal{R}_{\hat{p}_0,in}(\hat{p}(y,z|x))$*

*by using a more interesting choice of $U$ in addition to time-sharing.*[1] *Let $U \in \{0, 1, 2\}$ be correlated with $X$ in such a way that they are equal more often than one third of the time. Now restrict the domains of $Y$ and $Z$ based on $U$. The actions $Y$ and $Z$ are functions of $X$ and $U$ defined as follows:*

$$Y = \begin{cases} U + 1 \ mod\ 3, & X \neq U + 1 \ mod\ 3, \\ U, & X = U + 1 \ mod\ 3, \end{cases} \tag{2.4}$$

$$Z = \begin{cases} U - 1 \ mod\ 3, & X \neq U - 1 \ mod\ 3, \\ U, & X = U - 1 \ mod\ 3. \end{cases} \tag{2.5}$$

*This corresponds to sending a compressed description of $X$, represented by $U$, and then assigning default values to $Y$ and $Z$ centered around $U$. The actions $Y$ and $Z$ sit on both sides of $U$ and only move when $X$ tells them to get out of the way. The description rates needed for this method are*

$$\begin{aligned} R_1 &\geq I(X;U) + I(X;Y|U) \\ &= I(X;U) + H(Y|U). \\ R_2 &\geq I(X;U) + I(X;Z|U) \\ &= I(X;U) + H(Z|U). \end{aligned} \tag{2.6}$$

*Using a symmetric conditional distribution from $X$ to $U$, calculus provides the following parameters:*

$$P(U = u|X = x) = \begin{cases} \frac{1}{\sqrt{5}}, & u = x, \\ \frac{1}{\phi\sqrt{5}}, & u \neq x, \end{cases} \tag{2.7}$$

$$\tag{2.8}$$

*where $\phi = \frac{\sqrt{5}+1}{2}$ is the golden ratio. This level of compression results in a very low rate of description, $I(X;U) \approx 0.04$ bits, for sending $U$ to each of the nodes $Y$ and $Z$.*

*The description rates needed for this method are as follows, and are represented*

[1] Time-sharing is also lumped into $U$, but we ignore that here to simplify the explanation.

*by Point D in Figure 2.11:*

$$\begin{aligned} R_1 &\geq I(X;U) + H(Y|U) \\ &= \log 3 - \frac{1}{2}\log 5 - \frac{2}{\phi\sqrt{5}}\log\phi + H(Y|U) \\ &= \log 3 - \frac{1}{2}\log 5 - \frac{2}{\phi\sqrt{5}}\log\phi + H\left(\frac{1}{\phi\sqrt{5}}\right) \\ &= \log 3 - \frac{2}{\phi\sqrt{5}}\log\phi + \frac{1}{\phi\sqrt{5}}\log\phi - \frac{\phi}{\sqrt{5}}\log\phi \\ &= \log 3 - \left(\phi + \frac{1}{\phi}\right)\frac{1}{\sqrt{5}}\log\phi \\ &= \log 3 - \log\phi, \\ R_2 &\geq \log 3 - \log\phi, \end{aligned} \tag{2.9}$$

*where $H$ is the binary entropy function. The above calculation is assisted by observing that $\phi = \frac{1}{\phi} + 1$ and $\phi + \frac{1}{\phi} = \sqrt{5}$.*

### Cascade multiterminal

We now give bounds on the coordination capacity region for the cascade-multiterminal network of Figure 2.12. In this setting, node X and node Y each have an action specified by nature according to the joint distribution $p_0(x, y)$. Node X sends a message at rate $R_1$ to node Y. Based on its own action $Y$ and the incoming message about $X$, node Y sends a message to node Z at rate $R_2$. Finally, node Z produces an action based on the message from node Y.

The $(2^{nR_1}, 2^{nR_2}, n)$ coordination codes consist of an encoding function

$$i \;:\; \mathcal{X}^n \longrightarrow \{1, ..., 2^{nR_1}\},$$

a recoding function

$$j \;:\; \{1, ..., 2^{nR_1}\} \times \mathcal{Y}^n \longrightarrow \{1, ..., 2^{nR_2}\},$$

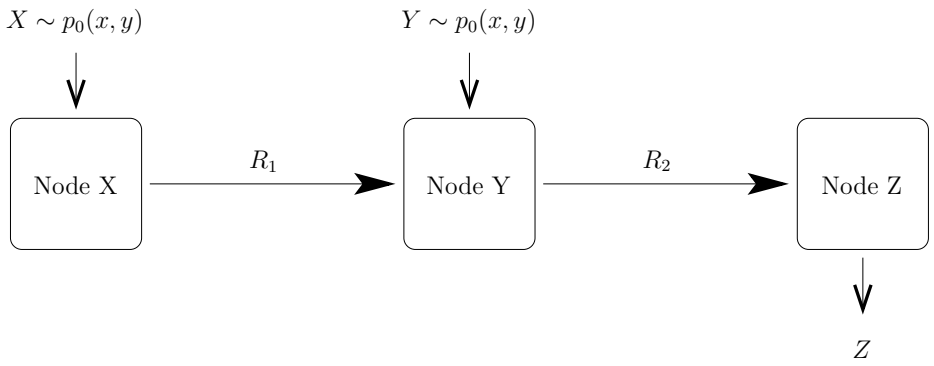


Figure 2.12: *Cascade multiterminal.* The actions $X$ and $Y$ are chosen by nature according to $p_0(x, y)$. A message is sent from node X to node Y at rate $R_1$. Node Y then constructs a message for node Z based on the received message from node X and its own action. Node Z produces an action based on the message it receives from node Y. Bounds on the coordination capacity region $\mathcal{C}_{p_0}$ are given in Theorem 8.

and a decoding function

$$z^n \quad : \quad \{1, ..., 2^{nR_2}\} \longrightarrow \mathcal{Z}^n.$$

The actions $X^n$ and $Y^n$ are chosen by nature i.i.d. according to $p_0(x, y)$, and the actions $Z^n$ are functions of $X^n$ and $Y^n$ given by implementing the coordination code as

$$Z^n \quad = \quad z^n(j(i(X^n), Y^n)).$$

Node Y is playing two roles in this network. It acts partially as a relay to send on the message from node X to node Z, while at the same time sending a message about its own actions to node Z. This situation applies to a variety of source coding scenarios. Nodes X and Y might both be sensors in a sensor network, or node Y can be thought of as a relay for connecting node X to node Z, with side information $Y$.

This network is similar to multiterminal source coding considered by Berger and Tung [19] in that two sources of information are encoded in a distributed fashion. In fact, the expansion to accommodate cooperative encoders [16] can be thought of as a generalization of our network. However, previous work along these lines is missing one

key aspect of efficiency, which is to partially relay the encoded information without changing it.

Vasudevan, Tian, and Diggavi [20] looked at a similar cascade communication system with a relay. In their setting, the relay's information $Y$ is a degraded version of the decoder's side information, and the decoder is only interested in recovering $X$. Because the relays observations contain no additional information for the decoder, the relay does not face the dilemma of mixing in some of the side information into its outgoing message. In our cascade multiterminal network, the decoder does not have side information. Thus, the relay is faced with coalescing the two pieces of information $X$ and $Y$ into a single message. Other research involving similar network settings can be found in [21], where Gu and Effros consider a more general network but with the restriction that the action $Y$ is a function of the action $X$, and [22], where Bakshi et. al. identify the optimal rate region for lossless encoding of independent sources in a longer cascade (line) network.

The set of rate-coordination tuples $\mathcal{C}_{p_0,in}$ is an inner bound on the coordination capacity region, given by

$$\mathcal{C}_{p_0,in} \triangleq \left\{ (R_1, R_2, p(z|x,y)) \; : \; \begin{array}{l} \exists p(u,v|x,y,z) \text{ such that} \\ p(x,y,z,u,v) = p_0(x,y)p(u,v|x)p(z|y,u,v) \\ R_1 \geq I(X;U,V|Y), \\ R_2 \geq I(X;U) + I(Y,V;Z|U). \end{array} \right\}.$$

The set of rate-coordination tuples $\mathcal{C}_{p_0,out}$ is an outer bound on the coordination capacity region, given by

$$\mathcal{C}_{p_0,out} \triangleq \left\{ (R_1, R_2, p(z|x,y)) \; : \; \begin{array}{l} \exists p(u|x,y,z) \text{ such that} \\ p(x,y,z,u) = p_0(x,y)p(u|x)p(z|y,u) \\ |\mathcal{U}| \leq |\mathcal{X}||\mathcal{Y}||\mathcal{Z}|, \\ R_1 \geq I(X;U|Y), \\ R_2 \geq I(X,Y;Z). \end{array} \right\}.$$

Also, define $\mathcal{R}_{p_0,in}(p(z|x,y))$ and $\mathcal{R}_{p_0,out}(p(z|x,y))$ to be the sets of rate pairs in $\mathcal{C}_{p_0,in}$

and $\mathcal{C}_{p_0,out}$ corresponding to the desired distribution $p(z|x,y)$.

**Theorem 8** (Coordination capacity region bounds)**.** *The coordination capacity region $\mathcal{C}_{p_0}$ for empirical coordination in the cascade multiterminal network of Figure 2.12 is bounded by*

$$\mathcal{C}_{p_0,in} \subset \mathcal{C}_{p_0} \subset \mathcal{C}_{p_0,out}.$$

*Discussion:* The regions $\mathcal{C}_{p_0,in}$ and $\mathcal{C}_{p_0,out}$ are convex. A time-sharing random variable can be lumped into the auxiliary random variable $U$ in the definition of $\mathcal{C}_{p_0,in}$ to show convexity.

The inner bound $\mathcal{C}_{p_0,in}$ is achieved by dividing the message from node X into two parts. One part, represented by $U$, is sent to all nodes, relayed by node Y to node Z. The other part, represented by $V$, is sent only to node Y. Then node Y recompresses $V$ along with $Y$.

The outer bound $\mathcal{C}_{p_0,out}$ is a combination of the Wyner-Ziv [23] bound for source coding with side information at the decoder, obtained by letting node Y and node Z fully cooperate, and the two-node bound of Theorem 3, obtained by letting node X and node Y fully cooperate.

For some distributions, the bounds in Theorem 8 are tight and the rate-coordination region $\mathcal{R}_{p_0} = \mathcal{R}_{p_0,in} = \mathcal{R}_{p_0,out}$. This is true for all distributions where $X - Y - Z$ form a Morkov chain or $Y - X - Z$ form a Markov chain. In the first case, where $X - Y - Z$ form a Morkov chain, choosing $U = V = \emptyset$ in the definition of $\mathcal{C}_{p_0,in}$ reduces the region to all rate pairs such that $R_2 \geq I(Y;Z)$, which meets the outer bound $\mathcal{C}_{p_0,out}$. In the second case, where $Y - X - Z$ form a Morkov chain, choosing $U = Z$ and $V = \emptyset$ reduces the region to all rate pairs such that $R_1 \geq I(X;Z|Y)$ and $R_2 \geq I(X;Z)$, which meets the outer bound. Therefore, we find as special cases that the bounds in Theorem 8 are tight if $X$ is a function of $Y$, if $Y$ is a function of $X$, or if the reconstruction $Z$ is a function of $X$ and $Y$ [24].

Table 2.2 shows choices of $U$ and $V$ from $\mathcal{R}_{p_0,in}$ that yield $\mathcal{R}_{p_0,in} = \mathcal{R}_{p_0,out}$ in each of the above cases. In case 3, $V$ is selected to minimize $R_1$ along the lines of [25].

Table 2.2: Known capacity region (cases where $\mathcal{R}_{p_0,in} = \mathcal{R}_{p_0,out}$).

| | Condition | Auxiliary |
|---|---|---|
| Case 1: | $X - Y - Z$ | $U = \emptyset, V = \emptyset$ |
| Case 2: | $Y - X - Z$ | $U = Z, V = \emptyset$ |
| Case 3: | $Z = f(X, Y)$ | $U = \emptyset$ |

**Example 5** (Task assignment)**.** *Consider again a task assignment setting similar to Example 3, where three tasks are to be assigned without duplication to the three nodes X, Y, and Z, and the assignments for nodes X and Y are chosen uniformly at random by nature among all pairs of tasks where $X \neq Y$. A distribution capturing this coordination behavior is the uniform distribution over the six permutations of task assignments. Let $\hat{p}_0(x, y)$ be the distributions obtained by sampling X and Y uniformly at random from the set $\{1, 2, 3\}$ without replacement, and let $\hat{p}(z|x, y)$ be the degenerate distribution where Z is the remaining unassigned task in $\{1, 2, 3\}$. Figure 2.13 illustrates a valid outcome of the task assignments.*

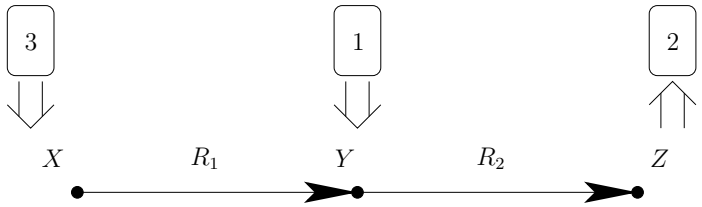


Figure 2.13: *Task assignment in the cascade multiterminal network.* Three tasks, numbered 1, 2, and 3, are distributed among three nodes X, Y, and Z in the cascade multiterminal network setting. The task assignments for X and Y are given randomly by nature but different from each other. What rates $R_1$ and $R_2$ are necessary to allow Z to choose a different task from both X and Y?

*Task assignment in the cascade multiterminal network amounts to computing a function $Z(X, Y)$, and the bounds in Theorem 8 are tight in such cases. The rate-coordination region $\mathcal{R}_{\hat{p}_0}(\hat{p}(z|x, y))$ is given by*

$$\mathcal{R}_{\hat{p}_0}(\hat{p}(z|x,y)) \;=\; \left\{ (R_1, R_2) \;:\; \begin{array}{l} R_1 \geq \log 2, \\ R_2 \geq \log 3. \end{array} \right\}.$$

*This is achieved by letting $U = \emptyset$ and $V = X$ in the definition of $\mathcal{C}_{p_0,in}$. To show that this region meets the outer bound $\mathcal{C}_{p_0,out}$, make the observation that $I(X;U|Y) \geq I(X;Z|Y)$ in relation to the bound on $R_1$, since $X - (Y,U) - Z$ forms a Markov chain.*

## 2.3 Rate-distortion theory

The challenge of describing random sources of information with the fewest bits possible can be defined in a number of different ways. Traditionally, source coding in networks follows the path of rate-distortion theory by establishing multiple distortion penalties for the multiple sources and reconstructions in the network. Yet, fundamentally, the rate-distortion problem is intimately connected to empirical coordination.

The basic result of rate-distortion theory for a single memoryless source states that in order to achieve any desired distortion level you must find an appropriate conditional distribution of the reconstruction $\hat{X}$ given the source $X$ and then use a communication rate larger than the mutual information $I(X;\hat{X})$. This lends itself to the interpretation that optimal encoding for a rate-distortion setting really comes down to coordinating a reconstruction sequence with a source sequence according to a selected joint distribution. Here we make that observation formal by showing that in general, even in networks, the rate-distortion region is a projection of the coordination capacity region.

The coordination capacity region $\mathcal{C}_{p_0}$ is a set of rate-coordination tuples. We can express rate-coordination tuples as vectors. For example, in the cascade network of Section 2.2.3 there are two rates $R_1$ and $R_2$. The actions in this network are $X$, $Y$, and $Z$, where $X$ is given by nature. Order the space $\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$ in a sequence $(x_1, y_1, z_1), ..., (x_m, y_m, z_m)$, where $m = |\mathcal{X}||\mathcal{Y}||\mathcal{Z}|$. The rate-coordination tuples $(R_1, R_2, p(y,z|x))$ can be expressed as vectors $[R_1, R_2, p(y_1, z_1|x_1), ..., p(y_m, z_m|x_m)]^T$.

The rate-distortion region $\mathcal{D}_{p_0}$ is the closure of the set of rate-distortion tuples that are achievable in a network. We say that a distortion $D$ is achievable if there exists a rate-distortion code that gives an expected average distortion less than $D$, using $d$ as a distortion measurement. For example, in the cascade network of Section 2.2.3 we

might have two distortion functions: The function $d_1(x, y)$ measures the distortion in the reconstruction at node Y; the function $d_2(x, y, z)$ evaluates distortion jointly between the reconstructions at nodes Y and Z. The rate-distortion region $\mathcal{D}_{p_0}$ would consist of tuples $(R_1, R_2, D_1, D_2)$, which indicate that using rates $R_1$ and $R_2$ in the network, a source distributed according to $p_0(x)$ can be encoded to achieve no more than $D_1$ expected average distortion as measured by $d_1$ and $D_2$ distortion as measured by $d_2$.

The relationship between the rate-distortion region $\mathcal{D}_{p_0}$ and the coordination capacity region $\mathcal{C}_{p_0}$ is that of a linear projection. Suppose we have multiple finite-valued distortion functions $d_1, ..., d_k$. We construct a distortion matrix $D$ using the same enumeration $(x_1, y_1, z_1), ..., (x_m, y_m, z_m)$ of the space $\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$ as was used to vectorize the tuples in $\mathcal{C}_{p_0}$:

$$D \triangleq \begin{bmatrix} d_1(x_1, y_1, z_1)p_0(x_1) & \cdots & d_1(x_m, y_m, z_m)p_0(x_m) \\ \vdots & \vdots & \vdots \\ d_k(x_1, y_1, z_1)p_0(x_1) & \cdots & d_k(x_m, y_m, z_m)p_0(x_m) \end{bmatrix}.$$

The distortion matrix $D$ is embedded in a block diagonal matrix $A$ where the upper-left block is the identity matrix $I$ with the same dimension as the number of rates in the network:

$$A \triangleq \begin{bmatrix} I & 0 \\ 0 & D \end{bmatrix}.$$

**Theorem 9** (Rate-distortion region)**.** *The rate-distortion region $\mathcal{D}_{p_0}$ for a memoryless source with distribution $p_0$ in any rate-limited network is a linear projection of the coordination capacity region $\mathcal{C}_{p_0}$ by the matrix $A$,*

$$\mathcal{D}_{p_0} = A\,\mathcal{C}_{p_0}.$$

*We treat the elements of $\mathcal{D}_{p_0}$ and $\mathcal{C}_{p_0}$ as vectors, as discussed, and the matrix multiplication by $A$ is the standard set multiplication.*

*Discussion:* The proof of Theorem 9 can be found in Section 2.4. Since the coordination capacity region $\mathcal{C}_{p_0}$ is a convex set, the rate-distortion region $\mathcal{D}_{p_0}$ is also a convex set.

Clearly we can use a coordination code to achieve the corresponding distortion in a rate-distortion setting. But the theorem makes a stronger statement. It says that there is not a more efficient way of satisfying distortion limits in any network setting with memoryless sources than by using a code that produces the same joint type for almost every observation of the sources. It is conceivable that a rate-distortion code for a network setting would produce a variety of different joint types, each satisfying the distortion limit, but varying depending on the particular source sequence observed. However, given such a rate-distortion code, repeated uses will produce a longer coordination code that consistently achieves coordination according to the expected joint type. The expected joint type of a good rate-distortion code can be shown to satisfy the distortion constraints.

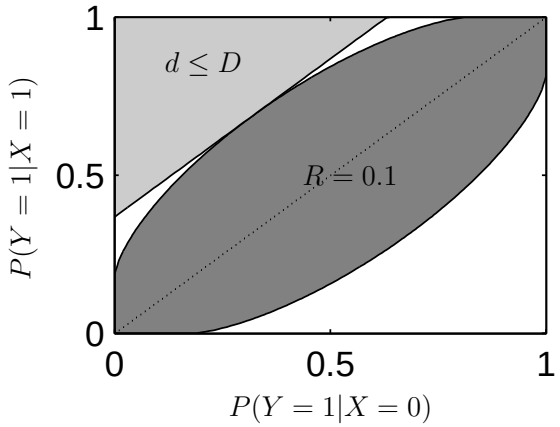


Figure 2.14: *Coordination capacity and rate-distortion.* The coordination-rate region for a uniform binary source $X$ and binary action $Y$, where $X$ is described at rate $R = 0.1$ bits to node Y in the two-node network. The shaded region shows distributions with Hamming distortion less than $D$, where $D$ is chosen to satisfy $R(D) = 0.1$ bits.

Geometrically, each distortion constraint defines a hyperplane that divides the coordination-rate region into two sets—one that satisfies the distortion constraint and one that does not. Therefore, minimizing the distortion for fixed rates in the

network amounts to finding optimal extreme points in the coordination-rate region in the directions orthogonal to these hyperplanes. Figure 2.14 shows the coordination-rate region for $R = 0.1$ bits in the two-node network of Section 2.2.1, with a uniform binary source $X$ and binary $Y$. The figure also shows the region satisfying a Hamming distortion constraint $D$.

## 2.4 Proofs

### 2.4.1 Achievability

For a distribution $p(x)$, define the *typical set* $\mathcal{T}_\epsilon^{(n)}$ with respect to $p(x)$ to be sequences $x^n$ whose types are $\epsilon$-close to $p(x)$ in total variation. That is,

$$\mathcal{T}_\epsilon^{(n)} \triangleq \{x^n \in \mathcal{X}^n : \|P_{x^n}(x) - p(x)\|_{TV} < \epsilon\}. \tag{2.10}$$

This definition is almost the same as the definition of the strongly typical set $\mathcal{A}_\epsilon^{*(n)}$ found in (10.106) of Cover and Thomas [26], and it shares the same important properties. The difference is that here we give a total variation constraint ($L_1$ distance) on the type of the sequence rather than an element-wise constraint ($L_\infty$ distance).[2] We deal with $\mathcal{T}_\epsilon^{(n)}$ since it relates more closely to the definition of achievability in Definition 5. However, the sets are almost the same, as the following sandwich suggests:

$$\mathcal{A}_\epsilon^{*(n)} \subset \mathcal{T}_\epsilon^{(n)} \subset \mathcal{A}_{\epsilon|\mathcal{X}|}^{*(n)}.$$

A jointly typical set with respect to a joint distribution $p(x, y)$ inherits the same definition as (2.10), where total variation of the type is measured with respect to the joint distribution. Thus, achieving empirical coordination with respect to a joint distribution is a matter of constructing actions that are $\epsilon$-*jointly typical* (i.e. in the jointly typical set $\mathcal{T}_\epsilon^{(n)}$) with high probability for arbitrary $\epsilon$.

[2] Additionally, our definition of the typical set handles the zero probability events more liberally, but this doesn't present any serious complications.

### Strong Markov lemma

The Markov Lemma [19] makes a statement about how likely a set of sequences will be jointly typical with respect to a Markov chain given that adjacent pairs in the chain are jointly typical. It is generally used to establish joint typicality in a source coding scheme where side information is not known to the encoder. However, as the network and encoding scheme become more intricate, the standard Markov Lemma lacks the necessary strength. Here we introduce a generalization.[3]

**Theorem 10** (Strong Markov lemma)**.** *Given a joint distribution $p(x, y, z)$ on the alphabet $\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$ that yields a Markov chain $X - Y - Z$ (i.e. $p(x, y, z) = p(y)p(x|y)p(z|y)$), let $x^n$ and $y^n$ be arbitrary sequences that are $\epsilon$-jointly typical. Suppose that $Z^n$ is chosen to be $\epsilon$-jointly typical with $y^n$ and additionally has a distribution that is permutation-invariant with respect to $y^n$, which is to say, for all $z^n$ and $\tilde{z}^n$,*

$$P_{y^n,z^n} = P_{y^n,\tilde{z}^n} \quad \Rightarrow \quad P(Z^n = z^n) = P(Z^n = \tilde{z}^n). \tag{2.11}$$

*(Notice that this condition is commonly satisfied in a variety of random coding proof techniques.) Then,*

$$\Pr\left((x^n, y^n, Z^n) \in \mathcal{T}_{4\epsilon}^{(n)}\right) \quad \rightarrow \quad 1$$

*exponentially fast as $n$ goes to infinity.*

*Proof.* The proof of Theorem 10 relies mainly on Lemma 11 (found in the sequel) and the repeated use of the triangle inequality. Suppose that (2.12), the high probability event of Lemma 11, holds true, namely

$$\|P_{x^n,y^n,Z^n} - P_{x^n,y^n}P_{Z^n|y^n}\|_{TV} \quad < \quad \epsilon.$$

[3]Through conversation we discovered that similar effort is being made by Young-Han Kim and Abbas El Gamal and may shortly be found in the Stanford EE478 Lecture Notes.

By the definition of total variation one can easily show that

$$\begin{aligned} \|P_{x^n,y^n}P_{Z^n|y^n} - p_{X,Y}P_{Z^n|y^n}\|_{TV} &= \|P_{x^n,y^n} - p_{X,Y}\|_{TV} \\ &< \epsilon. \end{aligned}$$

Similarly,

$$\begin{aligned} \|p_Y p_{X|Y}P_{Z^n|y^n} - P_{y^n}p_{X|Y}P_{Z^n|y^n}\|_{TV} &= \|p_Y - P_{y^n}\|_{TV} \\ &< \epsilon. \end{aligned}$$

And finally,

$$\begin{aligned} \|P_{y^n,Z^n}p_{X|Y} - p_{X,Y,Z}\|_{TV} &= \|P_{y^n,Z^n} - p_{Y,Z}\|_{TV} \\ &< \epsilon. \end{aligned}$$

Thus, the triangle inequality gives

$$\|P_{x^n,y^n,Z^n} - p_{X,Y,Z}\|_{TV} < 4\epsilon.$$

□

**Lemma 11** (Markov tendency)**.** *Let $x^n \in \mathcal{X}^n$ and $y^n \in \mathcal{Y}^n$ be arbitrary sequences. Suppose that the random sequence $Z^n \in \mathcal{Z}^n$ has a distribution that is permutation-invariant with respect to $y^n$, as in (2.11). Then with high probability which only depends on the sizes of the alphabets $\mathcal{X}$, $\mathcal{Y}$, and $\mathcal{Z}$, the joint type $P_{x^n,y^n,Z^n}$ will be $\epsilon$-close to the Markov joint type. That is, for any $\epsilon > 0$,*

$$\|P_{x^n,y^n,Z^n} - P_{x^n,y^n}P_{Z^n|y^n}\|_{TV} < \epsilon, \tag{2.12}$$

*with a probability of at least $1-2^{-\alpha n+\beta\log n}$, where $\alpha$ and $\beta$ only depend on the alphabet sizes and $\epsilon$.*

*Proof.* We start by defining two constants that simplify this discussion. The first

constant, $\alpha$, is the key to obtaining the uniform bound that Lemma 11 provides.

$$\begin{aligned}\alpha &\triangleq \min_{p(x,y,z)\in\mathcal{S}_{\mathcal{X},\mathcal{Y},\mathcal{Z}} \,:\, \|p(x,y,z)-p(x,y)p(z|y)\|_{TV}\geq\epsilon} I(X;Z|Y),\\ \beta &\triangleq 2|\mathcal{X}||\mathcal{Y}||\mathcal{Z}|.\end{aligned}$$

Here $\mathcal{S}_{\mathcal{X},\mathcal{Y},\mathcal{Z}}$ is the simplex with dimension corresponding to the product of the alphabet sizes. Notice that $\alpha$ is defined as a minimization of a continuous function over a compact set; therefore, by analysis we know that the minimum is achieved in the set. Since $I(X;Z|Y)$ is positive for any distribution that does not form a Markov chain $X-Y-Z$, we find that $\alpha$ is positive for $\epsilon>0$. The constants $\alpha$ and $\beta$ are functions of $\epsilon$ and the alphabet sizes $|\mathcal{X}|$, $|\mathcal{Y}|$, and $|\mathcal{Z}|$.

We categorize sequences into sets with the same joint type. The *type class* $T_{p(y,z)}$ is defined as

$$T_{p(y,z)} \triangleq \{(y^n,z^n) \,:\, P_{y^n,z^n}=p(y,z)\}.$$

We also define a *conditional type class* $T_{p(z|y)}(y^n)$ to be the set of $z^n$ sequences such that the pair $(y^n,z^n)$ are in the type class $T_{p(y,z)}$. Namely,

$$T_{p(z|y)}(y^n) \triangleq \{z^n \,:\, P_{y^n,z^n}=p(z|y)P_{y^n}\}.$$

We will show that the statement made in (2.12) is true conditionally for each conditional type class $T_{p(z|y)}(y^n)$ and therefore must be true overall.

Suppose $Z^n$ falls in the conditional type class $T_{P_{\bar{z}^n|y^n}}(y^n)$. By assumption (2.11), all $z^n$ in this type class are equally likely. Assessing probabilities simply becomes a matter of counting. From the method of types [26] we know that

$$\left|T_{P_{\bar{z}^n|y^n}}(y^n)\right| \geq n^{-|\mathcal{Y}||\mathcal{Z}|}2^{nH_{P_{y^n,\bar{z}^n}}(Z|Y)}.$$

We also can bound the number of $z^n$ sequences in $T_{P_{\bar{z}^n|y^n}}(y^n)$ that do not satisfy

(2.12). These sequences must fall in a conditional type class $T_{P_{\bar{z}^n|x^n,y^n}}(x^n, y^n)$ where

$$\left\| P_{x^n,y^n,\bar{z}^n} - P_{x^n,y^n} P_{\bar{z}^n|y^n} \right\|_{TV} \;\geq\; \epsilon.$$

For each such type class, the size can be bounded by

$$\begin{aligned}
\left| T_{P_{\bar{z}^n|x^n,y^n}}(x^n, y^n) \right| &\leq 2^{nH_{P_{x^n,y^n,\bar{z}^n}}(Z|X,Y)} \\
&= 2^{n\left(H_{P_{y^n,\bar{z}^n}}(Z|Y) - I_{P_{x^n,y^n,\bar{z}^n}}(X;Z|Y)\right)} \\
&\leq 2^{n\left(H_{P_{y^n,\bar{z}^n}}(Z|Y) - \alpha\right)}.
\end{aligned}$$

Furthermore, there are only polynomially many types, bounded by $n^{|\mathcal{X}||\mathcal{Y}||\mathcal{Z}|}$. Therefore, the probability that $Z^n$ does not satisfy (2.12) for any $P_{\bar{z}^n|y^n}$ is bounded by

$$\begin{aligned}
\Pr(\text{ not (2.12)} \mid Z^n \in T_{P_{\bar{z}^n|y^n}}(y^n)\ ) &= \frac{\left| \{ z^n \in T_{P_{\bar{z}^n|y^n}}(y^n) : \text{ not (2.12)} \} \right|}{\left| T_{P_{\bar{z}^n|y^n}}(y^n) \right|} \\
&\leq \frac{n^{|\mathcal{X}||\mathcal{Y}||\mathcal{Z}|} 2^{n\left(H_{P_{y^n,\bar{z}^n}}(Z|Y) - \alpha\right)}}{n^{-|\mathcal{Y}||\mathcal{Z}|} 2^{nH_{P_{y^n,\bar{z}^n}}(Z|Y)}} \\
&= n^{|\mathcal{Y}||\mathcal{Z}| + |\mathcal{X}||\mathcal{Y}||\mathcal{Z}|} 2^{-\alpha n} \\
&\leq 2^{-\alpha n + \beta \log n}.
\end{aligned}$$

□

**Generic achievability proof**

The coding techniques for achieving the empirical coordination regions in this chapter are familiar from rate distortion theory. We construct random codebooks (based on common randomness) and show that a particular randomized encoding scheme performs well on average, resulting in jointly-typical actions with high probability. Therefore, there must be at least one deterministic scheme that performs well. Here we prove one generally useful example to verify that the rate-distortion techniques actually do work for achieving empirical coordination.

Consider the two-node source coding setting of Figure 2.15 with arbitrary sequences $x^n$, $y^n$, and $z^n$ that are $\epsilon$-jointly typical according to a joint distribution $p(x, y, z)$. The sequences $x^n$ and $y^n$ are available to the encoder at node 1, while $y^n$ and $z^n$ are available to the decoder at node 2. We can think of $x^n$ as the source to be encoded and $y^n$ and $z^n$ as side information known to either both nodes or the decoder only, respectively. Communication from node 1 to node 2 at rate $R$ is used to produce a sequence $U^n$. Original results related to this setting in the context of rate-distortion theory can be found in the work of Wyner and Ziv [23]. Here we analyze a randomized coding scheme that attempts to produce a sequence $U^n$ at the decoder such that $(x^n, y^n, z^n, U^n)$ are $(8\epsilon)$-jointly typical with respect to a joint distribution of the form $p(x, y, z)p(u|x, y)$. We give a scheme that uses a communication rate of $R > I(X; U|Y, Z)$ and is successful with probability approaching one as $n$ tends to infinity for all jointly typical sequences $x^n$, $y^n$, and $z^n$.

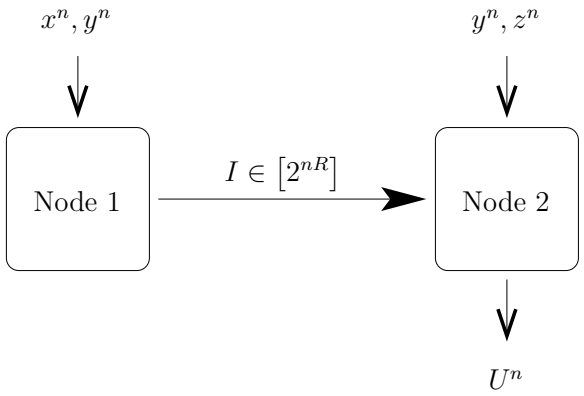


Figure 2.15: *Two nodes with side information.* This network represents a generic source coding setting encountered in networks and will illustrate standard encoding techniques. The sequences $x^n$, $y^n$, and $z^n$ are jointly typical with respect to $p_0(x, y, z)$. Only $x^n$ and $y^n$ are observed by the encoder at node 1. A message is sent to specify $U^n$ to node 2 at rate $R$. A randomized coding scheme can produce $U^n$ to be jointly typical with $(x^n, y^n, z^n)$ with respect to a Markov chain $Z - (X, Y) - U$ with high probability, regardless of the particular sequences $x^n$, $y^n$, and $z^n$, as long as the rate is greater than the conditional mutual information $I(X; U|Y, Z)$.

The $(2^{nR}, n)$ coordination codes consist of a randomized encoding function

$$i \;:\; \mathcal{X}^n \times \mathcal{Y}^n \times \Omega \longrightarrow \{1, ..., 2^{nR}\},$$

and a randomized decoding function

$$u^n \;:\; \{1,...,2^{nR}\} \times \mathcal{Y}^n \times \mathcal{Z}^n \times \Omega \longrightarrow \mathcal{U}^n.$$

These functions are random simply because the common randomness $\omega$ is involved for generating random codebooks.

The sequences $x^n, y^n$, and $z^n$ are arbitrary jointly typical sequences according to $p_0(x,y,z)$, and the sequence $U^n$ is a randomized function of $x^n, y^n$, and $z^n$ given by implementing the coordination code as

$$U^n \;=\; u^n(i(x^n,y^n,\omega),y^n,z^n,\omega).$$

**Lemma 12** (Generic coordination with side information)**.** *For the two-node network with side information of Figure 2.15 and any joint distribution of the form $p(x,y,z)p(u|x,y)$, a sequence of randomized coordination codes at rate $R > I(X;U|Y,Z)$ exists for which*

$$\Pr\left((x^n,y^n,z^n,U^n) \in \mathcal{T}_{8\epsilon}^{(n)}\right) \;\to\; 1$$

*as $n$ goes to infinity, uniformly for all $(x^n,y^n,z^n) \in \mathcal{T}_\epsilon^{(n)}$.*

*Proof.* Consider a joint distribution $p(x,y,z)p(u|x,y)$ and a value of $R$ such that $R - I(X;U|Y,Z) = \gamma > 0$. We first over-cover the typical set of $(x^n,y^n)$ using a codebook of size $2^{nR_c}$, where $R_c = I(X,Y;U) + \gamma/2$. We then randomly categorize the codebook sequences into $2^{nR}$ bins, yielding $2^{nR_b}$ sequences in each bin, where

$$\begin{aligned} R_b &= R_c - R \\ &= I(X,Y;U) - I(X;U|Y,Z) - \gamma/2 \\ &= I(X,Y,Z;U) - I(X;U|Y,Z) - \gamma/2 \\ &= I(Y,Z;U) - \gamma/2. \end{aligned}$$

*Codebook:* Generate a codebook $\mathbb{C}$ of $2^{n(I(X;Z|U)-\epsilon)}$ sequences $u^n(j)$ independently

according to the marginal distribution $p(u)$, namely $\prod_{i=1}^{n} p(u_i)$. Randomly and independently assign each one a bin number $b(u^n(j))$ in the set $\{1, ..., 2^{nR}\}$.

*Encoder:* The encoding function $i(x^n, y^n, \omega)$ can be explained as follows. Search the codebook $\mathbb{C}$ and identify an index $j$ such that $(x^n, y^n, u^n(j)) \in \mathcal{T}_{2\epsilon}^{(n)}$. If multiple exist, select the first such $j$. If none exist, select $j = 1$. Send the bin number $i(x^n, y^n) = b(u^n(j))$.

*Decoder:* The decoding function $u^n(i, y^n, z^n, \omega)$ can be explained as follows. Consider the codebook $\mathbb{C}$ and identify an index $j$ such that $(y^n, z^n, u^n(j)) \in \mathcal{T}_{8\epsilon}^{(n)}$ and $b(u^n(j)) = i$. If multiple exist, select the first such $j$. If none exist, select $j = 1$. Produce the sequence $U^n = u^n(j)$.

*Error Analysis:* We conservatively declare errors for any of the following, $E_1$, $E_2$, or $E_3$.

*Error 1: The encoder does not find a* $(2\epsilon)$*-jointly typical sequence in the codebook.* By the method of types one can show, as in Lemma 10.6.2 of [26], that each sequence in $\mathbb{C}$ is $(2\epsilon)$-jointly typical with $(x^n, y^n)$ with probability greater than $2^{-n(I(X,Y;U)+\delta(\epsilon))}$ for $n$ large enough, where $\delta(\epsilon)$ goes to zero as $\epsilon$ goes to zero.

Each sequence in the codebook $\mathbb{C}$ is generated independently, so the probability that none of them are jointly typical is bounded by

$$\begin{aligned} \Pr(E_1) &\leq (1 - 2^{-n(I(X,Y;U)+\delta(\epsilon))})^{2^{nR_c}} \\ &\leq e^{-2^{nR_c} 2^{-n(I(X,Y;U)+\delta(\epsilon))}} \\ &= e^{-2^{n(R_c - I(X,Y;U) - \delta(\epsilon))}} \\ &= e^{-2^{n(\gamma/2 - \delta(\epsilon))}}. \end{aligned}$$

If $\epsilon$ is chosen small enough, this tends to zero as $n$ tends to infinity.

*Error 2: The sequence identified by the encoder is not* $(8\epsilon)$*-jointly typical with* $(x^n, y^n, z^n)$*.* Assuming $E_1$ did not occur, because of the Markovity $Z - (X, Y) - U$ implied by $p(x, y, z)p(u|x, y)$ and the symmetry of our codebook construction, we can invoke Theorem 10 to verify that the conditional probability $Pr(E_2|E_1^c)$ is arbitrarily small for large enough $n$.

*Error 3: The decoder finds more than one eligible action sequences.* Assume that

$E_1$ and $E_2$ did not occur. If the decoder considers the same index $j$ as the encoder selected, then certainly $u^n(j)$ will be be eligible, which is to say it will be $(8\epsilon)$-jointly typical with $(y^n, z^n)$, and the bin index will match the received message. For all other sequences in the codebook $\mathbb{C}$, an appeal to the property of iterated expectation indicates that the probability of eligibility is slightly less than the a priori probability that a randomly generated sequence and bin number will yield eligibility, which is upper bounded by $2^{-nR}2^{-n(I(Z;U|Y)-\delta(8\epsilon))}$. Therefore, by the method of types and the union bound,

$$\begin{aligned}
\Pr(E_3|E_1^c, E_2^c) &\leq 2^{nR_c}2^{-n(R+I(Z,Y;U)-\delta(8\epsilon))} \\
&= 2^{-n(R-R_c+I(Z,Y;U)-\delta(8\epsilon))} \\
&= 2^{-n(I(Z,Y;U)-R_b-\delta(8\epsilon))} \\
&= 2^{-n(\gamma/2-\delta(8\epsilon))}.
\end{aligned}$$

This also tends to zero as $n$ tends to infinity for small enough $\epsilon$. □

With the result of Lemma 12 in mind, we can confidently talk about using communication to specify sequences across links in a network. Throughout the following explanations we will no longer pay particular attention to the $\epsilon$ in the $\epsilon$-jointly typical set. Instead, we will simply make reference to the generic jointly typical set.

### Two nodes - Theorem 3

It is clear from Lemma 12 that an action sequence $Y^n$ jointly typical with $X^n$ can be specified with high probability using any rate $R > I(X;Y)$. With high probability $X^n$ will be a typical sequence. Apply Lemma 12 with $Y = Z = \emptyset$.

### Isolated node - Theorem 4

No proof is necessary, as this is a special case of the cascade network with $R_2 = 0$.

**Cascade - Theorem 5**

The cascade network of Figure 2.6 has a sequence $X^n$ given by nature. The actions $X^n$ will be typical with high probability. Consider the desired coordination $p(y, z|x)$. A sequence $Z^n$ can be specified with rate $R_Z > I(X;Z)$ to be jointly typical with $X^n$. This communication is sent to node Y and forwarded on to node Z. Additionally, now that every node knows $Z^n$, a sequence $Y^n$ can be specified with rate $R_Y > I(X;Y|Z)$ and sent to node Y. The rates used are $R_1 = R_Y + R_Z > I(X;Y,Z)$ and $R_2 = R_Z > I(X;Z)$.

$$\begin{aligned} R_1 &= R_Y + R_Z &> I(X;Y,Z), \\ R_2 &= R_Z &> I(X;Z). \end{aligned}$$

**Degraded source - Theorem 6**

The degraded source network of Figure 2.8 has a sequence $X^n$ given by nature, known to node X, and another sequence $Y^n$, which is a letter-by-letter function of $X^n$, known to node Y. Incidentally, $Y^n$ is also known to node X because it is a function of the available information. The actions $X^n$ and $Y^n$ will be jointly typical with high probability.

Consider the desired coordination $p(z|x,y)$ and choose a distribution for the auxiliary random variable $p(u|x,y,z)$ to help achieve it. The encoder first specifies a sequence $U^n$ that is jointly typical with $X^n$ and $Y^n$. This requires a rate $R_U > I(X,Y;U) = I(X;U)$, but with binning we only need a rate of $R_1 > I(X;U|Y)$ to specify $U^n$ from node X to node Y. Binning is not used when $U^n$ is forwarded to node Z. Finally, after everyone knows $U^n$, the action sequence $Z^n$ jointly typical with $X^n$, $Y^n$, and $U^n$ is specified to node Z at a rate of $R_2 > I(X,Y;Z|U) = I(X;Z|U)$. Thus, all rates are achievable which satisfy

$$\begin{aligned} R_1 &> I(X;U|Y), \\ R_2 &> I(X;Z|U), \\ R_3 = R_U &> I(X;U). \end{aligned}$$

**Broadcast - Theorem 7**

The broadcast network of Figure 2.9 has a sequence $X^n$ given by nature, known to node X. The action sequence $X^n$ will be typical with high probability.

Consider the desired coordination $p(y, z|x)$ and choose a distribution for the auxiliary random variable $p(u|x, y, z)$ to help achieve it. We will focus on achieving one corner point of the pentagonal rate region. The encoder first specifies a sequence $U^n$ that is jointly typical with $X^n$ using a rate $R_U > I(X;U)$. This sequence is sent to both node Y and node Z. After everyone knows $U^n$, the encoder specifies an action sequence $Y^n$ that is jointly typical with $X^n$ and $U^n$ using rate $R_Y > I(X;Y|U)$. Finally, the encoder at node X, knowing both $X^n$ and $Y^n$, can specify an action sequence $Z^n$ that is jointly typical with $(X^n, Y^n, U^n)$ using a rate $R_Z > I(X, Y; Z|U)$. This results in rates

$$\begin{aligned} R_1 &= R_U + R_Y &> I(X;U) + I(X;Y|U) &= I(X;U,Y), \\ R_2 &= R_U + R_X &> I(X;U) + I(X,Y;Z|U). \end{aligned}$$

**Cascade multiterminal - Theorem 8**

The cascade multiterminal network of Figure 2.12 has a sequence $X^n$ given by nature, known to node X, and another sequence $Y^n$ given by nature, known to node Y. The actions $X^n$ and $Y^n$ will be jointly typical with high probability.

Consider the desired coordination $p(z|x, y)$ and choose a distribution for the auxiliary random variables $U$ and $V$ according to the inner bound in Theorem 8. That is, $p(x, y, z, u, v) = p(x, y)p(u, v|x)p(z|y, u, v)$. We specify a sequence $U^n$ to be jointly typical with $X^n$. By the Strong Markov Lemma (Theorem 10), in conjunction with the symmetry of our random coding scheme and the Markovity of the distribution $p(x, y)p(u|x)$, the sequence $U^n$ will be jointly typical with the pair $(X^n, Y^n)$ with high probability. Using binning, we only need a rate of $R_{U,1} > I(X;U|Y)$ to specify $U^n$ from node X to node Y (as in Lemma 12). However, we cannot use binning for the message to node Z, so we send the index of the codework itself at a rate of $R_{U,2} > I(X;U)$. Now that everyone knows the sequence $U^n$, it is treated as side

information.

A second auxiliary sequence $V^n$ is specified from node X to node Y to be jointly typical with $(X^n, Y^n, U^n)$. This scenario coincides exactly with Lemma 12, and a sufficient rate is $R_V > I(X; V|U, Y)$. Finally, an action sequence $Z^n$ is specified from node Y to node Z to be jointly typical with $(Y^n, V^n, U^n)$, where $U^n$ is side information known to the encoder and decoder. We achieve this using a rate $R_Z > I(Y, V; Z|U)$. Again, because of the symmetry of our encoding scheme, the Strong Markov Lemma (Theorem 10) tells us that $(X^n, Y^n, U^n, V^n, Z^n)$ will be jointly typical, and therefore, $(X^n, Y^n, Z^n)$ will be jointly typical.

The rates used by this scheme are

$$\begin{aligned} R_1 &= R_{U,1} + R_V > I(X; U, V|Y), \\ R_2 &= R_{U,2} + R_Z > I(X; U) + I(Y, V; Z|U). \end{aligned}$$

### 2.4.2 Converse

In proving outer bounds for the coordination capacity of various networks, a common *time mixing* trick is to make use of a random time variable $Q$ and then consider the value of a random sequence $X^n$ at the random time $Q$ using notation $X_Q$. We first make this statement precise and discuss the implications of such a construction.

Consider a coordination code for a block length $n$. We assign $Q$ to have a uniform distribution over the set $\{1, ..., n\}$, independent of the action sequences in the network. The variable $X_Q$ is simply a function of the sequence $X^n$ and the variable $Q$; namely, the variable $X_Q$ takes on the value of the $Q$th element in the sequence $X^n$. Even though all sequences of actions and auxiliary variables in the network are independent of $Q$, the variable $X_Q$ need not be independent of $Q$.

Here we list a couple of key properties of time mixing.

*Property 1: If all elements of a sequence $X^n$ are identically distributed, then $X_Q$ is independent of $Q$. Furthermore, $X_Q$ has the same distribution as $X_1$.* Verifying this property is easy when one considers the conditional distribution of $X_Q$ given $Q$.

*Property 2: For a collection of random sequences $X^n$, $Y^n$, and $Z^n$, the expected*

*joint type* $\mathbf{E}P_{X^n,Y^n,Z^n}$ *is equal to the joint distribution of the time-mixed variables* $(X_Q, Y_Q, Z_Q)$.

$$\begin{aligned}
\mathbf{E}\, P_{X^n,Y^n,Z^n}(x,y,z) &= \sum_{x^n,y^n,z^n} p(x^n,y^n,z^n) P_{X^n,Y^n,Z^n}(x,y,z) \\
&= \sum_{x^n,y^n,z^n} p(x^n,y^n,z^n) \frac{1}{n} \sum_{q=1}^{n} \mathbf{1}((x_q,y_q,z_q)=(x,y,z)) \\
&= \frac{1}{n} \sum_{q=1}^{n} \sum_{x^n,y^n,z^n} p(x^n,y^n,z^n) \mathbf{1}((x_q,y_q,z_q)=(x,y,z)) \\
&= \frac{1}{n} \sum_{q=1}^{n} p_{X_q,Y_q,Z_q}(x,y,z) \\
&= \sum_{q=1}^{n} p_{X_Q,Y_Q,Z_Q|Q}(x,y,z|q) p(q) \\
&= p_{X_Q,Y_Q,Z_Q}(x,y,z).
\end{aligned}$$

**Two nodes - Theorem 3**

Assume that a rate-coordination pair $(R, p(y|x))$ is in the interior of the coordination capacity region $\mathcal{C}_{p_0}$ for the two-node network of Figure 2.3 with source distribution $p_0(x)$. For a sequence of $(2^{nR}, n)$ coordination codes that achieves $(R, p(y|x))$, consider the induced distribution on the action sequences.

Recall that $I$ is the message from node X to node Y.

$$\begin{aligned} nR &\geq H(I) \\ &\geq I(X^n; Y^n) \\ &= \sum_{q=1}^{n} I(X_q; Y^n | X^{q-1}) \\ &= \sum_{q=1}^{n} I(X_q; Y^n, X^{q-1}) \\ &\geq \sum_{q=1}^{n} I(X_q; Y_q) \\ &= nI(X_Q; Y_Q | Q) \\ &\stackrel{a}{=} nI(X_Q; Y_Q, Q) \\ &\geq nI(X_Q; Y_Q). \end{aligned}$$

Equality $a$ comes from Property 1 of time mixing.

We would like to be able to say that the joint distribution of $X_Q$ and $Y_Q$ is arbitrarily close to $p_0(x)p(y|x)$ for some $n$. That way we could conclude, by continuity of the entropy function, that $R \geq I(X;Y)$.

The definition of achievability (Definition 5) states that

$$\|P_{X^n,Y^n,Z^n}(x,y,z) - p_0(x)p(y,z|x)\|_{TV} \longrightarrow 0 \text{ in probability.}$$

Because total variation is bounded, this implies that

$$\mathbf{E}\ \|P_{X^n,Y^n,Z^n}(x,y,z) - p_0(x)p(y,z|x)\|_{TV} \longrightarrow 0.$$

Furthermore, by the Jensen Inequality,

$$\mathbf{E}\ P_{X^n,Y^n,Z^n}(x,y,z) \longrightarrow p_0(x)p(y,z|x).$$

Now Property 2 of time mixing allows us to conclude the argument for Theorem 3.

**Isolated node - Theorem 4**

No proof is necessary, as this is a special case of the cascade network with $R_2 = 0$.

**Cascade - Theorem 5**

For the cascade network of Figure 2.6, apply the bound from the two-node network twice—once to show that the rate $R_1 \geq I(X;Y,Z)$ is needed even if node Y and node Z are allowed to fully cooperate, and once to show that the rate $R_2 \geq I(X;Z)$ is needed even if node X and node Y are allowed to fully cooperate.

**Degraded source - Theorem 6**

Assume that a rate-coordination quadruple $(R_1, R_2, R_3, p(z|x,y))$ is in the interior of the coordination capacity region $\mathcal{C}_{p_0}$ for the degraded source network of Figure 2.8 with source distribution $p_0(x)$ and the degraded relationship $Y_i = f_0(x_i)$. For a sequence of $(2^{nR_1}, 2^{nR_2}, 2^{nR_3}, n)$ coordination codes that achieves $(R_1, R_2, R_3, p(z|x,y))$, consider the induced distribution on the action sequences.

Recall that the message from node X to node Y at rate $R_1$ is labeled $I$, the message from node X to node Z at rate $R_2$ is labeled $J$, and the message from node Y to node Z at rate $R_3$ is labeled $K$. We identify the auxiliary random variable $U$ as the collection of random variables $(K, X^{Q-1}, Q)$.

$$
\begin{aligned}
nR_1 &\geq H(I) \\
&\geq H(I|Y^n) \\
&\overset{a}{=} H(I,K|Y^n) \\
&\geq H(K|Y^n) \\
&= I(X^n;K|Y^n) \\
&= \sum_{q=1}^{n} I(X_q;K|Y^n,X^{q-1}) \\
&= \sum_{q=1}^{n} I(X_q;K,X^{q-1},Y^{q-1},Y_{q+1}^n|Y_q) \\
&\geq \sum_{q=1}^{n} I(X_q;K,X^{q-1}|Y_q) \\
&= nI(X_Q;K,X^{Q-1}|Y_Q,Q) \\
&\overset{b}{=} nI(X_Q;K,X^{Q-1},Q|Y_Q) \\
&= nI(X_Q;U|Y_Q).
\end{aligned}
$$

Equality $a$ is justified because the message $K$ is a function of the message $I$ and the sequence $Y^n$. Equality $b$ comes from Property 1 of time mixing.

$$\begin{aligned}
nR_2 &\geq H(J) \\
&\geq H(J|K) \\
&\stackrel{a}{=} H(J, Z^n|K) \\
&\geq H(Z^n|K) \\
&= I(X^n; Z^n|K) \\
&\geq \sum_{q=1}^{n} I(X_q; Z^n|K, X^{q-1}) \\
&\geq \sum_{q=1}^{n} I(X_q; Z_q|K, X^{q-1}) \\
&= nI(X_Q; Z_Q|K, X^{Q-1}, Q) \\
&= nI(X_Q; Z_Q|U).
\end{aligned}$$

Equality $a$ is justified because the action sequence $Z^n$ is a function of the messages $J$ and $K$. Equality $b$ comes from Property 1 of time mixing.

$$\begin{aligned}
nR_3 &\geq H(K) \\
&= I(X^n; K) \\
&= \sum_{q=1}^{n} I(X_q; K|X^{q-1}) \\
&= nI(X_Q; K|X^{Q-1}, Q) \\
&\stackrel{a}{=} nI(X_Q; K, X^{Q-1}, Q) \\
&= nI(X_Q; U).
\end{aligned}$$

Equality $a$ comes from Property 1 of time mixing.

As seen in the proof for the two-node network, the joint distribution of $X_Q$, $Y_Q$, and $Z_Q$ is arbitrarily close to $p_0(x)\mathbf{1}(y = f_0(x))p(z|x, y)$. Therefore, the inequalities of the coordination capacity region stated in Theorem 6 must hold.

It remains to bound the cardinality of $U$. We can use the standard method rooted in the support lemma of [27]. The variable $U$ should have $|\mathcal{X}||\mathcal{Z}| - 1$ elements to preserve the joint distribution $p(x, z)$, which in turn preserves $p(x, y, z)$, $H(X)$, and $H(X|Y)$, and three more elements to preserve $H(X|U)$, $H(X|Y, U)$, and $H(X|Z, U)$.

**Broadcast - Theorem 7**

For the broadcast network of Figure 2.9, apply the bound from the two-node network three times—once to show that the rate $R_1 \geq I(X; Y)$ is needed and once to show that the rate $R_2 \geq I(X; Z)$ is needed, and finally a third time to show that the sum-rate $R_1 + R_2 = I(X; Y, Z)$ is needed even if node Y and node Z are allowed to fully cooperate.

**Cascade multiterminal - Theorem 8**

Assume that a rate-coordination triple $(R_1, R_2, p(z|x, y))$ is in the interior of the coordination capacity region $\mathcal{C}_{p_0}$ for the cascade multiterminal network of Figure 2.12 with source distribution $p_0(x, y)$. For a sequence of $(2^{nR_1}, 2^{nR_2}, n)$ coordination codes that achieves $(R_1, R_2, p(z|x, y))$, consider the induced distribution on the action sequences.

Recall that the message from node X to node Y at rate $R_1$ is labeled $I$, and the message from node Y to node Z at rate $R_2$ is labeled $J$. We identify the auxiliary random variable $U$ as the collection of random variables $(I, X^{Q-1}, Y^{Q-1}, Y_{Q+1}^n, Q)$. This is the same choice of auxiliary variable used by Wyner and Ziv [23]. Notice that $U$ satisfies the Markov chain properties $U - X_Q - Y_Q$ and $X_Q - (Y_Q, U) - Z_Q$

$$\begin{aligned}
nR_1 &\geq H(I) \\
&\geq H(I|Y^n) \\
&= I(X^n; I|Y^n) \\
&= \sum_{q=1}^{n} I(X_q; I|Y^n, X^{q-1}) \\
&= \sum_{q=1}^{n} I(X_q; I, X^{q-1}, Y^{q-1}, Y_{q+1}^n|Y_q) \\
&= nI(X_Q; I, X^{Q-1}, Y^{Q-1}, Y_{Q+1}^n|Y_Q, Q) \\
&\overset{a}{=} nI(X_Q; I, X^{Q-1}, Y^{Q-1}, Y_{Q+1}^n, Q|Y_Q) \\
&= nI(X_Q; U|Y_Q).
\end{aligned}$$

Equality $a$ comes from Property 1 of time mixing.

$$\begin{aligned}
nR_2 &\geq H(J) \\
&\geq I(X^n, Y^n; Z^n) \\
&= \sum_{q=1}^{n} I(X_q, Y_q; Z_q|X^{q-1}, Y^{q-1}) \\
&= \sum_{q=1}^{n} I(X_q, Y_q; Z_q, X^{q-1}, Y^{q-1}) \\
&\geq \sum_{q=1}^{n} I(X_q, Y_q; Z_q) \\
&= nI(X_Q, Y_Q; Z_Q|Q) \\
&\overset{a}{=} nI(X_Q, Y_Q; Z_Q, Q) \\
&\geq I(X_Q, Y_Q; Z_Q).
\end{aligned}$$

Equality $a$ comes from Property 1 of time mixing.

As seen in the proof for the two-node network, the joint distribution of $X_Q$, $Y_Q$,

and $Z_Q$ is arbitrarily close to $p_0(x,y)p(z|x,y)$. Therefore, the inequalities of the coordination capacity region stated in Theorem 8 must hold.

It remains to bound the cardinality of $U$. We can again use the standard method of [27]. Notice that $p(x,y,z|u) = p(x|u)p(y|x)p(z|y,u)$ captures all of the Markovity constraints of the outer bound. Therefore, convex mixtures of distributions of this form are valid for achieving points in the outer bound. The variable $U$ should have $|\mathcal{X}||\mathcal{Y}||\mathcal{Z}| - 1$ elements to preserve the joint distribution $p(x,y,z)$, which in turn preserves $I(X,Y;Z)$ and $H(X|Y)$, and one more element to preserve $H(X|Y,U)$.

### 2.4.3 Rate-distortion

We establish the relationship from Theorem 9 between the coordination capacity region and the rate-distortion region in two parts. First we show that $\mathcal{D}_{p_0}$ contains $A\mathcal{C}_{p_0}$ and then the other way around. To keep clutter to a minimum and without loss of generality, we only discuss a single distortion measure $d$, rate $R$, and a pair of sequences of actions $X^n$ and $Y^n$.

#### Coordination implies distortion ($\mathcal{D}_{p_0} \supset A\mathcal{C}_{p_0}$)

The distortion incurred with respect to a distortion function $d$ on a set of sequences of actions is a function of the joint type of the sequences. That is,

$$\begin{aligned} d^{(n)}(x^n,y^n) &= \frac{1}{n}\sum_{i=1}^{n} d(x_i,y_i) \\ &= \frac{1}{n}\sum_{i=1}^{n}\sum_{x,y} \mathbf{1}(x_i=x,y_i=y)d(x,y) \\ &= \sum_{x,y} d(x,y)\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}(x_i=x,y_i=y) \\ &= \sum_{x,y} d(x,y)P_{x^n,y^n}(x,y) \\ &= \mathbf{E}_{P_{x^n,y^n}} d(X,Y). \end{aligned} \tag{2.13}$$

When a rate-coordination tuple $(R, p(x,y))$ is in the interior of the coordination

capacity region $\mathcal{C}_{p_0}$, we are assured the existence of a coordination code for any $\epsilon > 0$ for which

$$\Pr(\|P_{X^n,Y^n} - p\|_{TV} > \epsilon) \;\; < \;\; \epsilon.$$

Therefore, with probability greater that $1 - \epsilon$,

$$\mathbf{E}_{P_{X^n,Y^n}} d(X,Y) \;\; \leq \;\; \mathbf{E}_p \; d(X,Y) + \epsilon d_{max}.$$

Recalling (2.13) yields,

$$\mathbf{E} \; d^{(n)}(x^n, y^n) \;\; \leq \;\; \mathbf{E}_p \; d(X,Y) + 2\epsilon d_{max}.$$

As expected, a sequence of $(2^{nR}, n)$ coordination codes that achieves empirical coordination for the joint distribution $p(x,y)$ also achieves the point in the rate-distortion region with the same rate and with distortion value $\mathbf{E}_p \; d(X,Y)$.

**Distortion implies coordination ($\mathcal{D}_{p_0} \subset A\mathcal{C}_{p_0}$)**

Suppose that a $(2^{nR}, n)$ rate-distortion codes achieves distortion $\mathbf{E} \; d^{(n)}(X^n, Y^n) \leq D$. Substituting from (2.13),

$$\mathbf{E} \left[ \mathbf{E}_{P_{X^n,Y^n}} d(X,Y) \right] \;\; \leq \;\; D.$$

However,

$$\mathbf{E} \left[ \mathbf{E}_{P_{X^n,Y^n}} d(X,Y) \right] \;\; = \;\; \mathbf{E}_{\mathbf{E}P_{X^n,Y^n}} d(X,Y)$$

by linearity.

We can achieve the rate-coordination pair $(R, \mathbf{E}P_{X^n,Y^n})$ by augmenting the rate-distortion code. If we repeat the use of the rate-distortion code over $k$ blocks of length $n$ each, then we induce a joint distribution on $(X^{kn}, Y^{kn})$ that consists of i.i.d. sub-blocks $(X^n, Y^n), ..., (X^{kn}_{kn-n+1}, Y^{kn}_{kn-n+1})$ denoted as $(X^{(1)n}, Y^{(1)n}), ..., (X^{(k)n}, Y^{(k)n})$.

By the weak law of large number,

$$\begin{aligned} P_{X^{kn},Y^{kn}} &= \frac{1}{k}\sum_{i=1}^{k} P_{X^{(i)n},Y^{(i)n}} \\ &\longrightarrow \mathbf{E}\, P_{X^n,Y^n} \text{ in probability.} \end{aligned}$$

Point-wise convergence in probability implies that as $k$ grows

$$\left\| P_{X^{kn},Y^{kn}} - \mathbf{E}\, P_{X^n,Y^n} \right\|_{TV} \longrightarrow 0 \text{ in probability.}$$

Thus, for any point $(R, D)$ in the rate-distortion region we have identified a point $(R, \mathbf{E}P_{X^n,Y^n})$ in the coordination capacity region that projects down to it.

# Chapter 3

# Strong Coordination

## 3.1 Introduction

Strong coordination addresses questions like this: If two players of a multiplayer game wish to collaborate, how should they best use communication to generate their actions?

So far we have examined coordination where the goal is to generate $Y^n$ so that the joint type $P_{X^n,Y^n}(x,y)$ is equal to $p_0(x)p(y|x)$. This goal relates to the joint behavior at the nodes in the network averaged over time. The order of the $(X_i, Y_i)$ pairs doesn't matter.

How different does the problem become if we actually want the actions at the various nodes in the network to be random according to a desired joint distribution? In this vein, we turn to a stronger notion which we call strong coordination. We want the induced distribution $\hat{p}(x^n, y^n)$ (induced by the coordination code) to be close to the true target distribution $p(x^n, y^n) = \prod_{i=1}^n p_0(x_i)p(y_i|x_i)$—so close that a statistician could not tell the difference, based on $(X^n, Y^n)$, of whether $(X^n, Y^n) \sim \hat{p}$ or $(X^n, Y^n) \sim p$. Thus, strong coordination yields an effective simulation of $n$ uses of the channel $p(y|x)$.

Clearly this new strong coordination objective is more demanding—after all, if one were to generate random actions, i.i.d. in time, according to the appropriate joint distribution, then the empirical distribution would also follow suit. However,

in some settings the added feature of coordinating random behavior can be crucial. For example, in situations where an adversary is involved, it might be important to maintain a mystery about the actions that will be generated in the network.

Strong coordination has applications in cooperative game theory, reminiscent of the framework investigated in [8]. Suppose a team shares the same payoff in a repeated game setting. An opponent tries to anticipate and exploit patterns in the team's combined actions, but a secure line of communication is available to help them coordinate. Of course, each player could communicate his randomized actions to the other players, but this is an excessive use of communication. Strong coordination according to a well-chosen joint distribution will be useful in this situation. This is explored in Section 3.4.

### 3.1.1 Problem specifics

Most of the definitions relating to empirical coordination in Chapter 2 carry over to strong coordination, including the notions of coordination codes and induced distributions. However, in the context of strong coordination, achievability has nothing to do with the joint type. Here we define strong achievability to mean that the distribution of the time-sequence of actions in the network is close in total variation to the desired joint distribution, i.i.d. in time. We discuss the strong coordination capacity region $\underline{\mathcal{C}}_{p_0}$ defined by this notion of strong achievability.

**Definition 9** (Strong achievability)**.** *A desired distribution $p(x, y, z)$ is strongly achievable if there exists a sequence of (non-deterministic) coordination codes such that the total variation between the induced distribution $p(x^n, y^n, z^n)$ and the i.i.d. desired distribution goes to zero. That is,*

$$\left\| p(x^n, y^n, z^n) - \prod_{i=1}^{n} p(x_i, y_i, z_i) \right\|_{TV} \longrightarrow 0.$$

Common randomness plays a crucial role in achieving strong coordination. For instance, in a network with no communication, only independent actions can be generated at each node without common randomness, but actions can be generated

according to any desired joint distribution if enough common randomness is available, as is illustrated in Figure 1.2 in Chapter 1. In addition, for each desired joint distribution we can identify a specific bit-rate of common randomness that must be available to the nodes in the network. This motivates us to deal with common randomness more precisely.

Aside from the communication in the network, we allow common randomness to be supplied to each node. However, to quantify the amount of common randomness, we limit it to a rate of $R_0$ bits per action. For an $n$-block coordination code, $\omega$ is uniformly distributed on the set $\Omega = \{1, ..., 2^{nR_0}\}$. In this way, common randomness is viewed as a resource alongside communication.

### 3.1.2 Preliminary observations

The strong coordination capacity region $\underline{\mathcal{C}}_{p_0}$ is not convex in general. This becomes immediately apparent when we consider a network with no communication and without any common randomness. An arbitrary joint distribution is not strongly achievable without communication or common randomness, but any extreme point in the probability simplex corresponds to a degenerate distribution that is trivially achievable. Thus we see that convex combinations of achievable points in the strong coordination capacity region are not necessarily strongly achievable, and cannot be achieved through simple time-sharing as was done for empirical coordination.

We use total variation as a measurement of fidelity for the distribution of the actions in the network. This has a number of implications. If two distributions have a small total variation between them, then a hypothesis test cannot reliably tell them apart. Additionally, the expected value of a bounded function of these random variables cannot differ by much. Steinberg and Verdú, for example, also use total variation as one of a handful of fidelity criteria when considering the simulation of random variables in [28].

Based on the success of random codebooks in information theory and source coding in particular, it seems hopeful that we might always be able to use common randomness to augment a coordination code intended for empirical coordination to result in

a randomized coordination code that achieves strong coordination. Bennett et. al. demonstrate this principle for the two-node setting with their reverse Shannon theorem [13]. They use common randomness to generate a random codebook. Then the encoder synthesizes a memoryless channel and finds a sequence in the codebook with the same joint type as the synthesized output. Will methods like this work in other network coordination settings as well? The following conjecture makes this statement precise and is consistent with both networks considered for strong coordination in this section of the paper.

**Conjecture 1** (Strong meets empirical coordination)**.** *With enough common randomness, for instance if* $\omega \sim Unif\{[0,1]\}$*, the strong coordination capacity region is the same as the empirical coordination capacity region for any specific network setting . That is,*

$$\textit{With unlimited common randomness:} \qquad \underline{\mathcal{C}}_{p_0} \;=\; \mathcal{C}_{p_0}.$$

## 3.2 No communication

Here we characterize the strong coordination capacity region $\underline{\mathcal{C}}$ for the no communication network of Figure 3.1. A collection of nodes X, Y, and Z generate actions according to the joint distribution $p(x, y, z)$ using only common randomness (and private randomization). The strong coordination capacity region characterizes the set of joint distributions that can be achieved with common randomness at a rate of $R_0$ bits per action.

Wyner considered a two-node setting in [29], where correlated random variables are constructed based on common randomness. He found the amount of common randomness needed and named the quantity "common information." Here we extend that result to three nodes, and the conclusion for any number of nodes is immediately apparent.

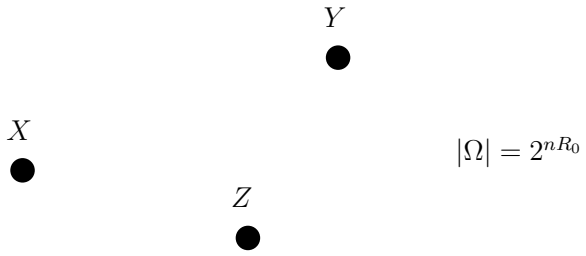


Figure 3.1: *No communication.* Three nodes generate actions $X$, $Y$, and $Z$ according to $p(x, y, z)$ without communication. The rate of common randomness needed is characterized in Theorem 13.

The $n$-block coordination codes consist of three non-deterministic decoding functions,

$$\begin{aligned} x^n &: \{1, ..., 2^{nR_0}\} \longrightarrow \mathcal{X}^n, \\ y^n &: \{1, ..., 2^{nR_0}\} \longrightarrow \mathcal{Y}^n, \\ z^n &: \{1, ..., 2^{nR_0}\} \longrightarrow \mathcal{Z}^n. \end{aligned}$$

Each function can use private randomization to probabilistically map the common random bits $\omega$ to action sequences. That is, the functions $x^n(\omega)$, $y^n(\omega)$, and $z^n(\omega)$ behave according to conditional probability mass functions $p(x^n|\omega)$, $p(y^n|\omega)$, and $p(z^n|\omega)$.

The rate region given in Theorem 13 can be generalized to any number of nodes.

**Theorem 13** (Strong coordination capacity region)**.** *The strong coordination capacity region $\underline{\mathcal{C}}$ for the no communication network of Figure 3.1 is given by*

$$\underline{\mathcal{C}} = \left\{ (R_0, p(x, y, z)) : \begin{array}{l} \exists p(u|x, y, z) \text{ such that} \\ p(x, y, z, u) = p(u)p(x|u)p(y|u)p(z|u) \\ |\mathcal{U}| \leq |\mathcal{X}||\mathcal{Y}||\mathcal{Z}|, \\ R_0 \geq I(X, Y, Z; U). \end{array} \right\}.$$

*Discussion:* The proof of Theorem 13 in Section 3.5 follows the same approach as Wyner's common information proof. This generalization can be interpreted as a

proposed measurement of common information between a group of random variables. Namely, the amount of common randomness needed to generate a collection of random variables at isolated nodes is the amount of common information between them. However, it would also be interesting to consider a richer problem by allowing each subset of nodes to have an independent common random variable and investigating all of the rates involved.

**Example 6** (Task assignment)**.** *Suppose there are tasks numbered* $1, ..., k$*, and three of them are to be assigned randomly to the three nodes X, Y, and Z without duplication. That is, the desired distribution* $\hat{p}(x, y, z)$ *for the three actions in the network is the distribution obtained by sampling* $X$*,* $Y$*, and* $Z$ *uniformly at random from the set* $\{1, ..., k\}$ *without replacement. The three nodes do not communicate but have access to common randomness at a rate of* $R_0$ *bits per action. We want to determine the infimum of rates* $R_0$ *required to strongly achieve* $\hat{p}(x, y, z)$*. Figure 3.2 illustrates a valid outcome of the task assignments.*

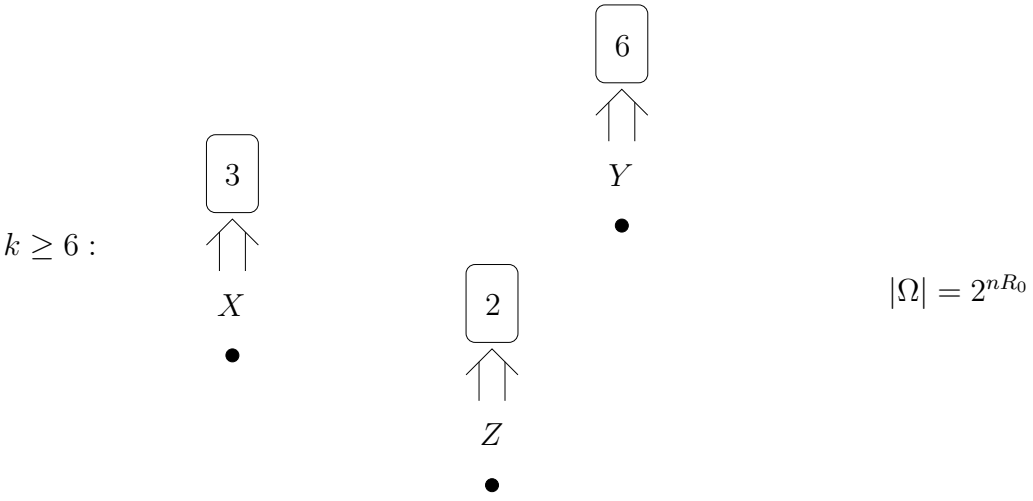


Figure 3.2: *Random task assignment with no communication.* A task from a set of tasks numbered $1, ..., k$ is to be assigned randomly but uniquely to each of the nodes X, Y, and Z without any communication between them. The rate of common randomness needed to accomplish this is roughly $R_0 \geq 3 \log 3$ for large $k$.

*Theorem 13 tells us which values of* $R_0$ *will result in* $(R_0, \hat{p}(x, y, z)) \in \underline{\mathcal{C}}$*. We must optimize over distributions of an auxiliary random variable* $U$*. Two things*

*come into play to make this optimization manageable: The variables $X$, $Y$, and $Z$ are all conditionally independent given $U$; and the distribution $\hat{p}$ has sparsity. For any particular value of $U$, the conditional supports of $X$, $Y$, and $Z$ must be disjoint. Therefore,*

$$\begin{aligned} I(X,Y,Z;U) &= H(X,Y,Z) - H(X,Y,Z|U) \\ &= H(X,Y,Z) - \mathbf{E}\left[H(X,Y,Z|U=u)\right] \\ &\geq H(X,Y,Z) - \mathbf{E}\left[\log(k_{1,U}k_{2,U}k_{3,U})\right], \end{aligned}$$

*where $k_{1,U}$, $k_{2,U}$, and $k_{3,U}$ are integers that sum to $k$ for all $U$. Therefore, we maximize $\log(k_{1,U}k_{2,U}k_{3,U})$ by letting the three integers be as close to equal as possible. Furthermore, it is straightforward to find a joint distribution that meets this inequality with equality.*

*If $k$, the number of tasks, is divisible by three, then we see that $(R_0, \hat{p}(x,y,z)) \in \underline{\mathcal{C}}$ for values of $R_0 > 3\log 3 - \log(\frac{k}{k-1}) - \log(\frac{k}{k-2})$. No matter how large $k$ is, the required rate never exceeds $R_0 > 3\log 3$.*

## 3.3 Two nodes

What is the intrinsic connection between correlated random variables? How much interaction is necessary to create correlation?

Many fruitful efforts have been made to quantify correlation between two random variables. Each quantity is justified by the operational questions that it answers. Covariance dictates the mean squared error in linear estimation. Shannon's mutual information is the descriptive savings from side information in lossless source coding and the additional growth rate of wealth due to side information in investing. Gács and Körner's common information [30] is the number of common random bits that can be extracted from correlated random variables. It is less than mutual information. Wyner's common information [29] is the number of common random bits needed to generate correlated random variables and is greater than mutual information.

We can revisit the two-node network from Section 2.2.1 and ask what communication rate is needed for strong coordination. This inquiry provides a fresh look at two of these quantities — mutual information and Wyner's common information. Both are found as extreme points of the strong coordination capacity region.

In this two-node network depicted in Figure 3.3, the action at node X is specified by nature according to $p_0(x)$, and a message is sent from node X to node Y at rate $R$. Common randomness is also available to both nodes at rate $R_0$. The common randomness is independent of the action $X$. In this setting we can think of strong coordination as synthesizing a memoryless channel from $X$ to $Y$. Indeed, if a memoryless channel $p(y|x)$ was available with input at node X and output at node Y, then strong coordination with respect to $p_0(x)p(y|x)$ would be achievable in a straightforward manner.

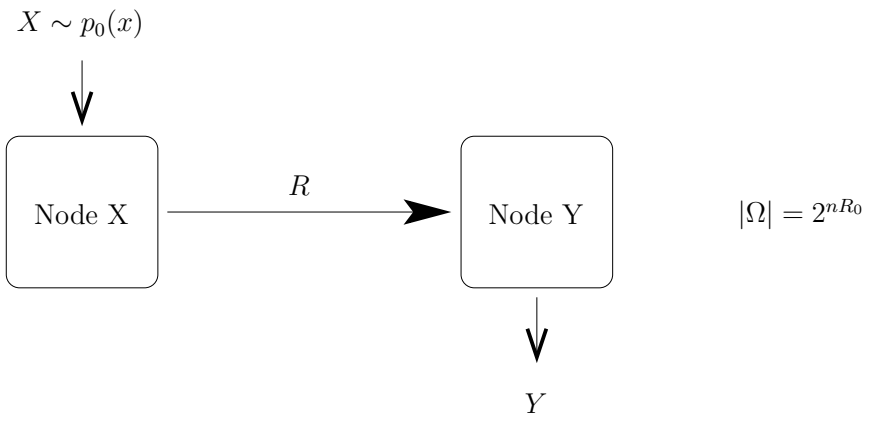


Figure 3.3: *Two nodes.* The action at node X is specified by nature according to $p_0(x)$, and a message is sent from node X to node Y at rate $R$. Common randomness is also available to both nodes at rate $R_0$. The common randomness is independent of the action $X$. The strong coordination capacity region $\underline{\mathcal{C}}_{p_0}$ depends on the amount of common randomness available. With no common randomness, $\underline{\mathcal{C}}_{p_0}$ contains all rate-coordination pairs where the rate is greater than the common information between $X$ and $Y$. With enough common randomness, $\underline{\mathcal{C}}_{p_0}$ contains all rate-coordination pairs where the rate is greater than the mutual information between $X$ and $Y$.

Synthesizing a memoryless channel is a form of random number generation. The variables $X^n$ come from an external source and $Y^n$ are generated to be correlated with $X^n$. The channel synthesis is successful if the total variation between the resulting

distribution of $(X^n, Y^n)$ and the i.i.d. distribution that would result from passing $X^n$ through a memoryless channel is small. This total variation requirement means that any hypothesis test that a statistician comes up with to determine whether $X^n$ was passed through a real memoryless channel or the channel synthesizer will be virtually useless.

Wyner's result implies that in order to generate $X^n$ and $Y^n$ separately as an i.i.d. source pair they must share bits at a rate of at least the common information $C(X;Y)$ of the joint distribution.[1] In the two-node network of Figure 3.3, the description of $X^n$ at rate $R$ serves the role of shared bits. However, the "reverse Shannon theorem" of Bennett et. al. [13] suggests that a description rate $R > I(X;Y)$ is all that is needed to successfully synthesize a channel. Does this entail a contradiction?

The resolution of Wyner's work and the work of Bennett et. al. is in the common randomness. Even though the common randomness at rate $R_0$ is independent of the action sequence $X^n$, it still assists in establishing a connection between the action sequences $X^n$ and $Y^n$ and opens the way for description rates smaller than the common information $C(X;Y)$.

We give the full characterization of the strong coordination capacity region $\underline{\mathcal{C}}_{p_0}$ for this two-node network[2], involving a tradeoff between the description rate $R$ and the rate of common randomness $R_0$, which indeed confirms the two extreme cases: If the encoder and decoder are provided with enough common randomness, sending $I(X;Y)$ bits per action suffices. On the other hand, in the absence of common randomness one must spend $C(X;Y)$ bits per action.

The $(2^{nR}, n)$ coordination codes consist of a non-deterministic encoding function,

$$i \;:\; \mathcal{X}^n \times \{1, ..., 2^{nR_0}\} \longrightarrow \{1, ..., 2^{nR}\}.$$

and a non-deterministic decoding function,

$$y^n \;:\; \{1, ..., 2^{nR}\} \times \{1, ..., 2^{nR_0}\} \longrightarrow \mathcal{Y}^n.$$

[1] To achieve strong coordination with a rate as low as the common information one must change Wyner's relative entropy requirement in [29] to a total variation requirement as used in this work.

[2] This result was independently discovered by Bennett et. al. [31].

Both functions can use private randomization to probabilistically map the arguments onto the range of the function. That is, the encoding function $i(x^n, \omega)$ behaves according to a conditional probability mass function $p(i|x^n, \omega)$, and the decoding function $y^n(i, \omega)$ behaves according to a conditional probability mass function $p(y^n|i, \omega)$.

The actions $X^n$ are chosen by nature i.i.d. according to $p_0(x)$, and the actions $Y^n$ are constructed by implementing the non-deterministic coordination code as

$$Y^n \;=\; y^n(i(X^n, \omega), \omega).$$

Let us define two quantities before stating the result. The first is Wyner's common information $C(X;Y)$ [29], which turns out to be the communication rate requirement for strong coordination in the two-node network when no common randomness is available:

$$C(X;Y) \;\triangleq\; \min_{U\,:\,X-U-Y} I(X,Y;U),$$

where the notation $X - U - Y$ represents a Markov chain from $X$ to $U$ to $Y$. The second quantity we call *necessary conditional entropy* $H(Y\dagger X)$, which we will show to be the amount of common randomness needed to maximize the strong coordination capacity region in the two-node network:

$$H(Y\dagger X) \;\triangleq\; \min_{f\,:\,X-f(Y)-Y} H(f(Y)|X).$$

**Theorem 14** (Strong coordination capacity region)**.** *The strong coordination capacity region $\underline{\mathcal{C}}_{p_0}$ for the two-node network of Figure 3.3 is given by*

$$\underline{\mathcal{C}}_{p_0} \;=\; \left\{ (R_0, R, p(y|x)) \;:\; \begin{array}{l} \exists p(u|x,y) \text{ such that} \\ |\mathcal{U}| \le |\mathcal{X}||\mathcal{Y}| + 1, \\ R \ge I(X;U), \\ R_0 + R \ge I(X,Y;U). \end{array} \right\}.$$

*Furthermore, the rate pair $(R_0, I(X;Y))$ is in the strong rate-coordination region $\underline{\mathcal{R}}_{p_0}$*

*if and only if the rate of common randomness* $R_0 \geq H(Y\dagger X)$.

### 3.3.1 Insights

Two extreme points of the strong coordination capacity region $\underline{\mathcal{C}}_{p_0}$ are manifested directly in the inequalities of Theorem 14. If $R_0 = 0$, the second inequality dominates. Thus, the minimum communication rate $R$ is Wyner's common information $C(X;Y)$. At the other extreme, using the data processing inequality on the first inequality yields $R \geq I(X;Y)$ no matter how much common randomness is available, and this is achieved when $R_0 \geq H(Y\dagger X)$. We show in Section 3.5 that in fact $H(Y\dagger X)$ is the minimum rate of common randomness for which $R = I(X;Y)$ is in the strong rate-coordination region $\underline{\mathcal{R}}_{p_0}$. For many joint distributions, the necessary conditional entropy $H(Y\dagger X)$ will simply equal the conditional entropy $H(Y|X)$.

This theorem is consistent with Conjecture 1—with enough common randomness, the strong coordination capacity region $\underline{\mathcal{C}}_{p_0}$ is the same as the coordination capacity region $\mathcal{C}_{p_0}$ found in Section 2.2.1.

The proof of Theorem 14 is found in Section 3.5.

**Example 7** (Task assignment)**.** *Consider again a task assignment setting similar to Example 6, where tasks are numbered* $1, ..., k$ *and are to be assigned randomly to the two nodes* $X$ *and* $Y$ *without duplication. The action* $X$ *is supplied by nature, uniformly at random, and the desired distribution* $\hat{p}(y|x)$ *for the action* $Y$ *is the uniform distribution over all tasks not equal to* $X$*. Figure 3.4 illustrates a valid outcome of the task assignments.*

*To apply Theorem 14 we must evaluate the three quantities* $I(X;Y)$*,* $C(X;Y)$*, and* $H(Y\dagger X)$*. For the joint distribution* $\hat{p}_0(x)\hat{p}(y|x)$*, the necessary conditional entropy* $H(Y\dagger X)$ *is exactly the conditional entropy* $H(Y|X)$*. The computation of the common information* $C(X;Y)$ *follows the same steps as the derivation found in Example 6.*

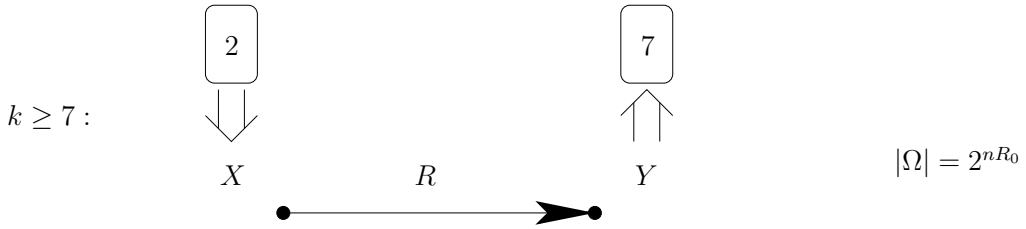


Figure 3.4: *Task assignment in the two-node network.* A task from a set of tasks numbered $1, ..., k$ is to be assigned randomly but uniquely to each of the nodes X and Y in the two-node network. The task assignment for X is given by nature. Common randomness at rate $R_0$ is available to both nodes, and a message is sent from node X to node Y at rate $R$. When no common randomness is available, the required communication rate is $R \geq 2 - \log(\frac{k}{k-1})$ bits (for even $k$). At the other extreme, if the rate of common randomness is greater than $\log(k-1)$, then $R \geq \log(\frac{k}{k-1})$ suffices.

*Let $\lceil k \rceil$ take the value of $k$ rounded up to the nearest even number.*

$$\begin{aligned} I(X;Y) &= \log\left(\frac{k}{k-1}\right), \\ C(X;Y) &= 2 \text{ bits } - \log\left(\frac{\lceil k \rceil}{\lceil k \rceil - 1}\right), \\ H(Y\dagger X) &= \log(k-1). \end{aligned}$$

*Without common randomness, we find that the communication rate $R \geq 2$ bits $- \log\left(\frac{\lceil k \rceil}{\lceil k \rceil - 1}\right)$ is necessary to strongly achieve $\hat{p}_0(x)\hat{p}(y|x)$. The strong coordination capacity region $\underline{\mathcal{C}}_{\hat{p}_0}$ expands as the rate of common randomness $R_0$ increases. Additional common randomness is no longer useful when $R_0 > \log(k-1)$. With this amount of common randomness, only the communication rate $R \geq \log(\frac{k}{k-1})$ is necessary to strongly achieve $\hat{p}_0(x)\hat{p}(y|x)$.*

**Example 8** (Binary erasure channel)**.** *For a Bernoulli-half source $X$, we demonstrate the strong rate-coordination region $\underline{\mathcal{R}}_{p_0}$ associated with synthesizing the binary erasure channel. The action $Y$ is an erasure with probability $P_e$ and is equal to $X$ otherwise. The distributions $p(x)p(u|x)p(y|u)$ that produce the boundary of the strong coordination capacity region are formed by cascading two binary erasure channels as*

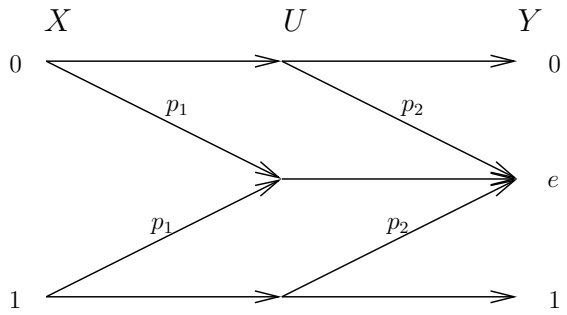


Figure 3.5: *Binary erasure channel synthesis.* The Markov chains $X - U - Y$ that give the boundary of the strong rate-coordination region for synthesizing the binary erasure channel from $X$ to $Y$ with a Bernoulli-half source $X$ are formed by cascading two erasure channels.

*shown in Figure 3.5, where*

$$\begin{aligned} p_2 &\in \left[0, \min\left\{\frac{1}{2}, P_e\right\}\right], \\ p_1 &= 1 - \frac{1 - P_e}{1 - p_2}. \end{aligned}$$

*The mutual information terms in Theorem 14 become*

$$\begin{aligned} I(X;U) &= 1 - p_1, \\ I(X,Y;U) &= H(P_e) + (1 - p_1)(1 - H(p_2)), \end{aligned}$$

*where $H$ is the binary entropy function.*

*Figure 3.6 shows the boundary of the strong rate-coordination region for erasure probability $P_e = 0.75$. The required description rate $R$ varies from $C(X;Y) = H(0.75) = 0.811$ bits to $I(X;Y) = 0.25$ bits as the rate of common randomness runs between 0 and $H(Y \dagger X) = H(0.75) = 0.811$ bits.*

## 3.4 Game theory

Consider a zero-sum repeated game between two teams. Team A consists of two players who on the $i$th iteration take actions $X_i \in \mathcal{X}$ and $Y_i \in \mathcal{Y}$. The opponents on

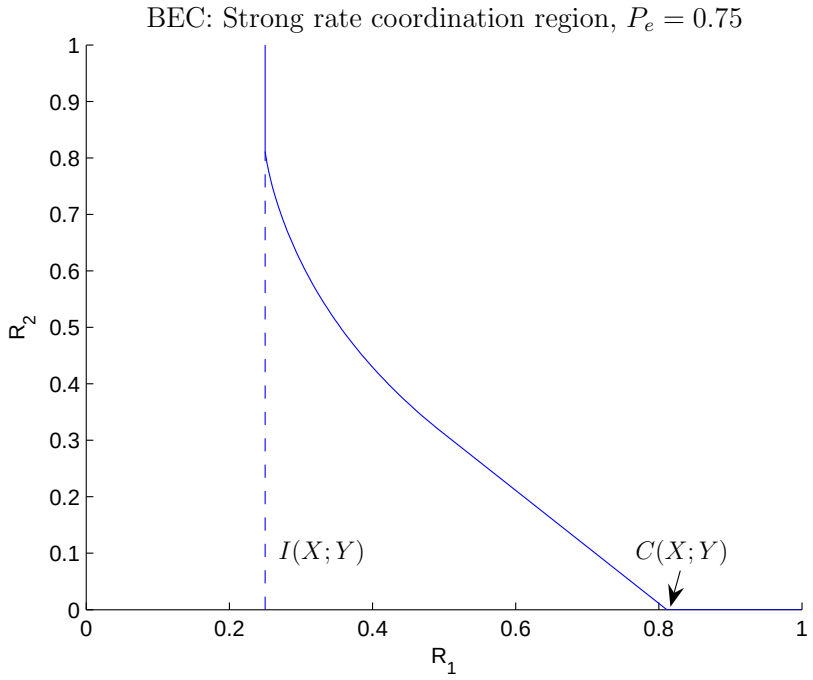


Figure 3.6: *General strong rate-coordination region for two nodes.* Boundary of the strong rate-coordination region for a binary erasure channel with erasure probability $P_e = 0.75$ and a Bernoulli-half input. In this figure, $R_1$ is the description rate and $R_2$ is the rate of common randomness. Without common randomness, a description rate of $C(X;Y)$ is required to simulate the channel. With unlimited common randomness, a description rate of $I(X;Y)$ suffices.

team B take combined action $Z_i \in \mathcal{Z}$. All action spaces $\mathcal{X}, \mathcal{Y}$, and $\mathcal{Z}$ are finite. The payoff for team A at each iteration is a time-invariant finite function $\Pi(X_i, Y_i, Z_i)$ and is the loss for team B. Each team wishes to maximize its time-averaged expected payoff.

Assume that team A plays conservatively, attempting to maximize the expected payoff for the worst-case actions of team B. The payoff at the $i$th iteration is

$$\Theta_i \triangleq \min_{z \in \mathcal{Z}} \mathbf{E}\left[\Pi(X_i, Y_i, z) | X^{i-1}, Y^{i-1}\right]. \tag{3.1}$$

Clearly, (3.1) could be maximized by finding an optimal mixed strategy $p^*(x, y)$ that maximizes $\min_{z \in \mathcal{Z}} \mathbb{E}\left[\Pi(X, Y, z)\right]$ and choosing independent actions each iteration. This would correspond to the minimax strategy. However, now we introduce a new constraint: The players on team A have a limited secure channel of communication. Player 1, who chooses the actions $X^n$, communicates at rate $R$ to Player 2, who chooses $Y^n$.

Let $I$ be the message passed from Player 1 to Player 2. We say a rate $R$ is achievable for payoff $\Theta$ if there exists a sequence of protocols described by random variable triples $(X^n, Y^n, I)$ that each form a Markov chain [3] $X^n - I - Y^n$ and such that $|\mathcal{I}| \leq 2^{nR}$ and

$$\lim_{n \to \infty} \mathbf{E}\left[\frac{1}{n} \sum_{i=1}^{n} \Theta_i\right] > \Theta. \tag{3.2}$$

Let $\mathcal{G}$ be the closure of the set of achievable pairs $(R, \Theta)$. We will show that optimality is obtained by producing strongly coordinated actions with respect to the appropriate joint distributions. Define,

$$\mathcal{G}_0 = \left\{ (R, \Theta) : \begin{array}{l} \exists p(x, y) \text{ such that} \\ R \geq C(X; Y), \\ \min_{z \in \mathcal{Z}} \mathbf{E}\, \Pi(X, Y, z) \geq \Theta. \end{array} \right\}.$$

[3] This Markov chain requirement can be relaxed to the more physically relevant requirement that $X_q - (I, X^{q-1}, Y^{q-1}) - Y_q$ forms a Markov chain for each $q$.

**Theorem 15** (Optimal cooperative play)**.**

$$\mathcal{G} \;=\; \textbf{ConvexHull}(\mathcal{G}_0).$$

## 3.5 Proofs

### Properties of total variation

Here we list a number of properties of total variation that are useful resources for both the achievability and converse proofs of this chapter.

**Lemma 16** (Total variation: marginal)**.** *For any two joint distributions $p(a, b)$ and $\hat{p}(a, b)$ on the same probability space, the total variation distance between them can only be reduced when attention is restricted to marginal distributions $p(a)$ and $\hat{p}(a)$. That is,*

$$\|p(a) - \hat{p}(a)\|_{TV} \;\leq\; \|p(a, b) - \hat{p}(a, b)\|_{TV}.$$

*Proof.*

$$\begin{aligned}
\|p(a) - \hat{p}(a)\|_{TV} &= \sum_a |p(a) - \hat{p}(a)| \\
&= \sum_a \left| \sum_b p(a, b) - \sum_b \hat{p}(a, b) \right| \\
&\leq \sum_a \sum_b |p(a, b) - \hat{p}(a, b)| \\
&= \|p(a, b) - \hat{p}(a, b)\|_{TV}.
\end{aligned}$$

□

**Lemma 17** (Total variation: same channel)**.** *When two random variables are passed through the same channel, the total variation between the resulting input-output joint*

*distributions is the same as the total variation between the input distributions. Specifically,*

$$\|p(a) - \hat{p}(a)\|_{TV} \quad = \quad \|p(a)p(b|a) - \hat{p}(a)p(b|a)\|_{TV}.$$

*Proof.*

$$\begin{aligned}
\|p(a) - \hat{p}(a)\|_{TV} &= \sum_a |p(a) - \hat{p}(a)| \\
&= \sum_a \left( \sum_b p(b|a) \right) |p(a) - \hat{p}(a)| \\
&= \sum_a \sum_b |p(b|a)(p(a) - \hat{p}(a))| \\
&= \sum_a \sum_b |p(a)p(b|a) - \hat{p}(a)p(b|a)| \\
&= \|p(a)p(b|a) - \hat{p}(a)p(b|a)\|_{TV}.
\end{aligned}$$

□

Let $Q \in \{1, ..., n\}$ be a random time index, independent of the random sequence $X^n$, and let the variable $X_Q$ be defined along the lines of time mixing in Section 2.4.2.

**Lemma 18** (Total variation: random time)**.** *The total variation between the distributions of two random sequences is an upper bound on the total variation between the distributions of the variables in the sequence at a random time index. That is,*

$$\|p_{X_Q}(x) - \hat{p}_{X_Q}(x)\|_{TV} \quad \leq \quad \|p(x^n) - \hat{p}(x^n)\|_{TV}.$$

*Proof.*

$$\begin{aligned}
\|p_{X_Q}(x) - \hat{p}_{X_Q}(x)\|_{TV} &= \left\| \sum_q p(q)(p_{X_q}(x) - \hat{p}_{X_q}(x)) \right\|_{TV} \\
&\leq \sum_{x,q} p(q)|p_{X_q}(x) - \hat{p}_{X_q}(x)| \\
&= \sum_q p(q)\|p_{X_q}(x) - \hat{p}_{X_q}(x)\|_{TV} \\
&\overset{a}{\leq} \sum_q p(q)\|p(x^n) - \hat{p}(x^n)\|_{TV} \\
&= \|p(x^n) - \hat{p}(x^n)\|_{TV},
\end{aligned}$$

where inequality $a$ comes from applying Lemma 16. □

### 3.5.1 Achievability

The following phenomenon was noticed both by Wyner [29] and by Han and Verdú [12]. Consider a memoryless channel $p(v|u)$. A channel input with distribution $p(u)$ induces an output with distribution $p(v) = \sum_u p(u)p(v|u)$. If the inputs are i.i.d. then the outputs are i.i.d. as well. Now suppose that instead a channel input sequence $U^n$ is chosen uniformly at random from a set $\mathbb{C}$ of $2^{nR}$ deterministic sequences. If $R > I(U;V)$ then the set $\mathbb{C}$ can be chosen so that the output distribution is arbitrarily close in total variation to the i.i.d. distribution $\prod_{i=1}^n p(v_i)$ for large enough $n$. This is the claim made by Lemma 19, and we provide our own simple proof.

#### Cloud mixing

We adopt the use of jointly typical sets $\mathcal{T}_\epsilon^{(n)}$ from Section 2.4.

Consider a memoryless channel from $U$ to $V$ defined by the conditional probability distribution $p(v|u)$. For an input distribution $p(u)$, we say that the desired output distribution is i.i.d. with respect to the distribution induced by the channel $p(v) =$

$\sum_u p(u)p(v|u)$. That is,

$$\begin{aligned} p(v^n) &= \prod_{i=1}^{n} p(v_i) \\ &= \prod_{i=1}^{n} \sum_u p(u)p(v_i|u). \end{aligned}$$

Consider an *input codebook* $\mathbb{C}$ of $2^{nR}$ sequences $u^n(1), ..., u^n(2^{nR})$, each constructed randomly and independently according to the i.i.d. input distribution $U^n(i) \sim \prod_{i=1}^n p(u_i)$. Also, let $M$ be a random selection of one of the sequences $u^n(M)$ in the codebook $\mathbb{C}$, uniformly at random.

The combination of selecting a random sequence from the input codebook $\mathbb{C}$ and using it as the input to the memoryless channel $p(v|u)$ results in an *induced output distribution*,

$$\hat{P}(v^n) = 2^{-nR} \sum_m p(v^n|U^n(m)).$$

where $p(v^n|u^n) = \prod_{i=1}^n p(v_i|u_i)$. We use capital letters $\hat{P}$ and $U^n(m)$ to indicate that they are random due to the random generation of the input codebook $\mathbb{C}$. Therefore, $\hat{P}(v^n)$ is a random probability mass function on $v^n$ sequences that depends on the particular codebook. The intention of Lemma 19 is to show that if $R$ is large enough, $\hat{P}(v^n)$ will be close in total variation to $p(v^n)$ for most codebooks.

**Lemma 19** (Cloud mixing)**.** *For a memoryless channel defined by $p(v|u)$, an input distribution $p(u)$, and an input codebook of size $2^{nR}$, the total variation distance between the induced output distribution and the desired output distribution goes to zero as $n$ goes to infinity if $R > I(U;V)$. That is,*

$$R > I(U;V) \implies \lim_{n\to\infty} \mathbf{E} \left\| \hat{P}(v^n) - p(v^n) \right\|_{TV} = 0.$$

*Proof.* We need to separate the jointly typical sequences from the rest. We start by

separating $\hat{P}(v^n)$ into two parts,

$$\begin{aligned}
\hat{P}_1(v^n) &\triangleq 2^{-nR} \sum_m p(v^n|U^n(m)) \mathbf{1}\left((U^n(m), v^n) \in \mathcal{T}_\epsilon^{(n)}\right), \\
\hat{P}_2(v^n) &\triangleq 2^{-nR} \sum_m p(v^n|U^n(m)) \mathbf{1}\left((U^n(m), v^n) \notin \mathcal{T}_\epsilon^{(n)}\right).
\end{aligned}$$

Notice a couple of observations. First, $\hat{P}(v^n) = \hat{P}_1(v^n) + \hat{P}_2(v^n)$. Second, the expected value of $\hat{P}(v^n)$ with respect to the randomly generated input codebook is the desired distribution $p(v^n)$. That is,

$$\begin{aligned}
\mathbf{E}\, \hat{P}(v^n) &\stackrel{a}{=} 2^{-nR} \sum_m \mathbf{E}\, p(v^n|U^n(m)) \\
&\stackrel{b}{=} \mathbf{E}\, p(v^n|U^n(1)) \\
&= \sum_{u^n} \prod_{i=1}^n p(u_i) \prod_{j=1}^n p(v_j|u_j) \\
&= p(v^n),
\end{aligned}$$

where equality $a$ is an application of the linearity of expectation to the definition of $\hat{P}(v^n)$, and equality $b$ is due to the i.i.d. nature of the input codebook $\mathbb{C}$.

We separate the total variation distance $\mathbf{E}\left\|\hat{P}(v^n) - p(v^n)\right\|_{TV}$ into three parts:

$$\begin{aligned}
\mathbf{E}\left\|\hat{P}(v^n) - p(v^n)\right\|_{TV} &= \mathbf{E}\left\|\hat{P}(v^n) - \mathbf{E}\,\hat{P}(v^n)\right\|_{TV} \\
&= \sum_{v^n} \mathbf{E}\left|\hat{P}(v^n) - \mathbf{E}\,\hat{P}(v^n)\right| \\
&= \sum_{v^n \in \mathcal{T}_\epsilon^{(n)}} \mathbf{E}\left|\hat{P}(v^n) - \mathbf{E}\,\hat{P}(v^n)\right| \\
&\quad + \sum_{v^n \notin \mathcal{T}_\epsilon^{(n)}} \mathbf{E}\left|\hat{P}(v^n) - \mathbf{E}\,\hat{P}(v^n)\right| \\
&\overset{a}{\leq} \sum_{v^n \in \mathcal{T}_\epsilon^{(n)}} \mathbf{E}\left|\hat{P}_1(v^n) - \mathbf{E}\,\hat{P}_1(v^n)\right| \\
&\quad + \sum_{v^n} \mathbf{E}\left|\hat{P}_2(v^n) - \mathbf{E}\,\hat{P}_2(v^n)\right| \\
&\quad + \sum_{v^n \notin \mathcal{T}_\epsilon^{(n)}} \mathbf{E}\left|\hat{P}(v^n) - \mathbf{E}\,\hat{P}(v^n)\right|, \qquad (3.3)
\end{aligned}$$

where inequality $a$ is due to the triangle inequality and adding more terms into the second sum.

The first sum in (3.3) is the interesting one to consider, so we save it for last. Both other sums are resolved by the A.E.P. For instance, the triangle inequality gives us,

$$\begin{aligned}
\sum_{v^n} \mathbf{E}\left|\hat{P}_2(v^n) - \mathbf{E}\,\hat{P}_2(v^n)\right| &\leq 2\sum_{v^n} \mathbf{E}\,\hat{P}_2(v^n) \\
&= 2\sum_{v^n} 2^{-nR} \sum_{m} \mathbf{E}\, p(v^n|U^n(m))\mathbf{1}\left((U^n(m), v^n) \notin \mathcal{T}_\epsilon^{(n)}\right) \\
&= 2\sum_{v^n} \mathbf{E}\, p(v^n|U^n(1))\mathbf{1}\left((U^n(1), v^n) \notin \mathcal{T}_\epsilon^{(n)}\right) \\
&= 2 \sum_{(u^n, v^n) \notin \mathcal{T}_\epsilon^{(n)}} p(u^n)p(v^n|u^n) \\
&= 2\left(1 - \Pr\left(\mathcal{T}_\epsilon^{(n)}\right)\right) \\
&\to 0 \text{ as } n \text{ goes to infinity,}
\end{aligned}$$

and

$$\begin{aligned}\sum_{v^n \notin \mathcal{T}_\epsilon^{(n)}} \mathbf{E}\left|\hat{P}(v^n) - \mathbf{E}\,\hat{P}(v^n)\right| &\leq 2\sum_{v^n \notin \mathcal{T}_\epsilon^{(n)}} \mathbf{E}\,\hat{P}(v^n) \\ &= 2\sum_{v^n \notin \mathcal{T}_\epsilon^{(n)}} p(v^n) \\ &\to 0 \text{ as } n \text{ goes to infinity.}\end{aligned}$$

The remaining term in (3.3) deals with only jointly typical sequences. We derive a uniform bound for all typical sequences $v^n \in \mathcal{T}_\epsilon^{(n)}$, starting with the Jensen inequality,

$$\begin{aligned}\mathbf{E}\left|\hat{P}_1(v^n) - \mathbf{E}\,\hat{P}_1(v^n)\right| &\leq \sqrt{\mathbf{E}\left(\hat{P}_1(v^n) - \mathbf{E}\,\hat{P}_1(v^n)\right)^2} \\ &= \sqrt{\mathbf{Var}\,\hat{P}_1(v^n)}\end{aligned}$$

Since the input codebook $\mathbb{C}$ is a collection of sequences generated i.i.d., the variance of $\hat{P}_1(v^n)$ breaks down nicely as

$$\begin{aligned}\mathbf{Var}\,\hat{P}_1(v^n) &\stackrel{a}{=} 2^{-nR}\mathbf{Var}\,\left(p(v^n|U^n(1))\mathbf{1}\left((U^n(1), v^n) \in \mathcal{T}_\epsilon^{(n)}\right)\right) \\ &\leq 2^{-nR}\mathbf{E}\left(p(v^n|U^n(1))\mathbf{1}\left((U^n(1), v^n) \in \mathcal{T}_\epsilon^{(n)}\right)\right)^2 \\ &= 2^{-nR}\sum_{u^n\,:\,(u^n,v^n)\in\mathcal{T}_\epsilon^{(n)}} p(u^n)p^2(v^n|u^n) \\ &\stackrel{b}{\leq} 2^{-n(R+H(V|U)-\delta_1(\epsilon))}\sum_{u^n} p(u^n)p(v^n|u^n) \\ &= 2^{-n(R+H(V|U)-\delta_1(\epsilon))}p(v^n) \\ &\leq 2^{-n(R+H(V|U)+H(V)-\delta_2(\epsilon))} \text{ for all typical } v^n \text{ sequences,}\end{aligned}$$

where $\delta_1(\epsilon))$ and $\delta_2(\epsilon))$ go to zero as $\epsilon$ goes to zero. Equality $a$ comes from the definition of $\hat{P}_1(v^n)$, and inequality $b$ is a result of continuity of the entropy function and the fact that $u^n$ and $v^n$ are $\epsilon$-jointly typical.

The conclusion with respect to the first term in (3.3) is,

$$\begin{aligned}\sum_{v^n \in \mathcal{T}_\epsilon^{(n)}} \mathbf{E}\left|\hat{P}_1(v^n) - \mathbf{E}\,\hat{P}_1(v^n)\right| &\leq \sum_{v^n \in \mathcal{T}_\epsilon^{(n)}} 2^{-\frac{n}{2}(R+H(V|U)+H(V)-\delta_2(\epsilon))} \\ &\leq 2^{-\frac{n}{2}(R+H(V|U)-H(V)-\delta_3(\epsilon))} \\ &= 2^{-\frac{n}{2}(R-I(U;V)-\delta_3(\epsilon))},\end{aligned}$$

where $\delta_3(\epsilon))$ goes to zero as $\epsilon$ goes to zero. Clearly, if $R > I(U;V)$ then a small enough choice of $\epsilon$ exists to make this term converge to zero as $n$ goes to infinity. □

**No communication - Theorem 13**

The network of Figure 3.1 with no communication generalizes Wyner's common information work [29] to three nodes. Figure 3.7 illustrates how to achieve the strong coordination capacity region $\underline{\mathcal{C}}$ of Theorem 13. Let each decoder simulate a memoryless channel from $U$ to $X$, $Y$, or $Z$, depending on the particular node. The common randomness $\omega$ is used to index a sequence $U^n(\omega)$ that is used as the inputs to the channels.

Notice that the action sequences $X^n$, $Y^n$, and $Z^n$ produced via these three separate channels are distributed the same as if they were generated as outputs of a single channel because $p(x,y,z|u) = p(x|u)p(y|u)p(z|u)$ according to the definition of $\underline{\mathcal{C}}$ in Theorem 13. Since $R > I(X,Y,Z;U)$ for points in the interior of $\underline{\mathcal{C}}$, Lemma 19 confirms that this scheme will achieve strong coordination.

**Two nodes - Theorem 14 - region**

Strong coordination is achieved in the two-node network of Figure 3.3 in much the same way as in the no-communication network. One key difference, however, is that the action $X$ is given randomly by nature in the two-node network.

We construct efficient coordination codes for the two-node network in a roundabout way. For analysis purposes, assume that we are allowed to generate the sequences $X^n$ at node X as well as $Y^n$ at node Y based on shared messages. The index $I$ and the index $J$ are both available to both nodes. We will derive legitimate

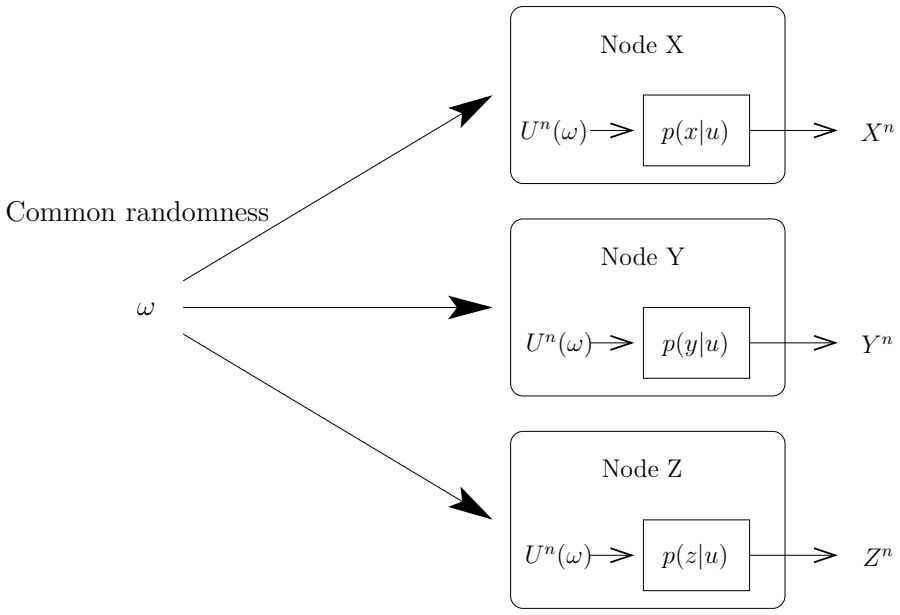


Figure 3.7: *Achievability for no-communication network.* The strong coordination capacity region $\underline{\mathcal{C}}$ of Theorem 13 is achieved in a network with no communication by using the common randomness to specify a sequence $U^n(\omega)$ that is then passed through a memoryless channel at each node using private randomness.

coordination codes from this hypothetical construction.

Any interior point of the strong coordination capacity region $\underline{\mathcal{C}}_{p_0}$ has an associated distribution $p(u)p(x|u)p(y|u)$ that satisfies the inequalities of Theorem 14. We generate the action sequences $X^n$ and $Y^n$ by letting the nodes X and Y simulate memoryless channels from $U$ to $X$ and $Y$, respectively, according to $p(x|u)$ and $p(y|u)$. Again, because of the Markovity $X - U - Y$, the separate channel simulations behave the same as if the nodes simulate the channel $p(x, y|u)$ jointly. The indices $I$ and $J$ are used to specify a sequence $U^n(I, J)$ uniformly at random from a randomly generated codebook of size $2^{n(R_0+R)}$ to be used as channel inputs by the nodes.

We identify the overall effect of this method for generating $X^n$ and $Y^n$ with the notation $\hat{P}(x^n|i, j)$ and $\hat{P}(y^n|i, j)$. Explicitly,

$$\begin{aligned} \hat{P}(x^n|i,j) &\triangleq p(x^n|U^n(i,j)), \\ \hat{P}(y^n|i,j) &\triangleq p(y^n|U^n(i,j)). \end{aligned}$$

These conditional distributions are random because the input codebook of $U^n(i,j)$ sequences is randomly generated. Furthermore, we define the *backward induced joint distribution*

$$\hat{P}(x^n, y^n, i, j) \triangleq 2^{-n(R_0+R)} \hat{P}(x^n|i,j)\hat{P}(y^n|i,j),$$

from which marginal and conditional distributions can be derived.

Lemma 19 allows us to draw two conclusions about this backward induced joint distribution. The fact that $R_0 + R > I(X,Y;U)$ tells us,

$$\lim_{n\to\infty} \mathbf{E} \left\| \hat{P}(x^n, y^n) - p(x^n, y^n) \right\|_{TV} = 0,$$

and $R > I(X;U)$ gives,

$$\begin{aligned}
\lim_{n\to\infty} \mathbf{E} \left\| \hat{P}(x^n, j) - p(x^n)p(j) \right\|_{TV} &= \lim_{n\to\infty} \sum_j p(j) \sum_{x^n} \mathbf{E} \left| \hat{P}(x^n|j) - p(x^n) \right| \\
&= \lim_{n\to\infty} \mathbf{E} \left\| \hat{P}(x^n|j=1) - p(x^n) \right\|_{TV} \\
&= 0.
\end{aligned}$$

The idea is that $R$ is large enough so that $J$ is roughly independent of $X^n$. For any particular value of $J$, there are enough sequences in the codebook due to the varying index $I$ that the dependence between $X^n$ and $J$ is washed away.

Finally, we specify the coordination codes for each $n$ for the two-node network by the conditional distributions derived from the backward induced joint distribution $\hat{P}(x^n, y^n, i, j)$. Given an input codebook $\mathbb{C}$, the encoder implements $\hat{P}(i|x^n, j)$ where it is well defined and otherwise sends the index 1. The decoder implements $\hat{P}(y^n|i,j)$.

Using these coordination codes, a new induced distribution $\widehat{P}$ will result. This distribution is given by

$$\widehat{P}(x^n, y^n, i, j) = p(x^n)p(j)\hat{P}(i|x^n, j)\hat{P}(y^n|i,j).$$

The induced joint distribution $\widehat{P}(x^n, y^n, i, j)$ and the backward induced joint distribution $\hat{P}(x^n, y^n, i, j)$ are both comprised of differing joint distributions on $X^n$ and $J$ passed through the same channel $\hat{P}(i|x^n, j)\hat{P}(y^n|i, j)$.

We conclude,

$$\begin{aligned}
\left\|\widehat{P}(x^n, y^n) - p(x^n, y^n)\right\|_{TV} &\overset{a}{\leq} \left\|\widehat{P}(x^n, y^n) - \hat{P}(x^n, y^n)\right\|_{TV} \\
&\quad + \left\|\hat{P}(x^n, y^n) - p(x^n, y^n)\right\|_{TV} \\
&\overset{b}{\leq} \left\|\widehat{P}(x^n, y^n, i, j) - \hat{P}(x^n, y^n, i, j)\right\|_{TV} \\
&\quad + \left\|\hat{P}(x^n, y^n) - p(x^n, y^n)\right\|_{TV} \\
&\overset{c}{=} \left\|p(x^n)p(j) - \hat{P}(x^n, j)\right\|_{TV} \\
&\quad + \left\|\hat{P}(x^n, y^n) - p(x^n, y^n)\right\|_{TV},
\end{aligned}$$

where $a$ comes from the triangle inequality, $b$ is an application of Lemma 16, and $c$ is a consequence of Lemma 17. The result is,

$$\lim_{n\to\infty} \mathbf{E} \left\|\widehat{P}(x^n, y^n) - p(x^n, y^n)\right\|_{TV} = 0,$$

so there exists a sequence of codebooks that sends total variation to zero as $n$ grows.

**Two nodes - Theorem 14 - extreme point**

In addition to characterizing the strong coordination capacity region for the two-node network, Theorem 14 states that $(R_0, I(X;Y))$ is in the strong rate-coordination region $\underline{\mathcal{R}}_{p_0}$ if $R_0 \geq H(Y\dagger X)$. This is a straightforward application of the definition of $H(Y\dagger X)$. We can verify this with the following choice of $U$:

$$U = \underset{f(Y)\,:\,X-f(Y)-Y}{\operatorname{argmin}} H(f(Y)|X).$$

Notice that this choice of $U$ separates $X$ and $Y$ into a Markov chain by definition. Also, the mutual information $I(X;U)$ is less than or equal to $I(X;Y)$, since $U$ is

a function of $Y$, thus satisfying the first rate inequality in $\underline{\mathcal{C}}_{p_0}$ as characterized by Theorem 14. The second inequality is satisfied because of the chain rule,

$$\begin{aligned} I(X,Y;U) &= I(X;U) + I(Y;U|X) \\ &= I(X;U) + H(Y\dagger X) \\ &\leq I(X;Y) + H(Y\dagger X). \end{aligned}$$

**Comments**

The efficient coordination codes presented in Section 3.5.1 have some resemblance to classical source coding in the rate-distortion sense. The difference here is that sequences in the codebook do not correspond to reconstruction sequences. Instead, randomization happens at both ends, both at the encoder via common randomness and at the decoder via private randomization.

### 3.5.2 Converse

**Mutual information bounds for nearly i.i.d. sequences**

If two finite-alphabet random variables have distributions that are close in total variation, then their entropies are also close (Theorem 16.3.2 of [26]). This leads to a couple of lemmas concerning a sequence of random variables $W^n$ with a distribution that is close to an i.i.d. distribution. First, the *total mutual information with the past* $\sum_{q=1}^{n} I(W_q; W^{q-1})$ is small. Second, the mutual information between an independent random time index $Q$ and the variable at that time index $W_Q$ is small.

**Lemma 20** (Information with the past)**.** *For any discrete random sequence $W^n \in \mathcal{W}^n$, if there exists a distribution $\hat{p}(w)$ on the alphabet $\mathcal{W}$ such that*

$$\left\| p(w^n) - \prod_{q=1}^{n} \hat{p}(w_q) \right\|_{TV} < \epsilon < 1/2,$$

*then the total mutual information with the past* $\sum_{q=1}^{n} I(W_q; W^{q-1})$ *is bounded by*

$$\sum_{q=1}^{n} I(W_q; W^{q-1}) \;\leq\; 2n\epsilon \left( \log |\mathcal{W}| + \log \frac{1}{\epsilon} \right).$$

*Proof.* First use Lemma 16 to claim that every symbol in $W^n$ must be close in distribution to $\hat{p}$. That is,

$$\left\| p_{W_q}(w) - \hat{p}(w) \right\|_{TV} \;<\; \epsilon \text{ for all } q \in \{1, ..., n\}.$$

Let $\hat{W}^n$ be distributed according to $\prod_{q=1}^{n} \hat{p}(w_q)$. Then by Theorem 17.3.3 of [26],

$$\left| H(W^n) - H(\hat{W}^n) \right| \;\leq\; \epsilon \log \left( \frac{|\mathcal{W}|^n}{\epsilon} \right), \tag{3.4}$$

$$\left| H(W_q) - H(\hat{W}_q) \right| \;\leq\; \epsilon \log \left( \frac{|\mathcal{W}|}{\epsilon} \right) \text{ for all } q \in \{1, ..., n\}. \tag{3.5}$$

Because $\hat{W}^n$ is an i.i.d. sequence, $H(\hat{W}^n) = \sum_{k=1}^{n} H(\hat{W}_k)$. Therefore,

$$\begin{aligned}
\sum_{q=1}^{n} I(W_q; W^{q-1}) &= \sum_{q=1}^{n} H(W_q) - H(W^n) \\
&= \sum_{q=1}^{n} \left( H(W_q) - H(\hat{W}_q) \right) + H(\hat{W}^n) - H(W^n) \\
&\leq n\epsilon \log \left( \frac{|\mathcal{W}|}{\epsilon} \right) + \epsilon \log \left( \frac{|\mathcal{W}|^n}{\epsilon} \right) \\
&= 2n\epsilon \log |\mathcal{W}| + (n+1)\epsilon \log \frac{1}{\epsilon} \\
&\leq 2n\epsilon \left( \log |\mathcal{W}| + \log \frac{1}{\epsilon} \right).
\end{aligned}$$

□

**Lemma 21** (Timing information)**.** *For any discrete random sequence* $W^n \in \mathcal{W}^n$*, if*

*there exists a distribution* $\hat{p}(w)$ *on the alphabet* $\mathcal{W}$ *such that*

$$\left\| p(w^n) - \prod_{q=1}^{n} \hat{p}(w_q) \right\|_{TV} \; < \; \epsilon \; < \; 1/2,$$

*then for a random time variable* $Q \in \{1, ..., n\}$ *independent of* $W^n$*,*

$$I(W_Q; Q) \;\; \leq \;\; 2\epsilon \left( \log |\mathcal{W}| + \log \frac{1}{\epsilon} \right).$$

*Proof.* Again, Lemma 16 implies that every symbol $W_q$ is close in distribution to $\hat{p}$,

$$\left\| p_{W_q}(w) - \hat{p}(w) \right\|_{TV} \;\; < \;\; \epsilon \text{ for all } q \in \{1, ..., n\}.$$

Also, again let $\hat{W}^n$ be distributed according to $\prod_{q=1}^{n} \hat{p}(w_q)$. Recall from Properties 1 of time mixing in Section 2.4.2 that $\hat{W}_Q$ has the same distribution as $\hat{W}_1$, namely $\hat{p}(w)$. Therefore, with the assistance of Theorem 16.3.2 of [26],

$$\begin{aligned}
\left| H(W_Q) - H(\hat{W}_Q) \right| &\overset{a}{\leq} \epsilon \log \left( \frac{|\mathcal{W}|^n}{\epsilon} \right), \\
\left| H(W_Q|Q) - H(\hat{W}_Q) \right| &= \left| \sum_q p(q) \left( H(W_q) - H(\hat{W}_q) \right) \right| \\
&\leq \sum_q p(q) \left| H(W_q) - H(\hat{W}_q) \right| \\
&\overset{b}{\leq} \sum_q p(q) \, \epsilon \log \left( \frac{|\mathcal{W}|}{\epsilon} \right) \\
&= \epsilon \log \left( \frac{|\mathcal{W}|}{\epsilon} \right),
\end{aligned}$$

where inequality $a$ uses Lemma 18, and inequality $b$ uses Lemma 16.

Finally, the triangle inequality gives,

$$\begin{aligned}
I(W_Q; Q) &= H(W_Q) - H(W_Q|Q) \\
&\leq 2\epsilon \left( \log |\mathcal{W}| + \log \frac{1}{\epsilon} \right).
\end{aligned}$$

□

In conjunction with Lemma 20 and Lemma 21, the following quantity will serve as a useful abbreviation:

$$g(\epsilon) \triangleq 2\epsilon \left( \log |\mathcal{X}| + \log |\mathcal{Y}| + \log |\mathcal{Z}| + \log \frac{1}{\epsilon} \right), \tag{3.6}$$

for $\epsilon > 0$. If the problem does not involve a random variable $Z$ then assume $|\mathcal{Z}| = 1$. Notice that $\lim_{\epsilon \to 0} g(\epsilon) = 0$.

**No communication - Theorem 13**

Let $\underline{\mathcal{C}}'$ represent the proposed strong coordination capacity region from Theorem 13, without any constraint on the cardinality $U$. We start with a rate-coordination pair $(R_0, p(x, y, z))$ that is known to be achievable for strong coordination and prove that it is in $\underline{\mathcal{C}}'$.

By the definition of achievability for strong coordination, there exists a sequence of coordination codes such that the induced distribution $p(x^n, y^n, z^n)$ is close to the i.i.d. desired distribution $p(x, y, z)$. That is,

$$\left\| p(x^n, y^n, z^n) - \prod_{i=1}^{n} p(x_i, y_i, z_i) \right\|_{TV} \longrightarrow 0.$$

For this sequence of coordination codes we can derive mutual information bounds. Let $\epsilon$ be an arbitrary (small) number greater than zero, and let $n$ be large enough that the above total variation is less than $\epsilon$. Identify the auxiliary variable $U$ as the pair $(\omega, Q)$, where $Q$ is a random time index uniformly distributed from 1 to $n$ (see time mixing in Section 2.4.2) and independent of $X^n$, $Y^n$, $Z^n$. Notice that $X_Q$, $Y_Q$,

and $Z_Q$ are conditionally independent given $\omega$ and $Q$.

$$\begin{aligned}
nR_0 &\geq H(\omega) \\
&\geq I(X^n, Y^n, Z^n; \omega) \\
&= \sum_{q=1}^{n} I(X_q, Y_q, Z_q; \omega | X^{q-1}, Y^{q-1}, Z^{q-1}) \\
&= \sum_{q=1}^{n} I(X_q, Y_q, Z_q; \omega, X^{q-1}, Y^{q-1}, Z^{q-1}) - \sum_{q=1}^{n} I(X_q, Y_q, Z_q; X^{q-1}, Y^{q-1}, Z^{q-1}) \\
&\overset{a}{\geq} \sum_{q=1}^{n} I(X_q, Y_q, Z_q; \omega, X^{q-1}, Y^{q-1}, Z^{q-1}) - ng(\epsilon) \\
&\geq \sum_{q=1}^{n} I(X_q, Y_q, Z_q; \omega) - ng(\epsilon) \\
&= nI(X_Q, Y_Q, Z_Q; \omega | Q) - ng(\epsilon) \\
&= nI(X_Q, Y_Q, Z_Q; \omega, Q) - nI(X_Q, Y_Q, Z_Q; Q) - ng(\epsilon) \\
&\overset{b}{\geq} nI(X_Q, Y_Q, Z_Q; \omega, Q) - 2ng(\epsilon). \\
&= nI(X_Q, Y_Q, Z_Q; U) - 2ng(\epsilon),
\end{aligned}$$

where $g$ is defined in (3.6), inequality $a$ is an application of Lemma 20, and inequality $b$ is an application of Lemma 21.

To summarize,

$$R_0 + 2g(\epsilon) \quad \geq \quad I(X_Q, Y_Q, Z_Q; U),$$

where $X_Q$, $Y_Q$, and $Z_Q$ are conditionally independent given $U$. This implies,

$$\big((R_0 + 2g(\epsilon), p_{X_Q, Y_Q, Z_Q}(x, y, z)\big) \quad \in \quad \underline{\mathcal{C}}'.$$

Finally, the function $g(\epsilon)$ is arbitrarily small for small enough $\epsilon$, and Lemma 18 states,

$$\big\| p(x, y, z) - p_{X_Q, Y_Q, Z_Q}(x, y, z) \big\|_{TV} \quad < \quad \epsilon.$$

Therefore, the rate-coordination pair $(R_0, p(x,y,z))$ is arbitrarily close to the closed set $\underline{\mathcal{C}}'$. We conclude,

$$(R_0, p(x,y,z)) \;\in\; \underline{\mathcal{C}}'.$$

It remains to bound the cardinality of $U$. We can again use the standard method of [27]. The variable $U$ should have $|\mathcal{X}||\mathcal{Y}||\mathcal{Z}| - 1$ elements to preserve the joint distribution $p(x,y,z)$, which in turn preserves $H(X,Y,Z)$, and one more element to preserve $H(X,Y,Z|U)$.

**Two nodes - Theorem 14 - region**

Let $\underline{\mathcal{C}}'_{p_0}$ represent the proposed strong coordination capacity region from Theorem 14, without any constraint on the cardinality $U$. We start with a rate-coordination triple $(R_0, R, p(y|z))$ that is known to be achievable for strong coordination and prove that it is in $\underline{\mathcal{C}}'_{p_0}$.

By the definition of achievability for strong coordination, there exists a sequence of coordination codes such that the induced distribution $p(x^n, y^n)$ is close to the i.i.d. desired distribution $p_0(x)p(y|x)$. That is,

$$\left\| p(x^n, y^n) - \prod_{i=1}^{n} p_0(x_i)p(y_i|x_i) \right\|_{TV} \;\longrightarrow\; 0.$$

For this sequence of coordination codes we can derive mutual information bounds. Let $\epsilon$ be an arbitrary (small) number greater than zero, and let $n$ be large enough that the above total variation is less than $\epsilon$. Recall that $I$ is the message from node X to node Y. Identify the auxiliary variable $U$ as the triple $(I, \omega, Q)$, where $Q$ is a random time index uniformly distributed from 1 to $n$ (see time mixing in Section 2.4.2) and independent of $X^n$ and $Y^n$. Notice that $X_Q$ and $Y_Q$ are conditionally independent

given $I$, $\omega$, and $Q$.

$$\begin{aligned}
nR &\geq H(I) \\
&\geq H(I|\omega) \\
&\geq I(X^n; I|\omega) \\
&\overset{a}{=} I(X^n; I, \omega) \\
&= \sum_{q=1}^{n} I(X_q; I, \omega | X^{q-1}) \\
&\overset{b}{=} \sum_{q=1}^{n} I(X_q; I, \omega, X^{q-1}) \\
&\geq \sum_{q=1}^{n} I(X_q; I, \omega) \\
&= nI(X_Q; I, \omega | Q) \\
&\overset{c}{=} nI(X_Q; I, \omega, Q) \\
&= nI(X_Q; U),
\end{aligned}$$

where equality $a$ arises from the independence between the common randomness $\omega$ and the action sequence $X^n$ given by nature, and equalities $b$ and $c$ are consequences of the i.i.d. distribution of $X^n$ (see Property 1 of time mixing in Section 2.4.2).

Furthermore,

$$\begin{aligned}
nR_0 + nR &\geq H(I,\omega) \\
&\geq I(X^n, Y^n; I, \omega) \\
&= \sum_{q=1}^{n} I(X_q, Y_q; I, \omega | X^{q-1}, Y^{q-1}) \\
&= \sum_{q=1}^{n} I(X_q, Y_q; I, \omega, X^{q-1}, Y^{q-1}) - \sum_{q=1}^{n} I(X_q, Y_q; X^{q-1}, Y^{q-1}) \\
&\overset{a}{\geq} \sum_{q=1}^{n} I(X_q, Y_q; I, \omega, X^{q-1}, Y^{q-1}) - ng(\epsilon) \\
&\geq \sum_{q=1}^{n} I(X_q, Y_q; I, \omega) - ng(\epsilon) \\
&= nI(X_Q, Y_Q; I, \omega | Q) \\
&= nI(X_Q, Y_Q; I, \omega | Q) - ng(\epsilon) \\
&= nI(X_Q, Y_Q; I, \omega, Q) - nI(X_Q, Y_Q; Q) - ng(\epsilon) \\
&\overset{b}{\geq} nI(X_Q, Y_Q; I, \omega, Q) - 2ng(\epsilon) \\
&= nI(X_Q, Y_Q; U) - 2ng(\epsilon),
\end{aligned}$$

where $g$ is defined in (3.6), inequality $a$ is a result of Lemma 20, and inequality $b$ is a result of Lemma 21.

To summarize,

$$\begin{aligned}
R &\geq I(X_Q; U), \\
R_0 + R + 2g(\epsilon) &\geq I(X_Q, Y_Q; U),
\end{aligned}$$

where $X_Q$ and $Y_Q$ are conditionally independent given $U$, and $X_Q$ is distributed according to $p_0(x)$. This implies,

$$\big((R_0 + 2g(\epsilon), R, p_{Y_Q|X_Q}(y|x)\big) \in \underline{\mathcal{C}}'_{p_0}.$$

Finally, the function $g(\epsilon)$ is arbitrarily small for small enough $\epsilon$, and Lemma 18 states,

$$\left\| p_0(x)p(y|x) - p_{X_Q,Y_Q}(x,y) \right\|_{TV} \;<\; \epsilon.$$

Without loss of generality, assume $\min_x p_0(x) = c > 0$. Then for all $x$,

$$\begin{aligned}
\left\| p_{Y|X=x}(y) - p_{Y_Q|X_Q=x}(y) \right\|_{TV} &\leq \sum_{x'} \| p_{Y|X=x'}(y) - p_{Y_Q|X_Q=x'}(y) \|_{TV} \\
&\leq \frac{1}{c} \sum_{x'} p_0(x') \; \| p_{Y|X=x'}(y) - p_{Y_Q|X_Q=x'}(y) \|_{TV} \\
&= \frac{1}{c} \sum_{x',y} p_0(x') \; \left| p(y|x') - p_{Y_Q|X_Q}(y|x') \right| \\
&= \frac{1}{c} \sum_{x',y} \left| p_0(x') \left( p(y|x') - p_{Y_Q|X_Q}(y|x') \right) \right| \\
&= \frac{1}{c} \sum_{x',y} \left| p_0(x')p(y|x') - p_{X_Q,Y_Q}(x',y) \right| \\
&= \frac{1}{c} \left\| p_0(x)p(y|x) - p_{X_Q,Y_Q}(x,y) \right\|_{TV} \\
&\leq \frac{1}{c} \, \epsilon.
\end{aligned}$$

Therefore, the rate-coordination triple $(R_0, R, p(y|x))$ is arbitrarily close to the closed set $\underline{\mathcal{C}}'_{p_0}$. We conclude,

$$(R_0, R, p(y|x)) \;\in\; \underline{\mathcal{C}}'_{p_0}.$$

It remains to bound the cardinality of $U$. We can again use the standard method of [27]. The variable $U$ should have $|\mathcal{X}||\mathcal{Y}| - 1$ elements to preserve the joint distribution $p_0(x)p(y|x)$, which in turn preserves $H(X,Y)$, and another element to preserve $H(X|U)$, and one more to preserve $H(X,Y|U)$.

**Two nodes - Theorem 14 - extreme point**

In addition to characterizing the strong coordination capacity region for the two-node network, Theorem 14 states that $(R_0, I(X;Y))$ is only in the strong rate-coordination region $\underline{\mathcal{R}}_{p_0}$ if $R_0 \geq H(Y\dagger X)$. In other words, this is the least amount of common randomness needed to fully expand the strong coordination capacity region.

To prove this, first consider the implications of $R = I(X;Y)$. This means that in order to satisfy the first rate inequality in the definition of $\underline{\mathcal{C}}_{p_0}$ in Theorem 14, we must have $I(X;U) \leq I(X;Y)$. However, because of the Markovity, $I(X;U) = I(X;U,Y)$. Therefore, $I(X;U|Y) = 0$, which implies a second Markov condition $X - Y - U$ in addition to $X - U - Y$.

We are concerned with minimizing the required rate of common randomness $R_0$. Since $R = I(X;Y)$, the second rate inequality in the definition of $\underline{\mathcal{C}}_{p_0}$ in Theorem 14 becomes $R_0 \geq I(Y;U|X)$. The conditional entropy $H(Y|X)$ is fixed, so we want to maximize the conditional entropy $H(Y|U,X)$.

With the distribution $p(x|y)$ in mind, we can clump values of $Y$ together for which the channel from $Y$ to $X$ is identical. Define a function $f$ with the property that

$$f(y) = f(\tilde{y}) \iff p(x|y) = p(x|\tilde{y}) \text{ for } \forall x \in \mathcal{X}. \tag{3.7}$$

Letting $U = f(Y)$ will be the choice of $U$ that simultaneously maximizes $H(Y|U,X)$ and satisfies the Markov conditions $X - U - Y$ and $X - Y - U$. We can compare $U$ to any other choice $\tilde{U}$ that satisfies the Markov conditions and show that the resulting conditional entropy $H(Y|\tilde{U},X)$ is smaller.

Another way to state the two Markov conditions is that for all values of $y$ and $\tilde{u}$ such that $p(y,\tilde{u}) > 0$, the conditional distributions $p(x|y)$ and $p(x|\tilde{u})$ are equal because $p(x|y) = p(x|y,\tilde{u}) = p(x|\tilde{u})$. Notice that the value of $U = f(Y)$, characterized in (3.7), only depends on the channel $p(x|y)$. However, with probability one this is the same as the channel $p(x|\tilde{u})$. This means that $U$ is a function of $\tilde{U}$, and clearly

$$\begin{aligned} H(Y|\tilde{U},X) &= H(Y|U,\tilde{U},X) \\ &\leq H(Y|U,X). \end{aligned}$$

### Game theory

The method of achieving pairs $(R, \Theta)$ in the interior of $\mathcal{G}_0$ is straightforward. Use the communication to share random bits for use as common randomness. Then generate actions by the method discussed in Section 3.2 that achieve strong coordination for a joint distribution $p(x, y)$. The resulting actions will have a distribution close in total variation to the i.i.d. distribution according to $p(x, y)$, which means the expected payoff in the game won't be far from the performance of $p(x, y)$.

Some pairs $(R, \Theta)$ in $\mathcal{G}$ may require splitting time between two Pareto optimal strategies in $\mathcal{G}_0$—one strong strategy and another one requiring lighter communication resources.

We must also prove that there is no better way to use the communication to generate actions than what is implied by the region $\mathcal{G}$. Assume that $(R, \Theta)$ is achievable. We again use the technique of time mixing from Section 2.4.2, where $Q$ is uniformly distributed on the set $\{1, ..., n\}$ and independent of the action sequences $X^n$ and $Y^n$. The important elements of the converse are the following inequalities:

$$\begin{aligned} nR &\geq H(I) \\ &\geq I(X^n, Y^n; I) \\ &= \sum_{i=q}^{n} I(X_q, Y_q; I | X^{q-1}, Y^{q-1}) \\ &= nI(X_Q, Y_Q; I | X^{Q-1}, Y^{Q-1}, Q). \end{aligned}$$

Identify the tuple $(X^{Q-1}, Y^{Q-1}, Q)$ as an auxiliary random variable $W$, and notice that $X_Q$ and $Y_Q$ are conditionally independent given $I$, $X^{Q-1}$, $Y^{Q-1}$, and $Q$. In fact, this Markov relationship would hold even if Player 1 and Player 2 were allowed to observe the past actions of their counterparts before choosing their next actions.

We can combine the requirement for achievability in (3.2) with the definition of

$\Theta_q$ in (3.1) and simplify the iterated expectation:

$$\begin{aligned}
\mathbf{E}\left[\frac{1}{n}\sum_{q=1}^{n}\Theta_q\right] &= \frac{1}{n}\sum_{q=1}^{n}\mathbf{E}\,\Theta_q \\
&= \sum_{q=1}^{n}p(q)\,\mathbf{E}\,\Theta_q \\
&= \mathbf{E}\,[\,\mathbf{E}\,[\Theta_Q|Q]] \\
&= \mathbf{E}\,\Theta_Q,
\end{aligned}$$

where $\Theta_Q$ is derived from the definition of $\Theta_q$ in (3.1) to be,

$$\Theta_Q = \min_{z\in\mathcal{Z}}\mathbf{E}\left[\Pi(X_Q,Y_Q,z)|X^{Q-1},Y^{Q-1},Q\right].$$

Since $(R,\Theta)$ is achievable, we can conclude that there exists a sequence of protocols such that

$$\begin{aligned}
\lim_{n\to\infty}\mathbf{E}\left[\min_{z\in\mathcal{Z}}\mathbf{E}\left[\Pi(X_Q,Y_Q,z)|W\right]\right] &= \mathbf{E}\,\Theta_Q \\
&= \lim_{n\to\infty}\mathbf{E}\left[\frac{1}{n}\sum_{q=1}^{n}\Theta_q\right] \\
&> \Theta.
\end{aligned}$$

To summarize, there exists as $n$ large enough so that

$$\begin{aligned}
\mathbf{E}\left[\min_{z\in\mathcal{Z}}\mathbf{E}\left[\Pi(X_Q,Y_Q,z)|W\right]\right] &> \Theta, \\
R &\geq I(X_Q,Y_Q;I|W), \\
X_Q- \;(I,W)\; -Y_Q &\text{ forms a Markov chain.}
\end{aligned}$$

But notice that this is exactly the list of requirements for $(R,\Theta)$ to be in the convex hull of $\mathcal{G}_0$.

# Chapter 4

# Extension

Rather than inquire about the possibility of moving data in a network, we have asked for the set of all achievable joint distribution on actions at the nodes. For some three-node networks we have fully characterized the answer to this question, while for others we have established bounds.

Some of the results discussed in this work extend nicely to larger networks. Consider for example an extended cascade network shown in Figure 4.1, where $X$ is given randomly by nature and $Y_1$ through $Y_{k-1}$ are actions based on a cascade of communication. Just as in the cascade network of Section 2.2.3, we can achieve rates $R_i \geq I(X; Y_i, ..., Y_k)$ for empirical coordination by sending messages to the last nodes in the chain first and conditioning later messages on earlier ones. These rates meet the cut-set bound. We now can make an interesting observation about assigning unique tasks to nodes in such a network. Suppose $k$ tasks are to be completed by the $k$ nodes in this cascade network, one at each node. Node X is assigned a task randomly, and the communication in the network is used to assign a permutation of all the tasks to the nodes in the network. The necessary rates in the network are $R_i \geq \log(\frac{k}{i})$. The sum of all the rates in the network, for large $k$, is then approximately $R_{total} \geq k$ nats, where $k$ is the number of tasks and nodes in the network.

Now consider the same task assignment scenario for an extended broadcast network shown in Figure 4.2. Here again $X$ is given randomly by nature, but $Y_1$ through $Y_{k-1}$ are actions based on individual messages sent to each of the nodes. Again, we

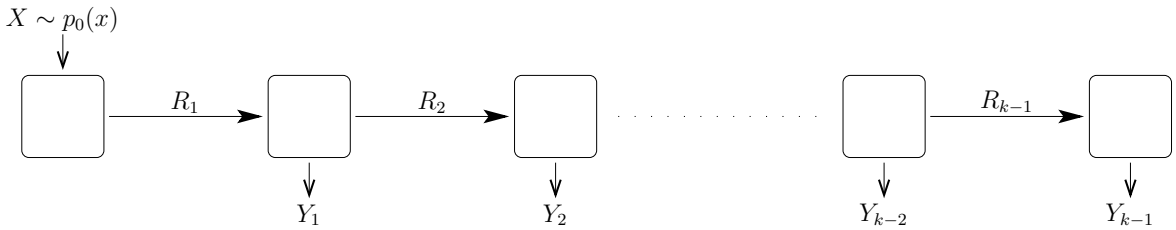


Figure 4.1: *Extended cascade network.* This is an extension of the cascade network of Section 2.2.3. Action $X$ is given randomly by nature according to $p_0(x)$, and a cascade of communication is used to produce actions $Y_1$ through $Y_{k-1}$. The coordination capacity region contains all rate-coordination tuples that satisfy $R_i \geq I(X; Y_i, ..., Y_k)$ for all $i$. In particular, the sum rate needed to assign a permutation of $k$ tasks to the $k$ nodes grows linearly with the number of nodes.

want to assign a permutation of all the $k$ tasks to all of the $k$ nodes. We can use ideas from the broadcast network results of Section 2.2.5. For example, let us assign default tasks to the nodes so that $Y_1 = 1, ..., Y_{k-1} = k-1$ unless told otherwise. Now the communication is simply used to tell each node when it must choose task $k$ rather than the default task, which will happen about one time out of $k$. The rates needed for this scheme are $R_i \geq H(1/k)$, where $H$ is the binary entropy function. For large $k$, the sum of all the rates in the network is approximately $R_{total} \geq \ln k + 1$ nats. The cut-set bound gives us a lower bound on the sum rate of $R_{total} \geq \ln k$ nats. Therefore, we can conclude that the optimal sum rate scales with the logarithm of the number of nodes in the network.

Even without explicitly knowing the coordination capacity region for the broadcast network, we are able to use bounds to establish the scaling laws for the total rate needed to assign tasks uniquely, and we can compare the efficiency of the broadcast network (logarithmic in the network size) with that of the cascade network (linear in the network size) for this kind of coordination.

We would also like to understand the coordination capacity region for a noisy network. For example, the communication capacity region for the broadcast channel $p(\tilde{y}_1, \tilde{y}_2|\tilde{x})$ of Figure 4.3 has undergone serious investigation. The standard question is, how many bits of independent information can be communicated from $X$ to $Y_1$ and from $X$ to $Y_2$. We know the answer if the broadcast channel is degraded; that is, if $Y_2$

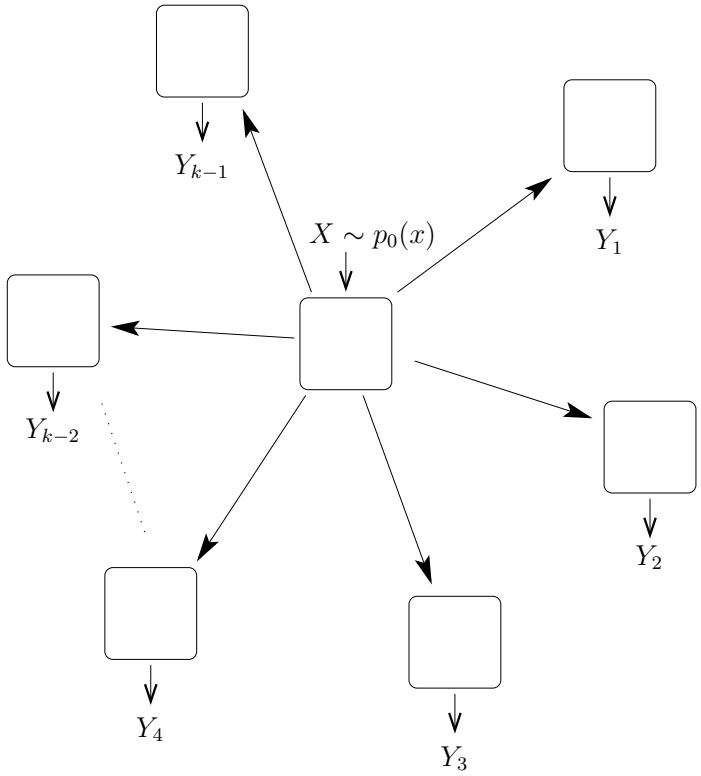


Figure 4.2: *Extended broadcast network.* This is an extension of the broadcast network of Section 2.2.5. Action $X$ is given randomly by nature according to $p_0(x)$, and each peripheral node produces an action $Y_i$ based on an individual message at rate $R_i$. Bounds on the coordination capacity region show that the sum rate needed to assign a permutation of $k$ tasks to the $k$ nodes grows logarithmically with the number of nodes.

can be viewed as a noisy version of $Y_1$. We also know the answer if the channel can be separated into two orthogonal channels or is deterministic. But what if instead we are trying to coordinate actions via the broadcast channel, similar to the broadcast network of Section 2.2.5? Now we care about the dependence between $Y_1$ and $Y_2$. The broadcast channel will impose a natural dependence between the channel outputs $\tilde{Y}_1$ and $\tilde{Y}_2$ that we abolish if we try to send independent information to the two nodes. After all, the communication capacity region for the broadcast channel depends only on the marginals $p(\tilde{y}_1|\tilde{x})$ and $p(\tilde{y}_2|\tilde{x})$. Here we are wasting a valuable resource—the natural conditional dependence between $\tilde{Y}_1$ and $\tilde{Y}_2$ given $\tilde{X}$.

Again, we are enlarging the focus from communication of independent information

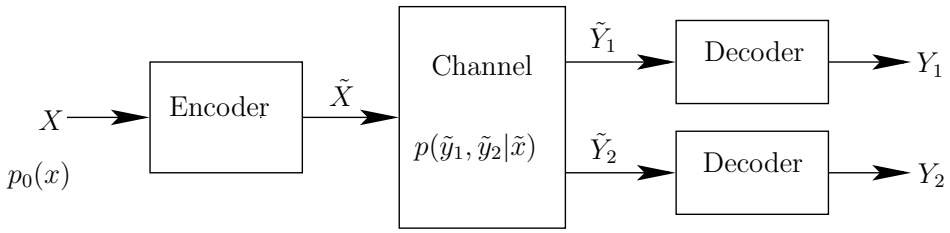


Figure 4.3: *Broadcast channel.* When a noisy channel is used to coordinate joint actions $(X, Y_1, Y_2)$, what is the resulting coordination capacity region? The broadcast network of Section 2.2.5 is a noiseless special case.

to the creation of coordinated actions. This larger question may force a simpler solution and illuminate the problem of independent information (the standard channel capacity formulation) as a special case. Presumably, information is being communicated for a reason—so future cooperative behavior can be achieved.[1]

[1]This dissertation draws on the author's collaborative work from [32], [24], [33], and [34].

# Bibliography


[1] R. Ahlswede, N. Cai, S.-Y. Li, and R. Yeung. Network information flow. *IEEE Trans. on Info. Theory*, 46(4):1204–1216, July 2000.

[2] J. Tsitsiklis, D. Bertsekas, and M. Athans. Distributed asynchronous deterministic and stochastic gradient optimization algorithms. *IEEE Trans. on Automatic Control*, 31(9):803–812, Sept. 1986.

[3] L. Xiao, S. Boyd, and S.-J. Kim. Distributed average consensus with least-mean-square deviation. *Journal of Parallel and Distributed Computing*, 67(1):33–46, Jan. 2007.

[4] B. Bollobas. *The Art of Mathematics: Coffee Time in Memphis*. Cambridge University Press, 2006.

[5] A. Yao. Some complexity questions related to distributive computing(preliminary report). In *ACM Symposium on Theory of Computing*, pages 209–213, 1979.

[6] A. Orlitsky and A. El Gamal. Average and randomized communication complexity. *IEEE Trans. on Info. Theory*, 36(1):3–16, Jan. 1990.

[7] O. Ayaso, D. Shah, and M. Dahleh. Distributed computation under bit constraints. In *IEEE Conference on Decision and Control*, pages 4837–4842, Dec. 2008.

[8] V. Anantharam and V. Borkar. Common randomness and distributed control; a counterexample. *Systems and Control Letters*, 56:568–572, 2007.

[9] H. Barnum, C. Caves, C. Fuchs, R. Jozsa, and B. Schumacher. On quantum coding for ensembles of mixed states. *Journal of Physics A: Mathematical and General*, 34:6767–6785, 2001.

[10] G. Kramer and S. Savari. Communicating probability distributions. *IEEE Trans. on Info. Theory*, 53(2):518–525, Feb. 2007.

[11] T. Weissman and E. Ordentlich. The empirical distribution of rate-constrained source codes. *IEEE Trans. on Info. Theory*, 51(11):3718–3733, Nov. 2005.

[12] T.S. Han and S. Verdú. Approximation theory of output statistics. *IEEE Trans. on Info. Theory*, 39(3):752–772, May 1993.

[13] C. Bennett, P. Shor, J. Smolin, and A. Thapliyal. Entanglement-assisted capacity of a quantum channel and the reverse shannon theorem. *IEEE Trans. on Info. Theory*, 48(10):2637–2655, Oct. 2002.

[14] C. Shannon. Coding theorems for a discrete source with fidelity criterion. In R. Machol, editor, *Information and Decision Processes*, pages 93–126. 1960.

[15] H. Yamamoto. Source coding theory for cascade and branching communication systems. *IEEE Trans. on Info. Theory*, 27:299–308, May 1981.

[16] A. Kaspi and T. Berger. Rate-distortion for correlated sources with partially separated encoders. *IEEE Trans. on Info. Theory*, 28:828–840, Nov. 1982.

[17] J. Barros and S. Servetto. A note on cooperative multiterminal source coding. In *Conference on Information Sciences and Systems*, March 2004.

[18] Z. Zhang and T. Berger. New results in binary multiple descriptions. *IEEE Trans. on Info. Theory*, 33:502–521, July 1987.

[19] T. Berger. Multiterminal source coding. In G. Longo, editor, *Information Theory Approach to Communications*, pages 171–231. CISM Course and Lecture, 1978.

[20] D. Vasudevan, C. Tian, and S. Diggavi. Lossy source coding for a cascade communication system with side-informations. In *Allerton Conference on Communication, Control, and Computing*, Sep. 2006.

[21] W. Gu and M. Effros. On multi-resolution coding and a two-hop network. In *Data Compression Conference*, 2006.

[22] M. Bakshi, M. Effros, W. Gu, and R. Koetter. On network coding of independent and dependent sources in line networks. In *IEEE International Symp. on Info. Theory*, Nice, 2007.

[23] A. Wyner and J. Ziv. The rate-distortion function for source coding with side information at the decoder. *IEEE Trans. on Info. Theory*, 22(1):1–10, Jan. 1976.

[24] P. Cuff, H. Su, and A. El Gamal. Cascade multiterminal source coding. In *IEEE International Symp. on Info. Theory*, Seoul, 2009.

[25] A. Orlitsky and J. Roche. Coding for computing. *IEEE Trans. on Info. Theory*, 47(3):903–917, March 2001.

[26] T. Cover and J. Thomas. *Elements of Information Theory*. Wiley, New York, 2nd edition, 2006.

[27] I. Csiszár and J. Körner. *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Academic, New York, 1981.

[28] Y. Steinberg and S. Verdú. Simulation of random processes and rate-distortion theory. *IEEE Trans. on Info. Theory*, 42(1):63–86, Jan. 1996.

[29] A. Wyner. The common information of two dependent random variables. *IEEE Trans. on Info. Theory*, 21(2):163–179, March 1975.

[30] P. Gács and J. Körner. Common information is far less than mutual information. *Problems of Control and Info. Theory*, 2:149–162, Jan. 1973.

[31] I. Devetak, A. Harrow, P. Shor, A. Winter, and C. Bennett. Quantum reverse shannon theorem. Presentation: http://www.research.ibm.com/people/b/bennetc/QRSTonlineVersion.pdf, 2007.

[32] P. Cuff. Communication requirements for generating correlated random variables. In *IEEE International Symp. on Info. Theory*, pages 1393–1397, Toronto, 2008.

[33] T. Cover and H. Permuter. Capacity of coordinated actions. In *IEEE International Symp. on Info. Theory*, Nice, 2007.

[34] P. Cuff, H. Permuter, and T. Cover. Coordination capacity. In preparation, 2009.