

# Communication and Distributional Complexity of Joint Probability Mass Functions

Sidharth Jaggi and Michelle Effros

Dept. of Electrical Engineering, Caltech, Pasadena, CA 91125, USA

{jaggi, effros}@z.caltech.edu

*Abstract* — The problem of truly-lossless ( $P_e = 0$ ) distributed source coding [1] requires knowledge of the joint statistics of the sources. In particular the locations of the zeroes of the probability mass functions (pmfs) are crucial for encoding at rates below  $(H(X), H(Y))$  [2]. We consider the distributed computation of the empirical joint pmf  $P_n$  of a sequence of random variable pairs observed at physically separated nodes of a network. We consider both worst-case and average measures of information exchange and treat both exact calculation of  $P_n$  and a notion of approximation. We find that in all cases the communication cost grows linearly with the size of the input. Further, we consider the problem of determining whether the empirical pmf has a zero in a particular location and show that in most cases considered this also requires a communication cost that is linear in the input size.

## I. INTRODUCTION

A source pmf  $Q = \{q_{xy}\}$  on the alphabet  $\mathcal{X} \times \mathcal{Y}$  generates an independent and identically distributed sequence of random variable pairs  $(X_i, Y_i)$ ,  $i \in \{1, \dots, n\}$ . Xavier observes only  $\{X_i\}_{i=1}^n$ ; Yvonne observes only  $\{Y_i\}_{i=1}^n$ . Let  $1(\cdot)$  be the indicator function. The empirical pmf  $P_n(X^n, Y^n) = \{p_{xy}\}$  of  $(X^n, Y^n)$  is defined as

$$p_{xy} = \frac{1}{n} \sum_{k=1}^n 1((X(k), Y(k)) = (x, y)), \text{ for } x \in \mathcal{X}, y \in \mathcal{Y}.$$

We study two problems. In the first, Xavier and Yvonne communicate with the goal of computing approximation  $\hat{P}_n$  of  $P_n$ . We define  $F(\hat{P}_n, \epsilon)$  as 1 if  $\|P_n - \hat{P}_n\|_1 < \epsilon$  and 0 otherwise, where  $\|P_n - \hat{P}_n\|_1 = \frac{1}{2} \sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} |P_n(x, y) - \hat{P}_n(x, y)|$  is the variational distance between  $P_n$  and  $\hat{P}_n$ . In the second problem, Xavier and Yvonne communicate with the goal of determining whether  $p_{ab} = 0$  for a fixed  $(a, b) \in \mathcal{X} \times \mathcal{Y}$ . We define  $G(P_n, (a, b))$  as 1 if  $p_{ab} = 0$ , and 0 otherwise.

We restate several definitions from [3]. Consider function  $f_n : \mathcal{X}^n \times \mathcal{Y}^n \rightarrow \mathcal{Z}$ . A communication protocol  $\Phi_n$  is a binary tree where each internal node  $v$  is labeled either by a function  $a_v : \mathcal{X}^n \rightarrow \{0, 1\}$  or by a function  $b_v : \mathcal{Y}^n \rightarrow \{0, 1\}$ , and each leaf is labeled with an element  $z \in \mathcal{Z}$ . The value of  $\Phi_n$  on input  $(x^n, y^n)$ , denoted by  $\Phi_n(x^n, y^n)$ , is the label of the leaf reached by starting from the root and walking on the tree in the following manner. At each internal node  $v$  labeled by  $a_v$ , walk left if  $a_v(x^n) = 0$  and right if  $a_v(x^n) = 1$ , and at each internal node  $v$  labeled by  $b_v$ , walk left if  $b_v(y^n) = 0$  and right if  $b_v(y^n) = 1$ . We say that  $\Phi_n$  computes function  $f$  if  $\Phi_n(x^n, y^n) = f(x^n, y^n)$  for all  $(x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n$ .

For each  $(x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n$ , let  $l_\Phi(x^n, y^n)$  be the length of the path from the root to the leaf for  $(x^n, y^n)$ . The cost of  $\Phi_n$  is  $C(\Phi_n) = \max_{(x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n} l_\Phi(x^n, y^n)$ . The communication complexity  $D(f)$  of  $f$ , is the minimum of  $C(\Phi_n)$ , over all protocols  $\Phi_n$  that compute  $f$ . Let  $\mu$  be a pmf on  $\mathcal{X}^n \times \mathcal{Y}^n$ . The average cost of  $\Phi_n$  over  $\mu$  is  $A(\Phi_n, \mu) = \sum_{(x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n} \mu(x^n, y^n) l_\Phi(x^n, y^n)$ . The distributional complexity  $D^\mu(f)$  of  $f$  for pmf  $Q$  is the minimum of  $A(\Phi_n, \mu)$  over all protocols  $\Phi_n$  that compute  $f$ .

## II. RESULTS

A  $z$ - $f$ -monochromatic rectangle (denoted by  $z$ - $f$ -mr) is a set  $R \subseteq \mathcal{X}^n \times \mathcal{Y}^n$  such that for all  $(x^n, y^n), (\bar{x}^n, \bar{y}^n) \in R$ ,  $f(x^n, y^n) = f(\bar{x}^n, \bar{y}^n) = z \in \mathcal{Z}$  and  $(x^n, \bar{y}^n), (\bar{x}^n, y^n) \in R$ . Let  $w_{Q^n}(f, z) = \max_{R: R \text{ is a } z\text{-}f\text{-mr}} \sum_{(x^n, y^n) \in R} Q^n(x^n, y^n)$  and  $W_{Q^n}(f, z) = \sum_{(x^n, y^n): f(x^n, y^n) = z} Q^n(x^n, y^n)$ .

**Lemma 1** Given a fixed pmf  $Q$ ,

$$D^{Q^n}(f) > - \sum_{z \in \mathcal{Z}} W_{Q^n}(f, z) \log w_{Q^n}(f, z).$$

Using combinatorial arguments and Lemma 1, Theorem 1 shows that the worst case and average case costs of calculating the empirical joint pmf are linear in  $n$ .

**Theorem 1** For any  $\epsilon > 0$ ,

$$\begin{aligned} nI_Q(X; Y) - 2^{-\Omega(n\epsilon^2)} - O(n\epsilon \log(1/\epsilon)) \\ \leq D^{Q^n}(F(\hat{P}_n, \epsilon)) \leq D(F(\hat{P}_n, 0)) \\ \leq \min\{\lceil n \log(|\mathcal{X}|) \rceil, \lceil n(\log(|\mathcal{Y}|)) \rceil\} + O(\log(n)). \end{aligned}$$

Using techniques similar to those in Theorem 1, Theorem 2 shows that the worst case cost of determining whether  $p_{ab}$  equals 0 is linear in  $n$ . If  $q_{ab} > 0$ , then the average cost is constant; otherwise it is also linear in  $n$ .

**Theorem 2**

$$\begin{aligned} D(G(P_n, (a, b))) &\geq n \log \left( \frac{|\mathcal{X}||\mathcal{Y}| - 1}{|\mathcal{X}||\mathcal{Y}| - \min\{|\mathcal{X}|, |\mathcal{Y}|\}} \right), \text{ if } \forall (a, b), q_{ab} > 0 \\ D^{Q^n}(G(P_n, (a, b))) &\begin{cases} \leq (\lceil \log |\mathcal{X}| \rceil + \lceil \log |\mathcal{Y}| \rceil) / q_{ab}, & \text{if } q_{ab} > 0 \\ \geq n(q_a + q_b)H(q_a/(q_a + q_b)) \\ \quad - O(\sqrt{n \log^2(n)}), & \text{if } q_{ab} = 0 \end{cases} \end{aligned}$$

where  $q_a = \sum_{y \in \mathcal{Y}} q_{ay}$  and  $q_b = \sum_{x \in \mathcal{X}} q_{xb}$ .

## REFERENCES

- [1] Q. Zhao and M. Effros, "Optimal Code Design for Lossless and Near Lossless Source Coding in Multiple Access Networks," *Proceedings of the IEEE Data Compression Conference, Snowbird, Utah*, March 2001.
- [2] H. S. Witsenhausen, "The Zero-Error Side Information Problem and Chromatic Numbers," *IEEE Transactions on Information Theory*, vol. 22, pp. 592–593, 1976.
- [3] E. Kushilevitz and N. Nisan, *Communication Complexity*, Cambridge University Press, 1997.

<sup>1</sup>This work was supported by NSF Grant CCR-0220039 and a grant from the Lee Center for Advanced Networking at Caltech.