On Rate Requirements for Achieving the Centralized Performance in Distributed Estimation

Mostafa El Gamal and Lifeng Lai

Abstract—We consider a distributed parameter estimation problem, in which multiple terminals send messages related to their local observations using limited rates to a fusion center who will obtain an estimate of a parameter related to the observations of all terminals. It is well known that if the transmission rates are in the Slepian-Wolf region, the fusion center can fully recover all observations and hence can construct an estimator having the same performance as that of the centralized case. One natural question is whether Slepian-Wolf rates are necessary to achieve the same estimation performance as that of the centralized case. In this paper, we show that the answer to this question is negative. We establish our result by explicitly constructing an asymptotically minimum variance unbiased estimator (MVUE) that has the same performance as that of the optimal estimator in the centralized case while requiring information rates less than the conditions required in the Slepian-Wolf rate region. The key idea is that, instead of aiming to recover the observations at the fusion center, we design universal schemes enabling the fusion center to compute a sufficient statistic using rates outside of the Selpian-Wolf region.

Index Terms—Distributed learning, MVUE, estimation algorithm, Slepian-Wolf rates, universal encoding/decoding scheme.

I. INTRODUCTION

Motivated by applications in sensor networks and other areas, the problem of distributed estimation has been extensively investigated from various perspective [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17]. As observations are distributed over multiple terminals in the distributed setting, the performances of distributed estimators are no better than those of centralized estimators who have access to all observations. The question we address in this paper is, to achieve the same performance as that of the centralized setup, how much information has to be exchanged in the distributed setting.

We consider this problem for the following setup. There are two random variables (X, Y) with a joint probability mass function (PMF) $P_\theta(X, Y)$ parameterized by an unknown parameter θ. Two terminals A and B observe X^n and Y^n respectively and send messages related to their own local observations with limited rates to terminal C, which will then obtain an estimate of the unknown parameter. It is well known that if the transmission rates from the terminals are inside the Slepian-Wolf rate region [18], there exists a universal

coding scheme [19] that enables terminal C to fully recover (X^n, Y^n). Hence, once the transmission rates are inside the Slepian-Wolf rate region, the performance of the best estimator for the distributed setup is the same as that of the best estimator for the centralized case.

One natural question is: are Slepian-Wolf rates *necessary* to achieve the same estimation performance as that of the centralized case? The answer to this question has significant implications in the distributed estimation. If the answer is yes, then to obtain the best estimate of the unknown parameter requires transmission rates to be so high that they are sufficient to fully recover the observations at the decoder, hence no rate reduction is possible. On the other hand, if the answer is no, then the observations can be compressed beyond the limits of source coding for full observation recovery. At a first glance, the answer to this question should be no as we are only interested in estimating a parameter related to the observations and are not interested in recovering the observations themselves. However, all existing related works indicate otherwise. For example, [20] addressed the same question and suggested that Slepian-Wolf rates might be necessary. In addition, the performance of the best known estimator by Han and Amari [21] does not match that of the centralized case when the information rates are outside of the Slepian-Wolf rate region. Furthermore, [22] showed that, under certain conditions, extracting even one bit of information from distributed sources is as hard as recovering full observations and hence requires the information rates to be in the Slepian-Wolf rate region.

In this paper, we show that the answer to this question is indeed *no*. We establish our result by explicitly constructing a distributed estimation algorithm that achieves the same performance as that of the optimal estimator for the centralized case while using information rates outside of the Slepian-Wolf region. The main observation is that, to construct an estimator that has the same performance as that of the centralized case, the fusion center needs only sufficient statistics not full data. Based on this observation, the key idea of our algorithm is that, instead of trying to fully recover the source observations, we design schemes that enable the fusion center to recover sufficient statistics using less information rates.

To illustrate the idea, we first consider binary symmetric sources (i.e., both X^n and Y^n are binary sequences) parameterized by an unknown parameter θ. For this model, in our algorithm, we first design a universal coding/decoding scheme that enables terminal C to compute component-wise module-two sum $Z^n = X^n \oplus Y^n$, which can be achieved using rates

outside of the Slepian-Wolf rate region, and then construct an estimator using Z^n. Here $\oplus$ denotes element-wise xor. We show that our estimator is an asymptotically minimum variance unbiased estimator (MVUE) [23] and achieves the same variance index as that of the best estimator in the centralized case. We then generalize our study to general binary sources models that are not necessarily symmetric anymore. Compared with the symmetric case, there are two additional challenges: 1) Z^n alone is not a sufficient statistic anymore; and 2) We do not have an MVUE to compare the performance to anymore, as it is not clear whether an MVUE exists and even if it exists its form is model dependent. To address the first issue, we modify our scheme and ask the transmitters to send additional information (more specifically, empirical marginal PMF) that requires diminishing rate. Combining Z^n with these additional information, the fusion center can then construct the empirical joint PMF, which is a sufficient statistic. To address the second issue, we show a stronger result that for any centralized estimator, we can construct a plugin estimator with the same performance by using the only decoded information at terminal C. We further extend our results to a more general class of non-binary sources and show that our algorithm can also achieve the same performance as that of the best estimator in the centralized case while using transmission rates less than the conditions required in the Slepian-Wolf rate region. Finally, although our estimation algorithm achieves the centralized performance at rates less than Slepian-Wolf rates, there is no optimality guarantee at very low rates which can be the case for a number of practical applications. To address this, we propose a practical design of our estimation algorithm and show that it outperforms the best known estimator by Han and Amari [21] at all rates. In [21], the authors established their estimation algorithm by introducing auxiliary random variables and solving the maximum-likelihood equation. They showed that their estimation algorithm achieves a smaller variance than the estimator by Zhang and Berger [24].

This paper builds and substantially expand upon our recent conference paper [1]. Compared with [1], we provide detailed proof for the case with binary symmetric sources, we extend our results to general binary sources in Section IV, we also propose a practical design of our estimator that achieves a good performance at all compression rates in Section VI, and we add more numerical results in Section VII.

The rest of the paper is organized as follows. We introduce the problem formulation in Section II. In Section III, we establish our main results for binary symmetric sources, then we generalize it to non-symmetric binary sources in Section IV. We extend our work to a more general class of information sources in Section V. We propose a practical design of our estimation algorithm in Section VI. We present the simulation results in Section VII. Finally, we conclude the paper in Section VIII.

II. PROBLEM FORMULATION

Consider two information sources X and Y taking values from the discrete alphabets $\mathcal{X}$ and $\mathcal{Y}$, respectively.

$(X^n, Y^n) = \{(X_i, Y_i)\}_{i=1}^n$ are n independently and identically distributed (i.i.d.) observations drawn according to the parametric joint PMF $P_\theta(X, Y)$ where $\theta \in \Theta$ is the unknown parameter. We assume that the range of Θ is bounded and hence $\theta_u \triangleq \max\{|\inf(\Theta)|, |\sup(\Theta)|\}$ is finite. We consider a distributed setup in which X^n are observed at terminal A and Y^n are observed at terminal B. Using limited rates, these two terminals send messages related to their own local observations to a fusion center (terminal C), which will then obtain an estimate $\hat{\theta}$ of θ using these messages. The setup is illustrated in Fig. 1.
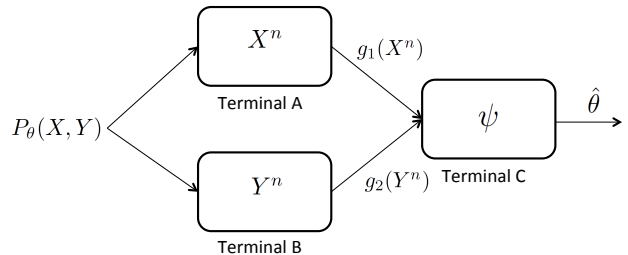

Fig. 1: System Model.

In particular, terminal A employs an encoding function $g_1 : X^n \to g_1(X^n)$, while terminal B employs an encoding function $g_2 : Y^n \to g_2(Y^n)$. The code rates are

$$R_X = \frac{\log ||g_1||}{n}, R_Y = \frac{\log ||g_2||}{n}, \tag{1}$$

where $||g_i||$ is the cardinality of the encoding function g_i.

From $g_1(X^n)$ and $g_2(Y^n)$, the decoder obtains an estimate $\hat{\theta}$ of the unknown parameter θ using estimator ψ:

$$\hat{\theta} = \psi(g_1(X^n), g_2(Y^n)). \tag{2}$$

To evaluate the quality of the estimator, we use the variance index that is defined as[1]

$$V_\theta[\hat{\theta}] = \lim_{n \to \infty} n \mathrm{Var}_\theta[\hat{\theta}] = \lim_{n \to \infty} n \mathbb{E}_\theta[(\hat{\theta} - \mathbb{E}[\hat{\theta}])^2]. \tag{3}$$

It is desirable to have an estimator that is asymptotically unbiased, i.e., $\mathbb{E}_\theta[\hat{\theta}] \to \theta$ as $n \to \infty$, range-preserving, i.e., the range of the estimation function ψ is Θ, and has a small variance index.

It is well-known that, if the coding rates satisfy (will be called Slepian-Wolf rates in the sequel)

$$R_X \geq H_\theta(X|Y), \tag{4}$$
$$R_Y \geq H_\theta(Y|X), \tag{5}$$
$$R_X + R_Y \geq H_\theta(X, Y), \tag{6}$$

there exists universal source coding schemes [19] (i.e., the coding scheme does not depends on the value of the unknown parameter θ) such that the decoder can reconstruct X^n and Y^n with a diminishing error probability. Here, $H_\theta(\cdot)$ and

---

[1]Throughout the paper, we use the subscript θ to emphasize the fact that value of the quantity of interest depends on the parameter θ.

$H_\theta(\cdot|\cdot)$ denote the entropy and conditional entropy respectively. Hence, if (4)-(6) are satisfied, we can obtain the same estimation performance as that of the centralized case.

The question we ask in this paper is: are Slepian-Wolf rates *necessary* to achieve the same estimation performance as that of the centralized case? [20] investigated the same question and suggested that Slepian-Wolf rates appear to be necessary for achieving the centralized estimation performance. In this paper, we show that Slepian-Wolf rates are *not* necessary. In particular, we show that there indeed exists a class of PMFs and the corresponding distributed estimators that require communication rates less than the Slepian-Wolf rates while still achieving the same performance as that of the best estimator for the centralized case.

Throughout the paper, we use an upper case letter U to denote a random variable, a lower case letter u to denote a realization of U, and $\mathcal{U}$ to denote the discrete alphabet from which U takes values. For any sequence $u^n = (u(1), \cdots, u(n)) \in \mathcal{U}^n$, the relative frequencies (empirical PMF) $\pi(a|u^n) \triangleq n(a|u^n)/n, \forall a \in \mathcal{U}$ of the components of u^n is called the type of u^n. Here $n(a|u^n)$ is the total number of indices t at which $u(t) = a$. Chapter 11 of [25] contains a comprehensive overview of useful properties of the type.

III. BINARY SYMMETRIC CASE

In this section, to illustrate our main idea, we first consider the case of binary symmetric sources with $|\mathcal{X}| = |\mathcal{Y}| = 2$ and a joint PMF of (X, Y) as given in Table I, in which the unknown parameter $\theta \in \Theta = (0, 1)$. The insights obtained here will be generalized to more general models in later sections.

X/Y	0	1
0	$\theta/2$	$(1-\theta)/2$
1	$(1-\theta)/2$	$\theta/2$

TABLE I: The joint PMF of binary symmetric sources.

Note that for this model, neither terminal A nor terminal B alone will be able to obtain a meaningful estimation of the value of θ, as the marginal distributions of X and Y are independent of θ. On the other hand, to estimate θ, the fusion center does not need to know (X^n, Y^n) fully. It is easy to check that the component-wise module-two sum $Z^n = X^n \oplus Y^n \triangleq [X_1 \oplus Y_1, \cdots, X_i \oplus Y_i, \cdots, X_n \oplus Y_n]$ is a sufficient statistic for estimating θ. Hence, as long as the fusion center can compute Z^n, it can construct an estimator that has the same performance as that of the centralized case. Based on this observation, we show that, to estimate θ for this class of PMFs, we can achieve the centralized estimation performance using rates that do not satisfy (4)-(6). We establish this result using two steps: 1) in the first step, we design a universal encoder at terminals A and B and universal decoder at terminal C to compute the modulo-two sum $Z^n = X^n \oplus Y^n$; 2) in the second step, we construct an estimator using Z^n and analyze its performance.

A. Step 1: Computing Z^n

Here, we discuss how to universally compute $Z^n = X^n \oplus Y^n$ at terminal C. Towards this goal, we will use the same linear code at both encoders and use a minimum entropy decoder at terminal C.

Since the encoders at terminals A and B are the same, we use the following simplified notation

$$\begin{aligned} f &= g_1 = g_2, \\ R &= R_X = R_Y. \end{aligned} \quad (7)$$

The following theorem shows that as long as $R \geq H_\theta(X|Y) = H_\theta(Y|X)$, the decoder can reconstruct Z^n with a diminishing error probability.

Theorem 1: If

$$R \geq H_\theta(X|Y) = H_\theta(Y|X), \quad (8)$$

there exist universal encoding/decoding functions to reconstruct $Z^n = X^n \oplus Y^n$ at terminal C with an exponentially decreasing error probability.

Proof: The proof follows a similar structure as the proofs in [26] and [19]. In particular, using the ideas in [19], we modify the proof of [26] to make it universal.

Random Code Generation: We use a linear code f with an encoding matrix A of size $n \times nR$ to map $\{0,1\}^n$ to $\{0,1\}^{nR}$. Hence $||f|| = 2^{nR}$. We independently generate each entry of A using a uniform binary distribution, i.e., each entry of A is 0 or 1 with probability 0.5.

Encoding: The encoded messages of the realizations $x^n \in \{0,1\}^n$ and $y^n \in \{0,1\}^n$ are

$$\begin{aligned} f(x^n) &= x^n A, \\ f(y^n) &= y^n A, \end{aligned} \quad (9)$$

in which the operations are all in binary field.

Decoding: The decoder first combines the messages into a single message as

$$f(x^n) \oplus f(y^n). \quad (10)$$

It follows from the code linearity that

$$f(x^n) \oplus f(y^n) = f(x^n \oplus y^n) = f(z^n). \quad (11)$$

From $f(x^n \oplus y^n)$, terminal C uses a minimum entropy decoder to obtain $\hat{z}^n$. In particular, for each $\bar{z}^n$ such that $f(\bar{z}^n) = f(x^n \oplus y^n)$, the minimum entropy decoder first calculates the entropy of its type, then picks the one that has the least entropy to be the decoded sequence. In the following, to simplify the notation, we use $\bar{Z}^{(n)}$ and $Z^{(n)}$ to denote dummy random variables whose PMFs $P_{\bar{Z}^{(n)}}$ and $P_{Z^{(n)}}$ are the same as the types of $\bar{z}^n$ and z^n, respectively. The final decoded message is denoted as

$$\hat{z}^n = \phi(f(z^n)), \quad (12)$$

where ϕ denotes the minimum entropy decoding function.

Error Probability Analysis: A decoding error occurs if and only if there exists a sequence $\hat{z}^n \neq z^n$ such that

$$f(\hat{z}^n) = f(z^n) \text{ and } H(\hat{Z}^{(n)}) \leq H(Z^{(n)}). \quad (13)$$

The error probability, averaging over all possible codebooks, is

$$P_e^{(n)} = \sum_{z^n \in \{0,1\}^n} P_\theta(z^n)\text{Pr}(\hat{z}^n \neq z^n) = \sum_f \text{Pr}(f)P_{e,f}^{(n)}, \quad (14)$$

in which $P_\theta(z^n) \triangleq \text{Pr}(Z^n = z^n)$, and $P_{e,f}^{(n)}$ denotes the error probability if a particular codebook f is used. By analyzing (14), we show that there exists a particular codebook f^* such that $P_{e,f^*}^{(n)} \to 0$ exponentially as $n \to \infty$ as long as the conditions in the theorem are satisfied. Detailed analysis can be found in Appendix A. This implies that if we use f^*, then the fusion center will be able to compute Z^n with an exponentially decreasing error probability. ∎

Theorem 1 implies that the required rates to decode $Z^n = X^n \oplus Y^n$ with a small error probability is

$$R_X \geq H_\theta(X|Y), \quad (15)$$
$$R_Y \geq H_\theta(Y|X). \quad (16)$$

This rate region is larger than the Slepian-Wolf region in (4)-(6), as the condition $R_X + R_Y \geq H_\theta(X,Y)$ is not necessary anymore.

B. Step 2: Estimation

After obtaining $\hat{Z}^n$, which is equal to Z^n with a probability converging to 1 exponentially, we then design an asymptotically MVUE of θ. Our estimator is

$$\hat{\theta} = \frac{n(0|\hat{Z}^n)}{n}, \quad (17)$$

in which the notation $n(\cdot|\cdot)$ is defined in Section II.

Theorem 2: If the conditions in Theorem 1 are satisfied, the estimator in (17) is an asymptotically MVUE and achieves the optimal variance index as that of the centralized case.

Proof: We establish this result by showing that the estimator (17) achieves the same performance as that of the optimal estimator in the centralized case.

Optimal Centralized Estimator: First consider the centralized case in which X^n and Y^n are both known perfectly. Let $\left(\frac{n_1}{n}, \frac{n_2}{n}, \frac{n_3}{n}, \frac{n_4}{n}\right)$ denote the joint type of the sequences x^n and y^n, where (n_1, n_2, n_3, n_4) are the frequencies of occurrence of the pairs $\{(0,0),(1,1),(0,1),(1,0)\}$, respectively. The joint PMF of (x^n, y^n) is

$$P_\theta(x^n, y^n) = \left(\frac{\theta}{2}\right)^{(n_1+n_2)} \left(\frac{1-\theta}{2}\right)^{(n_3+n_4)}. \quad (18)$$

Consider the centralized estimator

$$\hat{\theta}_c = \frac{(n_1 + n_2)}{n}. \quad (19)$$

This estimator is unbiased since

$$\mathbb{E}_\theta[\hat{\theta}_c] = \theta. \quad (20)$$

The variance of the estimator is calculated as

$$\begin{aligned} \text{Var}_\theta[\hat{\theta}_c] &= \frac{1}{n^2}\mathbb{E}_\theta[(n_1+n_2)^2] - \theta^2 \\ &= \frac{\theta(1-\theta)}{n}. \end{aligned} \quad (21)$$

The variance index is given by

$$V_\theta[\hat{\theta}_c] = \lim_{n\to\infty} n\text{Var}_\theta[\hat{\theta}_c] = \theta(1-\theta). \quad (22)$$

The Cramer-Rao lower bound (CRLB) of the centralized case is

$$\begin{aligned} \text{CRLB} &= -1/\mathbb{E}_\theta\left[\frac{\partial^2 \ln[P_\theta(x^n, y^n)]}{\partial^2\theta}\right] \\ &= \frac{\theta(1-\theta)}{n} = \text{Var}_\theta[\hat{\theta}_c]. \end{aligned} \quad (23)$$

This implies that $\hat{\theta}_c$ is an MVUE for the centralized case.

Comparison: Now, come back to our decentralized case, for which our estimator is

$$\hat{\theta} = \frac{n(0|\hat{Z}^n)}{n}. \quad (24)$$

We will compare the performance of $\hat{\theta}$ with that of the optimal centralized estimator $\hat{\theta}_c$.

For the codebook f^*, define $T_e^{(n)}$ as the set of sequences z^ns that are incorrectly decoded. Therefore,

$$P_{e,f^*}^{(n)} = \sum_{z^n \in T_e^{(n)}} P_\theta(z^n). \quad (25)$$

The expected value of our estimator is given by

$$\mathbb{E}_\theta[\hat{\theta}] = \sum_{z^n \in \{0,1\}^n} \text{Pr}(\hat{Z}^n = z^n)\frac{n(0|z^n)}{n}. \quad (26)$$

Note that $\text{Pr}(\hat{Z}^n = z^n)$ is not necessarily equal to $P_\theta(z^n)$, and the sum of the probability difference can be bounded as

$$\begin{aligned} &\sum_{z^n \in \{0,1\}^n} |\text{Pr}(\hat{Z}^n = z^n) - P_\theta(Z^n = z^n)| \\ &\leq 2\sum_{z^n \in T_e^{(n)}} P_\theta(z^n) = 2P_{e,f^*}^{(n)}. \end{aligned} \quad (27)$$

We have that

$$\begin{aligned} &|\mathbb{E}_\theta[\hat{\theta}] - \mathbb{E}_\theta[\hat{\theta}_c]| \\ &\leq \sum_{z^n \in \{0,1\}^n} |\text{Pr}(\hat{Z}^n = z^n) - P_\theta(Z^n = z^n)|\frac{n(0|z^n)}{n}. \end{aligned}$$

Since

$$0 \leq \frac{n(0|z^n)}{n} \leq 1, \quad (28)$$

then

$$\begin{aligned} |\mathbb{E}_\theta[\hat{\theta}] - \mathbb{E}_\theta[\hat{\theta}_c]| &\leq \sum_{z^n \in \{0,1\}^n} |\text{Pr}(\hat{Z}^n = z^n) - P_\theta(Z^n = z^n)| \\ &\leq 2P_{e,f^*}^{(n)}, \end{aligned} \quad (29)$$

in which the last inequality is due to (27).

As $P_{e,f^*}^{(n)}$ is shown to converge to zero exponentially fast in Section III-A, we have

$$\lim_{n\to\infty} \mathbb{E}_\theta[\hat{\theta}] = \mathbb{E}_\theta[\hat{\theta}_c] = \theta. \quad (30)$$

This shows that our estimator is asymptotically unbiased. Similarly, we have

$$
\begin{aligned}
&|\text{Var}_\theta[\hat\theta] - \text{Var}_\theta[\hat\theta_c]| \\
&\leq \quad |\mathbb{E}_\theta[\hat\theta^2] - \mathbb{E}_\theta[\hat\theta_c^2]| + |\mathbb{E}_\theta[\hat\theta] - \mathbb{E}_\theta[\hat\theta_c]| \\
&\leq \quad 4P_{e,f^*}^{(n)}.
\end{aligned} \tag{31}
$$

Hence,

$$
|V_\theta[\hat\theta] - V_\theta[\hat\theta_c]| \leq \lim_{n\to\infty} 4nP_{e,f^*}^{(n)}. \tag{32}
$$

As $n \to \infty$, $P_{e,f^*}^{(n)} \to 0$ exponentially, we have $4nP_{e,f^*}^{(n)} \to 0$. Therefore,

$$
V_\theta[\hat\theta] = V_\theta[\hat\theta_c] = \theta(1-\theta). \tag{33}
$$

This proves that our estimator is asymptotically unbiased and achieves the same minimum variance that can be achieved even in the centralized case. Hence, our estimator is optimal. ∎
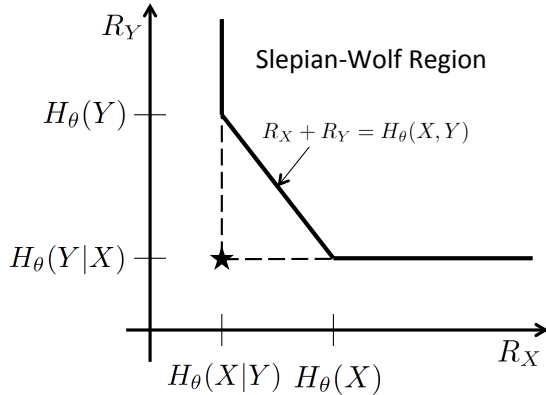

Fig. 2: ★: the rate pair required in our estimator, which is outside of the Slepian-Wolf rate region.

Combining Theorems 1 and 2, we conclude that, in the distributed parameter estimation, the Slepian-Wolf rates are not necessary to achieve the same optimal estimation performance as that of the centralized case. Fig. 2 illustrates the comparison between the Slepian-Wolf rate region and the rate pair used in our estimator.

IV. GENERAL BINARY CASE

In this section, we extend our study to the general binary source models $P_\theta(X, Y)$. Here, we do not make any particular assumption of the form of $P_\theta(X, Y)$. For example, $P_\theta(X, Y)$ could be a nonlinear function of θ. Similar to the previous section, we assume that $P_\theta(X = i, Y = j) > 0$ for all $\theta \in \Theta$ and $i, j \in \{0, 1\}$. Compared with the binary symmetric source model considered in Section III, there are two additional challenges. First, the component-wise module-two sum Z^n is not a sufficient statistic in general, hence recovering Z^n alone is not enough. Second, unlike the symmetric case in which we have an MUVE centralized estimator to compare to, we cannot do that anymore as we are considering general models whose optimal centralized is model specific (and in

some cases, MUVE may not exist). Despite these challenges, we prove the following result:

Theorem 3: For any binary source with a parametric PMF $P_\theta(X, Y)$, where $\theta \in \Theta$ is the unknown parameter and Θ is a bounded set, there exits an unbiased estimator $\hat{F}$ based on $Z^n = X^n \oplus Y^n$ that achieves the centralized performance asymptotically and requires communication rates of

$$
R_X = R_Y \geq H_\theta(Z). \tag{34}
$$

Proof: The proof consists of two main steps: 1) in the first step, we construct a scheme to enable the fusion center to compute a sufficient statistic with exponentially diminishing error probability; 2) in the second step, we establish an estimator using the computed statistics and show that the estimator achieves the performance of the centralized estimator.

Step 1: Computing a Sufficient Statistic

Different from the binary symmetric case considered in Section III, Z^n is not a sufficient statistic for the general binary case anymore. Now, we show the joint type $P_{X^{(n)}Y^{(n)}} = \left(\frac{n_1}{n}, \frac{n_2}{n}, \frac{n_3}{n}, \frac{n_4}{n}\right)$ of the observation sequences (x^n, y^n) is a sufficient statistic and show how to compute this statistics at the fusion center using rates (34).

Let $T_{P_{X^{(n)}Y^{(n)}}}$ be the set of all sequence pairs (x^n, y^n) that have the joint type $P_{X^{(n)}Y^{(n)}}$. The conditional PMF of (X^n, Y^n) given the joint type $P_{X^{(n)}Y^{(n)}}$ is

$$
P_\theta(x^n, y^n | P_{X^{(n)}Y^{(n)}}) = \begin{cases} 0, \text{ if } (x^n, y^n) \notin T_{P_{X^{(n)}Y^{(n)}}} \\ \frac{1}{||T_{P_{X^{(n)}Y^{(n)}}}||}, \text{ otherwise} \end{cases}, \tag{35}
$$

which is not a function of θ. Therefore, the joint type $P_{X^{(n)}Y^{(n)}}$ is a sufficient statistic of θ.

Now we show how to compute this statistic at the fusion center with rates in (34).

Encoding: At terminals A and B, we first encode X^n and Y^n using the same scheme presented in Section III-A. This will enable terminal C to compute Z^n. In addition, each terminal will send the marginal types $P_{X^{(n)}} \in \mathcal{P}_X^{(n)}$ and $P_{Y^{(n)}} \in \mathcal{P}_Y^{(n)}$ of the sequences x^n and y^n, respectively. The number of marginal types can be bounded as [25]

$$
||\mathcal{P}_X^{(n)}|| = ||\mathcal{P}_Y^{(n)}|| \leq (n+1)^2. \tag{36}
$$

Therefore each of the marginal types can be encoded using the rate $\frac{2\log(n+1)}{n}$, which goes to zero as n increases. Hence, sending these additional information requires diminishing additional rates.

Decoding: At terminal C, we first decode $\hat{Z}^n$ using the same scheme as discussed in Section III-A. Once $\hat{Z}^n$ is decoded, terminal C will compute the joint type $\hat{P}_{X^{(n)}Y^{(n)}} = \left(\frac{\hat{n}_1}{n}, \frac{\hat{n}_2}{n}, \frac{\hat{n}_3}{n}, \frac{\hat{n}_4}{n}\right)$ by combining $\hat{Z}^n$ along with the additional information $P_{X^{(n)}}, P_{Y^{(n)}}$ sent from terminals A and B re-

spectively. In particular, from these information, we have the following relationship

$$\hat{n}_1 + \hat{n}_2 = n(0|\hat{Z}^n), \tag{37}$$

$$\frac{\hat{n}_1 + \hat{n}_3}{n} = P_{X^{(n)}}(x=0), \tag{38}$$

$$\frac{\hat{n}_1 + \hat{n}_4}{n} = P_{Y^{(n)}}(y=0), \tag{39}$$

$$\sum_{i=1}^{4} \hat{n}_i = n. \tag{40}$$

From these four equations, we can easily obtain $\hat{P}_{X^{(n)}Y^{(n)}}$.

Error Probability: Define $P_e^{(n)}$ as

$$P_e^{(n)} = \Pr(\hat{P}_{X^{(n)}Y^{(n)}} \neq P_{X^{(n)}Y^{(n)}}). \tag{41}$$

As shown in Section III-A, Z^n can be decoded at the rates given in (34) with an exponentially decreasing probability of error. Furthermore, the marginal types can be perfectly recovered at asymptotically zero rates, then the joint type $P_{X^{(n)}Y^{(n)}}$ can be computed with an exponentially decreasing error probability $P_e^{(n)}$.

Step 2: Estimation

In the binary symmetric case considered in Section III, we have MVUE for the centralized case and hence we can compare our distributed estimator with this centralized MVUE. In the general binary model, this approach will not work as we don't know whether or not an MVUE exists. Furthermore, even if it exists, the form of MVUE is model specific. In the following, we show a stronger result that we can achieve the same performance for any centralized estimator.

First, as $P_{X^{(n)}Y^{(n)}}$ is a sufficient statistic for the centralized case, by Rao-Blackwell theorem [23], if we want to minimize the variance of unbiased estimators, we can focus on estimators that are functions of $P_{X^{(n)}Y^{(n)}}$, namely $F_c = F(P_{X^{(n)}Y^{(n)}})$, for the centralized case. For any unbiased F_c, we design the following simple plugin estimator

$$\hat{F} = F(\hat{P}_{X^{(n)}Y^{(n)}}). \tag{42}$$

In the following, we compare the performance of F_c and $\hat{F}$. We have that

$$\mathbb{E}_\theta[\hat{F}] = \sum_{P_{x^{(n)}y^{(n)}}} \Pr(\hat{P}_{X^{(n)}Y^{(n)}} = P_{x^{(n)}y^{(n)}}) F(P_{x^{(n)}y^{(n)}}), \tag{43}$$

and

$$\begin{aligned} &|\mathbb{E}_\theta[\hat{F}] - \mathbb{E}_\theta[F_c]| \\ &\leq \sum_{P_{x^{(n)}y^{(n)}}} |\Pr(\hat{P}_{X^{(n)}Y^{(n)}} = P_{x^{(n)}y^{(n)}}) - P_\theta(P_{x^{(n)}y^{(n)}})| \\ &\qquad \cdot |F(P_{x^{(n)}y^{(n)}})|. \end{aligned} \tag{44}$$

Since $F(P_{x^{(n)}y^{(n)}}) \in \Theta$ and Θ is bounded, we have $|F(P_{x^{(n)}y^{(n)}})| \leq \theta_u$. Furthermore, following similar steps as that of (27), we have

$$\sum_{P_{x^{(n)}y^{(n)}}} |\Pr(\hat{P}_{X^{(n)}Y^{(n)}} = P_{x^{(n)}y^{(n)}}) - P_\theta(P_{x^{(n)}y^{(n)}})| \leq 2P_e^{(n)}.$$

As the result, we have

$$|\mathbb{E}_\theta[\hat{F}] - \mathbb{E}_\theta[F_c]| \leq 2P_e^{(n)}\theta_u, \tag{45}$$

hence

$$\lim_{n\to\infty} \mathbb{E}_\theta[\hat{F}] = E_\theta[F_c] = \theta, \tag{46}$$

as $P_e^{(n)}$ goes to zero exponentially. Similarly,

$$|\mathrm{Var}_\theta[\hat{F}] - \mathrm{Var}_\theta[F_c]| \leq 2P_e^{(n)}(\theta_u^2 + \theta_u). \tag{47}$$

Therefore,

$$V_\theta[\hat{F}] = V_\theta[F_c]. \tag{48}$$

This implies that the plugin distributed estimator $\hat{F}$ achieves the same performance as that of the centralized estimator F_c if the rate condition (34) is satisfied. ∎

Depending on the PMF of the binary source, the required sum rate to achieve the optimal centralized performance $2H_\theta(Z)$ as obtained using our algorithm can be less than Slepian-Wolf sum rate $H_\theta(X,Y)$. As an example, consider a non-symmetric nonlinear binary source with the PMF shown in Table II.

X/Y	0	1
0	$1/4 + \theta^2$	$1/4 - \theta^2$
1	$1/4 - \theta$	$1/4 + \theta$

TABLE II: An example of a joint PMF of a non-symmetric binary source with $\theta \in \Theta = (0, 1/4)$.

Although the joint PMF given in Table II is not symmetric and nonlinear in θ, the required rates to obtain an unbiased estimator that achieves the centralized performance are still lower than Slepian-Wolf rates as shown in Fig. 3.
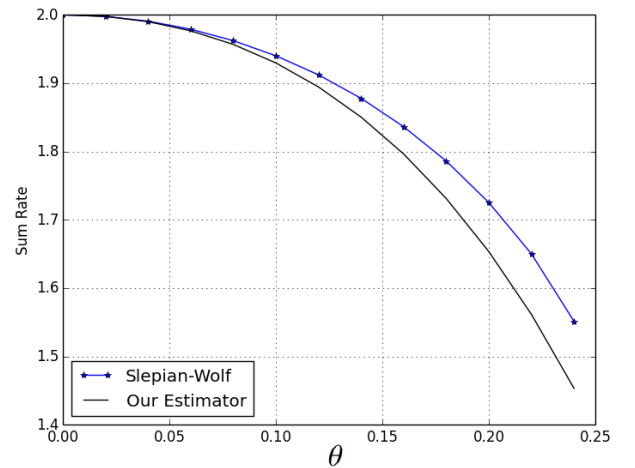

Fig. 3: The required rates to achieve the optimal centralized performance for the binary source given in Table II is lower than Slepian-Wolf rates.

V. NON-BINARY MODELS

In this section, we extend our results for binary models to more general class of non-binary models. Let $\mathcal{X} = \mathcal{Y} = \{0, 1, ..., M - 1\}$ and consider the class of PMFs

$$P_\theta(X = i, Y = j) = \begin{cases} \frac{\theta}{M} & \text{, if } (i+j) \neq M - 1 \\ \frac{1-\theta(M-1)}{M} & \text{, otherwise,} \end{cases} \quad (49)$$

where $\theta \in \Theta = (0, \frac{1}{(M-1)})$. Note that each information source has a uniform marginal PMF and setting $M = 2$ recovers the binary case.

Similar to the binary case, we first use a linear code and minimum entropy decoder to reconstruct $Z^n = (X^n + Y^n) \mod M$ at the decoder and then design an estimator from Z^n. In this section, we use $\mod M$ to denote element-wise mod operation,

In particular, we use a linear code f that maps $\{0, 1, ..., M - 1\}^n$ to $\{0, 1, ..., M - 1\}^k$. The encoded messages of the realizations $x^n \in \{0, 1, ..., M - 1\}^n$ and $y^n \in \{0, 1, ..., M - 1\}^n$ are

$$\begin{aligned} f(x^n) &= x^n A, \\ f(y^n) &= y^n A, \end{aligned} \quad (50)$$

in which the code matrix A has n rows and k columns with each entry taking values from $\{0, 1, ..., M - 1\}$. The coding rate is

$$R = \frac{k}{n} \log M. \quad (51)$$

The decoder first combines the encoded messages into a single message as

$$f(x^n) + f(y^n) \mod M. \quad (52)$$

The final decoded message is given by

$$\hat{z}^n = \phi(f(z^n)), \quad (53)$$

where ϕ the the minimum entropy decoding function. Following the same error probability analysis for the binary case, we can show that there exists a codebook f^* (and hence a particular encoding matrix A) that achieves a probability of decoding error $P_{e,f^*}^{(n)} \to 0$ exponentially as $n \to \infty$ if

$$R \geq H_\theta(Z) = H_\theta(X|Y) = H_\theta(Y|X). \quad (54)$$

Therefore, as long as

$$R_X \geq H_\theta(X|Y), \quad (55)$$
$$R_Y \geq H_\theta(Y|X), \quad (56)$$

we can reconstruct $Z^n = X^n + Y^n \mod M$ at the decoder with an exponentially diminishing error probability.

After obtaining $\hat{Z}^n$, which is equal to Z^n with a probability converging to 1 exponentially, our estimator is

$$\hat{\theta} = \frac{n - n(M - 1|\hat{Z}^n)}{n(M - 1)}. \quad (57)$$

Following similar steps as those in the binary case, we can show that, if (55)-(56) are satisfied, the estimator in (57) is asymptotically unbiased and achieves a variance index

$$V_\theta[\hat{\theta}] = \frac{\theta[1 - \theta(M - 1)]}{M - 1}. \quad (58)$$

We can further show that (58) is the best variance index that can be achieved even in the centralized case. This implies that our algorithm achieves the centralized performance using rates outside the Slepian-Wolf region.

VI. PRACTICAL APPROACH

In the previous sections, we established an unbiased estimator that achieves the centralized performance for a number of information sources, while requires less rates than Slepian-Wolf rates. For binary symmetric sources and its extension, our estimator achieves the CRLB within the combined regions of Slepian-Wolf and the dotted region as shown in Fig. 4, where

$$\begin{aligned} R_X &\geq H_\theta(X|Y), \\ R_Y &\geq H_\theta(Y|X). \end{aligned} \quad (59)$$
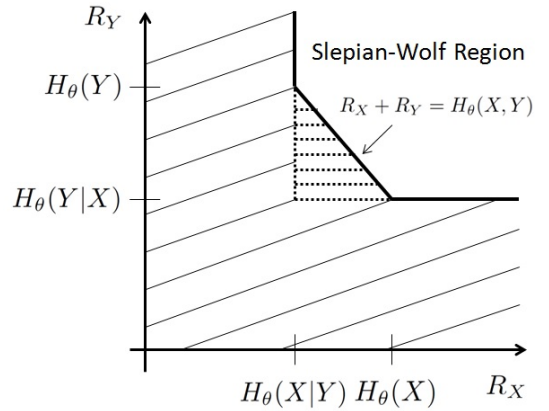

Fig. 4: The low rates inside the dashed region are considered in this section.

Our estimator is optimal if $Z^n = X^n \oplus Y^n$ is decoded with a vanishing probability of error. Otherwise, there is no optimality guarantee. In practical applications, the communications rates can be lower than our conditions (59). Therefore, we modify the design of our estimation algorithm in this section to ensure a good performance at all rates including the low rates inside the dashed region as shown in Fig. 4. We start with the case of binary symmetric sources then we extend the results to the general class of PMFs as presented in Section V. For binary symmetric sources, we assume that the unknown parameter θ takes values in $(0, t)$, where $t \in (0, 0.5)$ is known.

First, we apply the encoding/decoding scheme introduced in Section III to encode p observations (x^p, y^p) and decode $\hat{z}^p = x^p \oplus y^p$, where

$$p = \begin{cases} n, & \text{if } R \geq H(t) \\ \lfloor \frac{nR}{H(t)} \rfloor, & \text{otherwise,} \end{cases} \quad (60)$$

where $\lfloor \cdot \rfloor$ is an operator that maps its argument to the largest previous integer, then we modify our estimator as following:

$$\hat{\theta} = \frac{n(0|\hat{Z}^p)}{p}. \tag{61}$$

The following Theorem states the performance bounds of our estimator.

Theorem 4: If

$$R \geq H(t), \tag{62}$$

our estimator is an asymptotically MVUE. Otherwise, our estimator is asymptotically unbiased and its variance index is bounded as

$$V_\theta[\hat{\theta}] \leq \frac{H(t)\theta(1-\theta)}{R}. \tag{63}$$

Proof:

Case 1: $R \geq H(t)$

$$
\begin{aligned}
R &\geq H(t) \\
&\geq H_\theta(Z) = H_\theta(X|Y) = H_\theta(Y|X).
\end{aligned} \tag{64}
$$

In this case, $p = n$, and our estimator is given by

$$\hat{\theta} = \frac{n(0|\hat{Z}^n)}{n}. \tag{65}$$

As we proved in the previous sections, this estimator is an asymptotically MVUE if $R \geq H_\theta(Z)$.

Case 2: $R < H(t)$

In the centralized case, consider the estimator

$$\hat{\theta}_c = \frac{(n_1 + n_2)}{p}, \tag{66}$$

where n_1 and n_2 are the frequency of occurrence of the pairs $(0,0)$ and $(1,1)$ in the observations (x^p, y^p), respectively. We have that

$$\mathbb{E}_\theta[\hat{\theta}_c] = \frac{p\theta}{p} = \theta, \tag{67}$$

and

$$\mathrm{Var}_\theta[\hat{\theta}_c] = \frac{\theta(1-\theta)}{p}. \tag{68}$$

In the decentralized case, the effective rate per observation is given by

$$R_{eff} = \frac{nR}{p}. \tag{69}$$

Since

$$p \leq \frac{nR}{H(t)}, \tag{70}$$

then

$$R_{eff} \geq H(t) \geq H_\theta(Z). \tag{71}$$

For this range of rates, we showed that

$$\lim_{n \to \infty} \mathbb{E}_\theta[\hat{\theta}] = \mathbb{E}_\theta[\hat{\theta}_c] = \theta. \tag{72}$$

Therefore, our estimator is asymptotically unbiased. We also have that

$$
\begin{aligned}
V_\theta[\hat{\theta}] &= \lim_{n \to \infty} n \mathrm{Var}_\theta[\hat{\theta}_c] \\
&= \lim_{n \to \infty} \frac{n\theta(1-\theta)}{p}.
\end{aligned} \tag{73}
$$

It is obvious that

$$p \geq \frac{nR}{H(t)} - 1. \tag{74}$$

Hence,

$$V_\theta[\hat{\theta}] \leq \frac{H(t)\theta(1-\theta)}{R}. \tag{75}$$

$\square$

For the general class of PMFs given in (49), we assume that θ takes values in $(0, t)$, and $t \in (0, \frac{1}{2(M-1)})$. We establish our estimator as

$$\hat{\theta} = \frac{p - n(M-1|\hat{Z}^p)}{p(M-1)}. \tag{76}$$

Following similar steps to the proof of Theorem 4, we have that our estimator is an asymptotically MVUE if $R \geq H(t)$. Otherwise, our estimator is asymptotically unbiased and its variance index is bounded as

$$V_\theta[\hat{\theta}] \leq \frac{H(t)\theta[1 - \theta(M-1)]}{R(M-1)}. \tag{77}$$

For binary symmetric sources and its extension, we guarantee a worst case performance that is a function of the communication rate R. In the following section, we show that despite of a small performance degradation in the rate region $H(\theta) \leq R < H(t)$ as compared to our estimator in Section V, we managed to achieve a very good performance at low rates.

VII. NUMERICAL RESULTS

In this section, we use several numerical examples to illustrate the comparison between our estimators to the best known estimator by Han and Amari [21]. In the simulation, we fix the unknown parameter θ and change the encoding rates R_X and R_Y such that

$$R_X = R_Y = R \geq H_\theta(Z). \tag{78}$$

We conduct the comparison for $M = 2$ and $M = 4$, respectively.

For our estimator in Section V and $M = 2$, the variance index of our estimator is (33), while the variance index of the estimator by Han and Amari is calculated in example 3 of [21]

$$(V_\theta[\hat{\theta}])_{HA} \simeq \tag{79}$$
$$\frac{1}{16a^2b^2}\left\{ \frac{1}{4} - \left(\theta - \frac{1}{2}\right)^2 [1 - (1 - 4a^2)(1 - 4b^2)] \right\},$$

where a and b are functions of R_X and R_Y, whose expressions are given in (14.12) and (14.13) of [21], respectively.
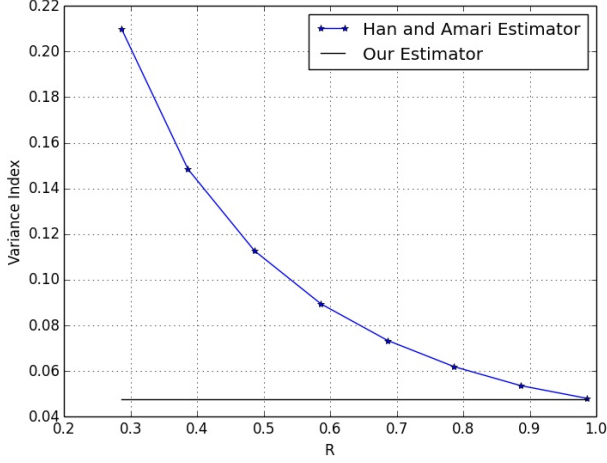
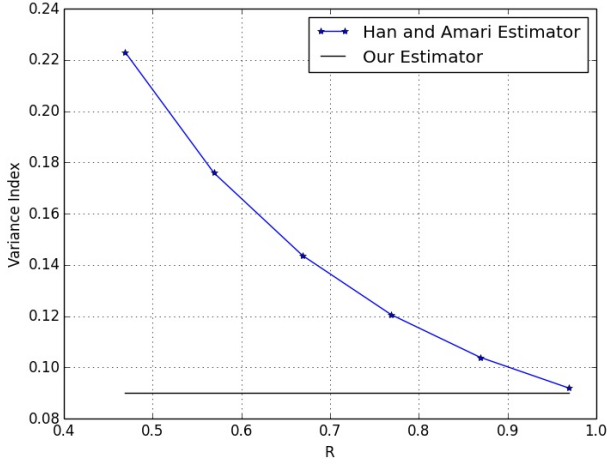Fig. 5: Performance Comparison: $\theta = 0.05$, $M = 2$

Fig. 6: Performance Comparison: $\theta = 0.9$, $M = 2$

Fig. 5 and Fig. 6 show the performance gain, in terms of the variance index, of our estimator over Han and Amari's estimator for binary symmetric sources ($M = 2$) at two different values of the unknown parameter, $\theta = 0.05$ and $\theta = 0.9$, respectively. The performance difference is more noticeable at low rates. For $\theta = 0.05$, the Slepian-Wolf sum rate is $R_X + R_Y = 1.29$ bits, while our estimator requires a sum rate of $R_X + R_Y = 2R = 0.57$ bits. For $\theta = 0.9$, the Slepian-Wolf sum rate is 1.47 bits, while our estimator requires a sum rate of 0.94 bits. Furthermore, for Han and Amari's estimator to achieve the centralized performance, the required sum-rate is 2 bits for both cases, which is not only much larger than the sum rate required in our estimator but also much larger than the sum-rate required by conditions specified in the Slepian-Wolf rate region.

For our estimator in Section V and $M = 4$, the variance index of our estimator is given in (58). The performance of

Han and Amari's estimator relies on the choice of the test channels. The authors did not specify an optimal choice of the test channels in order to extend example 3 in [21] to the case of $M = 4$. We find the following mapping to be a natural extension:

$$Q = \begin{cases} 0, & \text{if } X \in \{0,1\} \\ 1, & \text{if } X \in \{2,3\}, \end{cases} \quad T = \begin{cases} 0, & \text{if } Y \in \{0,1\} \\ 1, & \text{if } Y \in \{2,3\}. \end{cases} \quad (80)$$

Notice that (Q,T) are distributed according to a binary symmetric PMF with an unknown parameter $\alpha = 2\theta$. Using an estimator $\hat{\theta} = \frac{\hat{\alpha}}{2}$ leads to the following expression for the variance index:

$$(V_\theta[\hat{\theta}])_{HA} \simeq \qquad (81)$$
$$\frac{1}{64a^2b^2}\left\{ \frac{1}{4} - \left(2\theta - \frac{1}{2}\right)^2 [1 - (1 - 4a^2)(1 - 4b^2)] \right\}.$$
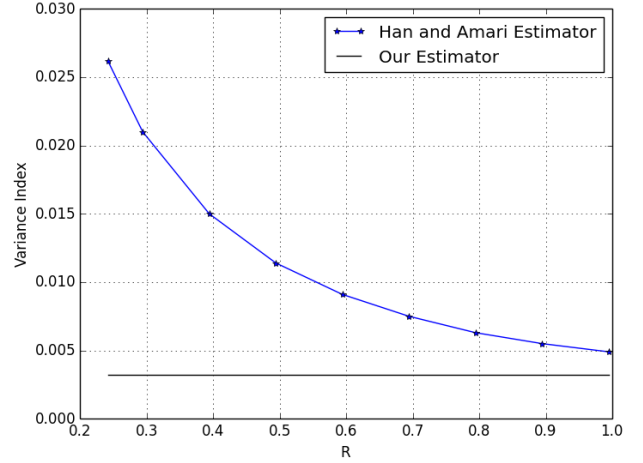

Fig. 7: Performance Comparison: $\theta = 0.01$, $M = 4$

Fig. 7 compares the variance indices achieved using our estimator and Han and Amari's estimator for $M = 4$ and $\theta = 0.01$. It is clear that our estimator outperforms that of Han and Amari's estimator. Furthermore, the performance difference is more noticeable at low rates. The Slepian-Wolf sum rate is 2.24 bits, while our estimator requires a sum rate of 0.48 bits.

For our practical estimator in Section VI and $M = 2$, the variance index of our estimator is bounded as in (77) if $R < H(t)$. Otherwise, it achieves the CRLB. The variance index of Han and Amari's estimator is (79).

For our practical estimator in Section VI and $M = 4$, the variance index of our estimator is bounded as in (75) if $R < H(t)$. Otherwise, it achieves the CRLB. The variance index of Han and Amari's estimator is (81).

Fig. 8 and Fig. 9 show that our estimator outperforms Han and Amari's estimator at all rates. The performance difference is more noticeable at very low rates, which makes our estimator a good choice for applications with low rate constraints. Our estimator performs better for smaller values of the range of θ, which is determined by t.
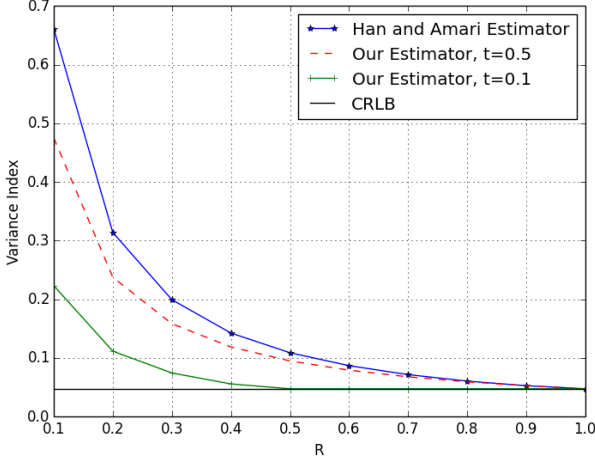
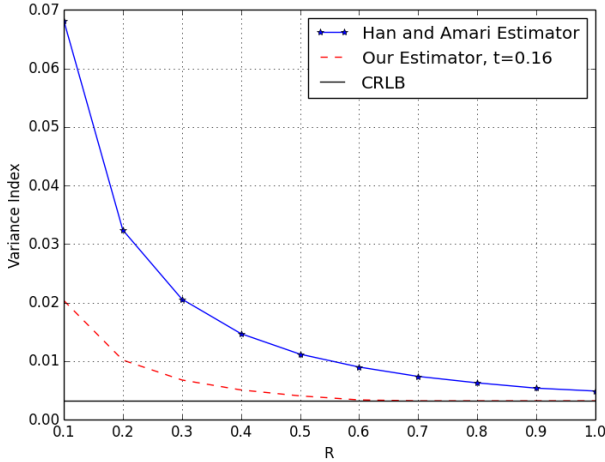Fig. 8: Performance Comparison: $\theta = 0.05$, $M = 2$, $t = 0.5$ and 0.1

Fig. 9: Performance Comparison: $\theta = 0.01$, $M = 4$, $t = 0.16$

VIII. CONCLUSION

In this paper, we have answered the question: Are Slepian-Wolf rates necessary to achieve the same estimation performance as that of the centralized case? We have showed that the answer to this question is negative by constructing an asymptoticly MVUE for binary symmetric sources using rates less than the conditions required in the Slepian-Wolf rate region. We have showed that our estimation algorithm can work for general binary sources to achieve the centralized estimation performance. We have also extended our work to a general class of non-binary information sources by modifying our estimation algorithm. We have further proposed a practical design of our estimation algorithm and compared our results to the best known estimator by Han and Amari to show the superiority of our estimator.

APPENDIX A
ERROR PROBABILITY ANALYSIS IN THE PROOF OF THEOREM 1

To analyze the probability of the decoding error, let $\tilde{z}^n \in \{0,1\}^n$ denote a sequence such that

$$\tilde{z}^n \neq z^n, \qquad f(\tilde{z}^n) = f(z^n). \tag{82}$$

Let $\tilde{Z}^{(n)}$ be a dummy random variable whose PMF $P_{\tilde{Z}^{(n)}}$ is the same as the type of $\tilde{z}^n$. Define $\mathcal{P}_{Z\tilde{Z}}^{(n)}$ as the set of all joint types between any two sequences z^n and $\tilde{z}^n$. For any given f (equivalently for a given encoding matrix A), define $N_f^n(Z\tilde{Z})$ as the number of sequences z^n such that there exists another sequence $\tilde{z}^n$ having the joint type $P_{Z^{(n)}\tilde{Z}^{(n)}} \in \mathcal{P}_{Z\tilde{Z}}^{(n)}$ and (82) holds.

Since each entry in A is uniformly distributed, then each element in $f(z^n)$ is uniformly distributed if z^n is a nonzero sequence. Therefore,

$$\Pr(f(z^n) = 0) = (0.5)^{nR} = \frac{1}{||f||}, \tag{83}$$

in which the probability is computed over all codebooks. This implies that

$$\Pr(f(\tilde{z}^n) = f(z^n)) = \Pr(f(\tilde{z}^n - z^n) = 0) = \frac{1}{||f||}. \tag{84}$$

Define $T_{P_{Z^{(n)}\tilde{Z}^{(n)}}}$ as the set of all sequence pairs $(z^n, \tilde{z}^n)$ that have the joint type $P_{Z^{(n)}\tilde{Z}^{(n)}}$, $T_{P_{Z^{(n)}}}$ as the set of all sequences z^n that have the marginal type $P_{Z^{(n)}}$, and $T_{P_{\tilde{Z}^{(n)}|Z^{(n)}}}(z^n)$ as the set of all sequences $\tilde{z}^n$ that have the joint type $P_{Z^{(n)}\tilde{Z}^{(n)}}$ with z^n. The sizes of the sets $T_{P_{Z^{(n)}}}$ and $T_{P_{\tilde{Z}^{(n)}|Z^{(n)}}}(z^n)$ are bounded as [27]

$$||T_{P_{Z^{(n)}}}|| \leq 2^{nH(Z^{(n)})},$$
$$||T_{P_{\tilde{Z}^{(n)}|Z^{(n)}}}(z^n)|| \leq 2^{nH(\tilde{Z}^{(n)}|Z^{(n)})+\epsilon}, \tag{85}$$

where ϵ is an arbitrary small number. Notice that, for any given $P_{Z^{(n)}\tilde{Z}^{(n)}}$, $N_f^n(Z\tilde{Z})$ is a random variable (random over f) that can be expressed as

$$N_f^n(Z\tilde{Z}) = \sum_{z^n \in T_{P_{Z^{(n)}}}} \mathbf{1}\big(\exists \tilde{z}^n \neq z^n : f(\tilde{z}^n) = f(z^n),$$
$$\text{and } (z^n, \tilde{z}^n) \in T_{P_{Z^{(n)}\tilde{Z}^{(n)}}}\big)$$
$$= \sum_{z^n \in T_{P_{Z^{(n)}}}} \mathbf{1}\big(\exists \tilde{z}^n \neq z^n : f(\tilde{z}^n) = f(z^n),$$
$$\text{and } \tilde{z}^n \in T_{P_{\tilde{Z}^{(n)}|Z^{(n)}}}(z^n)\big), \tag{86}$$

where $\mathbf{1}(\cdot)$ is the indication function. The expectation of $N_f^n(Z\tilde{Z})$ over all possible codebooks f is

$$\mathbb{E}[N_f^n(Z\tilde{Z})]$$
$$= \sum_{z^n \in T_{P_{Z^{(n)}}}} \mathbb{E}\big[\mathbf{1}\big(\exists \tilde{z}^n \neq z^n : f(\tilde{z}^n) = f(z^n),$$
$$\text{and } \tilde{z}^n \in T_{P_{\tilde{Z}^{(n)}|Z^{(n)}}}(z^n)\big)\big]$$
$$\leq \sum_{z^n \in T_{P_{Z^{(n)}}}} \sum_{\tilde{z}^n \in T_{P_{\tilde{Z}^{(n)}|Z^{(n)}}}(z^n)} \Pr(f(\tilde{z}^n) = f(z^n)). \tag{87}$$

(84), (85), and (87) imply that

$$\mathbb{E}[N_f^n(Z\tilde{Z})] \leq \frac{2^{n(H(Z^{(n)})+H(\tilde{Z}^{(n)}|Z^{(n)})+\epsilon)}}{||f||}. \tag{88}$$

Applying the Markov's inequality, we have

$$\Pr\left(N_f^n(Z\tilde{Z}) \geq \frac{2^{n(H(Z^{(n)})+H(\tilde{Z}^{(n)}|Z^{(n)})+\epsilon)}(||\mathcal{P}_{Z\tilde{Z}}^{(n)}||+\delta)}{||f||}\right)$$
$$\leq \frac{1}{||\mathcal{P}_{Z\tilde{Z}}^{(n)}||+\delta}, \tag{89}$$

where $||\mathcal{P}_{Z\tilde{Z}}^{(n)}||$ is the total number of possible joint types and δ is an arbitrary small number. To simplify the notation, let

$$B^n(Z\tilde{Z}) \triangleq \frac{2^{n(H(Z^{(n)})+H(\tilde{Z}^{(n)}|Z^{(n)})+\epsilon)}(||\mathcal{P}_{Z\tilde{Z}}^{(n)}||+\delta)}{||f||}. \tag{90}$$

Considering all joint types $P_{Z^{(n)}\tilde{Z}^{(n)}}$ simultaneously, the union bound and (89) imply that

$$\Pr\left(N_f^n(Z\tilde{Z}) \leq B^n(Z\tilde{Z}), \ \forall P_{Z^{(n)}\tilde{Z}^{(n)}} \in \mathcal{P}_{Z\tilde{Z}}^{(n)}\right)$$
$$\geq 1 - \sum_1^{||\mathcal{P}_{Z\tilde{Z}}^{(n)}||} \frac{1}{||\mathcal{P}_{Z\tilde{Z}}^{(n)}||+\delta}$$
$$> 0. \tag{91}$$

Since the probability in (91) is positive, then there exists a codebook f^* that the following equation holds for all joint types $P_{Z\tilde{Z}}$ simultaneously

$$N_{f^*}^n(Z\tilde{Z}) \leq \frac{2^{n(H(Z^{(n)})+H(\tilde{Z}^{(n)}|Z^{(n)})+\epsilon)}(||\mathcal{P}_{Z\tilde{Z}}^{(n)}||+\delta)}{||f^*||}. \tag{92}$$

As $||f^*|| = 2^{nR}$ and $||\mathcal{P}_{Z\tilde{Z}}^{(n)}|| \leq (n+1)^4$, we further have

$$N_{f^*}^n(Z\tilde{Z}) \tag{93}$$
$$\leq ((n+1)^4+\delta) \ 2^{n(H(Z^{(n)})+H(\tilde{Z}^{(n)}|Z^{(n)})+\epsilon-R)}.$$

In the following, we will focus on f^*.

Let $P_{e,f^*}^{(n)}(Z\tilde{Z})$ denote the portion of error probability associated with a fixed joint type $P_{Z^{(n)}\tilde{Z}^{(n)}}$

$$P_{e,f^*}^{(n)}(Z\tilde{Z}) \tag{94}$$
$$\triangleq \sum_{z^n \in T_{P_{Z^{(n)}}}} P_\theta(z^n)\mathbf{1}\big(\exists \tilde{z}^n \neq z^n : f^*(\tilde{z}^n) = f^*(z^n),$$
$$\text{and } (z^n, \tilde{z}^n) \in T_{P_{Z^{(n)}\tilde{Z}^{(n)}}}\big).$$

The total decoding error probability $P_{e,f^*}^{(n)}$, when using f^*, can be expressed as

$$P_{e,f^*}^{(n)} = \sum_{P_{Z^{(n)}\tilde{Z}^{(n)}}} P_{e,f^*}^{(n)}(Z\tilde{Z}). \tag{95}$$

Let $A_{\epsilon_1}^{(n)}$ denote the set of marginal types $P_{Z^{(n)}}$ such that $|P_{Z^{(n)}}(z=i) - P_\theta(z=i)| < \frac{\epsilon_1}{2}$ for $i \in \{0,1\}$, where ϵ_1 is

an arbitrarily small number. Using the definition of $A_{\epsilon_1}^{(n)}$, (95) can be rewritten as

$$P_{e,f^*}^{(n)} = \sum_{P_{Z^{(n)}\tilde{Z}^{(n)}}, P_{Z^{(n)}} \in A_{\epsilon_1}^{(n)}} P_{e,f^*}^{(n)}(Z\tilde{Z})$$
$$+ \sum_{P_{Z^{(n)}\tilde{Z}^{(n)}}, P_{Z^{(n)}} \in \bar{A}_{\epsilon_1}^{(n)}} P_{e,f^*}^{(n)}(Z\tilde{Z})$$
$$\triangleq S_1 + S_2, \tag{96}$$

where $\bar{A}_{\epsilon_1}^{(n)}$ denotes the complimentary set of $A_{\epsilon_1}^{(n)}$. For S_2, we have that

$$P_{e,f^*}^{(n)}(Z\tilde{Z}) \leq 2^{-n(D(P_{Z^{(n)}}||P_\theta(Z)))}, \tag{97}$$

where $D(P_{Z^{(n)}}||P_\theta(Z))$ is the KullbackLeibler divergence [25] between the marginal type $P_{Z^{(n)}}$ and the true PMF $P_\theta(Z)$ of $Z = X \oplus Y$. Using Pinsker's inequality [28], for $P_{Z^{(n)}} \in \bar{A}_{\epsilon_1}^{(n)}$, we have

$$D(P_{Z^{(n)}}||P_\theta(Z)) \geq 2\epsilon_1^2. \tag{98}$$

Therefore,

$$S_2 \leq \sum_{P_{Z^{(n)}\tilde{Z}^{(n)}}} 2^{-2n\epsilon_1^2}$$
$$\leq (n+1)^4 \ 2^{-2n\epsilon_1^2}. \tag{99}$$

(99) implies that $S_2 \to 0$ exponentially as $n \to \infty$.

For S_1, we have that

$$P_{e,f^*}^{(n)}(Z\tilde{Z}) \leq N_{f^*}^n(Z\tilde{Z}) \ 2^{-n(H(Z^{(n)})+D(P_{Z^{(n)}}||P_\theta(Z))}. \tag{100}$$

Using (93), we further have

$$P_{e,f^*}^{(n)}(Z\tilde{Z}) \leq \tag{101}$$
$$((n+1)^4+\delta) \ 2^{-n\big(D(P_{Z^{(n)}}||P_\theta(Z))+R-H(\tilde{Z}^{(n)}|Z^{(n)})-\epsilon\big)}.$$

As we use the minimum entropy decoder, we have $H(\tilde{Z}^{(n)}) \leq H(Z^{(n)})$, which implies $H(\tilde{Z}^{(n)}|Z^{(n)}) \leq H(\tilde{Z}^{(n)}) \leq H(Z^{(n)})$. Therefore,

$$P_{e,f^*}^{(n)}(Z\tilde{Z}) \tag{102}$$
$$\leq ((n+1)^4+\delta) \ 2^{-n\big(D(P_{Z^{(n)}}||P_\theta(Z))+R-H(Z^{(n)})-\epsilon\big)}.$$

Since $P_{Z^{(n)}} \in A_{\epsilon_1}^{(n)}$, it is easy to check that

$$|H(Z^{(n)}) - H_\theta(Z)| \leq D(P_{Z^{(n)}}||P_\theta(Z)) + \epsilon_2. \tag{103}$$

Here

$$\epsilon_2 = -\frac{\epsilon_1}{2}\sum_i \log P_\theta(z=i), \tag{104}$$

which can be made arbitrarily small as $\epsilon_1 \downarrow 0$ for $\theta \in (0,1)$.

Therefore,

$$P_{e,f^*}^{(n)}(Z\tilde{Z}) \leq ((n+1)^4+\delta) \ 2^{-n\big(R-H_\theta(Z)-\epsilon_3\big)}, \tag{105}$$

in which $\epsilon_3 = \epsilon + \epsilon_2$.

This implies that $S_1 \to 0$ exponentially as $n \to \infty$ if

$$R \geq H_\theta(Z). \tag{106}$$

Therefore, (106) is sufficient to guarantee that $P_{e,f^*}^{(n)} \to 0$ exponentially as $n \to \infty$. It is easy to check that $H_\theta(Z) = H_\theta(X|Y) = H_\theta(Y|X)$. The proof is complete.

REFERENCES

[1] M. ElGamal and L. Lai, "Are Slepian-Wolf rates necessary for distributed parameter estimation?," in *Proc. Allerton Conf. on Communication, Control, and Computing*, (Monticello, IL), pp. 1249–1255, Oct. 2015.

[2] G. Xu, S. Zhu, and B. Chen, "Decentralized data reduction with quantization constraints," *IEEE Trans. Signal Processing*, vol. 62, pp. 1775–1784, Apr. 2014.

[3] A. Vempaty, H. He, B. Chen, and P. K. Varshney, "On quantizer design for distributed Bayesian estimation in sensor networks," *IEEE Trans. Signal Processing*, vol. 62, pp. 5359–5369, Oct. 2014.

[4] J. Zhu, X. Lin, R. S. Blum, and Y. Gu, "Parameter estimation from quantized observations in multiplicative noise environments," *IEEE Trans. Signal Processing*, vol. 63, pp. 4037–4050, Aug. 2015.

[5] J. Zhang, R. S. Blum, X. Lu, and D. Conus, "Asymptotically optimum distributed estimation in the presence of attacks," *IEEE Trans. Signal Processing*, vol. 63, pp. 1086–1101, Mar. 2015.

[6] S. Kar, H. Chen, and P. K. Varshney, "Optimal identical binary quantizer design for distributed estimation," *IEEE Trans. Signal Processing*, vol. 60, pp. 3896–3901, Jul. 2012.

[7] J. Fang and H. Li, "Optimal/near-optimal dimensionality reduction for distributed estimation in homogeneous and certain inhomogeneous scenarios," *IEEE Trans. Signal Processing*, vol. 58, pp. 4339–4353, Aug. 2010.

[8] F. S. Cattivelli and A. H. Sayed, "Diffusion LMS strategies for distributed estimation," *IEEE Trans. Signal Processing*, vol. 58, pp. 1035–1048, Oct. 2009.

[9] S. Cui, J.-J. Xiao, A. J. Goldsmith, Z.-Q. Luo, and H. V. Poor, "Estimation diversity and energy efficiency in distributed sensing," *IEEE Trans. Signal Processing*, vol. 55, pp. 4683–4695, Sep. 2007.

[10] M. Raginsky, "Learning from compressed observations," in *Proc. IEEE Inform. Theory Workshop*, (Tahoe City, CA), pp. 420–425, Sep. 2007.

[11] A. Xu and M. Raginsky, "Converses for distributed estimation via strong data processing inequalities," in *Proc. IEEE Intl. Symposium on Inform. Theory*, (Hong Kong, China), Jun. 2015.

[12] Y. Zhang, J. Duchi, M. I. Jordan, and M. J. Wainwright, "Information-theoretic lower bounds for distributed statistical estimation with communication constraints," in *Advances in Neural Information Processing Systems*, (Stateline, NV), pp. 2328–2336, Dec. 2013.

[13] O. Shamir and N. Srebro, "Distributed stochastic optimization and learning," in *Proc. Allerton Conf. on Communication, Control, and Computing*, (Monticello, IL), pp. 850–857, Oct. 2014.

[14] J.-J. Xiao, S. Cui, Z.-Q. Luo, and A. J. Goldsmith, "Power scheduling of universal decentralized estimation in sensor networks," *IEEE Trans. Signal Processing*, vol. 54, pp. 413–422, Feb. 2006.

[15] A. Ribeiro and G. B. Giannakis, "Bandwidth-constrained distributed estimation for wireless sensor networks, Part I: Gaussian case," *IEEE Trans. Signal Processing*, vol. 54, pp. 1131–1143, Mar. 2006.

[16] A. Ribeiro and G. B. Giannakis, "Bandwidth-constrained distributed estimation for wireless sensor networks, Part II: Unknown probability density function," *IEEE Trans. Signal Processing*, vol. 54, pp. 2784–2796, Jul. 2006.

[17] J. Li and G. AlRegib, "Rate-constrained distributed estimation in wireless sensor networks," *IEEE Trans. Signal Processing*, vol. 55, pp. 1634–1643, May. 2007.

[18] D. Slepian and J. K. Wolf, "Noiseless coding of correlated information sources," *IEEE Trans. Inform. Theory*, vol. 19, pp. 471–480, Nov. 1973.

[19] I. Csiszár, "Linear codes for sources and source networks: Error exponents, universal coding," *IEEE Trans. Inform. Theory*, vol. 28, pp. 585–592, Jul. 1982.

[20] A. Zia, J. P. Reilly, and S. Shirani, "Distributed estimation; three theorems," in *Proc. IEEE Inform. Theory Workshop*, (Tahoe City, CA), pp. 517–522, Sep. 2007.

[21] T. S. Han and S.-I. Amari, "Parameter estimation with multiterminal data compression," *IEEE Trans. Inform. Theory*, vol. 41, pp. 1802–1833, Nov. 1995.

[22] R. Ahlswede and I. Csiszár, "To get a bit of information may be as hard as to get full information," *IEEE Trans. Inform. Theory*, vol. 27, pp. 398–408, Jul. 1981.

[23] H. V. Poor, *An introduction to signal detection and estimation*. New York: Springer Science & Business Media, 2013.

[24] Z. Zhang and T. Berger, "Estimation via compressed information," *IEEE Trans. Inform. Theory*, vol. 34, pp. 198–211, Mar. 1988.

[25] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Wiley Series in Telecommunications and Signal Processing, Wiley-Interscience, 2006.

[26] J. Körner and K. Marton, "How to encode the modulo-two sum of binary sources," *IEEE Trans. Inform. Theory*, vol. 25, pp. 219–221, Mar. 1979.

[27] I. Csiszár, "The method of types," *IEEE Trans. Inform. Theory*, vol. 44, pp. 2505–2523, Oct. 1998.

[28] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Cambridge University Press, 2011.