Inference under Information Constraints I: Lower Bounds from Chi-Square Contraction

Jayadev Acharya* *Member, IEEE* Clément L. Canonne† Himanshu Tyagi‡ *Senior Member, IEEE*

***Abstract*—Multiple players are each given one independent sample, about which they can only provide limited information to a central referee. Each player is allowed to describe its observed sample to the referee using a channel from a family of channels $\mathcal{W}$, which can be instantiated to capture, among others, both the communication- and privacy-constrained settings. The referee uses the players' messages to solve an inference problem on the unknown distribution that generated the samples. We derive lower bounds for the sample complexity of learning and testing discrete distributions in this information-constrained setting.**

Underlying our bounds is a characterization of the contraction in chi-square distance between the observed distributions of the samples when information constraints are placed. This contraction is captured in a local neighborhood in terms of chi-square and decoupled chi-square fluctuations of a given channel, two quantities we introduce. The former captures the average distance between distributions of channel output for two product distributions on the input, and the latter for a product distribution and a mixture of product distribution on the input. Our bounds are tight for both public- and private-coin protocols. Interestingly, the sample complexity of testing is order-wise higher when restricted to private-coin protocols.

I. Introduction

Large-scale distributed inference has taken a center stage in many machine learning tasks. In these settings, it is becoming increasingly critical to operate under limited resources at each player, where the players (who hold the data samples) may be limited in their computational capabilities, communication capabilities, or may restrict the information about their data to maintain privacy. Our focus in this work will be on the last two constraints of communication and privacy, and, in general, on local information constraints on each player's data.

We propose the following general framework for distributed statistical inference under local information constraints. There are n players observing independent samples $X_1, \dots, X_n$ from an unknown distribution $\mathbf{p}$ over a domain $\mathcal{X}$, with player i getting the sample $X_i \in \mathcal{X}$. The players want to enable a central referee $\mathcal{R}$ to complete an inference task about their data. However, the players are constrained in the amount of information they can reveal to $\mathcal{R}$ about their observations in the following manner: Player i must choose a channel W_i from a prespecified class of channels $\mathcal{W}$ whose input alphabet is $\mathcal{X}$ and output alphabet is $\mathcal{Y}$, and use it to report its observed sample to $\mathcal{R}$.[1] That is, player i passes its observation X_i as input to its chosen channel W_i and $\mathcal{R}$ receives the channel's output Y_i. The central referee then uses messages $Y_1, \dots, Y_n$ from the players to complete the inference task of interest, such as estimation or testing for the underlying distribution $\mathbf{p}$; Fig. 1 illustrates the setup.

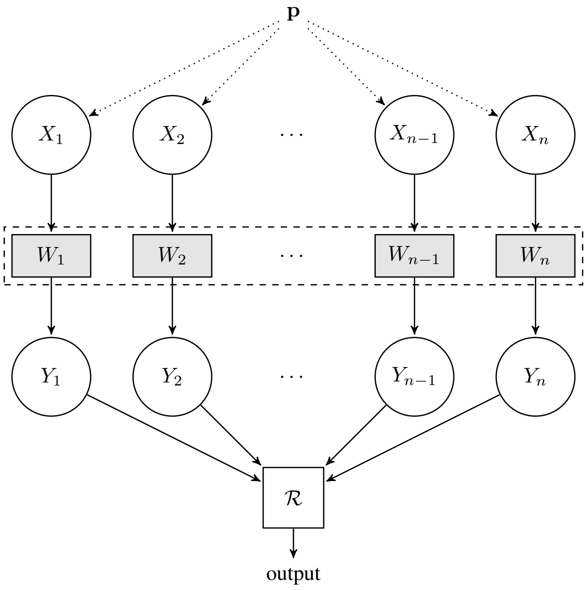

Fig. 1. The information-constrained distributed model. In the private-coin setting the channels $W_1, \dots, W_n$ are independent, while in the public-coin setting they are jointly randomized.

The family of allowed channels $\mathcal{W}$ serves as an abstraction of the local information constraints placed on each player's message to $\mathcal{R}$. Before moving ahead, we instantiate this abstraction with two important examples, local communication constraints and local privacy constraints, and specify the corresponding $\mathcal{W}$s.

(a) *Communication-Limited Inference.* Each player can only send ℓ bits about their sample, *i.e.*, player i can send a message $Y_i \in \{0,1\}^\ell$. This constraint is captured by considering the set of allowed channels $\mathcal{W} = \mathcal{W}_\ell :=$

*Cornell University. Email: acharya@cornell.edu.

†IBM Research. Email: ccanonne@cs.columbia.edu. Part of this work was performed while supported by a Motwani Postdoctoral Fellowship at Stanford University.

‡Indian Institute of Science. Email: htyagi@iisc.ac.in.

Jayadev Acharya is supported in part by the grant NSF-CCF-1846300 (CAREER), NSF-CCF-1815893, and a Google Faculty Fellowship. Part of this work was performed while Clément Canonne was supported by a Motwani Postdoctoral Fellowship at Stanford University. Himanshu Tyagi is supported in part by a research grant from the Robert Bosch Center for Cyberphysical Systems (RBCCPS), Indian Institute of Science, Bangalore.

[1] A channel W from $\mathcal{X}$ to $\mathcal{Y}$ is a randomized mapping $W\colon \mathcal{X} \to \mathcal{Y}$. We represent it by a $|\mathcal{Y}| \times |\mathcal{X}|$ *transition probability matrix* W whose rows and columns are indexed by $y \in \mathcal{Y}$ and $x \in \mathcal{X}$, respectively, and its (y,x)th entry $W(y \mid x) \coloneqq W_{y,x}$ is the probability of observing y when the input to the channel is x.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2020.3028440, IEEE Transactions on Information Theory

$\{W\colon \mathcal{X} \to \{0,1\}^\ell\}$, the family of channels with output alphabet $\{0,1\}^\ell$.

(b) *Locally Differentially Private Inference.* Each player seeks to maintain privacy of their own data. We adopt the notion of local differential privacy which, loosely speaking, requires that no output message from a player reveals too much about its sample. This is captured by restricting $\mathcal{W}$ to $\mathcal{W}_\rho$, the family of *ρ-locally differentially private* (ρ-LDP) channels $W\colon \mathcal{X} \to \{0,1\}^*$ that satisfy the following (*cf.* [23], [38], [9], [21]): For $\rho > 0$,

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

These specific cases of communication and privacy constraints have received significant attention in the literature, and we emphasize these cases separately in our results. We emphasize, however, that our results are valid for arbitrary families $\mathcal{W}$ and can handle other examples from the literature such as the t-step Markov transition matrices considered in [10].

Our proposed framework and the general tools we develop are applicable to statistical inference for any family of distributions, with the caveat that deriving concrete results for a general family will require more work. Our focus in this work will be on discrete distributions, *i.e.*, distributions on a finite alphabet $\mathcal{X}$. For this setting, we consider the canonical inference problems of estimating $\mathbf{p}$ and testing goodness-of-fit, under both communication and privacy constraints. Motivated by applications in distributed inference under resource constraints, we seek algorithms that enable the desired inference using the fewest samples possible, or equivalently, the least number of players. Our main results present a general approach for establishing lower bounds for the *sample complexity* of performing a given inference task under the aforementioned information-constrained setting. Underlying our lower bounds is a new quantitative characterization of contraction in the chi-square distance between two output message distributions, in comparison to that between the corresponding input distributions, as a function of the imposed information constraints represented by $\mathcal{W}$.

We allow randomized selection of Ws from $\mathcal{W}$ at each player and distinguish between *private-coin protocols*, where this randomized selection is done independently for each player, and *public-coin protocols*, where the players can use shared randomness. Interestingly, our chi-square contraction bounds provide a quantitative separation of sample complexity for private-coin and public-coin protocols, an aspect hitherto ignored in the distributed inference literature and which is perhaps the main contribution of our work.

We summarize our results below, after a formal description of our problem setting.

A. Information-constrained inference framework

We begin by recalling standard formulations for learning and testing discrete distributions in the classical non-distributed setting. Denote by Δ_k the set of all distributions over $[k] \coloneqq \{1,\dots,k\}$. We set $\mathcal{X}$ to be $[k]$ and the set of unknown distributions to Δ_k. Let $X^n \coloneqq (X_1,\dots,X_n)$ be independent samples from an unknown distribution $\mathbf{p} \in \Delta_k$. We focus on the following two inference tasks.

Distribution Learning. In the (k,ε)-distribution learning problem, we seek to estimate a distribution $\mathbf{p}$ in Δ_k to within ε in total variation distance. Formally, a (randomized) mapping $\hat{\mathbf{p}}\colon \mathcal{X}^n \to \Delta_k$ constitutes an (n,ε)-estimator if

$$\sup_{\mathbf{p}\in\mathcal{P}} \Pr_{X^n\sim\mathbf{p}}[\, \mathrm{d}_{\rm TV}(\hat{\mathbf{p}}(X^n), \mathbf{p}) > \varepsilon \,] < \frac{1}{12},$$

where $\mathrm{d}_{\rm TV}(\mathbf{p},\mathbf{q})$ denotes the total variation distance between $\mathbf{p}$ and $\mathbf{q}$ (see Section II for definition of total variation distance). Namely, $\hat{\mathbf{p}}$ estimates the input distribution $\mathbf{p}$ to within distance ε with probability at least $11/12$. This choice of probability is arbitrary and has been chosen for convenience; see Footnote 12 to see where it is exactly used.

The *sample complexity* of (k,ε)-distribution learning is the least n such that there exists an (n,ε)-estimator for $\mathbf{p}$. It is well-known that the sample complexity of distribution learning is $\Theta(k/\varepsilon^2)$, and the empirical distribution attains it.

Identity Testing. In the (k,ε)-identity testing problem, given a known reference distribution $\mathbf{q} \in \mathcal{P}$, 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 is given by a (randomized) mapping $\mathcal{T}\colon \mathcal{X}^n \to \{0,1\}$ such that

$$\Pr_{X^n\sim\mathbf{p}^n}[\, \mathcal{T}(X^n) = 1 \,] > \frac{11}{12} \text{ if } \mathbf{p} = \mathbf{q},$$
$$\Pr_{X^n\sim\mathbf{p}^n}[\, \mathcal{T}(X^n) = 0 \,] > \frac{11}{12} \text{ if } \mathrm{d}_{\rm TV}(\mathbf{p},\mathbf{q}) > \varepsilon.$$

In other words, upon observing independent samples X^n, the algorithm 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 $\mathbf{q}$. Note again that the choice of $1/12$ for probability of error is for convenience.[2]

The *sample complexity* of (k,ε)-identity testing is the least n for which there exists an (n,ε)-test for $\mathbf{p}$. Clearly, this quantity can depend on the reference distribution $\mathbf{q}$. However, it is customary to consider the sample complexity over the worst-case $\mathbf{q}$.[3] In this worst-case setting, while it has been known for some time that the most stringent sample requirement arises for $\mathbf{q}$ set to the uniform distribution, a recent result of [29] provides a formal reduction of arbitrary $\mathbf{q}$ to the uniform distribution case. It is therefore enough to restrict $\mathbf{q}$ to being the uniform distribution; identity testing for the uniform reference distribution is termed the (k,ε)-*uniformity testing* problem. The sample complexity of (k,ε)-uniformity testing was shown to be $\Theta\big(\sqrt{k}/\varepsilon^2\big)$ in [41].

We consider the two inference tasks above in our information-constrained setting. Let $\mathcal{W}$ be the set of allowed channels describing the constraints, and let as before $X_1,\dots,X_n$ be generated independently from an unknown distribution $\mathbf{p} \in$

[2] 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 $\mathcal{H}_1 = \{\, \mathbf{q}' \in \Delta_k : \mathrm{d}_{\rm TV}(\mathbf{q}',\mathbf{q}) > \varepsilon \,\}$ in a minmax setting, with both type-I and type-II errors set to $1/12$.

[3] The sample complexity for a fixed $\mathbf{q}$ has been studied under the "instance-optimal" setting (see [50], [12]): while the question is not fully resolved, nearly tight upper and lower bounds are known.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2020.3028440, IEEE Transactions on Information Theory

Δ_k. Player i chooses a channel $W_i \in \mathcal{W}$, passes its input X_i through $W_i \in \mathcal{W}$ and the output message Y_i constitutes information shared by player i with $\mathcal{R}$. For a given choice of channel W and $y \in \mathcal{Y}$, denote by $W\mathbf{p}$ the probability

$$W\mathbf{p}(y) := \sum_x \mathbf{p}(x) W(y \mid x) = \mathbb{E}_{\mathbf{p}}[W(y \mid X)]; \qquad (1)$$

namely, $W\mathbf{p}$ is the distribution of the output message for a choice $W \in \mathcal{W}$ of the channel. The referee $\mathcal{R}$, upon observing the messages Y^n from the players, seeks to solve the two inference tasks of (k, ε)-distribution estimation and (k, ε)-identity testing.

In choosing the channels W from $\mathcal{W}$, the players can be allowed to follow protocols with different information structures. In the most general case, the choice W_i of each player can depend on all the messages sent by the previous player and shared, *public coins* available to the players. In this work, we do not allow this most general class of protocols and restrict our attention to *simultaneous message passing* (SMP) protocols for communication. In an SMP protocol, the messages $Y_1, \ldots, Y_n$ from all players are transmitted simultaneously to the central server, and no other communication is allowed. However, we consider both the cases where public coins are and are not available. Note that this is equivalent to choosing the channels $W_1, \ldots, W_n$ simultaneously, possibly using public coins when they are available.

Note that the SMP setting forbids communication between the players, but does allow them to *a priori* agree on a strategy to select different mappings W_i from $\mathcal{W}$. In this context, the role of shared randomness available to the players is important and motivates us to distinguish the settings of *private-coin* and *public-coin* protocols. In fact, as pointed-out earlier, a central theme of this work is to demonstrate the role of shared randomness available as public coins in enabling distributed inference. We show that it is indeed a resource that can greatly reduce the sample complexity of distributed inference.

Formally, the private- and public-coin SMP protocols are described as follows.

Definition I.1 (Private-coin SMP Protocols)**.** Let $U_1, \ldots, U_n$ denote independent random variables, which are independent jointly of $(X_1, \ldots, X_n)$.[4] In a *private-coin* SMP protocol, player i is given access to U_i and the channel $W_i \in \mathcal{W}$ is chosen as a function of U_i. The central referee $\mathcal{R}$ does not have access to the realization of $U^n := (U_1, \ldots, U_n)$ used to generate the W_is; however, it knows the mapping from U_is to W_is.

Definition I.2 (Public-coin SMP Protocols)**.** Let U be a random variable independent of $(X_1, \ldots, X_n)$. In a *public-coin* SMP protocol, all players are given access to U, and they select their respective channels $W_i \in \mathcal{W}$ as a function of U. The central referee $\mathcal{R}$ is given access to the realization of U as well, and its estimator and test can depend on U.

Note that in a private-coin SMP protocol, the channels $W_1, \ldots, W_n$ are independent since the U_is are independent. Further, since X_is are independent samples from $\mathbf{p}$, the messages Y_is are also independent across the players. In particular, the distribution of $Y^n = (Y_1, \ldots, Y_n)$ is a product distribution. In contrast, in a public-coin SMP protocol, the channels W_i (and hence Y_is) are chosen as functions of the same random variable U and therefore need not be independent. Nonetheless, even for public-coin SMP protocols, the messages $Y_1, \ldots, Y_n$ are independent conditioned on the shared randomness U.

Remark I.3. Throughout we assume that some randomness is available to generate the output of the channel W_i given its input X_i. This randomness is assumed to be private as well (namely, it is independent across the players and is not available to $\mathcal{R}$). This assumption stands even for public-coin SMP protocols, implying the conditional independence of Y_is given U mentioned above, and is important in the context of privacy where the information available to $\mathcal{R}$ is seen as "leaked" and private randomness available only to the players is critical for enabling LDP channels.

We now define information-constrained discrete distribution estimation and testing at the referee.

For (k, ε)-distribution learning, an *estimator using* $\mathcal{W}$ comprises an SMP protocol that produces the messages Y^n and an estimator mapping $\hat{p}$ that is applied by $\mathcal{R}$ to the messages Y^n. An (n, ε)*-estimator using* $\mathcal{W}$ is defined analogously to the centralized setting, by replacing the input X^n of $\hat{p}$ with (Y^n, U) and Y^n, respectively, for public-coin and private-coin[5] SMP protocols. Similarly, for (k, ε)-identity testing, a test using $\mathcal{W}$ comprises an SMP protocol and a test mapping $\mathcal{T}$ applied by $\mathcal{R}$, and an (n, ε)*-test using* $\mathcal{W}$ is defined analogously to the centralized setting. We emphasize that the shared randomness used by the players (except that for realizing W) is available to $\mathcal{R}$, which only strengthens our lower bounds. Our main quantity of interest in this work is the following.

Definition I.4. The sample complexity of (k, ε)-distribution learning or (k, ε)-identity testing (or (k, ε)-uniformity testing) using $\mathcal{W}$ for public-coin protocols, respectively, is the least n such that there exists an (n, ε)-estimator or (n, ε)-test using $\mathcal{W}$ with a public-coin SMP protocol. The sample complexities of these tasks using $\mathcal{W}$ for private-coin protocols is defined analogously.

Since we are restricting to one sample per player, the sample complexity of these problems corresponds to the least number of players required to solve them as well. Our main objective in this line of work is the following:

Characterize the sample complexity for inference tasks using $\mathcal{W}$ for private- and public-coin protocols.

B. Summary of our techniques

We are initiating a systematic study of the distributed inference problems described in the previous section. In this paper, the first in our series, we shall focus on lower

[4] In this work, we are not concerned with the amount of private or public randomness used. Thus, we can assume U_is to be discrete random variables, distributed uniformly over a domain of sufficiently large cardinality.

[5] It is important to note that $\mathcal{R}$ does not have access to the private randomness U^n. In fact, our lower bounds for private-coin protocols may not hold if the output at $\mathcal{R}$ can depend on private randomness of the players.

bounds. For input distributions $\mathbf{p}$ and $\mathbf{q}$ over $\mathcal{X}$, the output messages are $W\mathbf{p}$ and $W\mathbf{q}$ over $\mathcal{Y}$. By data-processing inequalities, the output distributions $W\mathbf{p}$ and $W\mathbf{q}$ are "closer" than the corresponding input distributions $\mathbf{p}$ and $\mathbf{q}$, which makes it harder to perform inference using messages Y_is than using the inputs X_is themselves. We provide a quantitative characterization of this reduction in distance between the output distributions compared to the input distributions for the chi-square distance, which we term *chi-square contraction*, and use it to derive lower bounds for distributed inference problems.

In more detail, we study chi-square contractions for an ε-perturbed family (see Definition III.1), a collection of probability distributions that are obtained by perturbing a nominal distribution. The perturbed family of distributions is chosen carefully to ensure that in order to accomplish the given inference task, an algorithm must roughly distinguish the perturbed distributions. In particular, we relate the difficulty of inference problems using two notions of distances: (a) the average chi-square distance between the perturbed distributions to the nominal distribution and (b) the chi-square distance of the average perturbed distribution to the nominal distribution. For our distributed inference setting, we need to bound these two quantities for the *induced perturbed family* of distributions at the outputs of the chosen channels from $\mathcal{W}$.

We provide bounds for these two quantities for channel output distributions in terms of two new measures of average distance in a neighborhood: the *chi-square fluctuation* for the average distance and the *decoupled chi-square fluctuation* for the distance to the average. The former notion has appeared earlier in the literature, albeit in different forms, and recovers known bounds for distributed distribution learning problems. The second quantity, the decoupled chi-square fluctuation, is the main technical tool introduced in this work, and leads to new lower bounds for distributed identity testing. Heuristically, the smaller these quantities are, the closer are the distributions of a perturbed family to the center, which in turn makes it harder to distinguish them and results in a higher sample complexity.

Observe that the general approach sketched above can be applied to any perturbed family. We obtain lower bounds for private-coin protocols by a maxmin evaluation of these bounds, where the maximum is over the choice of channels from $\mathcal{W}$ and the minimum is over the choice of perturbed families. In other words, since the channels are chosen independently of each other, for any given choice of channels, we will design a specific perturbed family to give rise to a small fluctuation bound. In contrast, we obtain lower bounds for public-coin protocols by a minmax evaluation of these bounds where the minimum is over perturbed families and the maximum is over the choice of channels from $\mathcal{W}$. In this case, we design a perturbed family such that for any choice of channels (chosen using the shared randomness), we can upper bound the chi-squared fluctuations. Remarkably, we establish that the maxmin evaluation can be significantly smaller than the minmax evaluation, leading to a quantitative separation in performance of private-coin and public-coin protocols for testing problems.

This separation has a heuristic appeal: On the one hand, in public-coin protocols players can use shared randomness to sample channels that best separate the current point in the alternative hypothesis class from the null. On the other hand, for a fixed private-coin protocol, one can identify a perturbed family in a "direction" where the current choice of channels will face difficulty in distinguishing the elements of the perturbed family. Further, we remark that this separation only holds for testing problems. This, too, makes sense in light of the previous heuristic since learning problems require us to distinguish all distributions in a neighborhood around the current hypothesis, without any preference for a particular "direction of perturbation."

We develop these techniques systematically in Section III and Section IV. We begin by recasting the lower bounds for the standard, centralized, setting in our chi-square fluctuation language in Section III before extending these notions to the distributed setting in Section IV. Finally, we evaluate our general lower bounds for the distribution learning and identity testing problems.

Our lower bounds are obtained as a function of a channel-dependent matrix $H(W)$, defined below.

Definition I.5. For any channel $W \in \mathcal{W}$, define $H(W)$ as the $k/2 \times k/2$ positive semidefinite matrix given by

$$H(W)_{i_1,i_2} := \sum_{y\in\mathcal{Y}} \frac{(W(y \mid 2i_1-1) - W(y \mid 2i_1))}{\sum_{x\in[k]} W(y \mid x)} \times \\ (W(y \mid 2i_2-1) - W(y \mid 2i_2)), \quad (2)$$

for $i_1, i_2 \in [k/2]$.

This matrix roughly captures the ability of the channel output to distinguish between even and odd inputs. Our bounds are in terms of the Frobenius norm $\|H(W)\|_F$ and the nuclear norm $\|H(W)\|_*$ of the matrix $H(W)$; see Section II for definitions. In effect, our results characterize the informativeness of a channel W for distributed inference in terms of these norms of $H(W)$, and our final bounds for sample complexity involve the maximum of these norms over W in $\mathcal{W}$. The smaller these norms are, the lower is the ability to distinguish consecutive inputs, and the better are our lower bounds (see Table I).

We summarize in Table I our sample complexity lower bounds for the (k, ε)-distribution learning and (k, ε)-identity testing problems for any general information constraints $\mathcal{W}$ for public- and private-coin protocols. The form here is only indicative; formal statements for results for general channels are available in Corollaries IV.13, IV.16 and IV.20 in Section IV and implications for specific $\mathcal{W}$ are given in Section V, with results on the communication-limited setting in Theorems V.2 to V.4 and the LDP setting in Theorems V.6 to V.8. The terms in each cell denotes the $\Omega(\cdot)$ lower bound obtained by our approach. The first row contains our lower bounds for a general family $\mathcal{W}$. We are specifying the bounds in terms of the multiplicative factor increase with respect to the central setting in which the sample complexity for learning and testing is k/ε^2 and $\sqrt{k}/\varepsilon^2$ respectively. We can instantiate the centralized setting by choosing $\mathcal{W}$ to contain the identity channel, which leads to $\|H(W)\|_* = k$, and $\|H(W)\|_F^2 = k/2$ and retrieves the optimal bounds in the centralized setting.

As a corollary of these general bounds, we obtain $\Omega(k^2/(\varepsilon^2 2^\ell))$ and $\Omega(k^2/(\varepsilon^2\rho^2))$ lower bounds for both

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2020.3028440, IEEE Transactions on Information Theory

TABLE I
CHI-SQUARE CONTRACTION LOWER BOUNDS FOR LOCAL INFORMATION-CONSTRAINED LEARNING AND TESTING.

	Learning		Testing	
	Public-Coin	Private-Coin	Public-Coin	Private-Coin
General	$\frac{k}{\varepsilon^2} \cdot \frac{k}{\max_{W\in\mathcal{W}}\lVert H(W))\rVert_*}$		$\frac{\sqrt{k}}{\varepsilon^2} \cdot \frac{\sqrt{k}}{\max_{W\in\mathcal{W}}\lVert H(W)\rVert_F}$	$\frac{\sqrt{k}}{\varepsilon^2} \cdot \frac{k}{\max_{W\in\overline{\mathcal{W}}}\lVert H(W)\rVert_*}$
Communication	$\frac{k}{\varepsilon^2} \cdot \frac{k}{2^\ell}$		$\frac{\sqrt{k}}{\varepsilon^2} \cdot \sqrt{\frac{k}{2^\ell}}$	$\frac{\sqrt{k}}{\varepsilon^2} \cdot \frac{k}{2^\ell}$
Privacy	$\frac{k}{\varepsilon^2} \cdot \frac{k}{\rho^2}$		$\frac{\sqrt{k}}{\varepsilon^2} \cdot \frac{\sqrt{k}}{\rho^2}$	$\frac{\sqrt{k}}{\varepsilon^2} \cdot \frac{k}{\rho^2}$

private- and public-coin distribution learning using $\mathcal{W}_\ell$ (the communication-limited setting) and $\mathcal{W}_\rho$ (the LDP setting), respectively. In particular, the multiplicative increase is $k/2^\ell$ and k/ρ^2, respectively, for the communication-limited and LDP settings compared to the centralized setting. As discussed later, these bounds have also been obtained in previous works and are known to be tight.

We note that for communication-constrained identity testing, we obtain $\Omega(k/(\varepsilon^2 2^{\ell/2}))$ and $\Omega(k^{3/2}/(\varepsilon^2 2^\ell))$ lower bounds for public- and private-coin protocols respectively. Both these bounds are tight, thus establishing the first separation in sample complexity using public- and private-coin protocols for a natural distributed goodness-of-fit problem. In particular, when $\ell = 1$ (one bit of communication per player) the sample complexity for public- and private-coin protocols is $\Theta(k)$ and $\Theta(k^{3/2})$ respectively. Similarly, for LDP identity testing, we obtain $\Omega(k/(\varepsilon^2\rho^2))$ and $\Omega(k^{3/2}/(\varepsilon^2\rho^2))$ lower bounds for public- and private-coin protocols, respectively, which are both tight, too, exhibiting the role of shared randomness in reducing the sample complexity.

As an interesting consequence, our results show that shared randomness does not help for distribution learning under communication or LDP constraints, in contrast to identity testing. Moreover, note that for (k, ε)-identity testing under general constraints, the factor $k/(\max_{W\in\mathcal{W}}\lVert H(W)\rVert_*)$ increase in the lower bound for private-coin protocols is the same as the increase for (k, ε)-distribution learning under information constraints; but the corresponding factor increase for identity testing using public-coin protocols is only $\sqrt{k}/(\max_{W\in\mathcal{W}}\lVert H(W)\rVert_F)$, which in general can be much smaller.

In the subsequent papers in this series ([3], [1]), we present public-coin and private-coin protocols to match the bounds in the communication-limited and LDP settings, respectively, thereby establishing the optimality of these lower bounds.

C. Prior and related work

The statistical tasks of distribution learning and identity testing considered in this work have a rich history. The former requires no special techniques other than those used in parametric estimation problems with finite-dimensional parameter spaces, which are standard textbook material. The identity testing problem is the same as the classic goodness-of-fit problem. The latter goes beyond the discrete setting considered here, but often starts with a quantization to a uniform reference distribution (see [35], [39]). The focus in this line of research has always been on the relation of the performance to the support size (*cf.* [39]), with particular interest on the large-support and small-sample case where the usual normal approximations of statistics do not apply (*cf.* [40], [8]). Closer to our setting, Paninski [41] (see, also, [50]) established the sample complexity of uniformity testing, showing that it is sublinear in k and equal to $\Theta(\sqrt{k}/\varepsilon^2)$. As mentioned earlier, in this work we are following this sample complexity framework that has received attention in recent years. We refer the reader to surveys [18], [43], [15], [7] for a comprehensive review of recent results on discrete distribution learning and testing.

Distributed inference problems, too, have been studied extensively, although for the asymptotic, large-sample case and for simpler hypothesis classes. There are several threads here. Starting with [49], decentralized detection has received attention in the control and signal processing literature, with main focus on information structure, likelihood ratio tests and combining local decisions for global inference. In a parallel thread, distributed statistical inference under communication constraints was initially studied in the information theory community [5], [30], [31], with the objective to characterize the asymptotic error exponents as a function of the communication rate. Recent results in this area have focused on more complicated communication models [54], [53] and, more recently, on the minimum communication requirements for large sample sizes [44], [6].

Our focus is different from that of the works above. In our setting, independent samples are not available at one place, but instead information constraints are placed on individual samples. This is along the line of recent work on distributed mean estimation under communication constraints [58], [28], [45], [14], [55], although some of these works consider more general communication models than what we allow. The distribution estimation problem under communication constraints has been studied in [20],[6] and, subsequent to an earlier version of this work, the distribution testing problem has been considered in [19]. However, in these two papers the authors consider a blackboard model of communication and strive to minimize the total number of bits communicated, without placing any restriction on the number of bits per sample. A more closely related variant of the distribution testing problem is studied in [26] where players observe multiple samples and communicate their local test results to the central referee who is required to use simple aggregation rules such as

[6] To the best of our knowledge, only an extended abstract with the result statements is available, and a full version including the proofs is yet to appear.

AND. Interestingly, such setups have received a lot of attention in the sensor network literature where a fusion center combines local decisions using simple rules such as majority; see [51] for an early review.

Closest to our work and independent of it is [32], which studies the (k, ε)-distribution learning problem using ℓ bits of communication per sample. It was shown that the sample complexity for this problem is $\Theta(k^2/(\varepsilon^2 2^\ell))$. This paper in turn uses a general lower bound from [33], [34], which yields lower bounds for distributed parametric estimation under suitable smoothness conditions. For this special case, our general approach reduces to a similar procedure as [34], which was obtained independently of our work.

Distribution learning under LDP constraints has been studied in [21], [37], [56], [4], [52], all providing sample-optimal schemes with different merits. Our lower bound when specialized for this setting coincides with the one derived in [21].

In spite of this large body of literature closely related to our work, there are two distinguishing features of our approach. First, the methods for deriving lower bounds under local information constraints in all these works, while leading to tight bounds for distribution learning, do not extend to identity testing. In fact, our *decoupled chi-square fluctuation* bound fills this gap in the literature. We remark that distributed uniformity testing under LDP constraints has been studied recently in [46], however the lower bounds derived there are significantly weaker than what we obtain. Second, our approach allows us to prove a separation between the performances of public-coin and private-coin protocols. This qualitative lesson – namely that shared public randomness reduces the sample complexity – is in contrast to the prescription of [49] which showed that shared randomness does not help in distributed testing when the underlying problem is that of simple hypothesis testing.[7]

We observe that the unifying treatment based on chi-square distance is reminiscent of the lower bounds for learning under statistical queries (SQ) derived in [25], [24], [47]. On the one hand, the connection between these two problems can be expected based on the relation between LDP and SQ learning established in [38]. On the other hand, this line of work only characterizes sample complexity up to polynomial factors. In particular, it does not lead to lower bounds we obtain using our decoupled chi-square fluctuation bounds.

We close with a pointer to an interesting connection to the capacity of an arbitrary varying channel (AVC). At a high level, our minmax lower bound considers the worst perturbation for the best channel. This is semantically dual to the expression for capacity of an AVC with shared randomness, where the capacity is determined by the maxmin mutual information, with maximum over input distributions and minimum over channels (*cf.* [17]).

D. Organization

We specify our notation in Section II and recall some basic inequalities needed for our analysis. This is followed by a review of the existing lower bounds for sample complexity of distribution learning and identity testing in Section III. In doing so, we introduce the notions of chi-square fluctuation which will be central to our work, and cast existing lower bounds under our general formulation. In Section IV, we generalize these notions to capture the information-constrained setting. Further, we apply our general approach to distribution learning and identity testing in the information-constrained setting. Then, in Section V, we instantiate these results to the settings of communication-limited and LDP inference and obtain our order-optimal bounds for testing and learning under these constraints. We conclude with pointers to schemes matching our lower bounds which will be reported in the subsequent papers in this series.

II. Notation and preliminaries

Throughout this paper, we denote by $\log_2$ the logarithm to the base 2 and by $\log$ the natural logarithm. We use standard asymptotic notation $O(\cdot)$, $\Omega(\cdot)$, and $\Theta(\cdot)$ for complexity orders.[8]

Let $[k]$ be the set of integers $\{1, 2, \ldots, k\}$. Given a fixed (and known) discrete domain $\mathcal{X}$ of cardinality $|\mathcal{X}| = k$, we write Δ_k for the set of probability distributions over $\mathcal{X}$, *i.e.*,

$$\Delta_k = \{ \, \mathbf{p} \colon [k] \to [0,1] \, : \, \|\mathbf{p}\|_1 = 1 \, \} \, .$$

For a discrete set $\mathcal{X}$, we denote by $\mathbf{u}_{\mathcal{X}}$ the uniform distribution on $\mathcal{X}$ and will omit the subscript when the domain is clear from context.

The *total variation distance* between two probability distributions $\mathbf{p}, \mathbf{q} \in \Delta_k$ is defined as

$$\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{q}) \coloneqq \sup_{S \subseteq \mathcal{X}} \left(\mathbf{p}(S) - \mathbf{q}(S) \right) = \frac{1}{2} \sum_{x \in \mathcal{X}} |\mathbf{p}(x) - \mathbf{q}(x)| \, ,$$

namely, $\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{q})$ is equal to half of the ℓ_1 distance of $\mathbf{p}$ and $\mathbf{q}$. In addition to total variation distance, we will extensively rely on the chi-square distance $\mathrm{d}_{\chi^2}(\mathbf{p}, \mathbf{q})$ and Kullback–Leibler (KL) divergence $D(\mathbf{p}\|\mathbf{q})$ between distributions $\mathbf{p}, \mathbf{q} \in \Delta_k$, defined as

$$\mathrm{d}_{\chi^2}(\mathbf{p}, \mathbf{q}) \coloneqq \sum_{x \in \mathcal{X}} \frac{(\mathbf{p}(x) - \mathbf{q}(x))^2}{\mathbf{q}(x)}, \text{ and}$$
$$D(\mathbf{p}\|\mathbf{q}) \coloneqq \sum_{x \in \mathcal{X}} \mathbf{p}(x) \log \frac{\mathbf{p}(x)}{\mathbf{q}(x)} .$$

Using concavity of logarithms and Pinsker's inequality, we can relate these two quantities to total variation distance as follows:

$$2 \cdot \mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{q})^2 \le D(\mathbf{p}\|\mathbf{q}) \le \mathrm{d}_{\chi^2}(\mathbf{p}, \mathbf{q}) \, . \tag{3}$$

In our results, we will rely on the following norms for matrices. Given a real-valued matrix $A = (a_{ij})_{(i,j) \in [m] \times [n]}$

[7] As discussed earlier, identity testing is a composite hypothesis testing problem with null hypothesis $\mathbf{q}$ and alternative comprising all distributions $\mathbf{p}$ that are ε-far from $\mathbf{q}$ in total variation distance.

[8] Namely, for two non-negative sequences $(a_n)_n$ and $(b_n)_n$, we write $a_n = O(b_n)$ (resp., $a_n = \Omega(b_n)$) if there exist $C > 0$ and $N \ge 0$ such that $a_n \le C b_n$ (resp., $a_n \ge C b_n$) for all $n \ge N$. Further, we write $a_n = \Theta(b_n)$ when both $a_n = O(b_n)$ and $a_n = \Omega(b_n)$ hold.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2020.3028440, IEEE Transactions on Information Theory

with singular values $(\sigma_k)_{1\leq k\leq m\wedge n}$, the *Frobenius norm* (or *Schatten 2-norm*) of A is given by

$$\|A\|_F = \left(\sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}^2\right)^{1/2} = \left(\sum_{k=1}^{m\wedge n}\sigma_k^2\right)^{1/2} = \sqrt{\operatorname{Tr} A^T A}.$$

Similarly, the *nuclear norm* (also known as *trace* or *Schatten 1-norm*) of A is defined as

$$\|A\|_* := \sum_{k=1}^{m\wedge n}\sigma_k = \operatorname{Tr}\sqrt{A^T A},$$

where $\sqrt{A^T A}$ is the (principal) square root of the positive semi-definite matrix $A^T A$. For any A, the Frobenius and nuclear norms satisfy the following inequality

$$\|A\|_F \leq \|A\|_* \leq \sqrt{\operatorname{rank} A}\cdot\|A\|_F\,, \tag{4}$$

which can be seen to follow, for instance, from an ℓ_1/ℓ_2 inequality and Cauchy–Schwarz inequality. Finally, the *spectral radius* of complex square matrix $A \in \mathbb{C}^{n\times n}$ with eigenvalues $\lambda_1,\dots,\lambda_n$, is defined as $\rho(A) := \max_{1\leq i\leq n}|\lambda_i|$.

III. Perturbed Families, Chi-square Fluctuations, and Centralized Lower Bounds

To build basic heuristics, we first revisit the derivation of lower bounds for sample complexity of (k,ε)-distribution learning and (k,ε)-identity testing in the centralized setting. As mentioned previously, for the latter it suffices to derive a lower bound for (k,ε)-uniformity testing. For brevity, we will sometimes refer to distribution learning as learning and identity testing as testing. We review both proofs in a unifying framework which we will extend to our information-constrained setting in the next section.[9]

Lower bounds for both learning and testing can be derived from a local view of the geometry of product distributions around the uniform distribution. Let $\mathbf{u}$ be the uniform distribution on $[k]$ and $\mathbf{u}^n$ be the n-fold product distributions over $[k]^n$, denoting the distribution of n independent and identically distributed (i.i.d.) draws from $\mathbf{u}$. A typical lower bound proof involves designing an appropriate family of distributions close to $\mathbf{u}$ such that the underlying problem is information-theoretically difficult to solve even for this smaller family of distributions. We call such a family a *perturbed family* and define it next.

Definition III.1 (Perturbed Family)**.** For $0<\varepsilon<1$ and a given k-ary distribution $\mathbf{q}$, an *ε-perturbed family around* $\mathbf{q}$ is a finite collection $\mathcal{P}$ of distributions such that, for all $\mathbf{p}\in\mathcal{P}$, $\mathrm{d}_{\mathrm{TV}}(\mathbf{p},\mathbf{q})\geq\varepsilon$.

When ε is clear from context, we simply use the phrase *perturbed family around* $\mathbf{q}$.

As we shall see below, the bottleneck for learning distributions, which is a parametric estimation problem, arises from the difficulty in solving a multiple hypothesis testing problem with hypotheses given by the elements of a perturbed family around $\mathbf{u}$. Using Fano's inequality, we can show that this difficulty is captured by the average KL divergence between $\mathbf{u}$ and the elements of the perturbed family. In fact, for a unified treatment, we shall simply bound KL divergences by chi-square distances. This motivates the following definition.

Definition III.2 (Chi-square Fluctuation)**.** Given a perturbed family $\mathcal{P}$ around $\mathbf{q}$, the *chi-square fluctuation* of $\mathcal{P}$ is given by

$$\chi^2(\mathcal{P}) := \frac{1}{|\mathcal{P}|}\sum_{\mathbf{p}\in\mathcal{P}}\mathrm{d}_{\chi^2}(\mathbf{p},\mathbf{q})\,,$$

the average chi-square distance of the distributions in $\mathcal{P}$ from $\mathbf{q}$.

From Eq. (3), it follows that the average KL divergence mentioned above is upper bounded by the chi-square fluctuation of $\mathcal{P}$, which will be used to obtain a lower bound for the sample complexity of learning in the next section.

On the other hand, identity testing is a composite hypothesis testing problem, and a lower bound on the testing problem can be obtained using Le Cam's two-point method. Specifically, for an ε-perturbed family $\mathcal{P}$ around $\mathbf{u}$, consider a binary hypothesis testing between the following two distributions over $[k]^n$: $\mathbf{u}^n$ and the uniform mixture $\frac{1}{|\mathcal{P}|}\sum_{\mathbf{p}\in\mathcal{P}}\mathbf{p}^n$ of the n-fold product of elements of the perturbed family. Since each element of $\mathcal{P}$ is at a total variation distance ε from $\mathbf{u}$, it can be seen easily that any test for identity testing will yield a test for this binary hypothesis testing problem as well, with the same probability of error. In particular, a lower bound on the value of n required to solve this problem gives a lower bound on identity testing. Thus, our goal is to capture the difficulty of this binary hypothesis testing problem. This difficulty is captured by the total variation distance between these two distributions on $[k]^n$, for which a simple upper bound is $\sqrt{n}\cdot\sqrt{\chi^2(\mathcal{P})}$. However, this bound turns out to be suboptimal.

Instead, an alternative bound derived using a recipe of Ingster [36] (the form here is from [42]) was shown to be tight in [41]. To understand this bound, we let the perturbed family $\mathcal{P}$ be parameterized by a discrete set $\mathcal{Z}$, *i.e.*, for each $z\in\mathcal{Z}$, there is a $\mathbf{p}_z\in\mathcal{P}$. We will specify the choice of $\mathcal{Z}$ shortly. Denoting by $\delta_z\in\mathbb{R}^k$ the normalized perturbation with entries given by

$$\delta_z(x) := \frac{\mathbf{p}_z(x)-\mathbf{q}(x)}{\mathbf{q}(x)}, \qquad x\in[k]\,, \tag{5}$$

let $\|\delta_z\|_2^2 := \mathbb{E}_{X\sim\mathbf{q}}\left[\delta_z(X)^2\right] = \mathrm{d}_{\chi^2}(\mathbf{p}_z,\mathbf{q})$ be the second moment of the random variable $\delta_z(X)$ (for X drawn from $\mathbf{q}$). For Z uniform over $\mathcal{P}$, note that we can re-express $\chi^2(\mathcal{P})$ as

$$\chi^2(\mathcal{P}) = \mathbb{E}_Z\left[\mathrm{d}_{\chi^2}(\mathbf{p}_Z,\mathbf{q})\right] = \mathbb{E}_Z\left[\|\delta_Z\|_2^2\right].$$

Following [36], [42], we can bound the total variation distance mentioned above by a quantity we term the *decoupled chi-square fluctuation* of $\mathcal{P}$, instead of the weaker bound $\sqrt{n\cdot\chi^2(\mathcal{P})}$. This quantity appears by using the decoupling expression $\delta_Z^2 = \delta_Z\delta_{Z'}$, as will be seen below, and is defined next.

Definition III.3 (Decoupled Chi-square Fluctuation)**.** Given a k-ary distribution $\mathbf{p}$ and a perturbed family $\mathcal{P} =$

[9] Although we restrict ourselves to the discrete distributions over $[k]$ here, the framework extends to more general parametric families.

$\{\, \mathbf{p}_z \,:\, z \in \mathcal{Z} \,\}$ around $\mathbf{q}$, the n-fold *decoupled chi-square fluctuation* of $\mathcal{P}$ is given by

$$\chi^{(2)}(\mathcal{P}^n) := \log \mathbb{E}_{ZZ'}[\exp\left(n \cdot \langle \delta_Z, \delta_{Z'} \rangle\right)],$$

where

$$\langle \delta_z, \delta_{z'} \rangle := \mathbb{E}_{X\sim\mathbf{q}}[\delta_z(X)\delta_{z'}(X)]$$

denotes the correlation inner product with respect to $\mathbf{q}$, and the outer expectation is over Z and Z', which are independent and uniformly distributed uniformly over $\mathcal{Z}$.

While the quantities $n \cdot \chi^2(\mathcal{P})$ and $\chi^{(2)}(\mathcal{P}^n)$ are implicit in prior work, the abstraction here allows us to have a clear geometric view and lends itself to the more general local information-constrained setting. For completeness, we review the proofs of existing lower bounds using our chi-square fluctuation terminology.

A specific perturbed family used in our lower bounds. In the sections below, we will present the proofs of lower bounds for sample complexity of learning and testing using a specific perturbed family $\mathcal{P}$ and bring out the role of $\chi^2(\mathcal{P})$ and $\chi^{(2)}(\mathcal{P}^n)$ in these bounds. In particular, both bounds will be derived using the ε-perturbed family around $\mathbf{u}$ due to [41], consisting of distributions parameterized by $z \in \mathcal{Z} = \{-1,+1\}^{k/2}$ and comprising distributions $\mathbf{p}_z \in \Delta_k$ given by

$$\mathbf{p}_z = \frac{1}{k}(1+2\varepsilon z_1, 1-2\varepsilon z_1, \ldots, 1+2\varepsilon z_{\frac{k}{2}}, 1-2\varepsilon z_{\frac{k}{2}}), \quad (6)$$

for $z \in \{-1,+1\}^{\frac{k}{2}}$. The normalized perturbations for this perturbed family are given by

$$\delta_z(x) = \begin{cases} 2\varepsilon z_i, & x = 2i-1, \\ -2\varepsilon z_i, & x = 2i, \end{cases} \qquad i \in [k/2]. \quad (7)$$

Note that for any $x \in [k]$, $|\delta_z(x)| = 2\varepsilon$, and the chi-square fluctuation is given by

$$\chi^2(\mathcal{P}) = 4\varepsilon^2. \quad (8)$$

A. Chi-square fluctuation and the centralized learning lower bound

As a starting application, we recover the following well-known result for the sample complexity of distribution learning, using the notion of chi-squared fluctuation.

Theorem III.4. *For (k,ε)-distribution learning, if there exists an (n,ε)-estimator, then $n = \Omega(k/\varepsilon^2)$.*

To establish this bound, we consider the multiple hypotheses testing problem where the hypotheses are $\mathbf{p}_z$, $z \in \{-1,+1\}^{k/2}$, given in Eq. (6). Specifically, denote by Z the random variable distributed uniformly on $\mathcal{Z} = \{-1,+1\}^{k/2}$ and by X^n the random variable with distribution $\mathbf{p}_Z^n$ given Z. We can relate the accuracy of a probability estimate to the probability of error for the multiple hypothesis testing problem with hypotheses given by $\mathbf{p}_z$ using the standard Fano's method (*cf.* [57]). In particular, we can use a probability estimate $\hat{\mathbf{p}}$ to solve the hypothesis testing problem by returning as $\hat{Z}$ a $z \in \{-1,1\}^{k/2}$ that minimizes $\mathrm{d_{TV}}(\mathbf{p}_{\hat{z}}, \hat{\mathbf{p}})$. The difficulty here is that the total variation distance $\mathrm{d_{TV}}(\mathbf{p}_z, \mathbf{p}_{z'})$ may not be $\Omega(\varepsilon)$, and therefore, an (n,ε)-estimator may not return the correct hypothesis.

One way of circumventing this difficulty is to restrict to a perturbed family where pairwise-distances are $\Omega(\varepsilon)$. Note that for the perturbed family in Eq. (6)

$$\mathrm{d_{TV}}(\mathbf{p}_z, \mathbf{p}_{z'}) = \mathrm{dist}(z,z') \cdot \frac{2\varepsilon}{k}\,, \quad (9)$$

where $\mathrm{dist}(z,z')$ is the Hamming distance. This simple observation allows us to convert the problem of constructing a "packing" in total variation distance to that of constructing a packing in Hamming space. Indeed, a standard Gilbert–Varshamov construction of packing in Hamming space yields a subset $\mathcal{Z}_0 \subset \{-1,+1\}^{k/2}$ with $|\mathcal{Z}_0| \geq 2^{ck}$ such that $\mathrm{dist}(z,z') = \Omega(k)$ for every z, z' in $\mathcal{Z}_0$. Using Fano's inequality to bound the probability of error for this new perturbed family, we can relate the sample complexity of learning to $I(Z \wedge X^n)$. However, when later extending our bounds to the information-constrained setting, this construction would create difficulties in bounding $I(Z \wedge X^n)$ for public-coin protocols. We avoid this complication by relying instead on a slightly modified form of the classic Fano's argument from [22]; this form of Fano's argument was used in [34] as well to obtain a lower bound for the sample complexity of learning under communication constraints.

Specifically, in view of Eq. (9), it is easy to see that for an estimate $\hat{\mathbf{p}}$ such that $\mathbf{p}^n(\mathrm{d_{TV}}(\mathbf{p},\hat{\mathbf{p}}) > \varepsilon/3) \leq 1/12$ for all $\mathbf{p}$, we must have

$$\Pr\left[\mathrm{dist}\left(Z,\hat{Z}\right) > \frac{k}{6}\right] \leq \frac{1}{12}\,.$$

On the other hand, the proof of Fano's inequality in [16] can be extended easily to obtain (see, also, [22])

$$\Pr\left[\mathrm{dist}\left(Z,\hat{Z}\right) > \frac{k}{6}\right] \geq 1 - \frac{I(Z \wedge Y^n)+1}{\log_2|\mathcal{Z}| - \log_2 B_{k/6}}\,, \quad (10)$$

where B_t denotes the cardinality of Hamming ball of radius t. Noting that

$$\log_2 B_{k/6} \leq \frac{k}{2}\cdot h\left(\frac{1}{3}\right), \quad (11)$$

and combining the bounds above, if an (n,ε)-estimator exists, then we must have

$$I(Z \wedge X^n) + 1 \geq \frac{11k}{12\cdot 2\cdot(1-h(1/3))} \geq \frac{k}{30}\,. \quad (12)$$

Therefore, to obtain a lower bound for sample complexity it suffices to bound $I(Z \wedge X^n)$ from above. It is in this part that we bring in the role of chi-square fluctuations. Note that for a given value of Z, $X^n \sim \mathbf{p}_Z^n$. Therefore, we have

$$\begin{aligned} I(Z \wedge X^n) &= \mathbb{E}_Z[D(\mathbf{p}_Z^n \| \mathbb{E}_Z[\mathbf{p}_Z^n])] \\ &= \mathbb{E}_Z[D(\mathbf{p}_Z^n \| \mathbf{u}^n)] - D(\mathbb{E}_Z[\mathbf{p}_Z^n] \| \mathbf{u}^n) \\ &\leq \mathbb{E}_Z[D(\mathbf{p}_Z^n \| \mathbf{u}^n)] \\ &= n\,\mathbb{E}_Z[D(\mathbf{p}_Z \| \mathbf{u})] \\ &\leq n\,\mathbb{E}_Z\left[\mathrm{d}_{\chi^2}(\mathbf{p}_Z, \mathbf{u})\right] \\ &= n\cdot\chi^2(\mathcal{P})\,, \end{aligned} \quad (13)$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2020.3028440, IEEE Transactions on Information Theory

where the last inequality uses $D(\mathbf{p}\|\mathbf{q}) \leq \mathrm{d}_{\chi^2}(\mathbf{p}, \mathbf{q})$. Combining Eq. (12) and Eq. (13), we obtain that $n = \Omega\big(k/\chi^2(\mathcal{P})\big)$, which along with Eq. (8) proves Theorem III.4.

In fact, the argument above is valid for any perturbation (*i.e.*, around any nominal distribution $\mathbf{q}$) with the desired pairwise minimum total variation distance, namely any perturbed family satisfying an appropriate replacement for Eq. (11). In particular, it suffices to impose the following condition:

$$\max_{z\in\mathcal{Z}} \left| \left\{ z' \in \mathcal{Z} : \mathrm{d}_{\mathrm{TV}}(\mathbf{p}_z, \mathbf{p}_{z'}) \leq \frac{\varepsilon}{3} \right\} \right| \leq C_\varepsilon \,. \qquad (14)$$

Proceeding as in the proof of Theorem III.4 above, noting that $|\mathcal{Z}| = |\mathcal{P}|$ and replacing $B_{k/6}$ with the right-side of Eq. (14), we obtain the following.

Lemma III.5. *For $0 < \varepsilon < 1$ and a k-ary distribution $\mathbf{q}$, let $\mathcal{P}$ be an ε-perturbed family around $\mathbf{q}$ satisfying Eq. (14). Then, the sample complexity of $(k, \varepsilon/3)$-distribution learning must be at least*

$$\Omega\left(\frac{\log|\mathcal{P}| - \log C_\varepsilon}{\chi^2(\mathcal{P})}\right).$$

As a sanity check, when $\mathcal{P}$ is set to be Paninski's perturbed family given in Eq. (6), we have $|\mathcal{P}| = 2^{k/2}$, $C_\varepsilon = 2^{(1-h(1/3))k/2}$, and $\chi^2(\mathcal{P}) = 4\varepsilon^2$ from Eq. (8), recovering the $\Omega(k/\varepsilon^2)$ lower bound for sample complexity of distribution learning derived above.

B. Decoupled chi-square fluctuation and the centralized testing lower bound

In this section, we provide an alternative proof for the following result on uniformity testing, using the notion of decoupled chi-square fluctuation.

Theorem III.6 ([41])**.** *If there exists an (n, ε)-test for (k, ε)-uniformity testing, then $n = \Omega(\sqrt{k}/\varepsilon^2)$.*

Unlike for distribution learning, the binary hypothesis testing problems obtained from the pairs of distributions in the perturbed family $\mathcal{P}$ do not yield the desired dependence of sample complexity on k. As pointed earlier, the bottleneck in this case emerges by looking at a binary hypothesis testing problem between $\mathbf{u}^n$ and a uniform mixture of distributions from an ε-perturbed family. Specifically, by Le Cam's method, any test for uniformity also constitutes a test for $\mathbf{u}^n$ versus $\mathbb{E}[\mathbf{p}_Z^n] = \frac{1}{2^{k/2}} \sum_{z\in\{-1,+1\}^{k/2}} \mathbf{p}_z^n$, the uniform mixture of n-fold product distributions over the perturbed family from Eq. (6). Thus, another aspect of the geometry around $\mathbf{u}^n$ that enters our consideration is the distance between $\mathbf{u}^n$ and $\mathbb{E}[\mathbf{p}_Z^n]$. In particular, the key would be to prove a lower bound on the value of n to ensure that $\mathrm{d}_{\mathrm{TV}}(\mathbf{u}^n, \mathbb{E}[\mathbf{p}_Z^n])$ is at least $1/12$.

As a straightforward attempt towards bounding this quantity, by Pinsker's inequality and convexity of KL divergence, we get

$$\begin{aligned} \mathrm{d}_{\mathrm{TV}}(\mathbb{E}[\mathbf{p}_Z^n]\,, \mathbf{u}^n) &\leq \sqrt{\frac{1}{2} D(\mathbb{E}[\mathbf{p}_Z^n]\,\|\mathbf{u}^n)} \\ &\leq \sqrt{\frac{1}{2}\mathbb{E}[D(\mathbf{p}_Z^n\|\mathbf{u}^n)]} \\ &= \sqrt{\frac{n}{2}\mathbb{E}[D(\mathbf{p}_Z\|\mathbf{u})]} \\ &\leq \sqrt{\frac{n}{2}\cdot\chi^2(\mathcal{P})} \\ &= \sqrt{2n\varepsilon^2}, \end{aligned} \qquad (15)$$

where the last identity is by Eq. (8). Thus, this upper bound on the distance between $\mathbf{u}^n$ and $\mathbb{E}[\mathbf{p}_Z^n]$ in terms of the chi-square fluctuation only yields a sample complexity lower bound of $\Omega(1/\varepsilon^2)$, much lower than the $\Omega(\sqrt{k}/\varepsilon^2)$ bound that we strive for.

Interestingly, we obtain the desired improvement in the lower bound by taking recourse to the decoupled chi-square fluctuation $\chi^{(2)}(\mathcal{P}^n)$. Specifically, we have the following result.

Lemma III.7. *For $0 < \varepsilon < 1$ and a k-ary distribution $\mathbf{q}$, let $\mathcal{P}$ be an ε-perturbed family around $\mathbf{q}$. Then, the sample complexity $n = n(k, \varepsilon)$ for (k, ε)-identity testing with reference distribution $\mathbf{q}$ must satisfy*

$$\chi^{(2)}(\mathcal{P}^n) \geq \frac{1}{12}\,.$$

The proof is this result relies on the so-called Ingster's method to bound the chi-square distance between a mixture of product distributions and a product distribution. The complete proof is provided in Appendix B, and makes critical use of the following result from [42], which, too, is proved in the appendix.

Lemma III.8. *Consider a random variable θ such that for each $\theta = \vartheta$ the distribution Q_ϑ^n is defined as $Q_{1,\vartheta} \times \cdots \times Q_{n,\vartheta}$. Further, let $P^n = P_1 \times \cdots \times P_n$ be a fixed product distribution. Then,*

$$\chi^2(\mathbb{E}_\theta[Q_\theta^n], P^n) = \mathbb{E}_{\theta\theta'}\left[\prod_{j=1}^n (1 + H_j(\theta, \theta'))\right] - 1,$$

where θ' is an independent copy of θ, and with $\delta_j^\vartheta(X_j) = (Q_{j,\vartheta}(X_j) - P_j(X_j))/P_j(X_j)$,

$$H_j(\vartheta, \vartheta') := \left\langle \delta_j^\vartheta, \delta_j^{\vartheta'} \right\rangle = \mathbb{E}\left[\delta_j^\vartheta(X_j)\delta_j^{\vartheta'}(X_j)\right],$$

where the expectation is over X_j distributed according to P_j.

Proof of Theorem III.6. . We use Lemma III.7 to complete the proof. Specifically, we apply the lemma to Paninski's perturbed family given in Eq. (6). By Eq. (7),

$$\langle \delta_Z, \delta_{Z'} \rangle = \frac{8\varepsilon^2}{k} \sum_{i=1}^{\frac{k}{2}} Z_i Z_i' = \frac{8\varepsilon^2}{k} \sum_{i=1}^{\frac{k}{2}} V_i\,,$$

where $V_i := Z_i Z_i'$. Since Z, Z' are independently and uniformly distributed over $\{-1, +1\}^{k/2}$, $V_1, \ldots, V_{k/2}$ are independent and distributed uniformly over $\{-1, +1\}$. Therefore, we can bound the decoupled chi-square fluctuation using Hoeffding's Lemma (*cf.* [13]) as

$$\chi^{(2)}(\mathcal{P}^n) = \log\mathbb{E}\left[e^{\frac{8n\varepsilon^2}{k}\sum_{i=1}^{\frac{k}{2}} V_i}\right] \leq \frac{16n^2\varepsilon^4}{k}. \qquad (16)$$

Thus, Lemma III.7 implies that $\Omega(\sqrt{k}/\varepsilon^2)$ samples are needed for testing (in particular, for uniformity testing). □

In closing, we summarize the geometry captured by the bounds derived in this section in Fig. 2. This geometry is a local view in the neighborhood of the uniform distribution obtained using the perturbed family $\mathcal{P}$ in Eq. (6). Each $\mathbf{p}_z$ is at a total variation distance ε from $\mathbf{u}$. The mixture distribution we use is obtained by uniformly choosing the perturbation δ_z over $z \in \{-1,+1\}^{k/2}$.

The chi-square fluctuation of $\mathcal{P}$ is $O(n\varepsilon^2)$ whereby the average total variation distance to $\mathbf{u}^n$ is $O(\sqrt{n}\varepsilon)$. On the other hand, the decoupled chi-square fluctuation of $\mathcal{P}$ is $O(n^2\varepsilon^4/k)$ and thus the total variation distance of the mixture of $\mathbf{p}_z^n$ to $\mathbf{u}^n$ is $O(n\varepsilon^2/\sqrt{k})$. Note that for $n \ll k/\varepsilon^2$, the total variation distance between the mixture $\mathbb{E}[\mathbf{p}_Z^n]$ and $\mathbf{u}^n$ is much smaller than the average total variation distance.

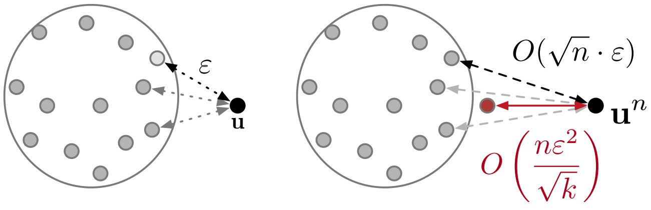
ε
$\mathbf{u}$
$O(\sqrt{n}\cdot\varepsilon)$
$\mathbf{u}^n$
$O\left(\frac{n\varepsilon^2}{\sqrt{k}}\right)$

Fig. 2. The figure depicts the distances in the probability simplex on the left and the n-fold distributions on the right. The mixture distribution $\mathbb{E}[\mathbf{p}_Z^n]$ is marked in red.

IV. Results: The chi-square contraction bounds with information constraints

We now extend our notions of chi-square fluctuation and decoupled chi-square fluctuation to the information-constrained setting. We follow the same notation as the previous section. Recall that in the information-constrained setting each player sends information about its sample by choosing a channel from a family $\mathcal{W}$ to communicate to the central referee $\mathcal{R}$. The perturbed family will now induce a distribution on the outputs of the chosen channels $W_1, \ldots, W_n$. By data-processing inequalities these induced distributions will be *closer* to each other than the distributions of the perturbed family itself, and in particular the *induced chi-squared* fluctuations of the induced distributions will be smaller than that of the input distributions. We term this reduction in fluctuations *chi-squared contraction*. We will relate the difficulty of the learning and testing problems to the chi-squared fluctuations on the *induced perturbed family*. Combining these two steps we obtain lower bounds on the sample complexity of the learning and testing problems with respect to the centralized setting in terms of the chi-squared contractions. In other words, the difficulty of inference gets amplified by information constraints since the induced distributions are closer than the original ones and the chi-square fluctuation decreases.

We extend the general results from previous section to general information constraints $\mathcal{W}$ under both private-coin and public-coin protocols. In particular, we will extend Lemma III.5 and Lemma III.7 to the information-constrained setting. We then specialize them for discrete distribution learning and testing by bounding the chi-squared fluctuations of the induced perturbed family corresponding to Paninski's perturbed family of Eq. (6), for a given $\mathcal{W}$; and use them to obtain tight lower bounds for testing and learning under communication and privacy constraints for both private- and public-coin protocols. Underlying these bounds is a precise characterization of the *contraction in chi-square fluctuation* owing to information constraints. One can view this as a bound for the minmax chi-square fluctuation for an induced perturbed family, where the minimum is taken over perturbed families and the maximum over all channels in $\mathcal{W}$. We will see that for public-coin protocols, the bottleneck is indeed captured by this *minmax chi-square fluctuation*.

On the other hand, for private-coin protocols the bottleneck can be tightened further by designing a perturbation specifically for each choice of channels from $\mathcal{W}$. In other words, in this case we can use a bound for *maxmin chi-square fluctuation*. Another main result of this section, perhaps our most striking one, is a tight bound for this maxmin chi-square fluctuation for the aforementioned induced perturbed family. This bound turns out to be more stringent than the minmax chi-square fluctuation bound, and leads to the separation for private- and public-coin protocols for the cases $\mathcal{W} = \mathcal{W}_\ell$ and $\mathcal{W} = \mathcal{W}_\rho$ considered in the next section.

We begin by defining induced chi-squared fluctuations. Throughout we assume that the family of channels $\mathcal{W}$ consists of channels $W \colon \mathcal{X} \to \mathcal{Y}$ where the input alphabet is $\mathcal{X}$ and the output alphabet $\mathcal{Y}$ is finite. Recall from Eq. (1) that for an input distribution $\mathbf{p}$ over $\mathcal{X}$, the output distribution of channel W is $\mathbf{p}(y) := \sum_x \mathbf{p}(x)W_j(y \mid x) = \mathbb{E}_{\mathbf{p}}[W(y \mid X)]$.

Let $\mathcal{P}$ be a perturbed family of distributions over $\mathcal{X}$ parameterized as $\{\, \mathbf{p}_z : z \in \mathcal{Z} \,\}$. As outlined above, our extension involves the notions of an induced perturbed family and its chi-square fluctuations, which is simply the family of distributions induced at the output for input distributions $\mathbf{p}_z$; formal definition follows.

Definition IV.1. For a perturbed family $\mathcal{P}$ and channels $W^n = (W_1, \ldots, W_n) \in \mathcal{W}^n$, the *induced perturbed family* $\mathcal{P}^{W^n}$ comprises distributions $W^n\mathbf{p}_z^n$ over $\mathcal{Y}^n$ given by

$$W^n\mathbf{p}_z^n(y^n) = \prod_{i=1}^n W_i\mathbf{p}_z(y_i).$$

We now are able to define the notion of chi-square fluctuation from that of induced perturbed families. We first extend the corresponding notion of normalized perturbation with respect to the nominal distribution $\mathbf{q}$, defined in Eq. (5) as $\delta(x) := (\mathbf{p}(x) - \mathbf{q}(x))/\mathbf{q}(x)$. The *induced normalized perturbation* of $\mathbf{p}$ and $\mathbf{q}$ with respect to a channel W is

$$\begin{aligned}\delta^W(y) &:= \frac{W\mathbf{q}(y) - W\mathbf{p}(y)}{W\mathbf{q}(y)} \\ &= \sum_{x\in\mathcal{X}} \frac{(\mathbf{p}(x) - \mathbf{q}(x))W(y \mid x)}{W\mathbf{q}(y)} \\ &= \frac{\sum_x \mathbf{q}(x)W(y \mid x)\delta(x)}{\sum_{x'} \mathbf{q}(x')W(y \mid x')}.\end{aligned}$$

Thus, the normalized perturbation for the induced perturbed family $\mathcal{P} = \{\, \mathbf{p}_z :\ z \in \mathcal{Z} \,\}$ is given by

$$\delta_Z^W(y) = \frac{1}{W\mathbf{q}(y)} \cdot \mathbb{E}_{\mathbf{q}}[\delta_Z(X)W(y \mid X)]\,,\ y \in \mathcal{Y}$$

where recall that $\delta_z(X) = (\mathbf{p}_z(x) - \mathbf{q}(x))/\mathbf{q}(x)$. As before, for a channel $W \in \mathcal{W}$, let $\|\delta_z^W\|_2^2 := \mathbb{E}_{Y\sim W\mathbf{q}}\left[\delta_z^W(Y)^2\right] = \mathrm{d}_{\chi^2}(W\mathbf{p}_z, W\mathbf{q})$.

Remark IV.2. An important observation that will be used in our proofs later is that the random variable δ_Z^W can be obtained as a (W-dependent) linear transform of δ_Z.

We now extend in the natural way the notions of chi-square fluctuations from the centralized setting from Definitions III.2 and III.3 to the induced chi-squared fluctuations of $\mathcal{P}^{W^n}$.

Definition IV.3. Consider a perturbed family $\mathcal{P} = \{\mathbf{p}_z : z \in \mathcal{Z}\}$ around $\mathbf{q}$ and a family of channels $\mathcal{W}$. The *induced chi-square fluctuation* of $\mathcal{P}$ for $W \in \mathcal{W}$ is given by

$$\chi^2\left(W \mid \mathcal{P}\right) := \mathbb{E}_Z\left[\mathrm{d}_{\chi^2}(W\mathbf{p}_z, W\mathbf{q})\right] = \mathbb{E}_Z\left[\|\delta_Z^W\|_2^2\right],$$

where Z is distributed uniformly over $\mathcal{Z}$. The n-fold *induced decoupled chi-square fluctuation* of $\mathcal{P}$ for $W^n \in \mathcal{W}^n$ is given by

$$\chi^{(2)}\left(W^n \mid \mathcal{P}\right) := \log \mathbb{E}_{ZZ'}\left[\exp\left(\sum_{i=1}^{n} \left\langle \delta_Z^{W_i}, \delta_{Z'}^{W_i}\right\rangle\right)\right],$$

where $\left\langle \delta_z^W, \delta_{z'}^W \right\rangle = \mathbb{E}_{Y\sim W\mathbf{q}}\left[\delta_z^W(Y)\delta_{z'}^W(Y)\right]$.

We now make two relaxations to these notions. The first, to the definition of chi-squared fluctuations, to allow for more general Z; and the other to that of perturbed family, to weaken the requirement on distance from the nominal distribution. These relaxations will be helpful in investigating the role of shared randomness as is discussed later, in the paragraph before Definition IV.5.

Our definitions until now have computed fluctuations by using a uniform distribution on the perturbed family $\mathcal{P} = \{\mathbf{p}_z : z \in \mathcal{Z}\}$. As can be seen from the results of the previous section, this is not required and all the results above extend to any distribution over $\mathcal{Z}$. We can consider a distribution ζ for Z, which need not even be independent across the coordinates Z_is (when Z is over $\{-1,+1\}^{k/2}$). For brevity, we will denote chi-square fluctuations for $\mathcal{P}$ when the expectation is computed using ζ by $\chi^2\left(W \mid \mathcal{P}_\zeta\right)$ and $\chi^{(2)}\left(W^n \mid \mathcal{P}_\zeta\right)$; when ζ is uniform, we omit the subscript ζ in $\mathcal{P}$.

In our definition of ε-perturbed family, we required $\mathrm{d}_{\mathrm{TV}}(\mathbf{p}_z, \mathbf{q})$ to be bounded below by ε for each $z \in \mathcal{Z}$. This requirement is imposed in view of Eq. (29) where it leads to the upper bound on probability of error. However, a nearly identical result can be obtained even if we relax this requirement to hold only with large probability. This motivates the next definition.

Definition IV.4 (Almost ε-Perturbation)**.** Fix $0 < \varepsilon < 1$, a family of distributions $\mathcal{P} = \{\mathbf{p}_z, z \in \mathcal{Z}\}$, and a distribution ζ on $\mathcal{Z}$. The pair $\mathcal{P}_\zeta = (\mathcal{P}, \zeta)$ is an *almost ε-perturbation (around $\mathbf{q}$)* if for some $\alpha \geq 1/10$

$$\Pr[\mathrm{d}_{\mathrm{TV}}(\mathbf{p}_Z, \mathbf{q}) \geq \varepsilon] \geq \alpha,$$

where the probability is over $Z \sim \zeta$. We denote the set of all almost ε-perturbations by Υ_ε (We omit $\mathbf{q}$ from the notation as it will be clear from context).

The choice of $1/10$ in the definition above is used to match the probability of error requirement of $1/12$ in our PAC formulations given in Section I; see Eq. (31) and Footnote 12 for justification for these choices.

The flexibility offered by approximate perturbations is required to obtain our results for private-coin protocol; in particular, it will be used to show the separation between the performance of private- and public-coin protocols for testing. Recall that in Lemma III.7 we showed that the existence of an identity test in the centralized setting implies that the chi-squared fluctuation must be bounded away from zero. We will establish an analogous result in the distributed setting, by defining two quantities which are the minmax and maxmin evaluations of the decoupled chi-squared fluctuation from Definition IV.3. The two notions will be used to derive the lower bounds for information-constrained testing using public- and private-coin protocols, respectively.

Definition IV.5 (Minmax and Maxmin Chi-square Fluctuations)**.** For a family of channels $\mathcal{W}$, the (n,ε)*-minmax decoupled chi-square fluctuation for* $\mathcal{W}$ is given by

$$\overline{\chi}^{(2)}(\mathcal{W}^n, \varepsilon) := \inf_{\mathcal{P}_\zeta \in \Upsilon_\varepsilon} \sup_{W^n \in \mathcal{W}^n} \chi^{(2)}\left(W^n \mid \mathcal{P}_\zeta\right),$$

and the (n,ε)*-maxmin decoupled chi-square fluctuation for* $\mathcal{W}$ is given by

$$\underline{\chi}^{(2)}(\mathcal{W}^n, \varepsilon) := \sup_{W^n \in \mathcal{W}^n} \inf_{\mathcal{P}_\zeta \in \Upsilon_\varepsilon} \chi^{(2)}\left(W^n \mid \mathcal{P}_\zeta\right),$$

where the infimum is over all almost ε-perturbations $\mathcal{P}_\zeta$.

In Section IV-A, we extend the proofs of Lemma III.5 and Lemma III.7 to provide general results on learning and testing distributions under information constraints. Note that the desired extension to product distributions for Lemma IV.8 requires Lemma III.8 in its full generality (non-identical marginals for both P^n and the Q_θ^ns), in contrast to the earlier usage in the proof of Lemma III.7.

Further, we observe that when obtaining bounds for public-coin protocols we can restrict ourselves to a smaller family of channels than $\mathcal{W}$. The following notions are needed to state our results in full strength.

Definition IV.6. For a family of channels $\mathcal{W}$, denote by $\overline{\mathcal{W}}$ its convex hull, namely the set of channels $\overline{\mathcal{W}} = \{\, \theta W_1 + (1-\theta) W_2 :\ \theta \in [0,1], W_1, W_2 \in \mathcal{W} \,\}$. A *generator family* for $\mathcal{W}$, denoted $\mathcal{W}_0$, is a minimal subset of $\mathcal{W}$ whose convex hull is $\mathcal{W}$.

Note that the channels in $\mathcal{W}$ can be generated from and can generate, respectively, channels in $\mathcal{W}_0$ and $\overline{\mathcal{W}}$ using randomness.

A. General chi-square fluctuation bounds

The bounds presented in this section are obtained by relating the notions of chi-square fluctuation for $\mathcal{W}$ developed above to average distances in a neighborhood of the probability simplex.

We present our bounds for learning and testing problems, but the recipe extends to many other inference problems. In the next section, we provide specific evaluations of these bounds for the Paninski perturbed family from Eq. (6), and some of its variants, which are tailored for the discrete distribution inference problems of learning and testing.

We begin with our bound for learning, which is a generalization of Lemma III.5 to the information-constrained setting; the proof is provided in Appendix C.

Lemma IV.7 (Chi-square fluctuation bound for learning)**.** *For $0 < \varepsilon < 1$ and a k-ary distribution $\mathbf{q}$, let $\mathcal{P}$ be an ε-perturbed family around $\mathbf{q}$ satisfying Eq. (14). Then, the sample complexity of (k, ε)-distribution learning using $\mathcal{W}$ for public-coin protocols is at least*

$$\Omega\left(\frac{\log|\mathcal{P}| - \log C_\varepsilon}{\max_{W\in\mathcal{W}_0} \chi^2(W \mid \mathcal{P})}\right).$$

Note that the numerator is the same as in Lemma III.5 which is the logarithm of a packing of distributions in total variation distance. The denominator, however, is now replaced with the induced chi-squared fluctuation.

The bounds above are for learning using public-coin protocols, and thus imply the same lower bound for learning with private-coin protocols. Interestingly, for testing, we obtain two different results. In Lemmas IV.8 and IV.10 we show two different conditions on the sample complexity of testing under public- and public-coin protocols in terms of the minmax and maxmin decoupled chi-squared fluctuations. We provide some insights into these bounds, whose proofs can be found in Appendix C. In the case of the maxmin decoupled chi-squared fluctuation, for any W^n we maximize over all possible perturbations: this is the setting of private-coin protocols where after the channels are decided, we can design the perturbations to *fool* the channels. In contrast, for public-coin protocols, we must commit on a given perturbed family first, and then choose the best channels W^n. This is because the shared randomness can be leveraged to choose the set of channels to be dependent on each other.

Lemma IV.8 (Minmax decoupled chi-square fluctuation bound for testing)**.** *For $0 < \varepsilon < 1$ and a k-ary reference distribution $\mathbf{p}$, the sample complexity $n = n(k, \varepsilon)$ of (k, ε)-identity testing using $\mathcal{W}$ for public-coin protocols must satisfy*

$$\overline{\chi}^{(2)}(\mathcal{W}_0^n, \varepsilon) \geq c\,, \tag{17}$$

for some constant $c > 0$ depending only on the probability of error.

Remark IV.9. Using calculations similar to Eq. (15), we can obtain the following counterpart of Eq. (17): For every ε-perturbed family $\mathcal{P}$, it must hold that $\chi^2(\mathcal{W}_0^n \mid \mathcal{P}) = \Omega(1)$. Interestingly, even this bound, although seemingly as weak as Eq. (15), leads to useful bounds in the information-constrained setting. In particular, it will be seen in Section V to yield tight lower bounds for communication-constrained testing for $\ell = 1$.

Lemma IV.10 (Maxmin decoupled chi-square fluctuation bound for testing)**.** *For $0 < \varepsilon < 1$ and a k-ary reference distribution $\mathbf{p}$, the sample complexity $n = n(k, \varepsilon)$ of (k, ε)-identity testing using $\mathcal{W}$ for private-coin protocols must satisfy*

$$\underline{\chi}^{(2)}(\overline{\mathcal{W}}^n, \varepsilon) \geq c\,,$$

for some constant $c > 0$ depending only on the probability of error.

B. Chi-square contraction bounds for learning and testing discrete distributions

All our main tools are in place. We now derive bounds for chi-square fluctuations for Paninski's perturbed family of Eq. (6) and a related almost ε-perturbation, for arbitrary channel families $\mathcal{W}$. These bounds in turn will be used to obtain bounds for maxmin and minmax chi-square fluctuation. Combined with the chi-square fluctuation lower bounds of the previous section, these bounds yield concrete lower bounds on the sample complexity of learning and testing using $\mathcal{W}$. In essence, our bounds precisely characterize the contraction in chi-square fluctuation in the information-constrained setting over the standard setting; we term these bounds the *chi-square contraction bounds*.

As noted in Remark IV.2, the normalized perturbation δ_Z^W is linear in δ_Z. Furthermore, for Paninski's perturbed family, it follows from Eq. (5) that δ_Z itself is linear in Z. This observation allows us to capture chi-square fluctuations in terms of the channel-dependent $(k/2) \times (k/2)$ matrix[10] $H(W)$ defined in Definition I.5, whose expression we recall below:

$$H(W)_{i_1,i_2} := \sum_{y\in\mathcal{Y}} \frac{(W(y \mid 2i_1 - 1) - W(y \mid 2i_1))}{\sum_{x\in[k]} W(y \mid x)} \times \\ (W(y \mid 2i_2 - 1) - W(y \mid 2i_2)),$$

for all $i_1, i_2 \in [k/2]$. An important property of $H(W)$ that will be used throughout is that it is positive semi-definite. Indeed, we can express $H(W)$ as $\sum_y b_y b_y^T$ where the b_ys are $(k/2)$-length vectors with entries given by

$$b_y(i) = \frac{W(y \mid 2i - 1) - W(y \mid 2i)}{\sqrt{\sum_{x\in[k]} W(y \mid x)}}, \quad i \in [k/2].$$

We are now in a position to state our main results. We start with a bound for chi-square fluctuation, which leads to a lower bound for sample complexity of learning.

Theorem IV.11. *For the ε-perturbed family $\mathcal{P}$ in Eq. (6) and any channel W, we have*

$$\chi^2(W \mid \mathcal{P}) = \frac{4\varepsilon^2}{k} \cdot \|H(W)\|_*\,.$$

Remark IV.12. A comparison of the bound above with Eq. (8) shows that the chi-square fluctuation contracts by a factor of at least $(4 \max_{W\in\mathcal{W}} \|H(W)\|_*)/k$ due to the local information constraints corresponding to $\mathcal{W}$.

[10] Specifically, note that $H(W)$ is defined such that, for this perturbed family, $\langle \delta_z^W, \delta_{z'}^W \rangle = \frac{4\varepsilon^2}{k} z^T H(W) z'$ for all z, z'.

Before we prove this theorem, we show how to use it to obtain a lower bound for learning. Recalling Eq. (11), note that the perturbed family $\mathcal{P}$ given in Eq. (6) satisfies

$$\log \frac{|\mathcal{P}|}{C_\varepsilon} \geq \frac{(1-h(1/3))k}{2}.$$

Thus, upon combining the chi-square fluctuation bound in Theorem IV.11 with Lemma IV.7, we obtain the following bound for the sample complexity of distribution learning.

Corollary IV.13 (Chi-square contraction bound for learning)**.** *For $0 < \varepsilon < 1$, the sample complexity of (k, ε)-distribution learning using $\mathcal{W}$ for public-coin protocols is at least*

$$\Omega\left(\frac{k}{\varepsilon^2} \cdot \frac{k}{\sup_{W \in \mathcal{W}_0} \|H(W))\|_*}\right).$$

Proof of Theorem IV.11. Using the expression of the normalized perturbation for $\mathcal{P}$ in Eq. (6), we get

$$\delta_z^W(y) = 2\varepsilon \cdot \frac{\sum_{i\in[k/2]} z_i[W(y \mid 2i-1) - W(y \mid 2i)]}{\sum_{x\in[k]} W(y \mid x)},$$

whereby

$$\begin{aligned}
&\chi^2(W \mid \mathcal{P}) \\
&= \mathbb{E}_Z\left[\|\delta_Z^W\|_2^2\right] \\
&= \frac{4\varepsilon^2}{k} \sum_y \frac{1}{\sum_{x\in[k]} W(y \mid x)} \times \\
&\qquad \mathbb{E}_Z\left[\left(\sum_{i\in[k/2]} Z_i[W(y \mid 2i-1) - W(y \mid 2i)]\right)^2\right] \\
&= \frac{4\varepsilon^2}{k} \sum_{i_1, i_2 \in [k/2]} \mathbb{E}[Z_{i_1} Z_{i_2}]\, H(W)_{i_1,i_2} \\
&= \frac{4\varepsilon^2}{k} \operatorname{Tr} H(W),
\end{aligned}$$

where we have used the definition of $H(W)$ and the fact that $\mathbb{E}[Z_{i_1} Z_{i_2}] = \mathbb{1}_{\{i_1 = i_2\}}$. The claim follows upon noting that $\operatorname{Tr} H(W) = \|H(W)\|_*$ since $H(W)$ is a positive semi-definite matrix. □

Next, we derive an upper bound for minmax chi-square fluctuation. As in the previous part, we obtain this bound by considering the perturbed family in Eq. (6).

Theorem IV.14. *Given $n \in \mathbb{N}$ and $\varepsilon \in (0,1)$, for a channel family $\mathcal{W}$ the minmax chi-square fluctuation is bounded as*

$$\overline{\chi}^{(2)}(\mathcal{W}^n, \varepsilon) = O\left(\frac{n^2\varepsilon^4}{k} \cdot \frac{\max_{W\in\mathcal{W}} \|H(W)\|_F^2}{k}\right),$$

whenever

$$n \leq \frac{k}{16\varepsilon^2 \max_{W\in\mathcal{W}} \|H(W)\|_F}. \tag{18}$$

Remark IV.15. Comparing the bound above with Eq. (16) shows that the decoupled chi-square fluctuation contracts by a factor of $(\max_{W\in\mathcal{W}} \|H(W)\|_F^2)/k$ due to the local information constraints corresponding to $\mathcal{W}$ with respect to the centralized setting (where it was roughly $(n^2\varepsilon^4/k)$).

Before we prove the previous theorem, we note that combining the minmax decoupled chi-square fluctuation bound for testing of Lemma IV.8 with Theorem IV.14 yields the following lower bound for the sample complexity of uniformity testing using public-coin protocols.

Corollary IV.16 (Chi-square contraction bound for testing using public-coin protocols)**.** *For $0 < \varepsilon < 1$, the sample complexity of (k, ε)-uniformity testing using $\mathcal{W}$ for public-coin protocols is at least*

$$\Omega\left(\frac{\sqrt{k}}{\varepsilon^2} \cdot \frac{\sqrt{k}}{\max_{W\in\mathcal{W}_0} \|H(W)\|_F}\right).$$

Proof of Theorem IV.14. We consider the ε-perturbed family $\mathcal{P}$ defined in Eq. (6) and evaluate the fluctuation $\chi^{(2)}(\mathcal{P}^n) = \log \mathbb{E}_{ZZ'}[\exp\left(n \cdot \langle \delta_Z, \delta_{Z'} \rangle\right)]$ for this perturbed family.[11]

We apply Lemma III.8 with $\vartheta = z$, $Q_{j,\vartheta} = W_j \mathbf{p}_z$, $P_j = W_j \mathbf{u}$, $1 \leq j \leq n$ and Z in the role of θ. For brevity, denote by $\rho_{j,y}^{\mathbf{u}}$ and $\rho_{j,y}^z$, respectively, the probability that the output of player j using channel W_j is y when the input distributions are $\mathbf{u}$ and $\mathbf{p}_z$. We have

$$\begin{aligned}
\rho_{j,y}^{\mathbf{u}} &= \sum_{i=1}^{n} \mathbf{u}(i) W_j(y \mid i) \\
&= \frac{2}{k} \sum_{i=1}^{k/2} \left(\frac{W_j(y \mid 2i-1) + W_j(y \mid 2i)}{2}\right),
\end{aligned}$$

and that for every $z \in \{-1,1\}^{k/2}$,

$$\rho_{j,y}^z = \rho_{j,y}^{\mathbf{u}} + \frac{2\varepsilon}{k} \sum_{i=1}^{k/2} z_i \left(W_j(y \mid 2i-1) - W_j(y \mid 2i)\right).$$

Therefore, the quantity δ_j^Z used in Lemma III.8 is given by

$$\begin{aligned}
\delta_j^z(y) &= \frac{\rho_{j,y}^z - \rho_{j,y}^{\mathbf{u}}}{\rho_{j,y}^{\mathbf{u}}} \\
&= \frac{2\varepsilon \sum_{i=1}^{k/2} z_i (W_j(y \mid 2i) - W_j(y \mid 2i-1))}{\sum_{i=1}^{k/2} (W_j(y \mid 2i) + W_j(y \mid 2i-1))},
\end{aligned}$$

whereby for $1 \leq j \leq n$ we get

$$H_j(z, z') = \mathbb{E}\left[\delta_j^z \delta_j^{z'}\right] = \sum_{y\in\mathcal{Y}} \rho_{j,y}^{\mathbf{u}} \delta_j^z(y) \delta_j^{z'}(y),$$

which upon substituting the expressions for $\rho_{j,y}$ and $\delta_j^z(y)$ from above yields

$$H_j(z, z') = \frac{4\varepsilon^2}{k} \cdot z^T H(W_j) z',$$

where the matrix $H(W_j)$ was introduced earlier in Eq. (2). Therefore,

$$\begin{aligned}
\chi^{(2)}\left(W^n \mid \mathcal{P}\right) &= \log \mathbb{E}_{ZZ'}\left[\exp\left(\sum_{j=1}^{n} \left\langle \delta_Z^{W_j}, \delta_{Z'}^{W_j} \right\rangle\right)\right] \\
&= \log \mathbb{E}_{ZZ'}\left[\exp\left(\sum_{j=1}^{n} \frac{4\varepsilon^2}{k} \cdot Z^T H(W_j) Z'\right)\right]
\end{aligned}$$

[11] We need not invoke the more general notion of almost ε-perturbation for this proof; it suffices to use uniform distribution over an ε-perturbed family.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2020.3028440, IEEE Transactions on Information Theory

$$= \log \mathbb{E}_{ZZ'}\left[\exp\left(\frac{4n\varepsilon^2}{k} \cdot Z^T \bar{H} Z'\right)\right], \tag{19}$$

where we denote

$$\bar{H} := \frac{1}{n}\sum_{j=1}^{n} H(W_j). \tag{20}$$

To prove the theorem, we need to bound the expression above in terms of the Frobenius norms of the matrices $H(W_j)$. To that end, we use the following result on Rademacher chaos, whose proof is deferred to Appendix A.

Claim IV.17. *Let θ, θ' be two independent random vectors, each distributed uniformly over $\{-1,1\}^{k/2}$. Then, for any positive semi-definite matrix H,*

$$\log \mathbb{E}_{\theta\theta'}\left[e^{\lambda\theta^T H\theta'}\right] \le \frac{\lambda^2}{2} \cdot \frac{\|H\|_F^2}{1 - 4\lambda^2\rho(H)^2}$$

for $0 \le \lambda < \frac{1}{2\rho(H)}$, where $\|\cdot\|_F$ denotes the Frobenius norm and $\rho(\cdot)$ the spectral radius.

With this result at our disposal, we are ready to complete our proof. Setting $\lambda := \frac{4n\varepsilon^2}{k}$, under assumption Eq. (18) we have

$$\begin{aligned} 1 &\ge \frac{16n\varepsilon^2 \cdot \max_{W\in\mathcal{W}}\|H(W)\|_F}{k} \ge \frac{16n\varepsilon^2 \cdot \|\bar{H}\|_F}{k} \\ &\ge \frac{16n\varepsilon^2 \cdot \rho(\bar{H})}{k} = 4\lambda\rho(\bar{H}), \end{aligned}$$

where the second inequality uses convexity of norm. Rearranging the terms we obtain that $\lambda^2/(1-4\lambda^2\rho(\bar{H})^2) \le 4\lambda^2/3$, which when applied along with Claim IV.17 to Eq. (19) further yields

$$\begin{aligned} \chi^{(2)}\left(W^n \mid \mathcal{P}\right) &\le \frac{8n^2\varepsilon^4}{k^2}\frac{\|\bar{H}\|_F^2}{1-4\lambda^2\rho(\bar{H})^2} \\ &\le \frac{8n^2\varepsilon^4}{k^2} \cdot \frac{4}{3} \cdot \|\bar{H}\|_F^2 \\ &\le \frac{32n^2\varepsilon^4}{3k^2} \cdot \frac{1}{n}\sum_{j=1}^{n}\|H(W_j)\|_F^2 \\ &\le \frac{32n^2\varepsilon^4}{3k^2} \cdot \max_{W\in\mathcal{W}}\|H(W)\|_F^2, \end{aligned}$$

where the penultimate inequality uses the convexity of x^2 in x; the proof is complete. □

Finally, we provide a bound for the maxmin chi-square fluctuation for a channel family $\mathcal{W}$.

Theorem IV.18. *Given $n \in \mathbb{N}$ and $\varepsilon \in (0,1)$, for a channel family $\mathcal{W}$ the (n,ε)-maxmin chi-square fluctuation is bounded as*

$$\underline{\chi}^{(2)}(\mathcal{W}^n, \varepsilon) = O\left(\frac{n^2\varepsilon^4}{k^3} \cdot \max_{W\in\mathcal{W}}\|H(W)\|_*^2\right),$$

whenever

$$n \le C \cdot \frac{k^{3/2}}{\varepsilon^2 \max_{W\in\mathcal{W}}\|H(W)\|_*}, \tag{21}$$

where $C > 0$ is a universal constant.

Remark IV.19. Comparing the bound above with Eq. (16) shows that the decoupled chi-square fluctuation contracts by a factor of $(1/k^2)\max_{W\in\mathcal{W}}\|H(W)\|_*^2$ due to local information constraints, when restricting to private-coin protocols, which is worse than the contraction for public-coin protocols in view of Eq. (4).

Note that combining the maxmin decoupled chi-square fluctuation bound for testing in Lemma IV.10 with Theorem IV.18 yields the following lower bound for the sample complexity of uniformity testing using private-coin protocols.

Corollary IV.20 (Chi-square contraction bound for testing using private-coin protocols)**.** *For $0 < \varepsilon < 1$, the sample complexity of (k,ε)-uniformity testing using $\mathcal{W}$ for private-coin protocols is at least*

$$\Omega\left(\frac{\sqrt{k}}{\varepsilon^2} \cdot \frac{k}{\max_{W\in\overline{\mathcal{W}}}\|H(W)\|_*}\right).$$

Before we provide a formal proof for Theorem IV.18, we summarize the high-level heuristics. In the proof of Theorem IV.14, we showed a bound for decoupled chi-square fluctuation of $\mathcal{P}$ for the induced perturbed family corresponding to the best choice of $W^n \in \mathcal{W}^n$. When only private-coin protocols are allowed, we can in fact design a perturbed family with the least decoupled chi-square fluctuation for the specific choice of W^n used. Furthermore, we identify this least favorable direction of perturbation for W^n by exploiting the spectrum of the positive semi-definite matrix $\bar{H}$ given in Eq. (20); details follow.

Proof of Theorem IV.18. To obtain the desired bound for maxmin chi-square fluctuation, we derive a bound for decoupled chi-square fluctuation for an appropriately chosen almost ε-perturbation $\mathcal{P}_\zeta$. Specifically, consider a random variable $Z = (Z_1, \ldots, Z_{k/2})$ taking values in $[-1,1]^{k/2}$ and with distribution ζ such that for some constants $\alpha \ge 1/10$ and $c > 0$,

$$\Pr\left[\|Z\|_1 \ge \frac{k}{c}\right] \ge \alpha. \tag{22}$$

For $\varepsilon \in (0, c^{-1})$, consider the perturbed family around $\mathbf{u}$ consisting of elements $\mathbf{p}_z$, $z \in [-1,1]^{k/2}$, given by

$$\mathbf{p}_z = \frac{1}{k}(1 + c\varepsilon z_1, 1 - c\varepsilon z_1, \ldots, 1 + c\varepsilon z_{k/2}, 1 - c\varepsilon z_{k/2}).$$

By the condition in Eq. (22) on Z, $\mathbf{p}_Z$ satisfies the following property with probability greater than α:

$$\mathrm{d_{TV}}(\mathbf{p}_Z, \mathbf{u}) = \frac{c}{2}\sum_{i=1}^{k/2}\frac{2\varepsilon|Z_i|}{k} = \frac{c\varepsilon}{k}\|Z\|_1 \ge \varepsilon.$$

Note that if we set $Z_i = Y_i$ for $Y_1, \ldots, Y_{k/2}$ independent Rademacher random variables and the constant $c = 2$, we recover the standard Paninski construction. However, we can do much more with this general construction. In particular, we can set Z_is to be dependent, which will be used crucially in our proof. For a fixed channel family $\mathcal{W}$, we bound its (n,ε)-maxmin decoupled chi-square fluctuation by fixing an arbitrary

$W^n \in \mathcal{W}^n$ and exhibit a perturbed family $\mathcal{P}_\varepsilon(\mathcal{W}) = \mathcal{P}_{\zeta_\mathcal{W}}$ by designing a specific distribution $\zeta_\mathcal{W}$ to "fool" it.

We proceed by bounding $\chi^{(2)}\left(W^n \mid \mathcal{P}_\zeta\right)$ for a distribution ζ satisfying Eq. (22). Following the proof of Theorem IV.14, we get

$$\chi^{(2)}\left(W^n \mid \mathcal{P}_\zeta\right) = \log \mathbb{E}_{ZZ'}\left[\exp\left(\frac{c^2\varepsilon^2}{k} \cdot Z^T \left(\sum_{j=1}^n H(W_j)\right) Z'\right)\right],$$

where Z, Z' are independent random variables with common distribution ζ and $H(W_j)$ is defined as in Eq. (2). Note that

$$\chi^{(2)}\left(W^n \mid \mathcal{P}_\zeta\right) = \log \mathbb{E}_{ZZ'}\left[\exp\left(\frac{c^2 n\varepsilon^2}{k} \cdot Z^T \bar{H} Z'\right)\right],$$

where the matrix $\bar{H}$ is from Eq. (20). Informally, the matrix $\bar{H}$ captures the directions of the input space where the n-fold channel W^n is the most informative; and thus, our goal is to design a distribution ζ which avoids these directions as much as possible.

To make this precise, let $0 \le \lambda_1 \le \lambda_2 \le \cdots \le \lambda_{k/2}$ be the eigenvalues of $\bar{H}$, and $\mathbf{v}^1, \dots, \mathbf{v}^{k/2}$ be corresponding (orthonormal) eigenvectors; in particular,

$$\bar{H} = \sum_{i=1}^{k/2} \lambda_i \mathbf{v}^i (\mathbf{v}^i)^T .$$

Denote by V the $(k/2) \times (k/4)$ matrix with columns given by $\mathbf{v}^i$ for $i \le k/4$, *i.e.*, the columns are the vectors corresponding to the $k/4$ smallest eigenvalues of $\bar{H}$. Let $Y_1 \dots Y_{k/4}$ be i.i.d. Rademacher random variables, and set ζ as the distribution of the random variable $Z := VY$.

The first claim below shows that ζ satisfies Eq. (22).[12]

Claim IV.21. *For $Z = VY$ described above, we have*

$$\Pr\left[\|Z\|_1 \ge \frac{k}{12\sqrt{2}}\right] \ge \frac{1}{9}.$$

Proof. For $m \in [k/2]$, we have $Z_m = \sum_{i=1}^{k/4} V_{m,i} Y_i$ where $V_{m,i}$ equals $\mathbf{v}^i_m$. Therefore, by Khintchine's inequality (*cf.* [48]),

$$\mathbb{E}[\|Z\|_1] = \sum_{m=1}^{k/2} \mathbb{E}[|Z_m|] \ge \frac{1}{\sqrt{2}} \sum_{m=1}^{k/2} \|\mathbf{v}_m\|_2,$$

where $\mathbf{v}_1, \dots, \mathbf{v}_{k/2}$ denote the row vectors of the matrix V.

Next, we note that $\|\mathbf{v}_m\|_2 \le 1$ for every $m \in [k/2]$. Indeed, denoting by V' the $(k/2) \times (k/2)$ matrix obtained by adding extra columns to V to obtain a complete orthonormal basis for $\mathbb{R}^{k/2}$, we have $V'^T V' = I$, whereby $V' V'^T = I$. Thus, each row $\mathbf{v}'_m$ of V' has $\|\mathbf{v}'_m\|_2 = 1$, which gives

$$\|\mathbf{v}_m\|_2^2 \le \|\mathbf{v}'_m\|_2^2 = 1.$$

Upon combining the bounds above, we obtain

$$\mathbb{E}[\|Z\|_1] \ge \frac{1}{\sqrt{2}} \sum_{m=1}^{k/2} \|\mathbf{v}_m\|_2 \ge \frac{1}{\sqrt{2}} \sum_{m=1}^{k/2} \|\mathbf{v}_m\|_2^2 = \frac{1}{\sqrt{2}} \sum_{m=1}^{k/2} \sum_{i=1}^{k/4} V_{m,i}^2 = \frac{1}{\sqrt{2}} \sum_{i=1}^{k/4} \|\mathbf{v}^i\|_2^2 = \frac{k}{4\sqrt{2}},$$

where in the second inequality we used $\|\mathbf{v}_m\|_2^2 \le 1$.

Moreover, note that

$$\mathbb{E}\left[\|Z\|_2^2\right] = \sum_{m=1}^{k/2} \sum_{i=1}^{k/4} V_{m,i}^2 = \frac{k}{4},$$

which further gives

$$\mathbb{E}\left[\|Z\|_1^2\right] \le \frac{k}{2} \mathbb{E}\left[\|Z\|_2^2\right] = \frac{k^2}{8}.$$

Therefore, by the Paley–Zygmund inequality, for any $\theta \in (0,1)$

$$\Pr\left[\|Z\|_1 \ge \frac{\theta}{4\sqrt{2}} k\right] \ge (1-\theta)^2 \frac{\mathbb{E}[\|Z\|_1]^2}{\mathbb{E}\left[\|Z\|_1^2\right]} \ge \frac{(1-\theta)^2}{4}.$$

The proof is completed by setting $\theta = 1/3$. □

We will also require the following property, which is ensured by our construction of the matrix V.

Claim IV.22. *For $V \in \mathbb{R}^{(k/2)\times(k/4)}$ defined as above, we have*

$$\|V^T \bar{H} V\|_F^2 \le \frac{4}{k} \|\bar{H}\|_*^2 .$$

Proof. Note that since for $i_1, i_2 \in [k/4]$, we have

$$\begin{aligned}(V^T \bar{H} V)_{i_1,i_2} &= (\mathbf{v}^{i_1})^T \left(\sum_{i=1}^{k/2} \lambda_i \mathbf{v}^i (\mathbf{v}^i)^T\right) \mathbf{v}^{i_2} \\ &= \sum_{i=1}^{k/2} \lambda_i (\mathbf{v}^{i_1})^T \mathbf{v}^i (\mathbf{v}^i)^T \mathbf{v}^{i_2} \\ &= \sum_{i=1}^{k/2} \lambda_i \left\langle \mathbf{v}^{i_1}, \mathbf{v}^i \right\rangle \left\langle \mathbf{v}^{i_2}, \mathbf{v}^i \right\rangle ,\end{aligned}$$

Thus, by the orthonormality of $\mathbf{v}^i$s, the matrix $V^T \bar{H} V$ is diagonal, with diagonal entries $\lambda_1, \dots, \lambda_{k/4}$. It follows that

$$\|V^T \bar{H} V\|_F^2 = \sum_{i=1}^{k/4} \lambda_i^2 \le \frac{k}{4} \cdot \lambda_{k/4}^2 .$$

On the other hand, we also have

$$\lambda_{k/4} \le \frac{4}{k} \sum_{i=k/4+1}^{k/2} \lambda_i \le \frac{4}{k} \operatorname{Tr} \bar{H}$$

and therefore,

$$\|V^T \bar{H} V\|_F^2 \le \frac{4}{k} (\operatorname{Tr} \bar{H})^2$$

which is what we sought. □

We proceed to bound $\chi^{(2)}\left(W^n \mid \mathcal{P}_\zeta\right)$. First, note that

$$\begin{aligned}\max_{W\in\mathcal{W}} \|H(W)\|_* &\ge \frac{1}{n} \sum_{j=1}^n \|H(W_j)\|_* = \frac{1}{n} \sum_{j=1}^n \operatorname{Tr} H(W_j) \\ &= \operatorname{Tr} \bar{H} = \|\bar{H}\|_*,\end{aligned}$$

where the first identity holds since $H(W)$ is positive semi-definite for every $W \in \mathcal{W}$. Using Claim IV.22 and the above

[12] The probability guarantees obtained in Claim IV.21 determined our choice $1/12$ for the probability of error in our formulations in Section I.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2020.3028440, IEEE Transactions on Information Theory

bound, with the view of using Claim IV.17 and setting $\lambda := (c^2 n\varepsilon^2)/k$, under assumption Eq. (21) (where $C = 1/(8c^2)$) we have

$$1 \geq \frac{8c^2 n\varepsilon^2 \cdot \max_{W\in\mathcal{W}}\|H(W)\|_*}{k^{3/2}} \geq \frac{8c^2 n\varepsilon^2 \|\bar{H}\|_*}{k^{3/2}} \\ \geq 4\lambda\|V^T\bar{H}V\|_F \geq 4\lambda\rho(V^T\bar{H}V)\,.$$

Rearranging the terms to obtain $\lambda^2/(1-4\lambda^2\rho(\bar{H})^2) \leq 4\lambda^2/3$ and applying Claim IV.17 to i.i.d. Rademacher random variables Y and the symmetric matrix $V^T\bar{H}V \in \mathbb{R}^{k/4\times k/4}$ gives

$$\begin{aligned}&\mathbb{E}_{ZZ'}\left[\exp\left(\frac{c^2 n\varepsilon^2}{k} Z^T\bar{H}Z'\right)\right]\\ &= \mathbb{E}_{YY'}[e^{\frac{c^2n\varepsilon^2}{k}Y^TV^T\bar{H}VY'}] - 1\\ &\leq e^{\frac{2c^4n^2\varepsilon^4}{3k^2}\|V^T\bar{H}V\|_F^2} - 1\,. \end{aligned}\tag{23}$$

It remains to bound the Frobenius norm on the right-side above. To do so, we invoke once more Claim IV.22 which, along with Eq. (23), gives

$$\begin{aligned}&\mathbb{E}_{ZZ'}\left[\exp\left(\frac{c^2 n\varepsilon^2}{k}\cdot Z^T\bar{H}Z'\right)\right]\\ &\quad\leq \exp\left(\frac{8c^4n^2\varepsilon^4}{3k^3}(\operatorname{Tr}\bar{H})^2\right) - 1\,,\end{aligned}$$

which completes the proof. □

On comparing Corollary IV.16 and Corollary IV.20, we note that the effective contraction in decoupled chi-square fluctuation due to private-coin protocols is roughly $\frac{k}{\max_{W\in\mathcal{W}}\|H(W)\|_*}$, which exceeds $\frac{\sqrt{k}}{\max_{W\in\mathcal{W}}\|H(W)\|_F}$ for public-coin protocol since $H(W)$ has rank $O(k)$ and so by Eq. (4), $\|H(W)\|_* \leq \sqrt{k}\cdot\|H(W)\|_F$.

Remark IV.23. Both channel families we consider in this paper, namely $\mathcal{W}_\ell$ for the communication-limited setting and $\mathcal{W}_\rho$ for the LDP setting, are convex and satisfy $\overline{\mathcal{W}} = \mathcal{W}$. Moreover, when evaluating bounds in Corollary IV.13 and Corollary IV.16 for these families, weaker bounds derived using $\mathcal{W}$ in place of $\mathcal{W}_0$ turn out to be optimal. Thus, our evaluations for these cases in the next section are based on $\mathcal{W}$ and do not require us to consider $\mathcal{W}_0$ or $\overline{\mathcal{W}}$. However, the more general form reported in this section may be useful elsewhere; in particular, in cases where one can identify a $\mathcal{W}_0$ that is more amenable to these bounds than $\mathcal{W}$ itself.

V. Examples and Applications

We now instantiate our general bounds for distribution learning and uniformity testing derived in the previous section to our two running examples of local information constraints, namely the communication-limited and LDP settings. We obtain tight lower bounds for sample complexity of learning and testing in these settings simply by bounding the Frobenius and trace norms of the associated matrices $H(W)$; see Table I for a summary of the results obtained. As mentioned earlier, we only focus on lower bounds here and delegate matching upper bounds to subsequent papers in this series.

A. *Communication-constrained inference*

Recall that in the communication-limited setting, each player can transmit at most ℓ bits, which can be captured by using $\mathcal{W} = \mathcal{W}_\ell$, the family of channels from $[k]$ to $\mathcal{Y} = \{0,1\}^\ell$. To derive lower bounds for sample complexity of learning and testing for this case, Corollaries IV.16 and IV.20 require us to obtain upper bounds for $\max_{W\in\mathcal{W}_0}\|H(W)\|_*$, $\max_{W\in\mathcal{W}_0}\|H(W)\|_*$ and $\max_{W\in\overline{\mathcal{W}}}\|H(W)\|_*$. We begin by observing that $\mathcal{W}$ is convex, whereby $\mathcal{W} = \overline{\mathcal{W}}$ which allows us to focus on $\|H(W)\|_*$ and $\|H(W)\|_F$ for $W \in \mathcal{W}$. Indeed, the convex combination of two ℓ-bit output channels is an ℓ-bit channel as well.

The next result provides bounds for the trace and Frobenius norms of the matrices $H(W)$ under communication constraints.

Lemma V.1. *For a channel $W\colon [k] \to \{0,1\}^\ell$ and $H(W)$ as in Eq. (2), we have*

$$\|H(W)\|_* \leq 2^\ell \text{ and } \|H(W)\|_F^2 \leq 2^{\ell+1}.$$

Proof. Since matrix $H(W)$ is a positive semi-definite matrix, by the definition of nuclear norm in Section II, we have

$$\begin{aligned}\|H(W)\|_* &= \operatorname{Tr} H(W)\\ &= \sum_{i=1}^{k/2}\sum_{y\in\mathcal{Y}}\frac{(W(y\mid 2i-1)-W(y\mid 2i))^2}{\sum_{i'\in[k]}W(y\mid i')}\\ &\leq \sum_{i=1}^{k/2}\sum_{y\in\mathcal{Y}}\frac{W(y\mid 2i-1)+W(y\mid 2i)}{\sum_{i'\in[k]}W(y\mid i')}\\ &= \sum_{y\in\mathcal{Y}}\frac{\sum_{i=1}^{k/2}W(y\mid 2i-1)+W(y\mid 2i)}{\sum_{i'\in[k]}W(y\mid i')}\\ &= 2^\ell\,.\end{aligned}$$

Moreover, for $y \in \mathcal{Y}$, denote by $\omega_y \in [0,1]^{[k/2]}$ the vector with the ith coordinate given by $\omega_{y,i} := W(y\mid 2i-1) + W(y\mid 2i)$. Then,

$$\begin{aligned}&\|H(W)\|_F^2\\ &= \sum_{i_1,i_2\in[k/2]}\Bigg(\sum_{y\in\mathcal{Y}}\frac{(W(y|2i_1-1)-W(y|2i_1))}{\sum_{i\in[k]}W(y|i)}\times\\ &\qquad\qquad (W(y\mid 2i_2-1)-W(y\mid 2i_2))\Bigg)^2\\ &\leq \sum_{i_1,i_2\in[k/2]}\left(\sum_{y\in\mathcal{Y}}\frac{\omega_{y,i_1}\omega_{y,i_2}}{\sum_{i\in[k/2]}\omega_{y,i}}\right)^2\\ &= \sum_{i_1,i_2\in[k/2]}\sum_{y_1,y_2\in\mathcal{Y}}\frac{\omega_{y_1,i_1}\omega_{y_1,i_2}\omega_{y_2,i_1}\omega_{y_2,i_2}}{\sum_{i\in[k/2]}\omega_{y_1,i}\cdot\sum_{i\in[k/2]}\omega_{y_2,i}}\\ &= \sum_{y_1,y_2\in\mathcal{Y}}\frac{\sum_{i_1\in[k/2]}\omega_{y_1,i_1}\omega_{y_2,i_1}\cdot\sum_{i_2\in[k/2]}\omega_{y_1,i_2}\omega_{y_2,i_2}}{\sum_{i\in[k/2]}\cdot\omega_{y_1,i}\sum_{i\in[k/2]}\omega_{y_2,i}}\\ &= \sum_{y_1,y_2\in\mathcal{Y}}\frac{\langle\omega_{y_1},\omega_{y_2}\rangle^2}{\langle\omega_{y_1},\mathbf{1}\rangle\langle\omega_{y_2},\mathbf{1}\rangle}\\ &\leq \sum_{y_1,y_2\in\mathcal{Y}}\frac{\langle\omega_{y_1},\omega_{y_2}\rangle}{\langle\omega_{y_1},\mathbf{1}\rangle}\end{aligned}$$

$$= 2 \sum_{y_1 \in \mathcal{Y}} \frac{\langle \omega_{y_1}, \mathbf{1} \rangle}{\langle \omega_{y_1}, \mathbf{1} \rangle}$$
$$= 2^{\ell+1},$$

where in the penultimate identity we used the observation that $\sum_{y \in \mathcal{Y}} \omega_{y,i} = 2$, for every $i \in [k/2]$. □

Plugging these bounds into Corollaries IV.13, IV.16 and IV.20 and recalling that $\mathcal{W} = \overline{\mathcal{W}}$ yield the following corollaries.

Theorem V.2 (Communication-limited learning using public coins)**.** *The sample complexity of (k, ε)-distribution learning using $\mathcal{W}_\ell$ for public-coin protocols is at least $\Omega\big(k^2/(2^\ell \varepsilon^2)\big)$.*

Theorem V.3 (Communication-limited testing using public coins)**.** *The sample complexity of (k, ε)-uniformity testing using $\mathcal{W}_\ell$ for public-coin protocols is at least $\Omega\big(k/(2^{\ell/2} \varepsilon^2)\big)$.*

Theorem V.4 (Communication-limited testing using private coins)**.** *The sample complexity of (k, ε)-uniformity testing using $\mathcal{W}_\ell$ for private-coin protocols is at least $\Omega\big(k^{3/2}/(2^\ell \varepsilon^2)\big)$.*

Thus, the blow-up in sample complexity for communication-limited learning with public-coin protocols is a factor of $k/2^\ell$, which is the same for testing with private-coin protocols. This blow-up is reduced to a factor of $\sqrt{k/2^\ell}$ for testing with public-coin protocols. In fact, these bounds are tight and match the upper bounds in [3], [34] for learning, with a private-coin protocol achieving the public-coin lower bound, and [3] for both testing using private- and public-coin protocols.

B. Local differential privacy constraints

Moving now to inference under LDP setting, recall that the information constraints here are captured by the family $\mathcal{W}_\rho$ of ρ-LDP channels $W\colon [k] \to \mathcal{Y}$ satisfying

$$\sup_{y \in \mathcal{Y}} \sup_{i_1, i_2 \in [k]} \frac{W(y \mid i_1)}{W(y \mid i_2)} \le e^\rho . \qquad (24)$$

As before, we seek bounds for $\|H(W)\|_*$ and $\|H(W)\|_F$. Observe that $\mathcal{W}_\rho$ is convex: indeed, for $W_1, W_2 \in \mathcal{W}_\rho$, and for any $\theta \in [0,1]$, $\in \mathcal{W}$, and $i \neq j$,

$$\theta W_1(y \mid i) + (1-\theta) W_2(y \mid i)$$
$$\le (\theta W_1(y \mid j) + (1-\theta) W_2(y \mid j)) \cdot e^\rho .$$

Thus, $\overline{\mathcal{W}_\rho} = \mathcal{W}_\rho$, and in the result below we may restrict to bounds for trace and Frobenius norms of $H(W)$ for $W \in \mathcal{W}_\rho$.

Lemma V.5. *For $\rho \in (0,1]$, a ρ-LDP channel $W \in \mathcal{W}_\rho$ and $H(W)$ as in Eq. (2), we have*

$$\|H(W)\|_* = O(\rho^2) \text{ and } \|H(W)\|_F^2 = O(\rho^4) .$$

Proof. For the symmetric matrix $H(W)$ with $W \in \mathcal{W}_\rho$, we have

$$\|H(W)\|_* = \operatorname{Tr} H(W)$$
$$= \sum_{i=1}^{k/2} \sum_{y \in \mathcal{Y}} \frac{(W(y \mid 2i-1) - W(y \mid 2i))^2}{\sum_{i' \in [k]} W(y \mid i')}$$
$$\le (e^\rho - 1)^2 \sum_{i=1}^{k/2} \sum_{y \in \mathcal{Y}} \frac{\left(\frac{1}{k} \sum_{i' \in [k]} W(y \mid i')\right)^2}{\sum_{i' \in [k]} W(y \mid i')}$$
$$= \frac{(e^\rho - 1)^2}{2k} \sum_{y \in \mathcal{Y}} \sum_{i' \in [k]} W(y \mid i')$$
$$= \frac{1}{2} (e^\rho - 1)^2 ,$$

where the first inequality as Eq. (24) implies that, for every $W \in \mathcal{W}_\rho$, $y \in \mathcal{Y}$, and $i_1, i_2, i_3 \in [k]$,

$$W(y \mid i_1) - W(y \mid i_2) \le (e^\rho - 1) W(y \mid i_3). \qquad (25)$$

To see why, observe that when $W(y \mid i_3) \ge W(y \mid i_2)$, by Eq. (24) we have

$$W(y \mid i_1) - W(y \mid i_2) \le (e^\rho - 1) W(y \mid i_2)$$
$$\le (e^\rho - 1) W(y \mid i_3) ,$$

and when $W(y \mid i_3) < W(y \mid i_2)$ we have

$$W(y \mid i_1) - W(y \mid i_2) \le e^\rho W(y \mid i_3) - W(y \mid i_2)$$
$$< (e^\rho - 1) W(y \mid i_3) ,$$

thereby establishing Eq. (25). Note that $\frac{1}{2}(e^\rho - 1)^2 = O(\rho^2)$ for $\rho \in (0,1]$, which completes the proof of the bound for $\|H(W)\|_*$. Moreover, from Eq. (4), we have $\|H(W)\|_F^2 \le \|H(W)\|_*^2 = O(\rho^4)$, concluding the proof of the lemma. □

Combining this with Corollaries IV.13, IV.16 and IV.20, respectively, we obtain the following lower bounds on learning and testing under LDP constraints.

Theorem V.6 (LDP learning using public coins)**.** *For $\rho \in (0,1]$, the sample complexity (k, ε)-distribution learning using $\mathcal{W}_\rho$ for public-coin protocols is at least $\Omega\big(k^2/(\rho^2 \varepsilon^2)\big)$.*

Theorem V.7 (LDP testing using public coins)**.** *For $\rho \in (0,1]$, the sample complexity of (k, ε)-uniformity testing using $\mathcal{W}_\rho$ for public-coin protocols is at least $\Omega\big(k/(\rho^2 \varepsilon^2)\big)$.*

Theorem V.8 (LDP testing using private coins)**.** *For $\rho \in (0,1]$, the sample complexity of (k, ε)-uniformity testing using $\mathcal{W}_\rho$ for private-coin protocols is at least $\Omega\big(k^{3/2}/(\rho^2 \varepsilon^2)\big)$.*

Similarly to the communication-limited setting, we see a separation between lower bounds for private- and public-coin protocols for testing under LDP constraints. In fact, the public-coin protocols for learning under LDP constraints from [21], [37], [56], [4], [52] match our lower bounds. Furthermore, [2], [1] provide private- and public-coin protocols for testing under LDP constraints that match our lower bounds here. Thus, indeed shared randomness strictly reduces sample complexity of testing when operating under LDP constraints.

VI. Future Directions and Upcoming Results

We have restricted our focus to lower bounds in this paper. Distributed inference schemes requiring number of players matching the lower bounds derived here will appear in two upcoming papers in this series. While these schemes will elaborate on the geometric view developed in this paper, the algorithms are new and tools needed for analysis are varied.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2020.3028440, IEEE Transactions on Information Theory

We chose to organize these closely related papers into three separate parts for ease of presentation and to disentangle the distinct ideas involved.

In [3], the second paper in this series, we focus on the communication-constrained setting and provide public- and private-coin protocols for distributed inference whose performance matches the lower bounds presented here. A general strategy of "simulate-and-infer," which is a private-coin protocol (and, in fact, a deterministic protocol), is used to achieve our bound learning as well as the bound for testing for private-coin protocols. On the other hand, a different scheme based on a random partition of inputs is used to attain bounds for testing with public-coin protocols. The efficacy of this latter scheme is closely tied to the geometric view developed here.

In [1], the third paper in this series, we provide schemes for testing under the LDP setting. For private-coin protocols, we simply use existing mechanisms such as RAPPOR and design sample-optimal tests for the $\mathcal{R}$. On the other hand, our bounds in this paper show that none of the existing LDP mechanisms, which are all private-coin protocols, can attain the public-coin lower bound. We present a new public-coin protocol that achieves our lower bounds here. Interestingly, our optimal public-coin protocol is similar to the one used in the communication-limited setting and draws on the geometric view developed here.

Finally, we point out that our framework readily extends to the high-dimensional and continuous settings, and can, for instance, be used to analyze the lower bounds for the problems of Gaussian mean testing and testing of product distributions under information constraints. We defer these interesting research directions to future work.

Appendix A
Proof of Claim IV.17

In this appendix, we prove Claim IV.17 which is recalled below for easy reference.

Claim A.1 (Claim IV.17, restated)**.** *Let θ, θ' be two independent random vectors, each distributed uniformly over $\{-1,1\}^{k/2}$. Then, for any positive semi-definite matrix H,*

$$\log \mathbb{E}_{\theta\theta'}\left[e^{\lambda\theta^T H\theta'}\right] \leq \frac{\lambda^2}{2} \cdot \frac{\|H\|_F^2}{1-4\lambda^2\rho(H)^2}$$

for $0 \leq \lambda < \frac{1}{2\rho(H)}$, where $\|\cdot\|_F$ denotes the Frobenius norm and $\rho(\cdot)$ the spectral radius.

Proof. The proof follows closely that of [27, Proposition 8.13], which derives tail bounds on a homogeneous Rademacher chaos of order 2 by bounding the moment-generating function. For θ, θ', and H as above and $\lambda \in \mathbb{R}$,

$$\begin{aligned}\mathbb{E}_{\theta\theta'}\left[e^{\lambda\theta^T H\theta'}\right] &= \mathbb{E}_\theta\left[\mathbb{E}_{\theta'}\left[e^{\lambda\sum_{i_1=1}^{k/2}\theta'_{i_1}\sum_{i_2=1}^{k/2}\theta_{i_2}H_{i_1i_2}}\right]\right]\\ &\leq \mathbb{E}_\theta e^{\frac{\lambda^2}{2}\sum_{i_1=1}^{k/2}\left(\sum_{i_2=1}^{k/2}\theta_{i_2}H_{i_1i_2}\right)^2}, \qquad (26)\end{aligned}$$

where to bound the inner expectation conditionally on θ we used the fact that Rademacher variables are sub-Gaussian and the sum of independent sub-Gaussian variables is sub-Gaussian. Since H is symmetric, we can rewrite $\sum_{i_1=1}^{k/2}\left(\sum_{i_2=1}^{k/2}\theta_{i_2}H_{i_1i_2}\right)^2 = \sum_{i_2,i_3}\theta_{i_2}\theta_{i_3}\sum_{i_1}H_{i_1i_2}H_{i_1i_3} = \theta^T H^2\theta$. Thus, for $M := H^2$ and $\mu \in \mathbb{R}$, we can consider

$$\begin{aligned}\mathbb{E}_\theta\left[e^{\mu\theta^T M\theta}\right] &= \mathbb{E}_\theta\left[e^{\mu\sum_{i=1}^{k/2}M_{ii}+\mu\sum_{i_1\neq i_2}M_{i_1i_2}\theta_{i_1}\theta_{i_2}}\right]\\ &= e^{\mu\operatorname{Tr}M}\mathbb{E}_\theta\left[e^{\mu\sum_{i_1\neq i_2}M_{i_1i_2}\theta_{i_1}\theta_{i_2}}\right]\\ &\leq e^{\mu\operatorname{Tr}M}\mathbb{E}_{\theta\theta'}\left[e^{4\mu\sum_{i_1,i_2\in[k/2]}M_{i_1i_2}\theta_{i_1}\theta'_{i_2}}\right]\\ &\leq e^{\mu\operatorname{Tr}M}\mathbb{E}_\theta\left[e^{8\mu^2\sum_{i_1=1}^{k/2}\left(\sum_{i_2=1}^{k/2}\theta_{i_2}M_{i_1i_2}\right)^2}\right],\end{aligned}$$

where the first inequality is by the decoupling inequality $\mathbb{E}\left[e^{\theta^T M\theta}\right] \leq \mathbb{E}\left[e^{\theta^T M\theta'}\right]$ (used in [27] as well) and the second uses sub-Gaussianity once again. Since $M = H^T H$ is positive semi-definite, we can rewrite

$$\sum_{i_1=1}^{k/2}\left(\sum_{i_2=1}^{k/2}\theta_{i_2}M_{i_1i_2}\right)^2 = \theta^T M^2\theta \leq \|M\|_2 \cdot \theta^T M\theta\,,$$

where $\|M\|_2 := \sup_{\|\mathbf{x}\|_2\leq 1}\langle M\mathbf{x},\mathbf{x}\rangle$ is the operator norm of M. For $8\mu\|M\|_2 \leq 1$, applying Jensen's inequality to the concave function $t \mapsto t^{8\mu\|M\|_2}$ we get

$$\begin{aligned}\mathbb{E}_\theta\left[e^{\mu\theta^T M\theta}\right] &\leq e^{\mu\operatorname{Tr}M}\mathbb{E}_\theta\left[e^{8\mu^2\|M\|_2\theta^T M\theta}\right]\\ &\leq e^{\mu\operatorname{Tr}M}\mathbb{E}_\theta\left[e^{\mu e^{\theta^T M\theta}}\right]^{8\mu\|M\|_2},\end{aligned}$$

which yields

$$\mathbb{E}_\theta\left[e^{\mu\theta^T M\theta}\right] \leq e^{\mu\frac{\operatorname{Tr}M}{1-8\mu\|M\|_2}}\,. \qquad (27)$$

Recalling that $\operatorname{Tr}M = \operatorname{Tr}(H^2) = \|H\|_F^2$ and $\|M\|_2 = \|H^2\|_2 = \rho(H)^2$, and choosing $\mu = \lambda^2/2$ (which satisfies $8\mu\|M\|_2 \leq 1$), we get from Eqs. (26) and (27) that

$$\mathbb{E}_{\theta\theta'}\left[e^{\lambda\theta^T H\theta'}\right] \leq \mathbb{E}_\theta\left[e^{\frac{\lambda^2}{2}\theta^T H^2\theta}\right] \leq e^{\frac{\lambda^2}{2}\frac{\|H\|_F^2}{1-4\lambda^2\rho(H)^2}},$$

which completes the proof. □

Appendix B
Proofs of Chi-square Fluctuation Bounds

Proof of Lemma III.7. The proof uses Le Cam's two-point method. We note first that

$$\mathrm{d_{TV}}(\mathbb{E}[\mathbf{p}_Z^n]\,,\mathbf{p}^n)^2 \leq \mathrm{d}_{\chi^2}(\mathbb{E}[\mathbf{p}_Z^n]\,,\mathbf{p}^n)\,,$$

and bound the right-side further using Lemma III.8 with θ replaced by z, $Q_\vartheta^n = \mathbf{p}_z^n$, and $P_i = \mathbf{p}$ to get

$$\begin{aligned}\mathrm{d_{TV}}(\mathbb{E}[\mathbf{p}_Z^n]\,,\mathbf{p}^n)^2 &\leq \mathbb{E}_{ZZ'}[(1+H_1(Z,Z'))^n]-1\\ &\leq \mathbb{E}_{ZZ'}\left[e^{nH_1(Z,Z')}\right]-1\\ &= \exp\left(\chi^{(2)}(\mathcal{P}^n)\right)-1, \qquad (28)\end{aligned}$$

since $H_1(Z,Z') = \langle\delta_Z,\delta_{Z'}\rangle$. Now, to complete the proof, consider an (n,ε)-test $\mathcal{T}$. By definition, we have $\Pr_{X^n\sim\mathbf{p}^n}[\mathcal{T}(X^n)=1] > 11/12$ and $\Pr_{X^n\sim\mathbf{p}_z^n}[\mathcal{T}(X^n)=1] > 11/12$ for every z, whereby

$$\frac{1}{2}\Pr_{X^n\sim\mathbf{p}^n}[\mathcal{T}(X^n)\neq 1] + \frac{1}{2}\Pr_{X^n\sim\mathbb{E}[\mathbf{p}_Z^n]}[\mathcal{T}(X^n)\neq 0] \leq \frac{1}{12}. \tag{29}$$

The left-hand-side above coincides with the Bayes error for test $\mathcal{T}$ for the simple binary hypothesis testing problem of $\mathbb{E}[\mathbf{p}_Z^n]$ versus $\mathbf{p}^n$, which must be at least

$$\frac{1}{2}\left(1-\mathrm{d}_{\mathrm{TV}}(\mathbb{E}[\mathbf{p}_Z^n], \mathbf{p}^n)\right).$$

Thus, we obtain $\mathrm{d}_{\mathrm{TV}}(\mathbb{E}[\mathbf{p}_Z^n], \mathbf{p}^n) \geq 5/6$, which together with Eq. (28) completes the proof. □

Proof of Lemma III.8. Using the definition of chi-square distance, we have

$$\begin{aligned}\chi^2(\mathbb{E}_\theta[Q_\theta^n], P^n) &= \mathbb{E}_{P^n}\left[\left(\mathbb{E}_\theta\left[\frac{Q_\theta^n(X^n)}{P^n(X^n)}\right]\right)^2\right] - 1\\ &= \mathbb{E}_{P^n}\left[\left(\mathbb{E}_\theta\left[\prod_{i=1}^n(1+\Delta_i^\theta)\right]\right)^2\right] - 1,\end{aligned}$$

where the outer expectation is for X^n using the distribution P^n. For brevity, denote by Δ_i^ϑ the random variable $\delta_i^\vartheta(X_i)$. The product in the expression above can be expanded as

$$\prod_{i=1}^n(1+\Delta_i^\theta) = 1 + \sum_{i\in[n]}\Delta_i^\theta + \sum_{i_1>i_2}\Delta_{i_1}^\theta\Delta_{i_2}^\theta + \dots,$$

whereby we get

$$\begin{aligned}\chi^2(\mathbb{E}_\theta[Q_\theta^n], P^n) &= \mathbb{E}_{P^n}\Bigg[\Bigg(1+\sum_i\mathbb{E}_\theta\left[\Delta_i^\theta\right]+\\ &\qquad\sum_{i_1>i_2}\mathbb{E}_\theta\left[\Delta_{i_1}^\theta\Delta_{i_2}^\theta\right]+\dots\Bigg)^2\Bigg]-1\\ &= \mathbb{E}_{P^n}\Bigg[\sum_i\mathbb{E}_\theta\left[\Delta_i^\theta\right]+\sum_j\mathbb{E}_{\theta'}\left[\Delta_j^{\theta'}\right]\\ &\qquad+\sum_{i,j}\mathbb{E}_{\theta,\theta'}\left[\Delta_i^\theta\Delta_j^{\theta'}\right]+\dots\Bigg].\end{aligned}$$

Observe now that $\mathbb{E}_{P^n}\left[\Delta_i^\vartheta\right]=0$ for every ϑ. Furthermore, θ' is an independent copy of θ and Δ_i^θ and $\Delta_j^{\theta'}$ are independent for $i\neq j$. Therefore, the expectation on the right-side above equals

$$\begin{aligned}&\mathbb{E}\left[\sum_i H_i(\theta,\theta')+\sum_{i_1>i_2}H_{i_1}(\theta,\theta')H_{i_2}(\theta,\theta')+\dots\right]\\ &\quad=\mathbb{E}\left[\prod_{i=1}^n(1+H_i(\theta,\theta'))\right]-1,\end{aligned}$$

which completes the proof. □

Appendix C
Proofs of Induced Chi-Square Fluctuation Bounds

Proof of Lemma IV.7. The proof is nearly identical to that of Lemma III.5, with few additional observations. Using Fano's inequality Eq. (10) and following the proof of Lemma III.5, it suffices to derive the counterpart of Eq. (13). Note that by definition of $\mathcal{W}_0$, any public-coin protocol can be realized by using a shared randomness U, together with $W_1,\dots,W_n$ from $\mathcal{W}_0$. Thus, considering observations (Y^n, U) and proceeding as in Eq. (13),

$$\begin{aligned}I(Z\wedge Y^nU) &= I(Z\wedge Y^n\mid U)\\ &\leq \max_{W^n\in\mathcal{W}_0^n}\mathbb{E}[D(W^n\mathbf{p}_Z^n\|W^n\mathbf{p}^n)]\\ &\leq \max_{W^n\in\mathcal{W}_0^n}\sum_{i=1}^n\mathbb{E}[D(W_i\mathbf{p}_Z\|W_i\mathbf{p})]\\ &\leq \max_{W^n\in\mathcal{W}_0^n}\sum_{i=1}^n\mathbb{E}\left[\mathrm{d}_{\chi^2}(W_i\mathbf{p}_Z, W_i\mathbf{p})\right]\\ &\leq n\cdot\max_{W\in\mathcal{W}_0}\chi^2(W\mid\mathcal{P}),\end{aligned}$$

which completes the proof together with Eq. (10). □

Proof of Lemma IV.8. Consider an almost ε-perturbation $\mathcal{P}_\zeta$. The proof of this extension is very similar to the proof of Lemma III.7, except that $\mathbb{E}[\mathbf{p}_Z^n]$ and $\mathbf{p}^n$ get replaced with $\mathbb{E}[W^n\mathbf{p}_Z^n]$ and $W^n\mathbf{p}^n$, respectively. The first part of the argument goes through verbatim, leading to

$$\mathrm{d}_{\mathrm{TV}}(\mathbb{E}[W^n\mathbf{p}_Z^n], W^n\mathbf{p}^n)^2 \leq \exp\left(\chi^{(2)}\left(W^n\mid\mathcal{P}\right)\right)-1, \tag{30}$$

for every choice of channels $W^n=(W_1,\dots,W_n)$. In the second step, we need to get a lower bound on the left-side above, while restricting to W_is in $\mathcal{W}_0$. Towards that, consider an (n,ε)-test $\mathcal{T}$ using a public-coin protocol. Denoting by U the public randomness and by $Y_1,\dots,Y_n$ the messages from each player and by $\mathcal{Z}_0$ the set of z such that $\mathrm{d}_{\mathrm{TV}}(\mathbf{p}_z,\mathbf{p})\geq\varepsilon$. Since $\mathcal{P}_\zeta$ is an almost ε-perturbation, $\Pr[Z\in\mathcal{Z}_0]\geq\alpha\geq 1/10$. Also, for the test $\mathcal{T}$ we have $\Pr_{X^n\sim\mathbf{p}^n}[\mathcal{T}(U,Y^n)=1]\geq 11/12$ and $\Pr_{X^n\sim\mathbf{p}_z^n}[\mathcal{T}(U,Y^n)=1]\geq 11/12$ for every $z\in\mathcal{Z}_0$. Thus, in the manner of Eq. (29) we obtain

$$\begin{aligned}&\frac{1}{2}\Pr_{X^n\sim\mathbf{p}^n}[\mathcal{T}(U,Y^n)=1]+\frac{1}{2}\Pr_{X^n\sim\mathbb{E}[\mathbf{p}_Z^n]}[\mathcal{T}(U,Y^n)=0]\\ &\qquad\geq\frac{11(1+\alpha)}{24}\geq\frac{121}{240},\end{aligned}$$

where in the last inequality we used $\alpha\geq 1/10$. Then, we can find a fixed realization $U=u$ such that

$$\begin{aligned}&\frac{1}{2}\Pr_{X^n\sim\mathbf{p}^n}[\mathcal{T}(U,Y^n)\neq 1\mid U=u]\\ &\quad+\frac{1}{2}\Pr_{X^n\sim\mathbb{E}[\mathbf{p}_Z^n]}[\mathcal{T}(U,Y^n)\neq 0\mid U=u]\leq\frac{119}{240}.\end{aligned} \tag{31}$$

An important remark here is that u may depend on $\mathcal{P}_\zeta$. Observe that by definition of $\mathcal{W}_0$, we can emulate the public-coin protocols by each player selecting its channel $W_i\in\mathcal{W}_0$ as a function of the shared randomness U. Denote by $W_u^n\in\mathcal{W}_0^n$ the channels chosen by the players when $U=u$. Then, conditioned on $U = u$, Y^n has distribution $W_u^n \mathbf{p}^n$ and $W_u^n \mathbf{p}_z^n$, respectively, when X^n has distribution $\mathbf{p}^n$ and $\mathbf{p}_z^n$. Thus, as in the proof of Lemma III.7, we can find $W_u^n \in \mathcal{W}_0^n$ such that

$$\mathrm{d}_{\mathrm{TV}}(\mathbb{E}[W_u^n \mathbf{p}_Z^n], W_u^n \mathbf{p}^n) \geq \frac{1}{120},$$

which along with Eq. (30) yields

$$\chi^{(2)}\left(W_u^n \mid \mathcal{P}_\zeta\right) \geq c, \tag{32}$$

where $c = \log(14401/14400)$. The result follows upon taking the maximum over $W_u^n \in \mathcal{W}_0^n$ and minimum over all almost ε-perturbations $\mathcal{P}_\zeta$. □

Proof of Lemma IV.10. The argument follows the same template as the proof of Lemma IV.8, but with an important difference. Instead of derandomizing as in Eq. (31), which leads to a choice of channels W_u^n that may depend on perturbation $\mathcal{P}_\zeta$ family, now in Eq. (32) we would like to take the minimum over $\mathcal{P}_\zeta \in \Upsilon_\varepsilon$ first. Observe that for private-coin protocols, the effective channel used by each player is a convex combination of channels from $\mathcal{W}$, namely it is a channel from $\overline{\mathcal{W}}$. Thus, when X^n has distribution either $\mathbf{p}^n$ and $\mathbf{p}_z^n$, respectively, Y^n has distribution $W^n \mathbf{p}^n$ and $W^n \mathbf{p}_z^n$ with $W^n \in \overline{\mathcal{W}}^n$. Therefore, following the steps in the proof of Lemma IV.8, we get $\chi^{(2)}\left(W^n \mid \mathcal{P}_\zeta\right) \geq c$, where $W^n \in \overline{\mathcal{W}}^n$ and the almost ε-perturbation $\mathcal{P}_\zeta$ is arbitrary. The claim then follows by taking the minimum over $\mathcal{P}_\varepsilon$ and maximum over $W^n \in \overline{\mathcal{W}}^n$. □

Acknowledgments

We thank Gautam Kamath and Vitaly Feldman for useful pointers to related work and Ayfer Özgür for sharing with us a draft of [34]. We thanks Lekshmi Ramesh for comments on an early draft and Prathamesh Mayekar for pointing a typo in an earlier version of Definition I.1. We are also grateful to Mayank Bakshi for pointing out the connections to the AVC setting, and to Chandra Nair for an illuminating discussion on data processing inequalities for chi-square distances. Finally, we would like to thank the organizers of the 2018 Information Theory and Applications Workshop (ITA), where the collaboration leading to this work started.

References

[1] J. Acharya, C. L. Canonne, C. Freitag, Z. Sun, and H. Tyagi, "Inference under information constraints III: Local privacy constraints," 2019, in submission. Preprint available at arXiv:abs/1808.02174. I-B, V-B, VI

[2] J. Acharya, C. L. Canonne, C. Freitag, and H. Tyagi, "Test without trust: Optimal locally private distribution testing," in *AISTATS*, ser. Proceedings of Machine Learning Research, vol. 89. PMLR, 2019, pp. 2067–2076. V-B

[3] J. Acharya, C. L. Canonne, and H. Tyagi, "Inference under information constraints II: Communication constraints and shared randomness," *IEEE Transactions on Information Theory*, 2020, to appear. Preprint available at arXiv:abs/1804.06952. I-B, V-A, VI

[4] J. Acharya, Z. Sun, and H. Zhang, "Hadamard response: Estimating distributions privately, efficiently, and with little communication," ser. Proceedings of Machine Learning Research, K. Chaudhuri and M. Sugiyama, Eds., vol. 89. PMLR, 16–18 Apr 2019, pp. 1120–1129. [Online]. Available: http://proceedings.mlr.press/v89/acharya19a.html I-C, V-B

[5] R. Ahlswede and I. Csiszár, "Hypothesis testing with communication constraints," *IEEE Transactions on Information Theory*, vol. 32, no. 4, pp. 533–542, July 1986. I-C

[6] A. Andoni, T. Malkin, and N. Shekel Nosatzki, "Two party distribution testing: communication and security," in *46th International Colloquium on Automata, Languages, and Programming*, ser. LIPIcs. Leibniz Int. Proc. Inform. Schloss Dagstuhl. Leibniz-Zent. Inform., Wadern, 2019, vol. 132, pp. Art. No. 15, 16. I-C

[7] S. Balakrishnan and L. Wasserman, "Hypothesis testing for high-dimensional multinomials: A selective review," *The Annals of Applied Statistics*, vol. 12, no. 2, pp. 727–749, 2018. [Online]. Available: https://doi.org/10.1214/18-AOAS1155SF I-C

[8] A. R. Barron, "Uniformly powerful goodness of fit tests," *The Annals of Mathematical Statistics*, vol. 17, pp. 107–124, 1989. I-C

[9] A. Beimel, K. Nissim, and E. Omri, "Distributed private data analysis: Simultaneously solving how and what," in *Proceedings of the 28th Annual International Cryptology Conference*, ser. CRYPTO '08. Berlin, Heidelberg: Springer, 2008, pp. 451–468. I

[10] Q. Berthet and V. Kanade, "Statistical windows in testing for the initial distribution of a reversible markov chain," in *AISTATS*, ser. Proceedings of Machine Learning Research, vol. 89. PMLR, 2019, pp. 246–255. I

[11] E. Blais, C. L. Canonne, and T. Gur, "Distribution testing lower bounds via reductions from communication complexity," in *Computational Complexity Conference*, ser. LIPIcs, vol. 79. Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik, 2017, pp. 28:1–28:40. 12

[12] ——, "Distribution testing lower bounds via reductions from communication complexity," *ACM Trans. Comput. Theory*, vol. 11, no. 2, pp. Art. 6, 37, 2019, journal version of [11]. [Online]. Available: https://doi.org/10.1145/3305270 3

[13] S. Boucheron, G. Lugosi, and P. Massart, *Concentration Inequalities: A Nonasymptotic Theory of Independence*. OUP Oxford, 2013. III-B

[14] 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 *Symposium on Theory of Computing Conference, STOC'16*. ACM, 2016, pp. 1011–1020. I-C

[15] C. L. Canonne, *A Survey on Distribution Testing: Your Data is Big. But is it Blue?*, ser. Graduate Surveys. Theory of Computing Library, 2020, no. 9. [Online]. Available: http://www.theoryofcomputing.org/library.html I-C

[16] T. M. Cover and J. A. Thomas, *Elements of information theory*, 2nd ed. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, 2006. III-A

[17] I. Csiszár and J. Körner, *Information theory: Coding theorems for discrete memoryless channels. 2nd edition*. Cambridge University Press, 2011. I-C

[18] I. Diakonikolas, "Learning structured distributions," in *Handbook of Big Data*. CRC Press, 2016. I-C

[19] I. Diakonikolas, T. Gouleakis, D. M. Kane, and S. Rao, "Communication and memory efficient testing of discrete distributions," in *Proceedings of the 32nd Conference on Learning Theory, COLT 2019*, ser. Proceedings of Machine Learning Research, vol. 99. PMLR, 2019, pp. 1070–1106. I-C

[20] I. Diakonikolas, E. Grigorescu, J. Li, A. Natarajan, K. Onak, and L. Schmidt, "Communication-efficient distributed learning of discrete distributions," in *Advances in Neural Information Processing Systems 30*, 2017, pp. 6394–6404. I-C

[21] J. C. Duchi, M. I. Jordan, and M. J. Wainwright, "Local privacy and statistical minimax rates," in *54th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2013*. IEEE Computer Society, 2013, pp. 429–438. I, I-C, V-B

[22] J. C. Duchi and M. J. Wainwright, "Distance-based and continuum Fano inequalities with applications to statistical estimation," *ArXiV*, vol. abs/1311.2669, 2013. III-A

[23] C. Dwork, "Differential privacy: A survey of results," in *Theory and Applications of Models of Computation*. Springer, 2008, vol. 4978, pp. 1–19. I

[24] V. Feldman, "A general characterization of the statistical query complexity," in *Proceedings of the 30th Conference on Learning Theory, COLT 2017*, ser. Proceedings of Machine Learning Research, S. Kale and O. Shamir, Eds., vol. 65. Amsterdam, Netherlands: PMLR, 07–10 Jul 2017, pp. 785–830. I-C

[25] V. Feldman, E. Grigorescu, L. Reyzin, S. S. Vempala, and Y. Xiao, "Statistical algorithms and a lower bound for detecting planted cliques," *Journal of the ACM*, vol. 64, no. 2, pp. 8:1–8:37, 2017. [Online]. Available: https://doi.org/10.1145/3046674 I-C

[26] O. Fischer, U. Meir, and R. Oshman, "Distributed uniformity testing," in *Proceedings of the 2018 ACM Symposium on Principles of Distributed Computing, PODC 2018*. ACM, 2018, pp. 455–464. I-C

[27] S. Foucart and H. Rauhut, *A mathematical introduction to compressive sensing*, ser. Applied and Numerical Harmonic

Analysis. Birkhäuser/Springer, New York, 2013. [Online]. Available: https://doi.org/10.1007/978-0-8176-4948-7 A, A

[28] A. Garg, T. Ma, and H. L. Nguyen, "On communication cost of distributed statistical estimation and dimensionality," in *Advances in Neural Information Processing Systems 27*, 2014, pp. 2726–2734. I-C

[29] O. Goldreich, "The uniform distribution is complete with respect to testing identity to a fixed distribution," in *Computational Complexity and Property Testing - On the Interplay Between Randomness and Computation*, ser. Lecture Notes in Computer Science, O. Goldreich, Ed. Springer, 2020, vol. 12050, pp. 152–172. [Online]. Available: https://doi.org/10.1007/978-3-030-43662-9_10 I-A

[30] T. S. Han, "Hypothesis testing with multiterminal data compression," *IEEE Transactions on Information Theory*, vol. 33, no. 6, pp. 759–772, November 1987. I-C

[31] T. S. Han and S.-I. Amari, "Statistical inference under multiterminal data compression," *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2300–2324, October 1998. I-C

[32] 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. I-C

[33] Y. Han, A. Özgür, and T. Weissman, "Geometric Lower Bounds for Distributed Parameter Estimation under Communication Constraints," *ArXiv e-prints*, vol. abs/1802.08417v1, Feb. 2018, first version (https://arxiv.org/abs/1802.08417v1). I-C

[34] Y. Han, A. Özgür, and T. Weissman, "Geometric lower bounds for distributed parameter estimation under communication constraints," in *Proceedings of the 31st Conference on Learning Theory, COLT 2018*, ser. Proceedings of Machine Learning Research, vol. 75. PMLR, 2018, pp. 3163–3188. I-C, III-A, V-A, C

[35] H. Hotelling, "The consistency and ultimate distribution of optimum statistics," *Transactions of the American Mathematical Society*, vol. 32, no. 4, pp. 847–859, October 1930. I-C

[36] Y. I. Ingster, "A minimax test of nonparametric hypotheses on the density of a distribution in L_p metrics," *Teor. Veroyatnost. i Primenen.*, vol. 31, no. 2, pp. 384–389, 1986. III, III

[37] P. Kairouz, K. Bonawitz, and D. Ramage, "Discrete distribution estimation under local privacy," in *Proceedings of the 33rd International Conference on Machine Learning, ICML 2016*, ser. JMLR Workshop and Conference Proceedings, vol. 48. JMLR.org, 2016, pp. 2436–2444. I-C, V-B

[38] S. P. Kasiviswanathan, H. K. Lee, K. Nissim, S. Raskhodnikova, and A. Smith, "What can we learn privately?" in *49th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2008*. IEEE, Oct 25–28 2008, pp. 531–540. I, I-C

[39] H. B. Mann and A. Wald, "On the choice of the number of class intervals in the application of the chi square test," *The Annals of Mathematical Statistics*, vol. 13, pp. 306–317, 1942. I-C

[40] Y. I. Medvedev, "Separable statistics in a polynomial scheme. I," *Theory of Probability and Its Applications*, vol. 22, pp. 1–15, 1977. I-C

[41] 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. I-A, I-C, III, III, III.6

[42] D. Pollard, "Asymptopia," 2003, manuscript. [Online]. Available: http://www.stat.yale.edu/~pollard/Books/Asymptopia/ III, III, III-B

[43] R. Rubinfeld, "Taming big probability distributions," *XRDS: Crossroads, The ACM Magazine for Students*, vol. 19, no. 1, p. 24, sep 2012. [Online]. Available: http://dx.doi.org/10.1145/2331042.2331052 I-C

[44] K. R. Sahasranand and H. Tyagi, "Extra samples can reduce communication for independence testing," in *Proceedings of the 2018 IEEE International Symposium on Information Theory (ISIT'18)*. IEEE, 2018. I-C

[45] O. Shamir, "Fundamental limits of online and distributed algorithms for statistical learning and estimation," in *Advances in Neural Information Processing Systems 27*, 2014, pp. 163–171. I-C

[46] O. Sheffet, "Locally private hypothesis testing," in *Proceedings of the 35th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, J. Dy and A. Krause, Eds., vol. 80. Stockholmsmässan, Stockholm Sweden: PMLR, 10–15 Jul 2018, pp. 4612–4621. I-C

[47] J. Steinhardt, G. Valiant, and S. Wager, "Memory, communication, and statistical queries," in *Proceedings of the 29th Conference on Learning Theory, COLT 2016*, ser. Proceedings of Machine Learning Research, V. Feldman, A. Rakhlin, and O. Shamir, Eds., vol. 49. New York, New York, USA: PMLR, 23–26 Jun 2016, pp. 1490–1516. I-C

[48] S. Szarek, "On the best constants in the khinchin inequality," *Studia Mathematica*, vol. 58, no. 2, pp. 197–208, 1976. IV-B

[49] J. N. Tsitsiklis, "Decentralized detection," in *Advances in Statistical Signal Processing*, H. V. Poor and J. B. Thomas, Eds., vol. 2. JAI Press, 1993, pp. 297–344. I-C

[50] G. Valiant and P. Valiant, "An automatic inequality prover and instance optimal identity testing," *SIAM Journal on Computing*, vol. 46, no. 1, pp. 429–455, 2017. 3, I-C

[51] R. Viswanathan and P. Varshney, "Distributed detection with multiple sensors: Part I – Fundamentals," *Proceedings of IEEE*, vol. 85, no. 1, pp. 54–63, January 1997. I-C

[52] S. Wang, L. Huang, P. Wang, Y. Nie, H. Xu, W. Yang, X. Li, and C. Qiao, "Mutual information optimally local private discrete distribution estimation," *ArXiV*, vol. abs/1607.08025, 2016. I-C, V-B

[53] M. Wigger and R. Timo, "Testing against independence with multiple decision centers," *IEEE International Conference on Signal Processing and Communications, IISc, Bangalore*, June 2016. I-C

[54] Y. Xiang and Y. H. Kim, "Interactive hypothesis testing against independence," in *Proceedings of the 2013 IEEE International Symposium on Information Theory (ISIT'13)*, 2013, pp. 1782–1786. I-C

[55] A. Xu and M. Raginsky, "Information-theoretic lower bounds on Bayes risk in decentralized estimation," *IEEE Transactions on Information Theory*, vol. 63, no. 3, pp. 1580–1600, 2017. I-C

[56] M. Ye and A. Barg, "Optimal schemes for discrete distribution estimation under locally differential privacy," *IEEE Transactions on Information Theory*, vol. 64, no. 8, pp. 5662–5676, 2018. I-C, V-B

[57] B. Yu, "Assouad, Fano, and Le Cam," in *Festschrift for Lucien Le Cam*. Springer, 1997, pp. 423–435. [Online]. Available: http://dx.doi.org/10.1007/978-1-4612-1880-7_29 III-A

[58] Y. Zhang, J. Duchi, M. I. Jordan, and M. J. Wainwright, "Information-theoretic lower bounds for distributed statistical estimation with communication constraints," in *Advances in Neural Information Processing Systems 26*, 2013, pp. 2328–2336. I-C

Jayadev Acharya (M., 2014) is an assistant professor in the School of Electrical and Computer Engineering at Cornell University. He received the Bachelor of Technology degree in Electronics and Electrical Communication Engineering from the Indian Institute of Technology, Kharagpur in 2007, and M.S. (2009) and Ph.D degree (2014) in Electrical and Computer Engineering from the University of California, San Diego. He was a postdoctoral associate in Electrical Engineering and Computer Science at Massachusetts Institute of Technology from 2014 to 2016.

Clément Canonne is a Goldstine Postdoctoral Fellow at IBM Research, and an upcoming Lecturer at the School of Computer Science of the University of Sydney, Australia. Prior to this, he was a Motwani Postdoctoral Fellow at Stanford University, 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.

Himanshu Tyagi (S'04–M'14–SM'19) received the B.Tech. degree in electrical engineering and the M.Tech. degree in communication and information technology, both from the Indian Institute of Technology, Delhi, India in 2007. He received the Ph.D. degree from the University of Maryland, College Park in 2013. From 2013 to 2014, he was a postdoctoral researcher at the Information Theory and Applications (ITA) Center, University of California, San Diego. Since January 2015, he has been a faculty member at the Department of Electrical Communication Engineering, Indian Institute of Science in Bangalore. His research interests broadly lie in information theory and its application in cryptography, statistics, machine learning, and computer science. Also, he is interested in communication and automation for city-scale systems.