Interactive Inference Under Information Constraints

Jayadev Acharya, *Member, IEEE*, Clément L. Canonne, Yuhan Liu, *Student Member, IEEE*, Ziteng Sun, *Student Member, IEEE*, and Himanshu Tyagi, *Senior Member, IEEE*

***Abstract*— We study the role of interactivity in distributed statistical inference under information constraints, e.g., communication constraints and local differential privacy. We focus on the tasks of goodness-of-fit testing and estimation of discrete distributions. From prior work, these tasks are well understood under noninteractive protocols. Extending these approaches directly for interactive protocols is difficult due to correlations that can build due to interactivity; in fact, gaps can be found in prior claims of tight bounds of distribution estimation using interactive protocols. We propose a new approach to handle this correlation and establish a unified method to establish lower bounds for both tasks. As an application, we obtain optimal bounds for both estimation and testing under local differential privacy and communication constraints. We also provide an example of a natural testing problem where interactivity helps.**

***Index Terms*— Distributed algorithms, inference algorithms, estimation, statistics, minimax techniques, data privacy, communication channels.**

I. Introduction

CLASSICAL statistics focuses on algorithms that are data-efficient. Recent years have seen revived interest in a different set of constraints for distributed statistics: local constraints on the amount of *information* that can be extracted from each data point. These local constraints can be communication constraints, where each data point must be expressed using a fixed number of bits; privacy constraints, where each user holding a sample seeks to reveal as little as possible about it; and many others, such as noisy communication channels, limited types of measurements, or quantization schemes. Our focus in this work is on statistical inference under such local constraints, when interactive protocols are allowed.

We study the strengths and limitations of interactivity for statistical inference under local information constraints for two fundamental inference tasks for discrete distributions: *learning* (density estimation) and *identity testing* (goodness-of-fit) under total variation distance. For these tasks, prior work gives a good understanding of the number of samples needed in noninteractive setting, including a precise dependence on the information constraints under consideration. However, the following question remains largely open:

> *Does interactivity help for learning and testing in total variation distance when the data is subject to local information constraints, and, if so, for which type of constraints?*

In this work, we resolve this question by establishing lower bounds that hold for general channel families (modeling local information constraints). We show that interaction does not help for learning and testing under communication constraints or local privacy constraints. Several prior works have claimed a subset of these results, but we exhibit technical gaps in most of them (with the important exceptions of [12] and [16], which both obtain a tight bound for testing under local privacy constraints). These gaps stem from the difficulty in handling the correlation that builds due to interaction. Our lower bound explicitly handles this correlation and is based on examining how effectively one can exploit this correlation in spite of the local constraints. Furthermore, our lower bounds allow us to identify a family of channels for which interaction strictly helps in identity testing, establishing the first separation between interactive and noninteractive protocols for distributed goodness-of-fit.

A. The Setting

We now describe the general framework of distributed inference under local information constraints and then specialize it to two canonical tasks: estimation and testing.

The general setting is captured in Fig. 1. There are n users, each of which observes an independent sample from an unknown distribution $\mathbf{p}$ over $[2k] = \{1, 2, \ldots, 2k\}$.[1] Each user is constrained in the amount of information they can reveal about their input. This constraint for user t is described by a channel $W_t \colon [2k] \to \mathcal{Y}$, which is a randomized function from $[2k]$ to the message space $\mathcal{Y}$.[2] In general, we will

Manuscript received January 6, 2021; accepted September 30, 2021. Date of publication October 27, 2021; date of current version December 23, 2021. The work of Jayadev Acharya was supported in part by the Grant NSF-CCF-1846300 (CAREER) and Grant NSF-CCF-1815893 and in part by Google Faculty Fellowship. Part of this work was performed while Clément L. Canonne was supported by the Goldstine Postdoctoral Fellowship, IBM Research. The work of Yuhan Liu and Ziteng Sun was supported by the Grant NSF-CCF-1846300 (CAREER). The work of Himanshu Tyagi was supported in part by the Research Grant from the Robert Bosch Center for Cyberphysical Systems (RBCCPS), Indian Institute of Science, Bengaluru. An earlier version of this paper was presented at the IEEE International Symposium on Information Theory (ISIT) 2021 [3] [DOI: 10.1109/ISIT45174.2021.9518069]. *(Corresponding author: Clément L. Canonne.)*

Jayadev Acharya, Yuhan Liu, and Ziteng Sun are with the School of Electrical and Computer Engineering, Cornell University, Ithaca, NY 14850 USA (e-mail: acharya@cornell.edu; yl2976@cornell.edu; zs335@cornell.edu).

Clément L. Canonne is with the School of Computer Science, The University of Sydney, Sydney, NSW 2006, Australia (e-mail: clement.canonne@sydney.edu.au).

Himanshu Tyagi is with the Department of Electrical Communication Engineering, Indian Institute of Science, Bengaluru 560012, India (e-mail: htyagi@iisc.ac.in).

Communicated by M. Raginsky, Associate Editor for Probability and Statistics.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TIT.2021.3123905.

Digital Object Identifier 10.1109/TIT.2021.3123905

[1] For convenience, we assume throughout the paper that the domain $\mathcal{X}$ has even cardinality; specifically $\mathcal{X} = [2k]$. This is merely for the ease of notation, and all results apply to any finite domain $\mathcal{X}$.

[2] Throughout, we use the information-theoretic notion of a channel and use the standard notation $W(y \mid x)$ for the probability with which the output is y when the input is x.

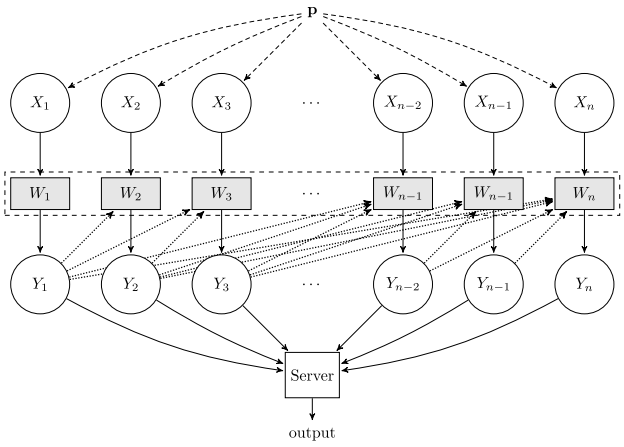

Fig. 1. The information-constrained distributed model. In the private-coin setting the channels $W_1, \ldots, W_n$ are independent, while in the public-coin setting they are jointly randomized, and in the interactive setting W_t can also depend on the previous messages $Y_1, \ldots, Y_{t-1}$ (dotted, upwards arrows).

consider a set of channels $\mathcal{W}$ from which each user's channel must be selected; this family of "allowed channels" models the local information constraints under consideration. This is a very general setting, which captures communication and local privacy constraints as special cases, as we elaborate next.

Communication constraints. Let $\mathcal{W}_\ell := \{W\colon [2k] \to \{0,1\}^\ell\}$ be the set of channels whose output alphabet $\mathcal{Y}$ is the set of all ℓ-bit strings. This captures the constraint where the message from each user is at most ℓ bits: that is, each user has a stringent bandwidth constraint.

Local differential privacy constraints. For a privacy parameter $\varrho > 0$, a channel $W\colon [2k] \to \{0,1\}^*$ is *ϱ-locally differentially private* [28], [30], [38] if

$$\frac{W(y \mid x_1)}{W(y \mid x_2)} \leq e^{\varrho}, \quad \forall x_1, x_2 \in [2k], \forall y \in \{0,1\}^*.$$

Loosely speaking, no output message from a user can reveal too much about their sample. We denote by $\mathcal{W}_\varrho$ the set of all ϱ-locally differentially private (ϱ-LDP) channels.

We emphasize that, although these two constraints will be our leading examples, our formulation of local information constraints captures many more settings. As an example, choosing message output $\mathcal{Y} = [2k] \cup \{\perp\}$ and $\mathcal{W}$ to be the set $W\colon [2k] \to \mathcal{Y}$ of the form $W(x \mid x) = \eta_x$, $W(\perp \mid x) = 1 - \eta_x$ for various sequences $(\eta_x)_{x\in[2k]}$ lets one model *erasure channels*. As another example, one can choose $\mathcal{Y} = \{0,1\}$, and let $\mathcal{W}$ to be the set of channels of the form $W(1 \mid x) = \mathbb{1}_{\{x \leq \tau\}}$, *i.e.*, of *threshold measurements*.

We now return to the description of distributed inference protocols under local information constraints described by $\mathcal{W}$. Once the channel $W_t \in \mathcal{W}$ at user t is decided, the message of user t is $y \in \mathcal{Y}$ with probability $W_t(y \mid X_t)$. The transcript of n messages, $Y^n = (Y_1, \ldots, Y_n)$, is observed by a server $\mathcal{R}$, whose goal is to perform some inference task based on the messages. We consider three classes of protocols, classified depending on how the channels are allowed to be chosen.[3]

[3]In what follows, "SMP" stands for *simultaneous-message passing*, *i.e.*, for noninteractive, one-shot protocol.

Private-coin noninteractive (SMP) protocols. Let $U_1, \ldots, U_n$ be independent random variables which are independent jointly of $(X_1, \ldots, X_n)$. U_t is available only at user t and W_t is chosen as a function of U_t. Therefore, the outputs of the channels are independent of each other.

Public-coin noninteractive (SMP) protocols. Let U be a random variable independent of $(X_1, \ldots, X_n)$. All users are given access to U, and they select their respective channels $W_t \in \mathcal{W}$ as a function of U. We note that the outputs of the channels are independent given U.

Sequentially interactive protocols. Let U be a random variable independent of $(X_1, \ldots, X_n)$. In an *interactive* protocol, all users are given access to U, and user t selects their respective channel $W_t \in \mathcal{W}$ as a function of $(Y_1, \ldots, Y^{t-1}, U)$. We will often make this dependence on previous messages explicit by writing $W^{Y^t,U}$ or as W^{Y^t} when U is fixed (see Section II).

Henceforth, we will interchangeably use "interactive" and "sequentially interactive," and will often omit to specify "noninteractive" when mentioning public- and private-coin protocols. Note that private-coin protocols are a subset of public-coin protocols which in turn are a subset of interactive protocols.

We now define information-constrained discrete distribution estimation and uniformity testing. For a discrete domain $\mathcal{X}$, let $\Delta_{\mathcal{X}}$ be the simplex of distributions over $\mathcal{X}$. Throughout this paper we consider $\mathcal{X} = [2k]$, and denote $\Delta_{[2k]}$ by Δ_{2k}.

Distribution learning. In the $(2k, \varepsilon)$-distribution learning problem (under constraints $\mathcal{W}$), we seek to estimate an unknown distribution $\mathbf{p}$ over $\mathcal{X} = [2k]$ to within ε in total variation distance (defined in Eq. (4)). Formally, a protocol $\Pi\colon [2k]^n \times \mathcal{U} \to \mathcal{Y}^n$ (using $\mathcal{W}$) and an estimator mapping $\hat{\mathbf{p}}\colon \mathcal{Y}^n \times \mathcal{U} \to \Delta_{2k}$ constitute an (n, ε)-estimator using $\mathcal{W}$ if

$$\sup_{\mathbf{p}\in\Delta_{2k}} \Pr_{X^n\sim\mathbf{p}} \left[\mathrm{d}_{\mathrm{TV}} \left(\hat{\mathbf{p}}(Y^n, U), \mathbf{p}\right) > \varepsilon \right] \leq \frac{1}{100}, \quad (1)$$

where $Y^n = \Pi(X^n, U)$ and $\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{q})$ denotes the total variation distance between $\mathbf{p}$ and $\mathbf{q}$. Namely, given the transcript (Y^n, U) of the protocol Π run on the samples X^n, $\hat{\mathbf{p}}$ estimates the input distribution $\mathbf{p}$ to within distance ε with probability at least $99/100$ (this choice of probability is arbitrary and has been chosen for convenience in the proof of Lemma 12). The *sample complexity* of (k, ε)-distribution learning using $\mathcal{W}$ is then the least n such that there exists an (n, ε)-estimator using $\mathcal{W}$.

Identity and uniformity testing. In the $(2k, \varepsilon)$-identity testing problem (under constraints $\mathcal{W}$), given a known reference distribution $\mathbf{q}$ over $[2k]$, and samples from an unknown $\mathbf{p}$, we seek to test if $\mathbf{p} = \mathbf{q}$ or if it is ε-far from $\mathbf{q}$ in total variation distance. Specifically, an (n, ε)-test using $\mathcal{W}$ is given by a protocol $\Pi\colon [2k]^n \times \mathcal{U} \to \mathcal{Y}^n$ (using $\mathcal{W}$) and a randomized decision function $T\colon \mathcal{Y}^n \times \mathcal{U} \to \{0,1\}$

TABLE I

LOWER BOUNDS FOR LOCAL INFORMATION-CONSTRAINED LEARNING AND TESTING. THE PUBLIC- AND PRIVATE-COIN BOUNDS WERE KNOWN FROM PREVIOUS WORK; THE INTERACTIVE BOUNDS ALL FOLLOW FROM OUR RESULTS. THE BOUND MARKED BY A (†) WAS PREVIOUSLY ESTABLISHED IN [12], [16]

	Learning			Testing		
	Private-Coin	Public-Coin	Interactive	Private-Coin	Public-Coin	Interactive
General	$\frac{k}{\varepsilon^2}\cdot\frac{k}{\|\mathcal{W}\|_*}$			$\frac{\sqrt{k}}{\varepsilon^2}\cdot\frac{k}{\|\mathcal{W}\|_*}$	$\frac{\sqrt{k}}{\varepsilon^2}\cdot\frac{\sqrt{k}}{\|\mathcal{W}\|_F}$	$\frac{\sqrt{k}}{\varepsilon^2}\cdot\frac{\sqrt{k}}{\sqrt{\|\mathcal{W}\|_*\|\mathcal{W}\|_{\mathrm{op}}}}$
Communication	$\frac{k}{\varepsilon^2}\cdot\frac{k}{2^\ell}$			$\frac{\sqrt{k}}{\varepsilon^2}\cdot\frac{k}{2^\ell}$	$\frac{\sqrt{k}}{\varepsilon^2}\cdot\sqrt{\frac{k}{2^\ell}}$	$\frac{\sqrt{k}}{\varepsilon^2}\cdot\sqrt{\frac{k}{2^\ell}}$
Privacy	$\frac{k}{\varepsilon^2}\cdot\frac{k}{\varrho^2}$			$\frac{\sqrt{k}}{\varepsilon^2}\cdot\frac{k}{\varrho^2}$	$\frac{\sqrt{k}}{\varepsilon^2}\cdot\frac{\sqrt{k}}{\varrho^2}$	$\frac{\sqrt{k}}{\varepsilon^2}\cdot\frac{\sqrt{k}}{\varrho^2}$ (†)
Leaky-Query	$\frac{k}{\varepsilon^2}\cdot\sqrt{k}$			$\frac{\sqrt{k}}{\varepsilon^2}\cdot\sqrt{k}$	$\frac{\sqrt{k}}{\varepsilon^2}\cdot\sqrt{k}$	$\frac{\sqrt{k}}{\varepsilon^2}\cdot\sqrt[4]{k}$

such that

$$\Pr_{X^n\sim\mathbf{q}^n}\left[T(Y^n,U)=0\right]\geq\frac{99}{100},$$
$$\inf_{\mathbf{p}:\mathrm{d}_{\mathrm{TV}}(\mathbf{p},\mathbf{q})\geq\varepsilon}\Pr_{X^n\sim\mathbf{p}^n}\left[T(Y^n,U)=1\right]\geq\frac{99}{100},\quad(2)$$

where $Y^n=\Pi(X^n,U)$. In other words, after running the protocol Π on independent samples X^n and public coins U, a decision function T is applied to the transcript (Y^n,U) of the protocol. Overall, the protocol should "accept" with high constant probability if the samples come from the reference distribution $\mathbf{q}$ and "reject" with high constant probability if they come from a distribution significantly far from $\mathbf{q}$. Once again, note that the choice of $1/100$ for probability of error is for convenience.[4] Identity testing for the uniform reference distribution $\mathbf{u}$ over $[2k]$ is termed the $(2k,\varepsilon)$-*uniformity testing* problem, and the *sample complexity* of $(2k,\varepsilon)$-uniformity testing using $\mathcal{W}$ is the least n for which there exists an (n,ε)-test using $\mathcal{W}$ for $\mathbf{u}$.

Remark 1: We note that our results are phrased in terms of sample complexity, *i.e.*, the number of users required to perform the corresponding task. Equivalently, this corresponds to minimax lower bounds on rates of convergence (for estimation) or critical radius (for testing).

B. Our Results

The lower bounds we develop associate to each channel $W\colon[2k]\to\mathcal{Y}$ a k-by-k positive semidefinite matrix $H(W)$, which we term the *channel information matrix* (see Eq. (6)), which captures the "informativeness" of the channel W. The spectrum of these matrices $H(W)$, for $W\in\mathcal{W}$, will play a central role in our results. In particular, for a given family of local constraints $\mathcal{W}$, the following quantities will be used:

$$\|\mathcal{W}\|_{\mathrm{op}}:=\max_{W\in\mathcal{W}}\|H(W)\|_{\mathrm{op}},\quad\text{(maximum operator norm)}$$
$$\|\mathcal{W}\|_{*}:=\max_{W\in\mathcal{W}}\|H(W)\|_{*},\quad\text{(maximum nuclear norm)}$$
$$\|\mathcal{W}\|_{F}:=\max_{W\in\mathcal{W}}\|H(W)\|_{F}.\quad\text{(maximum Frobenius norm)}$$

Two key inequalities to interpret our results are

$$\|\mathcal{W}\|_F^2\leq\|\mathcal{W}\|_{\mathrm{op}}\|\mathcal{W}\|_*\ \text{ and }\ \|\mathcal{W}\|_{\mathrm{op}}\leq\|\mathcal{W}\|_F\leq\|\mathcal{W}\|_*,\quad(3)$$

which follow from Hölder's inequality and monotonicity of norms, respectively.

Our results are summarized in Table I; we describe and discuss them in more detail below.

1) Learning: Our first result concerns distribution learning. We establish a new technical lemma which relates the mutual information between the parameters of the distribution to learn and the (adaptively chosen) messages sent by the users to the nuclear norm $\|\mathcal{W}\|_*$ of the local constraints (Theorem 13). This key result, combined with an Assouad-type bound for interactive protocols, yields the following:

Theorem 2: The sample complexity of $(2k,\varepsilon)$-distribution learning under local constraints $\mathcal{W}$ using interactive protocols is

$$\Omega\left(\frac{k^2}{\varepsilon^2\|\mathcal{W}\|_*}\right).$$

This bound matches the known lower bound for learning with noninteractive *private-coin* protocols in [7].

LDP and communication-limited learning. We now apply this to local differential privacy (LDP) and communication constraints. While bounds for these two constraints were presented in prior work, the proofs unfortunately break down for interactive protocols (see Section I-C).

Corollary 3: Let $\varrho\in(0,1]$. The sample complexity of interactive $(2k,\varepsilon)$-distribution learning under ϱ-LDP channels $\mathcal{W}_\varrho$ is

$$\Omega\left(\frac{k^2}{\varepsilon^2\varrho^2}\right).$$

Proof: This follows from $\|\mathcal{W}_\varrho\|_*=O(\varrho^2)$, which was seen in [7, Lemma V.5]. □

[4] In other words, we seek to solve the composite hypothesis testing problem with null hypothesis $\mathcal{H}_0=\{\mathbf{q}\}$ and composite alternative given by $\mathcal{H}_1=\{\,\mathbf{q}'\in\Delta_{2k}:\mathrm{d}_{\mathrm{TV}}(\mathbf{q}',\mathbf{q})\geq\varepsilon\,\}$ in a minimax setting, with both type-I and type-II errors set to $1/100$.

Corollary 4: For $1 \le \ell \le \log k$, the sample complexity of interactive $(2k, \varepsilon)$-distribution learning under communication constraints $\mathcal{W}_\ell$ is

$$\Omega\left(\frac{k^2}{\varepsilon^2 2^\ell}\right).$$

Proof: This follows from $\|\mathcal{W}_\ell\|_* \le 2^\ell$, which was seen in [7, Lemma V.1]. □

Both Corollaries 3 and 4 are optimal up to constant factors. In fact there exist noninteractive private-coin protocols that achieve these bounds (see references in Section I-C), showing that for learning with communication and LDP constraints, interactive protocols are no more powerful than noninteractive ones.

Learning under ℓ_2 distance. Finally, we note that one can instantiate the distribution learning question in Eq. (1) with other distance measures than total variation, *e.g.*, the ℓ_2 distance defined by $\ell_2(\mathbf{p}_1, \mathbf{p}_2) = \|\mathbf{p}_1 - \mathbf{p}_2\|_2$. Our results on total variation distance readily imply the following corollary for ℓ_2, which retrieves the two lower bounds from [13], [14] and matches the bounds of [15], [25] for LDP in the noninteractive case.

Corollary 5: For $1 \le \ell \le \log k$ and $\varrho \in (0, 1]$, the sample complexities of interactive $(2k, \varepsilon)$-distribution learning in ℓ_2 distance under constraints $\mathcal{W}_\ell$ and $\mathcal{W}_\varrho$ are

$$\Omega\left(\frac{k}{\varepsilon^2 2^\ell} \wedge \frac{1}{\varepsilon^4 2^\ell}\right) \quad \text{and} \quad \Omega\left(\frac{k}{\varepsilon^2 \varrho^2} \wedge \frac{1}{\varepsilon^4 \varrho^2}\right),$$

respectively.

Details can be found in Section IV.

2) Testing: Our next result, proved in Section V, is a general lower bound for uniformity testing (and thus, *a fortiori*, on the more general problem of identity testing).[5]

Theorem 6: The sample complexity of $(2k, \varepsilon)$-uniformity testing under local constraints $\mathcal{W}$ using interactive protocols is

$$\Omega\left(\frac{k}{\varepsilon^2 \sqrt{\|\mathcal{W}\|_{\mathrm{op}} \|\mathcal{W}\|_*}}\right).$$

[7] previously established an $\Omega\left(\frac{k}{\varepsilon^2 \|\mathcal{W}\|_F}\right)$ lower bound for (noninteractive) public-coin protocols.

a) LDP and communication-limited testing: We now apply this to common local constraints.

Corollary 7: Let $\varrho \in (0, 1]$. The sample complexity of interactive $(2k, \varepsilon)$-uniformity testing under ϱ-LDP channels $\mathcal{W}_\varrho$ is

$$\Omega\left(\frac{k}{\varepsilon^2 \varrho^2}\right).$$

Proof: This follows from Theorem 6 and the fact that $\|\mathcal{W}_\varrho\|_{\mathrm{op}} \asymp \|\mathcal{W}_\varrho\|_F \asymp \|\mathcal{W}_\varrho\|_* = O(\varrho^2)$ shown in [7, Lemma V.5]. □

Corollary 8: Let $1 \le \ell \le \log k$. The sample complexity of interactive $(2k, \varepsilon)$-uniformity testing under communication constraints $\mathcal{W}_\ell$ is

$$\Omega\left(\frac{k}{\varepsilon^2 2^{\ell/2}}\right).$$

[5]As uniformity testing is a special case of identity testing, lower bounds for the former problem imply worst-case lower bounds for the latter.

Proof: This follows from $\|\mathcal{W}_\ell\|_* \le 2^\ell$ ([7, Lemma V.1]) and $\|\mathcal{W}_\ell\|_{\mathrm{op}} \le 2$ from Lemma 19 in Section V-D. □

Both Corollaries 7 and 8 are tight up to constant factors, as they are in particular achieved by (noninteractive) public-coin protocols [1], [8]). This shows that for communication and local privacy constraints, interactive protocols are no more powerful than public-coin protocols, which are themselves more powerful than private-coin protocols.

b) A separation: By relations between matrix norms (3), it can be seen that the noninteractive public-coin lower bound of $\Omega\left(\frac{k}{\varepsilon^2 \|\mathcal{W}\|_F}\right)$ from [6] can be up to a $k^{1/4}$ factor smaller than the bound in Theorem 6 for interactive protocols. Guided by the analysis of the proof of Theorem 6, we show that this maximal separation is achievable, and in particular demonstrate a separation between noninteractive and interactive protocols for uniformity testing. (see Section V-D for details).

Theorem 9: There exists a natural family of constraints, which we term *leaky-query* channels, under which the sample complexity of $(2k, \varepsilon)$-uniformity testing for noninteractive public-coin protocols and interactive protocols are $\Theta(k/\varepsilon^2)$ and $\Theta(k^{3/4}/\varepsilon^2)$, respectively.

c) Power of the proof: Finally, we emphasize that $\sqrt{\|\mathcal{W}\|_{\mathrm{op}} \|\mathcal{W}\|_*}$ is a convenient, easy-to-apply bound which is optimal for the channel families considered above. However, the power of our techniques goes beyond that specific evaluation. To show this, we provide in Section V-C a family of partial erasure constraints $\mathcal{W}_\perp$ for which the bound given in Theorem 6 is loose, and for which interactivity does not help. Yet, while the general bound given in the statement of the theorem is not tight, the proof of Theorem 6, instantiated with this specific family $\mathcal{W}_\perp$ in mind, readily gives the correct bound.

C. Prior Work

There is a vast literature on statistical inference under LDP and communication constraints. We discuss some of these works below, focusing on those most relevant to ours.

Several protocols have been proposed for discrete distribution estimation and testing under communication and privacy constraints. To the best of our knowledge, all these schemes are noninteractive. [9], [10], [26], [29], [37], [43], [44] provide schemes under LDP, and [4], [5], [33], [34] provide estimation schemes under communication constraints. [15] considers estimation schemes under LDP in the ℓ_2 distance. [1], [12], [16] consider distribution testing under various privacy constraints, and [5]–[8], [31] study distribution testing under several communication constraints. [2] focuses on the role of shared randomness in distributed testing under information constraints. Most relevant to this paper is prior work by a subset of the authors [7] which provides a unifying view of lower bounds under information constraints in the noninteractive setting. We build on this work here.

1) Interactive Testing and Estimation of Discrete Distributions: We now describe prior work on distribution testing and learning for discrete distributions in the interactive setting. We focus on the papers that obtain or claim similar results

as ours. We point out the technical flaws in some of the prior work and outline the state of the art.

[27] (also, see preprint [25]) state lower bounds on distributed estimation of several families of distributions under LDP constraints. While their results hold true in one-dimensional settings and for noninteractive protocols, a crucial component of their proof of a private analogue of Assouad's method ([27, Proposition 3]) is their claim that, under a marginal mixture distribution they consider, the distribution of the sample is independent from the previous messages; in particular, this claim is used to show [27, Theorem 3, supplemental (12)]. This claim of independence simplifies the analysis and takes care of several dependency structures that might arise in sequentially interactive protocols. However, this key identity only holds for noninteractive protocols, not in general, even in the absence of any local constraint.

In another direction, [32] claim lower bounds on distributed estimation (in total variation distance) of several families of distributions under communication constraints. In [34], the authors claim lower bounds on distribution estimation under the ℓ_2 distance as well. The arguments of [32] and [34] appear to both rely on a particular flawed step stated as [34, Lemma 3], which essentially reduces their problem to the noninteractive setting. But this step does not hold in the interactive setting. Following an earlier version of this work, made available as preprint, the authors of [34] were able to mend their proofs by leveraging the techniques developed in the present paper.

Turning to testing, both [12] and [16] establish optimal lower bounds on uniformity testing under LDP constraints.[6,7] In particular, this implies that the separation between private-coin and public-coin noninteractive LDP protocols shown in [6] does not increase when allowing sequential interactivity. However, we note that the results in these paper do not extend to general constraints. Furthermore, even when we try to use their techniques to obtain general bounds, we only get bounds as a function of $\|\mathcal{W}\|_*$ alone. This turns out to be optimal for LDP constraints, but would lead a suboptimal bound for other types of constraints, such as communication constraints (where it would yield a denominator of 2^ℓ instead of the optimal $2^{\ell/2}$).

Several works have studied distribution estimation under the ℓ_2 distance [13], [14], [17] for parametric families of distributions. [13], [14] develop Fisher information-based methods to obtain these bounds, and one of the distribution families they consider is the class of discrete distributions. For sequentially interactive protocols and this particular class, our lower bounds for estimation under the total variation distance imply their results. We point out, nevertheless, that their bounds apply to a larger class of protocols (*blackboard* protocols, which allow for multiple rounds of messages from each user), and therefore hold in a more general setting than ours. However, it is important to remark that these techniques do not suffice for the total variation distance setting, and more importantly, unlike the bounds claimed in [34], cannot give bounds for testing.

Slightly further from the setting considered here, [22][8] and [21] also consider distributed estimation and identity testing of discrete distributions, respectively, under total variation distance. Although their results apply to blackboard protocols, their setting and results are incomparable to ours as they consider constraints on the *total* communication sent by all the users.

In a different direction, recent work of [18] considers the task of testing whether a quantum state is *maximally mixed*, the quantum analogue of uniformity testing. They focus on the setting where one is only allowed local measurements (*i.e.*, without entanglement) and provide a lower bound showing a separation between sequentially interactive "local measurement" protocols and the more general fully entangled ones. We note that, while the setting differs from the one we consider here, some of the considerations are similar, and there is a direct analogy between their techniques and those of [6].

2) General Interactive Testing and Estimation Bounds: Several papers have studied the role of interactivity for specific estimation tasks, establishing separation results under either local privacy or communication constraints.

The study of interactivity in LDP started with [38] who designed a learning task that requires exponentially fewer samples with interactive protocols than with noninteractive ones. Moreover, this separation can manifest itself in very natural optimization or learning problems [20], [41], [42]. [35] and [36] study the relation between sequentially interactive protocols and fully interactive protocols (where the same user can send multiple messages), establishing both relations and strong separations in sample complexity between the two settings. [24], drawing on machinery from the communication complexity literature, develop a lower bound results which apply to any locally private estimation protocol (regardless of the interactivity model). [40] studies various estimation tasks under a range of information constraints. Finally, [19] establish a separation between interactive and noninteractive learning for large-margin classifiers, under both local privacy and communication constraints.

II. Preliminaries

Hereafter, we write $\log$ and $\ln$ for the binary and natural logarithms, respectively. We will consider probability distributions over $[2k]$ which we identify with their probability mass functions $\mathbf{p}\colon [2k] \to [0,1]$ satisying $\sum_{x\in\mathcal{X}} \mathbf{p}(x) = 1$. We denote by Δ_{2k} the set of all such probability distributions. We denote by $\mathbf{u}$ the uniform distribution over $[2k]$.

[6]The conference version of [12] had a flaw on the claim that for any fixed setting of the previous messages Y^{t-1}, the random variables Y_t and Z_i (a parameter of their lower bound construction) are independent conditioned on $X_t \notin \{2i-1, 2i\}$. This is not true in general, as it overlooks some conditional dependencies that may arise. However, after this was brought to their attention, the authors were able fix the gap in their argument, using techniques that bear some resemblance to the ones used in the present paper [11].

[7]The result of [16] is actually phrased in a more general way, as they address the more general problem of *identity testing*, where the reference distribution need not be uniform and the lower bound quantitatively depends on the reference distribution itself.

[8]To the best of our knowledge, the details of the proofs of [22] have not been made publicly available, and as such we have not been able to assess correctness of the results claimed in this paper.

For two distributions $\mathbf{p}_1, \mathbf{p}_2$ over $\mathcal{X}$, denote their total variation distance by

$$\mathrm{d}_{\mathrm{TV}}(\mathbf{p}_1, \mathbf{p}_2) := \sup_{S \subseteq \mathcal{X}} (\mathbf{p}_1(S) - \mathbf{p}_2(S)), \tag{4}$$

and their Kullback–Leibler divergence and chi square divergence, respectively, by

$$\mathrm{D}(\mathbf{p}_1 \| \mathbf{p}_2) := \sum_{x \in \mathcal{X}} \mathbf{p}_1(x) \log \frac{\mathbf{p}_1(x)}{\mathbf{p}_2(x)}$$

and

$$\mathrm{d}_{\chi^2}(\mathbf{p}_1 \,||\, \mathbf{p}_2) := \sum_{x \in \mathcal{X}} \frac{(\mathbf{p}_1(x) - \mathbf{p}_2(x))^2}{\mathbf{p}_2(x)}.$$

By Pinsker's inequality and concavity of logarithm, these quantities obey the inequalities:

$$\mathrm{d}_{\mathrm{TV}}(\mathbf{p}_1, \mathbf{p}_2)^2 \leq \frac{\ln 2}{2} \mathrm{D}(\mathbf{p}_1 \| \mathbf{p}_2) \leq \frac{\ln 2}{2} \mathrm{d}_{\chi^2}(\mathbf{p}_1 \,||\, \mathbf{p}_2).$$

Throughout, We use the standard asymptotic notation $O(f)$, $\Omega(f)$, $\Theta(f)$. In addition, we will often write $a_n \lesssim b_n$ (resp. $a_n \gtrsim b_n$), to indicate there exists an absolute constant $C > 0$ such that $a_n \leq C \cdot b_n$ (resp. $a_n \geq C \cdot b_n$) for all n, and accordingly write $a_n \asymp b_n$ when both $a_n \lesssim b_n$ and $a_n \gtrsim b_n$.

A. Interactive Protocols

We set up some notation for sequentially interactive protocols, defined in Section I-A. When public-coin U is fixed constant, we will call the protocol a *deterministic protocol*.[9] Recall that in interactive protocols, user t selects its channel $W \in \mathcal{W}$ as a function of (Y^{t-1}, U). We denote this channel by $W^{Y^{t-1},U}$, or simply by $W^{Y^{t-1}}$ for deterministic protocols, and the corresponding output by Y_t.

We call (Y^n, U) the *transcript* of the protocol, which is used to complete the inference task. For a fixed protocol Π, when the input X^n has distribution $\mathbf{p}^n$, we denote the distribution of the transcript by $\mathbf{p}_\Pi^{Y^n,U}$. In fact, we often omit the dependence on the protocol from our notation (since it will be clear from the context) and simply use $\mathbf{p}^{Y^n,U}$. For deterministic protocols, $\mathbf{p}^{Y_t|Y^{t-1}}$ will be used to denote the conditional distribution of the message Y_t of the tth user, conditioned on the past messages $Y^{t-1} = (Y_1, \ldots, Y_{t-1})$.

B. Lower Bound Construction

Our lower bounds rely on a family of perturbed distributions around $\mathbf{u}$, a common starting point for establishing several statistical lower bounds. The particular construction we use is from [39] and consists of 2^k distributions parameterized by $\mathcal{Z} = \{-1, +1\}^k$. Specifically, for $z \in \mathcal{Z}$ the distribution $\mathbf{p}_z$ over $[2k]$ is given by

$$\mathbf{p}_z = \frac{1}{2k}(1 + 4\varepsilon z_1, 1 - 4\varepsilon z_1, \ldots, 1 + 4\varepsilon z_t, 1 - 4\varepsilon z_t, \ldots, 1 + 4\varepsilon z_k, 1 - 4\varepsilon z_k). \tag{5}$$

Each such $\mathbf{p}_z$ is therefore at total variation exactly 2ε from $\mathbf{u}$.

[9]This is a slight abuse of notation, since randomness is used by the channels to generate their (random) output.

C. The Channel Information Matrix

We capture the information revealed by a channel about the distribution of its input in terms of a matrix $H(W)$, which was defined in [7, Definition I.5].

Specifically, for a channel $W\colon [2k] \to \mathcal{Y}$, the associated *information matrix* $H(W)$ is the k-by-k positive semi-definite (p.s.d.) matrix $H(W)$ whose entry $H(W)_{i,j}$ is given by

$$\sum_{y \in \mathcal{Y}} \frac{(W(y \mid 2i-1) - W(y \mid 2i))(W(y \mid 2j-1) - W(y \mid 2j))}{\sum_{x \in [2k]} W(y \mid x)} \tag{6}$$

for $i, j \in [k]$. This matrix captures the ability of the channel output to distinguish between consecutive even and odd inputs, and is thus particularly tailored to the Paninski perturbed family defined above. However, the ordering of the elements is arbitrary and we can associate this matrix with any partition of the domain into equal parts.

D. Organization

The remainder of the paper is organized as follows. In Sections IV and V we prove our general results on learning and testing. In Section V-D we present a set of channels $\mathcal{W}$ for which interactivity helps when testing using $\mathcal{W}$.

III. Information-Loss Bounds

In this section, we present two bounds relating the loss for estimating Rademacher random variables using correlated observations to mutual information. The bounds are simple and highlighted separately here for easy reference – in essence, they say that *small loss implies large information*, the first step in any information-theoretic lower bound for statistical inference.

First, we consider estimation of a $\{-1, +1\}^k$-valued random vector under the average Hamming loss function $\mathrm{d}_{\mathrm{H}}(u, v) := \sum_{i=1}^k \mathbb{1}_{\{u_i \neq v_i\}}$. Note that $\mathbb{E}[\mathrm{d}_{\mathrm{H}}(U, V)] = \sum_{i=1}^k \Pr[U_i \neq V_i]$.

Lemma 10 (Hamming Loss): Consider random variables (Z, Y) with $Z \in \{-1, 1\}^k$ being a random vector with independent Rademacher entries. Let $\hat{Z}$ be a randomized function of Y, *i.e.*, such that the Markov relation Z—Y—$\hat{Z}$ holds. Then, for each $i \in [k]$, with $h(t) := -t \log t - (1-t)\log(1-t)$ denoting the binary entropy function, we get

$$I(Z_i \wedge Y) \geq 1 - h\left(\Pr\left[Z_i \neq \hat{Z}_i\right]\right),$$

whereby

$$\frac{1}{k} \sum_{i=1}^k I(Z_i \wedge Y) \geq 1 - h\left(\frac{1}{k} \sum_{i=1}^k \Pr\left[Z_i \neq \hat{Z}_i\right]\right).$$

Proof: Using the data processing inequality for mutual information,

$$I(Z_i \wedge Y) = 1 - H(Z_i \mid Y) \geq 1 - H(Z_i \mid \hat{Z}_i) \geq 1 - h\left(\Pr\left[Z_i \neq \hat{Z}_i\right]\right),$$

where the second inequality is by Fano's inequality. Upon taking average over i, we get

$$\frac{1}{k}\sum_{i=1}^{k} I\left(Z_i \wedge Y\right) \geq 1 - \frac{1}{k}\sum_{i=1}^{k} h(\Pr\left[Z_i \neq \hat{Z}_i\right]) \geq 1 - h\left(\frac{1}{k}\sum_{i=1}^{k}\Pr\left[Z_i \neq \hat{Z}_i\right]\right),$$

where the second inequality holds by concavity of $h(\cdot)$. □

Next, we consider the mean squared loss function $\|u-v\|_2^2$ for $u, v \in \{-1,+1\}^k$. Denote by $\mathrm{mmse}(Z \mid Y)$ the minimum mean squared error for estimating Z using Y given by

$$\mathrm{mmse}(Z \mid Y) := \min_{g\colon \mathcal{Y}\to\mathbb{R}^k} \mathbb{E}\Big[\|Z - g(Y)\|_2^2\Big], \tag{7}$$

where the minimum is taken over all randomized functions g of Y. It is well known that the minimum is attained by $g(Y) = \mathbb{E}[Z \mid Y]$.

Lemma 11 (Mean Squared Loss): Consider random variables (Z, Y) with $Z \in \{-1,1\}^k$ being a random vector with independent Rademacher entries. Then, for each $i \in [k]$, we get

$$I\left(Z_i \wedge Y\right) \geq \frac{1}{2\ln 2}\mathbb{E}\Big[\mathbb{E}[Z_i \mid Y]^2\Big],$$

whereby

$$\frac{1}{k}\sum_{i=1}^{k} I\left(Z_i \wedge Y\right) \geq \frac{1}{2\ln 2}\left(1 - \frac{1}{k}\,\mathrm{mmse}(Z \mid Y)\right).$$

Proof: By using $\mathbb{E}[Z_i\mathbb{E}[Z_i \mid Y]] = \mathbb{E}\Big[\mathbb{E}[Z_i \mid Y]^2\Big]$, we obtain

$$1 - \mathbb{E}\Big[\|Z_i - \mathbb{E}[Z_i|Y]\|_2^2\Big] = \mathbb{E}\Big[\mathbb{E}[Z_i \mid Y]^2\Big].$$

Thus, since $\mathrm{mmse}(Z \mid Y) = \sum_{i=1}^k \mathbb{E}\Big[\|Z_i - \mathbb{E}[Z_i|Y]\|_2^2\Big]$, the first inequality in the lemma implies the second.

To see the first inequality, note

$$I\left(Z_i \wedge Y\right) = 1 - H(Z_i \mid Y) = \mathbb{E}\big[D(P_{Z_i|Y}\|\mathtt{unif})\big],$$

where $\mathtt{unif}$ denotes the Rademacher distribution. Thus, by Pinsker's inequality,

$$I\left(Z_i \wedge Y\right) \geq \frac{2}{\ln 2}\mathbb{E}\left[\left(\frac{1}{2} - \Pr\left[Z_i = 1 \mid Y\right]\right)^2\right] = \frac{1}{2\ln 2}\mathbb{E}\Big[\mathbb{E}[Z_i \mid Y]^2\Big],$$

where in the final step we used the observation that for any $\{-1,+1\}$-valued random variable V, $\mathbb{E}[V] = 2\Pr[V=1]-1$. This completes the proof of the lemma. □

IV. Interactive Learning Under Information Constraints

We now prove Theorem 2, a lower bound on the sample complexity for learning using interactive protocols under general information constraints given by a channel family $\mathcal{W}$.

We proceed as in [7] and use the construction in Eq. (5). For each $z \in \{-1,+1\}^k$, let $\mathbf{p}_z \in \Delta_{2k}$ denote the $2k$-ary distribution given in Eq. (5). We use a uniform prior over these distributions to get our lower bound. Specifically, let Z be distributed uniformly over $\{-1,+1\}^k$. Conditioned on Z, let $X_1, \ldots, X_n$ denote independent samples from $\mathbf{p}_Z$. We run a (sequentially) interactive protocol Π which generates the messages (transcript) Y^n taking values in $\mathcal{Y}^n$. The distribution of messages over $\mathcal{Y}^n$ is given by

$$\mathbf{q}^{Y^n} := \frac{1}{2^k}\sum_{z\in\mathcal{Z}} \mathbf{p}_z^{Y^n} \tag{8}$$

In [7], Fano's inequality is used to derive the desired bound. However, this requires us to derive a bound for $I\left(Z \wedge Y^n\right)$, the joint information in the message about the vector Z. As noted in [27], this is a formidable task for interactive communication, since the correlation can be rather complicated. Instead, we exploit the additive structure of total variation distance to obtain an Assouad-type bound below, which relates the loss in total variation function to the *average information* $\frac{1}{k}\sum_{i=1}^k I\left(Z_i \wedge Y^n\right)$.[10]

Lemma 12 (Assouad-Type Bound): Consider local constraints $\mathcal{W}$ and $\varepsilon \in (0,1]$. Let $(\Pi, \hat{p})$ be an $(n, \varepsilon/12)$-estimator using $\mathcal{W}$ and (Y^n, U) be the corresponding transcript. Then, we must have

$$\sum_{i=1}^{k} I\left(Z_i \wedge Y^n \mid U\right) \geq \frac{k}{2}. \tag{9}$$

Proof: The proof involves relating PAC-style guarantees provided by Eq. (1) to an expected Hamming-loss guarantee and then applying the information-loss bound in Lemma 10. Specifically, let

$$\hat{Z} := \operatorname*{argmin}_{z\in\{-1,+1\}^k} \mathrm{d}_{\mathrm{TV}}\left(\mathbf{p}_z, \hat{\mathbf{p}}(Y^n, U)\right).$$

By the triangle inequality,

$$\mathrm{d}_{\mathrm{TV}}\left(\mathbf{p}_{\hat{Z}}, \mathbf{p}_Z\right) \leq \mathrm{d}_{\mathrm{TV}}\left(\hat{\mathbf{p}}(Y^n,U), \mathbf{p}_{\hat{Z}}\right) + \mathrm{d}_{\mathrm{TV}}\left(\hat{\mathbf{p}}(Y^n,U), \mathbf{p}_Z\right) \leq 2\mathrm{d}_{\mathrm{TV}}\left(\hat{\mathbf{p}}(Y^n,U), \mathbf{p}_Z\right)$$

which yields

$$\Pr\left[\,\mathrm{d}_{\mathrm{TV}}\left(\mathbf{p}_{\hat{Z}}, \mathbf{p}_Z\right) > \frac{\varepsilon}{12}\,\right] \leq \frac{1}{100},$$

since $\Pr\left[\,\mathrm{d}_{\mathrm{TV}}\left(\hat{\mathbf{p}}(Y^n,U), \mathbf{p}_Z\right) > \frac{\varepsilon}{10}\,\right] \leq 1/100$ by our assumption for the estimator $(\Pi, \hat{p})$. Noting that $\mathrm{d}_{\mathrm{TV}}\left(\mathbf{p}_z, \mathbf{p}_{z'}\right) \leq 2\varepsilon$ for every $z, z' \in \{-1,+1\}^k$, we get

$$\mathbb{E}\big[\mathrm{d}_{\mathrm{TV}}\left(\mathbf{p}_{\hat{Z}}, \mathbf{p}_Z\right)\big] \leq \frac{99}{100}\cdot\frac{2\varepsilon}{12} + \frac{1}{100}\cdot 2\varepsilon < \frac{\varepsilon}{5}.$$

Next, noting that $\mathrm{d}_{\mathrm{TV}}\left(\mathbf{p}_z, \mathbf{p}_{z'}\right) = (2\varepsilon/k)\sum_{i=1}^k \mathbb{1}_{\{z_i \neq z'_i\}}$, the previous inequality yields

$$\frac{1}{k}\sum_{i=1}^{k}\Pr\left[\,\hat{Z}_i \neq Z_i\,\right] < \frac{1}{10}.$$

The proof is now completed using Lemma 10. □

Upon combining the previous bound with Lemma 12, we obtain the proof of Theorem 2. Interestingly, the same bound will be useful for the testing problem as well, and

[10] Note that $\sum_{i=1}^k I\left(Z_i \wedge Y^n\right) \leq I\left(Z \wedge Y^n\right)$, suggesting that this bound is perhaps more stringent than the Fano-type bound in [7].

is one of the key components of our lower bound recipes in this paper. We provide its formal proof first, followed by remarks on its extension (which will be useful) and heuristics underlying the formal proof.

Theorem 13 (Average Information Bound): For $\varepsilon \in (0, 1/4]$, let (Y^n, U) be the transcript of an interactive protocol using $\mathcal{W}$, when the input is generated using $\mathbf{p}_Z$ from Eq. (5) with a uniform Z. Then, for every $1 \le t \le n$,

$$\frac{1}{k}\sum_{i=1}^{k} I\left(Z_i \wedge Y^t \mid U\right) \le \frac{8t\varepsilon^2}{k^2} \cdot \|\mathcal{W}\|_*.$$

Proof: Since $\sum_{i=1}^{k} I\left(Z_i \wedge Y^t \mid U\right) \le \max_u \sum_{i=1}^{k} I\left(Z_i \wedge Y^t \mid U = u\right)$, it suffices to establish the bound for every fixed realization of U; we will assume that U is a fixed constant and the protocol Π is a deterministic interactive protocol. Fix $1 \le t \le n$ and consider $i \in [k]$. For the distribution in Eq. (5) and $i \in [k]$, let

$$\mathbf{p}_{+i}^{Y^n} := \frac{1}{2^{k-1}} \sum_{z: z_i = +1} \mathbf{p}_z^{Y^n} \quad \text{and} \quad \mathbf{p}_{-i}^{Y^n} := \frac{1}{2^{k-1}} \sum_{z: z_i = -1} \mathbf{p}_z^{Y^n} \tag{10}$$

be distributions over n-message transcripts restricting z_i to be $+1$ or -1. Recalling the definition of $\mathbf{q}^{Y^t}$, from Eq. (8), we can rewrite

$$\mathbf{q}^{Y^t} = \frac{\mathbf{p}_{+i}^{Y^t} + \mathbf{p}_{-i}^{Y^t}}{2}. \tag{11}$$

By the convexity of KL divergence,

$$\begin{aligned} I\left(Z_i \wedge Y^t\right) &= \frac{\mathrm{D}\left(\mathbf{p}_{+i}^{Y^t} \| \mathbf{q}^{Y^t}\right) + \mathrm{D}\left(\mathbf{p}_{-i}^{Y^t} \| \mathbf{q}^{Y^t}\right)}{2} \\ &\le \frac{1}{4}\left(\mathrm{D}\left(\mathbf{p}_{+i}^{Y^t} \| \mathbf{p}_{-i}^{Y^t}\right) + \mathrm{D}\left(\mathbf{p}_{-i}^{Y^t} \| \mathbf{p}_{+i}^{Y^t}\right)\right). \end{aligned}$$

For $z \in \{-1, +1\}^k$, write $z^{\oplus i}$ for z with the ith coordinate flipped. Using the convexity of KL divergence and applying Jensen's inequality to the right-side of the previous bound, we get

$$I\left(Z_i \wedge Y^t\right) \le \frac{1}{2}\left(\frac{1}{2^k} \sum_{z \in \{-1,+1\}^k} \mathrm{D}\left(\mathbf{p}_z^{Y^t} \| \mathbf{p}_{z^{\oplus i}}^{Y^t}\right)\right). \tag{12}$$

Now for any z, z', by the chain rule for KL divergence we have

$$\mathrm{D}\left(\mathbf{p}_z^{Y^t} \| \mathbf{p}_{z'}^{Y^t}\right) = \sum_{r=1}^{t} \mathbb{E}_{\mathbf{p}_z^{Y^{r-1}}}\left[\mathrm{D}\left(\mathbf{p}_z^{Y_r|Y^{r-1}} \| \mathbf{p}_{z'}^{Y_r|Y^{r-1}}\right)\right]. \tag{13}$$

Next, we note that

$$\begin{aligned} &\Pr_{\mathbf{p}_z}\left[Y_r = y \mid Y^{r-1}\right] - \Pr_{\mathbf{p}_{z^{\oplus i}}}\left[Y_r = y \mid Y^{r-1}\right] \\ &\quad = \frac{2\varepsilon z_i}{k}\left(W^{Y^{r-1}}(y \mid 2i-1) - W^{Y^{r-1}}(y \mid 2i)\right). \end{aligned} \tag{14}$$

Indeed, this relation holds since for all z

$$\begin{aligned} &\Pr_{\mathbf{p}_z}\left[Y_r = y \mid Y^{r-1}\right] \\ &= \sum_{j=1}^{k}\Big(\mathbf{p}_z(2j-1)W^{Y^{r-1}}(y|2j-1) + \mathbf{p}_z(2j)W^{Y^{r-1}}(y|2j)\Big) \\ &= \sum_{j\neq i}\Big(\mathbf{p}_z(2j-1)W^{Y^{r-1}}(y|2j-1) + \mathbf{p}_z(2j)W^{Y^{r-1}}(y|2j)\Big) \\ &\quad + \left(\frac{1+4\varepsilon z_i}{2k}W^{Y^{r-1}}(y \mid 2i-1) + \frac{1-4\varepsilon z_i}{2k}W^{Y^{r-1}}(y|2i)\right) \\ &= \sum_{j\neq i}\Big(\mathbf{p}_{z^{\oplus i}}(2i-1)W^{Y^{r-1}}(y|2j-1) + \mathbf{p}_{z^{\oplus i}}(2j)W^{Y^{r-1}}(y|2j)\Big) \\ &\quad + \left(\frac{1-4\varepsilon z_i}{2k}W^{Y^{r-1}}(y|2i-1) + \frac{1+4\varepsilon z_i}{2k}W^{Y^{r-1}}(y|2i)\right) \\ &\quad + \frac{2\varepsilon z_i}{k}\left(W^{Y^{r-1}}(y \mid 2i-1) - W^{Y^{r-1}}(y \mid 2i)\right) \\ &= \Pr_{\mathbf{p}_{z^{\oplus i}}}\left[Y_r = y \mid Y^{r-1}\right] \\ &\quad + \frac{2\varepsilon z_i}{k}\left(W^{Y^{r-1}}(y \mid 2i-1) - W^{Y^{r-1}}(y \mid 2i)\right). \end{aligned}$$

Using (14), we bound $\mathrm{D}\left(\mathbf{p}_z^{Y_r|Y^{r-1}} \| \mathbf{p}_{z^{\oplus i}}^{Y_r|Y^{r-1}}\right)$ as follows. Since the KL divergence is bounded by the chi square distance, we have

$$\begin{aligned} &\mathrm{D}\left(\mathbf{p}_z^{Y_r|Y^{r-1}} \| \mathbf{p}_{z^{\oplus i}}^{Y_r|Y^{r-1}}\right) \\ &\le \sum_{y\in\mathcal{Y}} \frac{\left(\Pr_{\mathbf{p}_z}\left[Y_r = y \mid Y^{r-1}\right] - \Pr_{\mathbf{p}_{z^{\oplus i}}}\left[Y_r = y \mid Y^{r-1}\right]\right)^2}{\Pr_{\mathbf{p}_{z^{\oplus i}}}\left[Y_r = y \mid Y^{r-1}\right]} \\ &\le \frac{16\varepsilon^2}{k} \sum_{y\in\mathcal{Y}} \frac{\left(W^{Y^{r-1}}(y \mid 2i-1) - W^{Y^{r-1}}(y \mid 2i)\right)^2}{\sum_{x\in[2k]} W^{Y^{r-1}}(y \mid x)} \\ &= \frac{16\varepsilon^2}{k} H(W^{Y^{r-1}})_{i,i}, \end{aligned} \tag{15}$$

where we used the observation

$$\begin{aligned} \Pr_{\mathbf{p}_{z^{\oplus i}}}\left[Y_r = y \mid Y^{r-1}\right] &\ge \frac{1-2\varepsilon}{2k} \sum_{x\in[2k]} W^{Y^{r-1}}(y \mid x) \\ &\ge \frac{1}{4k} \sum_{x\in[2k]} W^{Y^{r-1}}(y \mid x). \end{aligned}$$

It follows that

$$\begin{aligned} \sum_{i=1}^{k} I\left(Z_i \wedge Y^t\right) &\le \frac{8\varepsilon^2}{k} \sum_{i=1}^{k}\left(\sum_{r=1}^{t} \mathbb{E}_{\mathbf{p}_z^{Y^{r-1}}}\left[H(W^{Y^{r-1}})_{i,i}\right]\right) \\ &= \frac{8\varepsilon^2}{k} \sum_{r=1}^{t}\left(\mathbb{E}_{\mathbf{p}_z^{Y^{r-1}}}\left[\sum_{i=1}^{k} H(W^{Y^{r-1}})_{i,i}\right]\right) \\ &= \frac{8\varepsilon^2}{k} \sum_{r=1}^{t}\left(\mathbb{E}_{\mathbf{p}_z^{Y^{r-1}}}\left[\left\|H(W^{Y^{r-1}})\right\|_*\right]\right) \\ &\le \frac{8t\varepsilon^2}{k} \cdot \|\mathcal{W}\|_*, \end{aligned}$$

concluding the proof. □

Remark 14 (Information bound for each coordinate): While we have stated the previous result as a bound for average information, our proof gives a bound for information $I\left(Z_i \wedge Y^t\right)$

about each coordinate contained in the message Y^t. Specifically, by Eq. (15) we get that

$$I\left(Z_i \wedge Y^t\right) \leq \frac{8\varepsilon^2}{k} \sum_{r=1}^{t} \mathbb{E}\Big[H(W^{Y^{r-1}})_{i,i}\Big].$$

This stronger form is useful; see Section V-C.

Remark 15 (Is This Bound Tight?): An examination of the proof above suggests that the only seemingly weak bound is Eq. (12). In this step, which is an important ingredient of our proof and perhaps allows us to circumvent the difficulty faced by prior works, we simplify the conditional distribution of Z_i given the past Y^t by conditioning additionally on all the other coordinates $Z^{-i} = (Z_1, \ldots, Z_{i-1}, Z_{i+1}, \ldots, Z_k)$. Our thesis is that until the time t when the ith bit of Z is determined by Y^t, the difficulty in determining Z_i using Y^t is not reduced much even when we condition on all the other bits Z^{-i}. This is a driving heuristic for the bound above.

We conclude this section by showing how our proof of Theorem 13 implies the claimed result on estimation under the ℓ_2 distance, Corollary 5.

Proof of Corollary 5: Note that, by the Cauchy–Schwarz inequality, a lower bound on estimation for distributions over domain $\mathcal{X}$ to total variation distance ε implies a lower bound to ℓ_2 distance $\varepsilon/\sqrt{|\mathcal{X}|}$, *i.e.*, with a square root of the domain size factor loss in the distance parameter. We will use this to derive our lower bounds under ℓ_2 distance: first, by the above it is easy to see that for $0 < \varepsilon \leq \frac{1}{4\sqrt{2k}}$, Theorem 2 implies a lower bound of $\Omega\left(\frac{k}{\varepsilon^2 \|\mathcal{W}\|_*}\right)$ users for learning under any set of constraints $\mathcal{W}$.

However, for larger values of ε, we cannot directly use the result, as $\sqrt{2k}\varepsilon > 1/4$ and our result does not apply. However, we can choose in that case a subset $\mathcal{X}' \subseteq [2k]$ of the domain of size $|\mathcal{X}| = 2\left\lfloor 1/(32\varepsilon^2)\right\rfloor$, and embed our (total variation) lower bound in this domain. One can check that this will indeed result in a lower bound of $\Omega\left(\frac{k}{\varepsilon^2 \|\mathcal{W}\|_*}\right)$ users, for a ℓ_2 distance parameter ε.

Combining the two cases yields a general lower bound for ℓ_2 estimation under $\mathcal{W}$; instantiating the bound to $\mathcal{W}_\varrho$ and $\mathcal{W}_\ell$ yields Corollary 5. □

V. Interactive Testing Under Information Constraints

A. The General Bound: Proof of Theorem 6

We proceed as in [7] and derive a lower bound for testing under information constraints using Le Cam's two-point method. Specifically, let Z be distributed uniformly over $\{-1, +1\}^k$. Note that for any (n, ε)-test (Π, T) for $(2k, \varepsilon)$-identity testing with transcript (Y^n, U), we must have

$$\frac{1}{2} \Pr_{\mathbf{u}^n}[\, T(Y^n, U) = 0\,] + \frac{1}{2}\mathbb{E}\left[\Pr_{\mathbf{p}_Z^n}[\, T(Y^n, U) = 1\,]\right] \geq \frac{99}{100},$$

where $\mathbf{p}_z$ is given by Eq. (5). It follows that we can find a fixed realization of U for which the same bound holds; thus, there exists a deterministic interactive protocol Π' for which the same bound holds. In the remainder of the section, we will assume that our protocol Π is deterministic and denote by $\mathbf{q}^{Y^n}$ and $\mathbf{u}^{Y^n}$, respectively, the probabilities distribution of the transcript under input distribution $\mathbb{E}[\mathbf{p}_Z^n]$ and $\mathbf{u}^n$.

Using standard relations between Bayesian error for binary hypothesis testing with uniform prior and the total variation distance, along with Pinsker's inequality, we get that $\mathrm{D}\left(\mathbf{q}^{Y^n} \| \mathbf{u}^{Y^n}\right) \geq c$ for a constant $c > 0$. It remains to bound this KL divergence, which we do after the following remark.

Remark 16 (Comparison With Decoupled Chi Square Bounds): Before proceeding, we draw contrast with the *decoupled chi square divergence* bound technique developed [7]. Their first step was to bound Kullback–Leibler divergence with chi square divergence and then handle the latter using the so-called "Ingster's method." While very powerful for SMP protocols, this technique requires us to handle the correlation of the vector Y^n directly, which is a formidable task for interactive protocols. Below, we proceed by first applying the chain rule to the Kullback–Leibler divergence to break it into contribution for each sample and then bounding it by the chi square divergence. As will be seen below, this allows us to work with one sample at a time. Further, switching to chi square divergence relates distances between distributions to a bilinear form involving $H(W)$s. Thus, we can relate distances between distributions to the spectrum of $H(W)$, a relation that was exploited to establish a separation between public- and private-coin protocols in [7]. But now we need to handle the posterior distribution of the message Y_t given the past Y^{t-1}, under the mixture distribution.

Proceeding with the proof, by the chain rule for Kullback–Leibler divergence, we can write

$$\mathrm{D}\left(\mathbf{q}^{Y^n} \| \mathbf{u}^{Y^n}\right) = \sum_{t=0}^{n-1} \mathbb{E}_{\mathbf{q}^{Y^t}}\left[\mathrm{D}\left(\mathbf{q}^{Y_{t+1}|Y^t} \| \mathbf{u}^{Y_{t+1}|Y^t}\right)\right] \quad (16)$$

We now present the key technical component of our testing bound in the result below.

Lemma 17 (Per-round divergence bound): For every $0 \leq t \leq n-1$, we have

$$\mathbb{E}_{\mathbf{q}^{Y^t}}\left[\mathrm{D}\left(\mathbf{q}^{Y_{t+1}|Y^t} \| \mathbf{u}^{Y_{t+1}|Y^t}\right)\right] \leq \frac{4(\ln 2)\varepsilon^2}{k} \|\mathcal{W}\|_{\mathrm{op}} \cdot \sum_{i=1}^{k} I\left(Z_i \wedge Y^t\right). \quad (17)$$

Proof: Fix t. As chi-squared divergence upper bounds KL divergence, we have

$$\begin{aligned}
&\mathbb{E}_{\mathbf{q}^{Y^t}}\left[\mathrm{D}\left(\mathbf{q}^{Y_{t+1}|Y^t} \| \mathbf{u}^{Y_{t+1}|Y^t}\right)\right] \\
&\leq \mathbb{E}_{\mathbf{q}^{Y^t}}\left[\mathrm{d}_{\chi^2}\left(\mathbf{q}^{Y_{t+1}|Y^t} \,||\, \mathbf{u}^{Y_{t+1}|Y^t}\right)\right] \\
&= 2k \cdot \mathbb{E}_{\mathbf{q}^{Y^t}}\left[\sum_{y \in \mathcal{Y}} \frac{\left(\sum_x W^{Y^t}(y \mid x)(\mathbf{q}_{X_{t+1}|Y^t}(x) - \frac{1}{2k})\right)^2}{\sum_x W^{Y^t}(y \mid x)}\right].
\end{aligned}$$

Upon noting that, for all $i \in [k]$,

$$\mathbf{q}_{X_{t+1}|Y^t}(2i-1) = \frac{1 + 2\varepsilon\mathbb{E}[Z_i \mid Y^t]}{2k},$$
$$\mathbf{q}_{X_{t+1}|Y^t}(2i) = \frac{1 - 2\varepsilon\mathbb{E}[Z_i \mid Y^t]}{2k},$$

we get

$$\begin{aligned}&\mathbb{E}_{\mathbf{q}^{Y^t}}\left[\mathrm{D}\left(\mathbf{q}^{Y_{t+1}|Y^t}\|\mathbf{u}^{Y_{t+1}|Y^t}\right)\right]\\&\le \frac{2\varepsilon^2}{k}\mathbb{E}_{\mathbf{q}^{Y^t}}\left[\sum_{y\in\mathcal{Y}}\frac{\left(\sum_{i=1}^k \mathbb{E}\left[Z_i \mid Y^t\right](W^{Y^t}(y|2i-1)-W^{Y^t}(y|2i))\right)^2}{\sum_x W^{Y^t}(y|x)}\right]\\&= \frac{2\varepsilon^2}{k}\mathbb{E}_{\mathbf{q}^{Y^t}}\left[\mathbb{E}\left[Z \mid Y^t\right]^T H(W^{Y^t})\mathbb{E}\left[Z \mid Y^t\right]\right]. \qquad (18)\end{aligned}$$

We can now bound[11]

$$\begin{aligned}&\mathbb{E}\left[Z \mid Y^t\right]^T H(W^{Y^t})\mathbb{E}\left[Z \mid Y^t\right]\\&\qquad\le \left\|H(W^{Y^t})\right\|_{\mathrm{op}}\cdot\left\|\mathbb{E}\left[Z \mid Y^t\right]\right\|_2^2, \qquad (19)\end{aligned}$$

where $\|\cdot\|_{\mathrm{op}}$ denotes the operator norm (or the maximum eigenvalue) of the p.s.d. matrix $H(W^{Y^t})$.

We now take recourse to the information-loss bound in Lemma 11 to relate $\|\mathbb{E}[Z \mid Y^t]\|_2^2$ to average information. By combining Eqs. (18) and (19) and using Lemma 11, we obtain

$$\begin{aligned}&\mathbb{E}_{\mathbf{q}^{Y^t}}\left[\mathrm{D}\left(\mathbf{p}^{Y_{t+1}|Y^t}\|\mathbf{u}^{Y_{t+1}|Y^t}\right)\right]\\&\qquad\le \frac{4(\ln 2)\varepsilon^2}{k}\left\|\mathcal{W}\right\|_{\mathrm{op}}\cdot\sum_{i=1}^k I\left(Z_i\wedge Y^t\right),\end{aligned}$$

proving the lemma. □

Remark 18 (Is the Bound Above Tight?): A key heuristic underlying our learning bound is the thesis that when the information gathered about each coordinate is small, the information revealed in the next iteration cannot be too much. The bound in Eq. (18) provides a quantitative counterpart for this heuristic. The crux of the previous bound is Eq. (19), which relates the Kullback–Leibler divergence to a per-coordinate information quantity $\sum_{i=1}^k \mathbb{E}\left[\mathbb{E}[Z_i \mid Y^t]^2\right]$. As for learning, this enables us to circumvent the difficulty in handling the joint correlation between Z_is, when conditioned on Y^t. In fact, this step can be weak, as we shall see in a later section below. Nonetheless, it allows us to relate the distance between message distribution induced by the mixture distribution and the uniform distribution to the average information quantity of Theorem 13. This connection between learning and testing bounds is interesting in its own right.

Upon combining Lemma 17 with Eq. (16), summing over t, and using the average information bound of Theorem 13, we get

$$\mathrm{D}\left(\mathbf{q}^{Y^n}\|\mathbf{u}^{Y^n}\right)\le\frac{16(\ln 2)\varepsilon^4 n^2}{k^2}\left\|\mathcal{W}\right\|_{\mathrm{op}}\left\|\mathcal{W}\right\|_*,$$

which gives the desired bound $n=\Omega(k/(\sqrt{\|\mathcal{W}\|_{\mathrm{op}}\|\mathcal{W}\|_*}\varepsilon^2))$ for $\mathrm{D}\left(\mathbf{q}^{Y^n}\|\mathbf{u}^{Y^n}\right)$ to be $\Omega(1)$. This proves Theorem 6.

[11]In view of Lemma 11, the right-side of Eq. (19) is large when the mean squared error in estimating Z from Y^t is small. Thus, if the divergence in Eq. (18) is large, we should be able to determine Z from Y^t.

B. A Bound for $\|H(W)\|_{\mathrm{op}}$

Next, we record a general property of the matrix $H(W)$, which is crucial for handling communication constraints but more generally holds for arbitrary information constraints.

Lemma 19 (Operator-Norm Bound): For any channel $W\colon\mathcal{X}\to\mathcal{Y}$, we have $\|H(W)\|_{\mathrm{op}}\le 2$.

Proof: By the Gershgorin circle theorem, the eigenvalue of a matrix is at most the largest sum of absolute entries of a row. Now, for any $i\in[k]$,

$$\begin{aligned}&\|H(W)\|_{\mathrm{op}}\\&\le\sum_{j=1}^k\left|\sum_{y\in\mathcal{Y}}\frac{(W(y|2i-1)-W(y|2i))(W(y|2j-1)-W(y|2j))}{\sum_{x\in[k]}W(y|x)}\right|\\&\le\sum_y\left|(W(y\mid 2i-1)-W(y\mid 2i))\right|\frac{\sum_j|W(y|2j-1)-W(y|2j)|}{\sum_{x\in[k]}W(y|x)}\\&\le\sum_y|W(y\mid 2i-1)-W(y\mid 2i)|\le 2,\end{aligned}$$

where in the last step we used the fact that $\sum_{y\in\mathcal{Y}}W(y\mid x)=1$ for all $x\in[2k]$. □

C. The General Bound Can Be Tightened

We now present a family of channels for which the general lower bound of Theorem 6 and the true sample complexity are a factor $k^{1/4}$ apart. Nonetheless, we can follow the proof of the lower bound instead of directly applying the statement and establish the tight lower bounds. In other words, the general proof methodology we have goes beyond the specific form in Theorem 6.

Let $\mathcal{X}=[2k]$, $\mathcal{Y}:=\mathcal{X}\cup\{\perp\}$, and $\eta\in(0,1)$. The family of *partial erasure* channels $\mathcal{W}_\perp^\eta$ consists of $2k$ channels from $\mathcal{X}$ to $\mathcal{Y}$, indexed by elements of $\mathcal{X}$ such that for $x^*\in\mathcal{X}$,

$$W_{x^*}(y\mid x)=\begin{cases}1, & \text{if } y=x=x^*,\\ \eta, & \text{if } y=x \text{ and } x\ne x^*,\\ 1-\eta, & \text{if } y=\perp \text{ and } x\ne x^*.\end{cases}$$

Namely, the channel W_{x^*} sends the symbol x^* exactly and erases every other symbol $x\ne x^*$ with probability $1-\eta$. Moreover, the channel matrix $H(W_{x^*})$ (see Eq. (6)) is diagonal with the ith diagonal entry equal to $1+\eta+\frac{1-\eta}{2k-1}$ for $x^*\notin\{2i-1,2i\}$, and is equal to 2η otherwise. For $\eta=1/\sqrt{k}$, we can verify that

$$\begin{aligned}2\le\|H(W_x)\|_F&\le 2\sqrt{2},\\2\sqrt{k}\le\|H(W_x)\|_*&\le 2\sqrt{k}+2,\\1\le\|H(W_x)\|_{\mathrm{op}}&\le 2.\end{aligned}\qquad(20)$$

Using these quantities to evaluate the lower bounds in Table I, we get a lower bound of $\Omega(k/\varepsilon^2)$ for the sample complexity of testing under SMP public-coin protocols. We now provide a simple SMP private-coin protocols that achieves this bound.

We set all the channels to be W_1, the channel that erases all symbols except symbol $x=1$. This can be converted into an erasure channel with erasure probability $1-\eta$ by simply converting the Y_ts that are equal to 1 to $\perp$ with probability $1-\eta$. With this modification, the channel output for the users are independent and identically distributed, and $\Pr[Y_t=x\mid Y_t\ne\perp]=\mathbf{p}(x)$, where $\mathbf{p}$ is the underlying

distribution. Therefore, with $O(\frac{\sqrt{k}}{\varepsilon^2}\cdot\frac{1}{\eta})$ users we can obtain $O(\frac{\sqrt{k}}{\varepsilon^2})$ samples and use a centralized uniformity test. Upon combining these bounds we get the following result.

Proposition 20: The sample complexity of noninteractive $(2k,\varepsilon)$-uniformity testing under local constraints $\mathcal{W}_\perp^{1/\sqrt{k}}$ is $\Theta(k/\varepsilon^2)$ for both public-coin and private-coin protocols.

Next, using the norm bounds in Eq. (20) to evaluate the general lower bound of Theorem 6 gives a lower bound of $\Omega(k^{3/4}/\varepsilon^2)$ for sample complexity of uniformity testing under $\mathcal{W}$, using interactive protocols.

Below we will see that this bound is not tight, showing that the general bound of Theorem 6 can be loose for specific families. Nonetheless, we show that the proof of Theorem 6 can be adapted easily to establish the optimal sample complexity $\Theta(k/\varepsilon^2)$ for interactive protocols, matching Proposition (21). This will be achieved by an improved evaluation for $\mathbb{E}[Z \mid Y^t]^T H(W^{Y^t})\mathbb{E}[Z \mid Y^t]$ in Eq. (19).

Proposition 21: Interactive $(2k,\varepsilon)$-uniformity testing under local constraints $\mathcal{W}_\perp^{1/\sqrt{k}}$ requires at least $\Omega(k/\varepsilon^2)$ users.

Proof: We proceed as in the proof of Lemma 17 until Eq. (18) and then replace the bound Eq. (19) with a more precise one. Specifically, we note that different choices of x simply allow us to permute the diagonal entries of the diagonal matrix $H(W_x)$. Therefore, we get

$$\begin{aligned}
&\mathbb{E}\big[Z \mid Y^t\big]^T H(W^{Y^t})\mathbb{E}\big[Z \mid Y^t\big] \\
&\le \left(1+\eta+\frac{1-\eta}{2k-1}\right)\max_{1\le i\le k}\mathbb{E}\big[Z_i \mid Y^t\big]^2 \\
&\quad+\eta\left(\big\|\mathbb{E}\big[Z \mid Y^t\big]\big\|_2^2-\max_{1\le i\le k}\mathbb{E}\big[Z_i \mid Y^t\big]^2\right) \\
&\le 2\big\|\mathbb{E}\big[Z \mid Y^t\big]\big\|_\infty^2+\frac{1}{\sqrt{k}}\big\|\mathbb{E}\big[Z \mid Y^t\big]\big\|_2^2.
\end{aligned}$$

Combining with Lemma 11, we get

$$\begin{aligned}
&\mathbb{E}\Big[\mathbb{E}\big[Z \mid Y^t\big]^T H(W^{Y^t})\mathbb{E}\big[Z \mid Y^t\big]\Big] \\
&\quad\le 4(\ln 2)\left(\max_{1\le i\le k} I\left(Z_i\wedge Y^t\right)+\frac{1}{\sqrt{k}}\sum_{i=1}^{k} I\left(Z_i\wedge Y^t\right)\right).
\end{aligned}$$

Next, we take recourse to Remark 14 to get a bound for information $I\left(Z_i\wedge Y^t\right)$ about each coordinate. We have

$$I\left(Z_i\wedge Y^t\right)\le\frac{8\varepsilon^2}{k}\sum_{j=1}^{t-1}\mathbb{E}\Big[H(W^{Y^j})_{i,i}\Big],$$

which when combined with the previous bound yields

$$\begin{aligned}
&\mathbb{E}\Big[\mathbb{E}\big[Z \mid Y^t\big]^T H(W^{Y^t})\mathbb{E}\big[Z \mid Y^t\big]\Big] \\
&\quad\le\frac{32(\ln 2)\varepsilon^2}{k}\sum_{j=1}^{t-1}\left(\max_{1\le i\le k}\mathbb{E}\Big[H(W^{Y^j})_{i,i}\Big]\right. \\
&\qquad\qquad\left.+\frac{1}{\sqrt{k}}\sum_{i=1}^{k}\mathbb{E}\Big[H(W^{Y^j})_{i,i}\Big]\right) \\
&\quad\le\frac{320(\ln 2)\varepsilon^2(t-1)}{k}.
\end{aligned}$$

It follows from (18) that

$$\mathrm{D}\left(\mathbf{q}^{Y^n}\|\mathbf{u}^{Y^n}\right)\le 320(\ln 2)\cdot\frac{\varepsilon^4 n^2}{k^2},$$

which completes the proof. □

We close by noting that the proof above provides yet another example of an application where our lower bound technique yields a tight bound; we believe there can be many more. We note that even in this example we related the distance to the per-coordinate information $I(Z_i\wedge Y^t)$. It will be interesting to seek examples where our technique yields a tight bound without using an upper bound for Eq. (18) in terms of per-coordinate information quantities.

D. A Separation Between Non-Interactive and Interactive Protocols

We will now show that there exists a "natural" family of local constraints for which the sample complexity of interactive protocols is much smaller than that of noninteractive protocols for $(2k,\varepsilon)$-uniformity testing. To the best of our knowledge, this is the first example of a separation between interactive and noninteractive protocols for a basic hypothesis testing problem.

1) The Search for a Suitable Family of Channels: To describe how we identify the family $\mathcal{W}$ that yields the desired separation, we first revisit the proof of Lemma 17. As noted before, the only possibly loose step in the argument is (19). As we saw in Section V-C, this bound can be improved by carefully examining the spectrum of $H(W)$ for different Ws. Our goal in this section is to construct an example where the bound in (19) is tight, but $\|\mathcal{W}\|_F$ is maximally separated from $\sqrt{\|W\|_*\|W\|_{\mathrm{op}}}$. Towards this, a key observation we have is that for (19) to be tight, we should have a channel family that such that for each $\mathbb{E}[Z \mid Y^t]$, we can find a channel W such that the maximum eigenvalue of W is roughly $\|W\|_{\mathrm{op}}$ and it corresponds to an eigenvector that is aligned with $\mathbb{E}[Z \mid Y^t]$.

Specifically, we seek a set of channels $\mathcal{W}$ for which (a) there is a large gap between $\|\mathcal{W}\|_F$ and $\sqrt{\|\mathcal{W}\|_{\mathrm{op}}\|\mathcal{W}\|_*}$; (b) there is a noninteractive protocol with sample complexity $O(k/\varepsilon^2\|\mathcal{W}\|_F)$; and (c) there is an interactive protocol with sample complexity $O(k/\varepsilon^2\sqrt{\|\mathcal{W}\|_{\mathrm{op}}\|\mathcal{W}\|_*})$. In view of the heuristic observation above, we seek $\mathcal{W}$ such that we can assign the maximum eigenvalue of $H(W)$ in any direction of our choice, by appropriately choosing $W\in\mathcal{W}$. In the previous section, we designed a set of channels that satisfy (a) and (b). However, (c) did not hold since (19) is not tight as we could only assign the maximum eigenvalue of $H(W)$ to one of the standard basis vectors, and to no other direction.

We meet the objectives above with channels that release membership queries for particular sets of our choice. We will also have a leakage component to introduce an eigenspace with a small eigenvalue to ensure a large gap in different norms of interest to us. We now formalize this family below.

For $\eta\in[0,1)$, $u\in[0,1]^{2k}$, and $\mathcal{Y}:=[2k]\cup\{\mathbf{1}^\star,\mathbf{0}^\star\}$, the *leaky-query* channel W_u^η for an input $x\in[2k]$ outputs x with probability η; otherwise it outputs $\mathbf{1}^\star$ and $\mathbf{0}^\star$ with probability u_x and $1-u_x$ respectively. For our scheme, we will use u_x

that has all the entries for coordinates in a set S equal to 1 and outside it equal to 0, corresponding to a membership query for S. Let $\mathcal{W}^{\eta}_{\in} = \{W^{\eta}_u : u \in [0,1]^{2k}\}$

$$W_u(y \mid x) = \begin{cases} \eta, & \text{if } y = x, \\ (1-\eta)u_x, & \text{if } y = \mathbf{1}^\star, \\ (1-\eta)(1-u_x), & \text{if } y = \mathbf{0}^\star. \end{cases}$$

Throughout this section, we will consider $\eta = 1/\sqrt{k}$. We begin by evaluating the required norms for this family.

Lemma 22:

$$\begin{aligned} 2 \le \left\|\mathcal{W}^{1/\sqrt{k}}_{\in}\right\|_F &\le 2\sqrt{2}, \\ 2\sqrt{k} \le \left\|\mathcal{W}^{1/\sqrt{k}}_{\in}\right\|_* &\le 2\sqrt{k}+2, \\ \left\|\mathcal{W}^{1/\sqrt{k}}_{\in}\right\|_{\mathrm{op}} &= 2. \end{aligned} \tag{21}$$

Proof: By the definition of W_u, we obtain for $i_1, i_2 \in [k]$ that

$$H(W_u)_{i_1,i_2} = \sum_{y \in [2k] \cup \{\mathbf{1}^\star, \mathbf{0}^\star\}} \frac{(W_u(y|2i_1-1) - W_u(y|2i_1))(W_u(y|2i_2-1) - W_u(y|2i_2))}{\sum_{x \in [2k]} W_u(y|x)}.$$

Note that for every $u \in [0,1]^{2k}$ and $y \in [2k]$, $\sum_{x \in [2k]} W_u(y \mid x) = W_u(y \mid y) = \eta$ and

$$(W_u(y \mid 2i_1-1) - W_u(y|2i_1))(W_u(y|2i_2-1) - W_u(y|2i_2)) = \begin{cases} \eta^2, & \text{if } i_1 = i_2 = \lceil y/2 \rceil, \\ 0, & \text{otherwise.} \end{cases}$$

Further, for $y \in \{\mathbf{1}^\star, \mathbf{0}^\star\}$, we have

$$(W_u(y \mid 2i_1-1) - W_u(y|2i_1))(W_u(y|2i_2-1) - W_u(y|2i_2)) = (1-\eta)^2 (u_{2i_1-1} - u_{2i_1})(u_{2i_2-1} - u_{2i_2}),$$

and

$$\sum_{x \in [2k]} W_u(\mathbf{1}^\star \mid x) = (1-\eta) \sum_{x \in [2k]} u_x,$$
$$\sum_{x \in [2k]} W_u(\mathbf{0}^\star \mid x) = (1-\eta) \sum_{x \in [2k]} (1-u_x).$$

Upon combining these bounds, we get

$$H(W_u) = 2\eta I_k + (1-\eta)\delta(u)\delta(u)^T,$$

where for all $i \in [k]$,

$$\delta(u)_i = (u_{2i-1} - u_{2i}) \sqrt{\frac{2k}{(\|u\|_1)(2k - \|u\|_1)}}.$$

From this, it can be verified that $H(W_u)$ has eigenvalues $2\eta + (1-\eta)\|\delta(u)\|_2^2$ with multiplicity one and 2η with multiplicity $k-1$. Also, we have $\|\delta(u)\|_2^2 \le 2$, and moreover that equality holds when $|u_{2i-1} - u_{2i}| = 1$ for all $i \in [k]$ and $\sum_{i \in [k]} u(2i) = k/2$. Setting the value of η to be $1/\sqrt{k}$ establishes the claimed bounds. □

The following two results establish our claim of a separation between noninteractive schemes and interactive scheme for $(2k, \varepsilon)$-uniformity testing under local constraints $\mathcal{W}^{1/\sqrt{k}}_{\in}$.

Proposition 23: Noninteractive $(2k, \varepsilon)$-uniformity testing under $\mathcal{W}^{1/\sqrt{k}}_{\in}$ has sample complexity $\Theta(k/\varepsilon^2)$, even when the unknown distribution $\mathbf{p}$ has bounded norm $\|\mathbf{p}\|_\infty \le 10/k$.

Proof: The lower bound can be shown by plugging $\|\mathcal{W}^{1/\sqrt{k}}_{\in}\|_F \le 2\sqrt{2}$ in the lower bound for noninteractive schemes obtained in [7] (see Table I). Crucially, this lower bound is established by considering the family of distributions given in Eq. (5), and thus still applies under the promise that $\|\mathbf{p}\|_\infty \le \frac{10}{k}$. For the upper bound, note that with probability $1/\sqrt{k}$ we observe a sample from $[2k]$ from the underlying distribution. Ignoring the binary responses and using the same argument as that in the previous section, we get a (matching) upper bound for the number of samples needed by this private-coin SMP protocol. □

Our next result provides the last piece to establish our separation, by showing that interactive protocols can do strictly better than the noninteractive ones.[12]

Proposition 24: For $\varepsilon \ge 8/k^{1/8}$, interactive $(2k, \varepsilon)$-uniformity testing under $\mathcal{W}^{1/\sqrt{k}}_{\in}$ under the promise that the unknown distribution $\mathbf{p}$ satisfies $\|\mathbf{p}\|_\infty \le 10/k$ has sample complexity $\Theta(k^{3/4}/\varepsilon^2)$.

Proof: The lower bound can be obtained by plugging the bounds $\|\mathcal{W}^{1/\sqrt{k}}_{\in}\|_* \le 2\sqrt{k}+2$ and $\|\mathcal{W}^{1/\sqrt{k}}_{\in}\|_{\mathrm{op}} = 2$ into Theorem 6 (noting that the lower bound instances (Eq. (5)) satisfy the ℓ_∞ promise). For the upper bound, we first present a high-level overview of our scheme, and then provide the details.

2) Sketch of The Scheme: Observe that when we sample $X \sim \mathbf{p}$ for a distribution $\mathbf{p}$ over $[2k]$, we have $\mathbb{E}[\mathbf{p}(X)] = \sum_x \mathbf{p}(x)^2 = \|\mathbf{p}\|_2^2$. If $\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{u}) \ge \varepsilon$, then by the Cauchy–Schwarz inequality $\|\mathbf{p}\|_2^2 \ge (1+4\varepsilon^2)/2k$, whereas $\|\mathbf{u}\|_2^2 = 1/k$. Therefore, when we sample from a distribution that is far from uniform, the expected probability of the observed sample is also larger than that under the uniform distribution. Our protocol exploits this and proceeds in two stages. In the first stage, we select any channel in $\mathcal{W}^{1/\sqrt{k}}_{\in}$ and use it for a fraction of the users. Let S be the set of outputs from these channels that are in $[2k]$. Now, $\mathbf{u}(S) = |S|/2k$, but using the motivation above we can hope that, for a $\mathbf{p}$ that is far from $\mathbf{u}$, $\mathbf{p}(S)$ will be noticeably larger. In the next stage, the remaining players choose $u_x = \mathbb{1}_{\{x \in S\}}$. Now, $\mathbf{u}(\mathbf{1}^\star) = (1-\eta)\mathbf{u}(S)$, and $\mathbf{p}(\mathbf{1}^\star) = (1-\eta)\mathbf{p}(S)$, and we will perform a binary hypothesis test to separate these two cases.

3) Detailed Argument: The rest of the argument makes the intuition above formal. We assume that there are $n = Ck^{3/4}/\varepsilon^2$ users, for some constant $C > 0$ which can be taken to be $C = 625$ and $k \ge \lceil 9C^2/2^{16} \rceil = 54$. Our protocol proceeds as follows:

1) For the first $n/2$ users, choose the channel $W_{\mathbf{0}}$, corresponding to a simple erasure channel with erasure probability $1-\eta$. Gather a set $S \subseteq [2k]$ of "leaked" samples in this stage.

[12]For simplicitly, we only provide a protocol for the case of $\varepsilon = \Omega(1/k^{1/8})$, which is enough for our purposes. We believe that handling smaller values of the distance parameter is possible, but would require a more involved protocol and analysis.

2) For the last $n/2$ users, choose the channel W_u for u corresponding to the indicator vector of S, in order to estimate $\mathbf{p}(S)$ to an additive accuracy $\frac{\varepsilon^2}{4}\mathbb{E}_{\mathbf{u}}[\mathbf{u}(S)]$, via the binary responses.

Step 1. Let N be the number of "leaked" samples in the first stage, namely N symbols are received without erasure in the first stage. It can be easily checked that $\mathbb{E}[N] = n/(2\sqrt{k})$ and $\mathrm{Var}(N) < n/(2\sqrt{k})$. Since our assumptions on C and k imply that $n \geq 800\sqrt{k}$, by Chebyshev's inequality we have $n/(4\sqrt{k}) \leq N \leq 3n/(4\sqrt{k})$ with probability at least $99/100$; below we proceed assuming this event holds and will account for the probability of its failure at the end.

Let $S \subseteq [2k]$ be the set of symbols of "leaked" samples in this step; note that $|S| \leq N$ (since some values may be repeated). By linearity of expectation, when the samples are generated from $\mathbf{p}$, we have

$$\mathbb{E}[\mathbf{p}(S)] = \sum_{i=1}^{2k} \mathbf{p}(i)\left(1 - (1-\mathbf{p}(i))^N\right),$$

and $\mathbf{u}(S) = 1 - (1 - 1/(2k))^N$. When $\mathbf{p}$ is ε-far from $\mathbf{u}$, we have $\|\mathbf{p}\|_2^2 \geq \frac{1+4\varepsilon^2}{2k}$, and

$$\begin{aligned}
&\mathbb{E}_{\mathbf{p}}[\mathbf{p}(S)] - \mathbb{E}_{\mathbf{u}}[\mathbf{u}(S)] \\
&\quad = \left(1 - \frac{1}{2k}\right)^N \sum_{i=1}^{2k} \mathbf{p}(i)\left(1 - \left(\frac{2k}{2k-1}(1-\mathbf{p}(i))\right)^N\right) \\
&\quad \geq \frac{N\varepsilon^2}{k},
\end{aligned}$$

where the last inequality follows by using almost the same analysis as that in the proof of [39, Lemma 1]. Further, we have

$$\mathbb{E}_{\mathbf{u}}[\mathbf{u}(S)] = 1 - \left(1 - \frac{1}{2k}\right)^N \geq \frac{N}{4k}.$$

since $N \leq k$ (which follows from our bound $N \leq 3n/(4\sqrt{k})$, along with $\varepsilon \geq 8/k^{1/8}$ and $k \geq 9C^2/2^{16}$), and therefore,

$$\mathbb{E}_{\mathbf{p}}[\mathbf{p}(S)] \geq (1 + 3\varepsilon^2/2)\mathbb{E}_{\mathbf{u}}[\mathbf{u}(S)]$$

whenever $\mathbf{p}$ is ε-far from uniform.

Turning to the variance, we can prove the following bound:

Claim 25: For any $\mathbf{p}$, we have

$$\mathrm{Var}_{\mathbf{p}}[\mathbf{p}(S)] \leq \|\mathbf{p}\|_\infty \mathbb{E}_{\mathbf{p}}[\mathbf{p}(S)].$$

Proof: Denote by $N_1, \ldots, N_k$ the sample counts, *i.e.*, N_i is the number of times element i is seen among the N samples. While N_i is distributed as a Binomial with parameters N and $\mathbf{p}(i)$, the $N_1, \ldots, N_k$ are not independent; however, they are negatively associated (see, *e.g.*, [23, Section 2.2]), which we will use below. We start with bounding the expected square:

$$\begin{aligned}
\mathbb{E}\left[\mathbf{p}(S)^2\right] &= \sum_{i=1}^k \sum_{j=1}^k \mathbf{p}(i)\mathbf{p}(j)\mathbb{E}\left[\mathbb{1}_{\{N_i \geq 1\}}\mathbb{1}_{\{N_j \geq 1\}}\right] \\
&= \sum_{i=1}^k \mathbf{p}(i)^2 \mathbb{E}\left[\mathbb{1}_{\{N_i \geq 1\}}\right] \\
&\quad + 2\sum_{i<j} \mathbf{p}(i)\mathbf{p}(j)\mathbb{E}\left[\mathbb{1}_{\{N_i \geq 1\}}\mathbb{1}_{\{N_j \geq 1\}}\right] \\
&\leq \sum_{i=1}^k \mathbf{p}(i)^2 \mathbb{E}\left[\mathbb{1}_{\{N_i \geq 1\}}\right] \\
&\quad + 2\sum_{i<j} \mathbf{p}(i)\mathbf{p}(j)\mathbb{E}\left[\mathbb{1}_{\{N_i \geq 1\}}\right]\mathbb{E}\left[\mathbb{1}_{\{N_j \geq 1\}}\right] \\
&= \sum_{i=1}^k \mathbf{p}(i)^2 \Pr\left[N_i \geq 1\right] + \left(\sum_{i=1}^k \mathbf{p}(i)\Pr\left[N_i \geq 1\right]\right)^2 \\
&\quad - \sum_{i=1}^k \mathbf{p}(i)^2 \Pr\left[N_i \geq 1\right]^2 \\
&= \sum_{i=1}^k \mathbf{p}(i)^2 \Pr\left[N_i \geq 1\right]\Pr\left[N_i = 0\right] + \mathbb{E}[\mathbf{p}(S)]^2,
\end{aligned}$$

where the inequality follows from negative associativity, and we got the third equality by completing the sum $2\sum_{i<j} x_{i,j} = \sum_{i,j} x_{i,j} - \sum_i x_{i,i}$. Rewriting, we have

$$\mathrm{Var}[\mathbf{p}(S)] \leq \sum_{i=1}^k \mathbf{p}(i)^2 \Pr\left[N_i \geq 1\right]\Pr\left[N_i = 0\right].$$

By upper bounding the last factor by 1, we then get

$$\mathrm{Var}[\mathbf{p}(S)] \leq \|\mathbf{p}\|_\infty \mathbb{E}\left[\mathbf{p}(S)\right],$$

concluding the proof. □

When $\mathbf{p} = \mathbf{u}$, this gives $\mathrm{Var}_{\mathbf{u}}[\mathbf{u}(S)] \leq \frac{1}{2k}\mathbb{E}_{\mathbf{u}}[\mathbf{u}(S)]$. By Chebyshev's inequality, using the chain of inequalities

$$\frac{2}{k\varepsilon^4 \mathbb{E}_{\mathbf{u}}[\mathbf{u}(S)]} \leq \frac{8}{\varepsilon^4 N} \leq \frac{32\sqrt{k}}{\varepsilon^4 n} \leq \frac{k^{3/4}}{2\varepsilon^2 n}$$

(where the second is due to our lower bound on N, and the third follows from our assumption $\varepsilon \geq 8/k^{1/8}$) we get that

$$\Pr_{X^n \sim \mathbf{u}^n}\left[\mathbf{u}(S) < (1 + \frac{\varepsilon^2}{2})\mathbb{E}_{\mathbf{u}}[\mathbf{u}(S)]\right] \geq 9/10 \tag{22}$$

as long as $C \geq 5$.

Now, consider the case where $\mathbf{p}$ is ε-far from uniform. By Chebyshev's inequality, using Claim 25 along with the promise that $\|\mathbf{p}\|_\infty \leq 10/k$, we get

$$\begin{aligned}
&\Pr_{X^n \sim \mathbf{p}^n}\left[\mathbf{p}(S) < (1+\varepsilon^2)\mathbb{E}_{\mathbf{u}}[\mathbf{u}(S)]\right] \\
&\quad \leq \Pr_{X^n \sim \mathbf{p}^n}\left[\mathbf{p}(S) < \frac{1+\varepsilon^2}{1 + \frac{3}{2}\varepsilon^2}\mathbb{E}_{\mathbf{p}}[\mathbf{p}(S)]\right] \\
&\quad \leq \frac{(2+3\varepsilon^2)^2}{\varepsilon^4}\frac{10}{k\mathbb{E}_{\mathbf{p}}[\mathbf{p}(S)]} \leq \frac{1000}{\varepsilon^4 N} \\
&\quad \leq \frac{125k^{3/4}}{2\varepsilon^2 n},
\end{aligned}$$

where the last inequality is derived as in the uniform case. This implies that, if $\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{u}) \geq \varepsilon$,

$$\Pr_{X^n \sim \mathbf{p}^n}\left[\mathbf{p}(S) > (1+\varepsilon^2)\mathbb{E}_{\mathbf{u}}[\mathbf{u}(S)]\right] \geq 9/10, \tag{23}$$

as long as $C \geq 625$.

Step 2. In the second stage, the $n/2$ users all choose the channel W_u, where $u \in \{0,1\}^{2k}$ is the indicator vector of S. We assume now that conditions Eq. (22) and Eq. (23), respectively, hold under the uniform and nonuniform distribution hypothesis. The goal of this stage is to distinguish between

the two cases $\mathbf{p}(S) < (1+\frac{1}{2}\varepsilon^2)\mathbb{E}_{\mathbf{u}}[\mathbf{u}(S)]$ and $\mathbf{p}(S) > (1+\varepsilon^2)\mathbb{E}_{\mathbf{u}}[\mathbf{u}(S)]$, which can be done by estimating $\mathbf{p}(S)$ to an additive $\frac{\varepsilon^2}{4}\mathbb{E}_{\mathbf{u}}[\mathbf{u}(S)]$ with probability at least $99/100$ (from $n/2$ users). Note that this is equivalent to estimating the mean of a Bernoulli random variable $p \geq \mathbb{E}_{\mathbf{u}}[\mathbf{u}(S)]$ to an additive $\varepsilon^2 p/8$ from $n/2$ observations. Using Chebyshev's inequality, we can check $n \geq 320k^{3/4}/\varepsilon^2$ suffices.

Overall. Accounting for the 3 good events above that hold with probability $99/100$, $9/10$, and $99/100$ respectively, this protocol is correct by a union bound with probability at least $22/25$, and involves $n = O(k^{3/4}/\varepsilon^2)$ users, as desired. By explicit computation of the Binomial distribution probabilities, repeating the protocol 7 times on independent subsets of samples (*i.e.*, groups of users) and taking the majority output can then boost the success probability from 22/25 to at least 99/100, only changing the total number of samples by this constant factor 7. □

REFERENCES

[1] J. Acharya, C. L. Canonne, C. Freitag, Z. Sun, and H. Tyagi, "Inference under information constraints III: Local privacy constraints," *IEEE J. Sel. Areas Inf. Theory*, vol. 2, no. 1, pp. 253–267, Mar. 2021.

[2] J. Acharya, C. L. Canonne, Y. Han, Z. Sun, and H. Tyagi, "Domain compression and its application to randomness-optimal distributed goodness-of-fit," in *Proc. 33rd Conf. Learn. Theory*, vol. 125, J. Abernethy and S. Agarwal, Eds. Graz, Austria: PMLR, Jul. 2020, pp. 3–40. [Online]. Available: http://proceedings.mlr.press/v125/acharya20a.html

[3] J. Acharya, C. L. Canonne, Y. Liu, Z. Sun, and H. Tyagi, "Interactive inference under information constraints," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Melbourne, VIC, Australia, Jul. 2021, pp. 326–331.

[4] J. Acharya, C. L. Canonne, and H. Tyagi, "Distributed simulation and distributed inference," 2018, *arXiv:1804.06952*.

[5] J. Acharya, C. Canonne, and H. Tyagi, "Communication-constrained inference and the role of shared randomness," in *Proc. 36th Int. Conf. Mach. Learn.*, vol. 97, K. Chaudhuri and R. Salakhutdinov, Eds. Long Beach, CA, USA: PMLR, Jun. 2019, pp. 30–39.

[6] J. Acharya, C. L. Canonne, and H. Tyagi, "Inference under information constraints: Lower bounds from chi-square contraction," in *Proc. 32nd Conf. Learn. Theory*, vol. 99, A. Beygelzimer and D. Hsu, Eds. Phoenix, AZ, USA: PMLR, Jun. 2019, pp. 3–17.

[7] J. Acharya, C. L. Canonne, and H. Tyagi, "Inference under information constraints I: Lower bounds from chi-square contraction," *IEEE Trans. Inf. Theory*, vol. 66, no. 12, pp. 7835–7855, Dec. 2020.

[8] 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.

[9] J. Acharya and Z. Sun, "Communication complexity in locally private distribution estimation and heavy hitters," in *Proc. 36th Int. Conf. Mach. Learn.*, vol. 97, K. Chaudhuri and R. Salakhutdinov, Eds. Long Beach, CA, USA: PMLR, Jun. 2019, pp. 51–60.

[10] J. Acharya, Z. Sun, and H. Zhang, "Hadamard response: Estimating distributions privately, efficiently, and with little communication," in *Proc. 22nd Int. Conf. Artif. Intell. Statist. (AISTATS)*, 2018, pp. 1120–1129.

[11] K. Amin, M. Joseph, and J. Mao, *Personal Communication*, Jul. 2020.

[12] K. Amin, M. Joseph, and J. Mao, "Pan-private uniformity testing," in *Proc. 33rd Conf. Learn. Theory*, vol. 125, J. Abernethy and S. Agarwal, Eds. Graz, Austria: PMLR, Jul. 2020, pp. 183–218. [Online]. Available: http://proceedings.mlr.press/v125/amin20a.html

[13] L. P. Barnes, Y. Han, and A. Ozgur, "Fisher information for distributed estimation under a blackboard communication protocol," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2019, pp. 2704–2708.

[14] L. P. Barnes, W.-N. Chen, and A. Ozgur, "Fisher information under local differential privacy," *IEEE J. Sel. Areas Inf. Theory*, vol. 1, no. 3, pp. 645–659, Nov. 2020.

[15] R. Bassily, "Linear queries estimation with local differential privacy," in *Proc. Int. Conf. Artif. Intell. Statist. (AISTATS)*, vol. 89. Naha, Japan: PMLR, 2019, pp. 721–729.

[16] T. Berrett and C. Butucea, "Locally private non-asymptotic testing of discrete distributions is faster using interactive mechanisms," in *Proc. Adv. Neural Inf. Process. Syst. 33, Annu. Conf. Neural Inf. Process. Syst. (NeurIPS)*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds. Red Hook, NY, USA: Curran Associates, Dec. 2020, pp. 3164–3173. [Online]. Available: https://proceedings.neurips.cc/paper/2020/hash/20b02dc95171540bc52912baf3aa709d-Abstract.html

[17] M. Braverman, A. Garg, T. Ma, H. L. Nguyen, and D. P. Woodruff, "Communication lower bounds for statistical estimation problems via a distributed data processing inequality," in *Proc. 48th Annu. ACM Symp. Theory Comput.*, Jun. 2016, pp. 1011–1020.

[18] S. Bubeck, S. Chen, and J. Li, "Entanglement is necessary for optimal quantum property testing," in *Proc. IEEE 61st Annu. Symp. Found. Comput. Sci.*, Los Alamitos, CA, USA, Nov. 2020, pp. 692–703.

[19] Y. Dagan and V. Feldman, "Interaction is necessary for distributed learning with privacy or communication constraints," in *Proc. 52nd Annu. ACM SIGACT Symp. Theory Comput. (STOC)*, New York, NY, USA, 2020, pp. 450–462.

[20] A. Daniely and V. Feldman, "Locally private learning without interaction requires separation," in *Proc. Adv. Neural Inf. Process. Syst.*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds. Brooklyn, NY, USA: Curran Associates, 2019, pp. 15001–15012.

[21] I. Diakonikolas, T. Gouleakis, D. M. Kane, and S. Rao, "Communication and memory efficient testing of discrete distributions," in *Proc. 32nd Conf. Learn. Theory*, vol. 99, A. Beygelzimer and D. Hsu, Eds. Phoenix, AZ, USA: PMLR, Jun. 2019, pp. 1070–1106.

[22] I. Diakonikolas, E. Grigorescu, J. Li, A. Natarajan, K. Onak, and L. Schmidt, "Communication-efficient distributed learning of discrete distributions," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6394–6404.

[23] D. Dubhashi and D. Ranjan, "Balls and bins: A study in negative dependence," *Random Struct. Algorithms*, vol. 13, no. 2, pp. 99–124, 1998.

[24] J. Duchi and R. Rogers, "Lower bounds for locally private estimation via communication complexity," in *Proc. 32nd Conf. Learn. Theory*, vol. 99, A. Beygelzimer and D. Hsu, Eds. Phoenix, AZ, USA: PMLR, Jun. 2019, pp. 1161–1191.

[25] J. C. Duchi, M. I. Jordan, and M. J. Wainwright, "Local privacy, data processing inequalities, and statistical minimax rates," 2014, *arXiv:1302.3203*.

[26] J. C. Duchi, M. I. Jordan, and M. J. Wainwright, "Local privacy and statistical minimax rates," in *Proc. IEEE 54th Annu. Symp. Found. Comput. Sci.*, Oct. 2013, pp. 429–438.

[27] J. C. Duchi, M. I. Jordan, and M. J. Wainwright, "Minimax optimal procedures for locally private estimation," *J. Amer. Statist. Assoc.*, vol. 113, no. 521, pp. 182–201, 2018.

[28] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Theory of Cryptography* (Lecture Notes in Computer Science), vol. 3876. Berlin, Germany: Springer, 2006, pp. 265–284.

[29] Ú. Erlingsson, V. Pihur, and A. Korolova, "RAPPOR: Randomized aggregatable privacy-preserving ordinal response," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, Nov. 2014, pp. 1054–1067.

[30] A. V. Evfimievski, J. Gehrke, and R. Srikant, "Limiting privacy breaches in privacy preserving data mining," in *Proc. 22nd ACM SIGMOD-SIGACT-SIGART Symp. Princ. Database Syst.*, 2003, pp. 211–222.

[31] O. Fischer, U. Meir, and R. Oshman, "Distributed uniformity testing," in *Proc. ACM Symp. Princ. Distrib. Comput.*, 2018, pp. 455–464.

[32] Y. Han, P. Mukherjee, A. Ozgur, and T. Weissman, "Distributed statistical estimation of high-dimensional and nonparametric distributions," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2018, pp. 506–510.

[33] Y. Han, P. Mukherjee, A. Özgür, and T. Weissman, "Distributed statistical estimation of high-dimensional and nonparametric distributions with communication constraints, February 2018," *Talk Given at ITA*, vol. 1, no. 1, p. 5, 2018. [Online]. Available: http://ita.ucsd.edu/workshop/18/files/abstract/abstract_2352.txt

[34] Y. Han, A. Ozgur, and T. Weissman, "Geometric lower bounds for distributed parameter estimation under communication constraints," 2018, *arXiv:1802.08417*.

[35] M. Joseph, J. Mao, S. Neel, and A. Roth, "The role of interactivity in local differential privacy," in *Proc. IEEE 60th Annu. Symp. Found. Comput. Sci.*, Los Alamitos, CA, USA, Nov. 2019, pp. 94–105.

[36] M. Joseph, J. Mao, and A. Roth, "Exponential separations in local differential privacy," in *Proc. ACM-SIAM Symp. Discrete Algorithms*. Philadelphia, PA, USA: SIAM, 2020, pp. 515–527.

[37] P. Kairouz, K. Bonawitz, and D. Ramage, "Discrete distribution estimation under local privacy," in *Proc. 33rd Int. Conf. Mach. Learn.*, vol. 48, Jun. 2016, pp. 2436–2444.
[38] S. P. Kasiviswanathan, H. K. Lee, K. Nissim, S. Raskhodnikova, and A. Smith, "What can we learn privately?" *SIAM J. Comput.*, vol. 40, no. 3, pp. 793–826, Jun. 2011.
[39] L. Paninski, "A coincidence-based test for uniformity given very sparsely sampled discrete data," *IEEE Trans. Inf. Theory*, vol. 54, no. 10, pp. 4750–4755, Oct. 2008.
[40] O. Shamir, "Fundamental limits of online and distributed algorithms for statistical learning and estimation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 163–171.
[41] A. Smith, A. Thakurta, and J. Upadhyay, "Is interaction necessary for distributed private learning?" in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2017, pp. 58–77.
[42] J. Ullman, "Tight lower bounds for locally differentially private selection," 2018, *arXiv:1802.02638*.
[43] S. Wang *et al.*, "Mutual information optimally local private discrete distribution estimation," 2016, *arXiv:1607.08025*.
[44] M. Ye and A. Barg, "Optimal schemes for discrete distribution estimation under locally differential privacy," *IEEE Trans. Inf. Theory*, vol. 64, no. 8, pp. 5662–5676, Aug. 2018.

Jayadev Acharya (Member, IEEE) received the Bachelor of Technology degree in electronics and electrical communication engineering from the Indian Institute of Technology, Kharagpur, in 2007, and the M.S. and Ph.D. degrees in electrical and computer engineering from the University of California at San Diego in 2009 and 2014, respectively. He was a Post-Doctoral Associate in electrical engineering and computer science at MIT from 2014 to 2016. He is currently an Assistant Professor with the School of Electrical and Computer Engineering, Cornell University.

Clément L. Canonne is currently a Lecturer with the School of Computer Science, The University of Sydney, Australia. Prior to this, he was a Motwani Post-Doctoral Fellow at Stanford University and a Goldstine Post-Doctoral Fellow at IBM Research, after graduating from Columbia University in 2017, where he was advised by Rocco Servedio. His research focuses on the fields of property testing and sublinear algorithms, and more broadly on computational aspects of learning and statistical inference.

Yuhan Liu (Student Member, IEEE) received the Bachelor of Engineering degree in automation from Tsinghua University. He is currently pursuing the Ph.D. degree in electrical and computer engineering with Cornell University. His research interests include distributed learning, statistical estimation with information constraints, differential privacy, and federated learning.

Ziteng Sun (Student Member, IEEE) received the B.S. degree from Tsinghua University. He is currently pursuing the Ph.D. degree in electrical and computer engineering with Cornell University. His research interests include studying the tradeoffs between different resources in modern data science, including samples, privacy, communication, memory, and computation.

Himanshu Tyagi (Senior Member, IEEE) received the B.Tech. degree in electrical engineering and the M.Tech. degree in communication and information technology from the Indian Institute of Technology, New Delhi, India, in 2007, and the Ph.D. degree from the University of Maryland, College Park, in 2013. From 2013 to 2014, he was a Post-Doctoral Researcher at the Information Theory and Applications (ITA) Center, University of California at San Diego. Since January 2015, he has been a Faculty Member at the Department of Electrical Communication Engineering, Indian Institute of Science, Bengaluru. His research interests broadly lie in information theory and its application in cryptography, statistics and computer science. Also, he is interested in communication and automation for city-scale systems.