A Few Interactions Improve Distributed Nonparametric Estimation, Optimally

Jingbo Liu, Member, IEEE

Abstract—Consider the problem of nonparametric estimation of an unknown $\beta$-Hölder smooth density $p_{XY}$ at a given point, where $X$ and $Y$ are both $d$ dimensional. An infinite sequence of i.i.d. samples $(X_i, Y_i)$ are generated according to this distribution, and two terminals observe $(X_i)$ and $(Y_i)$, respectively. They are allowed to exchange $k$ bits either in one-way or interactively in order for Bob to estimate the unknown density. We show that the minimax mean square risk is order $(k/\log k)^{-\frac{2\beta}{d+2\beta}}$ for one-way protocols and $k^{-\frac{2\beta}{d+2\beta}}$ for interactive protocols. The logarithmic improvement is nonexistent in the parametric counterparts, and therefore can be regarded as a consequence of nonparametric nature of the problem. Moreover, a few rounds of interactions achieve the interactive minimax rate; the number of rounds can grow as slowly as the super-logarithm (i.e., inverse tetration) of $k$. The proof of the upper bound is based on a novel multi-round scheme for estimating the joint distribution of a pair of biased Bernoulli variables, and the lower bound is built on a sharp estimate of a symmetric strong data processing constant for biased Bernoulli variables.

Index Terms—Density estimation, communication complexity, nonparametric statistics, learning with system constraints, strong data processing constant.

I. INTRODUCTION

The communication complexity problem was introduced in the seminal paper of Yao [50] (see also [26] for a survey), where two terminals (which we call Alice and Bob) compute a given Boolean function of their local inputs $(X = (X_i)_{i=1}^n)$ and $(Y = (Y_i)_{i=1}^n)$ by means of exchanging messages. The famous log-rank conjecture provides an estimate of the communication complexity of a general Boolean function, which is still open to date. Meanwhile, communication complexity of certain specific functions can be better understood. For example, the Gap-Hamming problem [12], [24] concerns testing $f(X, Y) > \frac{1}{2}$ against $f(X, Y) < -\frac{1}{2}$, where $f(X, Y) := \frac{1}{n} \sum_{i=1}^n X_i Y_i$ denotes the sample correlation and $X_i, Y_i \in \{+1, -1\}$. It was shown in [12] with a geometric argument that the communication complexity (for worst-case deterministic $X, Y$) is $\Theta(n)$, therefore a one-way protocol where Alice simply sends $X$ cannot be improved (up to a multiplicative constant) by an interactive protocol.

Gap-Hamming is closely related to the problem of estimating the joint distribution of a pair of binary or Gaussian random variables (using $n$ i.i.d. samples). Indeed, for $n$ large we may assume that Alice (resp. Bob) can estimate the marginal distributions of $X$ (resp. $Y$) very well, so that the joint distribution is parameterized by only one scalar which is the correlation. An information-theoretic proof of Gap-Hamming was previously provided in [19], building on a converse for correlation estimation for the binary symmetric distribution, and pinned down the exact prefactor in the risk-communication tradeoff. In particular, the result of [19] implies that the naive algorithm where Alice simply sends $X$ can be improved by a constant factor in the estimation risk by a more sophisticated scheme using additional samples. For the closely related problem of correlation (distribution) testing, [38] and [48] provided asymptotically tight bounds on the communication complexity under the one-way and interactive protocols when the null hypothesis is the independent distribution (zero correlation), which also implies that the error exponent can be improved by an algorithm using additional samples. The technique of [48] is based on the tensorization of internal and external information ([20] ahead), whereas the bound of [38] uses hypercontractivity. More recently, [20] derived bounds for testing against dependent distributions using optimal transport inequalities.

In this paper, we take the natural step of introducing nonparametric (NP) statistics to Alice and Bob, whereby two parties estimate a nonparametric density by means of sending messages interactively. It will be seen that this problem is closely related to a “sparse” version of the aforementioned Gap-Hamming problem, where interaction does help, in contrast to the usual Gap-Hamming problem.

For concreteness, consider the problem of nonparametric estimation of an unknown $\beta$-Hölder smooth density $p_{XY}$ at a given point $(x_0, y_0)$. For simplicity we assume the symmetric case where $X$ and $Y$ are both $d$ dimensional. An infinite sequence of i.i.d. samples $(X_i, Y_i)$ are generated according to $p_{XY}$, and Alice and Bob observe $(X_i)$ and $(Y_i)$, respectively. After they exchange $k$ bits (either in one-way or interactively), Bob estimates the unknown density at the given point. We successfully characterize the minimax rate in terms of the communication complexity $k$: it is order $(k/\log k)^{-\frac{2\beta}{d+2\beta}}$ for one-way protocols and $k^{-\frac{2\beta}{d+2\beta}}$ for interactive protocols.

© 2023 The Authors. This work is licensed under a Creative Commons Attribution 4.0 License.
For more information, see https://creativecommons.org/licenses/by/4.0/
Notably, allowing interaction strictly improves the estimation risk. Previously, separations between one-way and interactive protocols are known but in very different contexts: In [32, Corollary 1] (see also [31]), the separation was found in the rate region of common randomness generation from biased binary distributions, using certain convexity arguments, but this only implies a difference in the leading constant, rather than the asymptotic scaling. On the other hand, the example distribution in [42] is based on the pointer-chasing construction of [35], which appears to be a highly artificial distribution designed to entail a separation between the one-way and interactive protocols. Another example where interaction improves zero-error source coding with side information, based on a “bit location” algorithm, was described in [36], and it was shown that two-way communication complexity differs from interactivity communication complexity only by constant factors. In contrast, the logarithmic separation in the present paper arises from the nonparametric nature of the problem: If we consider the problem of correlation estimation for Bernoulli pairs with a fixed bias (a parametric problem), the risk will be order $k^{-\frac{1}{2}}$, and there will be no separation between one-way and interactive protocols (which is indeed the case in [19]).

In contrast, nonparametric estimation is analogous to Bernoulli correlation estimation where the bias changes with $k$ (since the optimal bandwidth adapts to $k$), which gives rise to the separation.

For the risk upper bound, in the one-way setting it is efficient for Alice to just encode the set of $i$’s such that $X_i$ falls within a neighborhood (computed by the optimal bandwidth for a given $k$) of the given point $x_0$. To achieve the optimal $k^{-\frac{1}{\alpha+2}}$ rate for interactive protocols, we provide a novel scheme that uses $r > 1$ rounds of interactions, where $r = r(k)$ grows as slowly as the super logarithm (i.e. the inverse of tetration) of $k$. With the sequence of $r(k)$ we use in Section V-C (and suppose that $\beta = d = 1$), while $r = 4$ interactions is barely enough for $k$ equal to the number of letters in a short sentence, $r = 8$ is more than sufficient for $k$ equal to the number of all elementary particles in the entire observable universe. Thus from a practical perspective, $r(k)$ is effectively a constant, although it remains an interesting theoretical question whether $r(k)$ really diverges (Conjecture 1).

For the lower bound, the proof is based on the symmetric data processing constant introduced in [32]. Previously, the data processing constant $s^*_{\epsilon}$ has been connected to two-party estimation and hypothesis testing in [19]; the idea was canonized as the following statement: “Information for hypothesis testing locally” is upper bounded by $s^*_{\epsilon}$ times “Information communicated mutually”. However, $s^*_{\epsilon}$ is not easy to compute, and previous bounds on $s^*_{\epsilon}$ are also not tight enough for our purpose. Instead, we first use an idea of simulation of continuous variables to reduce the problem to estimation of Bernoulli distributions, for which $s^*_{\epsilon}$ is easier to analyze. Then we use some new arguments to bound $s^*_{\epsilon}$. Let us emphasize that this paper concerns density estimation at a given point, rather than estimating the global density function. For the latter problem, it is optimal for Alice to just quantize the samples and send it to Bob, which we show in the companion paper [28]. The mean square error (in $\ell_2$ norm) of estimating global density function scales differently than the case of point-wise density estimation since the messages cannot be tailored to the given point.

A. Related Work

Besides function computation, distribution estimation and testing, other problems which have been studied in the communication complexity or privacy settings include lossy source coding [25] and common randomness or secret key generation [30], [32], [43], [47]. The key technical tool for interactive two-way communication models, namely the tensorization of internal and external information ([20] ahead), appeared in [25] for lossy compression, [9], [33] for function computation, [32] and [47] for common randomness generation, and [19], [48] for parameter estimation.

For one-way communication models, the main tool is a tensorization property related to the strong data processing constant (see (10) ahead), which was first used in [4] in the study of the error exponents in communication constrained hypothesis testing. The hypercontractivity method for single-shot bounds in one-way models was used in [27], [30] for common randomness generation and [38] for testing.

In statistics, communication-constrained estimation has received considerable attention recently, starting from [52], which considered a model where distributed samples are compressed and sent to a central estimator. Further works on this communication model include settings of Gaussian location estimation [10], [11], parametric estimation [22], nonparametric regression [53], Gaussian noise model [54], statistical inference [2], and nonparametric estimation [21] (with a bug fixed in [7]) [1]. Related problems solved using similar techniques include differential privacy [16] and data summarization [37], [46], [45]. Communication-efficient construction of test statistics for distributed testing using the divide-and-conquer algorithm is studied in [8]. Generally speaking, these works on statistical minimax rates concern the so-called horizontal partitioning of data sets, where data sets share the same feature space but differ in samples [18], [49]. In contrast, vertical distributed or federated learning, where data sets differ in features, has been used by corporations such as those in finance and medical care [18], [49]. It is worth mentioning that such horizontal partitioning model was also introduced in Yao’s paper [50] in the context of function computation under the name “simultaneous message model”, where different parties send messages to a referee instead of to each other. The direct sum property (similar to the tensorization property of internal and external information) of the simultaneous message model was discussed in [13].

B. Organization of the Paper

We review the background on nonparametric estimation, data processing constants and testing independence in Section II. The formulation of the two-party nonparametric estimation problem and the summary of main results are given in Section III. Section IV examines the problem of estimating a parameter in a pair of biased Bernoulli distributions, which
will be used as a building block in our nonparametric estimation algorithm. Section V proves some bounds on information exchanges, which will be the key auxiliary results for the proof of upper bound for Bernoulli estimation in Section VI, and for nonparametric estimation in Section VII. Finally, lower bounds are proved in Section VIII in the one-way case and in Section IX in the interactive case.

II. PRELIMINARIES

A. Notation

We use capital letters for probability measures and lower cases for the densities functions. We use the abbreviations 

\[ U_i := (U_{i1}, \ldots, U_{ir}) \]

\[ U^j := U^j \]

We use boldface letters to denote vectors, for example \( \mathbf{U} := (U_i(l)) \). Unless otherwise specified, the base of logarithms can be arbitrary but remain consistent throughout equations. The precise meaning of the Landau notations, such as \( O(\cdot) \), will be explained in each section or proofs of specific theorems. We use \( \sum_{i \leq S} \) to denote summing over \( i \in \{1, \ldots, r\} \setminus \mathbb{Z} \). For the vector representation of a binary probability distribution, we use the convention that \( P_U = [P_U(0), P_U(1)] \). For the matrix representation of a joint distribution of a pair of binary random variables, we use the convention that \( P_{XY} = \begin{pmatrix} P_{XY}(0,0) & P_{XY}(0,1) \\ P_{XY}(1,0) & P_{XY}(1,1) \end{pmatrix} \). For \( x \in [0,1] \), we use the shorthand \( \bar{x} := 1 - x \).

B. Nonparametric Estimation

Let us recall the basics about the problem of estimating a smooth density; more details may be found in [41] and [44].

Let \( d \geq 1 \) be an integer, and \( s = (s_1, \ldots, s_d) \in \{0,1,2,\ldots\}^d \) be a multi-index. For \( x = (x_1, \ldots, x_d) \in \mathbb{R}^d \), let \( D^s \) denote the differential operator

\[
D^s = \frac{\partial^{s_1+\cdots+s_d}}{\partial x_1^{s_1} \cdots \partial x_d^{s_d}}.
\]

Given \( \beta \in (0, \infty) \), let \( \lfloor \beta \rfloor \) be the maximum integer strictly smaller than \( \beta \) [44] (note the difference with the usual conventions). Given a function \( f \) whose domain includes a set \( A \subseteq \mathbb{R}^d \), define \( \|f\|_{A,\beta} \) as the minimum \( L \geq 0 \) such that

\[
|D^s f(x_1) - D^s f(x_2)| \leq L \|x_1 - x_2\|^{\lfloor \beta \rfloor - \lfloor \beta \rfloor}, \quad \forall x_1, x_2 \in A,
\]

for all multi-indices \( s \) such that \( s_1 + \cdots + s_d = \lfloor \beta \rfloor \). For example, \( \beta = 1 \) defines a Lipschitz function and an integer \( \beta \) defines a function with bounded \( \beta \)-th derivative.

Given \( L > 0 \), let \( P(\beta, L) \) be the class of probability density functions \( p \) satisfying \( \|p\|_{\mathbb{R}^d,\beta} \leq L \). Let \( x_0 \in \mathbb{R}^d \) be arbitrary. The following result on the minimax estimation error is well-known:

\[
\inf_{T_n} \sup_{p \in P(\beta, L)} \mathbb{E}[|T_n - p(x_0)|^2] = \Theta(n^{-\frac{2\beta}{\pi^2 \beta}})
\]

where the infimum is over all estimators \( T_n \) of \( p(x_0) \), i.e., measurable maps from i.i.d. samples \( X_1, \ldots, X_n \sim p \) to \( \mathbb{R} \). \( \Theta(\cdot) \) in (3) may hide constants independent of \( n \).

We say \( K: \mathbb{R}^d \to \mathbb{R} \) is a kernel of order \( l \) \((l \in \{1,2,\ldots\}) \) if \( \int K = 1 \) and all up to the \( l \)-th derivatives of the Fourier transform of \( K \) vanish at \( 0 \) [44, Definition 1.3]. Therefore the rectangular kernel, which is the indicator of a set, is order 1. A kernel estimator has the form

\[
T_n = \frac{1}{nh^n} \sum_{i=1}^n K \left( \frac{X_i - x_0}{h} \right)
\]

where \( h \in (0, \infty) \) is called bandwidth. If \( K \) is a kernel of order \( l = \lfloor \beta \rfloor \), then the kernel estimator (4) with appropriate \( h \) achieves the bound in (3) [44, Chapter 1]. In particular, the rectangular kernel is minimax optimal for \( \beta \in (0,2) \).

If \( K \) is compactly supported, then only local smoothness is needed, and density lower bound does not change the rate: we have

\[
\inf_{T_n} \sup_{p \in P_S(\beta, L, A)} \mathbb{E}[|T_n - p(x_0)|^2] = \Theta(n^{-\frac{2\beta}{\pi^2 \beta}})
\]

where \( S \) is any compact neighborhood of \( x_0 \), \( A \in [0, \frac{1}{\text{vol}(S)}) \) is arbitrary (with \( \text{vol}(S) \) denoting the volume of \( S \)), and \( P_S(\beta, L, A) \) denotes the non-empty set of probability density functions \( p \) satisfying \( \|p\|_{S,\beta} \leq L \) and \( \text{inf}_{x \in S} p(x) \geq A \).

C. Strong and Symmetric Data Processing Constants

The strong data processing constant has proved useful in many distributed estimation problems [4], [10], [16], [52]. In particular, it is strongly connected to two-party hypothesis testing under the one-way protocol. In contrast, the symmetric data processing constant [32] can be viewed as a natural extension to interactive protocols. This section recalls their definitions and auxiliary results, which will mainly be used in the proofs of lower bounds; however, the intuitions are useful for the upper bounds as well.

Given two probability measures \( P, Q \) on the same measurable space, define the KL divergence

\[
D(P||Q) := \int \log \left( \frac{dP}{dQ} \right) dP.
\]

Define the \( \chi^2 \)-divergence

\[
D_{\chi^2}(P||Q) := \int \left( \frac{dP}{dQ} - 1 \right)^2 dQ.
\]

Let \( X, Y \) be two random variables with joint distribution \( P_{XY} \). Define the mutual information

\[
I(X; Y) := D(P_{XY} \| P_X \times P_Y).
\]

Definition 1: Let \( P_{XY} \) be an arbitrary distribution on \( X \times Y \). Define the strong data processing constant

\[
s^*(X; Y) := \sup_{P_{UX}} \frac{I(U; Y)}{I(U; X)}
\]

where \( P_{UX} \) is the conditional distribution (with \( U \) being an arbitrary set), and the mutual informations are computed under the joint distribution \( P_{UXY} \).

Clearly, the value of \( s^*(X; Y) \) does not depend on the choice of the base of logarithm. A basic yet useful property
of the strong data processing constant is tensorization: if \((X, Y) \sim P_{XY}^m\) then
\[
s^*(X; Y) = s^*(X; Y).
\]
(10)

Now if \((X; Y)\) are the samples observed by Alice and Bob, \(\Pi_1\) denotes the message sent to Bob, then \(I(\Pi_1; X) \leq k\) implies that
\[
D(P_{\Pi_1, Y} \| P_{\Pi_1, P_Y}) \leq s^*(X; Y)k.
\]
(11)

The left side is the KL divergence between the distribution under the hypothesis that \((X, Y)\) follows some joint distribution, and the distribution under the hypothesis that \(X\) and \(Y\) are independent. Thus the error probabilities in testing against independence with one-way protocols can be lower bounded. This simple argument dates back at least to [3] and [4].

A similar argument can be extended to testing independence under interactive protocols [48]. The fundamental fact enabling such extensions is the tensorization of certain information-theoretic quantities, which appeared in various contexts [9], [25], [32].

Definition 2: Let \((X, Y) \sim P_{XY}\). For given \(r < \infty\), define \(s^*_r(X; Y)\) as the supremum of \(R/S\) such that there exists random variables \(U_1, \ldots, U_r\) satisfying
\[
R \leq \sum_{1 \leq i \leq r} I(U_i; Y|U_{i-1}) + \sum_{1 \leq i \leq r} I(U_i; X|U_{i-1});
\]
(12)
\[
S \geq \sum_{1 \leq i \leq r} I(U_i; X|U_{i-1}) + \sum_{1 \leq i \leq r} I(U_i; Y|U_{i-1}),
\]
(13)
and
\[
U_i - (X, U_{i-1}) - Y; \quad i \in \{1, \ldots, r\} \setminus 2\mathbb{Z}
\]
(14)
\[
U_i - (Y, U_{i-1}) - X; \quad i \in \{1, \ldots, r\} \cap 2\mathbb{Z}
\]
(15)
are Markov chains. We call \(s^*_\infty(X; Y)\) the symmetric data processing constant.

Let us remark that using the Markov chains we have the right side of (12)
\[
\sum_{1 \leq i \leq r} I(U_i; Y|U_{i-1}) + \sum_{1 \leq i \leq r} I(U_i; X|U_{i-1})
\]
(16)
\[
= I(X; Y) - I(X; Y|U^r)
\]
(17)
whereas the right side of (13)
\[
\sum_{1 \leq i \leq r} I(U_i; X|U_{i-1}) + \sum_{1 \leq i \leq r} I(U_i; Y|U_{i-1}) = I(U^r; XY).
\]
(18)

In the computer science literature [9], \(I(U^r; XY)\) is called the external information whereas \(I(U^r; X|Y) + I(U^r; Y|X)\) the internal information.

The symmetric strong data processing constant is symmetric in the sense that \(s^*_\infty(X; Y) = s^*_\infty(Y; X)\), since \(r = \infty\) in the definition. On the other hand, \(s^*_r(X; Y)\) coincides with the strong data processing constant which is generally not symmetric. Furthermore, a tensorization property holds for the internal and external information: denote by \(\mathcal{R}(X; Y)\) the set of all \((R, S)\) satisfying (12) and (13) for some \(U_1, \ldots, U_r\). Let \((X, Y) \sim P_{XY}^m\). Then
\[
\mathcal{R}(X; Y) = n\mathcal{R}(X; Y).
\]
(19)

In particular, taking the slope of the boundary at the original yields
\[
s^*_\infty(X; Y) = s^*_\infty(X; Y).
\]
(20)

A useful and general upper bound on \(s^*_\infty\) in terms of SVD was provided in [32, Theorem 4], which implies that \(s^*_\infty = s^*_1\) when \(X\) and \(Y\) are unbiased Bernoulli. However, that bound is not tight enough for the nonparametric estimation problem we consider, and in fact we adopt a new approach in Section IX for the biased Bernoulli distribution. Let us remark that \(s^*_\infty = s^*_1\) holds also for Gaussian \((X, Y)\), which follows by combining the result on unbiased Bernoulli distribution and a central limit theorem argument [32] (see also [19]). Moreover, it was conjectured in [32] that the set of possible \((R, S)\) satisfying (12)-(13) does not depend on \(r\) when \(X\) and \(Y\) are unbiased Bernoulli.

D. Testing Against Independence

Consider the following setting: \(P_{XY}\) is an arbitrary distribution on \(X \times Y\); \(P_{XY} := P_{XY}^m\), \(\Pi = (\Pi_0, \ldots, \Pi_r)\) is a sequence of random variables, with \(P_{\Pi_0|XY}\) being given and satisfying \(P_{\Pi_0|XY} = P_{\Pi_0}, P_{\Pi_i|XY|U_{i-1}} = P_{\Pi_i|XY|U_{i-1}}^1\) for \(i \in \{1, \ldots, r\} \setminus 2\mathbb{Z}\) and \(P_{\Pi_i|XY|U_{i-1}} = P_{\Pi_i|XY|U_{i-1}}^1\) for \(i \in \{1, \ldots, r\} \cap 2\mathbb{Z}\); \(P_{XY} = P_X P_Y\) is the distribution under the hypothesis of independence, and \(P_{\Pi XY} := P_{\Pi|XY} P_{XY}\).

The following result is known in [19], [20], and [48]:

Lemma 1: \(D(P_{\Pi XY} \| P_{\Pi|Y}) \leq I(X; Y) - I(X; Y|\Pi)\). Now by Definition 2, we immediately have
\[
s^*_r(X; Y) \geq \frac{I(X, Y) - I(X, Y|\Pi)}{I(X; Y|\Pi)}
\]
(21)
\[
\geq \frac{D(P_{\Pi XY} \| P_{\Pi|Y})}{I(\Pi)}
\]
(22)
which generalizes (11). Therefore, \(s^*_r(X; Y)\) can be used to bound \(D(P_{\Pi XY} \| P_{\Pi|Y})\), and in turn, the error probability in independence testing.

III. PROBLEM SETUP AND MAIN RESULTS

We consider estimating the density function at a given point, where the density is assumed to be Hölder continuous in a neighborhood of that point. It is clear that there is no loss of generality assuming such neighborhood to be the unit cube, and that the given point is its center. More precisely, the class of densities under consideration is defined as follows:

Definition 3: Given \(d \in \{1, 2, \ldots\}, L > 0, A \in [0, 1), \) and \(\beta > 0\), let \(\mathcal{H}(\beta, L, A)\) be the set of all probability density \(p_{XY}\) on \(X \times Y\) (where \(X = Y = \mathbb{R}^d\)) satisfying
\[
p_X(x), p_Y(y) \geq A, \quad \forall x, y \in [0, 1]^d,
\]
(23)
and
\[ \|p_{XY}\|_{[0,1]^{2^d}} \leq L. \] (24)

**Definition 4:** We say \( \mathcal{C} \) is a prefix code [14] if it is a subset of the set of all finite non-empty binary sequences satisfying the property that for any distinct \( s_1, s_2 \in \mathcal{C} \), \( s_1 \) cannot be a prefix of \( s_2 \).

The problem is to estimate the density at a given point of an unknown distribution from \( \mathcal{H}(\beta, L, A) \). More precisely,

- \( p_{XY} \) is a fixed but unknown distribution whose corresponding density \( p_{XY} \) belongs to \( \mathcal{H}(\beta, L, A) \) for some \( \beta \in (0, \infty), L \in (0, \infty), \) and \( A \in [0,1] \).
- Infinite sequence of pairs \((X(1), Y(1)), (X(2), Y(2))\)… are i.i.d. according to \( p_{XY} \).
- Alice (Terminal 1) observes \( X = (X(l))_{l=1}^\infty \) and Bob (Terminal 2) observes \( Y = (Y(l))_{l=1}^\infty \).
- Unlimited common randomness \( \Pi_0 \) is observed by both Alice and Bob. That is, an infinite random bit string independent of \((X, Y)\) shared by Alice and Bob.
- For \( i = 1, \ldots, r \) (\( r \) is an integer), if \( i \) is odd, then Alice sends to Bob a message \( \Pi_i \), which is an element in a prefix code, where \( \Pi_i \) is computed using the common randomness \( \Pi_0 \), the previous transcripts \( \Pi_{i-1} \), and \( X_i \); if \( i \) is even, then Bob sends to Alice a message \( \Pi_i \) computed using \( \Pi_0, \Pi_{i-1} \), and \( Y_i \).
- Bob computes an estimate \( \hat{p} \) of the true density \( p_{XY}(x_0, y_0) \), where \( x_0 = y_0 \) is the center of \([0,1]^d\).

**One-Way NP Estimation Problem:** Suppose that \( r = 1 \). Under the constraint on the expected length of the transcript (i.e. length of the bit string)
\[ E[|\Pi'|] \leq k, \] (25)
where \( k > 0 \) is a real number, what is the minimax risk
\[ R(k) := \min_{\hat{p} : H(\hat{p}) \in \mathcal{H}(\beta, L, A)} \max_{p_{XY} \in \mathcal{H}(\beta, L, A)} \mathbb{E}[|\hat{p} - p_{XY}(x_0, y_0)|^2]? \] (26)

**Interactive NP Estimation Problem.** Under the same constraint on the expected length of the transcript, but without any constraint on the number of rounds \( r \), what is the minimax risk?

**Remark 1:** The prefix condition ensures that Bob knows that the current round has terminated after finishing reading each \( \Pi_i \). Alternatively, the problem can be formulated by stating that \( \Pi_i \) is a random variable in an arbitrary alphabet, and replacing (25) by the entropy constraint \( H(\Pi') \leq k \). Furthermore, one may use the information leakage constraint \( I(X, Y; \Pi') \leq k \) instead. From our proofs it is clear that the minimax rates will not change under these alternative formulations.

**Remark 2:** There would be no essential difference if the problem were formulated with \( |\Pi| \leq k \) almost surely and \( |\hat{p} - p_{XY}(x_0, y_0)|^2 \leq R(k) \) with probability (say) at least \( 1/2 \). Indeed, for the upper bound direction, those conditions are satisfied with a truncation argument, once we have an algorithm satisfying \( \mathbb{E}[|\Pi'|] \leq k/4 \) and \( \mathbb{E}[|\hat{p} - p_{XY}(x_0, y_0)|^2] \leq R(k)/4 \), by Markov’s inequality and the union bound, therefore results only differ with a constant factor. For the lower bound, the proof can be extended to the high probability version, since we used the Le Cam style argument [51].

**Remark 3:** The common randomness assumption is common in the communication complexity literature, and, in some sense, is equivalent to private randomness [34]. In our upper bound proof, the common randomness is the randomness in the codebooks. Random codebooks give rise to convenient properties, such as the fact that the expectation of the distribution of the matched codewords equals exactly the product of idealized single-letter distributions (78). It is likely, however, that some approximate versions of these proofs steps, and ultimately the same asymptotic risk, should hold for some carefully designed deterministic codebooks.

**Theorem 1:** In one-way NP estimation, for any \( \beta \in (0, \infty), L \in (0, \infty), \) and \( A \in [0,1] \),
\[ R(k) = \Theta\left(\frac{k}{\log k}\right)^{-\frac{2\beta}{\pi^2+\beta}} \] (27)
where \( \Theta(\cdot) \) hides multiplicative factors depending on \( L, \beta, \) and \( A \).

The proof of the upper bound is in Section VII-B. Recall that nonparametric density estimation using a rectangular kernel is equivalent to counting the frequency of samples in a neighborhood of a given diameter, the bandwidth, which we denote as \( \Delta \). A naive protocol is for Alice to send the indices of samples in \( x_0 + [-\Delta, \Delta]^d \). Locating each sample in that neighborhood requires on average \( \Theta(\log \frac{1}{\Delta}) = \Theta(\log k) \) bits. Thus \( \Theta(k/\log k) \) samples in that neighborhood can be located. It turns out that the naive protocol is asymptotically optimal.

The proof of the lower bound (Section VIII) follows by a reduction to testing independence for biased Bernoulli distributions, via a simulation argument. Although some arguments are similar to [19], the present problem concerns biased Bernoulli distributions instead. The (KL) strong data processing constant turns out to be drastically different from the \( \chi^2 \) data processing constant, as opposed to the cases of many familiar distributions such as the unbiased Bernoulli or the Gaussian distributions.

As alluded, our main result is that the risk can be strictly improved when interactions are allowed:

**Theorem 2:** In interactive NP estimation, for any \( \beta \in (0, \infty), L \in (0, \infty), \) and \( A \in [0,1] \), we have
\[ R(k) = \Theta\left(k^{-\frac{2\beta}{\pi^2+\beta}}\right) \] (28)
where \( \Theta(\cdot) \) hides multiplicative factors depending on \( L, \beta, \) and \( A \).

To achieve the scaling in (28), \( r \) can grow as slowly as the super-logarithm (i.e., inverse tetration) of \( k \); for the precise relation between \( r \) and \( k \), see Section V-C.

The proof of the upper bound of Theorem 2 is given in Section VII-C, which is based on a novel multi-round estimation scheme for biased Bernoulli distributions formulated and analyzed in Sections IV.V.VI. Roughly speaking, the intuition is to “locate” the samples within neighborhoods of \((x_0, y_0)\) by successive refinements, which is more communication-efficient than revealing the location at once.

The lower bound of Theorem 2 is proved in Section IX. The main technical hurdle is to develop new and tighter bounds on...
the symmetric data processing constant in [32] for the biased binary cases.

IV. ESTIMATION OF BIASED BERNOULLI DISTRIBUTIONS

In this section, we shall describe an algorithm for estimating the joint distribution of a pair of biased Bernoulli random variables. The biased Bernoulli estimation problem can be viewed as a natural generalization of the Gap hammer problem [12, 24] to the sparse setting, and is the key component in both the upper and lower bound analysis for the nonparametric estimation problem. Indeed, we shall explain in Section VII that our nonparametric estimator is based on a linear combination of rectangle kernel estimators, which estimate the probability that X and Y fall into neighborhoods of x0 and y0. Indicators that samples are within such neighborhoods are Bernoulli variables, so that the biased Bernoulli estimator can be used. For the lower bound, we shall explain in Section VIII that the nonparametric estimation problem can be reduced to the biased Bernoulli estimation problem via a simulation argument.

For notational simplicity, we shall use X, Y for the Bernoulli variables in this section as well as Sections V–VI, although we should keep in mind that these are not the continuous variables in the original nonparametric estimation problem.

Bernoulli Estimation Problem:
- Fixed real numbers \( m_1, m_2 \in (10, \infty) \), and an unknown \( \delta \in [-1, \min\{m_1, m_2\} - 1] \).
- \((X, Y) = (X(l), Y(l))_{l=1}^{\infty} \) i.i.d. according to the distribution
  \[
  P_{XY}(\delta) := \left( \frac{1}{m_1} \right) \left( \frac{1 + \delta}{m_2} \right) \left( \frac{1 - \delta}{m_1 m_2} \right)
  \]
  \[
  \left( \frac{1 - \delta}{m_1 m_2} \right) \left( \frac{1 - \delta}{m_1 m_2} \right)
  \]
  \[
  \left( \frac{1}{m_1} \right) \left( \frac{1 - \delta}{m_2} \right) \left( \frac{1 - \delta}{m_1 m_2} \right)
  \]
  where we recall our convention that the upper left entry of the matrix denotes the probability that \( X = Y = 0 \).
- Alice observes \( (X(l), Y(l))_{l=1}^{\infty} \) and Bob observes \( Y(l)_{l=1}^{\infty} \).
- Unlimited common randomness \( \Pi_0 \).

Goal: Alice and Bob exchange messages in no more than \( r \) rounds in order to estimate \( \delta \).

Our algorithm is described as follows:

Input: \( m_1, m_2 \in (10, \infty) \); positive integer \( n \) and \( r \); a sequence of real numbers \( \alpha_1, \ldots, \alpha_r \in (1, \infty) \) satisfying

\[
\prod_{1 \leq i \leq r} \frac{\alpha_i}{10} \leq \frac{m_1}{10},
\]

\[
\prod_{1 \leq i \leq r} \frac{\alpha_i}{10} \leq \frac{m_2}{10},
\]

The \( \alpha_1, \ldots, \alpha_r \) can be viewed as parameters of the algorithm, and controls how much information is revealed about the locations of “common zeros” of X, Y in each round of communication. For example, setting \( \alpha_1 = \frac{m_1}{10} \) and all other \( \alpha_i = 1 \) yields a one-way communication protocol, whereas setting all \( \alpha_i > 1 \) yields a “successive refinement” algorithm which may incur smaller communication budget yet convey the same amount of information.

Before describing the algorithm, let us define a conditional distribution \( P_{U|X} \) by recursion, which will be used later in generation random codebooks.

Definition 5: For each \( i \in \{1, \ldots, r\} \setminus 2Z \), define

\[
P_{U_i|X=0, U^{i-1}_j=0} = [1, 0];
\]

\[
P_{U_i|X=1, U^{i-1}_j=0} = [\alpha_i^{-1}, 1 - \alpha_i^{-1}];
\]

\[
P_{U_i|X=0, U^{i-1}_j\neq 0} = P_{U_i|X=1, U^{i-1}_j\neq 0} = [0, 1].
\]

Then set \( P_{U_i|XU^{i-1}} = P_{U_i|XU^{i-1}} \). For \( i = 1, \ldots, r \) even, we use similar definitions, but with the roles of X and Y switched. This specifies \( P_{U_i|XU^{i-1}}, i = 1, \ldots, r \).

Note that by Definition 5, \( U_1 = 1 \) implies \( U_1 = 1 \) for each \( i = 1, \ldots, r - 1 \). In words, for \( i \) odd, \( U_1 \) marks all X = 0 as 0, and marks X = 1 as either 0 or 1; whenever \( U_i = 1 \) is marked, then X is definitely 1, and will be forgotten in all subsequent rounds. Now set

\[
P_{XYU}^{(i)} := P_{U|X} P_{XYU}^{(i)},
\]

where \( P_{U|X} \) is induced by \( (P_{U_i|XU^{i-1}})_{i=1}^{r} \) in Definition 5.

Initialization: By applying a common function to the common randomness, Alice and Bob can produce a shared infinite array \( (V_{i,j}(l)) \), where \( i \in \{1, \ldots, r\}, j \in \{1, 2, \ldots, m\}, l \in \{1, 2, \ldots, n\} \), such that the entries in the array are independent random variables, with \( V_{i,j}(l) \sim \text{Bern}(1 - \alpha_i^{-1}) \). Also set

\[
U_0(l) = 0, \quad \forall l = 1, \ldots, n.
\]

Iterations: Consider any \( i = 1, \ldots, r, \) where \( i \) is odd. We want to generate \( U_i \) by selecting a codeword so that \( (X, Y, U^i) \) follows the distribution of \( (P_{XYU}^{(i)} )_{i=1}^{\infty} \), where \( U^{i-1} \) is defined in previous rounds. Define

\[
A_0 := \{ l \leq n : X(l) = 0, U_{i-1}(l) = 0 \};
\]

\[
A_i := \{ l \leq n : X(l) = 1, U_{i-1}(l) = 0 \};
\]

\[
A := \{ l \leq n : U_{i-1}(l) = 0 \}.
\]

Note that Alice knows both \( A_0 \) and \( A_1 \), while Bob knows \( A \), since it will be seen from the recursion that Alice and Bob both know \( U_{1, \ldots, i-1} \) at the beginning of the i-th round. Alice chooses \( j_i \) as the minimum nonnegative integer \( j \) such that

\[
V_{i,j}(l) = 0, \quad \forall l \in A_0.
\]

Alice encodes \( j_i \) using a prefix code, e.g. Elias gamma code [17], and sends it to Bob. Then both Alice and Bob compute \( U_i = (U_i(l))_{l=1}^{n} \in \{0, 1\}^n \) by

\[
U_i(l) := V_{(i,j_i)}(l), \quad \forall l \in A;
\]

\[
U_i(l) := 1, \quad \forall l \in \{1, \ldots, n\} \setminus A.
\]

The operations in the i-th round for even i is similar, with the roles of Alice and Bob reversed. We will see later that the notation \( U_i \) is consistent in the sense of (49).

Estimator: Recall that in classical parametric statistics, one can evaluate the score function at the sample, compute its expectation and variation, and construct an estimator
achieving the Cramer-Rao bound asymptotically. Now for $i \in \{1, \ldots, r\} \setminus 2\mathbb{Z}$, define the score function
\begin{equation}
\Gamma_i(u^i, y) := \frac{\partial}{\partial y} \ln P_{U_i|YU_{i-1}}(u_i|y, u^{i-1}) \bigg|_{\delta = 0}
\end{equation}
respectively. Finally, Alice's and Bob's estimators are given
\begin{equation}
\Gamma_A := \sum_{1 \leq i \leq r} \sum_{l=1}^{n} \Gamma_i(U^i(l), X(l))
\end{equation}
and
\begin{equation}
\Gamma_B := \sum_{1 \leq i \leq r} \sum_{l=1}^{n} \Gamma_i(U^i(l), Y(l))
\end{equation}
respectively. Finally, Alice’s and Bob’s estimators are given by
\begin{equation}
\delta^A := \Gamma^A \cdot \left( \partial_\delta \bE^{(\delta)}[\Gamma^A] \right)^{-1}
\end{equation}
\begin{equation}
\delta^B := \Gamma^B \cdot \left( \partial_\delta \bE^{(\delta)}[\Gamma^B] \right)^{-1}
\end{equation}
where $\bE^{(\delta)}$ refers to expectation when the true parameter is $\delta$, and $\partial_\delta$ denotes the derivative in $\delta$. We will show that these estimators are well-defined: $\partial_\delta \bE^{(\delta)}[\Gamma^A]$ and $\partial_\delta \bE^{(\delta)}[\Gamma^B]$ are independent of $\delta$ (Lemma 3), and can be computed by Alice and Bob without knowing $\delta$.

**Lemma 2**: For each $i \in \{1, \ldots, r\} \setminus 2\mathbb{Z}$, conditioned on $X, Y, U_1, \ldots, U_{i-1}$, we have
\begin{equation}
U_i \sim P_{U_i|XYU_{i-1}}(\cdot|X, U_1, \ldots, U_{i-1})
\end{equation}
where $P_{U_i|XYU_{i-1}}$ is as defined in (32)-(34). A similar relation holds for even $i$.

**Proof**: Immediate from (41)-(42). □

**Lemma 3**: $\bE^{(\delta)}[\Gamma^A]$ and $\bE^{(\delta)}[\Gamma^B]$ are linear in $\delta$.

**Proof**: By (49),
\begin{equation}
\bE^{(\delta)}[\Gamma^A(U^i(l)), Y(l)] = \sum_{u^i, y} \Gamma_i(u^i, y) P_{X|YU_{i-1}}(u^i|y) P_{Y|U_{i-1}}(y|u^i)
\end{equation}
for each $i$ odd and $l \in \{1, \ldots, n\}$, and similar expressions hold for $i$ even. The claims then follow.

**Theorem 3**: $\delta^A$ and $\delta^B$ are unbiased estimators.

**Proof**: For $i$ odd, by (43) we have $\sum_{u_i} \Gamma_i(u_i, y) P_{U_i|YU_{i-1}}(u_i|y, u^{i-1}) = 0$ for any $(y, u^{i-1})$. Then $\bE^{(\delta)}[\Gamma^A(U^i(l)), Y(l)] = 0$ follows from (51). It follows that $\bE^{(0)}[\delta^A] = \bE^{(0)}[\delta^B] = 0$, and unbiasedness is implied by Lemma 3. □

**V. BOUNDS ON INFORMATION EXCHANGES**

In this section we prove key auxiliary results that will be used in the upper bounds.

**A. General ($\alpha$)**

The following Theorem is crucial for the achievability part of the analysis of the Bernoulli estimation problem described in Section IV (and hence for the nonparametric estimation problem). Specifically, (52)-(53) bounds the communication from Alice to Bob and in reverse, and (54)-(55) bounds the information exchanged which, in turn, will bound the estimation risk via Fano’s inequality.

**Theorem 4**: Consider any $m_1, m_2 > 0$, $\alpha_1, \ldots, \alpha_r \in (1, \infty)$ satisfying (30)-(31), and $P_{U^r|XY}$ as in (35). We have
\begin{equation}
\sum_{1 \leq i \leq r} P_{U_i|U_{i-1}}(0, 0) \log \alpha_i \leq \frac{1}{m_1} \sum_{1 \leq i \leq r} \log \alpha_i \prod_{2 \leq j \leq i-1} \alpha_j^{-1}
\end{equation}
\begin{equation}
\sum_{1 \leq i \leq r} P_{U_i|U_{i-1}}(0, 0) \log \alpha_i \leq \frac{1}{m_2} \sum_{1 \leq i \leq r} \log \alpha_i \prod_{1 \leq j \leq i-1} \alpha_j^{-1}
\end{equation}
and assuming the natural base of logarithms,
\begin{equation}
\lim_{\delta \to 0} \delta^{-2} \sum_{1 \leq i \leq r} I(U_i; Y|U^{i-1}) \geq \frac{1}{5m_1^2 m_2} \prod_{1 \leq j \leq r} \alpha_j^{-1}
\end{equation}
\begin{equation}
\lim_{\delta \to 0} \delta^{-2} \sum_{1 \leq i \leq r} I(U_i; X|U^{i-1}) \geq \frac{1}{5m_1 m_2^2} \prod_{1 \leq j \leq r} \alpha_j
\end{equation}
The proof can be found in Appendix A.

**Remark 4**: Since
\begin{equation}
P_{X|U_{i-1}}(0, 0) \log \alpha_i
= \max_{Q} P_{X|U_{i-1}}(0, 0) D(P_{U_i|X=0, U^{i-1}=0} || P_{U_i|X=1, U^{i-1}=0})
\end{equation}
\begin{equation}
\geq P_{U_i|U_{i-1}}(0, 0) \inf_Q \left[ P_{X|U_{i-1}}(0, 0) D(P_{U_i|X=1, U^{i-1}=0} || Q) \right]
\end{equation}
\begin{equation}
= P_{U_i|U_{i-1}}(1, 0) I(U_i; X|U^{i-1} = 0)
\end{equation}
\begin{equation}
= I(U_i; X|U^{i-1})
\end{equation}
Theorem 4 also implies the following bound on the external information (see (18)):
\begin{equation}
I(U^r; XY) \leq \frac{1}{m_1} \sum_{1 \leq i \leq r} \log \alpha_i \prod_{2 \leq j \leq i-1} \alpha_j^{-1}
+ \frac{1}{m_2} \sum_{1 \leq i \leq r} \log \alpha_i \prod_{1 \leq j \leq i-1} \alpha_j^{-1}
\end{equation}

**Remark 5**: Let us provide some intuition why interaction helps, assuming the case of $m_1 = m_2 = m$ for simplicity. From the proof of Theorem 4, it can be seen that up to a constant factor, $s^e_{\alpha \infty}(X; Y)$ is at least $\frac{d^2}{m^2} \int_0^1 \frac{\log m}{m} e^{-t} dt \left( \frac{1}{m} \int_0^1 \frac{\log m}{m} e^{-t} dt \right)^{-1}$ which is in fact sharp as will be seen from the upper bound on $s^e_{\alpha \infty}(X; Y)$ in Theorem 6. Moreover, lower bounds on $s^e_{\alpha}(X; Y)$ can be
computed by replacing the integrals with discrete sums with $r$ terms:

$$\frac{\delta^2 [r/2]}{m^3} \sum_{i=1}^{r} (e^{t_i} - e^{t_{i-1}}) \left( \frac{1}{m} \sum_{i=1}^{[r/2]} e^{-t_{i-1}}(t_i - t_{i-1}) \right)^{-1}$$  \hspace{1cm} (61)

where $1 = t_0 < t_1 < \cdots < t_{[r/2]} = \frac{\ln m}{100}$. In particular, when $r = 1$, we recover $s_1^*(X;Y) \sim \frac{\delta^2}{m^3}$, whereas choosing $t_i - t_{i-1} = 1$, $i = 1, \ldots, [r/2]$ shows that $r \sim \ln m$ achieves $s_r^*(X;Y) \sim \frac{\delta^2}{m}$. Even better, later we will take $t_k$ as the $k$-th iterated power of 2, and then $r$ will be the super logarithm of $m$.

Recall that $(\alpha_i)$ control the amount of information revealed in each round and serve as hyperparameters of the algorithm to be tuned. Next we shall explain how to select the value of $(\alpha_i)$ so that the optimal performance is achieved in the one-way and interactive settings.

**B. r = 1 Case**

Specializing Theorem 4 we obtain:

**Corollary 4:** For any $m_1, m_2 > 10$, with $r = 1$ and $\alpha_1 = \frac{m_1}{10}$ we have

$$P_X(0) \log \alpha_1 \leq \frac{1}{m_1} \log \frac{m_1}{10};$$

$$\lim_{\delta \to 0} \delta^{-2} I(U_1;Y) \geq \frac{1}{50m_1 m_2}. \hspace{1cm} (62)$$

**C. r = \infty Case**

Denote by $n_2$ the $n$-th tetration of 2, which is defined recursively by $0^2_2 = 1$ and

$$n^2_2 := 2^{(n-1)^2}, \forall n \geq 1. \hspace{1cm} (64)$$

Let $m := \min \{m_1, m_2\}$, and let $r_0$ be the minimum integer such that

$$\exp(e^{r_0} - 1) \geq \frac{m}{10} \hspace{1cm} (65)$$

For $m > 10$ we have $r_0 \geq 1$. Then we set

$$r := 2r_0; \hspace{1cm} (66)$$

$$\alpha_{2k-1} := \alpha_{2k} := \exp(e^{(k^2 - (k-1)^2)}), \forall k \in \{1, \ldots, r_0 - 1\}; \hspace{1cm} (67)$$

$$\alpha_{2r_0-1} := \alpha_{2r_0} := \frac{m}{10} \exp(e^{(1 - (r_0 - 1)^2)}), \hspace{1cm} (68)$$

which fulfills $\alpha_i > 1$. We see that

$$\sum_{1 \leq j \leq r} \ln \alpha_j \leq e \sum_{k=1}^{\infty} k^2 \exp(e^{-(k-1)^2}) \leq 5.$$  \hspace{1cm} (71)

The first inequality above follows by $\frac{m}{10} \exp_e(1 - (r_0 - 1)^2) \leq \exp(e^{r_0} - 1) \leq \exp(e^{r_0} - (r_0 - 1)^2)$. Similarly we also have $\sum_{1 \leq j \leq r} \ln \alpha_j \leq \ln m$. Moreover,

$$\prod_{1 \leq j \leq r} \alpha_j \geq \prod_{1 \leq j \leq r} \frac{m}{10}. \hspace{1cm} (73)$$

Summarizing, we have

**Corollary 5:** Consider $m_1, m_2 > 10$, $m := \min \{m_1, m_2\}$, and $(\alpha_i)$ defined in (66)-(68). We have

$$\sum_{1 \leq j \leq r} \ln \alpha_j \leq \frac{50m_1 m_2}{m}; \hspace{1cm} (74)$$

$$\sum_{1 \leq j \leq r} \frac{m_1}{m_2}; \hspace{1cm} (75)$$

$$\lim_{\delta \to 0} \delta^{-2} \sum_{1 \leq j \leq r} I(U_1;Y | U^{i-1}_1) \geq \frac{m}{50m_1 m_2}; \hspace{1cm} (76)$$

$$\lim_{\delta \to 0} \delta^{-2} \sum_{1 \leq j \leq r} I(U_1;X | U^{i-1}_1) \geq \frac{m}{50m_1 m_2}. \hspace{1cm} (77)$$

where $r = 2r_0$ and $r_0$ is defined in (65).

Let us remark that the sequence $(\alpha_i)$ we used in (66)-(68) is essentially optimal: Let $\beta_k := \prod_{2 \leq j \leq 2k} \alpha_j^{-1}$. In order for (69) to converge, we need $\sum_k \ln(\frac{\alpha_k}{\beta_k}) \beta_k^{-1}$ to be convergent. Therefore $\beta_k$ cannot grow faster than $\beta_k = \exp(\beta_{k-1})$ which is tetration. However this only amounts to a lower bound on $r$ for a particular design of $P_{U^i | XY}$ in Definition 5. Since tetration grows super fast, from a practical viewpoint $r$ is essentially a constant. Nevertheless, it remains an interesting theoretical question whether $r$ needs to diverge:

**Conjecture I:** If there is an algorithm (indexed by $k$) achieving the optimal risk (28) for nonparametric estimation, then necessarily $r \to \infty$ as $k \to \infty$.

**VI. ACHIEVABILITY BOUNDS FOR BERNOLLI ESTIMATION**

In this section we analyze the performance of the Bernoulli distribution estimation algorithm described in Section IV.

**A. Communication Complexity**

Consider any $i \in \{1, \ldots, r\}$. Denoting by $\tilde{P}_{X^i Y^i}$ the empirical distribution of $(X(i), Y(i), U_1(i), \ldots, U_{i}(i))_{i=1}^n$, we have from (49) that

$$\mathbb{E}(\delta) | \tilde{P}_{X^i Y^i} | X, Y, U^{i-1} = \tilde{P}_{X^i Y^i} P_{U^i | X^{i-1}}. \hspace{1cm} (78)$$

In particular,

$$\mathbb{E}(\delta) | \tilde{P}_{X^i Y^i} | X, Y = P_{XY} P_{U^i | X}. \hspace{1cm} (79)$$

Let $\ell(\hat{j}_i) := 2 \lfloor \log_2(\hat{j}_i) \rfloor + 1$ be the number of bits need to encode the positive integer $\hat{j}_i$ using the Elias gamma code [17].
For each $i \in \{1, \ldots, r\} \cap 2\mathbb{Z}$ we have
\[
\mathbb{E}^{(\delta)}[\ell(\hat{j}_i) | X, Y, U^{i-1}] \leq 2\mathbb{E}^{(\delta)}[\log \hat{P}_i | X, Y, U^{i-1}] + 1
\]  
(80)
\[
\leq 2 \log_2 \mathbb{E}^{(\delta)}[\hat{j}_i | X, Y, U^{i-1}] + 1
\]  
(81)
\[
= 2 \log_2 \alpha_i \Gamma_{XU^{i-1},(0,0)} + 1
\]  
(82)
\[
= 2n \Gamma_{XU^{i-1},(0,0)} \log_2 \alpha_i + 1
\]  
(83)
where (82) follows from the selection rule (40) and the formula of expectation of the geometric distribution. Then
\[
\mathbb{E}^{(\delta)}[\ell(\hat{j}_i)] \leq 2n P^{(\delta)}_{XU^{i-1},(0,0)} \log_2 \alpha_i + 1
\]  
(84)
\[
\leq 2(1 + \delta) n P^{(0)}_{XU^{i-1},(0,0)} \log_2 \alpha_i + 1
\]  
(85)
where (84) used (79); (85) used the fact that $P^{(\delta)}_{XY}$ is dominated by $(1 + \delta) P^{(0)}_{XY}$. Note that 0 in (84)-(85) denotes the value of the vector $U^{i-1}$.

B. Expectation of $\Gamma^B$

Recall that $\Gamma^B$ was defined in (44). Pick arbitrary $i \in \{1, \ldots, r\} \cap 2\mathbb{Z}$,
\[
P_{U^i Y}^{(\delta)} := P_{Y} \prod_{j=1}^{i} P_{U_j|U^{j-1}Y}^{(\delta)}
\]  
(86)
and since $P_{Y}$ and $(P_{U_j|U^{j-1}Y}^{(\delta)})_{j \in \{1, \ldots, r\} \cap 2\mathbb{Z}}$ are independent of $\delta$, we obtain
\[
\partial_{\delta} \ln P_{U^i Y}^{(\delta)}(u^i, y)_{\delta=0} = \sum_{1 \leq j \leq i} \text{odd} \, \Gamma_j(u^j, y). \tag{87}
\]
Next, observe that for any $l \in \{1, \ldots, n\}$,
\[
\mathbb{E}^{(0)}[\Gamma_1(U^l(1), Y(l)) | U^{l-1}(l), Y(l)]
\]  
\[
= \mathbb{E}^{(0)}\left[ \left. \frac{\partial_{\delta} P_{U^l|U^{l-1}(l)}^{(\delta)}(u^l|Y(l), U^{l-1}(l))}{P_{U^l|YU^{l-1}(l)}^{(0)}(u^l|Y(l), U^{l-1}(l))} \right| U^{l-1}(l), Y(l) \right] \tag{88}
\]  
\[
= \sum_{u^l, y} \partial_{\delta} P_{U^l|U^{l-1}(l)}^{(\delta)}(u^l|Y(l), U^{l-1}(l)) \big|_{\delta=0}
\]  
\[
= \sum_{u^l, y} \partial_{\delta} P_{U^l|U^{l-1}(l)}^{(\delta)}(u^l|Y(l), U^{l-1}(l)) \big|_{\delta=0}
\]  
\[
= 0. \tag{89}
\]
Moreover, for any $\delta \neq 0$,
\[
1 \sum_{l=1}^{n} \mathbb{E}^{(\delta)}[\Gamma_1(U^l(1), \ldots, U_l(l), Y(l))] \tag{90}
\]
\[
= \delta^{-1} \sum_{u^l, y} \Gamma_1(u^l, y) P_{U^l Y}^{(\delta)}(u^l, y) \tag{91}
\]
\[
= n \sum_{u^l, y} \Gamma_1(u^l, y) \frac{\partial}{\partial \delta} P_{U^l Y}^{(\delta)}(u^l, y)_{\delta=0} \tag{92}
\]
\[
= n \sum_{u^l, y} \Gamma_1(u^l, y) P_{U^l Y}^{(0)}(u^l, y) \sum_{1 \leq j \leq l} \Gamma_j(u^j, y) \tag{93}
\]
\[
= n \sum_{u^l, y} \Gamma_1^2(u^l, y) P_{U^l Y}^{(0)}(u^l, y) \tag{94}
\]
where (91) used (79); (92) used (90) and the linearity of $P_{U^{l-1}Y}^{(\delta)}$ in $\delta$; (93) used (87); (94) follows from (90). Thus
\[
\frac{1}{\delta} \mathbb{E}^{(\delta)}[\Gamma^B] = I^B, \quad \forall \delta \neq 0 \tag{95}
\]
where we defined
\[
I^B := n \sum_{1 \leq j \leq r} \text{odd} \, \mathbb{E}^{(0)}[\Gamma_1^2(U^j(1), Y(1))]. \tag{96}
\]

Lemma 6: Let $(U^i, Y) \sim P_{U^i Y}^{(\delta)}$. We have
\[
I^B \geq 2n \lim_{\delta \to 0} \delta^{-2} \sum_{1 \leq j \leq r} I(U_i; Y|U^{i-1}) \tag{97}
\]
where the logarithmic base of the mutual information is natural.

Proof: Consider any $i \in \{1, 2, \ldots, r\} \cap 2\mathbb{Z}$, we have
\[
\mathbb{E}^{(0)}[\Gamma_1^2(U^i, Y)]
\]  
\[
= \mathbb{E}^{(0)}\left[ \left. \left| \partial_{\delta} \ln P_{U_i|YU^{i-1}}^{(\delta)}(U_i|YU^{i-1}) \right|_{\delta=0} \right|^2 \right] \tag{98}
\]  
\[
= 2 \lim_{\delta \to 0} \delta^{-2} D(P_{U_i|YU^{i-1}}^{(\delta)} || P_{U_i|YU^{i-1}}^{(0)}) \tag{99}
\]  
\[
= 2 \lim_{\delta \to 0} \delta^{-2} D(P_{U_i|YU^{i-1}}^{(\delta)} || P_{U_i|U^{i-1}}^{(0)}) \tag{100}
\]  
\[
= 2 \lim_{\delta \to 0} \delta^{-2} D(P_{U_i|YU^{i-1}}^{(\delta)} || P_{U_i|U^{i-1}}^{(\delta)}) \tag{101}
\]  
\[
\geq 2 \lim_{\delta \to 0} \delta^{-2} D(P_{U_i|YU^{i-1}}^{(\delta)} || P_{U_i|U^{i-1}}^{(\delta)}) \tag{102}
\]  
\[
= \lim_{\delta \to 0} \delta^{-2} I(U_i; Y; U^{i-1}) \tag{103}
\]
where we defined the conditional KL divergence
\[
D(P_{Y|X} || Q_{Y|X}) := \int D(P_{Y|X=x} || Q_{Y|X=x}) dP_X(x);
\]  
(100) follows since $P_{U_i|YU^{i-1}}^{(0)} = P_{U_i|U^{i-1}}^{(0)}$; (101) follows since $\lim_{\delta \to 0} P_{U_i|YU^{i-1}}^{(\delta)} = P_{U_i|U^{i-1}}^{(0)}$. \[\Box\]

C. Variance of $\Gamma^B$

For any $\delta$, since $(U^r, X, Y) \sim (P_{U^r XY}^{(\delta)})^n$, we have
\[
\text{var}^{(\delta)}(\Gamma^B) = \sum_{l=1}^{n} \frac{\text{var}^{(\delta)}}{1 \leq j \leq r} \Gamma_1(U^l(1), \ldots, U_l(l), Y(l)) \tag{104}
\]
\[
= n \text{var}^{(\delta)} \left( \sum_{1 \leq j \leq r} \Gamma_1(U_1(1), \ldots, U_1(l), Y(1)) \right) \tag{105}
\]
However,
\[
\var(\delta) \left\{ \sum_{1 \leq i \leq r} \Gamma_i(U_i(1), \ldots, U_i(1), Y(1)) \right\} \\
\leq \mathbb{E}(\delta) \left\{ \left[ \sum_{1 \leq i \leq r} \Gamma_i(U_i(1), \ldots, U_i(1), Y(1)) \right]^2 \right\} \\
\leq (1 + \delta) \mathbb{E}(0) \left\{ \left[ \sum_{1 \leq i \leq r} \Gamma_i(U_i(1), \ldots, U_i(1), Y(1)) \right]^2 \right\} \\
\leq (1 + \delta) \sum_{1 \leq i \leq r} \mathbb{E}(0) \left[ \Gamma_i^2(U_i(1), \ldots, U_i(1), Y(1)) \right] \\
\leq (1 + \delta) \sum_{1 \leq i \leq r} \mathbb{E}(0) \left[ \Gamma_i^2(U_i(1), \ldots, U_i(1), Y(1)) \right] \tag{108}
\]
where (107) follows since \( P(\delta) \) is dominated by \( (1 + \delta) P(0) \); (108) used (90). Therefore
\[
\var(\delta)(\Gamma^B) \\
\leq n(1 + \delta) \sum_{1 \leq i \leq r} \mathbb{E}(0) \left[ \Gamma_i^2(U_i(1), \ldots, U_i(1), Y(1)) \right] \tag{109}
\]
\[
= (1 + \delta) \mathbb{E}(0).
\tag{110}
\]

**D. \( r = 1 \) Case**

We now prove achievability bounds for the Bernoulli distribution estimation algorithm.

**Corollary 7:** Given \( m_1, m_2 > 10 \), for \( r = 1 \) and \( \alpha_1 := \frac{m_1}{10} \), the mean square error \( \mathbb{E}[|\hat{\delta}^B - \delta|^2] \leq 25(1 + \delta)m_1m_2 \) and total communication cost \( \mathbb{E}[\|\Pi\|] \leq \frac{22(1 + \delta)^2m_1m_2}{m_1} \log_2 \frac{m_1}{10} + 1 \).

**Proof:** We have
\[
\mathbb{E}[|\hat{\delta}^B - \delta|^2] = \var(\delta) \cdot (\mathbb{E}(\delta^2) \cdot (\Pi^B))^{-2} \tag{111}
\]
\[
\leq (1 + \delta) I^B \cdot (I^B)^{-2} \tag{112}
\]
\[
\leq (1 + \delta) \left[ n \lim_{\delta \to 0} \delta^{-2} I(U_1; Y) \right]^{-1} \tag{113}
\]
\[
\leq 25 (1 + \delta)m_1m_2 \tag{114}
\]
where (111) follows since \( \hat{\delta}^B \) is unbiased (Theorem 3); (112) follows from (95) and (110); (113) follows from Lemma 6; lastly we used Corollary 4.

As for the communication cost
\[
\mathbb{E}[\|\Pi^{A-B}\|] \leq 2(1 + \delta)n \sum_{1 \leq i \leq 1} P_{XU_i-1}(0, 0) \log_2 \alpha_i + 1 \tag{115}
\]
\[
\leq \frac{22(1 + \delta)^2}{m_1} \log_2 \frac{m_1}{10} + 1 \tag{116}
\]
where we used (85) and Corollary 4. \( \square \)

**E. \( r = \infty \) Case**

**Corollary 8:** Let \( m_1, m_2 > 10 \), \( m := \min\{m_1, m_2\} \).

For \( r \), \( \alpha_i \) defined in Section V-C, the mean square error
\[
\mathbb{E}[|\hat{\delta}^B - \delta|^2] \leq \frac{25(1 + \delta)m_1m_2}{m_1} \text{ and total communication cost } \mathbb{E}[\|\Pi\|] \leq \frac{6(1 + \delta)n(m_1 + m_2 - 1) \log_2 e + \frac{r + 1}{2}}{m_1}.
\]

**Proof:** The bound on the mean square error is similar to the \( r = 1 \) case:
\[
\mathbb{E}[|\hat{\delta} - \delta|^2] \leq \var(\delta) \cdot (\mathbb{E}(\delta^2) \cdot (\Pi^B))^{-2} \tag{117}
\]
\[
\leq (1 + \delta) I^B \cdot (I^B)^{-2} \tag{118}
\]
\[
\leq (1 + \delta) \left[ n \lim_{\delta \to 0} \delta^{-2} \sum_{1 \leq i \leq r} I(U_i; Y | U_i^{-1}) \right]^{-1} \tag{119}
\]
\[
\leq 25 (1 + \delta)m_1^2m_2 \tag{120}
\]
except that we use Corollary 5 in the last step.

For the communication cost,
\[
\mathbb{E}[\|\Pi^{A-B}\|] \leq 2(1 + \delta)n \sum_{1 \leq i \leq r} P_{XU_i-1}(0, 0) \log_2 \alpha_i + \frac{r + 1}{2} \tag{121}
\]
\[
\leq 6(1 + \delta)n(m_1^{-1} + m_2^{-1}) \log_2 e + \frac{r + 1}{2} \tag{122}
\]
where used (85) and Corollary 5. \( \square \)

**VII. DENSITY ESTIMATION UPPER BOUNDS**

In this section we prove the upper bounds in Theorem 1 and Theorem 2, by building nonparametric density estimators based on the Bernoulli distribution estimator. For \( \beta \in (0, 2] \), the rectangular kernel is minimax optimal (Section II-B), so that the integral with the kernel can be directly estimated using the Bernoulli distribution estimator, which we explain in Section VII-B and VII-C. Extension to higher order kernels is possible using a linear combination of rectangular kernels, which is explained in Section VII-D.

**A. Density Lower Bound Assumption**

First, we observe the following simple argument showing that it suffices to consider \( A > 0 \). Define
\[
B := \sup_{x, y \in [0, 1]^4} \| p_{XY}(x, y) \|, \tag{123}
\]
where the supremum is over all density \( p_{XY} \) on \( \mathbb{R}^d \) satisfying \( \| p_{XY} \|_{[0,1]^2, \beta} \leq L \). Clearly \( B > 1 \) is finite and depends only on \( \beta, L, d \).

**Lemma 9:** In either the one-way or the interactive setting, suppose that there exists an algorithm achieving \( \max_{p_{XY} \in H(\beta, L, A)} \mathbb{E}[|\hat{p} - p_{XY}(x_0, y_0)|^2] \leq R \) for some \( R > 0 \) and \( A \in (0, 1) \). Then, there must be an algorithm achieving \( \max_{p_{XY} \in H(\beta, L, 0)} \mathbb{E}[|\hat{p} - p_{XY}(x_0, y_0)|^2] \leq \left( \frac{4 + A}{1 - A} \right)^2 \). \( \square \)

**Proof:** Pick one \( p_{XY} \) such that \( \| p_{XY} \|_{[0,1]^2, \beta} \leq L \) and \( \inf_{x, y \in [0, 1]^4} p_{XY}(x, y) \geq \frac{1 - A}{2} > A \). Since \( \frac{1 - A}{2} < 1 \), such \( p_{XY} \) exists. Consider an arbitrary \( q_{XY} \in H(\beta, L, 0) \), and suppose infinite pairs \( (X_1, Y_1), \ldots \) i.i.d. according to \( q_{XY} \) are available to Alice and Bob. Using the common randomness, Alice and Bob can simulate i.i.d. pairs \( (\hat{X}_1, \hat{Y}_1), \ldots \) according...
to \( \hat{p}_{XY} := \frac{2A}{A+1} p_{XY} + \frac{1-A}{A+1} q_{XY} \), by replacing each pair with probability \( \frac{2A}{A+1} \) with a new pair drawn according to \( p_{XY} \).

Clearly \( \hat{p}_{XY} \in \mathcal{H}(\beta, L, A) \), and by assumption, \( \hat{p}_{XY}(x_0, y_0) \) can be estimated with mean square risk \( R \). Since \( p_{XY} \) is known, this implies that \( g_{XY}(x_0, y_0) \) can be estimated with mean square risk \( \left( \frac{1+\delta}{1+2\delta} \right)^2 R \). \( \square \)

For the rest of the section, we will assume that there is a density lower bound \( A > 0 \) and \( p_{XY} \in \mathcal{H}(\beta, L, A) \), which is sufficient in view of Lemma 9. Consider bandwidth \( h > 0 \) (which will be specified later as an inverse polynomial of \( k \)).

Also introduce the notations

\[
\begin{align*}
A & := x_0 + h[-1, 1]^d; \\
B & := y_0 + h[-1, 1]^d.
\end{align*}
\]

Define \( P_{XY} \) as the probability distribution induced by \( p_{XY} \) and with

\[
\begin{align*}
\mathcal{X} & := 1\{X \notin A\}; \\
\mathcal{Y} & := 1\{Y \notin B\}.
\end{align*}
\]

Define

\[
\begin{align*}
m_1 & := P^{-1}(\mathcal{X} = 0); \\
m_2 & := P^{-1}(\mathcal{Y} = 0).
\end{align*}
\]

Note that \( m_1/m_2 \) is bounded above and below by positive constants depending on \( A, \beta, \) and \( L \) (see (132) and (137)).

Also, we can assume Alice and Bob both know \( m_1 \) and \( m_2 \), since with infinite samples Alice and Bob know their marginal densities \( p_X \) and \( p_Y \), and Alice can send \( m_1 \) to Bob with very high precision using negligible number of bits. Let \( \delta \geq -1 \) be the number such that \( \hat{p}_{XY} \) is the matrix

\[
\left( \frac{1}{m_2} \left( 1 + \frac{1}{m_2} - \frac{\delta}{m_1 m_2} \right) \right).
\]

Let \( \delta \) be Bob’s estimator of \( \delta \) in (48). Then we define Bob’s density estimator:

\[
\hat{p}_B := \frac{1 + \delta}{m_1 m_2 h^{2d}}.
\]

We next show that the smoothness of the density ensures that \( 1 + \delta \) is at most the order of a constant. Recall that \( A \) is a density lower bound on \( p_X \) and \( p_Y \). Define \( M := \max\{m_1, m_2\} \) and \( m := \min\{m_1, m_2\} \). The definition of \( (m_1, m_2) \) implies \( Ah^d \leq \frac{1}{M^d} \), and hence

\[
h \leq \left( \frac{1}{AM} \right)^{1/d}.
\]

Recall \( B \) defined in (123). We then have

\[
1 + \delta = m_1 m_2 P_{XY}(A \times B) \leq m_1 m_2 B h^{2d} \leq B m_1 m_2 A^2 m^2 \leq BM.
\]

Next, observe that \( 1 = m_1 P_X(A \times [-1, 1]^d) \leq m_1 B h^d \) which yields \( h^d \geq \frac{1}{m_1 B} \). Similarly we also have \( h^d \geq \frac{1}{m_2 B} \), therefore

\[
h^d \geq \frac{1}{mB}.
\]

Together with (132), we see that \( h^d = \Theta(1/m) = \Theta(1/M) \).

Next, the bias of the density estimator is

\[
E[\hat{p}_B] - P_{XY}(x_0, y_0) = \frac{P_{XY}(A \times B)}{h^{2d}} - P_{XY}(x_0, y_0)
\]

which is just the bias of the rectangular kernel estimator (with bandwidth \( h \) in each of the two subspaces). The rectangle kernel is order 1 [44, Definition 1.3] and compactly supported while, by assumption \( \beta \in (0, 2) \), and therefore the bias is bounded by (44, Proposition 1.2)

\[
|E[\hat{p}_B] - P_{XY}(x_0, y_0)| \leq Ch^\beta
\]

where \( C \) is a constant depending only on \( \beta, d \) and \( L \).

B. One-Way Case

By Corollary 7 and (137), we can bound the variance of the density estimator as

\[
\text{var}(\hat{p}_B) = \frac{1}{m_1 m_2 h^{2d}} \text{var}(\hat{p}_B) \leq \frac{1}{m_1 m_2 h^{2d}} \frac{25(1 + \delta) m_1 m_2}{n} \leq \frac{25(1 + \delta) B^4 m^3}{n M} + 1 \leq k.
\]

Now we can choose \( h \) so that \( m_1 = \left( \frac{k}{\log k} \right)^{\beta/2 \pi \pi} \) as defined by (128), and set

\[
n = \left( \frac{2.2(1 + \delta_{\text{max}})}{m_1} \log_2 \frac{m_1}{10} \right)^{-1} (k - 1)
\]

where \( \delta_{\text{max}} \) defined as the right side of (136) and hence depends only on \( A, \beta, L \), is an upper estimate of the true parameter \( \delta \). Then the communication constraint is satisfied if the following holds

\[
\frac{2.2(1 + \delta_{\text{max}})}{m_1} \log_2 \frac{m_1}{10} + 1 \leq k.
\]
C. Interactive Case

Choose $h$ such that $m$ as defined by $m := \min\{m_1, m_2\}$ and (128)-(129) satisfies

$$m := k^{\frac{d}{m + \Delta}}$$

and set

$$n := \left\lfloor m k \ln 2 \right\rfloor$$

where as before $\delta_{\max}$ is an upper bound on $\delta$ and only depends on $(A, \beta, L)$. By Corollary 8, for $k$ large enough we have $m > 10$, and the communication cost is bounded by $k$. Moreover from (146), the risk is bounded by $Dk^{-\frac{2d}{m + \Delta}}$ for some $D$ depending only on $\beta$, $L$, and $A$. This proves the upper bound in Theorem 2 when $\beta \in (0, 2]$.

D. Extension to $\beta > 2$

For $\beta > 2$, the rectangular kernel is no longer minimax optimal. However, observe the following:

Proposition 10: For any positive integers $d$ and $l$, there exists an order $l$-kernel in $\mathbb{R}^d$ which is a linear combination of $\left\lfloor 1/2 \right\rfloor + 1$ indicator functions.

Proof: In the following we prove for $d = 1$; the case of general $d$ will then follow by taking the tensor product of kernel functions in $\mathbb{R}$. Note that an $l$-th kernel must satisfy the following equations:

$$\int K(u)du = 1;$$

$$\int u^j K(u)du = 0, \quad j = 1, \ldots, l.$$  \hspace{1cm} (150)

Let us consider $K$ of the following form:

$$K(u) = \sum_{k=1}^{k_0} c_k 1_{[k,k]}$$  \hspace{1cm} (152)

where $k_0 := \left\lfloor 1/2 \right\rfloor + 1$. Since such $K(u)$ is an even function, (150)-(151) yield $k_0$ nontrivial equations for $c_1, \ldots, c_{k_0}$ (i.e., only when $j$ is even):

$$2 \sum_{k=1}^{k_0} k c_k = 1;$$

$$\sum_{k=1}^{k_0} \frac{2k+1}{j+1} c_k = 0, \quad j \in \{1, \ldots, l\} \cap 2\mathbb{Z}. \hspace{1cm} (154)$$

From the formula for the determinant of the Vandermonde matrix, we see that these equations have a unique solution for $c_1, \ldots, c_{k_0}$.

Now for general $\beta > 0$, we can take an order $l = \left\lfloor \beta \right\rfloor$ kernel as in Proposition 10. We can estimate

$$\frac{1}{n^2} \int p_{XY}(x,y) K_{\left\lfloor 1/2 \right\rfloor + 1}^{(x,y)} dx dy$$

by applying the Bernoulli distribution estimator repeatedly for $\left\lfloor 1/2 \right\rfloor + 1$ times. Therefore by the similar arguments as the preceding sections we see that the upper bounds in Theorem 1 and Theorem 2 hold for $\beta > 2$ as well.

VIII. ONE-WAY DENSITY ESTIMATION LOWER BOUND

A. Upper Bounding $s^*(X; Y)$

The pointwise estimation lower bound is obtained by lower bounding the risk for estimating $P_{XY}$ (with $X$ and $Y$ being indicators of neighborhoods of $x_0$ and $y_0$), and applying Le Cam’s inequality to the latter. Therefore we are led to considering the strong data processing constant for the biased Bernoulli distribution.

Theorem 5: Let $P_{XY}^{(i)}$ be as defined in (29), where $\delta \in (-1, 1)$ and $m > 1$. Then $s^*(X; Y) \leq \frac{\delta^2}{m \ln m - m + 1}$.

Proof: For this proof we can assume without loss of generality that the logarithms are natural. For any $Q_X$, let $Q_Y$ be the output through the channel $P_{Y|X}$. Then

$$s^*(X; Y) \leq \frac{D(Q_Y||P_Y)}{D(Q_X||P_X)} \leq \frac{D_{\chi^2}(Q_Y||P_Y)}{D_{\chi^2}(Q_X||P_X)} \leq \frac{\delta^2}{m \ln m - m + 1}$$

where we define the $\chi^2$ divergence in (7). The justification of the steps are as follows: (155) is well-known. (156) follows since the $\chi^2$ divergence dominates the KL divergence (see e.g. [39]). To see (157), note that $\frac{D_{\chi^2}(Q_Y||P_Y)}{D_{\chi^2}(Q_X||P_X)}$ is upper bounded by $\rho_m^2(X, Y)$, the square of the maximal correlation coefficient (see e.g. [5], [6]). As the operator norm of a linear operator, $\rho_m^2(X, Y)$ can be shown to equal the second largest singular value of

$$M := \left( \frac{1}{\sqrt{P_X(x)}} P_{XY}(x, y) \frac{1}{\sqrt{P_Y(y)}} \right)_{x,y} \hspace{1cm} (158)$$

see e.g. [6]. Since $M$ is a symmetric matrix, its singular values are its eigenvalues. The largest eigenvalue of $M$ is 1, corresponding to the eigenvector $\sqrt{P_X(0)}$, $\sqrt{P_X(1)}$ (which is evident from (158)), whereas the trace

$$\text{tr}(M) = \frac{1 + \delta}{m} + 1 - \frac{1}{m} + \frac{\delta}{m(m-1)} = 1 + \frac{\delta}{m-1} \hspace{1cm} (160)$$

which is evident from (159). Therefore $\rho_m^2(X; Y) = \frac{\delta}{m-1}$. Moreover, since $\chi^2$ and KL divergences are both f-divergences, their ratio can be bounded by the ratio of their corresponding f-functions (see e.g. [39]):

$$\frac{\chi^2(Q_X||P_X)}{\chi^2(Q_Y||P_Y)} \leq \sup_{0 < t < \epsilon_m} \frac{(t-1)^2}{t \ln t + t + 1} \leq \frac{(m-1)^2}{m \ln m - m + 1} \hspace{1cm} (162)$$

The constraint $t \leq m$ in (161) is because $\inf_{x \in (-1/m, 1)} \chi^2(Q_X(x) = 1/m$ and $\sup_{x \in (-1/m, 1)} \chi^2(Q_X(x) \leq m$. To show (162), it suffices to show that

$$\inf_{u \in (-1/m, 1)} \frac{(1 + u) \ln (1 + u) - u}{u^2} \leq 1$$
is achieved at \( u = m - 1 \). For this, it suffices to show that the derivative of the objective function, \( \frac{-(2+u) \ln(1+u)+2u}{u^2} \), is negative on \((-1, m-1)\). Indeed, define \( \phi(u) := (2+u) \ln(1+u)-2u \). We can check that \( \phi(0) = 0, \phi'(0) = 0, \) and \( \phi''(u) = \frac{u^2}{u^2} \), which imply that \( \phi(u) > 0 \) for \( u > 0 \) and \( \phi(u) < 0 \) for \( u < 0 \). Therefore (162), and hence (157), holds.

B. Lower Bounding One-Way NP Estimation Risk

Given \( k, d, \beta, L, A \), consider a distribution \( P_{XY} \) on \([0,1]^2\) with matrix

\[
\begin{pmatrix}
\frac{1}{m}(1+\delta) & \frac{1}{m}(1-\frac{1}{m}) - \frac{\delta}{m} \\
\frac{1}{m}(1-\frac{1}{m}) + \frac{\delta}{m} & \frac{1}{m}(1-\frac{1}{m})
\end{pmatrix},
\]

where \( m := \left(\frac{ak}{m} \right)^{\frac{2d}{d-1}} \) and \( \delta := m^{-\frac{d}{2}}, \) with \( a := \frac{16d+8d}{\ln 2} \) being a constant. We then need to “simulate” smooth distributions from \( P_{XY} \). Let \( f : \mathbb{R}^d \to [0, \infty) \) be a function satisfying the following properties:

- \( f \) has a compact support;
- \( \int_{\mathbb{R}^d} f = 1; \)
- \( f(0) > 0; \)
- \( f(x) \in [0, 1], \) for all \( x \in \mathbb{R}^d; \)
- \( \|f\|_{\mathbb{R}^d} \leq \frac{1}{4}; \)
- Define \( g(x, y) = f(x)f(y) \) as a function on \( \mathbb{R}^{2d}. \) Then

\[
\|g\|_{\mathbb{R}^{2d}} \leq \frac{1}{4}.
\]

Clearly, such a function exists for any given \( \beta, L, d. \) For sufficiently large \( m \) such that \( m^{-\frac{d}{4}} \text{supp}(f) + x_0 \in [0, 1]^d \) and \( m^{-\frac{d}{4}} \text{supp}(f) + y_0 \in [0, 1]^d \) (recall that \( (x_0, y_0) \) is the given point in the density estimation problem), define

\[
p_{X|Y=0}(x) := \frac{1}{P_X(1)} f(m^{\frac{1}{d}}(x-x_0)), \quad \forall x \in \mathbb{R}^d; \tag{164}
\]

\[
p_{Y|X=0}(y) := \frac{1}{P_Y(0)} f(m^{\frac{1}{d}}(y-y_0)), \quad \forall y \in \mathbb{R}^d. \tag{165}
\]

Since \( P_X(0) = P_Y(0) = \frac{1}{m}, \) clearly the above define valid probability densities supported on \([0, 1]^d\). Define

\[
p_{X|Y=1}(x) := \frac{1}{P_X(1)} \{x \in [0, 1]^d \} [1 - f(m^{1/d}(x-x_0))]; \tag{166}
\]

\[
p_{Y|X=1}(y) := \frac{1}{P_Y(1)} \{y \in [0, 1]^d \} [1 - f(m^{1/d}(y-y_0))], \tag{167}
\]

which are also probability densities supported on \([0, 1]^d\). Define \( P_{XY|XX} = P_{X|X}P_{Y|X}, \) where \( P_{X|X} \) and \( P_{Y|X} \) are conditional distributions defined by the densities above. Under the joint distribution \( P_{XY|XX}, \) we have

\[
p_X(x) = p_Y(y) = 1, \quad \forall x, y \in [0, 1]^d. \tag{168}
\]

Define

\[
\tilde{P}_{XY|XX} = P_{X|X}P_{Y|X}P_XP_Y. \tag{169}
\]

We now check that the density of \( P_{XY} \) satisfies

\[
\|P_{XY}\|_{(0,1)^{2d}, \beta} \leq L \quad \text{for } m \text{ sufficiently large. Indeed,}
\]

\[
\begin{align*}
\|P_{XY}(x,y)\| & = \sum_{i,j=0,1} p_{XY|X=0}(i,x)p_{XY|Y=0}(j,y)P_X(i,j) \\
& = \sum_{i,j=0,1} p_{XY|X=0}(i,x)p_{XY|Y=0}(j,y)P_X(i,j) \\
& + \sum_{i,j=0,1} p_{XY|X=1}(i,x)p_{XY|Y=1}(j,y)P_X(i,j) \\
& = 1 + \delta \left[ \frac{1}{m^2} (P_X|X=0(x) - P_X|X=1(x)) ight] \\
& + \delta \left[ \frac{1}{m^2} (P_Y|Y=0(y) - P_Y|Y=1(y)) \right] \\
& = 1 + \delta \left[ \frac{1}{m^2} f(m^{1/d}(x-x_0)) \right] \\
& + \delta \left[ \frac{1}{m^2} f(m^{1/d}(y-y_0)) \right] \\
& = \|P_{XY}\|_{(0,1)^{2d}, \beta} \leq L \quad \text{for } m \text{ sufficiently large. Indeed,}
\end{align*}
\]

\[
\|\tilde{P}_{XY}(x,y)\| & = 1 + \delta \left[ \frac{m}{m-1} f(0) - \frac{1}{m-1} \right] \tag{170}
\]

the latter following from (173). Now suppose that Bob declares \( P_{XY} \) if

\[
|\hat{p} - p_{XY}(x_0,y_0)| \leq |\hat{p} - 1|, \tag{179}
\]
and \( P_{XY} \) otherwise. By Chebyshev’s inequality, the error probability (under either hypothesis) is upper bounded by
\[
4\delta^{-2} \left[ \frac{m}{m-1} f(0) - \frac{1}{m-1} \right]^{-4} \sup_{p_{XY} \in \mathcal{H}(\beta, L, A)} \mathbb{E}[|\tilde{p} - p_{XY}(x_0, y_0)|^2].
\]
(180)

On the other hand, from (22) and Theorem 5 we have
\[
\frac{D(P_{YII} \| \bar{P}_{YII})}{H(\Pi)} \leq s^*(X; Y) \leq \frac{m \ln m - m + 1}{2\delta^2} \leq \frac{m \cdot \frac{d}{a} + 4d}{\text{dak}} \leq \frac{8\beta + 4d}{\text{dak}}
\]
(181)
when \( m \) is sufficiently large. However, it is known (from Kraft’s inequality, see e.g. [14]) that the expected length of a prefix code upper bounds the entropy. Thus
\[
k \geq \mathbb{E}[\|\Pi\|] \geq \frac{1}{\log 2} H(\Pi)
\]
(182)
and therefore
\[
D(P_{YII} \| \bar{P}_{YII}) \leq \frac{8\beta + 4d}{\text{dak}} \log 2.
\]
(183)

Then by Pinsker’s inequality (e.g. [44]),
\[
1 - \int dP_{YII} \wedge d\bar{P}_{YII} \leq \frac{1}{2 \log e} \sqrt{D(P_{YII} \| \bar{P}_{YII})}
\]
(184)
\[
= \frac{1}{2} \log 2
\]
(185)
where the last line follows from our choice \( a = \frac{16\beta + 4d}{d} \ln 2 \).

According to Theorem 2, to complete the proof of Theorem 6, the following technical result needs to be established:

A. Upper Bounding \( s^*_\infty(X; Y) \)

The proof is the following technical result:

Theorem 6: There exists \( c > 0 \) small enough such that the following holds: For any \( P_{XY} \) which is a distribution on \( \{0, 1\}^2 \) corresponding to the following matrix:
\[
\begin{pmatrix}
p^2(1 + \delta) & p\bar{p} - p^2\delta \\
p\bar{p} - p^2\delta & \bar{p}^2 + \beta \delta
\end{pmatrix}
\]
(186)
where \( p, |\delta| \in [0, c] \) and we used the notation \( \bar{p} := 1 - p \), we have
\[
s^*_\infty(X; Y) \leq c^{-1} p\delta^2.
\]
(187)

The proof can be found in Appendix B.

B. Lower Bounding Interactive NP Estimation Risk

The proof is similar to the one-way case (Section VIII-B). Consider the distribution \( P_{XY} \) on \( \{0, 1\}^2 \) as in (163). Let \( m := (ak)^{\frac{d}{2d+2}} \) and \( \delta := \frac{m^2}{2\sigma} \), where \( a = \frac{2\ln 2}{e} + \frac{c}{\text{dak}} \) with \( c \) being the absolute constant in Theorem 6. Pick the function \( f \), and define \( P_{X|XY} \) and \( P_{Y|XY} \) as before. Note that, as before, \( P_{XY} \) is uniform on \( [0,1]^{2d} \), while \( \|P_{XY}\|_{\{0,1\}^{2d}} \leq L \) for \( m \geq 10 \). \( P_{XY}(x_0, y_0) \) has the same formula (178), and Alice and Bob can convert a (now interactive) density estimation algorithm to an algorithm for testing \( P_{XY} \) against \( \bar{P}_{XY} \). With the same testing rule (179), the error probability under either hypothesis is again upper bounded by (180).

Changes arise in (181), where we shall apply Theorem 6 instead. Note that for the absolute constant \( c \) in Theorem 6, the condition \( \frac{1}{m}, |\delta| < c \) is satisfied for sufficiently large \( k \) (hence sufficiently large \( m \)).

\[
\frac{D(P_{YII} \| \bar{P}_{YII})}{H(\Pi)} \leq \frac{1}{2 \log e} \sqrt{D(P_{YII} \| \bar{P}_{YII})} \leq \frac{8\beta + 4d}{\text{dak}} \log 2
\]
(188)

Again using Kraft’s inequality to Bound \( H(\Pi) \), we obtain
\[
D(P_{YII} \| \bar{P}_{YII}) \leq \frac{\log 2}{ca}.
\]
(189)

Then Pinsker’s inequality yields
\[
1 - \int dP_{YII} \wedge d\bar{P}_{YII} \leq \frac{\sqrt{\ln 2}}{2ca} = \frac{1}{2}
\]
(190)

since we selected \( a = \frac{2\ln 2}{e} \). Again \( \int dP_{YII} \wedge d\bar{P}_{YII} \) lower bounds twice of (180), therefore
\[
\frac{D(P_{YII} \| \bar{P}_{YII})}{H(\Pi)} \leq \frac{1}{2 \log e} \sqrt{D(P_{YII} \| \bar{P}_{YII})} \leq \frac{8\beta + 4d}{\text{dak}} \log 2
\]
(191)

which is lower bounded by \( \frac{1}{17}m^{-2\beta/d}f^4(0) = \frac{f^4(0)}{17}(ak)^{\frac{d}{d+2}} \) for large enough \( k \). Since \( a \) and \( f(0) \) depend only on \( d, \beta, L \), this establishes the lower bound in Theorem 1.

IX. INTERACTIVE DENSITY ESTIMATION LOWER BOUND

In this section we prove the lower bound in Theorem 2.
APPENDIX I

PROOF OF THEOREM 4

First, assume that \( i \in \{1, 2, \ldots, r\} \setminus 2\mathbb{Z} \). By the definitions of \( P_{U_i|X^{U_i-1}} \) and \( P_{U_i|Y^{U_i-1}} \), we can verify that the following holds (for \( \delta = 0 \)):

\[
P_{X|U_i}(0|0) = \frac{P_X(0)}{P_X(1) \prod_{1 \leq j \leq i} \alpha_j^{-1} + P_X(0)}.
\]

(202)

Indeed, (202) follows by applying induction on the following

\[
P_{X|U_i}(0|0) = \frac{p_0}{p_0 + p_1}
\]

(203)

\[
= \frac{P_X(1) \prod_{1 \leq j \leq i} \alpha_j^{-1} + P_X(0)}{P_X(1) \prod_{1 \leq j \leq i} \alpha_j^{-1} + P_X(0)}
\]

(204)

where \( p_0 := P_{X|U_i}(0|0)P_{U_i|X^{U_i-1}}(0|0, 0) \), \( p_1 := P_{X|U_i}(1|0)P_{U_i|X^{U_i-1}}(1|0, 0) \), and we used \( P_{X|U_i}(0|0) = P_{X|U_i}(1|0) \) which in turn follows from the factorization

\[
P_{XYU_i|U_i} = P_{X|U_i}P_{Y|U_i}P_{U_i|Y^{U_i-2}U^{U_i-2}}
\]

(205)

\[
P_{X|U_i}(0|0) = P_{X|U_i}(1|0)
\]

(206)

Similarly, by switching the roles of \( X \) and \( Y \) we have

\[
P_{U_i|Y^{U_i-2}}(0|0) = P_Y(1) \prod_{2 \leq j \leq i} \alpha_j^{-1} + P_Y(0)
\]

(207)

\[
= P_Y(1) \prod_{2 \leq j \leq i} \alpha_j^{-1} + P_Y(0)
\]

(208)

\[
P_{Y|U_i}(1|0)
\]

(209)

Therefore,

\[
P_{X|U_i}(0|0) = \prod_{1 \leq j \leq i} P_{U_i|X^{U_i-1}}(0|0)\]

(210)

\[
= \left[ P_X(1) \prod_{1 \leq j \leq i} \alpha_j^{-1} + P_X(0) \right] \left[ P_Y(1) \prod_{2 \leq j \leq i} \alpha_j^{-1} + P_Y(0) \right]
\]

(211)

for any \( i = 1, \ldots, r \). We also see from (202) and (211) that for any \( i \) odd,

\[
\sum_{\text{even}} P_{X|U_i}(0|0) = P_X(0) \left[ P_Y(1) \prod_{2 \leq j \leq i-1} \alpha_j^{-1} + P_Y(0) \right] \]

(212)

\[
= \frac{1}{m_1} \left[ \frac{m_1 - 1}{m_2} \prod_{2 \leq j \leq i-1} \alpha_j^{-1} + \frac{1}{m_2} \right]
\]

(213)

\[
\leq \frac{1.1}{m_1} \prod_{2 \leq j \leq i-1} \alpha_j^{-1}
\]

(214)

where the last step follows since \( \prod_{2 \leq j \leq i-1} \alpha_j^{-1} \geq \frac{10}{m_2^2} \). Therefore, the claim (52) follows, and (53) is similar.

Next, consider any \( i \in \{1, \ldots, r\} \). Define

\[
\delta_i := \frac{P_{XY|U_i=0}(0, 0)}{P_{X|U_i=0}P_{Y|U_i=0}} - 1.
\]

(215)

Observe that the construction of \( P_{U_i|XY} \) fulfills the Markov chain conditions (14)-(15), implying that

\[
P_{XYU_i|U_i} = P_{XY|U_i}P_{U_i|Y^{U_i-1}}(1, 0)
\]

(216)

\[
\leq \frac{1 + \delta}{1 - \delta}(1 + \frac{\delta}{m_1-1})(1 - \frac{1}{m_2})
\]

(217)

where we defined

\[
\delta := \frac{P_{XY|U_i=0}(0, 0)}{P_{X|U_i=0}P_{Y|U_i=0}} - 1.
\]

(218)

By continuity, we have

\[
\delta = \frac{b(0) + o(1)}{b(0) + o(1)},
\]

(219)

as \( \delta \to 0 \). It is also easy to see from (217) that \( \delta_i = O(\delta) \) (for this proof, only \( \delta \) is the variable, and all other constants, such as \( m \) and \( (\alpha_i) \), can be hidden in the Landau notations).

Therefore (217)(220)(221) yield

\[
1 + \left( \frac{1}{m_1 - 1} \right) \left( \frac{1}{m_2 - 1} \right) \delta + o(\delta)
\]

(222)

\[
= 1 + \left( \frac{b(\delta)}{b(0)} \right) \left( \frac{c(\delta)}{c(0)} \right) \delta_i + o(\delta)
\]

(223)

Using the fact that \( X \) and \( Y \) are independent under \( P(0) \), noting (202) and the assumption \( \prod_{1 \leq j \leq i} \alpha_j^{-1} \geq \frac{10}{m_1^2} \), we have

\[
\delta = \frac{P_{X|U'}(0|0)}{P_{X|U'}(0|0) + \frac{m_1}{m_1 - 1} + \frac{1}{m_1}} \leq \frac{1}{10},
\]

(224)

where \( i' \) is the largest odd integer not exceeding \( i \). Similarly we also have \( c(0) \leq \frac{1}{10} \). Consequently, (223) yields

\[
\delta_i^2 \geq \delta^2 \left( \frac{1 + \frac{1}{m_1 - 1}}{1 + \frac{1}{m_2 - 1}} \right)^2 + o(\delta^2)
\]

(225)

\[
\geq 0.9^2 \delta^2 + o(\delta^2).
\]

(226)

Moreover, let us define

\[
\delta := \frac{P_{XY|U_i=1}(0, 0)}{P_{XY|U_i=1}(0, 0)P_{U_i=1}(1, 0)} - 1.
\]

(227)

1 We use the notation \( \delta := 1 - \delta \) for \( \delta \in [0, 1] \).
In the following paragraph we consider arbitrary $i \in \{1, 2, \ldots, r\} \setminus 2\mathbb{Z}$, and we shall omit the superscripts $(\delta)$ for $a^{(\delta)}$, $b^{(\delta)}$, $c^{(\delta)}$, unless otherwise noted. Then

$$I(U_i; Y|U^{i-1} = 0) = aD(P_{Y|U^{i-1}=0} \| P_{Y|U^{i-1}=0}) + \bar{a}D(P_{Y|U=1, U^{i-1}=0} \| P_{Y|U^{i-1}=0})$$

(228)

We can verify that $P_{X|U^{i}=0}(0) = \frac{b}{b + \alpha_i b}$. Therefore,

$$P_{Y|U^{i}=0}(0) = \frac{b}{b + \alpha_i b} \cdot c(1 + \delta_{i-1})$$

$$+ \frac{\alpha_i^{-1} b}{b + \alpha_i b} \cdot \bar{b}c - \delta_{i-1} bc$$

(229)

$$= \frac{bc(1 - \alpha_i^{-1})}{b + \alpha_i b} \cdot \delta_{i-1}.$$  (230)

Therefore as $\delta \to 0$,

$$D(P_{Y|U^{i}=0} \| P_{Y|U^{i-1}=0}) = d(P_{Y|U^{i}=0}(0) \| c)$$

(231)

$$= \frac{1}{2} c \left( \frac{b(\alpha_i - 1)}{\alpha_i b + b} \cdot \delta_{i-1} \right)^2 + o(\delta^2)$$

(232)

$$\geq \frac{1}{2} c \left( b(\alpha_i - 1)(0.92\delta) \right)^2 + o(\delta^2)$$

(233)

$$\geq \frac{0.9^4}{2} cb^2(\alpha_i - 1)\delta^2 + o(\delta^2)$$

(234)

where $d(p\|q) := p \log \frac{p}{q} + (1 - p) \log \frac{1-p}{1-q}$ denotes the binary divergence function, and recall that we assumed the natural base of logarithms. On the other hand, $P_{Y|U=1, U^{i-1}=0}(0) = P_{Y|X=1, U^{i-1}=0}(0) = c - \frac{\delta_{i-1} bc}{1 - b}$. Therefore

$$D(P_{Y|U=1, U^{i-1}=0} \| P_{Y|U^{i-1}=0}) = d \left( c - \frac{\delta_{i-1} bc}{1 - b} \right)$$

(235)

$$= \frac{1}{2} c \left( \frac{\delta_{i-1} b}{1 - b} \right)^2 + o(\delta^2)$$

(236)

$$\geq \frac{0.9^4}{2} cb^2\delta^2 + o(\delta^2).$$

(237)

Turning back to (228), we obtain

$$I(U_i; Y|U^{i-1} = 0)$$

$$= \frac{0.9^4}{2} \left[ ab^2 c(\alpha_i - 1)^2 \delta^2 + (1 - a)cb^2\delta^2 \right] + o(\delta^2)$$

(238)

$$\geq \frac{0.9^5}{2} \left( \frac{\alpha_i^{-1}(\alpha_i - 1)^2 + 1 - \alpha_i^{-1}}{2m_1 m_2} \right) \delta^2 + o(\delta^2)$$

(239)

$$\geq \frac{0.9^5}{2} (\alpha_i - 1)bc^2\delta^2 + o(\delta^2)$$

(240)

where (239) follows since (208) implies

$$a^{(\delta)} = a^{(0)} + a(1)$$

(241)

$$= \left( 1 - \frac{1}{m_1} \right) \prod_{1 \leq j \leq i - 2} \alpha_j^{-1} + \frac{1}{m_1} + o(1)$$

(242)

$$\geq \left( 1 - \frac{1}{m_1} \right) \prod_{1 \leq j \leq i - 2} \alpha_j^{-1} + o(1)$$

(243)

and

$$1 - a^{(\delta)} = 1 - a^{(0)} + o(1)$$

(244)

$$= \left( 1 - \frac{1}{m_1} \right) \prod_{1 \leq j \leq i - 2} \alpha_j^{-1} + o(1)$$

(245)

$$\geq \left( 1 - \frac{1}{m_1} \right) \prod_{1 \leq j \leq i - 2} \alpha_j^{-1} + o(1)$$

(246)

Moreover, by (202),

$$b^{(\delta)} = b^{(0)} + o(1)$$

(247)

$$= \left( 1 - \frac{1}{m_1} \right) \prod_{1 \leq j \leq i - 1} \alpha_j + o(1)$$

(248)

Similarly,

$$c^{(\delta)} \geq \frac{0.9}{m_1} \prod_{1 \leq j \leq i - 1} \alpha_j + o(1).$$

(249)

Therefore by (240) and (211),

$$I(U_i; Y|U^{i-1}) = I(U_i; Y|U^{i-1} = 0)P_{Y|U^{i-1} = 0}(0)$$

(250)

$$\geq \frac{0.9^5}{2} (\alpha_i - 1)b^2c\delta^2 \prod_{1 \leq j \leq i - 1} \alpha_j^{-1} + o(\delta^2)$$

(251)

$$\geq \frac{0.9^5 \delta^2}{2m_1 m_2} (\alpha_i - 1) \prod_{1 \leq j \leq i - 1} \alpha_j + o(\delta^2)$$

(252)

$$= \frac{0.9^5 \delta^2}{2m_1 m_2} \left( \prod_{1 \leq j \leq i - 1} \alpha_j - \prod_{1 \leq j \leq i - 2} \alpha_j \right) + o(\delta^2)$$

(253)

and hence

$$\sum_{1 \leq i \leq r} I(U_i; Y|U^{i-1}) \geq \frac{0.9^5 \delta^2}{2m_1 m_2} \prod_{1 \leq j \leq r} \alpha_j + o(\delta^2),$$

(254)

establishing the claim (54) of the theorem. The proof of (55) is similar.
APPENDIX II
PROOF OF THEOREM 6
We can choose the natural base of logarithms in this proof. Choose \( \mathbf{U} = (U_1, U_2, \ldots, U_r) \) satisfying the Markov chain conditions (14)-(15) and so that
\[
s_n^\infty(X; Y) \leq 2 \cdot \frac{I(X; Y) - I(X; Y|\mathbf{U})}{I(U; X, Y)} \quad (259)
\]
\[
\leq 4 \cdot \frac{I(X; Y) - I(X; Y|\mathbf{U})}{I(U; X) + I(U; Y)} \quad (260)
\]
which is possible by the definition of \( s_n^\infty(X; Y) \).
Given \( \alpha, \beta \in [0, 1] \), define by \( P^{\alpha, \beta} \) the unique distribution\(^2\) such that
\[
P^{\alpha, \beta}(x, y) = P_{XY}(x, y)f(x)g(y) \quad (261)
\]
for some functions \( f \) and \( g \), and such that the marginals are \( P^{\alpha} := [\alpha, \bar{\alpha}] \) and \( P^{\beta} := [\beta, \bar{\beta}] \). For the existence of \( P^{\alpha, \beta} \), see e.g. [23] and [40]. Define \( I(\alpha, \beta) \) as the mutual information of \( (X, Y) \) under \( P^{\alpha, \beta} \). Define \( \lambda = \lambda(\alpha, \beta) \) as the number such that \( P^{\alpha, \beta} \) is the matrix
\[
\left( \begin{array}{cc}
\alpha \beta + \lambda & \alpha \bar{\beta} - \lambda \\
\bar{\alpha} \beta - \lambda & \bar{\alpha} \bar{\beta} + \lambda
\end{array} \right). \quad (262)
\]
Given any \( u \), let \( \alpha_u \in [0, 1] \) be such that \( P_{X|U=u} = [\alpha_u, \bar{\alpha}_u] \). Define \( \beta_u \) similarly but for \( P_{Y|U=u} \). With these notations, note that
\[
E[\alpha_u] = E[\beta_u] = p; \quad (263)
\]
and
\[
I(X; Y) - I(X; Y|\mathbf{U}) = I(p, p) - E[I(\alpha_u, \beta_u)]; \quad (264)
\]
\[
I(U; X) + I(U; Y) = E[d(\alpha_u||p) + d(\beta_u||p)] \quad (265)
\]
where we recall that \( d(\cdot||\cdot) \) denotes the binary divergence function. Define
\[
\psi(\alpha, \beta) := d(\alpha||p) + d(\beta||p). \quad (266)
\]
Then note that \( \psi(\alpha, \beta) \) is a smooth nonnegative function on \([0, 1]^2\) with vanishing value and first derivatives at \((p, p)\). Also, define
\[
\phi(\alpha, \beta) := I(p, p) - I(\alpha, \beta) + I(\alpha, \beta)(\alpha - p) + I(\beta, \beta)(\beta - p) \quad (267)
\]
where \( I(\alpha, \beta) := \frac{\partial}{\partial \alpha} I(\alpha, \beta) \). Then \( \phi(\alpha, \beta) \) is also a smooth function on \([0, 1]^2\) with vanishing value and first derivatives at \((p, p)\). Moreover, due to (263), we have
\[
E[\phi(\alpha_u, \beta_u)] = I(p, p) - E[I(\alpha_u, \beta_u)]. \quad (268)
\]
Thus to prove the theorem it suffices to show the existence of sufficiently small \( c > 0 \), such that for any \( p, |\delta| \in (0, c) \), there is
\[
\sup_{\alpha, \beta} \phi(\alpha, \beta) \psi(\alpha, \beta) \leq c^{-1} p \delta^2 \quad (269)
\]
where the sup is over \( \alpha, \beta \in (0, 1) \).

\textbf{Case 1:} \( 0.1p < \alpha, \beta < 10p \).

\[\text{Since } \frac{\partial^2}{\partial \alpha^2} D(\alpha||p) = \left[ \frac{1}{\alpha} + \frac{1}{1-\alpha} \right] \geq \frac{1}{\alpha} \geq \frac{1}{10p} \text{ for } \alpha \in (0, 10p), \text{ we have} \]
\[
\psi(\alpha, \beta) \geq \frac{1}{20p} [(\alpha - p)^2 + (\beta - p)^2] \quad (270)
\]
for \( (\alpha, \beta) \in [0, 10p]^2 \). Now if we can show that
\[
\sup_{(\alpha, \beta) \in [0, 10p]^2} \|\partial^2 \phi(\alpha, \beta)\| = \sup_{(\alpha, \beta) \in [0, 10p]^2} \|\partial^2 I(\alpha, \beta)\| \quad (271)
\]
\[
\lesssim \delta^2, \quad (272)
\]
we will obtain \( \sup_{(\alpha, \beta) \in [0, 10p]^2} \frac{\phi(\alpha, \beta)}{\psi(\alpha, \beta)} \lesssim p \delta^2 \) which matches (269). Here and below, \( x \lesssim y \) means that there is an absolute constant \( C > 0 \) such that \( x \leq Cy \) when \( c \) in the theorem statement (and hence \( p \) and \( |\delta| \)) is sufficiently small.
Before explicitly computing \( \partial^2 \phi(\alpha, \beta) \), we give some intuitions why we should expect (272) to be true. For fixed \( \alpha, \beta, p \), we will show that
\[
I(\alpha, \beta) = \tilde{I}(\alpha, \beta) + o(\delta^2) \quad (273)
\]
as \( \delta \to 0 \), where we defined
\[
\tilde{I}(\alpha, \beta) := \frac{\delta^2}{2p^2} \alpha \bar{\alpha} \beta \bar{\beta}. \quad (274)
\]
If the difference between \( I(\alpha, \beta) \) and \( \tilde{I}(\alpha, \beta) \) could be neglected, then (272) should hold. To see (273), for given \( \alpha, \beta \in (0, 1) \), note that (261) implies,
\[
\frac{\frac{\lambda}{\alpha \bar{\alpha} \beta \bar{\beta}}}{(1 - \frac{\lambda}{\alpha \bar{\alpha}})(1 - \frac{\lambda}{\bar{\alpha} \bar{\beta}})} = \frac{(1 + \frac{\lambda}{\alpha \beta})(1 + \frac{\lambda}{\bar{\alpha} \beta})}{(1 - \frac{\lambda}{\alpha \bar{\alpha}})(1 - \frac{\lambda}{\beta \bar{\beta}})} \quad (275)
\]
Under the assumption \( \delta \to 0 \), the above linearizes to
\[
\frac{\lambda}{\alpha \bar{\alpha} \beta \bar{\beta}} = \frac{\delta}{p^2} + o(\delta). \quad (276)
\]
Moreover, note that
\[
D_{\chi^2}(P^{\alpha, \beta} \| P^{\alpha} \times P^{\beta}) = \frac{\lambda^2}{2 \alpha \bar{\alpha} \beta \bar{\beta}} \quad (277)
\]
\[
= \frac{\delta^2}{2p^2} \alpha \bar{\alpha} \beta \bar{\beta} + o(\delta^2) \quad (278)
\]
where the last step follows by comparing with (276). Since \( I(\alpha, \beta)/D_{\chi^2}(P^{\alpha, \beta} \| P^{\alpha} \times P^{\beta}) \to 1 \) as \( \delta \to 0 \), we see (273) holds. Of course, (273) does not really show (272) since approximation of function values generally does not imply approximation of the derivatives. However, we shall next explicitly take the derivatives to give a real proof, and the above observations are useful guides.
First, note that
\[
\frac{\partial I(\alpha, \beta)}{\partial \alpha} = \sum_{x,y \in \{0,1\}} \left( \frac{\partial}{\partial \alpha} P^{\alpha, \beta}(x, y) \right) \ln \frac{P^{\alpha, \beta}(x, y)}{P^{\alpha}(x) P^{\beta}(y)} \tag{279}
\]
\[
= (\beta + \lambda_\alpha) \ln(1 + \frac{\lambda}{\alpha \beta}) + (\beta - \lambda_\alpha) \ln(1 - \frac{\lambda}{\alpha \beta})
\]
\[
- \lambda_\alpha \ln(1 + \frac{\lambda}{\alpha \beta}) + (\lambda - \lambda_\alpha) \ln(1 + \frac{\lambda}{\alpha \beta}). \tag{280}
\]
where \(\lambda_\alpha := \frac{\partial}{\partial \alpha} \lambda_\alpha\). Next, we express the first and second derivatives, \(\lambda_\alpha\), \(\lambda_\beta\) and \(\lambda_{\alpha, \beta}\), in terms of \(\lambda\). Differentiating the logarithm of (275) in \(\beta\) yields
\[
\lambda_\beta \left[ \frac{1}{\alpha \beta + \lambda} + \frac{1}{\alpha \beta - \lambda} + \frac{1}{\alpha \beta - \lambda} \right] \tag{281}
\]
In the rest of the proof the notation \(f(t) = O(t)\) means \(|f(t)| \lesssim |t|\) (recall the definition of \(\lesssim\) in (272)), and \(f(t) = \Theta(t)\) if \(1 \lesssim |f(t)/t| \lesssim 1\). Note that for \(0.1 p < \alpha, \beta < 10 p\) we have
\[
\frac{\lambda}{\alpha \beta} = \Theta(\delta), \tag{282}
\]
since the right side of (275) clearly equals \(1 + \Theta(\delta)\). Then by (281),
\[
\lambda_\beta = \lambda \cdot \left( 1 + O\left( \frac{\lambda}{\alpha \beta} \right) \right) - \frac{\lambda}{\alpha \beta} \tag{283}
\]
and hence,
\[
\lambda_\beta = \lambda \cdot \frac{\beta - \lambda}{\beta \beta} \left( 1 + O\left( \frac{\lambda}{\alpha \beta} \right) \right) = O\left( \frac{\lambda}{\alpha \beta} \right). \tag{284}
\]
Expression of \(\lambda_\alpha\) can be found similarly. Moreover, differentiating (281) we get
\[
\lambda_{\alpha, \beta} \left[ \frac{1}{\alpha \beta + \lambda} + \frac{1}{\alpha \beta + \lambda} + \frac{1}{\alpha \beta - \lambda} \right] \tag{285}
\]
\[
+ \lambda_\alpha \lambda_\beta \left[ - \frac{1}{(\alpha \beta + \lambda)^2} + \frac{1}{(\alpha \beta - \lambda)^2} - \frac{1}{(\alpha \beta - \lambda)^2} \right] \tag{286}
\]
\[
+ \lambda_\alpha \lambda_\beta \left[ - \frac{1}{(\alpha \beta + \lambda)^2} + \frac{1}{(\alpha \beta + \lambda)^2} - \frac{1}{(\alpha \beta - \lambda)^2} \right] \tag{287}
\]
\[
+ \lambda_\beta \left[ - \frac{1}{(\alpha \beta + \lambda)^2} + \frac{1}{(\alpha \beta + \lambda)^2} - \frac{1}{(\alpha \beta - \lambda)^2} \right] \tag{288}
\]
\[
+ \lambda \left[ \frac{1}{(\alpha \beta + \lambda)^2} + \frac{1}{(\alpha \beta + \lambda)^2} - \frac{1}{(\alpha \beta - \lambda)^2} \right] = 0,
\]
from which we can deduce that
\[
\lambda_{\alpha, \beta} = O\left( \frac{\lambda}{\alpha \beta} \right). \tag{289}
\]
Now, taking the derivative in (280), we obtain
\[
\frac{\partial}{\partial \alpha} I(\alpha, \beta) \tag{290}
\]
\[
= \lambda_{\alpha, \beta} \lambda \cdot \frac{1}{\alpha \beta} + \lambda_\alpha \lambda_\beta + \lambda_\alpha \lambda \cdot \frac{\beta - \beta}{\alpha \beta} \tag{291}
\]
\[
+ \lambda_\beta \lambda \cdot \frac{\alpha - \alpha}{\alpha^2 \beta^2} + \frac{\lambda^2}{\alpha^2 \beta^2} \left( \alpha^2 \beta - \beta^2 \right) \tag{292}
\]
\[
+ O\left( \frac{\lambda^3}{\alpha^3 \beta^3} \right). \tag{293}
\]
In deriving (287), we applied the Taylor expansions of \(x \mapsto \ln(1 + x)\) and \(x \mapsto \frac{1}{x + 1}\). Plugging (284) and (286) into (287), we obtain
\[
|\partial_{\alpha, \beta} I(\alpha, \beta)| = O\left( \frac{\lambda^2}{\alpha \beta} \right) = O(\delta^2). \tag{294}
\]
Next, we control \(|\partial_{\alpha, \alpha} I(\alpha, \beta)|\). Similarly to (281), we have
\[
\lambda_\alpha \left[ \frac{1}{\alpha \beta + \lambda} + \frac{1}{\alpha \beta + \lambda} + \frac{1}{\alpha \beta - \lambda} + \frac{1}{\alpha \beta - \lambda} \right] \tag{295}
\]
\[
- \lambda_\alpha \left[ \frac{1}{\alpha \beta + \lambda} - \frac{1}{\alpha \beta + \lambda} - \frac{1}{\alpha \beta - \lambda} + \frac{1}{\alpha \beta - \lambda} \right] \tag{296}
\]
\[
+ \lambda_\beta \left[ \frac{1}{\alpha \beta + \lambda} - \frac{1}{\alpha \beta + \lambda} - \frac{1}{\alpha \beta - \lambda} + \frac{1}{\alpha \beta - \lambda} \right] \tag{297}
\]
\[
+ \lambda \left[ \frac{1}{\alpha \beta + \lambda} + \frac{1}{\alpha \beta + \lambda} - \frac{1}{\alpha \beta - \lambda} - \frac{1}{\alpha \beta - \lambda} \right] = 0. \tag{298}
\]
Further taking the derivative, we obtain
\[
\lambda_{\alpha, \alpha} \left[ \frac{1}{\alpha \beta + \lambda} + \frac{1}{\alpha \beta + \lambda} + \frac{1}{\alpha \beta - \lambda} + \frac{1}{\alpha \beta - \lambda} \right] \tag{299}
\]
\[
+ \lambda_\alpha \left[ - \frac{1}{\alpha \beta + \lambda} + \frac{1}{\alpha \beta + \lambda} + \frac{1}{\alpha \beta - \lambda} - \frac{1}{\alpha \beta - \lambda} \right] \tag{300}
\]
\[
+ \lambda_\alpha \left[ - \frac{1}{\alpha \beta + \lambda} + \frac{1}{\alpha \beta + \lambda} - \frac{1}{\alpha \beta - \lambda} + \frac{1}{\alpha \beta - \lambda} \right] \tag{301}
\]
\[
+ \lambda_\beta \left[ - \frac{1}{\alpha \beta + \lambda} + \frac{1}{\alpha \beta + \lambda} - \frac{1}{\alpha \beta - \lambda} + \frac{1}{\alpha \beta - \lambda} \right] \tag{302}
\]
\[
+ \lambda \left[ \frac{1}{\alpha \beta + \lambda} + \frac{1}{\alpha \beta + \lambda} - \frac{1}{\alpha \beta - \lambda} - \frac{1}{\alpha \beta - \lambda} \right] = 0. \tag{303}
\]
Next, we shall use the assumption of \(\alpha, \beta \in (0.1p, 10p)\) to simplify (303) as
\[
\lambda_{\alpha, \alpha} \cdot \Theta\left( \frac{1}{p^2} \right) - \lambda_\alpha \cdot \Theta\left( \frac{1}{p^2} \right) + \lambda_\beta \cdot \Theta\left( \frac{1}{p^2} \right) + \lambda \cdot \Theta\left( \frac{1}{p^2} \right) = 0. \tag{304}
\]
Since, analogous to (284), we have
\[
\lambda_\alpha = O\left( \frac{\lambda}{\alpha} \right), \tag{305}
\]
we see that (304) implies
\[
\lambda_{\alpha, \alpha} = O\left( \frac{\lambda}{p^2} \right). \tag{306}
\]
Tighter estimates of $\lambda_{\alpha,\alpha}$ are possible, but the above will suffice. We now take the derivative of (280) in $\alpha$:

$$\partial_{\alpha,\alpha}I(\alpha,\beta) = I_1 + I_2$$

(294)

where

$$I_1 := \lambda_{\alpha,\alpha} \left[ \ln(1 + \frac{\lambda}{\alpha\beta}) - \ln(1 - \frac{\lambda}{\alpha\beta}) - \ln(1 - \frac{\lambda}{\alpha\beta}) + \ln(1 + \frac{\lambda}{\alpha\beta}) \right]$$

(295)

and

$$I_2 := \lambda_{\alpha}^2 \left( \frac{1}{1 + \frac{\lambda}{\alpha\beta}} + \frac{1}{1 - \frac{\lambda}{\alpha\beta}} + \frac{1}{1 + \frac{\lambda}{\alpha\beta}} + \frac{1}{1 + \frac{\lambda}{\alpha\beta}} \right)$$

$$+ \lambda_{\alpha} \left( \frac{1}{1 + \frac{\lambda}{\alpha\beta}} - \frac{1}{1 - \frac{\lambda}{\alpha\beta}} + \frac{1}{1 + \frac{\lambda}{\alpha\beta}} - \frac{1}{1 - \frac{\lambda}{\alpha\beta}} \right)$$

$$+ \lambda_{\alpha} \lambda_{\alpha} \left( \frac{1}{1 + \frac{\lambda}{\alpha\beta}} - \frac{1}{1 - \frac{\lambda}{\alpha\beta}} + \frac{1}{1 + \frac{\lambda}{\alpha\beta}} - \frac{1}{1 - \frac{\lambda}{\alpha\beta}} \right)$$

$$+ \lambda \left( \frac{1}{1 + \frac{\lambda}{\alpha\beta}} + \frac{1}{1 - \frac{\lambda}{\alpha\beta}} + \frac{1}{1 + \frac{\lambda}{\alpha\beta}} + \frac{1}{1 + \frac{\lambda}{\alpha\beta}} \right)$$

(296)

Using (292) with $\lambda \leftarrow \delta^2$ and $\alpha \leftarrow \alpha$, we obtain

$$\lambda_{\alpha}(p, p) = O(\delta^2)$$

Thus Taylor expanding the above displayed, we obtain

$$\partial_{\alpha}I(p, p) \lesssim p\delta^2$$

(301)

Then

$$\phi(\alpha, \beta) := I(p, p) - I(\alpha, \beta) + I_\alpha(p, p)(\alpha - p)$$

$$+ I_\beta(p, p)(\beta - p)$$

(302)

$$\leq p^2\delta^2 - 0 + 2I_{\alpha}(p, p)(\alpha - p)$$

(303)

$$\lesssim p\alpha\delta^2$$

(304)

where we used the assumption that $\alpha \geq \beta$ and the fact that $I_{\alpha}(p, p) = I_{\alpha}(p, p)$. Now the assumption of $\alpha \geq 10p$ implies

$$\psi(\alpha, \beta) \geq d(\alpha\|p) \gtrsim \alpha.$$ 

(305)

To see the second inequality in (305), note that

$$\min_{\alpha \in [10p, p]} \frac{1}{\alpha} d(\alpha\|p) = \min_{\alpha \in [10p, p]} \left\{ \frac{\ln \alpha + 1 - \alpha}{\alpha} \right\}$$

(306)

$$= \frac{\delta(10p\|p)}{10p}$$

(307)

where the minimization is easily solved by checking that the derivative is positive for $\alpha \geq 10p$. Finally, combining (305) with (304), we obtain $\frac{\phi(\alpha, \beta)}{\psi(\alpha, \beta)} \lesssim \delta^2 p$, as desired.

- **Case 3:** $\max\{\alpha, \beta\} \leq 0.1p$. Assume without loss of generality that $\alpha \leq 0.1p$. In this case, using (301), we have

$$\phi(\alpha, \beta) := I(p, p) - I(\alpha, \beta) + I_\alpha(p, p)(\alpha - p)$$

$$+ I_\beta(p, p)(\beta - p)$$

(308)

$$\leq p^2\delta^2 - 0 + I_\alpha(p, p)(\alpha + \beta - 2p)$$

(309)

$$\leq p^2\delta^2 + I_\alpha(p, p) \cdot 18p$$

(310)

$$= O(p^2\delta^2).$$

(311)

On the other hand,

$$\psi(\alpha, \beta) \geq d(\alpha\|p) \geq d(0.1\|p)$$

(312)

$$\geq (0.9 - 0.1 \ln 10)p + O(p^2)$$

(313)

$$= \Theta(p).$$

(315)

Thus we once again obtain $\frac{\phi(\alpha, \beta)}{\psi(\alpha, \beta)} \lesssim \delta^2 p$, as desired.

**ACKNOWLEDGMENT**

The author would like to thank Prof. Venkat Anantharam, for bringing the reference [36] to his attention and some interesting discussions.

**REFERENCES**

[1] J. Acharya, C. Canonne, A. V. Singh, and H. Tyagi, “Optimal rates for nonparametric density estimation under communication constraints,” in Proc. Adv. Neural Inf. Process. Syst., vol. 34, 2021, pp. 26754–26766.

[2] J. Acharya, C. L. Canonne, and H. Tyagi, “Inference under information constraints II: Communication constraints and shared randomness,” IEEE Trans. Inf. Theory, vol. 66, no. 12, pp. 7856–7877, Dec. 2020.

[3] R. Ahlswede and M. V. Burnashev, “On minimax estimation in the presence of side information about remote data,” Ann. Statist., vol. 18, no. 1, pp. 141–171, Mar. 1990.

[4] R. Ahlswede and I. Csiszar, “Hypothesis testing with communication constraints,” IEEE Trans. Inf. Theory, vol. IT-32, no. 4, pp. 533–542, Jul. 1986.

[5] R. Ahlswede and P. Gacs, “Spreading of sets in product spaces and hypercontraction of the Markov operator,” Ann. Probab., vol. 4, no. 6, pp. 925–939, Dec. 1976.
