# Interactive Inference under Information Constraints

Jayadev Acharya
Cornell University
acharya@cornell.edu

Clément L. Canonne
University of Sydney
clement.canonne@sydney.edu.au

Yuhan Liu
Cornell University
yl2976@cornell.edu

Ziteng Sun
Cornell University
zs335@cornell.edu

Himanshu Tyagi
Indian Institute of Science
htyagi@iisc.ac.in

2021 IEEE International Symposium on Information Theory (ISIT) | 978-1-5386-8209-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/ISIT45174.2021.9518069

***Abstract*—We consider the problem of distributed estimation and testing of discrete distributions under *local* information constraints that include communication and privacy as special cases. Our main result is a unified method that establishes tight bounds for interactive protocols under both the constraints and both the problems. Our main technical contribution is an average information bound which connects learning and testing and handles correlations due to interactivity. While we establish that for learning and testing under both the constraints above, interactivity does not help, we also illustrate a natural family of "leaky query" local constraints under which interactive protocols strictly outperform the noninteractive ones for identity testing.**

## I. Introduction

Classical statistics focuses on algorithms that are data-efficient. However, modern applications entailing distributed statistics have added a new dimension to this objective with the introduction of local information constraints. Canonical examples of information constraints include communication constraints and local privacy constraints. In this paper, we study such modern distributed statistics problems where only limited information about each sample is available to the algorithm.

In particular, we consider the fundamental tasks of estimation and goodness-of-fit testing of discrete distributions. There is an unknown distribution $\mathbf{p}$ over $[2k] = \{1, 2, \ldots, 2k\}$,[1] and a set $\mathcal{W}$ of *allowed* channels with input domain $[2k]$, known a priori. The distributed setting comprises $n$ users, where the $t$th user observes an independent sample $X_t$ from $\mathbf{p}$. User $t$ chooses $W_t \in \mathcal{W}$, passes $X_t$ through $W_t$, and the output $Y_t$ is its message. This general setup is depicted in Fig. 1. This abstraction indeed captures, among others, communication and privacy constraints as special cases.

[0]Jayadev Acharya is supported in part by the grant NSF-CCF-1846300 (CAREER), NSF-CCF-1815893. Yuhan Liu and Ziteng Sun are supported by the grant NSF-CCF-1846300 (CAREER). Himanshu Tyagi was supported by a grant from Robert Bosch Center for Cyber Physical Systems (RBCCPS) at the Indian Institute of Science.

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

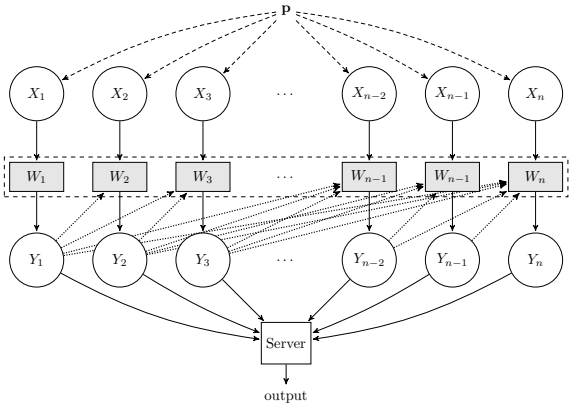


Fig. 1. The information-constrained distributed model. In the interactive setting, $W_t$ can also depend on the previous messages $Y_1, \ldots, Y_{t-1}$ (dotted, upwards arrows).

**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 $\ell$-bit strings.

**Local differential privacy constraints.** For a privacy parameter $\varrho > 0$, a channel $W \colon [2k] \to \{0,1\}^*$ is *$\varrho$-locally differentially private* ($\varrho$-LDP) [1, 2, 3] if

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

We denote by $\mathcal{W}_\varrho$ the set of all $\varrho$-LDP channels.

Based on how the users choose their channels, we consider two natural classes of protocols.

**Noninteractive (SMP[2]) protocols**: All users must select their channels without observing others' messages.

**(Sequentially) interactive protocols**[3]: User $t$ can select $W_t \in \mathcal{W}$ as a function of $Y^{t-1} := (Y_1, \ldots, Y_{t-1})$.

[2]"SMP" stands for *simultaneous-message passing*, *i.e.*, for non-interactive, one-shot protocol. SMP protocols can be divided into private-coin protocols and public-coin protocols based on whether a public string $U$ is available. Since this distinction is not the focus of this paper, we omit it and assume a public string $U$ is available for all SMP protocols.

[3]We focus on sequentially interactive protocols and use the terms "sequentially interactive" and "interactive" interchangeably.

Finally, upon observing the messages, the server performs the desired inference task. We focus on the following tasks.

**Distribution learning.** The goal is to estimate an unknown distribution $\mathbf{p}$ to within $\varepsilon$ in TV distance. Formally, a protocol $\Pi\colon [2k]^n \to \mathcal{Y}^n$ (using $\mathcal{W}$) and an estimator mapping $\hat{\mathbf{p}}\colon \mathcal{Y}^n \to \Delta_{2k}$ constitute an $(n,\varepsilon)$-estimator using $\mathcal{W}$ if

$$\sup_{\mathbf{p}\in\Delta_{2k}} \Pr_{X^n\sim\mathbf{p}}[\,\mathrm{d}_{\mathrm{TV}}(\hat{\mathbf{p}}(Y^n),\mathbf{p}) < \varepsilon\,] \geq 0.99. \tag{1}$$

where $Y^n := \Pi(X^n)$.

**Identity and uniformity testing.** The goal is to test if $\mathbf{p} = \mathbf{q}$ or if it is $\varepsilon$-far from $\mathbf{q}$ in TV distance. Specifically, an $(n,\varepsilon)$-test using $\mathcal{W}$ is given by a protocol $\Pi\colon [2k]^n \to \mathcal{Y}^n$ (using $\mathcal{W}$) and a randomized decision function $T\colon \mathcal{Y}^n \to \{0,1\}$ such that

$$\Pr_{X^n\sim\mathbf{q}^n}[\,T(Y^n) = 0\,] \geq 0.99$$

$$\inf_{\mathbf{p}:\mathrm{d}_{\mathrm{TV}}(\mathbf{p},\mathbf{q})\geq\varepsilon} \Pr_{X^n\sim\mathbf{p}^n}[\,T(Y^n) = 1\,] \geq 0.99.$$

Identity testing for the uniform reference distribution $\mathbf{u}$ over $[2k]$ is termed the $(2k,\varepsilon)$-*uniformity testing* problem.

The *sample complexity* of both learning and testing is then the least $n$ for which such estimators exist.

### A. The role of interactivity

While interactive protocols are arguably harder to implement in a practical setting, they may come with statistical benefits (e.g., smaller sample complexity). Our goal is thus to answer the following question:

*Does interactivity help for learning and testing under local information constraints?*

In this work, we resolve this question by establishing lower bounds that hold for general channel families (modeling local information constraints). As a direct corollary, we show that interaction does not help for learning and testing under either communication constraints or local privacy constraints.

Yet, our lower bounds allow us to identify a family of channels for which interaction strictly helps for testing, establishing the first separation between interactive and noninteractive protocols for distributed goodness-of-fit. We term the family as "leaky-query". Loosely speaking, each user answers a single-bit question about its sample, but with small probability leaks the true sample. Our result then shows that selecting those questions interactively yields a significant advantage.

Our results are summarized in Table I.

TABLE I
LOWER BOUNDS FOR LOCAL INFORMATION-CONSTRAINED LEARNING AND TESTING. THE NONINTERACTIVE BOUNDS WERE KNOWN IN PREVIOUS WORK; THE INTERACTIVE BOUNDS ALL FOLLOW FROM OUR RESULTS. THE BOUND (†) WAS PREVIOUSLY ESTABLISHED IN [4, 5].

| | Learning | Testing | |
|---|---|---|---|
| | Both | SMP | Interactive |
| General | $\frac{k}{\varepsilon^2}\cdot\frac{k}{\|\mathcal{W}\|_*}$ | $\frac{\sqrt{k}}{\varepsilon^2}\cdot\frac{\sqrt{k}}{\|\mathcal{W}\|_F}$ | $\frac{\sqrt{k}}{\varepsilon^2}\cdot\frac{\sqrt{k}}{\sqrt{\|\mathcal{W}\|_*\|\mathcal{W}\|_{\mathrm{op}}}}$ |
| $\ell$-bit | $\frac{k}{\varepsilon^2}\cdot\frac{k}{2^\ell}$ | $\frac{\sqrt{k}}{\varepsilon^2}\cdot\sqrt{\frac{k}{2^\ell}}$ | $\frac{\sqrt{k}}{\varepsilon^2}\cdot\sqrt{\frac{k}{2^\ell}}$ |
| $\varrho$-LDP | $\frac{k}{\varepsilon^2}\cdot\frac{k}{\varrho^2}$ | $\frac{\sqrt{k}}{\varepsilon^2}\cdot\frac{\sqrt{k}}{\varrho^2}$ | $\frac{\sqrt{k}}{\varepsilon^2}\cdot\frac{\sqrt{k}}{\varrho^2}$ (†) |
| Leaky-Query | $\frac{k}{\varepsilon^2}\cdot\sqrt{k}$ | $\frac{\sqrt{k}}{\varepsilon^2}\cdot\sqrt{k}$ | $\frac{\sqrt{k}}{\varepsilon^2}\cdot\sqrt[4]{k}$ |

### B. Our techniques

Central to our techniques is an average information bound (Theorem 3) which appears in both learning and testing bounds and connects them in an intriguing way. It is folklore that "a good learner must be able to test." Indeed, following a standard reduction from learning to testing, we obtain an Assouad-type argument which relates learning to the average mutual information between the messages and the unknown parameters. One of our main results is a bound for this average mutual information, which in turn leads to a lower bound for learning. Surprisingly, the very same average mutual information bound for learning also characterizes the difficulty of testing. Hence, "a good test must learn" at least something about the underlying distribution, a natural heuristic which we formalize in our bound for testing.

Our technical contributions are non-trivial: interactivity may give rise to complicated correlations which are absent in the noninteractive case and are difficult to handle. In fact, several prior works [6, 7, 8] have tried to handle these correlations and claimed a subset of the results presented in Table I. However, we found technical gaps in these works when handling the correlation.[4] We circumvent this difficulty by a careful application of convexity for KL divergence to break down the mixture distributions in Assouad's method.

We describe the lower bound construction in the next section, followed by related information quantities. We present bounds on these quantities in Sections III and IV for learning and testing, respectively. We conclude with the complete sample complexity bounds in Section V.

## II. LOWER BOUND CONSTRUCTION

Our lower bounds rely on a family of perturbed distributions around $\mathbf{u}$, the uniform distribution over $[2k]$. The construction [9] consists of $2^k$ distributions parameterized

[4]Important exceptions are [4] and [5], which both obtain a tight bound for testing under local privacy constraints.

by $\mathcal{Z} = \{-1,+1\}^k$. Specifically, for $z \in \mathcal{Z}$ the distribution $\mathbf{p}_z$ over $[2k]$ is given by

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

where $\forall i \in [k], z_i$ is the $i$th entry of $z$. Each such $\mathbf{p}_z$ is therefore at total variation distance exactly $\varepsilon$ from $\mathbf{u}$. Let $Z$ be distributed uniformly over $\{-1,+1\}^k$. Conditioned on $Z$, let $X^n$ denote independent samples from $\mathbf{p}_Z$, and $Y^n$ denote the messages sent to the server through the channels. We denote the distribution of $Y^n$ obtained from the generating process above by $\mathbf{q}^{Y^n}$. We use $\mathbf{u}^{Y^n}$ to denote the corresponding distribution if $X^n$ are generated from $\mathbf{u}$. Using Assouad's lemma and Le Cam's method, respectively, we can relate learning to the amount of information the server can extract about $Z$ using $Y^n$ and testing to the KL divergence between $\mathbf{q}^{Y^n}$ and $\mathbf{u}^{Y^n}$.

**Lemma 1** (Learning: 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 $Y^n$ be the corresponding transcript. Then, we must have*

$$\sum_{i=1}^{k} I(Z_i \wedge Y^n) \geq \frac{k}{2}. \quad (3)$$

**Lemma 2** (Testing: Le Cam's method)**.** *Consider local constraints $\mathcal{W}$ and $\varepsilon \in (0,1]$. If $(\Pi, T)$ solves $(2k, \varepsilon)$-uniformity testing using $\mathcal{W}$, then we must have*

$$\mathrm{D}\Big(\mathbf{q}^{Y^n} \| \mathbf{u}^{Y^n}\Big) \geq c, \quad (4)$$

*where $c > 0$ is an absolute constant.*

Our techniques focus on analyzing how the constraints on the channels $\mathcal{W}$ will limit these quantities in (3) and (4). The lower bounds we develop associate with 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 (5)). It captures the "informativeness" of the channel $W$. $\forall i, j \in [k]$, define $H(W)_{i,j}$ as

$$\sum_{y \in \mathcal{Y}} \frac{(W(y \mid 2i-1) - W(y \mid 2i))(W(y \mid 2j-1) - W(y \mid 2j))}{\sum_{x \in [2k]} W(y \mid x)} \quad (5)$$

The spectrum of these matrices $H(W)$, for $W \in \mathcal{W}$, will play a central role in our results.[5] In particular, for a given family of local constraints $\mathcal{W}$, the following quantities will be used:

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

[5]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 in Section II. However, the ordering of the elements is arbitrary and we can associate this matrix with any partition of the domain into equal parts.

Two key inequalities to interpret our results are

$$\|\mathcal{W}\|_F^2 \leq \|\mathcal{W}\|_{\mathrm{op}} \|\mathcal{W}\|_{*}, \ \|\mathcal{W}\|_{\mathrm{op}} \leq \|\mathcal{W}\|_F \leq \|\mathcal{W}\|_{*}, \quad (6)$$

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

## III. Interactive learning under information constraints

Our first main result, Theorem 3, is an upper bound for the average mutual information $\sum_{i=1}^{k} I(Z_i \wedge Y^n)$. Upon combining this bound with Lemma 1, we obtain the lower bound for distribution learning, stated in Theorem 8.

Interestingly, the same bound will be useful for the testing problem as well, and is one of the key components of our lower bound recipes in this paper. The proof of both theorems can be found in the full version [10].

**Theorem 3** (Average Information Bound)**.** *For $\varepsilon \in (0, 1/4]$, let $Y^n$ be the transcript of an interactive protocol using $\mathcal{W}$, when the input is generated using $\mathbf{p}_Z$ from (2) with a uniform $Z$. Then, for every $1 \leq t \leq n$,*

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

*Proof Sketch.* The main difficulty lies in dealing with the posterior distribution of the message $Y_t$ given the past $Y^{t-1}$ under the mixture distributions

$$\mathbf{p}_{+i}^{Y^t} := \frac{1}{2^{k-1}} \sum_{z : z_i = +1} \mathbf{p}_z^{Y^t}, \quad \mathbf{p}_{-i}^{Y^t} := \frac{1}{2^{k-1}} \sum_{z : z_i = -1} \mathbf{p}_z^{Y^t},$$

Previous works [6, 7, 8] mistakenly treated this as uniform mixtures of $\mathbf{p}_z^{Y_t}$, essentially reducing the problem to the noninteractive setting. In this work, we handle this correlation using a convexity argument, which reduces the problem of distinguishing two mixtures to distinguishing every pair of "neighboring points" in the mixture.

For $z \in \{-1,+1\}^k$, write $z^{\oplus i}$ for $z$ with the $i$th coordinate flipped. Using the convexity of KL divergence we get

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

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

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

Next, we note that since $z^{\oplus i}$ and $z$ only differ at the $i$th coordinate,

$$\Pr_{\mathbf{p}_z}\big[Y_r = y \mid Y^{r-1}\big] = \Pr_{\mathbf{p}_{z^{\oplus i}}}\big[Y_r = y \mid Y^{r-1}\big] + \frac{2\varepsilon z_i}{k}\Big(W^{Y^{r-1}}(y \mid 2i-1) - W^{Y^{r-1}}(y \mid 2i)\Big). \quad (7)$$

Using (7), and the fact that the KL divergence is bounded by the chi square distance, we have

$$\mathrm{D}\Big(\mathbf{p}_z^{Y_r|Y^{r-1}} \| \mathbf{p}_{z^{\oplus i}}^{Y_r|Y^{r-1}}\Big) \le \frac{16\varepsilon^2}{k} H(W^{Y^{r-1}})_{i,i}.$$

It follows that

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

concluding the proof. □

## IV. Interactive testing under information constraints

### A. The testing bound

In this section, we bound the KL divergence stated in (4). We proceed by first applying the chain rule for the KL divergence to break it into contribution for each sample,

$$\mathrm{D}\Big(\mathbf{q}^{Y^n} \| \mathbf{u}^{Y^n}\Big) = \sum_{t=0}^{n-1} \mathbb{E}_{\mathbf{q}^{Y^t}} \Big[ \mathrm{D}\Big( \mathbf{q}^{Y_{t+1}|Y^t} \| \mathbf{u}^{Y_{t+1}|Y^t} \Big) \Big]. \quad (8)$$

*Remark* 4 (Comparison with decoupled chi square bounds). Before proceeding, we draw contrast with the *decoupled chi square divergence* bound technique developed [11]. In that work, the first step was to bound KL 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. (8) allows us to work with one sample at a time. As we we will see below, switching KL 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 [11]. But now we need to handle the posterior distribution of the message $Y_t$ given the past $Y^{t-1}$, under the mixture distribution.

We now present the key technical component of our testing bound in the result below, which relates the per-round divergence to the spectrum of $H(W)$ and average information obtained at each round, which surprisingly, is the information quantity we bound for distribution learning in Theorem 3.

**Lemma 5** (Per-round divergence bound)**.** *For every $0 \le t \le n-1$, we have*

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

*Proof Sketch.* Fix $t$. As chi square divergence upper bounds KL divergence, we have

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

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

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

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

we get

$$\begin{aligned}&\mathbb{E}_{\mathbf{q}^{Y^t}} \Big[ \mathrm{D}\Big( \mathbf{q}^{Y_{t+1}|Y^t} \| \mathbf{u}^{Y_{t+1}|Y^t} \Big) \Big] \\ &\le \frac{2\varepsilon^2}{k} \mathbb{E}_{\mathbf{q}^{Y^t}} \Big[ \mathbb{E}\big[Z \bigm| Y^t\big]^T H(W^{Y^t}) \mathbb{E}\big[Z \bigm| Y^t\big] \Big] \quad (9) \\ &\le \frac{2\varepsilon^2}{k} \mathbb{E}_{\mathbf{q}^{Y^t}} \Big[ \Big\| H(W^{Y^t}) \Big\|_{\mathrm{op}} \big\| \mathbb{E}\big[Z \bigm| Y^t\big] \big\|_2^2 \Big]. \quad (10)\end{aligned}$$

Note that since $Z$ is uniformly sampled from $\{-1,+1\}^k$, we have $\mathbb{E}\Big[ \|\mathbb{E}[Z]\|_2^2 \Big] = 0$. Intuitively, $\mathbb{E}\Big[ \|\mathbb{E}[Z \mid Y^t]\|_2^2 \Big]$ indicates how well $Z$ can be estimated conditioned on observing $Y^t$.[6] The next lemma shows that it can be bounded by the mutual information between $Y^t$ and each $Z_i$, which might be of independent interest.

**Lemma 6.** *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*

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

Take $Y = Y^t$ in Lemma 6 and sum over $[k]$, we get

$$\mathbb{E}\Big[ \big\| \mathbb{E}\big[Z \bigm| Y^t\big] \big\|_2^2 \Big] \le 2 \ln 2 \sum_{i=1}^{k} I\big(Z_i \wedge Y^t\big).$$

Combining with (10), we obtain the result. □

*Remark* 7. The bound in (10) relates the per-sample divergence to the "uncertainty" about $Z$ given $Y^{t-1}$. A key heuristic underlying our analysis, formalized by this bound, is the thesis that when the information gathered about $Z$ is small, the "distance" contribued by the next sample cannot be too much. Since testing requires this distance to become large, our bound implies that to test, we must learn something about $Z$. This connection between learning and testing bounds is interesting in its own right.

[6] In fact, $\mathbb{E}\Big[ \mathbb{E}\big[Z_i \bigm| Y^t\big]^2 \Big]$ equals 1 minus the minimum mean squared error for estimating $Z_i$ from $Y^t$.

Upon combining Lemma 5 with (8), summing over $t$, and using the average information bound of Theorem 3, we get

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

For $\mathrm{D}\big(\mathbf{q}^{Y^n}\|\mathbf{u}^{Y^n}\big)$ to be $\Omega(1)$, this proves a bound $n = \Omega(k/(\sqrt{\|\mathcal{W}\|_{\mathrm{op}}\|\mathcal{W}\|_*}\varepsilon^2))$ for interactive testing, formally stated in Theorem 9.

### B. Our general bound

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 in this work. However, the power of our techniques goes beyond that specific evaluation. In fact, the step from (9) to (10) can be weak. For example, for a class of erasure channel consider in the full version [10], we can obtain stronger bounds using (9) directly.

The proof also gives insights to designing a family of channels that yields the desired separation. $\|\mathcal{W}\|_F$ should be maximally separated from $\sqrt{\|\mathcal{W}\|_*\|\mathcal{W}\|_{\mathrm{op}}}$, and for each vector $\mathbb{E}[Z \mid Y^t]$ we can find a $W$ such that the top eigenvector of $H(W)$ aligns with $\mathbb{E}[Z \mid Y^t]$. We state the family of channels below.

### C. The leaky-query channel

We now formalize the *leaky-query* channel which meet the objectives above. For $u \in [0,1]^{2k}$, and $\mathcal{Y} := [2k] \cup \{\mathbf{1}^\star, \mathbf{0}^\star\}$, given an input $x \in [2k]$, the channel $W_u$ outputs $x$ with probability $\eta = 1/\sqrt{k}$; otherwise it outputs $\mathbf{1}^\star$ and $\mathbf{0}^\star$ with probability $u_x$ and $1-u_x$ respectively. Let $\mathcal{W} = \{W_u : u \in [0,1]^{2k}\}$

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

The required norms satisfy

$$\|\mathcal{W}\|_* = \Theta(\sqrt{k}),\ \|\mathcal{W}\|_F = \Theta(1),\ \|\mathcal{W}\|_{\mathrm{op}} = \Theta(1).$$

The matching noninteractive scheme only uses the leaked samples and hence requires $k/\varepsilon^2$ samples. The matching interactive scheme consists of two stages: in the first stage we obtain a set $S$ of the symbols that are leaked. In the second stage, we set $u_x = \mathbb{1}_{\{x \in S\}}$. The idea is that $\mathbf{u}(S) = |S|/2k$ while $\mathbf{p}(S)$ is noticeably large when $\mathbf{p}$ is far from $\mathbf{u}$. We perform binary hypothesis test to separate the two cases. See the full version [10] for details.

## V. Lower bounds for learning and testing

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

### A. Learning

Combining Lemma 1 and Theorem 3, we obtain the following bound on distribution learning under channel constraint $\mathcal{W}$.

**Theorem 8.** *The sample complexity of $(2k,\varepsilon)$-distribution learning under local constraints $\mathcal{W}$ using interactive protocols is*

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

This bound matches the known lower bound for learning with noninteractive *private-coin* protocols in [11]. By properly bounding $\|\mathcal{W}\|_*$ for LDP and communication-limited setting, we obtain tight bounds for distribution learning under these settings. See Table I for the bounds and the full version [10] for detailed statements and references to achievability results.

### B. Testing

Combining Lemma 2, Theorem 3, and Lemma 5, we obtain a general lower bound for uniformity testing (and thus, *a fortiori*, on the more general problem of identity testing).

**Theorem 9.** *The sample complexity of $(2k,\varepsilon)$-uniformity testing under local constraints $\mathcal{W}$ using interactive protocols is*

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

[11] previously established an $\Omega\Big(\frac{k}{\varepsilon^2\|\mathcal{W}\|_F}\Big)$ lower bound for noninteractive protocols. By (6), we know the interactive lower bound is at least as small for any channel family $\mathcal{W}$. For LDP and communication-limited channels, the bounds are the same, and can be achieved using noninteractive protocols. The bounds are listed in Table I. The detailed analysis and references to achievability results can be found in [10].

### C. A separation

By relations between matrix norms (6), it can be seen that the noninteractive public-coin lower bound of $\Omega\Big(\frac{k}{\varepsilon^2\|\mathcal{W}\|_F}\Big)$ from [12] can be up to a $k^{1/4}$ factor smaller than the bound in Theorem 9 for interactive protocols. Guided by the analysis of the proof of Theorem 9, we show that this maximal separation is achievable for uniformity testing. See the full version [10] for details.

**Theorem 10.** *For the* leaky-query *channels, 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.*

[1] A. V. Evfimievski, J. Gehrke, and R. Srikant, "Limiting privacy breaches in privacy preserving data mining," in *PODS*. ACM, 2003, pp. 211–222.

[2] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Theory of cryptography*, ser. Lecture Notes in Comput. Sci. Springer, Berlin, 2006, vol. 3876, pp. 265–284.

[3] 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, 2011.

[4] K. Amin, M. Joseph, and J. Mao, "Pan-private uniformity testing," in *Proceedings of Thirty Third Conference on Learning Theory*, ser. Proceedings of Machine Learning Research, J. Abernethy and S. Agarwal, Eds., vol. 125. PMLR, 09–12 Jul 2020, pp. 183–218. [Online]. Available: http://proceedings.mlr.press/v125/amin20a.html

[5] T. B. Berrett and C. Butucea, "Locally private non-asymptotic testing of discrete distributions is faster using interactive mechanisms," in *Advances in Neural Information Processing Systems 33*, 2020.

[6] 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.

[7] Y. Han, P. Mukherjee, A. Özgür, and T. Weissman, "Distributed statistical estimation of high-dimensional and non-parametric distributions," in *Proceedings of the 2018 IEEE International Symposium on Information Theory (ISIT'18)*, 2018, pp. 506–510.

[8] Y. Han, A. Özgür, and T. Weissman, "Geometric lower bounds for distributed parameter estimation under communication constraints," in *Conference On Learning Theory*. PMLR, 2018, pp. 3163–3188.

[9] L. Paninski, "A coincidence-based test for uniformity given very sparsely sampled discrete data," *IEEE Transactions on Information Theory*, vol. 54, no. 10, pp. 4750–4755, 2008.

[10] J. Acharya, C. L. Canonne, Y. Liu, Z. Sun, and H. Tyagi, "Interactive inference under information constraints," *CoRR*, vol. abs/2007.10976, 2020.

[11] J. Acharya, C. L. Canonne, and H. Tyagi, "Inference under information constraints i: Lower bounds from chi-square contraction," *IEEE Transactions on Information Theory*, vol. 66, no. 12, pp. 7835–7855, 2020, full version of [12].

[12] ——, "Inference under information constraints: Lower bounds from chi-square contraction," in *Proceedings of the Thirty-Second Conference on Learning Theory*, ser. Proceedings of Machine Learning Research, A. Beygelzimer and D. Hsu, Eds., vol. 99. Phoenix, USA: PMLR, Jun. 2019, pp. 3–17.