Optimal Private and Communication Constraint Distributed Goodness-of-Fit Testing for Discrete Distributions in the Large Sample Regime

Lasse Vuursteen
Department of Statistics and Data Science
The Wharton School of the University of Pennsylvania
Philadelphia, PA 19104
`lassev@wharton.upenn.edu`

Abstract

We study distributed goodness-of-fit testing for discrete distribution under bandwidth and differential privacy constraints. Information constraint distributed goodness-of-fit testing is a problem that has received considerable attention recently. The important case of discrete distributions is theoretically well understood in the classical case where all data is available in one "central" location. In a federated setting, however, data is distributed across multiple "locations" (e.g. servers) and cannot readily be shared due to e.g. bandwidth or privacy constraints that each server needs to satisfy. We show how recently derived results for goodness-of-fit testing for the mean of a multivariate Gaussian model extend to the discrete distributions, by leveraging Le Cam's theory of statistical equivalence. In doing so, we derive matching minimax upper- and lower-bounds for the goodness-of-fit testing for discrete distributions under bandwidth or privacy constraints in the regime where number of samples held locally are large.

Keywords: distributed inference, goodness-of-fit testing, differential privacy, communication constraint, federated learning, statistical equivalence.

1 Introduction

Federated learning is a fundamental problem in statistics and machine learning, where data is distributed across multiple locations (e.g. servers) and cannot readily be shared due to e.g. bandwidth or privacy constraints that each server needs to satisfy. The primary goal in these distributed data settings is to perform a single global inference task, such as hypothesis testing, regression, or classification, by aggregating the local information from each server.

Starting a few decades ago, investigations into distributed settings with bandwidth and other information constraints originated in the electrical engineering community, under the names "decentralized decision theory / the CEO problem" e.g. [90, 12, 92, 18, 58, 87] or "inference under multiterminal compression" (see [88] for an overview). These were largely motivated by applications where data is by construction observed and processed locally, such as astronomy, meteorology, seismology, surveillance systems, wireless communication, military radar or air traffic control systems.

Modern federated learning often involves data distributed across siloed data centers (e.g., hospitals) or networks of cellphone users, applied in areas such as word prediction, facial and voice recognition, virtual assistants like Siri or Google Assistant, autonomous vehicles, and earthquake prediction [65, 56, 51, 73, 67, 31]. In these settings, bandwidth often becomes a limited or costly resource [57].

Similarly, with advances in electronic record keeping, privacy has become a more and more pressing issue. These issues are prominent in tech industry products [32], including many federated learning

applications mentioned earlier, as well as in scientific fields like medical sciences [60] and social sciences [69].

Methods that preserve privacy have been around in the statistics community for some time, starting in the 1980's [40, 41]. The current leading formal privacy framework is that of *differential privacy* (DP), as introduced in [42]. DP is a mathematical guarantee, describing whether results or data sets can be considered "privacy preserving" and hence can be openly published. Whilst many other privacy frameworks exist, this notion of privacy holds a prominent position both theoretically and practically, finding application within industry giants like Google [45], Microsoft [35], Apple [89], as well as governmental entities such as the US Census Bureau [74].

Quantifying the trade-off between privacy and statistical power means that researchers and data analysts can make an appropriate balance between data privacy and meaningful analysis. Similarly, by quantifying the impact of bandwidth constraints, systems can be designed to work as efficiently as possible within such bandwidth constraints.

The performance of distributed inference under bandwidth or differential privacy is well-studied for various estimation problems. For instance, distributed estimation under differential privacy has been studied for the many-normal-means model, discrete distributions and parametric models in [37, 38, 1, 98, 3], and density estimation [75, 59, 21], and nonparametric regression in [23]. Bandwidth constraints have been studied for the many-normal-means and parametric models in e.g. [101, 39, 77, 20, 97, 50, 26, 25], as well as nonparametric models, including Gaussian white noise [102], nonparametric regression [82], density estimation [17, 4], general, abstract settings [99] and online learning [95]. Distributed adaptive estimation methods under bandwidth constraints, where adaptation occurs to the unknown regularity of the functional parameter of interest, were derived in [82, 83, 25]. Testing simple hypotheses under bandwidth constraints has been studied by e.g. [92] and under differential privacy constraints by [29].

In this paper, we consider goodness-of-fit testing for discrete distributions (i.e. the multinomial model) in scenarios where the number of samples received by each server is large. Specifically, we study testing a simple null versus a composite alternative, in the setting where m servers receive n observations each from a distribution on a sample space of cardinality d, where n is large comparatively to m and d. Recently, such multinomial distributed data have found many applications in areas that handle very large samples over (possibly also large) discrete domains. For example, in population genetics [71, 86] and computer science; where it is used for e.g. information retrieval [100, 76], speech and text and classification [55], text mining [24] and large language models [72]. This has sparked recent interest in studying the statistical decision theoretic properties of the multinomial model, see [15] for an overview.

Deriving minimax rates for goodness-of-fit testing of discrete distributions under bandwidth and differential privacy constraints is particularly challenging when each server holds multiple observations. To date, matching rates have been established only when each machine observes a single observation [9, 10, 5] (see also our discussion of related work below). The techniques used to derive lower-bounds in the aforementioned paper heavily rely on the fact that each server contains only one observation, see [9]. Moreover, whilst tight lower-bounds for the multiple observations case exist for the Gaussian model, the functional analytic techniques used to derive these results heavily rely on Gaussianity, see [85] and [22] for the respective bandwidth constraint and DP lower-bounds. Additionally, lower-bound techniques developed for estimation problems generally do not yield tight impossibility results for goodness-of-fit testing problems (see also the discussion in Section B of the appendix).

We derive matching upper- and lower-bounds for goodness-of-fit testing for discrete distributions under bandwidth and differential privacy constraints in scenarios where the number of samples n held by each of the servers is large in comparison to d and m; $md \log d/\sqrt{n} = o(1)$. This is achieved by leveraging the theory of statistical equivalence, as introduced by Le Cam (see e.g. [62, 81] for an introduction). Leveraging existing results concerning statistical equivalence of multinomial data with a multivariate Gaussian model proven in [30] allow us to show, roughly speaking, that the distributed goodness-of-fit testing problem for discrete distributions is statistically equivalent a distributed goodness-of-fit testing problem for the mean of a multivariate Gaussian model, and hence the minimax rate for the former problem is the same as the minimax rate for the latter problem, which was established in [85] and [22], for bandwidth and differential privacy constraints respectively. Furthermore, we exploit the bandwidth constraint distributed setting in which these two models have

different minimax rates to show that, when n is small compared to d and m, the multinomial model and multivariate Gaussian model are statistically non-equivalent.

The rest of the paper is organized as follows. After a brief section on related work and notation, the article continues with a precise problem formulation in Section 2. Section 2.1 outlines the distributed framework, for both bandwidth and differential privacy constraints. In Section 2.2, we introduce the problem of distributed goodness-of-fit testing for discrete distributions under bandwidth and differential privacy constraints. In Section 3, the main results concerning the minimax rates are presented. Section 4 briefly outlines the main idea of the proof technique. Section 5 gives further insight into the comparison between discrete distributions and its comparable multivariate Gaussian model. The article ends with a brief discussion of the derived results. In the appendix, the tools of asymptotic equivalence within the distributed framework are presented and the technical proofs are provided.

1.1 Related work

Minimax goodness-of-fit testing knows a rich literature within the statistics and machine learning communities, see [46, 53, 64, 80, 52]. The d-ary discrete distribution uniformity testing problem bares a close relationship with "classical" nonparametric goodness-of-fit testing in the sense of [14, 79, 33, 96] and other nonparametric testing problems, see Section 1.4 in [54] and references therein.

For distributed goodness-of-fit testing specifically, much less is known. For multivariate Gaussian models under communication or differential privacy constraints, solutions have been established for the case where each server holds multiple observations. Communication constraints have been studied in [8, 84, 85] and differential privacy constraints in [22], with the authors deriving matching minimax upper- and lower-bounds for the goodness-of-fit testing for the mean of a multivariate Gaussian model.

For testing in discrete distributions, only the scenario where each server receives just one observation has been fully characterized in terms of the minimax rate in [9, 10, 5]. See also [28] for an overview. In these aforementioned works, the authors derive minimax rates goodness-of-fit testing for discrete distributions under bandwidth and differential privacy constraints. See [48, 78, 11, 2, 19] for investigations specifically under local DP (i.e. one observation per server with DP constraint). Nonparametric goodness-of-fit density testing for under local DP is considered in [36, 61], where in [61], the authors consider adaptation as well. For some investigations into the multiple observations per server case, see [34, 47].

For estimation, the bandwidth constraint estimation discrete distributions in the large sample-per-server case has been studied by [3], who derive matching upper and lower-bounds up to logarithmic factors. However, their technique does not extend to the goodness-of-fit testing problem.

1.2 Notation and notions

Throughout this paper, we shall use the following notation. For two positive sequences a_k and b_k, we use $a_k \lesssim b_k$ to mean that $a_k \leq C b_k$ for some universal positive constant C. We write $a_k \asymp b_k$ if both $a_k \lesssim b_k$ and $b_k \lesssim a_k$, and $a_k \ll b_k$ if $a_k/b_k = o(1)$.

We denote the maximum of a and b by $a \vee b$ and the minimum by $a \wedge b$. For $k \in \mathbb{N}$, $[k]$ represents the set $\{1, \ldots, k\}$. Universal constants c and C may vary between lines. The Euclidean norm of a vector $v \in \mathbb{R}^d$ is denoted by $\|v\|_2$. For a matrix $M \in \mathbb{R}^{d \times d}$, $\|M\|$ represents the spectral norm, and $\mathrm{Tr}(M)$ denotes its trace. I_d is the $d \times d$ identity matrix.

A non-negative sequence M_k is said to be of poly-logarithmic order in non-negative sequences a_k, b_k, c_k if there exists a constant $c > 0$ such that $M_k \lesssim (\log(a_k) \log(b_k) \log(c_k))^c$.

Given measurable spaces $(\mathcal{X}, \mathscr{X})$ and $(\mathcal{Y}, \mathscr{Y})$, a *Markov kernel K (between $(\mathcal{X}, \mathscr{X})$ and target $(\mathcal{Y}, \mathscr{Y})$)* is a map $K \equiv K(\cdot|\cdot) : \mathscr{Y} \times \mathcal{X} \to [0, 1]$ with the following two properties: The map $x \mapsto K(A|x)$ is measurable for all $A \in \mathscr{Y}$, and the map $A \mapsto K(A|x)$ is a probability measure on $\mathscr{Y}$ for every $x \in \mathcal{X}$.

If S is a random variable on a probability space $(\mathcal{X}, \mathscr{X}, \mathbb{P})$, we let $\mathbb{P}^S$ denote its *push-forward measure*, i.e. the measure defined by $\mathbb{P}^S(B) := \mathbb{P}(S^{-1}(B))$. We shall use $\mathbb{E}$ and $\mathbb{E}^S$ as the expectation

operator corresponding to $\mathbb{P}$ and $\mathbb{P}^S$. Random variables X, Y, Z form a *Markov chain $X \to Y \to Z$* whenever their joint distribution $\mathbb{P}^{(X,Y,Z)}$ disintegrates as $d\mathbb{P}^{(X,Y,Z)} = d\mathbb{P}^X d\mathbb{P}^{Y|X} d\mathbb{P}^{Z|Y}$.

2 Problem formulation

We begin by formally introducing the general framework of distributed inference.

2.1 The distributed framework

Consider a measurable space $(\mathcal{X}, \mathscr{X})$ with a statistical model $\mathcal{P} = \{P_f \ : \ f \in \mathcal{F}\}$ defined on it. In the distributed framework, we consider $j = 1, \ldots, m$ servers, each receiving data $X^{(j)}$ drawn from a given distribution $P_f \in \mathcal{P}$. Each of the servers communicates a transcript $Y^{(j)}$ based on the data to a central server, which in turn computes its solution to the testing problem $T(Y) \in \{0, 1\}$ based on the aggregated transcripts $Y = (Y^{(1)}, \ldots, Y^{(m)})$. We shall use the convention that $T(Y) = 1$ means rejecting the null hypothesis. The transcript generating mechanisms are then given by Markov kernels $\{K^j\}_{j=1,\ldots,m}$, with the Markov kernel (i.e. conditional distribution) of the transcript $Y^{(j)}$ given the data $X^{(j)}$ and the randomness U shared by the servers denoted by $K^j(\cdot | X^{(j)}, U)$. We formalize this in the following definition.

Definition 1. *A* distributed testing protocol *for the model $\mathcal{P}$ consists of a triplet* $\{T, \{K^j\}_{j=1}^m, (\mathcal{U}, \mathscr{U}, \mathbb{P}^U)\}$*, where $\{K^j\}_{j=1}^m$ is a collection of Markov kernels $K^j : \mathcal{Y}^{(j)} \times \mathcal{X} \times \mathcal{U} \to [0, 1]$ defined on a measurable space $(\mathcal{Y}^{(j)}, \mathscr{Y}^{(j)})$, $T : \bigotimes_{j=1}^m \mathcal{Y}^{(j)} \to \{0, 1\}$ is a measurable map and $(\mathcal{U}, \mathscr{U}, \mathbb{P}^U)$ is probability space.*

The probability space $(\mathcal{U}, \mathscr{U}, \mathbb{P}^U)$ is used to (possibly) generate a source of randomness (independent of the data) that is shared by the servers. The distributed protocol is said to have *no access to shared randomness* or to be a *local randomness protocol* if $\mathbb{P}^U$ is trivial[1]. In an abuse of notation, we shall often refer to the entire triplet $\{T, \{K^j\}_{j=1,\ldots,m}, (\mathcal{U}, \mathscr{U}, \mathbb{P}^U)\}$ using just T.

Given a distributed protocol and i.i.d. data from P_f we shall use $\mathbb{P}_f$ to denote the joint distribution of $Y = (Y^{(1)}, \ldots, Y^{(m)})$, the data X under P_f^m and the shared randomness $U \sim \mathbb{P}^U$. Writing $x = (x^{(1)}, \ldots, x^{(m)}) \in \mathcal{X}^m$, let $x \mapsto K(A|x, u)$ denote the Markov kernel $\bigotimes_{j=1}^m K^j(\cdot | x^{(j)}, u)$ (i.e. the product measure). The independence structure of the data yields that $P_f^m K = \bigotimes_{j=1}^m P_f K^j$ and the push-forward measure of Y can be seen to disintegrate as

$$\mathbb{P}_f^Y(A) = P_f^m \mathbb{P}^U K(A) = \mathbb{P}^U P_f^m K(A) = \int \int K(A|x, u) dP_f^m(x) d\mathbb{P}^U(u),$$

where the second equality follows from the independence of U with the data drawn from P_f. The above disintegration of the push-forward measure of Y and the product structure of K can be interpreted as $(X, Y, T(Y))$ forming a Markov chain given U, in the sense of the diagram

$$
\begin{array}{ccccc}
X^{(1)} & \longrightarrow & Y^{(1)}|U & \searrow & \\
\vdots & \longrightarrow & \vdots & \longrightarrow & T(Y). \\
X^{(m)} & \longrightarrow & Y^{(m)}|U & \nearrow &
\end{array}
\tag{1}
$$

The diagram indicates the flow of dependencies. The m servers each obtain data $X^{(j)}$ from P_f, and generate a transcript $Y^{(j)}$ based on the data and shared randomness U. The central server then makes a decision $T(Y)$ based on the aggregated transcripts Y. For a definition of Markov kernels and Markov chains, see Section 1.2.

Allowing transcript-generating mechanisms to access both shared and local randomness is important for our analysis, as shared randomness has been found to yield strictly better performance in distributed goodness-of-fit testing, see e.g. [9, 10, 5, 8, 84, 85, 22]. Shared randomness protocols can be seen as a subset of common interactive procedures, such as sequential and blackboard protocols

---

[1] $\mathscr{U}$ is the trivial sigma-algebra (meaning $U \sim \mathbb{P}^U$ is a degenerate random variable)

(see e.g. [6]). The aforementioned paper shows that for discrete distribution goodness-of-fit testing in the single observation per server case, sequential and blackboard protocols offer no benefit over shared randomness protocols. Similarly, for mean shift problems in the multivariate Gaussian case, no advantage of sequential protocols over shared randomness protocols is known, except in the case of estimation with unknown variance [**?**]. Since we study goodness-of-fit testing for discrete distributions in the large-number-of-observations case by comparing with a Gaussian model with known variance, we restrict the setting of the main article to local and shared randomness protocols only. Nevertheless, our theoretical framework is general enough to handle interactive protocols, which we discuss in Section A.2 of the appendix.

Next, we introduce the notion of a bandwidth constraint in the distributed setting.

Definition 2. *A distributed protocol is said to satisfy a* b-*bit bandwidth constraint if its kernels* $\{K^j\}_{j=1,\ldots,m}$ *are defined on measurable spaces* $(\mathcal{Y}^{(j)}, \mathscr{Y}^{(j)})$ *satisfying* $|\mathcal{Y}^{(j)}| \leq 2^b$ *for* $j = 1, \ldots, m$.

We use $\mathscr{T}_{\mathrm{LR}}^{(b)}$ and $\mathscr{T}_{\mathrm{SR}}^{(b)}$ to denote the classes of all local randomness and shared randomness distributed testing protocols with communication budget b per machine, respectively.

Lastly, we introduce the notion of differential privacy in the distributed setting. We will be focusing on the notion of differential privacy as put forward by [43, 44]. Differential privacy provides a mathematical framework that guarantees preservation of privacy in a notion akin to cryptographical guarantees. Formally, a differential privacy constraint on a transcript in our setting is formulated as follows.

Definition 3. *Let* $\epsilon \geq 0, \delta \geq 0$. *The transcript* $Y^{(j)}$ *generated from* $K^j, u \in \mathcal{U}$ *is said to be* (ϵ, δ)-*differentially private if*

$$K^j(A|x, u) \leq e^\epsilon K^j(A|x', u) + \delta \tag{2}$$

$$\text{for all } A \in \mathscr{Y}^{(j)}, \ x, x' \in \mathcal{X}, \ i \in \{1, \ldots, n\}.$$

A distributed testing protocol $\{T, \{K^j\}_{j=1}^m, (\mathcal{U}, \mathscr{U}, \mathbb{P}^U)\}$, *is said be a* distributed (ϵ, δ)-*differentially private testing protocol if* $\{K^j\}_{j=1,\ldots,m}$ *satisfies* (2) $\mathbb{P}^U$-*a.s.*

Small values of ϵ and δ ensure that, even when the transcript $Y^{(j)}$ is publicly available, the sample $X^{(j)}$ underlying $Y^{(j)}$ is unidentifiable. We stress that this type of differential privacy guarantee concerns the local data $X^{(j)}$ in full, even $X^{(j)}$ consists of multiple observations. This is often referred *local differential privacy*, where the privacy guarantee regards each server as essentially pertaining data to "one indiviual". For a thorough introduction on differential privacy guarantees, we refer the reader to [42]. We also note that the use of shared randomness does not affect the privacy guarantee provided by the protocol, as the guarantee holds even if the outcome of the shared randomness is known.

We use $\mathscr{T}_{\mathrm{LR}}^{(\epsilon,\delta)}$ and $\mathscr{T}_{\mathrm{SR}}^{(\epsilon,\delta)}$ to denote the classes of all local- and shared randomness (ϵ, δ)-differentially private distributed testing protocols, respectively. We note that the machinery developed in Section A.2 allows consideration of both types of constraints simultaneously. In the main text of the article, we shall focus on the bandwidth constraint and differential privacy constraint separately as minimax rates for the joint constraints are not known for the Gaussian model we use for comparison to the multinomial model in the main article.

2.2 Distributed goodness-of-fit testing

We start by giving a formal description of sampling from a discrete distribution in the distributed setting. Consider a set with cardinality d; for simplicity, we take $\tilde{\mathcal{X}} = \{1, \ldots, d\}$. Any probability distribution such a set can be characterized by an element of the $d - 1$-dimensional probability simplex $\mathbb{S}^d$, defined as

$$\left\{ q = (q_1, \ldots, q_d) \in [0, 1]^d : \sum_{i=1}^d q_i = 1 \right\}.$$

In our distributed framework, each server $j = 1, \ldots, m$ observes a data $\tilde{X}^{(j)}$ taking values in $\{1, \ldots, d\}^n$

$$\tilde{X}^{(j)} = (\tilde{X}_1^{(j)}, \ldots, \tilde{X}_n^{(j)}) \sim Q \equiv Q_{n,q}, \quad \tilde{X}_i^{(j)} \overset{i.i.d.}{\sim} \text{Multinomial}(1, q) \text{ for } q \in \mathbb{S}^d. \tag{3}$$

That is, each server obtains n i.i.d. draws from a multinomial distribution with parameter q.

The statistical decision problem of interest shall be that of *goodness-of-fit* or *uniformity testing*, i.e. distinguishing the hypotheses

$$H_0 : q = q_0 \quad \text{versus } H_1 : q \in \{q \in \mathcal{F} : \|q - q_0\|_1 \geq \rho\} =: H_\rho, \tag{4}$$

where $q_0 = (q_{01}, \ldots, q_{0d}) = (1/d, \ldots, 1/d) \in \mathbb{S}^d$ and

$$\mathcal{F} = \left\{ q \in \mathbb{S}^d \ : \ \frac{\max_i q_i}{\min_i q_i} \leq R \right\}, \tag{5}$$

for some fixed constant $R > 0$. The statistical model under consideration shall be denoted by $\mathcal{Q} = \{Q_q^n : q \in \mathcal{F}\}$.

We define the testing risk for a distributed testing protocol T, for the hypotheses (4) (and statistical model $\mathcal{Q}$) by sum of the type I and worst case type II error over the alternative class;

$$\mathcal{R}_\mathcal{Q}(T, H_\rho) := \mathbb{Q}_{q_0}^Y T(Y) + \sup_{f \in H_\rho} \mathbb{Q}_f^Y (1 - T(Y)).$$

The minimax testing risk over a class of distributed protocols $\mathscr{T}$ is then defined as $\inf_{T \in \mathscr{T}} \mathcal{R}_\mathcal{Q}(T, H_\rho)$.

It is clear that, as ρ tends to 0, the minimax testing risk should increase. We are interested in finding the so called *minimax separation rate*, or detection boundary, which is a sequence ρ^* depending on the model characteristics n, d, m and $\mathscr{T}$ such that the minimax testing risk converges to 0 if $\rho \ll \rho^*$ or 1 if $\rho \gg \rho^*$.

The minimax separation rate captures how the testing problem becomes easier, or more difficult, for different model characteristics. The minimax rate for the hypothesis above case is $\rho^2 \asymp \frac{\sqrt{d}}{mn}$ when $\mathscr{T}$ consists of the class of all testing protocols, as was established in [70] and [94].

When $\mathscr{T}$ is taken to be one of the bandwidth or privacy constraint classes of tests, i.e. $\mathscr{T}_{\mathrm{LR}}^{(b)}$ and $\mathscr{T}_{\mathrm{SR}}^{(b)}$ $\mathscr{T}_{\mathrm{LR}}^{(\epsilon,\delta)}$ and $\mathscr{T}_{\mathrm{SR}}^{(\epsilon,\delta)}$, it is sensible to expect ρ^* to depend on the bandwidth or differential privacy parameters, b and (ϵ, δ), respectively. In the distributed discrete distribution setupj described above with $n = 1$, such minimax rates have been derived in [9, 10]. We discuss these results in the next section, contrasting them with the minimax separation rate derived in this paper for the case where $md \log d / \sqrt{n} = o(1)$.

3 Minimax rates in the large sample regime

We now turn to the main results of this paper, which concern the minimax rates for goodness-of-fit testing for discrete distributions under bandwidth and differential privacy constraints in the large sample regime. We shall show that the minimax rates for the distributed multinomial model under bandwidth and differential privacy constraints are the same as the minimax rates for a d-dimensional distributed Gaussian model, as derived in [85] and [22], respectively.

The first theorem establishes the minimax rate for the distributed multinomial model under bandwidth constraints. A proof can be found in Section D of the appendix.

Theorem 1. *Consider sequences $m \equiv m_\nu$, $b \equiv b_\nu$, $d \equiv d_\nu$ and $n \equiv n_\nu$ such that $md \to \infty$ whilst*

$$md \log d / \sqrt{n} \overset{\nu \to \infty}{\to} 0.$$

Suppose that $\rho \equiv \rho_\nu$ is a nonnegative sequence satisfying

$$\rho^2 \asymp \left(\frac{d}{\sqrt{d \wedge b} mn} \right) \bigwedge \left(\frac{\sqrt{d}}{\sqrt{m} n} \right). \tag{6}$$

Then,

$$\inf_{T \in \mathscr{T}_{SR}^{(b)}} \mathcal{R}_\mathcal{Q}(T, H_{M_\nu \rho}) \to \begin{cases} 0 \ \textit{for any } M_\nu \to \infty, \\ 1 \ \textit{for any } M_\nu \to 0. \end{cases}$$

When considering the class of only local randomness protocols (i.e. replacing $\mathscr{T}_{SR}^{(b)}$ with $\mathscr{T}_{LR}^{(b)}$ in the above display), the minimax separation rate is given by

$$\rho^2 \asymp \left(\frac{d^{3/2}}{(d \wedge b)mn} \right) \bigwedge \left(\frac{\sqrt{d}}{\sqrt{mn}} \right). \tag{7}$$

The theorem above shows that the minimax rate for the distributed multinomial model under bandwidth constraints is given by (6) in the case of access to shared randomness, and (7) in the case of no access to shared randomness. Both rates are the same as those established for a signal detection problem in a d-dimensional distributed Gaussian model, as derived in [85], Theorems 3.1 and 3.2. In Section 4, we shall provide a proof of this result through the notion of statistical equivalence, where we explicitly use that the multinomial model is asymptotically similar to a specific Gaussian model and a corresponding signal detection problem.

The distributed b-bit bandwidth constraint minimax rate for the hypotheses (4) in the multinomial model with $n = 1$ is established in [9, 10]. Specifically, they find that

$$\rho^2 \asymp \begin{cases} \frac{d}{m\sqrt{2^b \wedge d}} & \text{in case of access to shared randomness,} \\ \frac{d\sqrt{d}}{m(2^b \wedge d)} & \text{without access to shared randomness.} \end{cases} \tag{8}$$

Several aspects of this minimax rate are intriguing. First, unlike in the "large n case" for the same model and hypothesis ((6) and (7)), there is no elbow effect. Secondly, the benefit (i.e. "efficiency gain") from an increase in bandwidth is exponential, whereas in the large sample scenario of Theorem 4 it is sub-linear. We shall comment on this "communication super-efficiency" phenomenon further below.

We now turn to the distributed multinomial model under differential privacy constraints. As in the case of the bandwidth constraint uniformity testing problem, we shall show that the minimax rate for the distributed multinomial model under differential privacy constraints is the same as the minimax rate for a d-dimensional distributed Gaussian model, as derived in [22].

The following theorem describes that the above rates are the minimax rates for uniformity testing in the distributed multinomial model under differential privacy constraints, for shared randomness and local randomness only protocols, respectively.

Theorem 2. *For any sequences $m \equiv m_\nu$, $d \equiv d_\nu$ and $n \equiv n_\nu$ such that $md \to \infty$, $\frac{md \log d}{\sqrt{n}} \overset{\nu \to \infty}{\to} 0$, $n^{-1/4} \ll \epsilon \equiv \epsilon_\nu \leq 1$, $\delta \equiv \delta_\nu \asymp (md)^{-p}$ for some $p > 1$. The minimax separation rate in the distributed multinomial model $\mathcal{Q}$ for testing the hypotheses (4) using locally (ϵ, δ)-differentially private protocols is*

$$\rho^2 \asymp \text{poly-log}(d, m, n) \begin{cases} \frac{d}{mn\epsilon^2} & \text{if } \epsilon \geq \frac{\sqrt{d}}{\sqrt{m}}, \\ \frac{\sqrt{d}}{\sqrt{mn}\epsilon} & \text{if } \frac{1}{\sqrt{md}} \leq \epsilon < \frac{\sqrt{d}}{\sqrt{m}} \end{cases} \tag{9}$$

in the case of having access to shared randomness. In the case of having only access to local randomness, it is given by

$$\rho^2 \asymp \text{poly-log}(d, m, n) \begin{cases} \frac{d\sqrt{d}}{mn\epsilon^2} & \text{if } \epsilon \geq \frac{d}{\sqrt{m}}, \\ \frac{\sqrt{d}}{\sqrt{mn}\epsilon} & \text{if } \frac{1}{\sqrt{md}} \leq \epsilon < \frac{d}{\sqrt{m}}. \end{cases} \tag{10}$$

We provide a proof of the theorem in Section D of the appendix. As with the bandwidth constraint case, the minimax separation rates for the distributed multinomial model under differential privacy constraints are derived by comparing the model and hypothesis test to a signal detection problem for the d-dimensional distributed Gaussian model. The rates for the latter problem follow from the proofs of Theorems 4 and 5 in [22], who describe a more general setup which includes signal detection in the d-dimensional distributed Gaussian model as a special case[2].

Also in the case of privacy, there is a difference between the one observation per server case minimax rate ($n = 1$) and the multiple observations per server with local differential privacy case. The

---

[2]The above rates do not observe all phase transitions present in the more general setup of [22], as it pertains to the local differential privacy case in that paper, with $\sigma = 1/\sqrt{n}$ in their notation.

minimax rate in the multinomial model for $n = 1$ is derived in [9, 5];

$$\rho^2 \asymp \begin{cases} \frac{d}{m\epsilon^2} & \text{in case of access to shared randomness,} \\ \frac{d\sqrt{d}}{m\epsilon^2} & \text{without access to shared randomness.} \end{cases} \tag{11}$$

Comparing this rate to the rate obtained in Theorem 2, we observe phase transitions in the distributed testing problem for multinomial model under local differential privacy constraints which only occurs if the number of observations locally is large compared to the cardinality of the sample space.

4 Deriving the minimax rates through statistical equivalence

The minimax rates for the distributed multinomial model under bandwidth and differential privacy constraints are derived through the notion of statistical equivalence (Le Cam theory), which is a powerful tool for establishing minimax rates in statistical decision theoretic problems. In this section, we shall provide a brief introduction to statistical equivalence, and show how it can be used to derive the minimax rates for the distributed multinomial model under bandwidth and differential privacy constraints. Further details on the statistical equivalence and a detailed proof are deferred Section A of the appendix.

Le Cam theory is a general framework for decision problems. At the core of this theory is the notion of a distance between statistical models, known as Le Cam's deficiency distance. The objective of this distance is to quantify the extent to which a complex statistical model can be approximated by a more simple one. If a model is close to another model in Le Cam's distance, then there is a mapping of solutions to decision theoretic problems from one model to the other. Whenever the risk of the decision problem is bounded, this means that similar performance can be achieved in the two models. Consequently, studying the complex model can be reduced to studying the corresponding simple model. For an extensive introduction to Le Cam theory, see e.g. [62, 81]. For a brief introduction; [63, 66].

Consider a model $\mathcal{P} = \{P_f : f \in \mathcal{F}\}$ (a collection of probability distributions) on a measurable space $(\mathcal{X}, \mathscr{X})$ (the sample space). For this article, we consider only models with Polish sample spaces and corresponding Borel sigma-algebras and dominated models, meaning that there exists a sigma-finite measure μ such that $P_f \ll \mu$ for all $f \in \mathcal{F}$. This greatly simplifies the definition of deficiency, given next.

Given another model $\mathcal{Q} = \{Q_f : f \in \mathcal{F}\}$ indexed by the same set $\mathcal{F}$ and sample space $(\tilde{\mathcal{X}}, \tilde{\mathscr{X}})$, we define the *deficiency of $\mathcal{P}$ with respect to $\mathcal{Q}$* as

$$\mathfrak{d}(\mathcal{P}; \mathcal{Q}) = \inf_C \sup_{f \in \mathcal{F}} \|P_f C - Q_f\|_{\mathrm{TV}}. \tag{12}$$

where the infimum is taken over all Markov kernels $C : \tilde{\mathscr{X}} \times \mathcal{X} \to [0, 1]$ and the probability measure $P_f C : \tilde{\mathscr{X}} \to [0, 1]$ is understood as $P_f C(A) := \int_{x \in \mathcal{X}} C(A|x) dP_f(x)$. This is equivalent to the more general notion of deficiency of [27] for dominated models on Polish spaces (see Proposition 9.2 in [68]).

Le Cam's deficiency distance between $\mathcal{P}$ and $\mathcal{Q}$ is then defined as $\Delta(\mathcal{P}, \mathcal{Q}) = \max\{\mathfrak{d}(\mathcal{P}; \mathcal{Q}), \mathfrak{d}(\mathcal{Q}, \mathcal{P})\}$. This semi-metric becomes a metric whenever $\mathcal{P}$ and $\mathcal{Q}$ are identified whenever $\mathfrak{d}(\mathcal{P}; \mathcal{Q}) + \mathfrak{d}(\mathcal{Q}, \mathcal{P}) = 0$. Two sequences of experiments $\mathcal{P}_\nu$ and $\mathcal{Q}_\nu$ are called *asymptotically equivalent* if their difference $\Delta(\mathcal{P}_\nu, \mathcal{Q}_\nu)$ tends to zero as ν approaches infinity. Conversely, such sequences shall be called *asymptotically nonequivalent* if $\Delta(\mathcal{P}_\nu, \mathcal{Q}_\nu) > c$ as $\nu \to \infty$ for a fixed constant $c > 0$.

In Section A.2, we prove that models that are close in the Le Cam metric (compared to m) have similar testing risks in the distributed setup. We leverage this result in combination with the fact that the distributed multinomial model is asymptotically equivalent to a d-dimensional distributed Gaussian model, which we describe next.

Consider for $q \in \mathcal{F}$ and $i = 1, \ldots, d$ the random variables

$$X_i^{(j)} = \sqrt{q_i} + \frac{1}{\sqrt{2n}} Z_i^{(j)} \quad \text{with} \quad Z^{(j)} = (Z_1^{(j)}, \ldots, Z_d^{(j)}) \sim N(0, I_d). \tag{13}$$

Let $P_f \equiv P_f^n$ denote the distribution of $X^{(j)} = (X_1^{(j)}, \ldots, X_d^{(j)})$. Let $\mathcal{P}$ denote the corresponding experiment. It is shown in [30] that $\mathcal{Q}$ is close to $\mathcal{P}$ in the Le Cam metric when d is relatively small compared to n. More precisely, it follows from Theorem 1 and Section 7 in [30] that

$$\Delta(\mathcal{P}, \mathcal{Q}) \leq C_R \frac{d \log d}{\sqrt{n}}, \tag{14}$$

where $C_R > 0$ is a constant depending only on R. For the testing problem in Gaussian model, with hypotheses (4), the minimax rate can be derived using the results of [85] in case of bandwidth constraints and [22] in case of differential privacy constraints. The key tool from which the minimax rates can then be derived for the multinomial model is the following lemma, which allows comparison of the minimax testing risks for the multinomial and Gaussian models in regimes where the Le Cam distance is small. Its proof is given in Section A.2 of the appendix.

Lemma 1. *Suppose $m\Delta(\mathcal{Q}; \mathcal{P}) \leq \varrho$ for $\varrho > 0$. Then, it holds that*

$$\left| \inf_{T \in \mathscr{T}(\mathcal{P})} \mathcal{R}_\mathcal{P}(T, H_1) - \inf_{T \in \mathscr{T}(\mathcal{Q})} \mathcal{R}_\mathcal{Q}(T, H_1) \right| \leq 2\varrho,$$

where $\mathscr{T}$ is either $\mathscr{T}_{SR}^b$, $\mathscr{T}_{LR}^b$, $\mathscr{T}_{SR}^{(\epsilon, \delta)}$ or $\mathscr{T}_{LR}^{(\epsilon, \delta)}$.

5 Statistical non-equivalence of discrete and multivariate Gaussian distributions

Theorem 1 describing uniformity testing in the large sample regime and the result derived for $n = 1$, as displayed in (8), shows a striking difference terms of the role of the communication budget. Specifically, in the $n = 1$ regime, an exponential communication efficiency is observed, whereas in the large sample regime, the benefit is only linear. In this section, we shall provide some explanation for this phenomenon, and shall actually leverage this difference to show that the distributed multinomial model and the distributed Gaussian model are asymptotically non-equivalent: Two models are considered *asymptotically nonequivalent* if their Le Cam distance remains bounded away from zero, even as the amount of data increases in both models.

The multinomial model is equivalent to a model in which one observes $N^{(j)} = (N_1^{(j)}, \ldots, N_d^{(j)})$ taking values in $\{1, \ldots, n\}^d$, where $N_k^{(j)} \equiv N_k^{(j)}\left(\tilde{X}^{(j)}\right) = \left|\left\{i : \tilde{X}_i^{(j)} = k\right\}\right|$. Let $\mathcal{Q}'$ denote the model generated by the observations $N^{(j)}$. This model is equivalent to $\mathcal{Q}$, meaning $\Delta(\mathcal{Q}, \mathcal{Q}') = 0$. To see this, note that for $x = (x_1, \ldots, x_n) \in \{1, \ldots, d\}^n$,

$$Q\left(\tilde{X}^{(j)} = x\right) = \prod_{i=1}^n Q\left(\tilde{X}_i^{(j)} = x_i\right) = \prod_{k \in \{1, \ldots, d\}} q_k^{|\{i : x_i = k\}|} \text{ for all } Q \in \mathcal{Q},$$

after which the aforementioned equivalence follows by the Neyman-Fisher factorization criterion, e.g. Lemma 2 in the appendix. When n is large compared to d, one could standardize the count statistics $N^{(j)}$ to obtain a statistic that tends towards a d-dimensional Gaussian random vector. When d and m are not too large with respect to n, one can obtain transcripts and corresponding test statistics from these approximately Gaussian vectors, that resemble those one would consider in the Gaussian model, and attain the corresponding minimax rates.

Since the observation $N^{(j)}$ takes values in $\{1, \ldots, n\}^d$, the full data can be transmitted whenever there are at least $d \log_2 n$-bits are available per server. However, recalling that the observation $\tilde{X}^{(j)}$ takes values in the space $\{1, \ldots, d\}^n$, which has cardinality bounded above by $2^{n \log_2 d}$, we also obtain that the full data can be transmitted whenever $n \log_2 d$-bits are available. Consequently, whenever

$$b \gtrsim d \log_2(n + 1) \wedge n \log_2 d,$$

the distributed problem has the same minimax separation rate for the hypothesis in (4) as the unconstrained problem with nm observations; $\rho_\mathcal{Q}^2 \asymp \frac{\sqrt{d}}{mn}$. For the Gaussian problem, this is only the case whenever $b \gtrsim d$, as can be seen from Theorem 4. This indicates a kind of "tipping point" occuring whenever n gets small compared to d, where in a bandwidth constraint distributed setting, the testing problem in for the Gaussian model starts to exhibit very different behavior.

Interestingly, this does not imply that the multinomial model is "easier" from a distributed testing under bandwidth constraints perspective, as there are sub-regimes in which the Gaussian model has a solution whereas the multinomial model does not and vice versa. It indicates that the "communication complexity" of the sample space matters in the respective decision problems. We can leverage this fact to obtain a lower-bound on the Le Cam distance between the multinomial model and the Gaussian model; which is the content of the next theorem.

Theorem 3. *There exists constants $C > 0$ and $c > 0$ such that for any $n, d \in \mathbb{N}$ with*

$$\frac{d}{n \log(d)} \geq C \quad and \quad n \geq \sqrt{d} \log(d), \tag{15}$$

it holds that

$$\mathfrak{d}(\mathcal{Q}, \mathcal{P}) \geq c, \tag{16}$$

where $\mathcal{P}$ is the experiment generated by the observations in (13)*, $\mathcal{Q}$ is generated according to* (3)*, both indexed by $\mathcal{F}$ as given in* (5)*.*

The proof of the theorem is given in Section D. It leverages that there exist distributed, b-bit bandwidth constraint settings in which the (distributed) multinomial model allows for consistent goodness-of-fit testing, whereas the (distributed) Gaussian model does not. The result then readily follows from the distributed equivalence results derived in Section A.2. The fact that the separation in the respective (distributed) testing risks occurs for a constant number of servers, yields that the two models are asymptotically nonequivalent whenever $\sqrt{d}/\log^2(d) \geq d/n \gg 1$. This reasoning crucially exploits the differing minimax rates that occur under the bandwidth constraint, since without such a constraint, the same goodness-of-fit testing problem of (4) would have similar minimax performance for both of the models.

6 Discussion

We have derived minimax separation rates for uniformity testing in the distributed multinomial model under bandwidth and differential privacy constraints, in the large sample regime where $md \log d/\sqrt{n} = o(1)$. When contrasted with existing results for large sample regimes, the minimax rates show that the large sample regime is subject to distinctly different phenomena.

The applicability of our results is somewhat constrained by the requirement that $md \log d/\sqrt{n} = o(1)$, which limits the range of model characteristics we can consider. Consequently, further work is needed to understand the behavior of the distributed multinomial model in other regimes. The non-equivalence result in Theorem 3 indicates that the distributed multinomial model and the distributed Gaussian model are fundamentally different regarding distributed statistical decision problems when the sample size is small. Therefore, direct analysis of the distributed multinomial model might be necessary, requiring new techniques to derive minimax rates. We note, however, that this pertains to the specific Gaussian model formulated in (13), and there might be a different Gaussian model that is equivalent to the distributed multinomial model even in the small n regime.

The results in this paper are derived through the notion of statistical equivalence, which is a powerful tool for establishing minimax rates in statistical decision theoretic problems. The results and techniques can be applied more generally to other distributed inference problems, and proving more general results concerning statistical equivalence and distributed inference is an interesting avenue for future research.

A downside of leveraging statistical equivalence is that it generally does not provide a direct path to obtain methods that are minimax rate optimal. However, Theorem 1 and Section 7 in [30] provide a specific transformation that converts the local multinomial sample into a statistic approximately distributed as a Gaussian random vector. Such a transformation, combined with the rate optimal methods given in [85] and [22], provide guidance to construct methods that attain the minimax rates described in this article.

Acknowledgments and Disclosure of Funding

References

[1] ACHARYA, J., BONAWITZ, K., KAIROUZ, P., RAMAGE, D., AND SUN, Z. Context aware local differential privacy. In *Proceedings of the 37th International Conference on Machine Learning* (13–18 Jul 2020), H. D. III and A. Singh, Eds., vol. 119 of *Proceedings of Machine Learning Research*, PMLR, pp. 52–62.

[2] ACHARYA, J., CANONNE, C., FREITAG, C., AND TYAGI, H. Test without trust: Optimal locally private distribution testing. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics* (16–18 Apr 2019), K. Chaudhuri and M. Sugiyama, Eds., vol. 89 of *Proceedings of Machine Learning Research*, PMLR, pp. 2067–2076.

[3] ACHARYA, J., CANONNE, C., LIU, Y., SUN, Z., AND TYAGI, H. Distributed estimation with multiple samples per user: Sharp rates and phase transition. In *Advances in Neural Information Processing Systems* (2021), M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., vol. 34, Curran Associates, Inc., pp. 18920–18931.

[4] ACHARYA, J., CANONNE, C., SINGH, A. V., AND TYAGI, H. Optimal rates for nonparametric density estimation under communication constraints. *Advances in Neural Information Processing Systems 34* (2021), 26754–26766.

[5] ACHARYA, J., CANONNE, C. L., FREITAG, C., SUN, Z., AND TYAGI, H. Inference under information constraints iii: Local privacy constraints. *IEEE Journal on Selected Areas in Information Theory 2*, 1 (2021), 253–267.

[6] ACHARYA, J., CANONNE, C. L., LIU, Y., SUN, Z., AND TYAGI, H. Interactive inference under information constraints. *IEEE Transactions on Information Theory 68*, 1 (2022), 502–516.

[7] ACHARYA, J., CANONNE, C. L., SUN, Z., AND TYAGI, H. Unified lower bounds for interactive high-dimensional estimation under information constraints. *Advances in Neural Information Processing Systems 36* (2024).

[8] ACHARYA, J., CANONNE, C. L., AND TYAGI, H. Distributed signal detection under communication constraints. In *Conference on Learning Theory* (2020), PMLR, pp. 41–63.

[9] ACHARYA, J., CANONNE, C. L., AND TYAGI, H. Inference under information constraints i: Lower bounds from chi-square contraction. *IEEE Transactions on Information Theory 66*, 12 (2020), 7835–7855.

[10] ACHARYA, J., CANONNE, C. L., AND TYAGI, H. Inference under information constraints ii: Communication constraints and shared randomness. *IEEE Transactions on Information Theory 66*, 12 (2020), 7856–7877.

[11] ACHARYA, J., SUN, Z., AND ZHANG, H. Differentially private testing of identity and closeness of discrete distributions. *Advances in Neural Information Processing Systems 31* (2018).

[12] AHLSWEDE, R., AND CSISZÁR, I. Hypothesis testing with communication constraints. *IEEE transactions on information theory 32*, 4 (1986), 533–542.

[13] AHN, S., CHEN, W.-N., AND ÖZGÜR, A. Estimating sparse distributions under joint communication and privacy constraints. In *2022 IEEE International Symposium on Information Theory (ISIT)* (2022), pp. 3144–3149.

[14] AN, K. Sulla determinazione empirica di una legge didistribuzione. *Giorn Dell'inst Ital Degli Att 4* (1933), 89–91.

[15] BALAKRISHNAN, S., AND WASSERMAN, L. Hypothesis testing for high-dimensional multinomials: A selective review. *The Annals of Applied Statistics 12*, 2 (2018), 727 – 749.

[16] BARNES, L. P., HAN, Y., AND ÖZGÜR, A. Fisher information for distributed estimation under a blackboard communication protocol. In *2019 IEEE International Symposium on Information Theory (ISIT)* (2019), IEEE, pp. 2704–2708.

[17] BARNES, L. P., HAN, Y., AND ÖZGÜR, A. Lower bounds for learning distributions under communication constraints via fisher information. *The Journal of Machine Learning Research 21*, 1 (2020), 9583–9612.

[18] BERGER, T., AND ZHANG, Z. On the ceo problem. In *Proceedings of 1994 IEEE International Symposium on Information Theory* (1994), pp. 201–.

[19] BERRETT, T., AND BUTUCEA, C. Locally private non-asymptotic testing of discrete distributions is faster using interactive mechanisms. *Advances in Neural Information Processing Systems 33* (2020), 3164–3173.

[20] BRAVERMAN, M., GARG, A., MA, T., NGUYEN, H. L., AND WOODRUFF, D. P. Communication lower bounds for statistical estimation problems via a distributed data processing inequality. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing* (2016), pp. 1011–1020.

[21] BUTUCEA, C., DUBOIS, A., KROLL, M., AND SAUMARD, A. Local differential privacy: Elbow effect in optimal density estimation and adaptation over Besov ellipsoids. *Bernoulli 26*, 3 (2020), 1727 – 1764.

[22] CAI, T. T., CHAKRABORTY, A., AND VUURSTEEN, L. Federated nonparametric hypothesis testing with differential privacy constraints: Optimal rates and adaptive tests, 2024. arXiv:2406.06749 [math.ST].

[23] CAI, T. T., CHAKRABORTY, A., AND VUURSTEEN, L. Optimal federated learning for nonparametric regression with heterogeneous distributed differential privacy constraints, 2024. arXiv:2406.06755 [math.ST].

[24] CAI, T. T., KE, Z. T., AND TURNER, P. Testing high-dimensional multinomials with applications to text analysis. *Journal of the Royal Statistical Society Series B: Statistical Methodology 86*, 4 (02 2024), 922–942.

[25] CAI, T. T., AND WEI, H. Distributed adaptive gaussian mean estimation with unknown variance: Interactive protocol helps adaptation. *The Annals of Statistics 50*, 4 (2022), 1992–2020.

[26] CAI, T. T., AND WEI, H. Distributed gaussian mean estimation under communication constraints: Optimal rates and communication-efficient algorithms. *Journal of Machine Learning Research 25*, 37 (2024), 1–63.

[27] CAM, L. L. Sufficiency and Approximate Sufficiency. *The Annals of Mathematical Statistics 35*, 4 (1964), 1419 – 1455.

[28] CANONNE, C. L., ET AL. Topics and techniques in distribution testing: A biased but representative sample. *Foundations and Trends® in Communications and Information Theory 19*, 6 (2022), 1032–1198.

[29] CANONNE, C. L., KAMATH, G., MCMILLAN, A., SMITH, A., AND ULLMAN, J. The structure of optimal private tests for simple hypotheses. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing* (2019), pp. 310–321.

[30] CARTER, A. V. Deficiency distance between multinomial and multivariate normal experiments. *The Annals of Statistics 30*, 3 (2002), 708 – 730.

[31] COCHRAN, E. S., LAWRENCE, J. F., CHRISTENSEN, C., AND JAKKA, R. S. The quake-catcher network: Citizen science expanding seismic horizons. *Seismological Research Letters 80*, 1 (2009), 26–30.

[32] CRAIN, M. The limits of transparency: Data brokers and commodification. *new media & society 20*, 1 (2018), 88–104.

[33] CRAMÉR, H. On the composition of elementary errors: First paper: Mathematical deductions. *Scandinavian Actuarial Journal 1928*, 1 (1928), 13–74.

[34] DIAKONIKOLAS, I., GOULEAKIS, T., KANE, D. M., AND RAO, S. Communication and memory efficient testing of discrete distributions. In *Proceedings of the Thirty-Second Conference on Learning Theory* (Phoenix, USA, 25–28 Jun 2019), A. Beygelzimer and D. Hsu, Eds., vol. 99 of *Proceedings of Machine Learning Research*, PMLR, pp. 1070–1106.

[35] DING, B., KULKARNI, J., AND YEKHANIN, S. Collecting telemetry data privately. *Advances in Neural Information Processing Systems 30* (2017).

[36] DUBOIS, A., BERRETT, T., AND BUTUCEA, C. Goodness-of-Fit Testing for Hölder Continuous Densities Under Local Differential Privacy. In *Foundations of Modern Statistics*, vol. PROMS-425 of *Springer Proceedings in Mathematics & Statistics*. Springer International Publishing, 2023, pp. 53–119.

[37] DUCHI, J. C., JORDAN, M. I., AND WAINWRIGHT, M. J. Local privacy and statistical minimax rates. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science* (2013), IEEE, pp. 429–438.

[38] DUCHI, J. C., JORDAN, M. I., AND WAINWRIGHT, M. J. Minimax optimal procedures for locally private estimation. *Journal of the American Statistical Association 113*, 521 (2018), 182–201.

[39] DUCHI, J. C., JORDAN, M. I., WAINWRIGHT, M. J., AND ZHANG, Y. Optimality guarantees for distributed statistical estimation. *arXiv preprint arXiv:1405.0782* (2014).

[40] DUNCAN, G., AND LAMBERT, D. The risk of disclosure for microdata. *Journal of Business & Economic Statistics 7*, 2 (1989), 207–217.

[41] DUNCAN, G. T., AND PEARSON, R. W. Enhancing access to microdata while protecting confidentiality: Prospects for the future. *Statistical Science 6*, 3 (1991), 219–232.

[42] DWORK, C. Differential privacy. In *International colloquium on automata, languages, and programming* (2006), Springer, pp. 1–12.

[43] DWORK, C., MCSHERRY, F., NISSIM, K., AND SMITH, A. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3* (2006), Springer, pp. 265–284.

[44] DWORK, C., AND SMITH, A. Differential privacy for statistics: What we know and what we want to learn. *Journal of Privacy and Confidentiality 1*, 2 (2010).

[45] ERLINGSSON, U., PIHUR, V., AND KOROLOVA, A. Rappor: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security* (New York, NY, USA, 2014), CCS '14, Association for Computing Machinery, p. 1054–1067.

[46] ERMAKOV, M. Asymptotically minimax tests for nonparametric hypotheses concerning the distribution density. *Journal of Soviet Mathematics 52* (1990), 2891–2898.

[47] FISCHER, O., MEIR, U., AND OSHMAN, R. Distributed uniformity testing. In *Proceedings of the 2018 ACM Symposium on Principles of Distributed Computing* (New York, NY, USA, 2018), PODC '18, Association for Computing Machinery, p. 455–464.

[48] GABOARDI, M., LIM, H., ROGERS, R., AND VADHAN, S. Differentially private chi-squared hypothesis testing: Goodness of fit and independence testing. In *Proceedings of The 33rd International Conference on Machine Learning* (New York, New York, USA, 20–22 Jun 2016), M. F. Balcan and K. Q. Weinberger, Eds., vol. 48 of *Proceedings of Machine Learning Research*, PMLR, pp. 2111–2120.

[49] HAN, Y., MUKHERJEE, P., OZGUR, A., AND WEISSMAN, T. Distributed statistical estimation of high-dimensional and nonparametric distributions. In *2018 IEEE International Symposium on Information Theory (ISIT)* (2018), IEEE, pp. 506–510.

[50] HAN, Y., ÖZGÜR, A., AND WEISSMAN, T. Geometric lower bounds for distributed parameter estimation under communication constraints. In *Conference On Learning Theory* (2018), PMLR, pp. 3163–3188.

[51] HARD, A., RAO, K., MATHEWS, R., RAMASWAMY, S., BEAUFAYS, F., AUGENSTEIN, S., EICHNER, H., KIDDON, C., AND RAMAGE, D. Federated learning for mobile keyboard prediction. *arXiv preprint arXiv:1811.03604* (2018).

[52] INGSTER, Y., AND SUSLINA, I. A. *Nonparametric goodness-of-fit testing under Gaussian models*, vol. 169. Springer Science & Business Media, 2003.

[53] INGSTER, Y. I. Asymptotically minimax hypothesis testing for nonparametric alternatives. i, ii, iii. *Math. Methods Statist 2*, 2 (1993), 85–114.

[54] INGSTER, Y. I., AND SUSLINA, I. A. *Nonparametric Goodness-of-Fit Testing Under Gaussian Models*, vol. 169 of *Lecture Notes in Statistics*. Springer New York, New York, NY, 2003.

[55] JURAFSKY, D., AND MARTIN, J. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, vol. 2. 02 2008.

[56] KONEČNÝ, J., MCMAHAN, H. B., RAMAGE, D., AND RICHTÁRIK, P. Federated optimization: Distributed machine learning for on-device intelligence. *arXiv preprint arXiv:1610.02527* (2016).

[57] KONEČNÝ, J., MCMAHAN, H. B., YU, F. X., RICHTÁRIK, P., SURESH, A. T., AND BACON, D. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492* (2016).

[58] KREIDL, O. P., TSITSIKLIS, J. N., AND ZOUMPOULIS, S. I. On Decentralized Detection With Partial Information Sharing Among Sensors. *IEEE Transactions on Signal Processing 59*, 4 (Apr. 2011), 1759–1765. Number: 4.

[59] KROLL, M. On density estimation at a fixed point under local differential privacy. *Electronic Journal of Statistics 15*, 1 (2021), 1783 – 1813.

[60] KULYNYCH, J., AND GREELY, H. T. Clinical genomics, big data, and electronic medical records: reconciling patient rights with research when privacy and science collide. *Journal of Law and the Biosciences 4*, 1 (2017), 94–132.

[61] LAM-WEIL, J., LAURENT, B., AND LOUBES, J.-M. Minimax optimal goodness-of-fit testing for densities and multinomials under a local differential privacy constraint. *Bernoulli 28*, 1 (2022), 579–600.

[62] LE CAM, L. *Asymptotic methods in statistical decision theory*. Springer Science & Business Media, 2012.

[63] LE CAM, L., AND YANG, G. L. *Asymptotics in statistics: some basic concepts*. Springer Science & Business Media, 2000.

[64] LEPSKII, O. Asymptotically minimax adaptive estimation. i: Upper bounds. optimally adaptive estimates. *Theory of Probability & Its Applications 36*, 4 (1992), 682–697.

[65] LI, T., SAHU, A. K., TALWALKAR, A., AND SMITH, V. Federated learning: Challenges, methods, and future directions. *IEEE signal processing magazine 37*, 3 (2020), 50–60.

[66] MARIUCCI, E. Le cam theory on the comparison of statistical models. *arXiv preprint arXiv:1605.03301* (2016).

[67] NGUYEN, A., DO, T., TRAN, M., NGUYEN, B. X., DUONG, C., PHAN, T., TJIPUTRA, E., AND TRAN, Q. D. Deep federated learning for autonomous driving. In *2022 IEEE Intelligent Vehicles Symposium (IV)* (2022), IEEE, pp. 1824–1830.

[68] NUSSBAUM, M. Asymptotic equivalence of density estimation and Gaussian white noise. *The Annals of Statistics 24*, 6 (1996), 2399 – 2430.

[69] OBERSKI, D. L., AND KREUTER, F. Differential privacy and social science: An urgent puzzle. *Harvard Data Science Review 2*, 1 (2020), 1–21.

[70] PANINSKI, L. A Coincidence-Based Test for Uniformity Given Very Sparsely Sampled Discrete Data. *IEEE Transactions on Information Theory 54*, 10 (Oct. 2008), 4750–4755.

[71] PRITCHARD, J. K., STEPHENS, M., AND DONNELLY, P. Inference of population structure using multilocus genotype data. *Genetics 155*, 2 (2000), 945–959.

[72] RADFORD, A., WU, J., CHILD, R., LUAN, D., AMODEI, D., SUTSKEVER, I., ET AL. Language models are unsupervised multitask learners. *OpenAI blog 1*, 8 (2019), 9.

[73] RAMASWAMY, S., MATHEWS, R., RAO, K., AND BEAUFAYS, F. Federated learning for emoji prediction in a mobile keyboard, 2019.

[74] RODRIGUEZ, I. M., SEXTON12, W. N., SINGER, P. E., AND VILHUBER, L. The modernization of statistical disclosure limitation at the us census bureau.

[75] SART, M. Density estimation under local differential privacy and Hellinger loss. *Bernoulli 29*, 3 (2023), 2318 – 2341.

[76] SCHÜTZE, H., MANNING, C. D., AND RAGHAVAN, P. *Introduction to information retrieval*, vol. 39. Cambridge University Press Cambridge, 2008.

[77] SHAMIR, O. Fundamental limits of online and distributed algorithms for statistical learning and estimation. *Advances in Neural Information Processing Systems 27* (2014), 163–171.

[78] SHEFFET, O. Locally private hypothesis testing. In *International Conference on Machine Learning* (2018), PMLR, pp. 4605–4614.

[79] SMIRNOV, N. Table for Estimating the Goodness of Fit of Empirical Distributions. *The Annals of Mathematical Statistics 19*, 2 (1948), 279 – 281.

[80] SPOKOINY, V. G. Adaptive hypothesis testing using wavelets. *The Annals of Statistics 24*, 6 (Dec. 1996).

[81] STRASSER, H. *Mathematical theory of statistics: statistical experiments and asymptotic decision theory*. No. 7 in De Gruyter studies in mathematics. W. de Gruyter, Berlin ; New York, 1985.

[82] SZABO, B., AND VAN ZANTEN, H. Adaptive distributed methods under communication constraints. *The Annals of Statistics 48*, 4 (2020), 2347–2380.

[83] SZABÓ, B., AND VAN ZANTEN, H. Distributed function estimation: Adaptation using minimal communication. *Mathematical Statistics and Learning 5*, 3 (2022), 159–199.

[84] SZABÓ, B., VUURSTEEN, L., AND VAN ZANTEN, H. Optimal distributed composite testing in high-dimensional gaussian models with 1-bit communication. *IEEE Transactions on Information Theory 68*, 6 (2022), 4070–4084.

[85] SZABÓ, B., VUURSTEEN, L., AND VAN ZANTEN, H. Optimal high-dimensional and nonparametric distributed testing under communication constraints. *The Annals of Statistics 51*, 3 (2023), 909–934.

[86] TAI, Y. C., AND SPEED, T. P. A multivariate empirical Bayes statistic for replicated microarray time course data. *The Annals of Statistics 34*, 5 (2006), 2387 – 2412.

[87] TARIGHATI, A., GROSS, J., AND JALDEN, J. Decentralized Hypothesis Testing in Energy Harvesting Wireless Sensor Networks. *IEEE Transactions on Signal Processing 65*, 18 (Sept. 2017), 4862–4873.

[88] TE SUN HAN, AND AMARI, S. Statistical inference under multiterminal data compression. *IEEE Transactions on Information Theory 44*, 6 (Oct. 1998), 2300–2324. Number: 6.

[89] TEAM, A. D. P. Learning with privacy at scale.

[90] TENNEY, R. R., AND SANDELL, N. R. Detection with Distributed Sensors. *IEEE Transactions on Aerospace and Electronic Systems AES-17*, 4 (July 1981), 501–510. Number: 4.

[91] THORISSON, H. *Coupling, Stationarity, and Regeneration*. Probability and Its Applications. Springer New York, 2000.

[92] TSITSIKLIS, J. N. Decentralized detection by a large number of sensors. *Mathematics of Control, Signals, and Systems 1*, 2 (June 1988), 167–182. Number: 2.

[93] TSYBAKOV, A. B. *Introduction to nonparametric estimation*. Springer series in statistics. Springer, New York ; London, 2009. OCLC: ocn300399286.

[94] VALIANT, G., AND VALIANT, P. An automatic inequality prover and instance optimal identity testing. *SIAM Journal on Computing 46*, 1 (2017), 429–455.

[95] VAN DER HOEVEN, D., HADIJI, H., AND VAN ERVEN, T. Distributed online learning for joint regret with communication constraints. In *Proceedings of The 33rd International Conference on Algorithmic Learning Theory* (29 Mar–01 Apr 2022), S. Dasgupta and N. Haghtalab, Eds., vol. 167 of *Proceedings of Machine Learning Research*, PMLR, pp. 1003–1042.

[96] VON MISES, R. Statistik und wahrheit. *Julius Springer 20* (1928).

[97] XU, A., AND RAGINSKY, M. Information-Theoretic Lower Bounds on Bayes Risk in Decentralized Estimation. *arXiv:1607.00550 [cs, math, stat]* (July 2016). arXiv: 1607.00550.

[98] YE, M., AND BARG, A. Optimal schemes for discrete distribution estimation under locally differential privacy. *IEEE Transactions on Information Theory 64*, 8 (2018), 5662–5676.

[99] ZAMAN, A., AND SZABÓ, B. Distributed nonparametric estimation under communication constraints. *arXiv preprint arXiv:2204.10373* (2022).

[100] ZHAI, C., AND LAFFERTY, J. A study of smoothing methods for language models applied to information retrieval. *ACM Transactions on Information Systems (TOIS) 22*, 2 (2004), 179–214.

[101] ZHANG, Y., DUCHI, J., JORDAN, M. I., AND WAINWRIGHT, M. J. Information-theoretic lower bounds for distributed statistical estimation with communication constraints. *Advances in Neural Information Processing Systems 26* (2013).

[102] ZHU, Y., AND LAFFERTY, J. Distributed nonparametric regression under communication constraints. In *Proceedings of the 35th International Conference on Machine Learning* (10–15 Jul 2018), J. Dy and A. Krause, Eds., vol. 80 of *Proceedings of Machine Learning Research*, PMLR, pp. 6009–6017.

A   Le Cam theory in distributed setting

We introduce some formal notions of Le Cam theory first in Section A.1. Then, in Section A.2, we study the equivalence of models in the distributed setting. The theoretical developments presented in this section apply to general models denoted by $\mathcal{P}$ and $\mathcal{Q}$; although the main text specifically focuses on the Gaussian location model for $\mathcal{P}$ and the multinomial model for $\mathcal{Q}$, the machinery developed here is applicable to general statistical models.

A.1   Preliminary notions of Le Cam theory

A *statistical experiment* is a set of probability distributions $\mathcal{P} = \{P_f : f \in \mathcal{F}\}$ (a model) on a measurable space $(\mathcal{X}, \mathscr{X})$ (the sample space). For the purpose of simplification, we shall consider only statistical experiments with Polish sample spaces and corresponding Borel sigma-algebras. Furthermore, we shall only consider dominated models, meaning that there exists a sigma-finite measure μ such that $P_f \ll \mu$ for all $f \in \mathcal{F}$. In a slight abuse of terminology, we shall sometimes refer to $\mathcal{P}$ as the experiment, suppressing the presence of the sample space and indexing set.

Given another statistical experiment with model $\mathcal{Q} = \{Q_f : f \in \mathcal{F}\}$ indexed by the same set $\mathcal{F}$ and sample space $(\tilde{\mathcal{X}}, \tilde{\mathscr{X}})$, we define the *deficiency of $\mathcal{P}$ with respect to $\mathcal{Q}$* as

$$\mathfrak{d}(\mathcal{P}; \mathcal{Q}) = \inf_C \sup_{f \in \mathcal{F}} \|P_f C - Q_f\|_{\mathrm{TV}}. \tag{17}$$

where the infimum is taken over all Markov kernels $C : \tilde{\mathscr{X}} \times \mathcal{X} \to [0,1]$ and the probability measure $P_f C : \tilde{\mathscr{X}} \to [0,1]$ is understood as

$$P_f C(A) := \int_{x \in \mathcal{X}} C(A|x) dP_f(x). \tag{18}$$

This is equivalent to the more general notion of deficiency of [27] for dominated models on Polish spaces (see Proposition 9.2 in [68]).

The deficiency $\mathfrak{d}(\mathcal{P}; \mathcal{Q})$ quantifies the degree to which $\mathcal{Q}$ can be approximated by an experiment $\mathcal{P}$. If $\mathfrak{d}(\mathcal{P}; \mathcal{Q}) \leq \varrho$, it implies that for bounded loss functions, each decision procedure within $\mathcal{Q}$ has an associated procedure in $\mathcal{P}$ that achieves nearly the same risk, up to a multiple of ϱ.

To make this precise, let $\mathcal{F}$ be a measurable space and consider a function $\ell : \mathcal{F} \times \mathcal{D} \to [0,1]$ on a measurable space $(\mathcal{D}, \mathscr{D})$, such that $t \mapsto \ell(f, t)$ is measurable for all $f \in \mathcal{F}$, which we shall refer to a *loss functions*. We shall consider a *decision procedure* for $(\mathcal{Q}, \mathcal{D})$ to be a Markov kernel $D : \mathscr{D} \times \tilde{\mathcal{X}} \to [0,1]$. If $\mathfrak{d}(\mathcal{P}; \mathcal{Q}) \leq \varrho$, there exists $C : \tilde{\mathscr{X}} \times \mathcal{X} \to [0,1]$ such that for all decision procedures D for $(\mathcal{Q}, \mathcal{D})$ we have that

$$\int \ell(f, \varphi) dP_f C D(\varphi) \leq \int \ell(f, \varphi) dQ_f D(\varphi) + 2\varrho, \quad \text{for all } f \in \mathcal{F}.$$

Here, the Markov kernel $Q_f D$ is to be understood in the sense of (18) and $CD : \mathscr{D} \times \mathcal{X} \to [0,1]$ as

$$CD(A|x) = \int D(A|\tilde{x}) dC(\tilde{x}|x).$$

There is also the following reverse implication; suppose that there exists a loss function $\ell : \mathcal{F} \times \mathcal{D} \to [0,1]$ on a measurable space $(\mathcal{D}, \mathscr{D})$, and

$$\inf_C \inf_D \sup_{f \in \mathcal{F}} \left| \int \ell(f, \varphi) dQ_f D(\varphi) - \int \ell(f, \varphi) dP_f C D(\varphi) \right| > 2\varrho,$$

where the two infimums are over all decision procedures D and Markov kernels $C : \tilde{\mathscr{X}} \times \mathcal{X} \to [0,1]$. Then, $\mathfrak{d}(\mathcal{Q}, \mathcal{P}) > \varrho$. This follows immediately from e.g. Lemma 12 in the appendix, since $x \mapsto \int \ell(f, \varphi) dD(\varphi|x)$ is measurable. In the more extensive framework considered in e.g. [27], such a reverse implication for risk functions fully characterizes the deficiency between two models, but this framework is not needed in what follows.

Le Cam's deficiency distance between $\mathcal{P}$ and $\mathcal{Q}$ is then defined as

$$\Delta(\mathcal{P}, \mathcal{Q}) = \max \{\mathfrak{d}(\mathcal{P}; \mathcal{Q}), \mathfrak{d}(\mathcal{Q}, \mathcal{P})\}.$$

This semi-metric becomes a metric whenever $\mathcal{P}$ and $\mathcal{Q}$ are identified whenever $\mathfrak{d}(\mathcal{P};\mathcal{Q}) + \mathfrak{d}(\mathcal{Q},\mathcal{P}) = 0$. Two sequences of experiments $\mathcal{P}_\nu$ and $\mathcal{Q}_\nu$ are called *asymptotically equivalent* if their difference $\Delta(\mathcal{P}_\nu, \mathcal{Q}_\nu)$ tends to zero as ν approaches infinity. Conversely, such sequences shall be called *asymptotically nonequivalent* if $\Delta(\mathcal{P}_\nu, \mathcal{Q}_\nu) > c$ as $\nu \to \infty$ for a fixed constant $c > 0$.

The final notion we shall recall is that of sufficiency. A statistic $S : \mathcal{X} \to \tilde{\mathcal{X}}$ is *sufficient for the model* $\mathcal{P}$ if for any $A \in \mathscr{X}$ there exists a measurable map $\psi_A : \tilde{\mathcal{X}} \to \mathbb{R}$ such that

$$P_f\left(A \cap S^{-1}(B)\right) = \int_B \psi_A(\tilde{x}) dP_f^S(\tilde{x}) \ \text{ for all } B \in \tilde{\mathscr{X}} \ \text{ and } f \in \mathcal{F}.$$

Here, the measure P_f^S is to be understood as the push-forward measure $P_f^S(B) = P_f(S^{-1}(B))$. A sufficient statistic allows for transforming observations from one model to another, "sufficient" model which is equivalent in the sense of Le Cam distance. That is, if S is a sufficient statistic for $\mathcal{P}$, then the model $\mathcal{P}' := \{P_f^S : f \in \mathcal{F}\}$ satisfies $\Delta(\mathcal{P}, \mathcal{P}') = 0$.

The next lemma is the Neyman-Fisher factorization theorem gives a useful characterization of sufficiency of a statistic for models that admit densities with respect to the same dominating measure.

Lemma 2. *Suppose that $P_f \ll \mu$ for all $P_f \in \mathcal{P}$ with μ a sigma-finite measure. A statistic $S : \mathcal{X} \to \tilde{X}$ is sufficient for $\mathcal{P}$ if and only if there exists measurable functions $g_f : \mathbb{R} \to \mathbb{R}$ and $h : \mathcal{X} \to \mathbb{R}$ such that*

$$\frac{dP_f}{d\mu}(x) = g_f(S(x))h(x) \ \text{ for almost every } x \in \mathcal{X} \text{ and every } f \in \mathcal{F}. \tag{19}$$

A proof for both the lemma and the last statement of the previous paragraph can be found in Chapter 5 of [62].

A.2 Equivalence of distributed decision problems

We now turn to the distributed setting considered in the paper, where $j = 1, \ldots, m$ servers each receive data $X^{(j)}$ drawn from a distribution P_f and sample space $(\mathcal{X}, \mathscr{X})$. Each of the servers communicates a transcript based on the data to a central server, which based on the aggregated transcripts computes its solution to the decision problem at hand.

The tools developed in this section apply to wider range of distributed architectures than the one considered in the main text of the paper, as introduced in Section 2. The framework introduced here accommodates various forms of interaction between servers, including sequential and blackboard protocols (see e.g. [6, 16]). In contrast, the main text focuses on servers that either do not communicate (local randomness protocols) or utilize a shared randomness source (a special case of sequential or blackboard communication).

A *distributed protocol for the experiment* $\mathcal{P}$ with decision space $(\mathcal{D}, \mathscr{D})$ consists of a triplet $\{D, \mathcal{K}, (\mathcal{U}, \mathscr{U}, \mathbb{P}^U)\}$, a Markov kernel $D : \mathscr{D} \times \bigotimes_{j=1}^m \mathcal{Y}^{(j)} \to [0,1]$ and a probability space $(\mathcal{U}, \mathscr{U}, \mathbb{P}^U)$, and $\mathcal{K}$ is a collection of Markov kernels.

To unpack all this notation: the Markov kernel D takes the role of the decision procedure, where the decision is to made on the basis of the transcripts generated by the Markov kernels $\mathcal{K}$. The transcripts are in turn generated based on the data and a source of shared randomness independent of the data. The probability space $(\mathcal{U}, \mathscr{U}, \mathbb{P}^U)$ plays the role of the source of randomness that is shared by the servers. The distributed protocol is said to have *no access to shared randomness* or to be *a local randomness protocol* if $\mathscr{U}$ is the trivial sigma-algebra.

In this section, we shall consider three types of communication architectures:

- *One shot protocols*: $\mathcal{K} = \{K^j\}_{j=1,\ldots,m}$ where $K^j : \mathscr{Y}^{(j)} \times (\mathcal{X} \times \mathcal{U}) \to [0,1]$. These protocols are what are considered in the main text of the paper.
- *Sequential protocols*: $\mathcal{K} = \{K^j\}_{j=1,\ldots,m}$ where $K^j : \mathscr{Y}^{(j)} \times (\mathcal{X} \times \mathcal{U} \times \mathcal{Y}^{(1)} \times \cdots \times \mathcal{Y}^{(j-1)}) \to [0,1]$. That is, the transcript generated by server j is based on the data, the shared randomness and the transcripts of the previous servers.
- *Blackboard protocols*:

- *Blackboard protocols*: $\mathcal{K} = \{K_t^j\}_{j=1,\ldots,m,t=1,\ldots,T}$ where $K_1^j : \mathcal{Y}^{(j)} \times (\mathcal{X} \times \mathcal{U}) \to [0,1]$
  and
  $$K_t^j : \mathcal{Y}^{(j)} \times \left(\mathcal{X} \times \mathcal{U} \times (\mathcal{Y}^{(1)} \times \cdots \times \mathcal{Y}^{(m)})^{\otimes(t-1)}\right) \to [0,1],$$
  for $t = 2,\ldots,T$. That is, the transcript generated by server j is based on the data, the shared randomness, and the transcripts of *all* the servers from the previous round.

For one shot protocols, we have in terms of random variables that $X^{(j)} \sim P_f$, $U \sim \mathbb{P}^U$, $Y^{(j)}|(X^{(j)}, U) \sim K^j(\cdot|X^{(j)}, U)$ for $j = 1,\ldots,m$ and $\varphi \sim D(\cdot|Y)$ with $Y = (Y^{(1)},\ldots,Y^{(m)})$. This gives rise to a Markov chain

$$
\begin{array}{ccccc}
X^{(1)} & \longrightarrow & Y^{(1)}|U & \searrow & \\
\vdots & \longrightarrow & \vdots & \longrightarrow & \varphi. \\
X^{(m)} & \longrightarrow & Y^{(m)}|U & \nearrow &
\end{array}
\tag{20}
$$

For $x = (x^{(1)},\ldots,x^{(m)}) \in \mathcal{X}^m$, $u \in \mathcal{U}$ and $\{K^j\}_{j=1,\ldots,m}$, let $x \mapsto K(A|x)$ be the Markov kernel product distribution $\bigotimes_{j=1}^m K^j(\cdot|x^{(j)}, u)$. Given a distributed protocol and i.i.d. data from P_f we shall use $\mathbb{P}_f$ to denote the joint distribution the data $X \sim P_f^m$, the shared randomness $U \sim \mathbb{P}^U$ and $Y = (Y^{(1)},\ldots,Y^{(m)})$ with $Y|(X,U) \sim K(Y|X,U)$. We have that $P_f^m K = \bigotimes_{j=1}^m P_f K^j$ and the push-forward measure of Y then disintegrates as

$$\mathbb{P}_f^Y(A) = P_f^m \mathbb{P}^U K(A) = \mathbb{P}^U P_f^m K(A) = \int d\bigotimes_{j=1}^m P_f K^j(\cdot|X^{(j)}, u)(A) d\mathbb{P}^U(u), \tag{21}$$

where the second equality follows from the independence of U with the data $X := (X^{(1)},\ldots,X^{(m)})$ drawn from P_f.

For sequential protocols, the push-forward measure of Y instead disintegrates as

$$\mathbb{P}_f^Y(A) = \int_{\mathcal{U}} \left[\int \cdots \int \mathbb{1}_A(y) dP_f K^m\left(y^m \mid X^{(m)}, u, (y)_{j=1}^{m-1}\right) \cdots dP_f K^1(y^1 \mid X^{(1)}, u)\right] d\mathbb{P}^U(u), \tag{22}$$

For blackboard protocols, a similar disintegration applies for each of the rounds.

A one shot or sequential distributed protocol is said to satisfy a *b-bit bandwidth constraint* if its kernels $\{K^j\}_{j=1,\ldots,m}$ are defined on spaces satisfying $|\mathcal{Y}^{(j)}| \leq 2^b$. For blackboard protocols, various bandwidth constraints can be imposed, such as a b-bit bandwidth constraint for each round $t = 1,\ldots,T$.

Given Markov Kernels $C^j : \mathcal{X} \times \tilde{\mathcal{X}} \to [0,1]$, $j = 1,\ldots,m$, a distributed one shot protocol $\{D, \{K^j\}_{j=1,\ldots,m}, (\mathcal{U}, \mathcal{U}, \mathbb{P}^U)\}$ for the model $\mathcal{P}$, yields a distributed protocol for the model $\mathcal{Q}$: $\{D, \{CK^j\}_{j=1,\ldots,m}, (\mathcal{U}, \mathcal{U}, \mathbb{P}^U)\}$. If $\{K^j\}_{j=1,\ldots,m}$ is b-bit bandwidth constraint, the collection of kernels $\{C^j K^j\}_{j=1,\ldots,m}$ do so too, as each $C^j K^j$ is defined on $\mathcal{Y}^{(j)} \times \tilde{\mathcal{X}}$.

Similarly, for a sequential protocol, the Markov kernels C^j and K^j yield a distributed sequential protocol for the model $\mathcal{Q}$ with kernels $\{C^j K^j\}_{j=1,\ldots,m}$. If each K^j is b-bit bandwidth constraint, so is each $C^j K^j$. For blackboard protocols, the same reasoning applies to each round $t = 1,\ldots,T$. That is, type of protocol defined by the kernels $\mathcal{K}$ is "closed under composition" with kernels C^j between $\tilde{\mathcal{X}}$ with target space $\mathcal{X}$, where bandwidth constraints are preserved.

We shall consider the notion of local ϵ-differential privacy of Definition 3. A Markov kernel $K : \mathcal{Y} \times \mathcal{X} \to [0,1]$ is called *locally (ϵ, δ)-differentially private* if

$$K(A|x) \leq e^\epsilon K(A|x') + \delta \quad \text{for all } A \in \mathcal{Y} \text{ and } x, x' \in \mathcal{X}. \tag{23}$$

Since the definition of differential privacy depends heavily on what one defines as the sample space, it is difficult to obtain a similar "transfer of distributed protocols" that respects the (ϵ, δ)-differential privacy constraint, hence the choice to consider local constraints only.

A one shot or sequential distributed protocol shall be called locally ϵ-differentially private if (23) holds for each K^j; $j = 1, \ldots, m$. For blackboard protocols, one can impose a (ϵ, δ)-differential privacy constraint for each round $t = 1, \ldots, T$, or for the entire output over $t = 1, \ldots, T$ rounds. The following lemma shows that local ϵ-differential privacy, just like bandwidth constraints, carry over from one model to the other.

Lemma 3. *Let $(\mathcal{X}, \mathscr{X})$ and $(\tilde{\mathcal{X}}, \tilde{\mathscr{X}})$ be measurable spaces and consider Markov kernels $C : \tilde{\mathcal{X}} \times \mathcal{X} \to [0,1]$ and $K : \mathcal{Y} \times \mathcal{X} \to [0,1]$. If K is b-bit bandwidth constraint, so is the Markov kernel $CK : \mathcal{Y} \times \tilde{\mathcal{X}} \to [0,1]$. If K is locally ϵ-differentially private, so is CK. Furthermore, for any collection of Markov kernels $\mathcal{K}$, the same reasoning applies to the collection $\{CK\}_{K \in \mathcal{K}}$, preserving bandwidth constraints and local differential privacy, as well as the protocol's architecture.*

Proof. The first statement has been remarked on earlier in the section. For the second statement, consider arbitrary $\tilde{x}, \tilde{x}' \in \tilde{\mathcal{X}}$ and $A \in \mathcal{Y}$. Using that C is a Markov kernel and applying (23) to K yields

$$CK(A|\tilde{x}) = \int K(A|x)dC(x|\tilde{x}) = \int \int K(A|x)dC(x|\tilde{x})dC(x'|\tilde{x}')$$

$$\leq e^\epsilon \int K(A|x')dC(x'|\tilde{x}') + \delta = e^\epsilon CK(A|\tilde{x}') + \delta,$$

which shows CK is locally (ϵ, δ)-differentially private. The above argument applies pointwise for all other conditional arguments in the Markov kernel, hence the same reasoning applies to shared randomness, sequential and blackboard protocols. $\square$

In an abuse of notation, let D denote the entire distributed protocol (triplet)

$$\{D, \{K^j\}_{j=1,\ldots,m}, (\mathcal{U}, \mathscr{U}, \mathbb{P}^U)\}$$

for the experiment $\mathcal{P}$ (indexed by $\mathcal{F}$) with decision space $(\mathcal{D}, \mathscr{D})$. Given D and a loss function $\ell : \mathcal{F} \times \mathcal{D} \to [-1,1]$, we define the *distributed risk of D in $\mathcal{P}$ for ℓ* as

$$\mathcal{R}_{\mathcal{P}}(D, \ell) := \sup_{f \in \mathcal{F}} \int \int \int \ell(f, \varphi)dD(\varphi|y) \, d\bigotimes_{j=1}^{m} P_f K^j(\cdot|X^{(j)}, u)(y)d\mathbb{P}^U(u).$$

We are now ready to formulate a straightforward consequence for the distributed risk, following from models being close in Le Cam distance. This finding, formulated in Lemma 4, shall serve as one of the main tools for deriving the main results. It states roughly that, whenever there is a b-bit bandwidth constrained distributed protocol that achieves a certain risk is one model and there is small deficiency with the other model relative to the number of servers, there exists a b-bit distributed protocol that achieves comparable risk for the other model. A similar statement holds under local differential privacy constraints. If there is a locally (ϵ, δ)-differentially private distributed procedure in the one model and there is small deficiency with another model, it means that there is comparable risk for the privacy constraint distributed decision problem.

Lemma 4. *Let $m \in \mathbb{N}$. Consider two experiments $\mathcal{P}$ and $\mathcal{Q}$ with indexing set $\mathcal{F}$, satisfying $m\mathfrak{d}(\mathcal{Q}; \mathcal{P}) \leq \varrho$ for some $\varrho > 0$. Let $\mathscr{J}_{\mathcal{P}}$ and $\mathscr{J}_{\mathcal{Q}}$ denote the class of b-bit bandwidth constraint shared randomness protocols for the models $\mathcal{P}$ and $\mathcal{Q}$ respectively.*

Then, for any loss function $\ell : \mathcal{F} \times \mathcal{D} \to [0,1]$,

$$\inf_{D \in \mathscr{J}_{\mathcal{Q}}} \mathcal{R}_{\mathcal{Q}}(D, \ell) - \inf_{D \in \mathscr{J}_{\mathcal{P}}} \mathcal{R}_{\mathcal{P}}(D, \ell) \leq \varrho.$$

where in the infimum, in an abuse of notation, D denotes the entire distributed protocol triplet $\{D, \{K^j\}_{j=1,\ldots,m}, (\mathcal{U}, \mathscr{U}, \mathbb{P}^U)\}$.

The same statement is holds for $\mathscr{J}_{\mathcal{P}}$ and $\mathscr{J}_{\mathcal{Q}}$ denoting either classes of b-bit bandwidth constraint local randomness, sequential protocols, or any of these distributed protocols satisfying local (ϵ, δ)-differential privacy constraints, for the respective models $\mathcal{P}$ and $\mathcal{Q}$. If $Tm\mathfrak{d}(\mathcal{Q}; \mathcal{P}) \leq \varrho$, the same statement holds for blackboard protocols with T rounds.

Remark 1. This exemplifies also that, even though models $\mathcal{P}^m = \{P_f^m : f \in \mathcal{F}\}$ and $\mathcal{Q}^m = \{Q_f^m : f \in \mathcal{F}\}$ are close in Le Cam distance, distributed decision problems formulated in terms the models $\mathcal{P}$ and $\mathcal{Q}$, can have greatly different performance in terms of associated risks.

Proof. By e.g. Theorem 2 in [62], $m\mathfrak{d}(\mathcal{Q};\mathcal{P}) \leq \varrho$ implies that there exists a kernel $C : \mathscr{X} \times \tilde{\mathscr{X}} \to [0,1]$ such that

$$\sup_{f \in \mathcal{F}} \|P_f - Q_f C\|_{\mathrm{TV}} \leq \varrho/m. \tag{24}$$

By Lemma 3, the kernels $\tilde{\mathcal{K}} = \{CK : K \in \mathcal{K}\}$ satisfy the a b-bit bandwidth constraint or local (ϵ, δ)-differential privacy constraint if the collection $\{K^j\}_{j=1,\dots,m}$ does. To illustrate this further, consider a one shot distributed protocol for $\mathcal{P}$, $\{D, \{K^j\}_{j=1,\dots,m}, (\mathcal{U}, \mathscr{U}, \mathbb{P}^U)\} \in \mathscr{J}_{\mathcal{P}}$, the distributed protocol $\tilde{D} = \{D, \{CK^j\}_{j=1,\dots,m}, (\mathcal{U}, \mathscr{U}, \mathbb{P}^U)\}$ is then an element of $\mathscr{J}_{\mathcal{Q}}$.

Using the fact that ℓ is bounded by one and Lemma 13 in the appendix, it follows that

$$\mathcal{R}_{\mathcal{Q}}(\tilde{D}, \ell) - \mathcal{R}_{\mathcal{P}}(D, \ell) \leq \|\mathbb{P}^U \bigotimes_{j=1}^m P_f K^j - \mathbb{P}^U \bigotimes_{j=1}^m Q_f CK^j\|_{\mathrm{TV}}$$

$$\leq \sum_{j=1}^m \|\mathbb{P}^U P_f K^j - \mathbb{P}^U Q_f CK^j\|_{\mathrm{TV}}.$$

By Lemma 14 in the appendix,

$$\|\mathbb{P}^U P_f K^j - \mathbb{P}^U Q_f CK^j\|_{\mathrm{TV}} \leq \|\mathbb{P}^U P_f - \mathbb{P}^U Q_f C\|_{\mathrm{TV}} = \|P_f - Q_f C\|_{\mathrm{TV}},$$

which combined with (24) finishes for one shot protocols.

Similar reasoning applies to sequential protocols. We start by noting that

$$\|P_f K^1(\cdot \mid u) - Q_f CK^1(\cdot \mid u)\|_{\mathrm{TV}} \leq \|P_f - Q_f C\|_{\mathrm{TV}} \leq \varrho/m,$$

Lemma 11 implies that (for any u) there exists a coupling $\mathbb{P}$ of $Y^{(1)} \sim P_f K^1(\cdot \mid u)$ and $\tilde{Y}^{(1)} \sim Q_f CK^1(\cdot \mid u)$ such that $2\mathbb{P}(Y^{(1)} \neq \tilde{Y}^{(1)}) \leq \varrho/m$. Write

$$K^{m:2}(\cdot \mid x^m, \dots, x^2, u, y^1) := \int \cdots \int dK^m(y^m \mid x^m, u, y^{m-1}, \dots, y^1) \cdots dK^2(y^2 \mid x^2, u, y^1).$$

Using the disintegration relationship of (22), we obtain

$$\|\mathbb{P}^U P_f^m K - \mathbb{P}^U Q_f CK\|_{\mathrm{TV}} \leq \int \|P_f^m K(\cdot \mid u) - Q_f CK(\cdot \mid u)\|_{\mathrm{TV}} d\mathbb{P}^U(u)$$

$$\leq \varrho/m + \int \mathbb{E}_f^{Y^{(1)} \mid U=u} \|P_f^m K^{m:2}(\cdot \mid u, Y^{(1)}) - Q_f CK^{m:2}(\cdot \mid u, Y^{(1)})\|_{\mathrm{TV}} d\mathbb{P}^U(u).$$

Iterating the above argument $m-1$ times, we the statement for sequential protocols. The statement for blackboard protocols follows by combining the above arguments for each round $t = 1, \dots, T$. $\square$

In the remainder of this text, we shall constrain ourselves to a particular bounded risk function and distributed decision problem; distributed hypothesis testing. The following corollary formalizes the statement at the start of the paragraph for the testing a simple null versus a compositive alternative hypothesis in the distributed setting. To that extent, consider a test of the hypotheses

$$H_0 : f = f_0 \quad \text{versus the alternative hypothesis} \quad f \in H_1 \tag{25}$$

and an experiment $\mathcal{P}$ with indexing set $\mathcal{F}$ satisfying $\{f_0\} \cup H_1 \subset \mathcal{F}$. Consider for $m \in \mathbb{N}$ a *distributed testing protocol* for the model $\mathcal{P}$ to be a distributed protocol $T \equiv \{D, \{K^j\}_{j=1,\dots,m}, (\mathcal{U}, \mathscr{U}, \mathbb{P}^U)\}$, where in a slight abuse of notation, we shall also use T to denote the (possibly randomized) test $T|Y \sim D(\cdot|Y)$. Recalling the notation $\mathbb{P}_f^Y = P_f^m \mathbb{P}^U K$ as given in (21), define the distributed testing risk for the hypotheses in (25) and the model $\mathcal{P}$ as

$$\mathcal{R}_{\mathcal{P}}(T, H_1) := \mathbb{P}_{f_0}^Y D(T|Y) + \sup_{f \in H_1} \mathbb{P}_f^Y (1 - D(T|Y)).$$

Here, $D(T|Y) := D(\{1\}|Y)$, but one can equivalently consider a deterministic measurable map $T : \prod_{j=1}^{m} \mathcal{Y}^{(j)} \to [0,1]$ without loss of generality. Let $\mathscr{T}_{SR}^{b}(\mathcal{P})$ (resp. $\mathscr{T}_{LR}^{b}(\mathcal{P})$) denote the set of shared randomness (resp. local randomness) distributed testing protocols for $\mathcal{P}$ satisfying a b-bit bandwidth constraint. Similarly, let $\mathscr{T}_{SR}^{(\epsilon,\delta)}(\mathcal{P})$ (resp. $\mathscr{T}_{LR}^{(\epsilon,\delta)}(\mathcal{P})$) denote the set of shared randomness (resp. local randomness) distributed testing protocols for $\mathcal{P}$ satisfying a local (ϵ, δ)-differential privacy constraint. Define the same classes for the model $\mathcal{Q}$ in the obvious way. Using Lemma 4, we obtain the following result, which is a more general version of Lemma 1.

Lemma 5. *Consider experiments $\mathcal{P}, \mathcal{Q}$ such that $m\mathfrak{d}(\mathcal{Q};\mathcal{P}) \leq \varrho$ for $\varrho > 0$. It holds that*

$$\inf_{T \in \mathscr{T}(\mathcal{P})} \mathcal{R}_{\mathcal{P}}(T, H_1) \leq \inf_{T \in \mathscr{T}(\mathcal{Q})} \mathcal{R}_{\mathcal{Q}}(T, H_1) + 2\varrho,$$

where $\mathscr{T}$ is either $\mathscr{T}_{SR}^{b}, \mathscr{T}_{LR}^{b}, \mathscr{T}_{SR}^{(\epsilon,\delta)}$ or $\mathscr{T}_{LR}^{(\epsilon,\delta)}$.

Proof. Given $\{T, \{K^j\}_{j=1,\dots,m}, (\mathcal{U}, \mathscr{U}, \mathbb{P}^U)\} \in \mathscr{T}(\mathcal{P})$, Lemma 4 applied to the loss function

$$\ell(f, t) := t\mathbb{1}_{\{f_0\}}(f) + (1-t)\mathbb{1}_{H_1}(f)$$

and using that $\{f_0\} \cup H_1 \subset \mathcal{F}$ gives

$$\mathbb{P}_{f_0}^Y D(T|Y) < \mathbb{Q}_{f_0}^Y D(T|Y) + \varrho \quad \text{and} \quad \sup_{f \in H_1} \mathbb{P}_f^Y (1 - D(T|Y)) < \sup_{f \in H_1} \mathbb{Q}_f^Y (1 - D(T|Y)) + \varrho$$

for some distributed testing protocol $\{D, \{\tilde{K}^j\}_{j=1,\dots,m}, (\mathcal{U}, \mathscr{U}, \mathbb{P}^U)\}$ in $\mathscr{T}(\mathcal{Q})$, which yields the first statement. The second statement follows by symmetry of the argument. $\square$

The implications of Lemma 4 have implications beyond the testing framework. Whilst in distributed estimation settings, the loss function under consideration is typically not bounded, rates can still be derived in probability. That is, if the minimax rate for the distance d on $\mathcal{F}$ in the model $\mathcal{P}_\nu$ is ρ_ν, the bounded loss function

$$\ell_\nu(f, g) = \mathbb{1}\{d(f, g) \leq C\rho_\nu\} \quad \text{for } C > 0$$

can be used describe minimax estimation rates (in probability) between models $\mathcal{P}$ and $\mathcal{Q}$. Since the paper is about testing, we shall not pursue this direction any further beyond this remark.

In the next sections, we will explore the consequences of Lemma 5 for minimax distributed testing rates for both bandwidth- and privacy constraints.

B Difficulties in direct analysis of the multinomial model under information constraints

Lower-bounds for both estimation and testing problems are typically established by bounding divergence measures between probability distributions, such as the chi-square divergence, mutual information, or total variation; [3, 6, 7, 16**?** , 49, 9, 99, 13, 84, 85, 23, 22].

The proof techniques used for discrete distribution estimation employed in [3, 6] and [7], tight lower-bounds can often be obtained by "tensorizing" the divergence—breaking the problem into a sum of local divergences. The inferential cost incurred due to bandwidth or privacy constraints are then captured via data processing arguments. Similar tensorization techniques are employed in other estimation problems, see for example [16, 99].

However, this tensorization approach does not yield tight bounds for testing problems. For example, [84] uses mutual information in a tensorization framework for testing but only recovers optimal rates when each server communicates a single bit. Similarly, [8] attempts an estimation-based approach for goodness-of-fit testing but obtains tight lower-bounds only under 1-bit constraints, as detailed in Section 4 of their paper.

To achieve tight lower-bounds in testing problems, especially under communication and privacy constraints, different techniques are required. Papers [9, 85, 22] employ methods that significantly diverge from those used in estimation. In [9], a combinatorial expansion of the likelihood is used, effective for small sample sizes in the multinomial model but not generalizable to large numbers of

observations. [85] and [22] address this limitation in the Gaussian setting by utilizing the Brascamp-Lieb inequality from functional analysis, which explicitly leverages the Gaussian properties of the log-likelihood. This approach is not directly applicable to discrete models, due to the lack of quadratic structure in the log-likelihood of the multinomial model.

C   Separation rates for the Gaussian model

For completeness, we provide the relevant results for the Gaussian model studied in [85] and [22].

The first two results come in the form of lower-bounds for the minimax detection thresholds under bandwidth- and differential privacy constraints for the distributed signal detection problem presented in the introduction. We recall that in this problem, each local machine $j \in \{1, \ldots, m\}$ observes

$$X_i^{(j)} = f + Z_i^{(j)}, \tag{26}$$

with $f \in \mathbb{R}^d$ and $Z_i^{(j)} \sim N(0, I_d)$, i.i.d. for $i = 1, \ldots, n$. The null hypothesis constitutes that $f = 0$ versus the alternative hypothesis that

$$f \in H_\rho := \left\{ f \in \mathbb{R}^d \ : \ \|f\|_2 \geq \rho \right\}. \tag{27}$$

The following is Theorem 3.1 from [85].

Theorem 4. *For each $\alpha \in (0, 1)$ there exists a constant $c_\alpha > 0$ (depending only on α) such that if*

$$\rho^2 < c_\alpha \frac{\sqrt{d}}{n} \left(\sqrt{\frac{d}{b \wedge d}} \bigwedge \sqrt{m} \right), \tag{28}$$

then in the shared randomness protocol case

$$\inf_{T \in \mathscr{T}_{SR}^{(b)}} \mathcal{R}(H_\rho, T) > \alpha \ \text{for all} \ n, m, d, b \in \mathbb{N}.$$

Similarly, for

$$\rho^2 < c_\alpha \frac{\sqrt{d}}{n} \left(\frac{d}{b \wedge d} \bigwedge \sqrt{m} \right), \tag{29}$$

we have under the local randomness protocol that

$$\inf_{T \in \mathscr{T}_{LR}^{(b)}} \mathcal{R}(H_\rho, T) > \alpha \ \text{for all} \ n, m, d, b \in \mathbb{N}.$$

Following the proof Theorem 3.1 from [85], we obtain the following lemma.

Lemma 6. *Let $\mathscr{T}^b$ denote the class of b-bit bandwidth constrained shared- or local randomness distributed testing protocols and let ρ satisfy either (28) or (29), respectively. For any $\alpha \in (0, 1)$, there exists $c_\alpha > 0$ such that for all $T \in \mathscr{T}^{(b)}$ it holds that*

$$\inf_{T \in \mathscr{T}} \mathcal{R}(H_\rho, T) > \alpha - \pi(H_\rho^c),$$

where $\pi = N(0, c_\alpha^{-1/2} d^{-1} \rho^2 \bar{\Gamma})$ for a symmetric, idempotent matrix $\bar{\Gamma} \in \mathbb{R}^{d \times d}$ with $d/2 \leq rank(\bar{\Gamma}) \leq d$.

Similarly, the following result can be derived from the proof of Theorem 5 in [22], by taking $s > 0$ in the theorem such that $d_{L_s} = d$.

Theorem 5. *For each $\alpha \in (0, 1)$ there exists a constant $c_\alpha > 0$ (depending only on α), such that for any $n, m, d \in \mathbb{N}$ and*

$$0 < \epsilon \leq 1 \ \text{and} \ 0 \leq \delta \leq \left(c_\alpha m^{-3/2} \wedge n d^{-1} \epsilon^2 \wedge n^{1/2} d^{-1/2} \epsilon^2 \right)^{1+p} \ \text{for some} \ p > 0, \tag{30}$$

the condition

$$\rho^2 < c_\alpha \left(\frac{d}{mn\sqrt{n\epsilon^2 \wedge 1}\sqrt{n\epsilon^2 \wedge d}} \bigwedge \left(\frac{\sqrt{d}}{\sqrt{m}n\sqrt{n\epsilon^2 \wedge 1}} \bigvee \frac{1}{mn^2\epsilon^2} \right) \right), \tag{31}$$

implies

$$\inf_{T \in \mathscr{T}_{SR}^{(\epsilon,\delta)}} \mathcal{R}(H_\rho, T) > \alpha.$$

Similarly, for any $n, m, d \in \mathbb{N}$ and ϵ, δ satisfying (30), the condition

$$\rho^2 < c_\alpha \left(\frac{d\sqrt{d}}{mn(n\epsilon^2 \wedge d)} \bigwedge \left(\frac{\sqrt{d}}{\sqrt{m}n\sqrt{n\epsilon^2 \wedge 1}} \bigvee \frac{1}{mn^2\epsilon^2} \right) \right), \tag{32}$$

implies that

$$\inf_{T \in \mathscr{T}_{LR}^{(\epsilon,\delta)}} \mathcal{R}(H_\rho, T) > \alpha.$$

Following its proof, we obtain the following the following sub-result.

Lemma 7. *Let $\mathscr{T}$ denote the class of shared- or local randomness distributed testing protocols satisfying an (ϵ, δ)-differential privacy constraint for $0 < \epsilon \leq 1$, $0 \leq \delta \leq \left(c_\alpha m^{-1} \wedge c_\alpha \epsilon m^{-1/2} \wedge n\epsilon^2 \wedge n^2 d^{-1}\epsilon^2 \wedge n^{3/2}d^{-1/2}\epsilon^2 \right)$ and let ρ satisfy either (28) or (29), respectively. For any $\alpha \in (0,1)$, there exists $c_\alpha > 0$ such that for all $T \in \mathscr{T}^{(\epsilon,\delta)}$ it holds that*

$$\mathcal{R}(H_\rho, T) > \alpha - \pi(H_\rho^c),$$

where $\pi = N(0, c_\alpha^{-1/2}d^{-1}\rho^2\bar{\Gamma})$ for a symmetric, idempotent matrix $\bar{\Gamma} \in \mathbb{R}^{d \times d}$ with $\text{rank}(\bar{\Gamma}) \asymp d$.

D   Proofs of Theorems 1, 2 and 3

Proof of Theorems 1 and 2. In what follows, let $\mathscr{T}$ denote a class of distributed protocols satisfying either a $b \equiv b_\nu$-bit bandwidth constraint or a local (ϵ, δ)-differential privacy constraint for $\epsilon \equiv \epsilon_\nu$, $\delta \equiv \delta_\nu$, allowing either for shared randomness or only local randomness.

For any sequences $m \equiv m_\nu$, $d \equiv d_\nu$ and $n \equiv n_\nu$ with $C_R\, md \log d / \sqrt{n} = o(1)$, it follows from Lemma 5 and the bound (14) that the testing risks satisfy

$$\inf_{T \in \mathscr{T}_\mathcal{Q}} \mathcal{R}_{\mathcal{Q}_\nu}(H_{\rho_\nu}, T) = \inf_{T \in \mathscr{T}_\mathcal{P}} \mathcal{R}_{\mathcal{P}_\nu}(H_{\rho_\nu}, T) + o(1). \tag{33}$$

Let $\rho^* \equiv \rho_\nu^*$ be the minimax rate of the $\mathcal{P}$-distributed problem, over the class $\mathscr{T}_\mathcal{P}$, in the sense that ρ^* equals (up to constants) the right-hand side of (6), (7), (9) or (10). We split the proof into showing that ρ^* is an upper and lower-bound for the $\mathcal{Q}$-distributed problem over the class $\mathscr{T}_\mathcal{P}$.

The rate ρ^* is an upper-bound (up to a poly-logarithmic factor) for the minimax rate in $\mathcal{Q}$: Write, for $q \in \mathcal{F}$, $\sqrt{q} = (\sqrt{q_i})_{i \in [d]}$. Since $X^{(j)} - \sqrt{q_0}$ is a sufficient statistic for $X^{(j)}$, the model (13) is equivalent in the Le Cam sense the one generated by

$$X^{(j)} = \sqrt{q} - \sqrt{q_0} + \frac{1}{\sqrt{2n}}Z^{(j)} \text{ with } Z^{(j)} \sim N(0, I_d), \tag{34}$$

for $q \in \mathcal{F}$, which we shall denote by $\tilde{\mathcal{P}}$. Consequently, by another application of Lemma 5, it suffices to show

$$\inf_{T \in \mathscr{T}_{\tilde{\mathcal{P}}}} \mathcal{R}_{\tilde{\mathcal{P}}}(H_{\rho_\nu}, T) \to 0.$$

If $\|q - q_0\|_1 \geq \rho$, Lemma 15 implies that $\|\sqrt{q} - \sqrt{q_0}\|_2 \geq \rho/2$. Consequently, if $\rho \equiv \rho_\nu \gg M_\nu \rho^*$ where ρ^* is of equal order of the minimax rate for the respective class of distributed protocols $\mathscr{T}_\mathcal{P}$ and M_ν is an appropriately large factor (of poly-logarithmic order in case of differential privacy constraints), a distributed protocol $T \in \mathscr{T}_\mathcal{P}$ exists for the Gaussian model that achieves the separation rate for whenever $H_0 : \sqrt{q} - \sqrt{q_0} = 0$ versus $H_\rho : \|\sqrt{q} - \sqrt{q_0}\|_2 \geq \rho/2$. By the established equivalence of the minimax risks (33), this implies that a protocol $T \in \mathscr{T}_\mathcal{Q}$ exists for the multinomial model as well. Thus, ρ_ν is an upper-bound for the minimax separation rate for the class of distributed protocols $\mathscr{T}_\mathcal{Q}$ of the multinomial model.

The rate ρ^* is a lower-bound for the minimax rate in $\mathcal{Q}$: Suppose that $\rho \equiv \rho_\nu$ is of smaller order than the minimax rate ρ^* of the class $\mathscr{T}_\mathcal{P}$, in the sense that $\rho^*/\rho \to \infty$ as $\nu \to \infty$. We aim to use the Bayes risk lower-bound of Lemmas 6 and 7, which apply to a Gaussian prior. To accommodate

a Gaussian prior with sufficient mass on the alternative hypothesis, we first need to address the "constraint on the signal" imposed by $\sum_{i=1}^{d} q_i = 1$ for $q \in \mathcal{F}$.

To that extent, consider without loss of generality d to be divisible by two. Let $I_R := [-(R-1)/(R+1), (R-1)/(R+1)]$. For all $(f_i)_{i \in [d/2]} \in I_R^{d/2}/\sqrt{d}$, there exists a $q^f := (q_i^f)_{i \in [d]} \in \mathcal{F}$ such that $q_i^f = 1/d + f_i/\sqrt{d}$ for $i = 1, \ldots, d/2$ and $q_i^f = 1/d - f_i/\sqrt{d}$ for $i = d/2+1, \ldots, d$. To see that $q^f \in \mathcal{F}$, note that $\sum_{i=1}^{d} q_i^f = 1$, $q^f \geq 0$ and

$$\max_{1 \leq i,k \leq d} \frac{q_i^f}{q_k^f} \leq \max_{c \in I_R} \frac{1+c}{1-c} = R.$$

Define $\mathcal{F}'$ as the set

$$\left\{ (q_i)_{i \in [d]} \in \mathcal{F} \; : \; (f_i)_{i \in [d/2]} \in \frac{I_R^{d/2}}{\sqrt{d}} \text{ s.t. } q_i^f = 1/d + (1 - 2\mathbb{1}_{i > d/2}) \frac{f_i}{\sqrt{d}} \text{ for } i = 1, \ldots, d \right\}$$

and

$$H_\rho' := \{q \; : \; q \in \mathcal{F}', \|q - q_0\|_1 \geq \rho\}.$$

We have $\mathcal{F}' \subset \mathcal{F}$, which in turn implies that $H_\rho' \subset H_\rho$. Combined with the fact that the testing risk decreases by considering smaller alternative hypotheses, this results in

$$\inf_{T \in \mathscr{T}_\mathcal{P}} \mathcal{R}_\mathcal{P}(T, H_\rho) \geq \inf_{T \in \mathscr{T}_\mathcal{P}} \mathcal{R}_\mathcal{P}(T, H_\rho'). \tag{35}$$

Define $g_f = (1/2)(f, -f) \in \mathbb{R}^d$. By Pinsker's inequality,

$$\left\| P_{\sqrt{q^f}-\sqrt{q_0}}^{nm} - P_{g_f}^{nm} \right\|_{\mathrm{TV}} \leq 1 \wedge \sqrt{\frac{mn}{4} D_{\mathrm{KL}}(P_{\sqrt{q}-\sqrt{q_0}}; P_{g_f})}$$

$$= 1 \wedge \frac{\sqrt{mn}}{2} \left\| \sqrt{q_0 + 2g_f/\sqrt{d}} - \sqrt{q_0} - g_f \right\|_2 =: D_f,$$

where $P_{\sqrt{q}-\sqrt{q_0}}^{n}$ denotes the distribution of (34) and the square root is to be understood as applied coordinate wise.

Let $\pi = N(0, d^{-1}(\rho^*)^2 \bar{\Gamma})$ for a symmetric, idempotent matrix $\bar{\Gamma} \in \mathbb{R}^{d/2 \times d/2}$ with $d/4 \leq \mathrm{rank}(\bar{\Gamma}) \leq d/2$.

We have that

$$\inf_{T \in \mathscr{T}_\mathcal{P}} \mathcal{R}_\mathcal{P}(T, H_\rho') \geq \inf_{T \in \mathscr{J}_\mathcal{P}} \left[\mathbb{P}_0 T(Y) + \int \mathbb{P}_{g_f}(1 - T(Y)) d\pi(f) \right] - 2 \int D_f d\pi(f)$$

$$- \pi \left(f : f \notin (I_R/\sqrt{d})^{d/2} \text{ or } \left\| (q_i^f)_{i \in [d]} - q_0 \right\|_1 < \rho \right).$$

By Lemma 8, the model $\{P_{g_f} \; : \; f \in I_R^{d/2}/\sqrt{d}\}$ is equivalent to the model generated by the observations

$$S_i^{(j)} := f_i + \frac{1}{\sqrt{n}} Z_i^{(j)} \tag{36}$$

for $i = 1, \ldots, d/2$. Since $\mathcal{F}'$ is bijective with $(I_R/\sqrt{d})^{d/2}$, Lemma 5 implies that

$$\inf_{T \in \mathscr{J}_\mathcal{P}} \left[\mathbb{P}_0 T(Y) + \int \mathbb{P}_{g_f}(1 - T(Y)) d\pi(f) \right] = \inf_{T \in \mathscr{J}_{\tilde{\mathcal{P}}}} \left[\mathbb{P}_0' T(Y) + \int \mathbb{P}_f'(1 - T(Y)) d\pi(f) \right] \tag{37}$$

where $\tilde{\mathcal{P}}$ is the model generated by the observations in display (36) for $i = 1, \ldots, d/2$ and $\mathbb{P}_f'$ denotes the distribution of the distributed protocol with data generated from $f \in \tilde{\mathcal{P}}$.

It follows from Lemma 6 in the case of bandwidth constraints or Lemma 7 in the case of privacy constraints (using that $\rho \ll \rho^*$ in both cases) that the latter distributed testing risk is lower-bounded by

$$1 - o(1) - \pi \left(f \in \mathbb{R}^{d/2} \; : \; f \notin (I_R/\sqrt{d})^{d/2} \text{ or } \left\| (q_i^f)_{i \in [d]} - q_0 \right\|_1 < \rho \right) - 2 \int D_f d\pi(f). \tag{38}$$

Addressing the third term in the display above; the theorem(s) assume that $md \log d/\sqrt{n} \stackrel{\nu \to \infty}{\to} 0$, $b \geq 1$ and $\epsilon \gg n^{-1/4}$, we have that $\rho^* \ll 1/\sqrt{\log(d)}$, which gives $\|\sqrt{d}f_i\|_\infty \to 0$ with π-probability tending to one (see e.g. Lemma 16), which in turn implies that

$$f_i \in I_R/\sqrt{d} \quad \text{for all } i = 1, \ldots, d/2. \tag{39}$$

Next, we show that $\|(q_i^f)_{i \in [d]} - q_0\|_1 \geq \rho$ with π-probability tending to one. Since $\sum_{i=1}^d |q_i^f - q_0| = 2\sum_{i=1}^{d/2} |f_i/\sqrt{d}|$, we have that for some constants $c, c' > 0$,

$$\pi\left(\left\|q^f - q_0\right\|_1 < \rho\right) \leq \pi\left(\left\|f/\sqrt{d}\right\|_1 < \rho\right) \leq 1 - \Pr\left(\|\bar{\Gamma}Z\|_1 \geq c'd\frac{\rho}{\rho^*}\right).$$

where in the expression on the right-hand side, $Z \sim N(0, I_{d/2})$. Since $\rho \ll \rho^*$ and $\bar{\Gamma}$ is idempotent with rank of the order d, we can conclude that the expression vanishes. This takes care of the third term in (38).

For the last term in (38), the Taylor approximation $\sqrt{1+y} - 1 = y/2 - y^2/8 + \frac{y^3}{16(1+\eta_y^{5/2})}$ for some $\eta \in [0, y]$, combined with the fact that $\|\sqrt{d}f\|_\infty = o_\pi(1)$ yields that

$$2\int D_f d\pi(f) \leq \int 1 \wedge \sqrt{mnd} \left\|(f_i^2)_{i \in [d/2]}\right\|_2 d\pi(f) \lesssim \sqrt{mn}\rho^2.$$

Since the theorem(s) assume that $md\log d/\sqrt{n} \stackrel{\nu \to \infty}{\to} 0$, $b \geq 1$ and $\epsilon \gg n^{-1/4}$, the right-hand side of the above display vanishes. $\square$

Proof of Theorem 3. Let $\mathscr{T}_Q, \mathscr{T}_P$ denote the class of distributed b-bit bandwidth constrained testing protocols with $b \in \mathbb{N}$, $m \in \mathbb{N}$, $d \in 2\mathbb{N}$ and $n \in \mathbb{N}$ and no access to shared randomness for the models Q and P, respectively. We note here that under the conditions of the theorem, we can assume d and n are both larger than some constant; and in particular we can assume $d \in 2\mathbb{N}$ without loss of generality. Assume d and n satisfy (15), for a constant C to be set later. The proof follows by the fact that the distributed testing problems have different minimax testing rates, for certain values of b and m.

Consider the hypothesis test given in (4), with $H_0 : q_0 = (1/d, \ldots, 1/d) \in \mathbb{S}^d$ and H_ρ as in the display.

Set $b = \lceil n \log_2(d) \rceil$. When $b \geq n\log_2(d)$, the observations $\tilde{X}^{(j)}$ in the multinomial model as given in (3) are valid b-bit transcripts, since $|\{1, \ldots, d\}^n| \leq n\log_2(d)$. These transcripts are therefore sufficient for the nondistributed / unconstrained model Q^m, i.e. corresponding to observations

$$\tilde{X} \sim Q_{q,nm} \quad \text{for } q \in \mathcal{F}.$$

Consequently, the distributed, b-bit bandwidth constraint testing risk for Q is equal to the testing risk Q^m;

$$\inf_{T \in \mathscr{T}_Q} \mathcal{R}_Q(H_\rho, T) = \inf_T \mathcal{R}_{Q^m}(H_\rho, T).$$

This means that, for all $\alpha \in (0,1)$, there exists $C_\alpha > 0$ and a distributed protocol T satisfying a b-bandwidth constraint for distributed experiment Q such that

$$\inf_{T \in \mathscr{T}_Q} \mathcal{R}_Q(H_\rho, T) < \alpha \text{ whenever } \rho^2 \geq C_\alpha \frac{\sqrt{d}}{mn}$$

where H_ρ as defined in (4), as the minimax rate for the unconstrained problem with mn observations is $\rho_{Q^m}^2 := \sqrt{d}/(mn)$ (see e.g. Theorem 3 in [70]).

On the other hand, whenever $mb = m\lceil n\log_2(d)\rceil \leq d$, the minimax rate for the distributed testing risk of P for the (comparable) hypotheses

$$H_0 : q = q_0 \quad \text{versus} \quad \tilde{H}_\rho : \|\sqrt{q} - \sqrt{q_0}\|_2 \geq \rho$$

is bounded from below by $\rho_P^2 \asymp \sqrt{d}/(\sqrt{m}n)$, as a consequence of Theorem 4. Specifically, following the proof of Theorem 1 above, we have that

$$\inf_{T \in \mathscr{T}_P} \mathcal{R}_P(H_\rho, T) \geq \inf_{T \in \mathscr{T}_{\tilde{P}}} \mathcal{R}_{\tilde{P}}(\tilde{H}_\rho, T),$$

where
$$\tilde{H}_\rho := \left\{ f \in (I_R/\sqrt{d})^{d/2} : \|f\|_1 \geq \rho \right\},$$

for $I_R := [\sqrt{2}(1-\sqrt{R})/\sqrt{1+R}, \sqrt{2}(\sqrt{R}-1)/\sqrt{1+R}]$, $\tilde{\mathcal{P}}$ is generated by the observations

$$X^{(j)} = f + \frac{1}{\sqrt{n}} Z^{(j)}$$

for $Z^{(j)} \sim N(0, I_{d/2})$, indexed by $f \in (I_R/\sqrt{d})^{d/2}$ and the class $\mathscr{T}_{\tilde{\mathcal{P}}}$ is to be understood as the b-bit bandwidth constraint distributed testing protocols for the model $\tilde{\mathcal{P}}$ and $j = 1, \ldots, m$ machines.

Lemma (6) implies that for all $\alpha \in (0,1)$ the latter is bounded by

$$\alpha - N(0, c_\alpha^{-1/2} d^{-1} \rho^2 \bar{\Gamma}) \left(\tilde{H}_\rho^c \right),$$

for a symmetric, idempotent matrix $\bar{\Gamma} \in \mathbb{R}^{d/2 \times d/2}$ with $d/4 \leq \mathrm{rank}(\bar{\Gamma}) \leq d/2$ $\alpha \in (0,1)$, whenever $\rho^2 \leq c_\alpha \frac{\sqrt{d}}{\sqrt{mn}}$ for some small enough constant $c_\alpha > 0$. By the same analysis as conducted in the proof of Theorem 1 above (using that $n \leq \log(d)$), we find that the second term is at most $\alpha/2$ for $c_\alpha > 0$ small enough. Summarizing, we find in particular that for some constant $c_\alpha > 0$,

$$\inf_{T \in \mathscr{T}_{\mathcal{P}}} \mathcal{R}_{\mathcal{P}}(T, H_\rho) > 1/3,$$

for all $\rho^2 \leq c\sqrt{d}/(\sqrt{m}n)$ and m, n, b, d such that $mb \leq d$, where the number $1/3$ is chosen without particular significance.

Whenever $mb = m\lceil n \log_2(d) \rceil \leq d$,

$$\inf_{T \in \mathscr{T}_{\mathcal{Q}}} \mathcal{R}_{\mathcal{Q}}(H_\rho, T) < 1/6 < 1/3 < \inf_{T \in \mathscr{T}_{\mathcal{P}}} \mathcal{R}_{\mathcal{P}}(H_\rho, T). \tag{40}$$

for some $C_\alpha > 0$ large enough and $c_\alpha > 0$ small enough. Take the constant $C = \lceil C_\alpha^2/c_\alpha^2 \rceil$ such that if $m = C$, it holds that

$$C_\alpha \frac{\sqrt{d}}{mn} \leq \rho^2 \leq c_\alpha \frac{\sqrt{d}}{\sqrt{mn}}, \quad \text{with } \rho^2 := C_\alpha \frac{\sqrt{d}}{\sqrt{M}\sqrt{mn}}.$$

Now suppose that $C\mathfrak{d}(\mathcal{Q}, \mathcal{P}) \leq 1/6$. Lemma 5 then implies in that

$$\inf_{T \in \mathscr{T}_{\mathcal{P}}} \mathcal{R}_{\mathcal{P}}(H_\rho, T) \leq \inf_{T \in \mathscr{T}_{\mathcal{Q}}} \mathcal{R}_{\mathcal{Q}}(H_\rho, T) + 1/6 < 1/3.$$

This contradicts (40). We conclude that

$$C\mathfrak{d}(\mathcal{Q}, \mathcal{P}) > 1/6, \tag{41}$$

whenever $d/\lceil n \log_2(d) \rceil > C$. The result now follows with $c = 1/(6C)$. $\qquad \square$

E   Auxilliary lemmas

The following lemma is used in the comparison of the multinomial model to the many-normal-means model.

Lemma 8. *Let $d \in 2\mathbb{N}$, $\mathcal{F} \subset \mathbb{R}^{d/2}$, and consider for $i = 1, \ldots, d$ independent random variables $X_i = h_i + \sigma Z_i$ with $\sigma > 0$ and $Z_i \sim N(0,1)$ satisfying*

$$h_i = \begin{cases} a_i f_i & \text{if } i \leq d/2, \\ -a_i f_{i-d/2} & \text{if } i > d/2, \end{cases}$$

for some $f \in \mathcal{F}$ and $a = (a_i)_{i \in [d]} \in \mathbb{R}^d$. Let $\mathcal{P}$ denote the model generated by the observations $X := (X_1, \ldots, X_d) \sim P_f$, $f \in \mathcal{F}$ and let $\mathcal{Q}$ denote the model generated by

$$\tilde{X}_i = (a_i + a_{d/2+i}) f_i + \sqrt{2}\sigma Z_i, \quad \text{for } i = 1, \ldots, d/2,$$

with $Z_i \overset{i.i.d.}{\sim} N(0,1)$ and $f \in \mathcal{F}$.

Then, $\Delta(\mathcal{P}, \mathcal{Q}) = 0$.

Proof. The statistic $S = (a_i X_i - a_i X_{d/2+1})_{i \in [d]}$ is sufficient for the model $\mathcal{P}$ by using Neyman-Fisher (Lemma 2). We have

$$\frac{dP_f}{dP_0}(X) = \prod_{i=1}^{d} \exp\left(\sigma^{-1} X_i h_i - \frac{1}{2\sigma^2} h_i^2\right)$$

$$= \prod_{i=1}^{d/2} \exp\left(\sigma^{-1}(a_i X_i - a_i X_{d/2+1})f_i - \frac{1}{\sigma^2}f_i^2\right) \qquad = e^{\sigma^{-1} S^\top f - \frac{1}{\sigma^2}\|f\|_2^2}.$$

In distribution, $\tilde{X} = (\tilde{X}_i)_{i \in [d]}$ is equal to S, which implies $\Delta(\mathcal{P}, \mathcal{Q}) = 0$ per Lemma 2. □

The following lemmas are well known but included for completeness.

Lemma 9. *Let P_f denote the distribution of a $N(f, \sigma I_d)$ distributed random vector for $f \in \mathbb{R}^d$ and let P_f^n denote the distribution of n i.i.d. draws (i.e. $P_f^n = \bigotimes_{i=1}^n P_f$).*

It holds that

$$\left\|P_f^n - P_g^n\right\|_{\mathrm{TV}} \leq \frac{n}{2\sigma}\|f - g\|_2.$$

Proof. By Pinsker's inequality,

$$\|P_f^n - P_g^n\|_{\mathrm{TV}} \leq \sqrt{\frac{n}{2} D_{\mathrm{KL}}(P_f; P_g)}.$$

A straightforward calculation gives that the latter is bounded by $\frac{\sqrt{n}}{2\sigma}\|f - g\|_2$. □

The following lemma relates the total variation distance between P, Q to the L_1-distance between corresponding densities.

Lemma 10. *Let P, Q be probability measures dominated by a sigma-finite measure μ with corresponding probability densities $p = \frac{dP}{d\mu}$ and $q = \frac{dQ}{d\mu}$. It holds that*

$$\|P - Q\|_{\mathrm{TV}} = \frac{1}{2}\int |p(x) - q(x)|d\mu(x).$$

Proof. See e.g. Section 2.4 in [93]. □

Lemma 11. *For any two probability measures P and Q on a measurable space $(\mathcal{X}, \mathcal{X})$ with $\mathcal{X}$ a Polish space and $\mathcal{X}$ its Borel sigma-algebra. There exists a coupling $\mathbb{P}^{X, \tilde{X}}$ such that*

$$\|P - Q\|_{\mathrm{TV}} = 2\mathbb{P}^{X, \tilde{X}}\left(X \neq \tilde{X}\right).$$

Proof. See e.g. Section 8.3 in [91]. □

The next lemma gives a useful characterization of the total variation distance between two probability measures.

Lemma 12. *Let P be a signed, bounded measure defined on measurable space $(\mathcal{X}, \mathcal{X})$ and suppose that $P \ll \nu$ for a sigma-finite measure ν. It holds that*

$$\|P\|_{\mathrm{TV}} = \frac{1}{2}\sup\left\{\int f dP : |f| \leq 1 \text{ and } f : \mathcal{X} \to \mathbb{R} \text{ is measurable}\right\}. \tag{42}$$

Proof. Consider the Jordan measure decomposition $P = P^+ - P^-$, where P^+, P^- are both positive, bounded measures such that $P^+ \perp P^-$. For any measurable f, $\{f \geq 0\}, \{f \leq 0\} \in \mathcal{X}$, so $|f| \leq 1$ means that

$$\int f dP \leq \int f \mathbb{1}_{\{f \geq 0\}} dP^+ - \int f \mathbb{1}_{\{f \leq 0\}} dP^-$$

$$\leq \int \mathbb{1}_{\{f \geq 0\}} dP^+ + \int \mathbb{1}_{\{f \leq 0\}} dP^-$$

$$\leq \|P^+\|_{\mathrm{TV}} + \|P^-\|_{\mathrm{TV}} \leq 2\|P\|_{\mathrm{TV}}.$$

For the other direction, note that $f = \text{sign}(p - q)$ is measurable and bounded by 1, which gives

$$\frac{1}{2} \int f dP = \int |p - q| d\nu = \|P - Q\|_{\text{TV}},$$

where the last equality follows from Lemma 10. □

Lemma 13. *Let $P = \bigotimes_{j=1}^{m} P_j$ and $Q = \bigotimes_{j=1}^{m} Q_j$ for probability measures P_j, Q_j defined on a common measurable space $(\mathcal{X}, \mathcal{X})$, with probability densities p_j, q_j for $j = 1, \ldots m$. It holds that*

$$\|P - Q\|_{\text{TV}} \leq \sum_{j=1}^{m} \|P_j - Q_j\|_{\text{TV}}.$$

Proof. The measures P_j and Q_j admit densities with respect to $P_j + Q_j$, which we shall denote by p_j and q_j, respectively, with

$$p := \prod_{j=1}^{m} p_j = \frac{d \bigotimes_{j=1}^{m} P_j}{d \bigotimes_{j=1}^{m} (P_j + Q_j)} \quad \text{and} \quad q := \prod_{j=1}^{m} q_j = \frac{d \bigotimes_{j=1}^{m} Q_j}{d \bigotimes_{j=1}^{m} (P_j + Q_j)}.$$

Writing $\mu = \bigotimes_{j=1}^{m} (P_j + Q_j)$ and applying Lemma 10 we obtain

$$\|P - Q\|_{\text{TV}} = \frac{1}{2} \int |\prod_{j=1}^{m} p_j(x_j) - \prod_{j=1}^{m} q_j(x_j)| d\mu(x_1, \ldots, x_m). \tag{43}$$

By the telescoping product identity

$$a_1 \cdot a_2 \cdots a_m - b_1 \cdot b_2 \cdots b_m = \sum_{j=1}^{m} (a_j - b_j) \prod_{k=1}^{j-1} a_k \prod_{k=j+1}^{m} b_k \tag{44}$$

and Fubini's Theorem, the right-hand side of (43) is bounded by

$$\sum_{j=1}^{m} \frac{1}{2} \int |p_j(x_j) - q_j(x_j)| d(P_j + Q_j)(x_j) = \sum_{j=1}^{m} \|P_j - Q_j\|_{\text{TV}}.$$

□

The following lemma can be seen as a data processing inequality for the total variation distance.

Lemma 14. *Let $(\mathcal{X}, \mathcal{X})$ and $(\mathcal{Y}, \mathcal{Y})$ be two measurable spaces and let $K : \mathcal{Y} \times \mathcal{X} \to [0, 1]$ be a Markov kernel. For any probability measures P, Q defined on $\mathcal{X}$ it holds that*

$$\|PK - QK\|_{\text{TV}} \leq \|P - Q\|_{\text{TV}}.$$

Proof. This follows immediately from the representation in Lemma 12 combined with the fact that, for $|f| \leq 1$, $x \mapsto \int f(y) dK(y|x)$ is a measurable function bounded by 1, since K is Markov kernel. Hence,

$$\sup_A |PK(A) - QK(A)| = \frac{1}{2} \sup_f \int \int f(y) dK(y|x) d(P - Q)(x)$$

$$\leq \frac{1}{2} \sup_f \int f(x) d(P - Q)(x).$$

□

The next lemma bounds the L_1-distance $\|p - q\|_1$ between densities with a multiple of the Hellinger distance $2^{-1/2} \|\sqrt{p} - \sqrt{q}\|_2$.

Lemma 15. *For two probability densities p, q with respect to μ, it holds that*

$$\frac{1}{2} \int |p(x) - q(x)| d\mu(x) \leq \sqrt{\int \left(\sqrt{p(x)} - \sqrt{q(x)} \right)^2 d\mu(x)}.$$

Proof. The result follow from the Cauchy-Schwarz inequality and the fact that $\int p d\mu = \int q d\mu = 1$. See e.g. [93] for details. $\qquad\square$

Lemma 16. *Let $K \in \mathbb{N}$ and $M \in \mathbb{R}^{K \times K}$ be symmetric and positive definite. Consider the random vector $G = (G_1, \ldots, G_K) \sim N(0, M)$. It holds that $\mathbb{E} \max\limits_{1 \leq i \leq K} |G_i| \leq 3\|M\|\sqrt{\log(K) \vee \log(2)}$ and*

$$Pr\left(\max_{1 \leq i \leq K} G_i^2 \geq \|M\|^2 x\right) \leq \frac{2K}{e^{x/4}},$$

for all $x > 0$.

Proof. It holds that

$$G \overset{d}{=} \sqrt{M}Z, \quad \text{with} \quad Z \sim N(0, I_K).$$

Since M is symmetric, positive definite, it has SVD decomposition $M = V\text{Diag}(\lambda_1, \ldots, \lambda_K)V^\top$. Since V is orthonormal,

$$\sqrt{M}Z = V\sqrt{\text{Diag}(\lambda_1, \ldots, \lambda_K)}(V^\top Z) \overset{d}{=} V\sqrt{\text{Diag}(\lambda_1, \ldots, \lambda_K)}Z.$$

Writing $V = [v_1 \ \ldots \ v_K]$ where v_k are orthogonal unit vectors, the latter display equals

$$\sum_{k=1}^K \sqrt{\lambda_k} v_k Z_k \ \sim \ N\left(0, \text{Diag}(\lambda_1, \ldots, \lambda_K)\right).$$

Consequently,

$$\max_{k \in [K]} |G_k| \overset{d}{=} \max_{k \in [K]} |\lambda_k Z_k| \leq \|M\| \max_{k \in [K]} |Z_k|.$$

Hence, it suffices to show that

$$\Pr\left(\max_{1 \leq i \leq K} Z_i^2 \geq x\right) \leq \frac{2K}{e^{x/4}}.$$

The case where $K = 1$ follows by standard Gaussian concentration properties. Assume $K \geq 2$. For $0 \leq t \leq 1/4$,

$$\mathbb{E}e^{t \max_i (Z_i)^2} = e^t \mathbb{E} \max_i e^{t(Z_i^2 - 1)} \leq K e^{2t^2 + t}.$$

Taking $t = 1/4$ and applying Markov's inequality yields the second statement of the lemma. Furthermore, in view of Jensen's inequality

$$\mathbb{E} \max_i (Z_i)^2 \leq \frac{\log(K)}{t} + 2t + 1,$$

which in turn yields $\mathbb{E} \max_i |Z_i| \leq 3\sqrt{\log(K)}$. $\qquad\square$

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS paper checklist",**
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: Yes.

   Justification: Proofs of all theory are provided.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: Yes.

   Justification: See discussion

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: Yes.

   Justification:

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: NA

   Justification: No code / experiments are provided.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: NA

   Justification: No code.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: NA

   Justification:

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: NA

   Justification:

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: NA

   Justification:

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.

- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: Yes

Justification:

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: NA

Justification: The paper focuses on theoretical developments in statistical models without negative societal applications.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: NA

Justification: All the results are of theoretical nature; no data or models are released.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: NA

Justification:

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: NA

Justification:

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: NA

Justification:

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: NA

    Justification:

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.