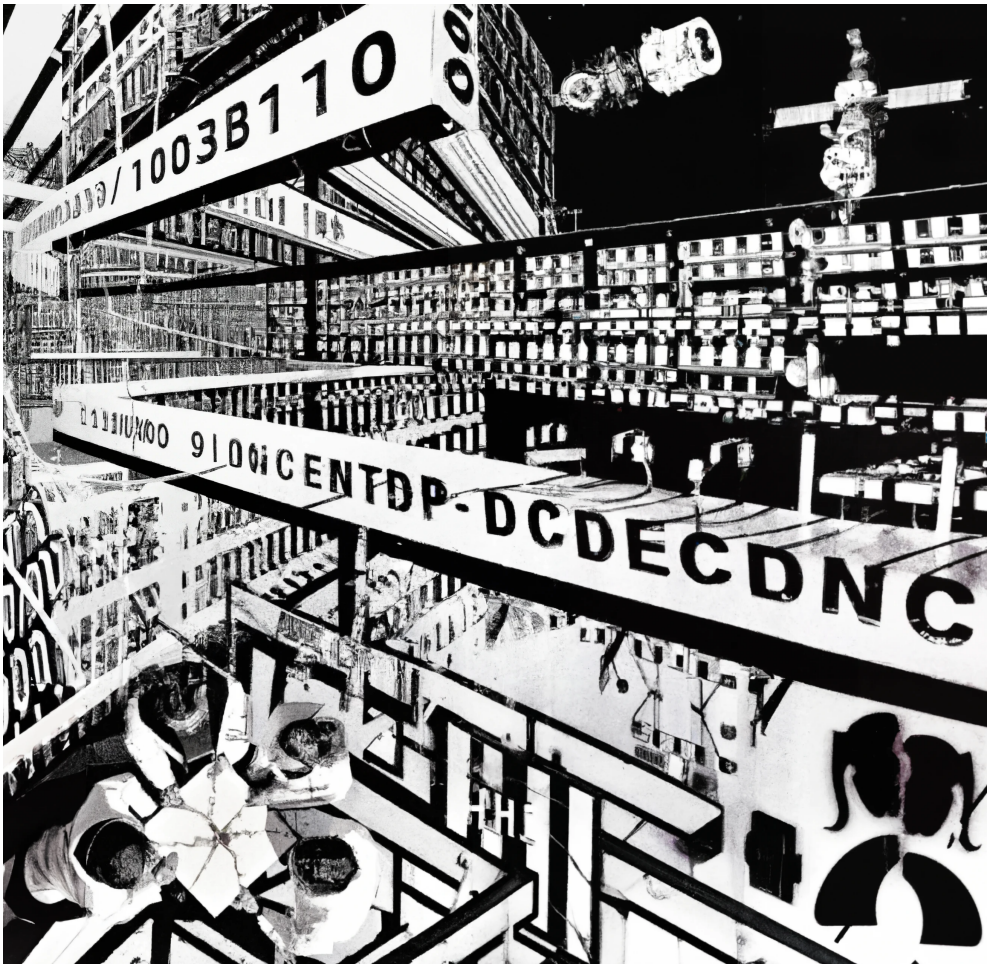
Distributed Hypothesis Testing under Privacy and Bandwidth Constraints

Lasse Vuursteen



Distributed Hypothesis Testing under Privacy and Bandwidth Constraints

Proefschrift

ter verkrijging van de graad van doctor
aan de Technische Universiteit Delft,
op gezag van de Rector Magnificus Prof.dr.ir. T.H.J.J. van der Hagen,
voorzitter van het College voor Promoties,
in het openbaar te verdedigen op
datum om tijdstip.

door

Lasse VUURSTEEN

Master of Science in Stochastics and Financial Mathematics
Vrije Universiteit Amsterdam, Nederland
geboren te Groningen, Nederland

Samenstelling promotiecommissie:

Rector Magnificus	Voorzitter
Prof.dr. A.W. van der Vaart	Technische Universiteit Delft, promotor
Prof.dr. J.H. van Zanten	Vrije Universiteit Amsterdam, promotor
Dr. B.T. Szabó	Bocconi University, promotor

*Dit onderzoek werd gedeeltelijk gefinancierd door de Spinozapremie, toegekend aan A.W. van der Vaart door de Nederlandse Organisatie voor Wetenschappelijk Onderzoek (*NWO*).*

COLOPHON

Lasse Vuursteen
Distributed Hypothesis Testing under Privacy and Bandwidth Constraints, Delft, 2024

Een electronische versie van dit proefschrift is beschikbaar op:

http://repository.tudelft.nl/

Contents

Chapter 1

Introduction

In this thesis, we study distributed hypothesis testing under bandwidth and privacy constraints. Hypothesis testing is concerned with making a decision on the truth or falsehood of a statement on the basis of data that may provide evidence in support of, or against said statement. In a distributed setting, however, such data is not readily available in one "location", meaning that there is no "central" access to the complete data. This scenario commonly occurs when data is observed or stored at multiple locations, such as hospitals, sensors or servers, from which the data cannot be shared in full because of communication limitations. Constraints on the sharing of data arise for various reasons, such as limited bandwidth, issues concerning privacy or ownership of the data.

Hypothesis testing enjoys a large body of classical literature studying theoretically optimal performance in terms of statistical power, the ability to correctly discern the falsehood of a statement based on data. Nevertheless, in the presence of communication constraints, classical statistical methods that are designed with having full access to the data in mind, no longer apply. Distributed methods aim to overcome these barriers by providing mechanisms that operate within these limitations on what can bed communicated, by e.g. preserving privacy or using only a limited amount of bandwidth.

When it comes to the performance of these methods, the communication constraints might severely affect the quality of the statistical inference, for instance by diminishing the statistical power that could be obtained under availability of the complete data. For example, a technique that does not abide by privacy constraints has full utility in the sense that it can give the classically optimal "full data" answer, whereas maintaining full privacy may prevent the data owner from disclosing anything about the data at all. Thus, there is often a trade-off between the quality of statistical inference and privacy. Similarly, in order to satisfy a bandwidth constraint, the data may need to be compressed, which could result in loss of information and consequently a

worse performance. Another example can be found in meta-analysis, where combining test statistics or test outcomes from independent trials or experiments is a popular method when only the outcome of studies are published (e.g. a test outcome or p-value). This can be seen as a communication constraint and a form of compression also, which similarly might result in a loss of statistical power.

This introduces an important question that lies at the heart of this thesis: What is the anticipated loss of statistical power under communication constraints? What are the possibilities and limitations when operating under bandwidth constraints, when assuring a certain level of privacy, or when performing meta-analysis solely on the basis of study outcomes? The answers to these questions, i.e. knowing what is theoretically possible and quantifying the impact of communication constraints on performance is of great importance when conducting statistical analysis within such contexts.

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. It enables a conscious choice in terms of the amount of privacy that is sacrificed for the sake of accuracy in the data analysis and gives insight in how to design studies such that high statistical power can be combined with adequate privacy guarantees.

The capacity to transmit data does not match the capacity to generate or process data in many modern applications. By knowing the bounds of what can and cannot be achieved, systems can be designed to work as efficiently as possible within such bandwidth constraints. It allows organizations to make informed decisions about where and how much to invest in infrastructure. Furthermore, for inherently bandwidth constraint settings such as voting systems or meta-analysis on the basis of test outcomes, it is important to understand to what extent substantial statistical power can be expected at all.

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. [199, 20, 203, 33, 133, 197] or "inference under multiterminal compression" (see [198] 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. With the advent of the internet and big data, interest in distributed methods with bandwidth constraints increased further. Modern machine learning settings often concern settings where inference is centralized, while training data remains distributed over numerous clients. Examples of such "federated learning" or "edge computing" settings are siloed data centers, such as hospitals, or networks of cellphone users, in applications such as word prediction, face or voice recognition, Siri or Google Assistant, driverless cars or even earthquake prediction [142, 130, 114, 31, 160, 63]. In such settings, bandwidth often forms a limited/costly resource, or an outright bottle-

neck [131].

Similarly, with advances in electronic record keeping, privacy has become a more and more pressing issue in the modern era. In various scientific fields, there is an increased awareness of privacy issues, for example in the medical sciences [135] or social sciences [162]. In the internet era, also societal awareness towards privacy issues has been heightened, paralleling the rise of consumer engagement with tech industry products [66], including many of the federated learning applications mentioned in the previous paragraph.

Methods that preserve privacy have been around in the statistics community for some time, starting in the 1980s [80, 81]. The current leading formal privacy framework is that of *differential privacy*, as introduced in [82]. Differential privacy 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 [93], Microsoft [74], Apple [25], as well as governmental entities such as the US Census Bureau [175].

Rigorous study of performance under bandwidth constraints has been mostly conducted for estimation problems. Bandwidth constraints have been studied for the many-normal-means and parametric models in e.g. [229, 77, 180, 39, 219, 113, 46, 45], as well as nonparametric models, including Gaussian white noise [230], nonparametric regression [187], density estimation [30], general, abstract settings [191] and online learning [207]. Distributed adaptive estimation methods under bandwidth constraints, where adaptation occurs to the unknown regularity of the functional parameter of interest, were derived in [187, 188, 47].

For distributed testing under bandwidth constraints, much less is known. In [6], the authors consider a setting in which each machine obtains a single observation from a distribution on a finite sample space and derive lower bounds for testing uniformity of this distribution. Similar distributed uniformity testing is considered in [7], where matching upper bounds are exhibited for this setting. [12] derives bounds that are optimal for the Gaussian setting in case of 1-bit bandwidth constraints for a single observation of the many-normal-means model.

The literature on the theoretical properties of differential privacy can be mostly divided into those studying *local* differential privacy or *central* differential privacy. In local differential privacy, the privacy protection is applied at the level of individual data entries or observations. This is a stringent form of differential privacy because each "item" of data is independently given privacy protection. In the other extreme, central differential privacy, only the inference output needs to satisfy the differential privacy constraint, meaning that if the output is a test, only the final decision needs to satisfy a differential privacy constraint.

Distributed estimation under local differential privacy has been studied for the many-normal-means model, discrete distributions and parametric models in [78, 79, 5, 223],

and density estimation [176, 134, 43]. Testing under local differential privacy has been studied for discrete distributions in [103, 181, 3, 4, 34, 9, 15] and nonparametric densities in [136]. In [44], the authors consider estimation of a quadratic functional under local differential privacy constraints, which has connections to goodness-of-fit testing. Estimation in the central setting was investigated in [127, 48, 50]. Private testing in the central setting, where all data is available on a single machine, has been studied by [2, 54, 157] for discrete distributions and the many-normal means model.

In-between local differential privacy and central differential privacy, there are general distributed settings. Here, privacy is applied at a level 'sample' level (sometimes called 'user' level), where for example different entities, such as hospitals, are concerned about sharing their data with other entities due to privacy concerns for their patients. In such settings, differential privacy needs to apply at the level of the local sample, e.g. the patient pool of each hospital. Investigations into this more general setting have been much more limited, with only estimation being considered in [147, 18], which study estimation for discrete distributions and [141, 158] which study mean estimation.

Contrary to some literature that defines "communication constraints" solely as bandwidth constraints, we we broaden the term to include both bandwidth and privacy constraints. This use of the terminology is appropriate when considering that both types of constraints, both limit information sharing, albeit in different ways. Furthermore, both types of communication constraints will be studied using similar mathematical techniques.

This brings us to specifying the aim of this thesis.

1.0.1 The aim of the thesis

The aim of this thesis is to mathematically characterize and quantify the impact of bandwidth and differential privacy communication constraints in distributed hypothesis testing. That is to say, we wish to gain insight into how the statistical problem of testing gets more difficult depending on the severity of the communication constraint. The statistical task this thesis centers around is that of hypothesis testing, where the type of hypothesis test we shall consider will be that of "signal detection" or "goodness-of-fit" testing, where we wish to decide between the *null hypothesis* that the data is generated by a particular specified "null" probability distribution, versus the *alternative hypothesis* that the data is generated by some other probability distribution belonging to a family of alternatives.

The theory shall be derived in an abstract setting in which we have m locations (e.g. hospitals, sensors, servers, etc.) which we shall refer to as machines. Each of the $j = 1, \dots, m$ machines communicates a transcript $Y^{(j)}$ on the basis of a local independent sample of $i = 1, \dots, n$ data points $X_i^{(j)}$ drawn from an unknown distribution. Each transcript $Y^{(j)}$ has to satisfy a certain communication constraint. In case of a bandwidth constraint, the transcript $Y^{(j)}$ may contain at most b-bits of information

(details are given in Section 1.2.1). In case of privacy constraint, the transcript $Y^{(j)}$ must satisfy a differential privacy constraint governed by two parameters ϵ and δ, where smaller values for ϵ and δ give stronger privacy guarantees (we defer the details to Section 1.2.2).

In modern applications, it is common that the number of data samples is small compared to the dimension of statistical model, such as in functional data or large histograms. For some settings, the dimensionality is a known, finite number, say $d \in \mathbb{N}$. For nonparametric models, data is infinite dimensional and appropriate statistical techniques in such cases require *adaptation*. In such adaptive settings, the certain *hyperparameters* such as the "effective dimension" or "regularity" are unknown, and in order achieve optimal performance, the statistical procedure should be able to *adapt* to the unknown hyperparameter in a data driven way.

The theory in this thesis aims to capture the difficulty of the testing problem in terms of the characteristics of the underlying model and statistical setting. That is to say, to capture the difficulty as a function of b in the bandwidth constraint setting and ϵ and δ in the differential privacy constraint setting, as well as m, n and d (or in terms of the regularity hyperparameter "s" if the model is nonparametric).

In order to effectively do so, we restrict ourselves to certain canonical statistical models. The main focus is the d-dimensional many-normal-means model, which offers the benefit of tractable analysis whilst also capturing the principle phenomena of testing under communication constraints. As a canonical nonparametric setting, we shall consider *goodness-of-fit testing* in the signal-in-white-noise model, where we investigate the case where the regularity of an underlying signal is unknown. Here, goodness-of-fit testing is to be understood in the sense of [123], which bares a close relationship with "classical" nonparametric goodness-of-fit testing in the sense of [23, 182, 67, 211] (see e.g. Section 1.4 in [123] and the introduction to Chapter 5).

These models typically serve as benchmark models for many other models in the statistical literature, as it has been a long-standing and consistent finding that models that describe seemingly very different data and dynamics, can still be subject to very similar phenomena, such as the asymptotic minimax risk coinciding as the number of samples grows. That this holds in the distributed communication constraint setting as well finds mathematical substantiation in Chapter 6. Leveraging existing results on distance between statistical models, it is shown that the detection boundary for the Gaussian model occurs in certain other models, such as for discrete distributions but also more complicated statistical models as well.

Lastly, throughout the thesis, we contrast the results in the thesis for testing with known results for estimation under communication constraints. There are many connections between estimation and testing. However, classically the high-dimensional testing problem is fundamentally different from estimation. We uncover even more fundamental differences between estimation and testing which occur under the presence of communication constraints.

1.0.2 On the structure of the thesis

This section provides an overview of the layout of the thesis and describes how the progression of ideas is structured.

The introduction continues with formally outlining the problem of hypothesis testing under communication constraints. Here, some pages are spent outlining and motivating the minimax paradigm for goodness-of-fit hypothesis testing. Then, the distributed setting is formally introduced, where we formally specify what the bandwidth constraints and privacy constraints that will be considered in the thesis are, and give a motivation for them.

Chapters 2 and 3 together establish the key optimality results which lay the foundation for the concepts and results in the subsequent chapters. Chapter 2 establishes the fundamental boundaries on what is possible performance wise within the statistical problem of signal detection in the Gaussian model. It provides insights into what levels of performance under communication constraints are theoretically unattainable, with an emphasis on the techniques that are required to show this. On the other hand, Chapter 3 presents various methods that successfully achieve the performance boundary outlined in Chapter 2, demonstrating approaches and techniques that are optimal in terms of the bounds laid out in Chapter 2. These two chapters can be read in any order, but together form the theoretical basis for much of the later chapters.

Chapter 4 establishes a link between the bandwidth constrained setting and meta-analysis, from which optimality results for meta-analysis are obtained. Chapter 5 concerns nonparametric goodness-of-fit testing, in which the dimension of the model is infinite and the quality of inference is dependent on an unknown hyperparameter which requires so called 'adaptive methods' in order to attain optimal inference. Chapters 4 and 5 can be read independently of each other.

The final chapter, Chapter 6, is dedicated to extending the results from the first four chapters to other statistical models using Le Cam theory. It is perhaps best saved for last as this chapter uses results from Chapters 2, 3 and 5, unless one wants a sneak peek into the broader applicability of the results beyond the Gaussian models of the earlier chapters.

In each chapter, proofs are given for results whenever the result is key, the technique novel, or when the proof is short and (perhaps) insightful. Some proofs are moved to the appendix, of which each chapter has its own. These proofs are typically either of a (purely) technical nature or concern well known results and therefore only included for completeness.

Unless indicated otherwise, all results in this thesis are original in the sense that they are based on original proofs or original combinations of existing results by the author and collaborators. They are based on joint works [193, 195] with Botond Szabó and Harry van Zanten,[194] with the aforementioned authors and Aad van der Vaart, and

the works [52, 51] with T. Tony Cai and Abhinav Chakraborty. The final chapter is based [215], which benefited from comprehensive proofreading by Aad van der Vaart.

1.1 Statistical hypothesis testing: formal preliminaries

> "*A good technical writer, trying not to be obvious about it, says everything twice; formally and informally. Or maybe three times.*" -Donald E. Knuth

In this thesis, we shall study hypothesis testing through the lens of statistical decision theory as developed by [159, 216]. Statistical decision theory is concerned with decision-making under uncertainty, where the decision is to be made on the basis of observed data. For a more general introduction see e.g. [32]. In what follows, let the random variable X, defined on some measurable space $(\mathcal{X}, \mathscr{X})$, denote the observed data. We shall consider a family of possible probability distributions of the data, $\mathcal{P} = \{P_f : \mathscr{X} \to [0,1] \,|\, f \in \mathcal{F}\}$, which we refer to as *the statistical model*. The indexing set $\mathcal{F}$ shall be called the *parameter space*.

The statistician is concerned with designing a rule of how to make the decision based on the data. In order to assess the quality of the decision rule, we shall be concerned with the distribution of outcomes from the decision rule under the distributions $P_f \in \mathcal{P}$ as if P_f was the true distribution truly generating the data.

A hypothesis test is a particular kind of statistical decision problem, where the statistician needs to decide whether to reject a statement called the *null hypothesis*, which is a statement about the distribution of the data. We shall be concerned with *goodness-of-fit testing for a simple null hypothesis*, where the null hypothesis is a statement of the form: "X follows the distribution P_{f_0}", where $f_0 \in \mathcal{F}$ is a fixed element. Such a null hypothesis could represent the scenario where a disease or defect is absent, or the case where a treatment has no effect.

The opposing statement of the null hypothesis is called the *alternative hypothesis*, which states that data comes from some other distribution P_f, belonging to a class of alternative distributions $H_1 \subset \mathcal{P} \backslash \{P_{f_0}\}$. The alternative hypothesis is sometimes what one hopes or might expect to establish, for example that a treatment has an effect of a certain magnitude or the presence of a signal. When H_1 consists of more than one element, it is called a *composite alternative hypothesis*.

This type of *goodness-of-fit* testing is what we shall be concerned with in this thesis. It corresponds to situations in which one wishes to assess how well one particular "explanation" of the data fares against a class of alternative explanations. For example, in linear regression, one might want to test whether a group of explanatory variables with corresponding vector of coefficients f have an effect on outcome variable, which could be expressed as testing the null hypothesis that $f = 0$ versus the alternative hypothesis of $f \neq 0$. Or, one might want to test whether the data is distributed

according to some particular density, $H_0 : f = f_0$, versus the alternative hypothesis that f belongs some specified (nonparametric) class of densities. For a more general discussion of hypothesis testing problems, we refer the reader to [139, 123].

When testing a null hypothesis, making the incorrect decision comes in two flavors. In one case, the statistician could decide to reject the null hypothesis whilst it is true: X does in fact follow the distribution P_{f_0}. This mistake; of incorrectly rejecting the null hypothesis, shall be referred to as a *Type I error*. The other kind of mistake, is to not reject the null hypothesis, even though it is false. This shall be referred to as a *Type II error*.

A decision rule deciding between the two hypotheses shall be referred to as a test. Formally, a *test* T is a random variable, possibly depending on the data, taking values in a space of cardinality two, e.g. $\{\text{DO NOT REJECT}, \text{REJECT}\}$. Given a test of a null hypothesis "X follows the distribution P_{f_0}", the *Type I error probability of* T or the *level of* T is $\mathbb{P}_{f_0}(T = \text{REJECT})$. Here, we use $\mathbb{P}_f$ to denote the probability distribution governing both T and X, where X is marginally distributed according to P_f. The probability of making the correct decision given a test T and $P_f \in H_1$, i.e. $\mathbb{P}_f(T = \text{REJECT})$, is called the *power of the test* T *against* P_f. Similarly, the *Type II error probability of the test* T *under* P_f is $\mathbb{P}_f(T = \text{DO NOT REJECT})$. To assess the power or probability of making a mistake of the second kind against the alternative hypothesis as a whole, one needs to specify which "alternative" distribution $P_f \neq P_{f_0}$ is under consideration, since $\mathbb{P}_f(T = \text{DO NOT REJECT})$ might vary across different $P_f \in H_1$. Given the class H_1 and a test T, the *worst-case Type II error probability* is

$$\sup_{P_f \in H_1} \mathbb{P}_f(T = \text{DO NOT REJECT}).$$

We shall concern ourselves with studying how well an alternative hypothesis can be distinguished from the null hypothesis in terms of achieving "minimal" worst-case Type II error probability with tests that also have a "minimal" Type I error probability. While this approach may appear cautious, doing so provides assurances for differentiating between the null hypothesis and the entire set of alternative hypotheses, without being tied to a specific distribution within the alternative class. This *statistically guarantees* validity of the procedure, irrespective of the a priori unknown truth.

Given a class of alternatives, it is natural to ask what is the best possible testing performance, in terms of worst-case Type II error probability. Given a level $\alpha \in (0, 1)$, we quantify the best possible performance by the *minimax Type II error probability for tests of level* α, which we define as

$$\beta_{\mathcal{P}}(\alpha, H_0, H_1) = \inf_T \sup_{P_f \in H_1} \mathbb{P}_f(T = \text{DO NOT REJECT}),$$

where the infimum is over all tests of level at most α.

The choice of the class of alternatives H_1 is important to the statistical analysis. If one takes "too small" of a class, it might mean that it does not include, or is in some sense

"too far" from distributions which in reality could explain the data. When testing a theory, it is desirable to rule out alternatives that are "close to the theory" or may produce behavior that "looks like the theory" but that are in fact not the theory at all. On the other hand, if one chooses "too large" of a class for the alternative hypothesis, it might not be distinguishable from the null hypothesis at all, which in turn leads to unduly conservative expectations regarding the optimal performance of a test. This brings us to the concept of minimax separation between the hypotheses.

In what follows, consider a collection of alternative hypotheses H_ρ indexed $\rho > 0$, such that $H_{\rho'} \subset H_\rho$ for $\rho \leqslant \rho'$. We will call ρ the *separation between* H_0 *and the alternative hypothesis* (H_ρ). The map $\rho \mapsto \beta_{\mathcal{P}}(\alpha, H_0, H_\rho)$ is a decreasing function, meaning that the testing problem increases in difficulty as ρ decreases. Now, let us consider a collection of models $\{\mathcal{P}_\nu : \nu \in \mathcal{I}\}$, where each model $\mathcal{P}_\nu$ has a corresponding null hypothesis $H_0 \equiv H_{0;\nu}$ and alternative hypotheses, $H_\rho \equiv H_{\rho;\nu}$, $\rho > 0$. The *minimax separation at* $\alpha, \beta \in (0,1)$ is given by

$$\rho^*_{\nu,\alpha,\beta} := \inf\left\{\rho > 0 : \beta_{\mathcal{P}_\nu}(\alpha, H_0, H_\rho) \leqslant \beta\right\}, \quad \nu \in \mathcal{I}.$$

The minimax separation effectively captures what is the smallest degree of separation at which the null hypothesis can be distinguished from the alternative hypothesis, as the alternative hypothesis "grows closer" to H_0 as ρ decreases. The index ν can be related to certain features of the model, such as the number of observations, the dimension of the data or other characteristics corresponding to the model $\mathcal{P}_\nu$. This is exemplified at the end of the section, where we present the many-normal-means model. In this model, the relevant characteristics are the dimension of the data d and the number of observations n. The minimax separation for the many-normal-means model as a function of d and n tells us how the statistical problem depends on these model characteristics. For instance, it could tell us what the gain in terms of power is when we obtain additional observations, or how much more complicated the problem becomes as the dimension grows.

Often, the minimax separation is characterized by giving upper and lower bounds on it, which we shall express here as a function of ν. This provides a coarser lens, but in somewhat complex statistical models, such upper and lower bounds on the minimax separation are the best that one can hope for. We shall call a nonnegative function $\nu \mapsto \rho_{\nu,\alpha,\beta}$ an *upper bound for the minimax separation rate* whenever $\rho^*_{\nu,\alpha,\beta} \lesssim \rho_{\nu,\alpha,\beta}$. It is a *lower bound for the minimax separation rate* if $\rho_{\nu,\alpha,\beta} \lesssim \rho^*_{\nu,\alpha,\beta}$. It is 'the' *minimax separation rate* whenever $\rho^*_{\nu,\alpha,\beta} \asymp \rho_{\nu,\alpha,\beta}$. In slight abuse of terminology, we shall sometimes simply refer to the minimax testing rate, or the minimax rate, whenever the context should be clear. Grasping the minimax rate offers insights into the dynamics of a statistical problem as its attributes change. The minimax testing framework, as developed in [94, 121, 140, 118], provides a robust foundation for addressing hypothesis testing within complex statistical contexts, such as high-dimensional and nonparametric settings.

In problems where data consists of multiple independent, identically distributed observations that increase as (a function of) ν, it (typically) makes sense to analyze the *testing risk of a test* T,

$$\mathcal{R}_{\mathcal{P}_\nu}(H_{\rho_\nu}, T) := \mathbb{P}_{f_0;\nu}(T = \text{reject}) + \sup_{P_{f;\nu} \in H_{\rho_\nu}} \mathbb{P}_f(T = \text{do not reject})$$

and study sequences of models $\mathcal{P}_\nu$ and ρ_ν such that the *minimax testing risk* $\inf_T \ \mathcal{R}_{\mathcal{P}_\nu}(H_{\rho_\nu}, T)$ tends to either zero or one. This is easier to analyze instead of the minimax separation at different $\alpha, \beta \in (0,1)$ and yields (typically) the same conclusion. To see this, suppose that $\beta_{\mathcal{P}_\nu}(\alpha, H_0, H_{\rho_\nu}) \leqslant \beta$ for all ν large enough, for some $\beta \in (0,1)$ such that $\alpha + \beta < 1$. Under (typical) assumptions on the minimax rate, any other desired level α' and level worst-case Type II error probability $\beta' \leqslant \beta$ can be achieved by repeating a testing procedure of level α and worst-case Type II error probability β a constant number of times (see e.g. Lemma 3.25 in the appendix of Chapter 3 for a more precise statement).

We exemplify the minimax framework outlined above with a statistical problem that is canonical for goodness-of-fit-testing; signal detection in the many-normal-means model. The many-normal-means model postulates that we observe data X satisfying

$$X = f + \frac{1}{\sqrt{n}} Z, \tag{1.1}$$

where $f \in \mathbb{R}^d$ is the unknown signal, Z is an unobserved noise vector with a d-dimensional standard Gaussian distribution, and $1/n$ is the signal-to-noise ratio. Note that this is equivalent to observing n independent copies of a $N_d(f, I_d)$-vector. The corresponding statistical model $\mathcal{P}_{n,d}$, indexed by the parameter set $\mathbb{R}^d$, consists of distributions P_f such that X is governed by (1.1) given $f \in \mathbb{R}^d$. Testing for the presence of a signal in the normal-means model translates to testing the null hypothesis $H_0 : f = 0$ that the sequence is identically equal to 0. Rejecting this hypothesis means declaring that there is a non-zero signal. The difficulty of distinguishing between the two hypotheses depends on signal strength, the noise ratio n and dimension d. Given separation $\rho > 0$, one could translate this to the test of hypotheses

$$H_0 : f = 0 \ \text{ versus } H_\rho : \|f\|_2 \geqslant \rho, \tag{1.2}$$

The separation ρ tells us for what signal size (by which we mean the Euclidean norm of f) a signal can be meaningfully distinguished from 0. For this testing problem, it is known that the minimax separation satisfies

$$c_{\alpha,\beta} \frac{\sqrt{d}}{n} \leqslant \left(\rho^*_{n,d,\alpha,\beta}\right)^2 \leqslant C_{\alpha,\beta} \frac{\sqrt{d}}{n},$$

where $c_{\alpha,\beta}, C_{\alpha,\beta} > 0$ are constants depending only on the tolerated Type I and Type II errors α and β. This can be found in e.g. [29], but it also follows as a corollary to results in this thesis. The (squared) minimax rate for this problem is consequently given by $\sqrt{d}/n$.

The quantity $\sqrt{d}/n$ indicates how the difficulty of the statistical problem changes as the dimension d and number of observations n change. The factor $\sqrt{d}/n$ in this case, indicates that the as the number of observations increases faster than the square root of the dimension, we can guarantee detection of signals of a smaller size, since every $f \in \mathbb{R}^d$ with $\|f\|_2^2 \geqslant C_{\alpha,\beta}\frac{\sqrt{d}}{n}$ can be detected with probability at least β by a test of level at most α, where $C_{\alpha,\beta} > 0$ is a constant depending only on α and β. On the other hand, the class of signals with $\|f\|_2 \leqslant c_{\alpha,\beta}\frac{\sqrt{d}}{n}$ cannot be distinguished from 0 with testing risk smaller than $\alpha + \beta$, whenever $c_{\alpha,\beta} > 0$ is small enough. This means in particular that, given sequences $n \equiv n_\nu$, $d \equiv d_\nu$ and $\rho \equiv \rho_\nu$, we cannot *consistently* distinguish classes of signals f such that $\|f\|_2^2 \leqslant \rho^2 = o(\sqrt{d}/n)$ from $f = 0$. The square root of the dimension here is particularly interesting, as it implies that certain small signals can still be detected even when the dimension is larger than the number of observations. This is not the case for estimation in the Euclidean norm, where we find that we cannot consistently estimate if d is of larger order than n, as we shall discuss next.

In the canonical estimation problem in the many-normal-means model, one is concerned with finding an estimate of the true parameter f when the data X is generated by P_f. An *estimator* is a measurable function $\hat{f}$ on the sample space[1] taking values in $\mathcal{F} = \mathbb{R}^d$. Loosely speaking, a good estimator uses the data to produce an estimate that is (hopefully) "close" to the true parameter f. The minimax estimation risk for the Euclidean norm is given by

$$\inf_{\hat{f}} \sup_{f \in \mathbb{R}^d} \mathbb{E}_f \left\| \hat{f}(X) - f \right\|_2^2,$$

where the infimum is taken over all measurable functions $\hat{f} : \mathcal{X} \to \mathcal{F}$. It can be shown that the above expression is bounded between cd/n and Cd/n for fixed constants $c, C > 0$ (see e.g. [204]). This means that, for sequences $n \equiv n_\nu$, $d \equiv d_\nu$ such that d/n converges to a constant (or diverges), no estimator is *guaranteed* to converge to the true underlying parameter, even as n tends to infinity. Interestingly, even as the estimation problem becomes easier with larger n, the dimension increasing at an equal or faster rate prevents the estimation problem from having a consistent solution. However, as previously discussed, for such sequences d and n, the testing problem *is* solvable in the sense that signals above the $\sqrt{d}/n$ threshold can be detected as n tends to infinity. In particular, when $\sqrt{d}/n \to 0$ whilst $d/n \gtrsim 1$, consistent testing is possible, whilst consistent estimation is not. The testing problem is *easier* than the estimation problem in the sense that it is possible to distinguish between the null-hypothesis and the alternative, even when the seperation between them is much smaller than the possible accuracy of estimation. Thus, optimal testing requires different procedures than those used for estimating unknown parameters.

[1] Or a possibly enlarged probability space allowing for $\hat{f}$ to be a random measurable function of the data.

We shall continue the study of signal detection in the many-normal-means model in the distributed setting under communication constraints in Chapters 2, 3 and 4. The reason for studying this Gaussian problem, besides it being of practical importance, is that it is simple from technical point of view yet allows for exhibition of various principle phenomena that occur in the distributed setting. Using Le Cam's theory of experiments and asymptotic equivalence, Chapter 6 shows that some of the main results in the many-normal-means model translate to other, more complicated statistical models, such as regression and density testing, both parametric and nonparametric.

1.1.1 Notation and notions

Most of the notation used in the thesis shall be introduced throughout the text itself. Certain notions will be used so frequently that it makes sense to briefly go through them here.

We shall frequently abuse notation when speaking about sequences, where "a sequence a_k" is used to refer both to the collection $\{a_k\}_{k\in\mathbb{N}}$ as well as individual elements of the sequence a_k, where the reader is to discern which is which based on the context. For two positive sequences a_k and b_k, let $a_k \ll b_k$ denote that $a_k/b_k = o(1)$. For two nonnegative functions f, g defined on the same domain D, let $f \lesssim g$ if the inequality $f(x) \leqslant Cg(x)$ holds for all $x \in D$ for some universal positive constant C. Similarly, we write $f \asymp g$ if $f \lesssim g$ and $g \lesssim f$ hold simultaneously. Every once in a while, such notations will be used in the thesis without providing proper function notation, for example by saying "when $mb \lesssim d$". Such a statement is then to be understood as: "for all sequences $m \equiv m_k$, $b \equiv b_k$ and $d \equiv d_k$ such that $mb \lesssim d$".

We use the notations $a \vee b$ and $a \wedge b$ for the maximum and minimum, respectively, between two real numbers a and b. For $k \in \mathbb{N}$, $[k]$ shall denote the set $\{1, \dots, k\}$. Throughout, c and C denote universal positive constants which value can differ from line to line. The Euclidean norm of a vector $v \in \mathbb{R}^d$ is denoted by $\|v\|_2$ and its i-th coordinate by $(v)_i$. Throughout, $I_d \in \mathbb{R}^{d\times d}$ denotes the identity matrix and $\iota_d \in \mathbb{R}^d$ the vector of all ones. For a subset V of a vector space and scalars $\gamma \in \mathbb{R}$, the set γV is to be understood as $\{\gamma v : v \in V\}$. For vectors $v = (v_1, \dots, v_k) \in V^k$, let $\overline{v}$ denote their average, i.e. $k^{-1}\sum_{i=1}^k v_i$. Given a matrix $M \in \mathbb{R}^{d\times d}$, the norm $M \mapsto \|M\|$ is the spectral norm and $\mathrm{Tr}(M)$ is its trace.

We shall define the total *variation distance* between two probability measures P and Q defined on a measurable space $(\mathcal{X}, \mathscr{X})$ as

$$\|P - Q\|_{\mathrm{TV}} := \sup_{A\in\mathscr{X}} |P(A) - Q(A)|. \tag{1.3}$$

For two sigma algebras $\mathscr{X}, \mathscr{Y}$, we let $\mathscr{X} \otimes \mathscr{Y}$ denote the smallest sigma algebra containing $\mathscr{X} \times \mathscr{Y}$. 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 $A \in \mathscr{Y}$.
- 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 $\mathbb{P}^S$. For statistical models defined on a measurable space $(\mathcal{X}, \mathscr{X})$, we shall use regular capital letters such as P_f, with f typically indicating the indexing parameter $f \in \mathcal{F}$. In these cases, for a measurable function $h : \mathcal{X} \to \mathbb{R}$, $P_f(h)$ is to be understood as the expectation of h, i.e. $\int h(x) dP_f(x)$. 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

$$d\mathbb{P}^{(X,Y,Z)} = d\mathbb{P}^X d\mathbb{P}^{Y|X} d\mathbb{P}^{Z|Y}.$$

In displays such as the one above this sentence, we shall use the short hands "left-hand side" and "right-hand side" to refer to the left-hand side and right-hand side of the relational operator (e.g. the equality sign) in the display, respectively.

Certain terminology shall be used purely to indicate the "role" of the object within the statistical framework considered. For example, the words *sample space* and *decision space* are both to be understood "just" as measurable spaces; the terminology indicating the role they play in the statistical decision problem. In that light, a *statistic* is nothing more than a measurable map between two measurable spaces, the measurable space of the data is called the *sample space*. When we are concerned with hypothesis testing, we shall consider the decision space $\{0, 1\}$, where 0 corresponds to DO NOT REJECT and 1 with REJECT. A *test* is then simply to be understood as a statistic taking values in $\{0, 1\}$.

We shall use the notions of σ-sub-Gaussian and σ-subExponential random variables as defined in [210] and use stochastic-O-notation "O_P" and "o_P" as defined in [205].

1.2 Distributed inference

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 $j = 1, \ldots, m$ machines, each receiving data $X^{(j)}$ drawn from a given distribution $P_f \in \mathcal{P}$. Each of the machines 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. In case of a hypothesis test, we shall call the combination of the process generating the transcripts and the test based on the transcripts a distributed testing protocol.

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 $K^j : \mathscr{Y}^{(j)} \times \mathcal{X} \times \mathcal{U} \to [0, 1]$ 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 test T decides on the basis of the transcripts generated from the data by the kernels $\{K^j\}_{j=1,\dots,m}$, which form the conditional distribution of transcripts given the data. The transcript from kernel K^j, which takes values in $\mathcal{Y}^{(j)}$, shall be denoted by $Y^{(j)}$. 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 machines. 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[2]. In an abuse of notation, we shall often refer to the entire triplet $\{T, \{K^j\}_{j=1,\dots,m}, (\mathcal{U}, \mathscr{U}, \mathbb{P}^U)\}$ using just T.

The statistical model underlying the distributed protocol plays an (ambient) role in the definition of a distributed protocol. In most of the thesis, the statistical model under consideration is clear from the context, and we shall simply say "distributed testing protocol" without stating for which underlying model. An exception to this is Chapter 6, where multiple models are under consideration in the same context.

Given a distributed protocol and i.i.d. data from P_f we shall use $\mathbb{P}_f$ to denote the joint distribution of $Y = (Y^{(1)}, \dots, Y^{(m)})$, the data X under P_f^m and the shared randomness $U \sim \mathbb{P}^U$. Writing $x = (x^{(1)}, \dots, x^{(m)}) \in \mathcal{X}^m$, 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

$$\begin{array}{ccccc} X^{(1)} & \longrightarrow & Y^{(1)}|U & \searrow & \\ \vdots & \longrightarrow & \vdots & \longrightarrow & T(Y). \\ X^{(m)} & \longrightarrow & Y^{(m)}|U & \nearrow & \end{array} \tag{1.4}$$

In Chapter 6, the above definition is generalized further to general decision problems. This generalization is straightforward. Because it is interesting to draw parallels with certain estimation problems throughout the thesis, we informally describe distributed estimation problems. A *distributed estimation protocol* consists of a similar triplet as a distributed testing protocol, differing only in the decision function; which we shall call *estimator* and denote by $\hat{f} : \bigotimes_{j=1}^m \mathcal{Y}^{(j)} \to \mathcal{F}$. We shall consider $\mathcal{F}$ to be equipped with some metric ℓ and the corresponding Borel sigma-algebra. The estimator $\hat{f}$ is required to be measurable. The *estimation risk* is then defined as

$$\sup_{f \in \mathcal{F}} \mathbb{E}_f \ell(\hat{f}, f).$$

[2]Meaning that $U \sim \mathbb{P}^U$ is a degenerate random variable / $\mathscr{U}$ is the trivial sigma-algebra.

The next section introduces a specific kind of distributed protocol, namely those that are bandwidth constrained.

1.2.1 Bandwidth constrained distributed protocols

> "*2019: Now 325 Mbps. The regression line has $R^2 = .99$, meaning that Nielsen's Law explains 99% of the variability in the data. Beyond uncanny. One small change is that when I first wrote about this in 1998, the best-fit growth rate for the 1984–1998 data was 53% (which I rounded to 50%), whereas the best-fit growth rate for the larger data set of 1984–2019 is 49% per year (which still rounds to 50%).*" - Jakob Nielsen

The kind of bandwidth constraints considered in this thesis is a limitation on the number of bits that a transcript can consist of. That is, when a distributed protocol satisfies a bandwidth constraint, each of the machines can only communicate a limited number of bits to the central server.

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

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.

Such a constraint is of concern in settings where data is observed, stored and/or processed locally and then required to be "compressed" when communicated to a central server. Note that the bandwidth constraint as defined here does not involve any notion of time or back-and-forth communication between the machines. In settings where it makes sense to consider bandwidth per unit of time (such as when describing up- and download speed), b should be interpreted as the total number of bits allowed to be communicated over a fixed amount of time.

Historically, computational power has increased[3] at a faster rate than bandwidth[4]. To speed up computation, it could make sense to distribute computation across servers, but for large data one might consequently run into bandwidth limitations as a bottleneck. In such settings, it is natural to consider the bandwidth and the data to be rather large, where the latter could be large in terms of dimensionality.

In other settings, bandwidth might be naturally scarce or costly. One could think of cellphone networks, where the total bandwidth is to be divided across all users, or sensors which gather data at a much higher resolution than they have the capacity to transmit, such as low energy sensors. Very low bandwidth settings capture for example

[3]Through e.g. the doubling of transistors every two years by "Moore's law" [154], which is stipulated to result in a roughly 60% increase in computational power year over year.

[4]The roughly 50% year of year increase in bandwidth is sometimes referred to as "Nielsen's law".

voting problems, where decisions are made on the basis of "yes" or "no" outcomes only, which can be seen as a 1-bit bandwidth constraint. A classical example of such a setting is meta-analysis on the basis of test outcomes. In fact, in Chapter 4, we shall employ bandwidth constraint bounds to obtain rates for meta-analysis in settings where only a study outcome in the form of a single real valued test statistic such as a p-value is available.

1.2.2 Privacy constrained distributed protocols

"*Anonymized data isn't.*" - Cynthia Dwork

The notion privacy considered in the thesis is that of differential privacy as put forward by [85]. Differential privacy provides a mathematical framework that guarantees preservation of privacy, in a notion akin to cryptographical guarantees [84].

What differentiates privacy from cryptographic protocols is that in the latter the goal is to protect data by giving selective access. The goal of a differential privacy constraint is that individuals within a dataset have their identity protected whilst (a part or version of) the dataset is publicly accessible.

The latter goal is the same as anonymization, which aims to to protect the identity of individuals by obfuscating certain information. Differential privacy, however, provides much stronger framework than anonymization. Anonymization, in principle, allows to reconstruct the identity of individuals given enough data or side information, see e.g. [156, 83, 87]. Differential privacy on the other hand, guarantees that the identity of individuals cannot be reconstructed with certainty, regardless of the amount of side information available.

Formally, a differential privacy constraint on a transcript in our setting is formulated as follows.

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

$$K^j(A|x_1, \ldots, x_i, \ldots, x_n, u) \leqslant e^{\epsilon} K^j(A|x_1, \ldots, x_i', \ldots, x_n, u) + \delta \qquad (1.5)$$
$$\text{for all } A \in \mathscr{Y}^{(j)},\ x_i', x_1, \ldots, x_i, \ldots, x_n \in \mathcal{X},\ i \in \{1, \ldots, n\}.$$

Small values of ϵ and δ ensure that, even when the transcript $Y^{(j)}$ is publicly available, the individuals within the sample (i.e. $x_1, \ldots, x_n$) underlying $Y^{(j)}$ are unidentifiable. The notion of differential privacy offers a very strong privacy guarantee: even when the entire sample is known, up to one individual, that one individual remains unidentifiable.

We shall elucidate the latter statement formally: Even if an entire sample $(x_1, \ldots, x_n)$ is known except for the individual corresponding to the index $i = 1$, deciding between $H_0 : x_1 = v$ or $H_1 : x_1 = w$ cannot be done with a testing risk less than $1-(e^{\epsilon}-1)-\delta$.

To see this, note that test $T \equiv T(Y^{(j)})$ of level α (i.e. $K^j(T=1|v,x_2,\ldots,x_n,u)=\alpha$), has Type II error

$$K^j(T=0|w,x_2,\ldots,x_n,u) = 1-K^j(T=1|w,x_2,\ldots,x_n,u) \tag{1.6}$$
$$\geqslant 1-e^\epsilon K^j(T=1|v,x_2,\ldots,x_n,u)-\delta = 1-e^\epsilon\alpha-\delta.$$

That is, any test with a nontrivial level, has close to trivial power when $\epsilon > 0$ and $\delta > 0$ are small. This extends to the alternative hypothesis $x_1 \neq w$, as the power calculation holds uniformly over the sample space. Such a guarantee is sometimes called *plausible deniability:* each individual can plausibly deny their presence or absence in a dataset, thereby protecting their identity even if the rest of the data is known.

The sizes of ϵ and δ depend heavily on the application. Typically, $\epsilon \lesssim 1$ is considered, as a value of $\epsilon = 3$ already allows arbitrary power for tests of level 0.05. For this thesis, we shall take $\epsilon \leqslant 1$ for simplicity, but minimax rate results do not change when considering any range for which $\epsilon = O(1)$.

Protocols satisfying the above definition for $\delta = 0$ are often called "pure" differentially private protocols, whereas for $\delta > 0$ the protocols are sometimes called "approximately" or "impure" differentially private protocols. In this text, we shall sometimes write ϵ-differentially private protocols instead of $(\epsilon, 0)$-differentially private protocols, and use DP as a shorthand for "differentially private". The δ parameter allows for catastrophic privacy breaches: with probability at most δ transcripts which that reveal the identity of individuals in the sample could be released. When δ is small, this may still be acceptable.

Typically, δ decaying polynomially in the number of observations is deemed acceptable, e.g. $\delta \ll (mn)^{-p}$ with $p > 1$. Larger values of δ can permit somewhat pathological situations. For example, $\delta \gtrsim 1/n$ permits privacy protocols that with positive probability violate the privacy of a random individual. In such a "bad-luck-lottery", one data point of the machine can be released with probability $1/n$. This is $(0, 1/n)$-DP, yet exposes a person's information with probability $1-(1-1/n)^n$. Similarly, $\delta \gtrsim 1/m$ allows for at least one machine to give up privacy of its sample with nonvanishing probability.

We note that for the definition of differential privacy, the sigma-algebra underlying the space of transcripts is important. The larger the sigma-algebra, the stronger the privacy constraint. For our purposes, when deriving differentially private testing protocols, it suffices to consider $\mathbb{R}^d$ equipped with the Borel sigma-algebra. The lower bounds hold for general sigma-algebras as well, as long as the quantities considered in the proof are appropriately measurable.

A distributed differentially private (testing) protocol is one in which the transcripts generated satisfy (3), or more specifically, as in the following definition.

Definition 4. 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 (1.5)

for all $u \in \mathcal{U}$.

Lastly, we note that the use of shared randomness does not affect the privacy guarantee provided by the protocol, as under the current definition, the guarantee of (1.6) is not affected 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.

1.3 Main results for the many-normal-means model under bandwidth and privacy constraints

> "*A model is some way of reducing the actuality of the world, to something where you can readily give a narrative for what actually occurs. Where you can make an abstraction of what is happening and answer questions that you care about.*" - Stephen Wolfram

In this section, we describe the minimax rate for this distributed signal detection problem under bandwidth- and differential privacy constraints. We revisit the many-normal-means model considered in Section 1.1, now formulated in the distributed setting. In the distributed version of the above normal-means model, the mn observations are divided over m machines. Equivalently, each local machine $j \in \{1, \ldots, m\}$ observes

$$X_i^{(j)} = f + Z_i^{(j)}, \tag{1.7}$$

with $Z_i^{(j)}$ i.i.d. $N(0, I_d)$ for $i = 1, \ldots, n$, with $f \in \mathbb{R}^d$.

The hypotheses remain unchanged; we wish to test the null hypothesis that $f = 0$ versus the alternative hypothesis that

$$f \in H_\rho := \left\{ f \in \mathbb{R}^d \,:\, \|f\|_2 \geqslant \rho \right\}. \tag{1.8}$$

The test is to be conducted on the transcripts $Y = (Y^{(1)}, \ldots, Y^{(m)})$, where each of machine $j = 1, \ldots, m$ has generated its transcript $Y^{(j)}$ on the basis of the underlying data $X^{(j)} = (X_1^{(j)}, \ldots, X_n^{(j)})$ and possibly a shared source of randomness U. Following the framework outlined in the previous section, the test is to be conducted by using a distributed testing protocol, $\{T, \{K^j\}_{j=1}^m, (\mathcal{U}, \mathscr{U}, \mathbb{P}^U)\}$, where the Markov kernels map from the underlying sample space $\mathbb{R}^{n \times d}$.

In this section, we present the main results for this model where the distributed protocol is either satisfying a b-bit bandwidth constraint or a (ϵ, δ)-differential privacy constraint. These are spread across two results, Theorems 1.1 and 1.2. The first, which describes the detection boundary under bandwidth constraints, is given in Section 1.3.1. The second describes the detection boundary under (ϵ, δ)-differential privacy constraints, presented in Section 1.3.2. Deriving these results is the focus of Chapters 2 and 3.

The phenomena occurring within the many-normal-means model extend to other models and testing problems, such as the nonparametric signal-white-noise model discussed in Chapter 5, the multinomial model, regression or density testing (see Chapter 6) or to meta-analysis (Chapter 4). We give an interpretation of these phenomena in this section, to provide the reader a flavor of the intricacies encountered in distributed, communication constrained settings.

1.3.1 Detection boundary under bandwidth constraints

The following theorem captures the detection boundary corresponding to the goodness-of-fit test of (1.8) in the many-normal-means problem under bandwidth constraints. Recall that $\mathscr{T}_{\mathrm{LR}}^{(b)}$ and $\mathscr{T}_{\mathrm{SR}}^{(b)}$, respectively denote the classes of all local randomness and shared randomness distributed testing protocols with a communication budget of b-bits per machine.

Theorem 1.1. *Consider any sequences $n \equiv n_\nu$, $d \equiv d_\nu$, $m \equiv m_\nu$ and $b \equiv b_\nu$ in $\mathbb{N}$. For any sequence of nonnegative numbers $\rho \equiv \rho_\nu$ such that*

$$\rho^2 \asymp \frac{\sqrt{d}}{mn}\left(\sqrt{\frac{d}{b \wedge d}} \bigwedge \sqrt{m}\right), \tag{1.9}$$

it holds that

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

Similarly, in the case of only local randomness, if

$$\rho^2 \asymp \frac{\sqrt{d}}{mn}\left(\frac{d}{b \wedge d} \bigwedge \sqrt{m}\right), \tag{1.10}$$

we have that

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

When $m \equiv 1$, we obtain the non-distributed (or unconstrained) minimax testing rate of $\rho^2 = \sqrt{d}/(nm)$. This makes sense, as even one bit of communication allows for the single machine to conduct and communicate an optimal test. Furthermore, when $b \gtrsim d$, enough information about the coefficients can be communicated to obtain the non-distributed minimax rate also, for both shared- and local randomness distributed protocols. When the communication budget is smaller than the dimension ($b = o(d)$), the class of shared randomness protocols starts to exhibit strictly better performance than the local randomness ones in scenarios as long as $d = o(mb)$. That is, as long as the total communication budget mb of the system exceeds the dimension d of the parameter, shared randomness protocols achieve a strictly better rate than the local randomness ones. This remarkable phenomenon is further explored in Chapter 3,

particularly in Section 3.3. This feature disappears when the dimension is larger than the total communication budget (i.e. $mb = o(d)$), at which point there exists a one-bit local randomness protocol achieving the optimal rate of $\rho^2 \asymp \sqrt{d}/(\sqrt{m}n)$ in both cases.

Consistent distributed testing turns out to be possible even for small values of b and m or n, as long as $\sqrt{m}n$ is large enough compared to $\sqrt{d}$. This stands in contrast to estimation in the d-dimensional Gaussian mean model, where in the squared L_2-loss as considered in Section 2.1 is subject to a lower bound rate of $\rho^2 \gtrsim d^2/(bmn)$, as exhibited in Theorem 2.1. Comparing the latter estimation rate with the testing rate shows that there are multiple scenarios in which distributed estimation is not possible consistently whereas consistent distributed testing is. What is not necessarily unique to the distributed setting, is that estimation is more difficult in terms of dimensional dependence. However, this phenomenon is further exacerbated in the distributed setup under bandwidth constraints. For example, when $b \asymp 1$, consistent estimation under bandwidth constraints requires $m \gg d$, whereas distributed testing only requires $mn^2 \gg d$.

Another stark difference with estimation is that, as long as $mb = o(d)$ in the shared randomness case or $mb^2 = o(d^2)$ in the local randomness case, an increase in communication budget does not lead to a better rate in testing. However, in estimation, an increase in small budgets can lead to an exponential improvement in convergence rate when the budget is very small, as found in [46].

Finally, we remark that the phenomenon of shared randomness offering improvement in terms of error rate is also not observed in the estimation problem considered in Section 2.1. In Chapter 3, Section 3.3.1 we go into why this is.

1.3.2 Detection boundary under differential privacy constraints

The following theorem describes the detection boundary in the many-normal-means problem under (ϵ, δ)-differential privacy constraints. The goodness-of-fit test we shall consider here is null hypothesis $f = 0$ versus the alternative hypothesis

$$f \in H_\rho := \left\{ f \in \mathbb{R}^d \,:\, M \geqslant \|f\|_2 \geqslant \rho \right\},$$

where $M > 0$ is a constant that can be taken arbitrarily large. Such a restriction is commonplace in the differential privacy landscape, see e.g. [128]. We recall that $\mathscr{T}_{\mathrm{LR}}^{(\epsilon,\delta)}$ and $\mathscr{T}_{\mathrm{SR}}^{(\epsilon,\delta)}$ denote the classes of all local- and shared randomness (ϵ, δ)-differentially private distributed testing protocols, respectively.

Theorem 1.2. *Consider any sequences of natural numbers $n \equiv n_\nu$, $m \equiv m_\nu$ such that $mn \to \infty$, $d \equiv d_\nu$, $\epsilon \equiv \epsilon_\nu \in ((mn)^{-1}, 1]$ and $\delta \equiv \delta_\nu \lesssim (mnd)^{-p}$ for some constant $p \geqslant 2$. Let $\rho \equiv \rho_\nu$ be a sequence of positive numbers such that*

$$\rho^2 \asymp \left(\frac{d}{mn\sqrt{n\epsilon^2 \wedge 1}\sqrt{n\epsilon^2 \wedge d}} \bigwedge \left(\frac{\sqrt{d}}{\sqrt{m}n\sqrt{n\epsilon^2 \wedge 1}} \bigvee \frac{1}{mn^2\epsilon^2} \right) \right). \tag{1.11}$$

Then,

$$\inf_{T\in\mathscr{T}_{SR}^{(\epsilon,\delta)}} \mathcal{R}(H_{M_\nu\rho},T)\to\begin{cases}0 & \text{for any } M_\nu\gg\sqrt{\log(1/\delta)}\log^3(mnd),\\ 1 & \text{for any } M_\nu\to 0.\end{cases}$$

Similarly, for

$$\rho^2\asymp\left(\frac{d\sqrt{d}}{mn(n\epsilon^2\wedge d)}\bigwedge\left(\frac{\sqrt{d}}{\sqrt{m}n\sqrt{n\epsilon^2\wedge 1}}\bigvee\frac{1}{mn^2\epsilon^2}\right)\right), \tag{1.12}$$

we have that

$$\inf_{T\in\mathscr{T}_{LR}^{(\epsilon,\delta)}} \mathcal{R}(H_{M_\nu\rho},T)\to\begin{cases}0 & \text{for any } M_\nu\gg\sqrt{\log(1/\delta)}\log^3(mnd),\\ 1 & \text{for any } M_\nu\to 0.\end{cases}$$

The derived rate indicates that the distributed testing problem under privacy constraints undergoes multiple phase transitions, resulting in different regimes where ϵ affects the detection boundary differently. Specifically, a smaller ϵ, which implies a stronger privacy guarantee, leads to an increased detection threshold. When δ decreases as a polynomial of d, m and n, its impact on the detection boundary is limited to a logarithmic factor, making its effect on the error rate minor compared to that of ϵ.

For $m=1$, the theorem describes the optimal separation rate for the testing problem in the central DP setting; where all data is available on a single machine. In this case, our theorem recovers the result of [157]. When $\epsilon\lesssim 1/\sqrt{n}$, the privacy constraint affects the rate polynomially. In contrast, for $\epsilon\gtrsim 1/\sqrt{n}$, the rate approximates the classical minimax rate, up to logarithmic factors. Thus, the privacy constraint significantly impacts the rate only when ϵ is relatively small compared to the number of observations n whenever $m=1$. That testing is more difficult under central DP constraints might be surprising, since in the central setting, a transcript can consist of just a binary outcome. However, when ϵ is relatively small enough ($\epsilon\lesssim 1/\sqrt{n}$), the privacy constraint still forms a bottleneck. This is in contrast to the bandwidth constraint setting, where the detection boundary is not affected by the communication budget when $m=1$.

When $n=1$, we establish the optimal separation rate for the testing problem in the local DP setting. Here, ϵ can be seen to always have a pronounced effect on the rate. This makes sense, as in this case, the privacy constraint is applied at the observation level, which is comparatively costly. The optimal rate unconstraint rate of $\sqrt{d}/mn$ cannot be reached for values of ϵ that align with conventional differential privacy considerations (i.e. $\epsilon\lesssim 1$) in this case.

In the general federated setting, with $m\gg 1$, we see that m and n come into play with different powers in the minimax rate whenever $\epsilon^2\lesssim d/n$. This means that if one

distributes $N = mn$ observations across m machines, the task becomes more challenging as the N observations are spread over a greater number of machines, rather than having many observations on a smaller number of machines. This phenomenon of benefitting from large local samples is not unique to testing, as we shall see in the estimation rates derived in Section 2.5.6. Here, it is found that the L_2-risk minimax estimation rate under (ϵ, δ)-differential privacy is $d^2/(mn^2\epsilon^2)$ whenever privacy constraints are binding ($\epsilon \lesssim \sqrt{d/n}$). The main difference with the estimation problem, is that the estimation rate does not exhibit the phase transitions that are observed in the testing problem.

The distributed testing problem under privacy constraints can be seen to be subject to multiple phase transitions, resulting in different "regimes" where ϵ affects the detection boundary differently. We shall interpret these regimes and phase transitions below. In most regimes, ϵ has a polynomial impact on the detection boundary. The impact of δ on the detection boundary is no more than a poly-logarithmic factor in n, m and d. This is true for the entire range of δ's that decrease faster than $(nmd)^{-p}$, with $p \geqslant 2$, where the power of the poly-logarithmic factor is unaffected by the choice of p. Whilst this means that the effect of δ on the error rate is minor compared to that of ϵ, fully capturing the impact of δ on the error rate is an interesting future area of research, but beyond the scope of this thesis.

When the privacy constraints are binding ($\epsilon \lesssim \sqrt{d/n}$), the first phase transition in the testing problem occurs whenever $\epsilon \asymp \sqrt{d/(mn)}$ in the case of shared randomness, $\epsilon \asymp d/\sqrt{mn}$ in case of local randomness protocols with $\epsilon \leqslant 1/\sqrt{n}$ or $\epsilon \asymp d/(\sqrt{m}n)$ in case of local randomness protocols with $\epsilon \geqslant 1/\sqrt{n}$, respectively. These particular phase transition corresponds to shifting from a "high-privacy-budget", in the sense that ϵ is relatively large compared to d, $1/m$ and $1/n$, to a "low-privacy-budget". In the high-privacy-budget regime, the relatively lenient privacy constraint enables a distinct testing strategy from the one in the low-privacy-budget regime. Whenever $\epsilon \gtrsim \sqrt{d/n}$, the optimal unconstrained rate of $\sqrt{d}/mn$ is achieved by these differentially private methods. There are certain values of d, m, n, where some of these regimes do not occur, for any value of $\epsilon \leqslant 1$. When $\epsilon \geqslant 1/\sqrt{n}$, the phase transitions between the high-privacy-budget regime and low-privacy-budget regimes still occur, but at different values of ϵ in case of local randomness protocols.

In the high-privacy-budget regime, there is an improvement in the minimax rate when shared randomness is allowed. This highlights a phenomenon that is remarkably similar to the bandwidth constraint setting; the delineation into high- and low-budget regimes, where only in the high budget regime, there is benefit to having access to shared randomness. The root cause of this advantage bears resemblance across both types of constraint settings. In the high-privacy-budget regime, the optimal strategy mirrors that of the optimal high-bandwidth approach. Specifically, this entails transmitting transcripts that essentially allow (partial) reconstruction of the original data at the central machine (see Chapter 3 for details). For both types of constraints, there is benefit to the increased coordination between the machines that the shared

randomness allows.

Another similarity between the privacy and bandwidth constraint settings is that the shared randomness advantage disappears in the low-budget scenario. The optimal testing strategies under both types of constraints bear a resemblance as well, where in both cases, locally optimal tests (or their corresponding test statistics) are essentially averaged, yielding a testing strategy that can essentially be viewed as a type of "majority vote" globally, on the basis of the locally optimal tests. Besides the testing strategies reflecting this phenomenon, the attentive reader will also find that the strategy in the proof the lower bound does so too.

Within the low-privacy-budget regime ($\epsilon \lesssim \sqrt{d/(mn)}$ in case of shared randomness, $\epsilon \lesssim d/\sqrt{mn}$ in case of only local randomness), the best possible rate that can be attained is $\sqrt{d}/(\sqrt{m}n)$. What is striking here, is that this rate is achieved for $\epsilon \gtrsim 1/\sqrt{n}$, whilst it does not depend on ϵ. Thus, in this regime, one should opt for the smaller $\epsilon \asymp 1/\sqrt{n}$, which obtains a stronger privacy guarantee "for free". This phenomenon can be explained as follows. Whereas in the high-privacy-budget regime, the strategy for attaining the corresponding minimax rate is to create synthetic data which retains certain aspects of the underlying true data, the strategy in the low-privacy-budget regime is to conduct testing procedures on the local data and combine only the m outcomes of the local test statistics. When $\epsilon \gtrsim 1/\sqrt{n}$, the "artificial noise" needed to guarantee the privacy of the local test statistics is negligible in terms of its effect on minimax rate. Whenever $\epsilon \lesssim 1/\sqrt{n}$, second phase transition occurs where this "artificial noise" that is added to the locally optimal private test is no longer negligible. A third phase transition occurs around $\epsilon \asymp 1/\sqrt{mnd}$, for both shared- and local randomness protocols. Here, a striking phenomenon occurs whenever $\epsilon \lesssim 1/\sqrt{mnd}$: dimension ceases to be of influence in the minimax rate. Essentially, this reveals that there is no difference between the one dimensional problem and the multivariate problem whenever the required privacy guarantee is stringent in relation to m,n and d.

The latter phenomenon can be explained as follows. Given the condition $\epsilon \lesssim 1/\sqrt{mnd}$, signals of size larger than $(mn^2\epsilon^2)^{-1}$ are in particular larger than d/n, which is the local estimation rate. When signals can be estimated consistently locally, dimensionality seizes to be a bottleneck. Loosely speaking, given a (very) accurate estimate of the mean vector f in each machine, the problem almost reduces to a univariate testing problem in the sense that $\langle f, X^{(j)} \rangle$ can be accurately estimated locally. However, even if $\langle f, X^{(j)} \rangle$ can be computed locally, it cannot be communicated without adding substantial noise due to the stringent privacy constraint stemming from $\epsilon \lesssim 1/\sqrt{mnd}$. Roughly speaking, retaining privacy for the univariate test statistic $\langle f, X^{(j)} \rangle$ is *easier* than retaining privacy for a local estimator of the signal (e.g. $n^{-1}\sum X_i^{(j)}$) as a whole, which is d-dimensional. The minimax rate reflects this, as it can be seen to be much smaller than the estimation rate whenever d is large in this regime. This regime also exemplifies privacy folklore: retaining privacy is easier in testing than in estimation, as the inference outcome is inherently low-dimensional. Interestingly,

our findings show that this is not always the case (i.e. this is not observed in the high-privacy-budget regime).

1.4 Beyond the many-normal-means model

Whilst the many-normal-means model is a canonical benchmark model for continuous distributions on $\mathbb{R}^d$, the multinomial model describes independent draws from distributions on discrete spaces. Consider a sample space $\mathcal{X}$ with cardinality d, such that a probability distribution q can be identified with an element of the $d-1$-dimensional simplex; $q : \mathcal{X} \to [0,1]$ and $\sum_{x\in\mathcal{X}} q(x) = 1$.

Recently, there have been numerous applications in areas that handle large samples of multinomial data over extensive domains (i.e. large d and n). For example, in population genetics [166, 196] and computer science; where it is used for e.g. information retrieval [228, 177], speech and text and classification [126], text mining [49] and large language models [168].

In the distributed analogue of this model, each machine $j = 1, \dots, m$ observes $i = 1, \dots, n$ i.i.d. observations $X_i^{(j)}$ taking values in a discrete space $\mathcal{X}$ of cardinality d. Given some candidate distribution q_0 on $\mathcal{X}$, a goodness-of-fit test would then be

$$H_0 : q(x) = q_0(x) \text{ for all } x \in \mathcal{X} \quad \text{versus} \quad H_1 : \|q_0 - q\|_{\mathrm{TV}} \geqslant \rho.$$

The natural comparison to the signal detection problem as described in Section 1.3 with the null hypothesis $f = 0$, is to test for uniformity, $q_0(x) = 1/d$ for all $x \in \mathcal{X}$. In Chapter 6, we shall demonstrate that, depending on the values of d and n, the communication constraint phenomena observed in the many-normal-means model described in the previous section extend to goodness-of-fit testing for these discrete distributions too. In particular, we derive the minimax testing rates for the above hypothesis under bandwidth constraints and differential privacy constraints whenever n is large enough compared to d and m. At the time of writing, minimax rates only having been obtained for the case of having just one draw from a discrete distribution per machine in [9, 10, 15], so the results here contribute to the literature by deriving the rates for the large sample regime (i.e. large n compared to d and m).

The many-normal-means model allows for extensions to nonparametric settings too. In Chapter 5, we shall consider the infinite dimensional *signal-in-white-noise model*, which serves as a canonical benchmark model for nonparametric goodness-of-fit testing and has been extensively studied outside of the distributed setting, see [94, 121, 140, 184, 118]. In the distributed setting, the $j = 1, \dots, m$ machines observe i.i.d. $X^{(j)}$ taking values in $\mathcal{X} \subset L_2[0,1]$ and subject to the stochastic differential equation

$$dX_{t;i}^{(j)} = f(t)dt + dW_{t;i}^{(j)} \tag{1.13}$$

under P_f, with $\{W_{\cdot;i}^{(j)} : i \in [n], j \in [m]\}$ i.i.d. Brownian motions and $f \in L_2[0,1]$. Besides the difference in the local observations, the distributed setup considered for

this model remains exactly the same. The results derived for the alternatives H_ρ in the finite dimensional model translate to testing in the infinite dimensional model against the alternative hypotheses

$$f \in H_\rho^{s,R} := \{f \in \mathcal{H}^{s,R}[0,1] : \|f\|_{L_2} \geqslant \rho \text{ and } \|f\|_{\mathcal{H}^s} \leqslant R\}. \tag{1.14}$$

Here, $\mathcal{H}^{s,R} = \mathcal{H}^{s,R}([0,1])$ denotes the Sobolev ball of radius R in the space of s-smooth Sobolev functions and $\|\cdot\|_{\mathcal{H}^s}$ the Sobolev norm, see Section 5.5.3 in Chapter 5 for the definition.

The problem bares a close relationship with "classical" nonparametric goodness-of-fit testing in the sense of [23, 182, 67, 211], in which we aim to distinguish the null hypothesis that an i.i.d. sample is generated from a cumulative distribution function $F = F_0$ versus the alternative hypothesis that $F \neq F_0$. To briefly (and roughly) illustrate this relationship, consider a CDF $t \mapsto F_0(t)$ and $F(t) := F_0(t) + n^{-1/2}\int_0^t f(s)ds$ for bounded $|f|$ such that $\int_0^1 f(s)ds = 0$ and an i.i.d. sample $\zeta_1, \ldots, \zeta_n \sim F$ taking values in $[0,1]$. A natural statistic for this problem is the function mapping $t \in [0,1]$ to

$$\sqrt{n}\left(n^{-1}\sum_{i=1}^n \mathbb{1}\{\zeta_i \leqslant t\} - F_0(t)\right) = \sqrt{n}\left(n^{-1}\sum_{i=1}^n \mathbb{1}\{\zeta_i \leqslant t\} - F(t)\right) + \int_0^t f(s)ds.$$

The first term on the right-hand side converges weakly to an F-Brownian bridge (see e.g. Section 19.1 in [205]). This motivates the Gaussian model described by (5.1) and test of the hypotheses $H_0 : f = 0$ and alternative (5.2) as a "benchmark problem" for nonparametric goodness-of-fit testing, with a class of alternatives of the form $F(t) := F_0(t) + n^{-1/2}\int_0^t f(s)ds$ with $f \in H_\rho^{s,R}$.

The smoothness parameter $s > 0$ determines the difficulty of the classical (non-distributed, $m = 1$) nonparametric testing problem as considered in e.g. [123]. Typically, the regularity of the function is not known in practice and one has to use data driven methods to find the best testing strategies. In Chapter 5, we derive upper and lower bounds for distributed tests adapting to unknown regularity under both bandwidth constraints and differential privacy constraints for the signal-in-white-noise model. The bounds are tight up to a log-log factor in the case of bandwidth constraints and up to poly-logarithmic factors in case of differential privacy.

The results for the nonparametric signal-in-white-noise models are extended to various other nonparametric models in Chapter 6. One such extension is distributed nonparametric regression, where each machine $j = 1, \ldots, m$ observes $i = 1, \ldots, n$ i.i.d. samples

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

where $Z_i^{(j)}$ are i.i.d. standard Gaussian and $\zeta_i^{(j)}$ are either fixed or random design points. When $\zeta_i^{(j)} = i/n$, the model can be seen as a discretized version of (1.13).

Lastly, we shall extend the results to nonparametric density testing. Here, each machine $j = 1, \dots, m$ observes $i = 1, \dots, n$ i.i.d. observations from a probability density f on $[0, 1]$, say

$$X^{(j)} = (X_1^{(j)}, \dots, X_n^{(j)}) \overset{\text{i.i.d.}}{\sim} f.$$

The goodness-of-fit test we shall consider is over the class $\mathcal{F}$ of all probability densities $f \in \mathcal{H}^{s,R}[0, 1]$ such that $f \gtrsim n^{-1/(2s+1)}$, where we consider a hypothesis test of the form

$$H_0 : f = f_0 \text{ versus } H_1 : f \in \mathcal{F}, \|f - f_0\|_1 \geqslant \rho,$$

for some fixed density $f_0 \in \mathcal{F}$. For simplicity, we shall restrict here also test to tests of uniformity ($f_0 \equiv 1$), but with more technical work the results can be extended to testing for general Sobolev smooth densities f_0 bounded away from 0. Minimax distributed testing rates for nonparametric density testing for the above hypothesis have been studied in [75, 136], but only for the case of privacy constraints with a single observation per machine. The results in Chapter 6 compliment these results by providing minimax rates for bandwidth constraints and privacy constraints for the distributed setting when the number of observations per machine is large.

The extensions from the many-normal-means model to these more complicated models is one of the reasons warranting the thorough study of the many-normal-means model, to which we now return in Chapters 2, 3 and 4.

Chapter 2

Impossibility theorems for distributed testing

> "*Once you eliminate the impossible, whatever remains, no matter how improbable, must be the truth.*" – Arthur Conan Doyle

The main results in this chapter come in the form of lower bounds for the minimax detection thresholds under bandwidth- and 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

$$X_i^{(j)} = f + Z_i^{(j)}, \tag{2.1}$$

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

$$f \in H_\rho := \left\{ f \in \mathbb{R}^d \,:\, \|f\|_2 \geqslant \rho \right\}. \tag{2.2}$$

The first of these main results is to be found Section 2.3 in the form of Theorem 2.3, which establishes the lower bounds for the detection threshold for both the shared- and local randomness distributed testing protocols under bandwidth constraints. Theorem 2.4 in Section 2.4 establishes the lower bounds for both the shared- and local randomness distributed testing protocols under (ϵ, δ)-differential privacy constraints. The lower bounds established in each of these theorems are *tight* (up to log-factors in the case of Theorem 2.4), in the sense that the lower bound rates can be attained by distributed testing protocols within their respective classes. This is established in Chapter 3, by providing methods which attain the respective rates posed by the lower bounds of Theorem 2.3 and Theorem 2.4. Together with the results from Chapter 3, we obtain the minimax rates as posed by Theorems 1.1 and 1.2.

We note that the aforementioned results are not asymptotic in nature as they hold for every combination of b, n, m and d under bandwidth constraints and every n, m, d and $0 < \epsilon \leqslant 1$ under privacy constraints, hence going beyond the classical parametric framework. We do not present explicit constants for each of the theorems, but these could in principle be determined through the chosen methods of proof.

However, before deriving the aforementioned lower bounds, we will first take a brief detour to explore the use of mutual information in obtaining lower bounds for the distributed testing problem, based on the approach of [193]. The mutual information technique has been successful in deriving lower bounds for distributed estimation problems, see e.g. [78, 77, 39, 188, 46, 47, 226]. It is natural to consider this successful approach for the distributed testing problem as well.

It turns out that the mutual information technique is only partially successful in deriving the optimal minimax rates for the distributed testing problem. To understand the limitations of the mutual information technique for the distributed testing problem, and to fully understand the necessity of the novel approach based on a Brascamp-Lieb type inequality of Section 2.2, we shall first explore its use in the context lower bounds in the many-normal-means model.

In Section 2.1.1, we start by deriving a mutual information based lower bound for an estimation problem closely aligned with the testing problem under consideration. We then turn to the testing lower bound using mutual information in Section 2.1.2.

Not only does this approach illustrate the difference between the estimation and testing problems, but it also serves as a warm-up exercise in terms of understanding the general approach in deriving lower bounds for both problems. The Brascamp-Lieb type inequality based proof of Section 2.2 is more lengthy and technical, which means that some of the intuition might be lost.

2.1 Lower bounds through mutual information

In this section, we shall explore the use of mutual information in obtaining a testing lower bound. Mutual information is a concept in information theory that measures the amount of information shared between two random variables. It quantifies the dependency between the variables and provides a way to understand how much knowing one variable can tell you about the other.

For random variables X, Y we define the *mutual information between X and Y* as the Kullback-Leibler distance between the joint distribution and the product of the marginal distribution:

$$I(X;Y) = D_{\mathrm{KL}}(\mathbb{P}^{(X,Y)} \| \mathbb{P}^X \times \mathbb{P}^Y).$$

When X and Y are independent, the mutual information between them is 0, a positive value for the mutual information indicates dependence between X and Y, where the dependence is stronger for larger values.

The concept of mutual information is used in distributed systems to derive lower bounds on various problems, such as data compression and source coding (see e.g. [65]) or interactive communication [38], but also general minimax theory (see e.g. [221, 225]).

Lower bounds based on mutual information enjoy success as an approach to distributed estimation for two main reasons. Firstly, its tensorization properties allows exploitation of the Markov chain structure of (1.4). Secondly, data processing arguments allow quantitative capture of the loss of bandwidth constraints. These properties of mutual information are proven in Section 2.5.1 of the chapter appendix.

Before we turn to the testing lower bound using mutual information in Section 2.1.2, we make a small detour to the corresponding estimation problem, deriving a mutual information based lower bound for an estimation problem closely aligned with the testing problem under consideration. This estimation lower bound obtained is not novel, but serves to exemplify an approach commonly taken for distributed estimation lower bounds.

2.1.1 A mutual information based lower bound for estimation under bandwidth constraints

Below, we exemplify the mutual information approach to obtaining a distributed estimation lower bound, culminating into Theorem 2.1. The proof of the theorem is based on [77] and [39]. The proof is structured in the general framework of the distributed testing lower bounds in this thesis: the estimation risk is lower bounded by a type of Bayes risk, which is then further lower bounded, in this case by a variation of Fano's inequality. The final step uses "data processing arguments", e.g. arguments that capture the loss of information due to the communication restriction.

Theorem 2.1 focusses on bandwidth constraints only, although for local differential privacy ($n = 1$) such a bound can easily be derived using the approach below (with a different data processing argument). However, we shall defer the reader to Section 2.5.6, in which we present a novel method to derive a tight (ϵ, δ)-differential privacy lower bound in the same estimation setting for a full range of $n \in \mathbb{N}$ values.

In what follows for the formulation and proof of Theorem 2.1 below, we consider no shared randomness. Specifically, we take U to be degenerate and ignore it completely in notation. The motivation for this is given in Section 3.3.1, in which we show that for convex loss functions, distributed protocols do not benefit from shared randomness (Theorem 3.3). The proof is essentially that of [77] combined with the data processing arguments of [193].

Theorem 2.1. *Let $Y = (Y^{(1)}, \ldots, Y^{(m)})$ be generated according to a b-bit constrained distributed estimation protocol (see Section 1.2 and Section 1.2.1). There exists con-*

stant $c, c' > 0$ such that whenever $\rho^2 \leqslant c\frac{d^2}{(b \wedge d)nm}$, it holds that

$$\sup_{f \in \mathbb{R}^d} \mathbb{E}_f \rho^{-2} \left\| \hat{f}(Y) - f \right\|_2^2 \geqslant c' \quad \textit{for all } d, b, n, m \in \mathbb{N}. \tag{2.3}$$

Remark 1. This is a (log-factor) tighter version of the lower bound of [77]. The lower bound is tight for $mb/d \gtrsim \log n$, where notably $mb/d \to 0$ implies that the estimate on a local machine starts outperforming the central estimate. To uncover the optimal lower bound rate for the regime when $mb/d \lesssim \log(n)$ requires a much more extensive argument, see [46]. Here, it is shown in that whenever $b \lesssim d/m$, $\rho^2 \asymp d$ is the minimax squared L_2-norm estimation rate: for very small bandwidth budgets, the estimation error blows up linearly in d.

We consider a prior distribution on the parameter f in (2.1); given by $F = d^{-1/2}\rho R$ with R a d-dimensional vector of independent Rademacher random variables (R_i takes values -1 or 1 with probability $1/2$ each). Such choices are typically considered as "least favorable priors" supported on signals that are difficult to detect, see for instance Section 3.2 of [118]. Let $\mathbb{P}_f = \mathbb{P}^{(X,Y)|F=f}$, such that under $\mathbb{P}^{(F,X,Y)}$ we have the Markov chain structure

$$\begin{array}{ccccccc} & \nearrow & X^{(1)} & \longrightarrow & Y^{(1)} & \searrow & \\ F & \longrightarrow & \vdots & \longrightarrow & \vdots & \longrightarrow & \hat{f}(Y), \\ & \searrow & X^{(m)} & \longrightarrow & Y^{(m)} & \nearrow & \end{array} \tag{2.4}$$

where $\hat{f} : \bigotimes_{j=1}^m \mathcal{Y}^{(j)} \to \mathbb{R}^d$ is some estimator. The following lemma provides lower bound on the squared L_2-norm estimation risk in terms of the mutual information between F and Y. The proof, which we provide for completeness, is based on [77] and employs a standard multiple testing argument combined with a version of Fano's inequality (Lemma 2.22 in the appendix).

Lemma 2.1. *Consider F as above. The following lower bound holds for the estimation risk;*

$$\sup_{f \in \mathbb{R}^d} \mathbb{E}_f \|\hat{f}(Y) - f\|_2^2 \geqslant \frac{\rho^2}{2}\left(1 - 2 \cdot \frac{I(Y;F) + \log 2}{d}\right). \tag{2.5}$$

Remark 2. The unconstrained estimation lower bound follows by a data processing inequality $I(Y;F) \leqslant I(X;F)$ (Lemma 2.19 in the appendix) and by showing that $I(X;F) \leqslant n\rho^2 \log 2$, which follows by the arguments below. Plugging this bound into the right-hand side of (2.5), we see that if $\rho^2 \ll d/n$, the estimation risk is strictly bounded away from 0, which yields $\rho^2 \gtrsim d/n$ as a lower bound estimation rate. This lower bound rate is tight, as it can be seen to be attained by the sample mean through a simple calculation.

Proof. Lower bounding the supremum over a set by an integral over the same set and using Markov's inequality, we obtain for $t \geqslant 0$ that

$$\sup_{f \in \mathbb{R}^d} \mathbb{E}_f \|\hat{f}(Y) - f\|_2^2 \geqslant \mathbb{E}^F \mathbb{E}^{Y|F} \|\hat{f}(Y) - F\|_2^2 \geqslant t^2 \mathbb{P}\left(\|\hat{f}(Y) - F\|_2 \geqslant t \right).$$

Define

$$\phi(Y) = \operatorname*{arg\,min}_{r \in \frac{\rho}{\sqrt{d}}\{-1,1\}^d} \|\hat{f}(Y) - r\|_2.$$

By definition of $\phi(Y)$, $\|\hat{f}(Y) - \phi(Y)\|_2 \leqslant \|\hat{f}(Y) - F\|_2$. Therefore, on the event that $\|\hat{f}(Y) - F\|_2 < t$, we also have that

$$\|\phi(Y) - F\|_2 \leqslant \|\hat{f}(Y) - F\|_2 + \|\hat{f}(Y) - \phi(Y)\|_2 < 2t.$$

Thus, $\|\phi(Y) - F\|_2 \geqslant 2t \implies \|\hat{f}(Y) - F\|_2 \geqslant t$, which in turn implies that

$$\mathbb{P}\left(\|\hat{f}(Y) - F\|_2 \geqslant t \right) \geqslant \mathbb{P}\left(\|\phi(Y) - F\|_2 \geqslant 2t \right).$$

Combining the above chain of inequalities with Fano's inequality (Lemma 2.22) – with $\mathcal{V} = d^{-1/2}\rho\{-1,1\}^d$ and $t = \rho/2\sqrt{2}$ – we obtain (2.5), following the fact that $|\{v \in \{-1,1\}^d : \|v - v'\|_2 \geqslant \sqrt{d/2}\}| \leqslant 2^{d/2}$ for all $v' \in \{-1,1\}^d$. □

Next, we employ data processing arguments to capture the loss of information in the Markov chain $F \to X \to Y$ that necessarily occurs. To do so, we start with the following "tensorization" upper bound,

$$I(F;Y) \leqslant \sum_{j=1}^m I\left(F; Y^{(j)}\right),$$

which follows from applying Lemma 2.21 (with $V = F$ and U in the lemma degenerate). Writing $F_k = d^{-1/2}\rho R_k$ for $k = 1, \ldots, d$, the chain rule for mutual information (see (2.72) in the appendix) gives

$$I(F;Y^{(j)}) = \sum_{k=1}^d I\left(F_k; Y^{(j)} | R_{1:k-1}\right) = \sum_{k=1}^d I(F_k; Y^{(j)}), \tag{2.6}$$

where the second equality follows from the fact that $F_1, \ldots, F_d$ are independent random variables. Loosely speaking, this identity combined with (2.6) effectively reduces the distributed estimation problem to the sum of the information loss of the Markov chain $F_k \to (X_{1k}^{(j)}, \ldots, X_{nk}^{(j)}) \to Y^{(j)}$, as $(X_{1k}^{(j)}, \ldots, X_{nk}^{(j)})$ is independent of F_l for $l \neq k$.

The mutual information is necessarily decreasing as we move further along a Markov chain. Lemma 2.19 in the appendix captures this effect and yields

$$I(F_k; Y^{(j)}) \leqslant \gamma I((X_{1k}^{(j)}, \dots, X_{nk}^{(j)}); Y^{(j)}) \tag{2.7}$$

with $\gamma = 1$. The above inequality (with $\gamma = 1$) is referred to as the *data processing inequality for the mutual information*. The Markov chain $F \to X^{(j)} \to Y^{(j)}$, is said to satisfy a *$\gamma$-strong data-processing inequality* if the above inequality holds for $0 < \gamma < 1$. Here, γ captures a "strict" loss of information. The following lemma states that, if the likelihood ratio (with respect to the mixture distribution) of the data is $\sqrt{\gamma/2}$-sub-Gaussian, the mutual information satisfies a strong data processing inequality. A proof is provided in the appendix (Lemma 2.23), based on the proof of [170] who shows the same result for discrete sample spaces.

Lemma 2.2. *Consider random vectors $V, W, \hat{V}$ forming a Markov chain $V \to W \to \hat{V}$. Suppose that $\mathbb{P}^{W|V=v} \ll \mathbb{P}^W$ and that the random variables*

$$\frac{d\mathbb{P}^{W|V=v}}{d\mathbb{P}^W}(W)$$

are $\sqrt{\gamma/2}$-sub-Gaussian for $0 < \gamma < 1$, $\mathbb{P}^V$-almost surely. Then, the Markov chain $V \to W \to \hat{V}$ satisfies the γ-strong data-processing inequality (2.7).

With some effort, it can be shown that the likelihoods of $W_{jk} = (X_{1k}^{(j)}, \dots, X_{nk}^{(j)})$;

$$\frac{d\mathbb{P}^{W_{jk}|F_k=v}}{d\mathbb{P}^{W_{jk}}}(W) \quad \text{for} \quad v \in \frac{\rho}{\sqrt{d}}\{-1, 1\} \tag{2.8}$$

are $\sqrt{Cn\rho^2/d}$-sub-Gaussian for a universal constant $C > 0$ (Lemma 2.25 in the appendix). Putting the above together, we have obtained that

$$I(F; Y) \leqslant \sum_{j=1}^{m} \sum_{k=1}^{d} 2C \frac{n\rho^2}{d} I((X_{1k}^{(j)}, \dots, X_{nk}^{(j)}); Y^{(j)}).$$

So far, it has not been used that the transcript $Y^{(j)}$ is bandwidth constraint. At this point, the lower bound without communication constraints of $\rho^2 \lesssim d/(mn)$ could be obtained by showing that $I((X_{1k}^{(j)}, \dots, X_{nk}^{(j)}); Y^{(j)}) = O(1)$. Under communication constraints, a better bound is available for the above display whenever $b \lesssim d$. Using once more that the vectors $(X_{1k}^{(j)}, \dots, X_{nk}^{(j)})$ are independent (since $F_1, \dots, F_d$ are independent), we obtain

$$\sum_{k=1}^{d} I((X_{1k}^{(j)}, \dots, X_{nk}^{(j)}); Y^{(j)}) = I(X^{(j)}; Y^{(j)}).$$

The transcript $Y^{(j)}$ takes values in a discrete space $\mathcal{Y}^{(j)}$ of cardinality at most 2^b under a b-bit bandwidth constraint. Consequently, by standard results for the mutual

information, $I(X^{(j)}; Y^{(j)}) \leqslant b \log(2)$ (see Lemma 2.20 in the chapter appendix). To conclude the proof of the estimation lower bound, we have obtained that

$$\sup_{f \in \mathbb{R}^d} \mathbb{E}_f \|\hat{f}(Y) - f\|_2^2 \geqslant \frac{\rho^2}{2} \left(1 - 2 \cdot \frac{\log(2) \left(\frac{2Cbmn\rho^2}{d} + 1 \right)}{d} \right),$$

which yields (2.3) whenever $\rho^2 \leqslant cd^2/(mnb)$ for a small enough constant $c > 0$. To obtain the "classical", unconstrained rate, Lemma 2.19 and the chain rule for mutual information (see (2.72) in the appendix) also yield that

$$I\left(F; Y^{(j)}\right) \leqslant I\left(F; X^{(j)}\right) = \sum_{k=1}^{d} \sum_{i=1}^{n} I\left(F_k; X_{ik}^{(j)}\right).$$

Whenever $d^{-1/2}\rho \leqslant 1/4$,

$$\begin{aligned} I\left(F_k; X_{ki}^{(j)}\right) &= \sum_{r \in \{-1,1,\}} D_{\mathrm{KL}}\left(N(rd^{-1/2}\rho, 1); \frac{1}{2}(N(d^{-1/2}\rho, 1) + N(-d^{-1/2}\rho, 1)) \right) \\ &= \mathbb{E} \log \cosh\left(d^{-1/2}\rho N(0,1) \right) \leqslant 2\frac{\rho^2}{d}. \end{aligned}$$

This yields the statement of the theorem, since we now also have that

$$\sup_{f \in \mathbb{R}^d} \mathbb{E}_f \|\hat{f}(Y) - f\|_2^2 \geqslant \frac{\rho^2}{2} \left(1 - 2 \cdot \frac{\log(2) \left(2mn\rho^2 + 1 \right)}{d} \right).$$

2.1.2 A mutual information based lower bound for testing under bandwidth constraints

The following theorem establishes a detection threshold for the bandwidth constraint distributed signal detection problem of (2.1) using the mutual information approach as exhibited for estimation in the previous section. The theorem is tight for bandwidth constrained shared randomness protocols when $b = 1$, otherwise the technique cannot successfully capture the tight testing lower bounds, as we shall argue in the next section. Its proof is described in the remainder of this section.

Theorem 2.2. *For any $\alpha \in (0,1)$ there exists $c_\alpha > 0$ small enough such that whenever*

$$\rho^2 < c_\alpha \frac{\sqrt{d(\sqrt{m} \wedge d)}}{nmb}, \tag{2.9}$$

it holds that

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

Remark 3. It should be noted that we do not optimize for the value of the constant c_α in the proof below and the statement is likely to be still true for larger values of c_α.

The proof of the theorem relies on three key lemmas, which we state below after introducing some necessary notations. As a first step, we use the basic fact that the supremum of the probability of a type two error of a test can be lower bounded by a Bayesian type two error, i.e. for any prior distribution π supported on H_ρ

$$\sup_{f\in H_\rho} \mathbb{P}_f\left(T=0\right) \geqslant \int_{H_\rho} \mathbb{P}_f\left(T=0\right)\, d\pi(f).$$

To further lower bound the risk we construct an appropriate Markov chain and relate the testing problem to an information transfer problem through the chain. Consider $V \sim \mathrm{Ber}(1/2)$, i.e. V is 0 or 1, each with probability $1/2$, independent of the shared random vector U, such that the random vectors $X^{(j)}|(V=0)$, $j=1,\dots,m$ follow (2.1) with $f=0$ and $X^{(j)}|(V=1)$ follows a Gaussian mixture P_π defined as $P_\pi(A) = \int P_f(A)\, d\pi(f)$ for all Borel sets $A \subset \mathbb{R}^d$. Let us denote by $\mathbb{P}$ the joint probability measure describing the corresponding Markov dynamics

$$V \longrightarrow F \begin{array}{ccccc} \nearrow & (X^{(1)},U) & \longrightarrow & Y^{(1)} & \searrow \\ \longrightarrow & \vdots & \longrightarrow & \vdots & \longrightarrow \\ \searrow & (X^{(m)},U) & \longrightarrow & Y^{(m)} & \nearrow \end{array} T, \tag{2.10}$$

where $F \sim \pi$. We then have that for any distributed test T,

$$\mathcal{R}(H_\rho, T) \geqslant \mathbb{P}(T=1|V=0) + \mathbb{P}(T=0|V=1) = 2\mathbb{P}(T \neq V). \tag{2.11}$$

The right-hand side of (2.11) can be further bounded from below using the *mutual information* between T and V in the chain (2.10), defined by

$$I(V,T) = D_{\mathrm{KL}}\left(\mathbb{P}^{V\times T} \,\|\, \mathbb{P}^V \times \mathbb{P}^T\right),$$

where $\mathbb{P}^V$, $\mathbb{P}^T$ and $\mathbb{P}^{V\times T}$ denote marginal- and joint distributions of V and T. Informally, the mutual information measures how much knowing T reduces uncertainty about V and vice versa. The following lemma fulfills this role, similarly to Fano's inequality in the estimation problem.

Lemma 2.3. *Let π be a prior on H_ρ and consider the dynamics* (2.10)*. For any $T \in \mathcal{T}^b_{SR}$ we have*

$$\mathcal{R}(H_\rho, T) \geqslant 1 - \sqrt{2I(V,T)}.$$

Proof. In view of (2.11) we have

$$\begin{aligned} \mathcal{R}(H_\rho, T) &\geqslant 1 - \left(\mathbb{P}(T=0|V=0) - \mathbb{P}(T=0|V=1)\right) \\ &\geqslant 1 - \left|\mathbb{P}^{T|V=0} - \mathbb{P}^{T|V=1}\right|(T=0) \\ &\geqslant 1 - \|\mathbb{P}^{T|V=0} - \mathbb{P}^{T|V=1}\|_{\mathrm{TV}}. \end{aligned}$$

By the triangle inequality,

$$\|\mathbb{P}^{T|V=0} - \mathbb{P}^{T|V=1}\|_{\mathrm{TV}} \leqslant \|\mathbb{P}^{T|V=0} - \mathbb{P}^{T}\|_{\mathrm{TV}} + \|\mathbb{P}^{T} - \mathbb{P}^{T|V=1}\|_{\mathrm{TV}}.$$

Applying the second Pinsker bound to the two terms on the right-hand side and using that $2ab \leqslant a^2 + b^2$,

$$\|\mathbb{P}^{T|V=0} - \mathbb{P}^{T|V=1}\|_{\mathrm{TV}}^2 \leqslant D_{\mathrm{KL}}(\mathbb{P}^{T|V=0}\|\mathbb{P}^{T}) + D_{\mathrm{KL}}(\mathbb{P}^{T|V=1}\|\mathbb{P}^{T}) = 2I(V,T),$$

which completes the proof of the lemma. □

In view of the usual data processing inequality (Lemma 2.19 in the appendix) we have $I(V,T) \leqslant I(V,(Y^{(1)},\dots,Y^{(m)}))$. The following lemma asserts that, up to an additional term, this further tensorizes conditionally on the shared randomness.

Lemma 2.4. *Consider the dynamics* (2.10)*. We have*

$$I(V,(Y^{(1)},\dots,Y^{(m)})) \leqslant \sum_{j=1}^{m} I(V,Y^{(j)}|U) + \sum_{j=1}^{m} I(F,Y^{(j)}|U,V). \tag{2.12}$$

The proof of this lemma is given in Section 2.5.1, restated as Lemma 2.21. This bound, combined with Lemma 2.3, allows us to break down the difficulty of the 'global' testing problem in terms of the difficulty of the m 'local' testing problems, captured by the quantities $I(V,Y^{(j)}|U)$. These conditional local mutual informations quantify the capacity of the local tests to distinguish a signal drawn from the prior π from the zero signal. The second sum in the display of the lemma captures dependency between the transcripts and the prior draw $F \sim \pi$. The terms $I(F,Y^{(j)}|U,V)$ are similar quantities to the ones appearing in the estimation setting of the previous section. Essentially, the second sum captures how well the signal can be *estimated* by the local tests.

We now discuss the choice of prior distribution π. Let $\varrho := \rho/\sqrt{d}$ and let R be a d-dimensional vector of independent Rademacher random variables, and define the prior π as the distribution of $\rho/\sqrt{d}R$. Note that π has support contained in H_ρ. Since V, F and $X^{(j)}$ are independent of U, conditioning on U does not disrupt the Markov chain property: we have the chain $V|U \to F|U \to X^{(j)}|U \to Y^{(j)}|U$.

As a consequence of this choice of prior distribution, the "estimation term" $I(F,Y^{(j)}|U,V)$ can be handled using strong data processing techniques employed in distributed estimation as in the previous section. Writing $R_1,\dots,R_d$ for the coordinates of R and write for $k \leqslant d$, $R_{1:k} := (R_1,\dots,R_k)$ and $X^{(j)}_{i;1:k} = (X^{(j)}_{i1},\dots,X^{(j)}_{ik})$. Conditionally on $V=0$, $F=0$ with probability 1, so $I(F;Y^{(j)}|V=0)=0$. Conditionally on $V=1$, $F=\varrho R$. Combining these facts with the chain rule for mutual information (see (2.72)

in the appendix),

$$I(F;Y^{(j)}|V) = \frac{1}{2}I(F;Y^{(j)}|V=1) = \frac{1}{2}\sum_{k=1}^{d} I(F_k;Y^{(j)}|V=1,R_{1:k-1}) \\ = \frac{1}{2}\sum_{k=1}^{d} I(F_k;Y^{(j)}|V=1),$$

where the last equality follows from the fact that the coordinates of R are independent. Furthermore, $R_k|V=1 \to (X_{1k}^{(j)},\dots,X_{nk}^{(j)})|V=1 \to Y^{(j)}|V=1$ forms a Markov chain with $(X_{ik}^{(j)}|R_k,V=1) \sim N(\varrho R_k,1)$. Consequently, by applying Lemma 2.2 in conjunction with (2.8), we obtain that

$$I(F,Y^{(j)}|U,V) \leqslant 96\frac{n\rho^2}{d} I(X^{(j)},Y^{(j)}|U,V=1).$$

Using that $Y^{(j)}$ is supported on a set of cardinality at most 2^b, we obtain that the second term in (2.12) is bounded above by $96\frac{bmn\rho^2}{d}$.

The loss of information about V resulting from the compression of $X^{(j)}|U$ into $Y^{(j)}|U$ in this Markov chain is quantified by inequality (2.14) below, which is a strong data processing inequality for the information contained on V. Similarly to the approach in estimation, we prove the strong data processing inequality through proving sub-Gaussianity of the conditional likelihood ratio and then employing Lemma 2.2. The aforementioned sub-Gaussianity is described by the following lemma.

Lemma 2.5 (Public Coin Strong Data Processing Inequality)**.** *The likelihood ratios*

$$\frac{d\mathbb{P}^{X^{(j)}|V=0}}{d\mathbb{P}^{X^{(j)}}}(X^{(j)}) \quad \textit{and} \quad \frac{d\mathbb{P}^{X^{(j)}|V=1}}{d\mathbb{P}^{X^{(j)}}}(X^{(j)})$$

are $\sqrt{C\beta}$*-sub-Gaussian with*

$$\beta = \begin{cases} \frac{n^2\rho^4}{d} & \text{if } n\rho^2 > 2, \\ \frac{2n\rho^2}{d} & \text{if } n\rho^2 \leqslant 2, \end{cases} \tag{2.13}$$

and $C > 0$ *a universal constant.*

We obtain the following strong data processing inequality for the local testing problem, capturing its difficulty of the local testing problem in terms of n, d and ρ;

$$I(V,Y^{(j)}|U) \leqslant (48\beta \wedge 1) I(X^{(j)},Y^{(j)}|U). \tag{2.14}$$

By combining the information theoretic inequalities above with the fact that

$$I(X^{(j)},Y^{(j)}|U) \leqslant H(Y^{(j)}|U) \leqslant b,$$

we get that

$$I(V,T) \leqslant \sum_{j=1}^{m} I(V,Y^{(j)}|U) + \sum_{j=1}^{m} I(F,Y^{(j)}|U,V) \leqslant 48\beta mb + 96\frac{bmn\rho^2}{d}.$$

Therefore, in view of Lemma 2.3,

$$\mathcal{R}(H_\rho,T) \geqslant 1 - C\sqrt{\frac{bmn\rho^2}{d}(\max\{n\rho^2,2\}+1)},$$

for a universal constant $C > 0$. For ρ satisfying (2.9), the right-hand side is bounded from below by α for an arbitrary distributed test.

2.2 The Brascamp-Lieb inequality and testing lower bound

> "*I don't have any particular recipe... Doing research is challenging as well as attractive. It is like being lost in a jungle and trying to use all the knowledge that you can gather to come up with some new tricks, and with some luck you might find a way out.*" – Maryam Mirzakhani

In this section, we develop an approach that proves fruitful in deriving tight testing lower bounds for both bandwidth and differential privacy constraints.

Before describing the novel method, let us reflect on why the usual "estimation approach" through mutual information does not lead to a tight distributed testing lower bound. The mutual information approach fails to capture a tight testing communication constrained lower bound for the full range of bits for multiple reasons. For one, the bound depends on b even when upper bound methods suggest that there is no benefit to having additional communication budget (i.e. the regime where $mb \lesssim d$). In this regime, majority voting (see Section 3.1.1), which requires only 1-bit of information per machine, turns out to be optimal. In addition, it fails to capture the increase in testing error when no shared randomness is available, due to the limited options in choosing the prior due to the requirement of coordinate wise independence. As we shall see in this section, *least favorable priors in the distributed testing setting exploit the local randomness distributed protocol's limitations in terms of the extent to which each dimension of the data is sufficiently "covered" by the protocol's transcripts*[1]. Furthermore, in the case of differential privacy constraints, adequate data processing techniques are not available for the mutual information whenever $n \gg 1$. Another approach to obtain a distributed testing lower bound is through directly Taylor expanding the (local) likelihoods and bounding the resulting polynomials directly, see [12]. This approach suffers from the same fate as mutual information; see Section 4 of the aforementioned paper for a description of the issues of this specific approach.

[1] We further explore this idea in generality in Section 3.3.1.

Similarly to the mutual information based approach, the approach taken in this section consists of three steps. The first step is standard; we put a prior on the alternative hypothesis and lower bound the testing risk by the Bayes risk. Through standard arguments, the Bayes risk is shown to be lower bounded by a quantity tending to one if the variance of the *Bayes factor* (i.e. the likelihood ratio of the prior mixture with respect to the null distribution) tends to zero. As a second step, we bound the variance of the Bayes factor using a type of *Brascamp-Lieb inequality*, which is a generalization of Young's inequality [37]. Roughly speaking, the Brascamp-Lieb inequality we derive "factorizes" the second moment of the Bayes factor into m-times the second moment of the "local Bayes factors" and a factor that can be interpreted as the Fisher information of the distributed protocol. The third step consists of data processing arguments, i.e. inequalities that capture the loss of information resulting from the privacy and bandwidth constraints. These data processing techniques differ for bandwidth- and privacy constraints. Hence, they are discussed separately in Section 2.3 and Section 2.4, respectively. In these sections, we also formally state the main theorems that are consequently obtained or bandwidth- and privacy constraints.

The rest of this section is dedicated to step one and two as outlined in the previous paragraph. As a first step, we introduce a prior distribution π on $\mathbb{R}^d$ and lower bound the testing risk by a type of Bayes risk and the mass of π that resides outside of the alternative hypothesis H_ρ, akin to e.g. [123]. Recall that $\mathbb{P}_f$ denotes the joint distribution of Y, U and X where $X^{(j)}$ follows (1.7) and $Y \sim \mathbb{E}_f^{X,U} K(\cdot|X,U) =: \mathbb{P}_{f,K}^Y = \mathbb{P}_f^Y$. For π a given distribution on $\mathbb{R}^d$, define the mixture distribution $\mathbb{P}_\pi^X = P_\pi$ on $\mathbb{R}^{md}$ by $P_\pi(A) = \int P_f(A) d\pi(f)$, where we recall the notational convention $\mathbb{P}_f^X = P_f$. For an arbitrary distributed testing protocol $T \equiv \{T, \{K^j\}_{j=1}^m, \mathbb{P}^U\}$, using that $T \leqslant 1$, we can lower bound the testing risk $\mathcal{R}(H_\rho, T)$ by the Bayes risk as follows:

$$\mathbb{P}_0(T(Y)=1) + \sup_{f \in H_\rho} \mathbb{P}_f(T(Y)=0) \geqslant \mathbb{P}_0(T(Y)=1) + \int \mathbb{P}_f(T(Y)=0) d\pi(f) - \pi(H_\rho^c). \tag{2.15}$$

Consequently, the minimax testing risk satisfies

$$\inf_{T \in \mathscr{T}} \mathcal{R}(H_\rho, T) \geqslant \inf_{T \in \mathscr{T}} \sup_{\pi} \left(\mathbb{P}_0(T(Y)=1) + \int \mathbb{P}_f(T(Y)=0) d\pi(f) - \pi(H_\rho^c) \right), \tag{2.16}$$

where the supremum is taken over all probability distributions on $\mathbb{R}^d$. We note that the above display means we can adversarially choose π contingent on $\{T, \{K^j\}_{j=1}^m, \mathbb{P}^U\}$, but not the outcome of the source shared randomness U. Let $L_\pi^{Y|U=u}(Y)$ denote the Bayes factor of the Bayesian testing problem corresponding to the Bayes risk above; that is,

$$L_\pi^{Y|U=u}(Y) = \frac{d\mathbb{P}_\pi^{Y|U=u}}{d\mathbb{P}_0^{Y|U=u}}(Y).$$

To lower bound the Bayes risk, in light of the Neyman-Pearson lemma, it should suffice to show that $L_\pi^{Y|U=u}(Y)$ is close to 1 with high probability. Lemma 2.28 in

Section 2.5.3 makes this precise, showing that the right-hand side of (2.16) is further bounded from below by

$$1 - \sup_K \inf_\pi \left(\sqrt{(1/2) \int \mathbb{E}_0^{Y|U=u} \left(L_\pi^{Y|U=u}(Y) - 1 \right)^2 d\mathbb{P}^U(u)} + \pi(H_\rho^c) \right). \tag{2.17}$$

To lower bound the testing risk further, it suffices to show that for some prior π on $\mathbb{R}^d$ with little mass outside of H_ρ, the variance of the Bayes factor conditionally on the shared randomness U, is small while integrating over $\mathbb{P}^U$.

This brings us to the crux of the proof, Lemma 2.6 below. We first introduce some notation. Denote the "local" and "global" likelihoods of the data as

$$\mathscr{L}_f^j(X^{(j)}) = \frac{dP_f^n}{dP_0}(X^{(j)}), \quad \mathscr{L}_f(X) := \prod_{j=1}^m \frac{dP_f^n}{dP_0}(X^{(j)}),$$

and the mixture likelihoods as

$$\mathscr{L}_\pi^j(X) = \int \mathscr{L}_f(X^{(j)}) d\pi(f) \text{ and } \mathscr{L}_\pi(X) = \int \mathscr{L}_f(X) d\pi(f)$$

In view of the Markov chain structure, the probability measure $d\mathbb{P}_\pi(x,u,y)$ disintegrates as $d\mathbb{P}^{Y|(X,U)=(x,u)} d\mathbb{P}_f^X(x) d\mathbb{P}^U(u) d\pi(f)$. Using this, $\mathbb{E}_0^{Y|U=u} \left(L_\pi^{Y|U=u}(Y) \right)^2$ can be seen to equal

$$\mathbb{E}_0^{Y|U=u} \mathbb{E}_0 \left[\mathscr{L}_\pi(X) \middle| Y, U = u \right]^2 = \int \left(\int \mathscr{L}_\pi(x) \frac{dK(\cdot|x,u)}{d\mathbb{P}_0^{Y|U=u}}(y) d\mathbb{P}_0^X(x) \right)^2 d\mathbb{P}_0^{Y|U=u}(y), \tag{2.18}$$

where it is used that $K(\cdot|x,u) \ll \mathbb{P}_0^{Y|U=u}(\cdot)$, $\mathbb{P}_f^{(X,U)}$-almost surely (Lemma 2.27). Using Fubini's theorem ("decoupling" in X), we can write the above display as

$$\int \mathscr{L}_\pi(x_1) \mathscr{L}_\pi(x_2) q_u(x_1, x_2) d(\mathbb{P}_0^X \times \mathbb{P}_0^X)(x_1, x_2), \tag{2.19}$$

where

$$q_u(x_1, x_2) := \int \frac{dK(\cdot|x_1,u)}{d\mathbb{P}_0^{Y|U=u}}(y) \frac{dK(\cdot|x_2,u)}{d\mathbb{P}_0^{Y|U=u}}(y) d\mathbb{P}_0^{Y|U=u}(y). \tag{2.20}$$

Since $K(\cdot|x,u)$ and $\mathbb{P}_0^{Y|U=u}$ are product measures on $\mathcal{Y} = \bigotimes_{j=1}^m \mathcal{Y}^{(j)}$, we can write $q_u(x_1,x_2) = \Pi_{j=1}^m q_u^j(x_1,x_2)$ where

$$q_u^j(x_1,x_2) = \int \frac{K^j(y^j|x_1^j,u) K^j(y^j|x_2^j,u)}{\mathbb{P}_0^{Y^{(j)}|U=u}(y^j)} d\mathbb{P}_0^{Y^{(j)}|U=u}(y). \tag{2.21}$$

The map $(x_1,x_2) \mapsto q_u(x_1,x_2)$ can be seen as capturing the dependence between the original data X and a random variable X' with conditional distribution

$$X'|X = x \sim \int d\mathbb{P}_0^{X|(Y,U)=(y,u)} d\mathbb{P}^{Y|(X,U)=(x,u)}, \tag{2.22}$$

which is sometimes referred to as the "forward-backward channel", stemming from the fact that $X \to Y \to X'$ forms a Markov chain. An easy computation using the law of total expectation shows that the covariance of $q_u(x_1, x_2)d(P_0 \times P_0)(x_1, x_2)$,

$$\int \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \begin{pmatrix} x_1^\top & x_2^\top \end{pmatrix} q_u(x_1, x_2)d(P_0 \times P_0)(x_1, x_2) \in \mathbb{R}^{2mnd \times 2mnd}, \tag{2.23}$$

is equal to $\Sigma_u := \mathrm{Diag}\left(\Sigma_u^{11}, \ldots, \Sigma_u^{1n}, \ldots, \Sigma_u^{m1}, \ldots, \Sigma_u^{mn}\right) \in \mathbb{R}^{2mnd \times 2mnd}$ for

$$\Sigma_u^{ji} := \begin{pmatrix} I_d & \Xi_u^{ji} \\ \Xi_u^{ji} & I_d \end{pmatrix},$$

with

$$\Xi_u^{ji} := \mathbb{E}_0^{Y^{(j)}|U=u} \mathbb{E}_0\left[X_i^{(j)} \middle| Y, U = u\right] \mathbb{E}_0\left[X_i^{(j)} \middle| Y^{(j)}, U = u\right]^\top.$$

Define also

$$\Xi_u^{j} := \mathbb{E}_0^{Y^{(j)}|U=u} \mathbb{E}_0\left[\sum_{i=1}^n X_i^{(j)} \middle| Y^{(j)}, U = u\right] \mathbb{E}_0\left[\sum_{i=1}^n X_i^{(j)} \middle| Y^{(j)}, U = u\right]^\top. \tag{2.24}$$

We are now ready to state the lemma that forms the crux of our distributed testing lower bound proofs, both in the case of bandwidth and differential privacy constraints.

Lemma 2.6. *Suppose that $(x_1, x_2) \mapsto q_u(x_1, x_2)$ is bounded and that π is a centered Gaussian distribution on $\mathbb{R}^d$. Then,*

$$\frac{\int \mathscr{L}_\pi(x_1)\mathscr{L}_\pi(x_2)q_u(x_1, x_2)d(\mathbb{P}_0^X \times \mathbb{P}_0^X)(x_1, x_2)}{\prod_{j=1}^m \int \mathscr{L}_\pi^j(x_1^j)\mathscr{L}_\pi^j(x_2^j)q_u^j(x_1^j, x_2^j)d(\mathbb{P}_0^{X^{(j)}} \times \mathbb{P}_0^{X^{(j)}})(x_1^j, x_2^j)} \tag{2.25}$$

is bounded above by

$$\frac{\int \mathscr{L}_\pi(x_1)\mathscr{L}_\pi(x_2)dN(0, \Sigma_u)(x_1, x_2)}{\prod_{j=1}^m \int \mathscr{L}_\pi^j(x_1^j)\mathscr{L}_\pi^j(x_2^j)dN(0, \Sigma_u^j)(x_1^j, x_2^j)}.$$

The lemma has the following interpretation: the ratio of the second moment of the Bayes factor of the "global Bayesian hypothesis test" that of the product of second moments of the "local Bayes factors", is maximized over the class of forward-backward channel with covariance Σ when the forward-backward channel is Gaussian.

There is an existing literature on Brascamp-Lieb inequality in relation to information theoretical problems, in relation to mutual information [56, 145, 146]. The proof of Lemma 2.25 relies on a different method of proof however, namely that of [143]. The fact that the prior π is Gaussian is vital to the proof technique, which exploits

the conjugacy between the prior and the model which enables the use of techniques from [143].

The proof of the lemma is presented at the end of the section. We first describe how it is of consequence to the testing lower bound, which is the content of Lemma 2.7 below.

Lemma 2.7. *Define*

$$\mathrm{A}_u^\pi := \int e^{f^\top \sum_{j=1}^m \Xi_u^j g} d(\pi \times \pi)(f,g) \tag{2.26}$$

and

$$\mathrm{B}_u^\pi := \prod_{j=1}^m \mathbb{E}_0^{Y^{(j)}|U=u} \mathbb{E}_0 \left[\mathscr{L}_\pi \left(X^{(j)} \right) \middle| Y^{(j)}, U = u \right]^2. \tag{2.27}$$

If $(x_1, x_2) \mapsto q_u(x_1, x_2)$ is bounded and if π is a centered Gaussian distribution on $\mathbb{R}^d$, it holds that

$$\mathbb{E}_0^{Y|U=u} \left(L_\pi^{Y|U=u}(Y) \right)^2 \leqslant \mathrm{A}_u^\pi \cdot \mathrm{B}_u^\pi.$$

The above lemma describes how the variance of the Bayes factor given U is bounded by two factors. One factor depends on the Fisher information of the transcripts' likelihood at $f = 0$ given $U = u$; $\Xi_u := \sum_{j=1}^m \Xi_u^j$. In this sense, A_u^π captures how well the transcript allows for "estimation" of f. The second factor can be seen as the m-fold product of the local Bayes factors, capturing essentially the power of combining the locally most powerful test statistics; the likelihood ratios.

Proof of Lemma 2.7. We start by noting that B_u^π is equal to the denominator of (2.25). By Lemma 2.6,

$$\mathbb{E}_0^{Y|U=u} \left(L_\pi^{Y|U=u}(Y) \right)^2 \leqslant \frac{\int \mathscr{L}_\pi(x_1) \mathscr{L}_\pi(x_2) dN(0,\Sigma)(x_1,x_2)}{\prod_{j=1}^m \int \mathscr{L}_\pi^j(x_1^j) \mathscr{L}_\pi^j(x_2^j) dN(0,\Sigma^j)(x_1^j, x_2^j)} \cdot \mathrm{B}_u^\pi.$$

By the block diagonal matrix structure of Σ, the denominator in the first factor of the right-hand side satisfies

$$\begin{aligned}
\prod_{j=1}^m \int \mathscr{L}_\pi^j(x_1^j) \mathscr{L}_\pi^j(x_2^j) dN(0,\Sigma^j)(x_1^j, x_2^j) &= \prod_{j=1}^m \int e^{\left(\|\sqrt{\Sigma^j}(f,g)\|_2^2 - \|(f,g)\|_2^2 \right)} d(\pi \times \pi)\,(f,g) \\
&= \prod_{j=1}^m \int e^{f^\top \Xi_u^j g} d(\pi \times \pi)(f,g) \\
&\geqslant \prod_{j=1}^m e^{\int f^\top \Xi_u^j g \, d(\pi \times \pi)(f,g)} = 1.
\end{aligned}$$

Through the expression for the moment generating function of the Gaussian, the numerator of A_u^π is equal to

$$\int \mathscr{L}_\pi(x_1) \mathscr{L}_\pi(x_2) dN(0,\Sigma)(x_1,x_2) = \int e^{f^\top \sum_{j=1}^m \Xi_u^j g} d(\pi \times \pi)(f,g).$$

□

What is left to show in this step, is that for $\pi = N(0, \Gamma)$, $\Gamma \in \mathbb{R}^{d\times d}$ can be chosen such that the $\mathrm{A}_u^\pi \cdot \mathrm{B}_u^\pi$ is small enough whilst also ensuring that $\pi(H_\rho^c)$ is controlled. We start with the latter. For a given $c_\alpha > 0$, set $\varrho := \rho c_\alpha^{-1/4} d^{-1/2}$ and $\Gamma := \varrho^2 \overline{\Gamma}$ for some $\overline{\Gamma} \in \mathbb{R}^{d\times d}$ to be specified later, separately for the shared- and non-shared randomness protocols. The remaining mass $\pi(H_\rho^c)$ can now be seen to equal

$$\pi(f : \|f\|_2^2 \leqslant \rho^2) = \Pr\left(Z^\top \overline{\Gamma} Z \leqslant \sqrt{c_\alpha} d\right), \tag{2.28}$$

where Z is a d-dimensional standard normal vector. If $\overline{\Gamma}$ is symmetric, idempotent and has rank (proportional to) d, the concentration inequality in Lemma 3.28 yields that the probability on the right-hand side of the above display can be made arbitrarily small for small enough choice of $c_\alpha > 0$.

Suppose that for some constant $c > 0$,

$$\varrho^2 \left\| \sqrt{\overline{\Gamma}}^\top \Xi_u \sqrt{\overline{\Gamma}} \right\| \leqslant c. \tag{2.29}$$

If $\overline{\Gamma} \in \mathbb{R}^{d\times d}$ is symmetric, idempotent with rank proportional to d and $\pi = N(0, \varrho^2\overline{\Gamma})$, standard results for the Gaussian chaos, e.g. Lemma 6.2.2 in [210] combined with (2.29) and the fact that $\|\sqrt{\overline{\Gamma}}\| \leqslant 1$, yield that

$$\mathrm{A}_u^\pi \leqslant \exp\left(C\varrho^4 \mathrm{Tr}\left((\sqrt{\overline{\Gamma}}^\top \Xi_u \sqrt{\overline{\Gamma}})^2\right)\right),$$

for a constant $C > 0$ depending only on c. As a final step of the testing risk lower bound technique, we use essentially a geometric argument to sharpen this bound in case the distributed protocol does not enjoy shared randomness. The $d \times d$ matrix $\Xi_u := \sum_{j=1}^m \Xi_u^j$ geometrically captures how well Y allows to "reconstruct" the compressed sample X. When U is degenerate, Ξ_u is "known" to the prior, and $\overline{\Gamma}$ can be chosen to exploit "direction" in which Ξ_u contains the least information. This brings us to the Lemma 2.8 below, which summarizes our testing risk lower bound up to the data processing step, with which we continue in Section 2.3 for bandwidth constraints and Section 2.4 for privacy constraints. We finish this section by proving the lemma below and then, last but not least, by proving the Brascamp-Lieb type inequality of Lemma 2.6.

Lemma 2.8. *Let $\alpha \in (0,1)$ and suppose that the map $(x_1, x_2) \mapsto q_u(x_1, x_2)$ defined in (2.20) is bounded for all distributed testing protocols in $\mathscr{T}$. Let $\pi = N(0, \varrho^2\overline{\Gamma})$, with $\varrho := \frac{\rho}{c_\alpha^{1/4} d^{1/2}}$ and $\overline{\Gamma} \in \mathbb{R}^{d\times d}$ a symmetric, idempotent with rank$(\overline{\Gamma}) \in [\lfloor d/2 \rfloor, d]$. Assume that ρ is such that $\varrho^2\|\Xi_u\| \leqslant c\, \mathbb{P}^U$-a.s. for some constant $c > 0$. It then holds that $\inf_{T\in\mathscr{T}} \mathcal{R}(H_\rho, T)$ and in particular the Bayes risk*

$$\inf_{T\in\mathscr{T}} \sup_\pi \left(\mathbb{P}_0(T(Y) = 1) + \int \mathbb{P}_f(T(Y) = 0) d\pi(f) - \pi(H_\rho^c)\right)$$

are lower bounded by

$$1 - \sup_{K} \sqrt{(1/2) \int (\mathrm{A}_u \mathrm{B}_u^{\pi} - 1)\, d\mathbb{P}^U(u)} - \sup_{\overline{\Gamma}} \pi(H_\rho^c), \tag{2.30}$$

where the supremum is taken over all kernels corresponding to distributed testing protocols in $\mathscr{T}$, the second supremum over all symmetric, idempotent $\overline{\Gamma} \in \mathbb{R}^{d\times d}$ with $\lfloor d/2 \rfloor \leqslant rank(\overline{\Gamma}) \leqslant d$, B_u^{π} is as in (2.27) *and*

$$\mathrm{A}_u = \exp\left(C \frac{\rho^4}{c_\alpha d^2} \|\Xi_u\| \, Tr(\Xi_u) \right), \tag{2.31}$$

for some fixed constant $C > 0$ depending only on $c > 0$. Furthermore, if U is degenerate and $\frac{2\rho^2}{\sqrt{c_\alpha} d^2} Tr(\Xi) \leqslant c$, (2.30) *holds with*

$$\mathrm{A}_u = \exp\left(C \frac{\rho^4}{c_\alpha d^3} Tr(\Xi_u)^2 \right), \tag{2.32}$$

for some fixed constant $C > 0$ depending only on $c > 0$.

Proof. In case of shared randomness (i.e. U not being degenerate), simply taking $\overline{\Gamma} = I_d$, noting that $\mathrm{Tr}(\Xi_u^2) = \|\Xi_u\| \, \mathrm{Tr}(\Xi_u)$ and combining the results earlier in the section (from (2.17) onwards) leads to (2.30) and (2.31).

Now assume U is degenerate and write $\Xi_u = \Xi$. The matrix Ξ is positive definite and symmetric, therefore it possesses a spectral decomposition $V^\top \mathrm{Diag}(\xi_1, \ldots, \xi_d) V$. Without loss of generality, assume that $\xi_1 \geqslant \xi_2 \geqslant \cdots \geqslant \xi_d$ with corresponding eigenvectors $V = \begin{pmatrix} v_1 & \cdots & v_d \end{pmatrix}$. Let $\check{V}$ denote the $d \times \lceil d/2 \rceil$ matrix $\begin{pmatrix} v_{\lfloor d/2 \rfloor + 1} & \cdots & v_d \end{pmatrix}$. The choice of prior may depend on Ξ, to see this, note the order of the supremum and infimum in (2.17) and the fact that Ξ solely depends on the choice of kernel. To that extent, set $\overline{\Gamma} = \check{V}\check{V}^\top$. It holds that

$$\mathrm{Tr}(\check{V}\check{V}^\top) = \sum_{i=1}^{d} \sum_{k=\lfloor d/2 \rfloor + 1}^{d} (v_k)_i^2 = \lceil d/2 \rceil.$$

The choice $\overline{\Gamma}$ is thus seen to satisfy the conditions of symmetry and positive definiteness and is idempotent with rank $\lceil d/2 \rceil$.

Since the eigenvalues are decreasingly ordered,

$$\xi_{\lfloor d/2 \rfloor} \leqslant \frac{2}{d} \sum_{i=1}^{\lfloor d/2 \rfloor} \xi_i \leqslant \frac{2}{d} \mathrm{Tr}(\Xi).$$

By orthogonality of the columns of V, $\check{V}^\top \Xi \check{V} = \mathrm{Diag}(\xi_{\lfloor d/2 \rfloor + 1}, \ldots, \xi_d)$. The condition of (2.29) reduces to

$$\varrho^2 \|\sqrt{\overline{\Gamma}}^\top \Xi_u \sqrt{\overline{\Gamma}}\| \leqslant \varrho^2 \xi_{\lfloor d/2 \rfloor} \leqslant 2 \frac{\rho^2}{\sqrt{c_\alpha} d^2} \mathrm{Tr}(\Xi).$$

Note that

$$\mathrm{Tr}\big((\sqrt{\overline{\Gamma}}^{\top}\Xi_u\sqrt{\overline{\Gamma}})^2\big) = \mathrm{Tr}\left((\check{V}^{\top}\Xi\check{V})^2\right) = \sum_{i=\lfloor d/2\rfloor+1}^{d}\xi_i^2 \leqslant d\xi_{\lfloor d/2\rfloor}^2 \leqslant \frac{4}{d}\mathrm{Tr}(\Xi)^2,$$

which implies in turn that

$$\varrho^4\mathrm{Tr}\left((\check{V}^{\top}\Xi\check{V})^2\right) \leqslant 4\frac{\rho^4}{c_\alpha d^3}\mathrm{Tr}(\Xi)^2.$$

□

2.2.1 Proof of the Brascamp-Lieb type inequality

We shall prove Lemma (2.9), which is a slightly more general version of Lemma 2.6. To see this, note that for $k = 1, 2$ in display (2.21), x_k^j are projections of x_k on the coordinates indexed by $\{(j-1)dn+1, \ldots, jdn\}$, respectively. Since K is a Markov kernel, the function $q_u \in L_1(\mathbb{R}^{2dnm}, P_0 \times P_0)$ is nonnegative and it is bounded by the assumption of Lemma 2.6. Furthermore,

$$\begin{aligned}\int q_u(x_1,x_2)\,dP_0(x_1) &= \int \frac{dK(\cdot|x_2,u)}{d\mathbb{P}_0^{Y|U=u}}(y)\int dK(y|x_1,u)\,dP_0(x_1)\\ &= \int \frac{dK(y|x_2,u)}{d\mathbb{P}_0^{Y|U=u}(y)}d\mathbb{P}_0^{Y|U=u}(y) = 1,\end{aligned}$$

$\int q_u(x_1,x_2)\,dP_0(x_2) = 1$ and

$$\int x_i q_u(x_1,x_2)d(P_0\times P_0)(x_1,x_2) = \int x_i dP_0(x_i) = 0 \in \mathbb{R}^{mnd} \tag{2.33}$$

for $i = 1, 2$.

Lemma 2.9. *For $x \in \mathbb{R}^{mk}$, let $x^j \in \mathbb{R}^k$, $j = 1, \ldots, m$, denote the projection of x on the coordinates $\{(j-1)k+1, \ldots, jk\}$. Let $\Lambda \in \mathbb{R}^{k\times k}$ a positive definite symmetric matrix and $\Lambda^{\otimes m} = Diag(\Lambda, \ldots, \Lambda) \in \mathbb{R}^{mk\times mk}$. For $h \in \mathbb{R}^k$, let p_h denote the density of a $N(h,\Lambda)$ distribution with respect to the Lebesgue measure on $\mathbb{R}^k$, let $p_h^m(x) := \Pi_{j=1}^m p_h(x^j)$ and let P_h^m denote the probability measure corresponding to the Lebesgue density p_h^m. Define for $M > 0$,*

$$\begin{aligned}\mathcal{Q}(M,\Sigma) := \Bigg\{ q \in L_1(\mathbb{R}^{mk}, P_0^m) :\ & q \geqslant 0,\ \ \frac{q}{\int q(x)dP_0^m(x)} \leqslant M\ \ P_0^m - a.e.,\\ & \int x\,q(x)dP_0^m(x) = 0,\ \text{and}\ \ \frac{\int xx^{\top}\,q(x)dP_0^m(x)}{\int q(x)dP_0^m(x)} = \Sigma\Bigg\}.\end{aligned}$$

Furthermore, let H a $N(0,\Upsilon)$-distributed random vector in $\mathbb{R}^k$ for some nonnegative definite matrix $\Upsilon\mathbb{R}^{k\times k}$. Then,

$$\sup_{q\in\mathcal{Q}}\frac{\int \mathbb{E}^H \Pi_{j=1}^m \frac{p_H}{p_0}\left(x^j\right)\,q(x)p_0^m(x)dx}{\int \Pi_{j=1}^m \mathbb{E}^H \frac{p_H}{p_0}\left(x^j\right)\,q(x)p_0^m(x)dx} = \frac{\int \mathbb{E}^H \Pi_{j=1}^m \frac{p_H}{p_0}\left(x^j\right)\,dN(0,\Sigma)(x)}{\int \Pi_{j=1}^m \mathbb{E}^H \frac{p_H}{p_0}\left(x^j\right)\,dN(0,\Sigma)(x)}.$$

Proof. We start by introducing some short-hand notations for convenience. Write, for $x \in \mathbb{R}^{vk}$, $v \in \{1, m\}$,

$$\phi_v(x) = \mathbb{E}^H \frac{p_H^v}{p_0^v}(x) = \mathbb{E}^H e^{H^\top (\sum_{j=1}^v \Lambda^{-1} x^j) - \frac{v}{2}\|\Lambda^{-1/2} H\|_2^2},$$

with $\phi_m(x) p_0^m(x) = \mathbb{E}^H p_H^m(x)$, $x = (x^1, \dots, x^m)$, and $\Pi_{j=1}^m \phi_1(x^j) = \Pi_{j=1}^m \mathbb{E}^H p_H(x^j)$.

Let $\lambda \equiv \lambda_{mk}$ denote the Lebesgue measure on $\mathbb{R}^{mk}$, define for $r \in L_1(\mathbb{R}^{mk}, \lambda)$ nonnegative,

$$F(r) := \frac{\int \phi_m(x)\, r(x) dx}{\int \prod_{j=1}^m \phi_1(x^j)\, r(x) dx} \in [0, \infty], \tag{2.34}$$

and set $G(q) := F(q p_0^m)$. Let $\mathcal{Q} \equiv \mathcal{Q}(1, \Sigma)$. Since $G(cq) = G(q)$ for any constant $c \in \mathbb{R}$, it suffices to show that

$$\overline{G} = \sup_{q \in \mathcal{Q}} G(q) = \frac{\int \phi_m(x)\, dN(0,\Sigma)(x)}{\int \prod_{j=1}^m \phi_1(x^j)\, dN(0,\Sigma)(x)}.$$

We will proceed through the following steps.

1. First, we show that the supremum $\overline{G}$ is finite and attained in $\mathcal{Q}$, i.e. by the Banach–Alaoglu theorem there exists $q \in \mathcal{Q}$ such that $G(q) = \overline{G}$.

2. We will then consider $\mathcal{Q}_2$, the class of all $Q \in L_1(\mathbb{R}^{2km}, \lambda)$ such that $x_1 \mapsto Q(x_1, x_2)$ is in $\mathcal{Q}$ for P_0^m-almost every $x_2 \in \{x_1 \mapsto Q(x_1, x_2) \not\equiv 0\}$ and $x_2 \mapsto Q(x_1, x_2)$ is in $\mathcal{Q}$ for P_0^m-almost every $x_1 \in \{x_2 \mapsto Q(x_1, x_2) \not\equiv 0\}$. It holds that

$$G_2(Q) := \frac{\int \phi_m(x_1)\phi_m(x_2)\, p_0^m(x_1) p_0^m(x_2) Q(x_1, x_2) d(x_1, x_2)}{\int \prod_{j=1}^m \phi_1(x_1^j)\phi_1(x_2^j)\, p_0^m(x_1) p_0^m(x_2) Q(x_1, x_2) d(x_1, x_2)}$$

   satisfies $\sup_{Q \in \mathcal{Q}_2} G_2(Q) = \overline{G}^2$.

3. Next, we show that $(x_1, x_2) \mapsto q(\frac{x_1 - x_2}{\sqrt{2}}) q(\frac{x_1 + x_2}{\sqrt{2}})$ is a maximizer of G_2 whenever $q \in \mathcal{Q}$ is a maximizer of G. This is a consequence of the "conjugacy" between the distribution P_H^m and the distribution of H.

4. Then it will be shown that for any maximizer Q of G_2, $x_1 \mapsto Q(x_1, x_2)$ maximizes G for P_0^m-almost every x_2.

5. Combining the above steps, we obtain that for any maximizer q, an appropriately rescaled convolution of q with itself is also a maximizer, i.e.

$$F(\sqrt{2}(q p_0^m) * (q p_0^m)(\sqrt{2}\,\cdot)) = \overline{G},$$

   where $*$ denotes convolution.

6. By repeated application of Step 5 and the central limit theorem, the result follows.

Step 1. For $q \in \mathcal{Q}$, define the normalizing constant as $C_q := (\int q dP_0^m)^{-1}$. As linear combinations and products of nonnegative convex functions are convex, the mapping

$$x \mapsto \prod_{j=1}^{m} \mathbb{E}^H e^{H^\top \Lambda^{-1} x^j - \frac{1}{2}\|\Lambda^{-1/2} H\|_2^2}$$

is convex. The latter fact and Jensen's inequality imply that

$$\frac{\int \mathbb{E}^H e^{H^\top (\sum_{j=1}^m \Lambda^{-1} x^j) - \frac{1}{2}\|\Lambda^{-1/2} H\|_2^2} q(x) dP_0^m(x)}{\int \prod_{j=1}^{m} \mathbb{E}^H e^{H^\top \Lambda^{-1} x^j - \frac{1}{2}\|\Lambda^{-1/2} H\|_2^2}\, q(x) dP_0^m(x)} \leqslant$$
$$\frac{C_q \int \mathbb{E}^H e^{H^\top (\Lambda^{-1} \sum_{j=1}^m x^j) - \frac{1}{2}\|\Lambda^{-1/2} H\|_2^2} q(x) dP_0^m(x)}{\prod_{j=1}^{m} \mathbb{E}^H e^{C_q \int H^\top \Lambda^{-1} x^j q(x) dP_0^m(x) - \frac{1}{2}\|\Lambda^{-1/2} H\|_2^2}}.$$

Since $X = (X_1, \ldots, X_m) \sim q dP_0^m$ has mean 0, the denominator on the left-hand side is equal to $(\mathbb{E}^H e^{-\frac{1}{2}\|\Lambda^{-1/2} H\|_2^2})^m > 0$. This means that the denominator in the above display is bounded away from 0 over q. Since $qC_q \leqslant M$ a.e., the numerator is bounded above by $M \int \mathbb{E}^H p_H^m(x) dx = M$. We can conclude that the supremum of (2.34) over qp_0^m, $q \in \mathcal{Q}$ is finite. It is easy to construct a $q^* \in \mathcal{Q}$ such that $G(q^*) > 0$, so we can conclude that $0 < \overline{G} < \infty$.

Let q_t be a maximizing sequence for G, rescale q_t such that $\int q_t P_0^m = 1$ and note that $q_t \in \mathcal{Q}$ and q_t is contained in the $L_\infty(\mathbb{R}^{mk})$ ball of radius M. By the Banach–Alaoglu theorem the $L_\infty(\mathbb{R}^{mk})$ ball of radius M is weak-$*$-compact (associating the dual of $L_1(\mathbb{R}^{mk}, \lambda)$ with $L_\infty(\mathbb{R}^{mk})$). Therefore, there exists a subsequence, again denoted by q_t, along which $q_t \overset{\text{wk}-*}{\to} q$ for some q in the $L_\infty(\mathbb{R}^{mk})$ ball of radius M. Since $x = (x^1, \ldots, x^m) \mapsto \phi_m(x)$ is in $L_1(\mathbb{R}^{mk}, P_0^m)$, the weak-$*$-convergence implies that

$$\int \phi_m(x)\; q_t(x) dP_0^m(x) \to \int \phi_m(x)\; q(x) dP_0^m(x).$$

Similarly,

$$\int \Pi_{j=1}^m \phi_1\left(x^j\right)\; q_t(x) dP_0^m(x) \to \int \Pi_{j=1}^m \phi_1\left(x^j\right)\; q(x) dP_0^m(x) \in (0, \infty),$$

where the boundedness away from 0 has been concluded earlier on in the proof. We have now obtained that

$$\overline{G} = \lim_{t\to\infty} \frac{\int \phi_m(x)\; q_t(x) dP_0^m(x)}{\int \Pi_{j=1}^m \phi_1\left(x^j\right)\; q_t(x) dP_0^m(x)} = \frac{\int \phi_m(x)\; q(x) dP_0^m(x)}{\int \Pi_{j=1}^m \phi_1\left(x^j\right)\; q(x) dP_0^m(x)}. \tag{2.35}$$

Since $q_t \in \mathcal{Q}$, we have

$$\int x\, q_t(x) dP_0^m(x) = 0 \;\text{ and }\; \int x x^\top q_t(x) dP_0^m(x) = \Sigma \;\text{ for all } t.$$

As $x \mapsto 1$, $x \mapsto x$ and $x \mapsto xx^\top$ are all P_0^m integrable, the weak-$*$-convergence yields that $\int q(x)dP_0^m(x) = 1$, $\int x\, q(x)dP_0^m(x) = 0$ and $\Sigma = \int xx^\top q(x)dP_0^m(x)$. Since we have that $\int \zeta(x)q_t(x)dP_0^m(x) \to \int \zeta(x)q_t(x)dP_0^m(x)$ for every continuous and bounded function $\zeta : \mathbb{R}^{mk} \to \mathbb{R}^{mk}$, the Portmanteau lemma yields that $\int_B q dP_0^m \geqslant 0$ for all open sets B so $q \geqslant 0$ almost everywhere. We conclude that $G(q) = \overline{G}$ and $q \in \mathcal{Q}$.

Step 2. Let $Q \in \mathcal{Q}_2$ be given. By definition, the marginals $x_1 \mapsto Q(x_1, x_2)$, $x_2 \mapsto Q(x_1, x_2)$ are in $\mathcal{Q}$ P_0^m-a.e. and $\mathbb{E}^H p_H(x)dx = \phi_m(x)p_0^m(x)dx$ is equivalent to the Lebesgue measure, hence

$$\begin{aligned} G_2(Q) &= \int \phi_m(x)p_0^m(x_1) \int \phi_m(x)\, p_0^m(x_2)Q(x_1, x_2)dx_2dx_1 \\ &\leqslant \overline{G} \int \phi_m(x)p_0^m(x_1) \int \Pi_{j=1}^m \phi_1(x_2^j)p_0^m(x_2)Q(x_1, x_2)dx_2dx_1 \\ &\leqslant \overline{G}^2 \int \Pi_{j=1}^m \phi_1(x_2^j)p_0^m(x_2) \int \Pi_{j=1}^m \phi_1(x_1^j)p_0^m(x_1)Q(x_1, x_2)dx_1dx_2. \end{aligned}$$

Let $q \in \mathcal{Q}$ be a maximizer of G. Then, the above steps hold with equality for $Q(x_1, x_2) := q(x_1)q(x_2)$. For almost every $x_1 \in \{q \neq 0\} \equiv \{x_2 \mapsto Q(x_1, x_2) \not\equiv 0\}$,

$$\frac{Q(x_1, x_2)}{\int Q(x_1, x_2)dP_0^m(x_2)} = \frac{q(x_2)}{\int q(x_2)dP_0^m(x_2)} \leqslant M.$$

By similar calculations, the rescaled marginal has the correct mean and covariance. By symmetry, we conclude that the marginals of $(x_1, x_2) \mapsto q(x_1)q(x_2)$ belong to $\mathcal{Q}$, and it is a maximizer of G_2 over $\mathcal{Q}_2$.

Step 3. Consider a maximizer $q \in \mathcal{Q}$ of G. By a change of variables $w_1 = (x_1 - x_2)/\sqrt{2}$ and $w_2 = (x_1 + x_2)/\sqrt{2}$,

$$\begin{aligned} &\int \phi_m(x_1)\phi_m(x_2)q\left(\frac{x_1 - x_2}{\sqrt{2}}\right) q\left(\frac{x_1 + x_2}{\sqrt{2}}\right) p_0^m(x_1)p_0^m(x_2)d(x_1, x_2) = \\ &\int \phi_m\Big(\frac{w_1 + w_2}{\sqrt{2}}\Big)\phi_m\Big(\frac{w_1 - w_2}{\sqrt{2}}\Big)q(w_1)q(w_2)p_0^m\left(\frac{w_1 - w_2}{\sqrt{2}}\right) p_0^m\left(\frac{w_1 + w_2}{\sqrt{2}}\right) d(w_1, w_2). \end{aligned}$$

Since p_0^m is a Gaussian density, $p_0^m\left(\frac{w_1 - w_2}{\sqrt{2}}\right) p_0^m\left(\frac{w_1 + w_2}{\sqrt{2}}\right) = p_0^m(w_1)p_0^m(w_2)$. This follows from direct computation, but it characterizes Gaussian functions in general, see e.g. Theorem 1 in [55]. Likewise, for H' an independent copy of the centered Gaussian random vector H, $\frac{H - H'}{\sqrt{2}}$ and $\frac{H + H'}{\sqrt{2}}$ are independent and furthermore equal

in distribution to H. Therefore,

$$\begin{aligned}
&\phi_m\Big(\frac{w_1+w_2}{\sqrt{2}}\Big)\phi_m\Big(\frac{w_1-w_2}{\sqrt{2}}\Big)\\
&\quad= \mathbb{E}^{(H,H')} e^{H^\top\Lambda^{-1}\sum_{j=1}^m \frac{w_1^j+w_2^j}{\sqrt{2}} + (H')^\top\Lambda^{-1}\sum_{j=1}^m \frac{w_1^j-w_2^j}{\sqrt{2}} - \frac{m}{2}\|\Lambda^{-1/2}H\|_2^2 - \frac{m}{2}\|\Lambda^{-1/2}H'\|_2^2}\\
&\quad= \mathbb{E}^{(H,H')} e^{\left(\frac{H+H'}{\sqrt{2}}\right)^\top\Lambda^{-1}\sum_{j=1}^m w_1^j - \frac{m}{2}\|\Lambda^{-1/2}\frac{H+H'}{\sqrt{2}}\|_2^2 + \left(\frac{H-H'}{\sqrt{2}}\right)^\top\Lambda^{-1}\sum_{j=1}^m w_2^j - \frac{m}{2}\|\Lambda^{-1/2}\frac{H-H'}{\sqrt{2}}\|_2^2}\\
&\quad= \phi_m(w_1)\phi_m(w_2).
\end{aligned}$$

Since $(x_1,x_2)\mapsto q(x_1)q(x_2)$ was established to be a maximizer of G_2 in the second step, the above establishes that $(x_1,x_2)\mapsto q(\frac{x_1-x_2}{\sqrt{2}})q(\frac{x_1+x_2}{\sqrt{2}})$ is a maximizer of G_2 also.

Step 4. Next, we will show that for a maximizer $Q\in\mathcal{Q}_2$ of G_2, $x\mapsto Q(x,w)$ is in $\mathcal{Q}$ and is a maximizer of G for almost every w. We prove this by contradiction. Take an arbitrary measurable set $A\subset\mathbb{R}^{mk}$ s.t. $\lambda(A)>0$. Note that Gaussian measures are equivalent to the Lebesgue measure, so both $\mathbb{E}^H P_H^m(A)$ and $\Pi_{j=1}^m\mathbb{E}^H P_H^1(A)$ are bounded away from zero. Suppose that for $Q\in\mathcal{Q}_2$ a maximizer of G_2 it holds that

$$\begin{aligned}
&\int_A \phi_m(w)\int \phi_m(x)\,Q(x,w)dP_0^m(x)dP_0^m(w)\\
&\qquad< \overline{G}\int_A \phi_m(w)\int \Pi_{j=1}^m\phi_1\left(x^j\right)\,Q(x,w)dP_0^m(x)dP_0^m(w).
\end{aligned}\tag{2.36}$$

Since the marginal $w\mapsto Q(x,w)$ is in $\mathcal{Q}$ for almost every $x\in\{w\mapsto Q(x,w)\not\equiv 0\}$,

$$\begin{aligned}
&\overline{G}^2\int \Pi_{j=1}^m\phi_1\left(w^j\right)\Pi_{j=1}^m\phi_1\left(x^j\right)\,Q(x,w)(dP_0^m\times P_0^m)(x,w)\\
&\qquad\geqslant \overline{G}\int \Pi_{j=1}^m\phi_1\left(x^j\right)\int \phi_m(w)\,Q(x,w)dP_0^m(w)dP_0^m(x).
\end{aligned}$$

Likewise, $x\mapsto Q(x,w)$ is in $\mathcal{Q}$ for almost every $u\in A^c\cap\{x\mapsto Q(x,w)\not\equiv 0\}$, so

$$\begin{aligned}
&\overline{G}\int \Pi_{j=1}^m\phi_1\left(x^j\right)\int_{A^c} \phi_m(w)\,Q(x,w)dP_0^m(w)dP_0^m(x)\\
&\qquad\geqslant \int_{A^c}\phi_m(w)\int \phi_m(x)\,Q(x,w)dP_0^m(x)dP_0^m(w).
\end{aligned}$$

Together with (2.36) and the second to last display, we obtain that

$$\begin{aligned}
&\overline{G}^2\int \Pi_{j=1}^m\phi_1\left(w^j\right)\Pi_{j=1}^m\phi_1\left(x^j\right)\,Q(x,w)(dP_0^m\times P_0^m)(x,w)\\
&\qquad> \int\int \phi_m(x)\,\phi_m(w)\,Q(x,w)dP_0^m(w)dP_0^m(x),
\end{aligned}$$

which contradicts Q maximizing G_2.

Step 5. Let $q \in \mathcal{Q}$ be a maximizer of G over $\mathcal{Q}$, where q is normalized such that $\int q dP_0^m = 1$. Define q_2 as

$$q_2(x) := \int q\left(\frac{x-w}{\sqrt{2}}\right) q\left(\frac{x+w}{\sqrt{2}}\right) dP_0^m(w).$$

The map $x \mapsto q\left(\frac{x-w}{\sqrt{2}}\right) q\left(\frac{x+w}{\sqrt{2}}\right) := Q(x,w)$ is in $\mathcal{Q}$ for almost all w s.t. $Q(x,w) \not\equiv 0$ and as a consequence of the previous step, it is a maximizer of G for such w. Hence, $q_2(x)$ is a maximizer of G:

$$\begin{aligned}\int \phi_m(x) q_2(x) dP_0^m(x) &= \int\int \phi_m(x) q\left(\frac{x-w}{\sqrt{2}}\right) q\left(\frac{x+w}{\sqrt{2}}\right) dP_0^m(x)\, dP_0^m(w) \\ &= \overline{G} \int \Pi_{j=1}^m \phi_1\left(x^j\right) q_2(x) dP_0^m(x).\end{aligned}$$

Let $h \in L_1(\mathbb{R}^{mk}, p_0^m)$. Using again that $p_0^m\left(\frac{w_1-w_2}{\sqrt{2}}\right) p_0^m\left(\frac{w_1+w_2}{\sqrt{2}}\right) = p_0^m(w_1) p_0^m(w_2)$ and applying a change of variable $w = \sqrt{2}w - x$, we get

$$\begin{aligned}\int h(x) q_2(x) p_0^m(x) dx &= \int\int h(x) q\left(\frac{x-w}{\sqrt{2}}\right) q\left(\frac{x+w}{\sqrt{2}}\right) p_0^m\left(\frac{x-w}{\sqrt{2}}\right) p_0^m\left(\frac{x+w}{\sqrt{2}}\right) dxdw \\ &= \int\int h(x) q\left(\sqrt{2}x - w\right) q(w)\; p_0^m\left(\sqrt{2}x - w\right)\; p_0^m(w)\; dx\sqrt{2}dw \\ &= \int h(x)\sqrt{2}(qp_0^m) * (qp_0^m)(\sqrt{2}x) dx,\end{aligned}$$

where $f * g$ denotes convolution. Therefore, qp_0^m being a probability density with mean 0 and covariance Σ implies that $q_2 p_0^m$ is too. So, $q_2 \in \mathcal{Q}$ and maximizes G.

Step 6. Consider now $q_4 \in \mathcal{Q}$ defined by $q_4(x) := \int q_2\left(\frac{x-w}{\sqrt{2}}\right) q_2\left(\frac{x+w}{\sqrt{2}}\right) dP_0^m(w)$. Since $q_2 \in \mathcal{Q}$ is a maximizer, the above steps imply that $G(q_4) = \overline{G}$ and by a similar computation as above,

$$q_4(x) p_0^m(x) = \sqrt{4} \overset{4}{\circledast} (qp_0^m)(\sqrt{4}x),$$

where $\overset{4}{\circledast} r$ denotes $r * r * r * r$. Repeating the above steps, we obtain a maximizer $q_{2^N} \in \mathcal{Q}$ of G for $N \in \mathbb{N}$ which satisfies

$$\begin{aligned}r_{2^N}(x) := q_{2^N}(x) p_0^m(x) &= \int q_{2^{N-1}}\left(\frac{x-w}{\sqrt{2}}\right) q_{2^{N-1}}\left(\frac{x+w}{\sqrt{2}}\right) p_0^m(x) p_0^m(w) dxdw \\ &= \sqrt{2} \int q_{2^{N-1}}\left(\sqrt{2}x - w\right) p_0^m\left(\sqrt{2}x - w\right) q_{2^{N-1}}(w)\, dP_0^m(w) \\ &= \sqrt{2}\left(q_{2^{N-1}} p_0^m\right) * \left(q_{2^{N-1}} p_0^m\right)(\sqrt{2}x).\end{aligned}$$

We conclude that

$$r_{2^N}(x) = 2^{N/2} \overset{2^N}{\circledast} (qp_0^m)(2^{N/2}x)$$

and

$$\frac{\int \phi_m(x) r_{2^N}(x)dx}{\int \Pi_{j=1}^m \phi_1(x^j) r_{2^N}(x)dx} = G(q_{2^N}) = \overline{G}$$

for all $N \in \mathbb{N}$. Let $r = qp_0^m$. The characteristic function of r_{2^N} equals, for $s \in \mathbb{R}^{mk}$,

$$\begin{aligned}\Psi r_{2^N}(s) &:= \int e^{-is^\top x} r_{2^N}(x)dx = \int e^{-i\frac{s^\top}{2^{N/2}}x} \overset{2^N}{\circledast} r(x)\,dx = \left(\int e^{-i\frac{s^\top}{2^{N/2}}x} r(x)dx\right)^{2^N} \\ &= \left(\int \left(1 - i\frac{s}{2^{N/2}}x - \frac{(s^\top x)^2}{2^{N+1}} + O\left(\frac{(s^\top x)^3}{2^{3N/2}}\right)\right) r(x)dx\right)^{2^N}.\end{aligned}$$

Since r has mean 0, covariance Σ and bounded third moment (by the boundedness of q and $p_0^m d\lambda$ possessing a third moment), $\Psi r_{2^N}(s) \to e^{-\frac{1}{2}s^\top \Sigma s}$. Consequently, $r_{2^N} d\lambda$ converges weakly to a Gaussian distribution with mean 0 and covariance Σ. In particular, $\int \phi r_{2^N} d\lambda \to \int \phi dN(0,\Sigma)$ for all $\phi \in C^\infty(\mathbb{R}^{mk})$, so

$$\overline{G} = \lim_{N\to\infty} \frac{\int \phi_m(x)\, r_{2^N}(x)dx}{\int \Pi_{j=1}^m \phi_1(x^j)\, r_{2^N}(x)dx} = \frac{\int \phi_m(x)\, dN(0,\Sigma)(x)}{\int \Pi_{j=1}^m \phi_1(x^j)\, dN(0,\Sigma)(x)},$$

which finishes the proof. □

2.3 A complete lower bound for testing under bandwidth constraints

The main results of this section come in the form of a single theorem describing the lower bounds for the detection threshold distributed testing protocols that satisfy a bandwidth constraint, both with and without shared randomness. The optimality of the theorem is established in Section 3.1, by providing both a shared randomness and local randomness distributed testing protocol, which attain the respective rates posed by the lower bounds. Together, these results yield Theorem 1.1.

Theorem 2.3. *For each $\alpha \in (0,1)$ there exists a constant $c_\alpha > 0$ (depending only on α) such that if*

$$\rho^2 < c_\alpha \frac{\sqrt{d}}{n}\left(\sqrt{\frac{d}{b \wedge d}} \bigwedge \sqrt{m}\right), \tag{2.37}$$

then in the shared randomness protocol case

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

Similarly, for

$$\rho^2 < c_\alpha \frac{\sqrt{d}}{n}\left(\frac{d}{b \wedge d} \bigwedge \sqrt{m}\right), \tag{2.38}$$

we have under the local randomness protocol that

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

Remark 4. The proof of the theorem reveal that the theorem holds for other classes of alternatives as well. In particular, the lower bounds hold for any H_ρ such that $N(0, c_\alpha^{-1/2} d^{-1} \rho^2 I_d)(H_\rho) \leqslant \alpha$ for $c_\alpha > 0$ small enough.

The above theorem implies that if (2.37) holds, no consistent shared randomness distributed testing protocol with communication budget b bits per machine exists for the hypotheses $H_0 : f = 0$ versus the alternative $H_1 : \|f\|_2 \geqslant \rho$. In other words, no shared randomness distributed test manages to consistently distinguish all signals from 0 if the signals are smaller than the right-hand side of (2.37). When considering distributed testing protocols with only local randomness, the detection threshold (2.38) is more stringent than the shared randomness threshold (2.37) for certain values of d, m and b. Theorem 3.1 in Section 3.1 affirms that, in these cases, the best local randomness protocols have a strictly worse performance compared to the best shared randomness protocols.

Next, we provide a proof of the theorem. As a starting point, we aim to apply Lemma 2.12. To that extent, we will verify its conditions and as a "data processing" step, we bound the quantities A_u and B_u^π. To start of, note that if the Markov kernels $\{K^j\}_{j=1}^m$ are bandwidth constraint in the sense of Definition 2, the product kernel is a measure on the finite space $\mathcal{Y} := \bigotimes_{j=1}^m \mathcal{Y}^{(j)}$ and consequently its corresponding forward-backward kernel equals

$$q_u(x_1, x_2) = \sum_{y \in \mathcal{Y}} \frac{K(y|x_1, u)}{\mathbb{P}_0^{Y|U=u}(y)} K(y|x_2, u),$$

which is clearly bounded since $K \leqslant 1$.

Next, we turn to bounding the factor B_u^π, which functions the same for the shared- and local randomness classes of distributed protocols. No "strong" data processing argument is required here: The proof boils down to using the fact that conditional expectation contracts the L_2-norm and straightforward calculations.

Lemma 2.10. *Consider* B_u^π *as in* (2.27) *with* $\pi = N(0, \rho^2 c_\alpha^{-1/2} d^{-1} \overline{\Gamma})$ *as in Lemma 2.8. It holds that*

$$\mathrm{B}_u^\pi \leqslant \exp\left(C \frac{m n^2 \rho^4}{c_\alpha d}\right).$$

Proof. Since conditional expectation contracts the $L_2(\mathbb{P}_0)$-norm,

$$\prod_{j=1}^{m} \mathbb{E}_0^{Y^{(j)}|U=u} \mathbb{E}_0 \left[\mathscr{L}_\pi \left(X^{(j)} \right) \middle| Y^{(j)}, U = u \right]^2 \leqslant \prod_{j=1}^{m} \mathbb{E}_0^{X^{(j)}} \left[\mathscr{L}_\pi \left(X^{(j)} \right)^2 \right].$$

We now proceed to bound the first factor in the product on the right-hand side of the display above, which for a positive semi-definite choice of $\overline{\Gamma}$ equals

$$\int \mathbb{E}_0^{X^{(j)}} \exp \left((\sqrt{\overline{\Gamma}}(f+g))^\top \sum_{i=1}^{n} X_i^{(j)} - \frac{n}{2} \left\| \sqrt{\overline{\Gamma}} f \right\|_2^2 - \frac{n}{2} \left\| \sqrt{\overline{\Gamma}} g \right\|_2^2 \right) dN \left(0, \frac{\rho^2}{\sqrt{c_\alpha} d} I_{2d} \right) (f, g).$$

By direct computation involving the moment generating function of the normal distribution, the latter display equals

$$\int \exp \left(\frac{n\rho^2}{\sqrt{c_\alpha} d} z^\top \overline{\Gamma} z' \right) dN(0, I_{2d})(z, z').$$

We aim to apply the moment generating function of the Gaussian chaos, e.g. Lemma 6.2.2 in [210] to the above display. Using that ρ^2 satisfies (2.37) or (2.38) and since $\|\overline{\Gamma}\| = 1$ by the fact that $\overline{\Gamma}$ is idempotent, $\frac{n\rho^2}{c_\alpha^{1/2} d} \leqslant \sqrt{c_\alpha}$ with $c_\alpha > 0$ chosen small enough, the aforementioned result yields that there exists a constant $C > 0$

$$\prod_{j=1}^{m} \mathbb{E}_0^{X^{(j)}|U=u} \left[\mathscr{L}_\pi \left(X^{(j)} \right)^2 \right] \leqslant \exp \left(C c_\alpha^{-1} \frac{mn^2\rho^4}{d} \right), \tag{2.39}$$

where $C > 0$ is universal. □

The information lost by compressing a d dimensional observation $X^{(j)}$ into a b-bit transcript $Y^{(j)}$ is captured in a data processing inequality for the matrix Ξ_u and its trace, which comes in the form of Lemma 2.11. This can be seen as a "matrix analogue" of the (strong) data processing arguments for the mutual information used in Section 2.1.1 and Section 2.1.2.

Lemma 2.11. *Consider the matrix Ξ_u^j given in* (2.24)*. It holds that $\Xi_u^j \leqslant nI_d$ and*

$$Tr(\Xi_u^j) \leqslant 2\log(2) n (\log_2 |\mathcal{Y}^{(j)}|).$$

In particular, for $\log_2 |\mathcal{Y}^{(j)}| \leqslant b$,

$$Tr(\Xi_u^j) \leqslant \left(2\log(2) \frac{b}{d} \bigwedge 1 \right) nd.$$

Both statements of the lemma are known results, see e.g. Lemma 3 in [227] and Theorem 2 of [30], respectively. The "strong" data processing part of the lemma concerns the trace of the covariance, where the loss of information due to $Y^{(j)}$ being constrained to take values in a b-bit sample space is captured. When $b \ll d$, the latter

data processing inequality is stronger than the data processing inequality implied by the statement $\Xi_u^j \leqslant nI_d$. We provide a proof that is adapted to our setting in Section 2.5.3 of the chapter appendix, for the sake of completeness. The proof can be seen to it crucially rely on sub-Gaussianity, this time of the data, which is reminiscent of the relationship between sub-Gaussianity of the mixture likelihood and the strong data-processing inequalities for the mutual information of Section 2.1.

Combining Lemma 2.8 with the above assertions, we obtain the following lower bound.

Lemma 2.12. *Let $\mathscr{T}^b$ denote the class of b-bit bandwidth constrained shared- or local randomness distributed testing protocols and let ρ satisfy either* (2.37) *or* (2.38)*, 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 $\pi = N(0, c_\alpha^{-1/2} d^{-1} \rho^2 \overline{\Gamma})$ for a symmetric, idempotent matrix $\overline{\Gamma} \in \mathbb{R}^{d \times d}$ with $d/2 \leqslant rank(\overline{\Gamma}) \leqslant d$.

Remark 5. The lemma, combined with the earlier drawn conclusion that for any $\alpha \in (0,1)$ there exists $c_\alpha > 0$ small enough such that $\pi(H_\rho^c) \leqslant \alpha$ (recall (2.28)) finishes the proof of Theorem 2.3. The lemma also allows us to derive lower bounds for other alternative hypotheses H_ρ, as long as $\pi(H_\rho^c)$ can be shown to be small. For example, for the class of alternatives $\{f \in \mathbb{R}^d : \|f\|_1 \geqslant \rho\}$.

Proof. By Lemma 2.8, what left to show is that, for ρ satisfying (2.37) in the case of shared randomness and (2.38) in the case of local randomness, with $c_\alpha > 0$ small enough, the conditions required to obtain (2.31) and (2.32) hold (respectively) and the respective expressions for A_u and B_u^π are sufficiently close to 1. The latter follows from Lemma 2.10. By the first assertion of Lemma 2.11,

$$\|\Xi_u\| \leqslant \sum_{j=1}^m \|\Xi_u^j\| \leqslant mn.$$

For shared randomness protocols, ρ^2 is assumed to satisfy (2.37), which yields

$$\varrho^2 \|\Xi_u\| \leqslant \frac{mn\rho^2}{\sqrt{c_\alpha} d} \leqslant \sqrt{c_\alpha}. \tag{2.40}$$

By the third assertion of Lemma 2.11,

$$\mathrm{Tr}(\Xi_u) = \sum_{j=1}^m \mathrm{Tr}(\Xi_u^j) \leqslant \min\{2 \log 2 \cdot \frac{b}{d}, 1\} mnd. \tag{2.41}$$

For local randomness protocols, (2.38) and Lemma 2.41 implies

$$\frac{2\varrho^2}{\sqrt{c_\alpha} d^2} \mathrm{Tr}(\Xi_u) \leqslant \frac{\min\{2 \log 2 \cdot \frac{b}{d}, 1\} mn\rho^2}{\sqrt{c_\alpha} d} \leqslant 2 \log 2 \sqrt{c_\alpha}. \tag{2.42}$$

This verifies the conditions of Lemma 2.8. We also can use (2.41) to bound A_u in 2.8, from which we obtain that

$$\begin{aligned}\inf_{T\in\mathcal{T}_{pub}(b)}\mathcal{R}(H_\rho,T) &\geqslant 1-\sqrt{2(e^{C(\frac{mn^2\rho^4}{c_\alpha d}+\frac{m^2n^2\rho^4(b\wedge d)}{c_\alpha d^2})}-1)}-\pi(H_\rho^c)\\ &\geqslant 1-\sqrt{2(e^{2Cc_\alpha}-1)}-\pi(H_\rho^c)>\alpha-\pi(H_\rho^c),\end{aligned}$$

whenever ρ^2 satisfies (2.37) for $c_\alpha>0$ small enough. This finishes the proof for the shared randomness case. In the case of local randomness, we similarly obtain that

$$\begin{aligned}\inf_{T\in\mathcal{T}_{priv}(b)}\mathcal{R}(H_\rho,T) &\geqslant 1-\sqrt{2(e^{C(\frac{mn^2\rho^4}{c_\alpha d}+\frac{m^2n^2\rho^4(b\wedge d)^2}{c_\alpha d^3})}-1)}-\pi(H_\rho^c)\\ &\geqslant 1-\sqrt{2(e^{2Cc_\alpha}-1)}-\pi(H_\rho^c)>\alpha-\pi(H_\rho^c),\end{aligned}$$

for ρ^2 satisfying (2.38) and $c_\alpha>0$ small enough. □

2.4 A complete lower bound for testing under differential privacy constraints

The primary outcome of this section is presented as a theorem outlining the lower bound rate for the detection threshold for distributed testing protocols that adhere to differential privacy constraints, with and without the use of shared randomness. The optimality of the lower bounds is confirmed in Chapter 3, by introducing distributed differentially private testing protocols for both the shared randomness and local randomness classes that achieve the rates specified by the theorem (up to poly-logarithmic terms).

The methods constructed in Section 3.2 that attain the rates of the theorem are $(\epsilon,0)$-differentially private protocols matching the rate (up to poly-logarithmic terms) for the range $(nm)^{-1}<\epsilon\leqslant n^{-1/2}$. Whenever $\epsilon\gtrsim 1/\sqrt{n}$, within the class of distributed (ϵ,δ)-differential privacy protocols, we derive matching upper bounds for Theorem 2.4 for δ satisfying $\log(1/\delta)\asymp nmd$. Together with the upper bound of Theorem 3.2, the theorem below yields Theorem 1.2. The lower bound applies to all (ϵ,δ)-differentially private protocols where δ is small enough in comparison to m,d,n and ϵ. The range ϵ considered in the upper bound guarantees that we can set $\log(1/\delta)\asymp nmd$.

Theorem 2.4. *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*

$$0<\epsilon\leqslant 1 \text{ and } 0\leqslant\delta\leqslant\left(c_\alpha m^{-3/2}\wedge nd^{-1}\epsilon^2\wedge n^{1/2}d^{-1/2}\epsilon^2\right)^{1+p} \text{ for some } p>0, \tag{2.43}$$

the condition

$$\rho^2 < c_\alpha \left(\frac{d}{mn\sqrt{n\epsilon^2 \wedge 1}\sqrt{n\epsilon^2 \wedge d}} \bigwedge \left(\frac{\sqrt{d}}{\sqrt{m}n\sqrt{n\epsilon^2 \wedge 1}} \bigvee \frac{1}{mn^2\epsilon^2} \right) \right), \tag{2.44}$$

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 (2.43)*, the condition*

$$\rho^2 < c_\alpha \left(\frac{d\sqrt{d}}{mn(n\epsilon^2 \wedge d)} \bigwedge \left(\frac{\sqrt{d}}{\sqrt{m}n\sqrt{n\epsilon^2 \wedge 1}} \bigvee \frac{1}{mn^2\epsilon^2} \right) \right), \tag{2.45}$$

implies that

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

Remark 6. As in the bandwidth constraint case, the proof of the theorem reveals that the theorem holds for other classes of alternatives as well. In particular, the lower bounds hold for any H_ρ such that $N(0, c_\alpha^{-1/2} d^{-1} \rho^2 I_d)(H_\rho) \leqslant \alpha$ for $c_\alpha > 0$ small enough, whilst ρ satisfies (2.44) or (2.45).

Next, we provide a proof for the theorem. The first part of the proof follows a similar structure as that of the bandwidth constrained problem of the previous section, whose notation we shall also use here. We aim to use the Brascamp-Lieb type inequality of Section 2.2, by employing Lemma 2.8. To that extent, consider $\pi = N(0, c_\alpha^{1/2} d^{-1/2} \rho^2 \overline{\Gamma})$ for a symmetric idempotent matrix $\overline{\Gamma} \in \mathbb{R}^{d \times d}$ and the corresponding Bayes risk

$$P_0^m \breve{K} T + \int P_f^m \breve{K} (1 - T) d\pi(f), \tag{2.46}$$

where $\breve{K}$ is the product kernel, suppressing (integrating out) the shared randomness in the notation, corresponding to a distributed protocol T with (ϵ, δ)-DP Markov kernels $\{\breve{K}^j\}_{j=1}^m$. A first obstacle to deploying Lemma 2.8 to the Bayes risk above is that the forward-backward channel corresponding to $\breve{K}$, $(x_1, x_2) \mapsto \breve{q}_u(x_1, x_2)$ as defined in (2.20) is not necessarily bounded. This issue is specific to $\delta > 0$, as for $(\epsilon, 0)$-DP protocols the induced Radon-Nikodym derivatives are always bounded, see Lemma 2.34. Lemma 2.32 combined with Lemma 2.30 in the chapter appendix yield that for all $\alpha \in (0, 1)$ there exists $(\epsilon, 3\delta)$-DP Markov kernels $\{\tilde{K}^j\}_{j=1}^m$ such that the Bayes risk is bounded from below by

$$P_0^m \tilde{K} T + \int P_f^m \tilde{K} (1 - T) d\pi(f) - \alpha, \tag{2.47}$$

and with a bounded forward-backward channel.

Another issue suffered by (ϵ, δ)-DP Markov kernels with $\delta > 0$, is that one has very poor control over the higher moments of the local likelihoods

$$L^j_{\pi,u}(y) := \frac{dP_\pi K^j(\cdot|X^{(j)}, u)}{dP_0 K^j(\cdot|X^{(j)}, u)}(y),$$

which are required to sufficiently bound the corresponding quantity B_u, as defined in (2.27). In order to overcome this, we will use a coupling argument that allows a comparison of P_0K^j with $P_\pi K^j$. This lemma forms an essential building block for our data processing argument for B_u, but also allows us to overcome the aforementioned hurdles preventing the direct application of Lemma 2.8.

Lemma 2.13. *Let K^j satisfy a (ϵ, δ)-differential privacy constraint for $0 < \epsilon \leqslant 1$. Consider $\pi = N(0, c_\alpha^{1/2} d^{-1/2} \rho^2 \overline{\Gamma})$, with ρ^2 satisfying (2.44) or (2.45), with $\epsilon \leqslant 1/\sqrt{n}$ and $\delta \leqslant c_\alpha(m^{-1} \wedge \epsilon)$.*

For all measurable sets A it holds that

$$P_\pi K^j\left(A|X^{(j)}, u\right) \leqslant \left(1 + c_\alpha^{1/4} m^{-1/2}\right) P_0 K^j\left(A|X^{(j)}, u\right) + 2\delta + \frac{c_\alpha}{m^{3/2}} \tag{2.48}$$

and

$$P_\pi K^j\left(A|X^{(j)}, u\right) \geqslant \left(1 - c_\alpha^{1/4} m^{-1/2}\right) P_0 K^j\left(A|X^{(j)}, u\right) - 2\delta - \frac{c_\alpha}{m^{3/2}} \tag{2.49}$$

for all $c_\alpha > 0$ small enough.

We defer the proof of the lemma to Section 2.4.2. The lemma that follows can be seen as a consequence of the previous lemma.

Lemma 2.14. *Let $\pi = N(0, c_\alpha^{-1/2} d^{-1/2} \rho^2 \overline{\Gamma})$ for an arbitrary positive semidefinite $\overline{\Gamma}$ and let $\{K^j\}_{j=1}^m$ correspond to a (ϵ, δ)-DP distributed protocol T for the testing problem of (2.1) (i.e. K^j satisfies (1.5)). Furthermore, assume that $\epsilon \leqslant 1/\sqrt{n}$ and define for $j = 1, \ldots, m$ the events*

$$A_{j,u} := \left\{ y : |L^j_{\pi,u}(y) - 1| \leqslant \frac{4m^{1/2}}{\alpha} \right\}$$

and define

$$\tilde{K}^j(B|x, u) := K^j(B \cap A_{j,u}|x, u) + K^j(A^c_{j,u}|x, u) \frac{P_0 K^j(B \cap A_{j,u}|X^{(j)}, u)}{P_0 K^j(A_{j,u}|X^{(j)}, u)}.$$

Suppose in addition that $\delta \leqslant c_\alpha/m$. Then,

(a) The collection $\{\tilde{K}^j\}_{j=1}^m$ are $(\epsilon, 2\delta)$-DP Markov kernels.

(b) It holds $P_0\tilde{K}^j(\cdot|X^{(j)}, u)$-a.s. that

$$\tilde{L}^j_{\pi,u}(y) := \int \frac{d\tilde{K}^j(y|x, u)}{dP_0\tilde{K}^j(y|X^{(j)}, u)} dP^n_\pi(x)$$

satisfies

$$|\tilde{L}^j_{\pi,u}(y) - 1| \leqslant \frac{5m^{1/2}}{\alpha}. \tag{2.50}$$

(c) It holds that

$$P_0^m KT + \int P_f^m K(1-T)d\pi(f) \geqslant P_0^m \tilde{K}T + \int P_f^m \tilde{K}(1-T)d\pi(f) - \alpha,$$

where $\tilde{K}$ is the product kernel corresponding to $\{\tilde{K}^j\}_{j=1}^m$.

Applying the lemma above to the Bayes risk in (2.47), (with the roles of K^j and $\tilde{K}^j$ swapped) we obtain that the aforementioned expression is lower bounded by

$$P_0^m KT + \int P_f^m K(1-T)d\pi(f) - 2\alpha, \tag{2.51}$$

where K denotes the product kernel of an $(\epsilon, 6\delta)$-DP distributed protocol $\{T, \{K^j\}_{j=1}^m, \mathbb{P}^U\}$. Note that the shared randomness component is still suppressed (integrated out) in the notation above. Since $\tilde{K}$ has a bounded forward-backward channel, so does K, also when conditioned on the shared randomness component. That is, $K(\cdot|x,u)$ has a $\mathbb{P}_0^{Y|U=u}$-a.s. uniformly bounded Radon-Nikodym derivative

$$x \mapsto \frac{dK(\cdot|x,u)}{dP_0K(\cdot|X,u)}(y).$$

Consequently, the Brascamp-Lieb machinery applies (in particular Lemma 2.8), from which we obtain that for all (ϵ,δ)-DP kernels $\{\breve{K}^j\}_{j=1}^m$, there exist $(\epsilon, 6\delta)$-DP kernels $\{K^j\}_{j=1}^m$ such that

$$P_0^m \breve{K}T + \sup_{f\in H_\rho} P_f^m \breve{K}(1-T) \geqslant 1 - \sqrt{(1/2)\int (\mathrm{A}_u \mathrm{B}_u^\pi - 1)\, d\mathbb{P}^U(u)} + \pi(H_\rho^c) - \alpha, \tag{2.52}$$

where $\{K^j\}_{j=1}^m$ satisfies (2.50), with $\pi = N(0, c_\alpha^{-1/2} d^{-1/2}\rho^2\overline{\Gamma})$ with $\overline{\Gamma}$ any symmetric, idempotent matrix with rank proportional to d. We highlight here that the quantities A_u and B_u^π correspond to the quantities as defined in (2.26) and (2.27), respectively, with the underlying kernels $\{K^j\}_{j=1}^m$ corresponding to the kernels "approximating" $\{\breve{K}^j\}_{j=1}^m$. Next, we aim to apply Lemma 2.8, for which we need to sufficiently bound B_u^π and A_u for shared- and local randomness protocols.

We start with the bound on A_u, for which we proceed by a data processing argument for the matrix Ξ_u under the $(\epsilon, 6\delta)$-differential privacy constraint. This comes in the guise of Lemma 2.15 below. Its proof is deferred to the end of the section.

Lemma 2.15. *Let $0 < \epsilon \leqslant 1$ and let $Y^{(j)}$ be a transcript generated by a (ϵ, δ)-differential privacy constraint distributed protocol, with $0 < \epsilon \leqslant 1$ and $0 \leqslant \delta \leqslant \left(\left(nd^{-1} \wedge n^{1/2}d^{-1/2}\right)\epsilon^2\right)^{1+p}$ for some $p > 0$. The matrix Ξ_u^j as defined in (2.24) satisfies*

$$Tr\left(\Xi_u^j\right) \leqslant (Cn^2\epsilon^2) \wedge (nd)$$

for a fixed constant $C > 0$.

The lemma, combined with the first assertion of Lemma 2.11 implies in particular that $\|\Xi_u^j\| \leqslant (Cn^2\epsilon^2) \wedge n$, as Ξ_u^j is symmetric and positive definite. Combining this with (2.44) and the triangle inequality, we obtain

$$\varrho^2\|\Xi_u\| \leqslant \varrho^2 \sum_{j=1}^{m} \|\Xi_u^j\| \leqslant \frac{m\left((Cn^2\epsilon^2) \wedge n\right)\rho^2}{\sqrt{c_\alpha}d} \leqslant C\sqrt{c_\alpha}. \tag{2.53}$$

Similarly, (2.45) yields

$$\frac{2\rho^2}{\sqrt{c_\alpha}d^2}\mathrm{Tr}\left(\Xi_u^j\right) \leqslant \frac{m\left((Cn^2\epsilon^2) \wedge (nd)\right)\rho^2}{\sqrt{c_\alpha}d^2} \leqslant C\sqrt{c_\alpha}/\sqrt{d}. \tag{2.54}$$

The last two displays together finish the verification of the conditions of Lemma 2.8. The above data processing inequalities for Ξ_u^j and bounds on ρ^2 also yield a bound on A_u as defined in Lemma 2.8. In case of shared randomness protocols, using (2.31), (2.53), Lemma 2.15 and (2.44), we obtain

$$\mathrm{A}_u \leqslant \exp\left(C^2 c_\alpha\right).$$

In case of local randomness protocols, combining (2.32) with (2.54) and (2.45) yields the above bound on A_u.

Next, we turn to B_u^π. Lemma 2.10 proven in the previous section implies $\mathrm{B}_u^\pi \leqslant \exp(C\frac{mn^2\rho^4}{d})$. Whenever $\epsilon > n^{-1/2}$, this bound is actually tight in terms of rate in the exponent, but whenever $\epsilon \leqslant n^{-1/2}$ a much more involved data processing argument is needed than the one used in the bandwidth constraint case, in conjunction with Lemma 2.14. We provide a bound in the form of Lemma 2.16 below. Both proofs are based on coupling arguments, where the two different couplings constructed result in the different rates observed in the condition of the theorem.

Lemma 2.16. *Let $\pi = N(0, d^{-1}\rho^2\overline{\Gamma})$, with $\overline{\Gamma} \in \mathbb{R}^{d\times d}$ a symmetric idempotent matrix,*

$$\rho^2 \leqslant c_\alpha d^{1/2}/(\sqrt{m}n^{\frac{3}{2}}\epsilon) \vee c_\alpha/(mn^2\epsilon^2)$$

and $\{K^j\}_{j=1}^m$ correspond to a (ϵ, δ)-DP distributed protocol with transcripts $Y^{(j)}$ such that $0 < \epsilon \leqslant 1$, $\delta \lesssim c_\alpha(m^{-1} \wedge \epsilon)$ and

$$|L_{\pi,u}^j(y) - 1| \leqslant \frac{5m^{1/2}}{\alpha} \quad P_0K^j(\cdot|X^{(j)}, u)\text{-a.s.}$$

Then, there exists a universal constant $C > 0$ such that

$$\mathrm{B}_u^\pi \leqslant e^{C\sqrt{c_\alpha}}. \tag{2.55}$$

Combining the lemma above with the bound $\mathrm{B}_u^\pi \leqslant \exp(C\frac{mn^2\rho^4}{d})$ (which follows from Lemma 2.10), we obtain that (2.55) holds whenever ρ satisfies (2.37) or (2.38). Combining this with the bounds on A_u derived earlier, and following a similar calculation as in the proof of Lemma 2.12, we obtain the lemma below. The proof of Theorem 2.4 is finished by noticing that for the alternative hypothesis in question, $\pi(H_\rho^c) \leqslant \alpha$ for c_α small enough (see also (2.28)). After stating Lemma 2.17 below, we finish the section by providing the proofs for the Lemmas 2.14, 2.15 and 2.16.

Lemma 2.17. *Let $\mathscr{T}$ denote the class of shared- or local randomness distributed testing protocols satisfying a (ϵ, δ)-differential privacy constraint for $0 < \epsilon \leqslant 1$, $0 \leqslant \delta \leqslant \left(c_\alpha m^{-1} \wedge c_\alpha \epsilon m^{-1/2} \wedge n\epsilon^2 \wedge n^2 d^{-1}\epsilon^2 \wedge n^{3/2}d^{-1/2}\epsilon^2\right)$ and let ρ satisfy either (2.37) or (2.38), 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 $\pi = N(0, c_\alpha^{-1/2} d^{-1}\rho^2 \overline{\Gamma})$ for a symmetric, idempotent matrix $\overline{\Gamma} \in \mathbb{R}^{d\times d}$ with rank$(\overline{\Gamma}) \asymp d$.

Remark 7. The lemma also allows us to derive lower bounds for other alternative hypotheses H_ρ, as long as $\pi(H_\rho^c)$ can be shown to be small, e.g. the class of alternatives $\{f \in \mathbb{R}^d : \|f\|_1 \geqslant \rho\}$.

Proof. Putting the results of the section together, the Lemmas 2.15 and 2.16 with the condition on ρ^2, we obtain that in the shared randomness case, there exists a symmetric, idempotent matrix $\overline{\Gamma} \in \mathbb{R}^{d\times d}$ with $\mathrm{rank}(\overline{\Gamma}) \asymp d$ such that $\log \mathrm{A}_u\mathrm{B}_u^\pi$ is bounded by $C\sqrt{c_\alpha}$ for a universal constant $C > 0$, whenever $c_\alpha > 0$ is small enough. We conclude that,

$$\begin{aligned}\inf_{T\in\mathscr{T}_{SR}^{(\epsilon,\delta)}} \mathcal{R}(H_\rho, T) &\geqslant 1 - \sqrt{2(\mathrm{A}_u\mathrm{B}_u^\pi - 1)} - \pi(H_\rho^c) \\ &\geqslant 1 - \sqrt{2(e^{C\sqrt{c_\alpha}} - 1)} - \pi(H_\rho^c) > \alpha - \pi(H_\rho^c),\end{aligned}$$

whenever ρ^2 satisfies (2.37) for $c_\alpha > 0$ small enough. This finishes the proof for the shared randomness case. In the case of local randomness, $\log \mathrm{A}_u\mathrm{B}_u^\pi$ is bounded by $C\sqrt{c_\alpha}$ when ρ^2 satisfies (2.38) for $c_\alpha > 0$ small enough, yielding

$$\inf_{T\in\mathscr{T}_{LR}^{(\epsilon,\delta)}} \mathcal{R}(H_\rho, T) \geqslant 1 - \sqrt{2(e^{C\sqrt{c_\alpha}} - 1)} - \pi(H_\rho^c) > \alpha - \pi(H_\rho^c).$$

□

2.4.1 Proof of the differential privacy data processing Lemmas 2.14 and 2.16

For some of the results in this section, we suppress the dependence of the Markov kernels on the draw of the shared randomness u, as it bares no relevance to the results here.

At the heart of the proofs of Lemmas 2.14 and 2.16, is the Lemma 2.13. A proof of this lemma is provided in Section 2.4.2.

2.4.1.1 Proof of Lemma 2.14

Proof. The first statement follows by Lemma 2.31 in the chapter appendix. For the second statement, we first note that by Lemma 2.13

$$\begin{aligned}\frac{c_\alpha^{1/4}}{m^{1/2}} + \delta + \frac{c_\alpha}{m^{3/2}} &\geqslant (P_\pi - P_0)K^j\left(\{|L^j_{\pi,u} - 1| \geqslant 4m^{1/2}/\alpha\}|X^{(j)}, u\right)\\ &= P_0K^j\left((L^j_{\pi,u} - 1)\mathbb{1}\{|L^j_{\pi,u} - 1| \geqslant 4m^{1/2}/\alpha\}|X^{(j)}, u\right)\\ &\geqslant 4\frac{m^{1/2}}{\alpha}P_0K^j\left(|L^j_{\pi,u} - 1| \geqslant 4m^{1/2}/\alpha|X^{(j)}, u\right),\end{aligned}$$

where the second inequality follows from the fact that $m^{1/2}/\alpha \geqslant 1$ and $L^j_{\pi,u} \geqslant 0$. Using that $\delta \leqslant c_\alpha/m$, we obtain that

$$P_0K^j(A^c_{j,u}|X^{(j)}, u) \leqslant \frac{\alpha(c_\alpha^{1/4} + c_\alpha(1 + m^{-1}))}{4m} := \eta_\alpha. \tag{2.56}$$

Since $K^j(B|x, u) \leqslant \tilde{K}^j(B|x, u)$ for all measurable $B \subset A_{j,u}$ and $P_0\tilde{K}^j(\cdot|X^{(j)}, u)$ has no support outside of $A_{j,u}$, it holds that

$$\frac{dK^j(\cdot|x, u)}{dP_0\tilde{K}^j(\cdot|X^{(j)}, u)}(y) \leqslant \frac{dK^j(\cdot|x, u)}{dP_0K^j(\cdot|X^{(j)}, u)}(y),$$

for all $y \in A_{j,u}$ (and hence $P_0\tilde{K}^j(\cdot|X^{(j)}, u)$-a.s.). Similarly, we have for P_π-a.s. all x's that

$$\frac{K^j(A^c_{j,u}|x, u)}{P_0K^j(A_{j,u}|X^{(j)}, u)}\frac{dP_0K^j(\cdot \cap A_{j,u}|X^{(j)}, u)}{dP_0\tilde{K}^j(\cdot|X^{(j)}, u)}(y) \leqslant \frac{K^j(A^c_{j,u}|x, u)}{P_0K^j(A_{j,u}|X^{(j)}, u)} \leqslant \frac{1}{1 - \eta_\alpha},$$

using that $K^j \leqslant 1$ and $P_0K^j(A_{j,u}|X^{(j)}, u) \geqslant 1 - \eta_\alpha$. By standard arguments and the above two statements, it follows that

$$\begin{aligned}\int \frac{d\tilde{K}^j(y|x, u)}{dP_0\tilde{K}^j(y|X^{(j)}, u)}dP^n_\pi(x) &\leqslant \mathbb{1}_{A_{j,u}}(y)\int \frac{dK^j(y|x, u)}{dP_0\tilde{K}^j(y|X^{(j)}, u)}dP^n_\pi(x) + \frac{1}{1 - \eta_\alpha}\\ &= \mathbb{1}_{A_{j,u}}(y)L^j_{\pi,u}(y) + \frac{1}{1 - \eta_\alpha}.\end{aligned}$$

Applying the definition of the event $A_{j,u}$ and using that $\alpha \leqslant 1$, we obtain that for $c_\alpha > 0$ small enough

$$\tilde{L}^j_{\pi,u} - 1 \leqslant \frac{4m^{1/2}}{\alpha} + \frac{1}{1-\eta_\alpha} - 1 \leqslant \frac{5m^{1/2}}{\alpha}.$$

Using that $\tilde{L}^j_{\pi,u} - 1 \geqslant -1$, we obtain (2.50), proving statement *(b)*.

For the third statement, we will aim to apply Lemma 2.30. By the construction of $\tilde{K}^j$ and the triangle inequality,

$$\|P_0 K^j(\cdot|X^{(j)}, u) - P_0\tilde{K}^j(\cdot|X^{(j)}, u)\|_{\mathrm{TV}} \leqslant 2\left\|P_0 K^j(\cdot \cap A^c_{j,u}|X^{(j)}, u)\right\|_{\mathrm{TV}}.$$

The latter is bounded by $\alpha/(2m)$ (see (2.56)). Similarly,

$$\|P_\pi K^j(\cdot|X^{(j)}, u) - P_\pi\tilde{K}^j(\cdot|X^{(j)})\|_{\mathrm{TV}} \leqslant 2P_\pi K^j(A^c_{j,u}|X^{(j)}, u).$$

By Lemma 2.13,

$$P_\pi K^j(A^c_{j,u}|X^{(j)}, u) \leqslant \left(1 + c_\alpha^{1/4} m^{-1/2}\right) P_0 K^j(A^c_j|X^{(j)}, u) + \delta + \frac{c_\alpha}{m^{3/2}}.$$

Again using (2.56) and the fact that $\delta \leqslant c_\alpha/m$ yield that the latter is also bounded by $\alpha/4m$ for $c_\alpha > 0$ small enough. The condition and small enough choice of $c_\alpha > 0$ yields that the conditions of Lemma 2.30 and the conclusion of (c) follows. □

2.4.1.2 Proof of Lemma 2.16

Proof. Write $L^j_{\pi,u}(Y^{(j)}) \equiv L^j_\pi$ and let $V_\pi \equiv V^j_\pi := L^j_\pi - 1$. Using that $\mathbb{E}_0 \mathscr{L}_\pi(\tilde{X}^{(j)}) = 1$ and that by the law of total probability

$$\mathbb{E}_0^{Y^{(j)}|U=u}\mathbb{E}_0\left[\mathscr{L}_\pi(\tilde{X}^{(j)})\middle|Y^{(j)}, U = u\right] = 1,$$

it follows that $\mathbb{E}_0^{Y^{(j)}|U=u} V_\pi = 0$ and

$$\mathbb{E}_0^{Y^{(j)}|U=u}\left(L^j_\pi\right)^2 = 1 + \mathbb{E}_0^{Y^{(j)}|U=u} L^j_\pi(L^j_\pi - 1) = 1 + \mathbb{E}_\pi^{Y^{(j)}|U=u} V_\pi.$$

Define $V^+_\pi := 0 \vee V_\pi$ and let $V^-_\pi = -(0 \wedge V_\pi)$, which are both nonnegative random variables, with $V_\pi = V^+_\pi - V^-_\pi$. We have

$$\begin{aligned}\mathbb{E}_\pi^{Y^{(j)}|U=u} V^+_\pi &= \int_0^\infty \mathbb{P}_\pi^{Y^{(j)}|U=u}\left(V^+_\pi \geqslant t\right) dt \\ &= \int_0^T \mathbb{P}_\pi^{Y^{(j)}|U=u}\left(V^+_\pi \geqslant t\right) dt + \int_T^\infty \mathbb{P}_\pi^{Y^{(j)}|U=u}\left(V^+_\pi \geqslant t\right) dt. \qquad (2.57)\end{aligned}$$

Taking $T = \frac{5m^{1/2}}{\alpha}$, the second term is equal to zero as

$$V_\pi^+ \leqslant |L^j_{\pi,u}(y) - 1| \leqslant \frac{5m^{1/2}}{\alpha} \quad P_0K^j(\cdot|X^{(j)}, u)\text{-a.s.}$$

and $P_\pi \sim P_0$ (which in turn implies $\mathbb{P}_\pi^{Y^{(j)}|U=u} \sim \mathbb{P}_0^{Y^{(j)}|U=u}$). The integrand of the first term equals $P_\pi^n K^j(\{V_\pi \geqslant t\}|X^{(j)}, u)$. By Lemma 2.13, it holds that

$$P_\pi K^j(\{V_\pi \geqslant t\}|X^{(j)}, u) \leqslant \left(1 + c_\alpha^{1/4} m^{-1/2}\right) P_0K^j(\{V_\pi \geqslant t\}|X^{(j)}, u) + \delta + \frac{c_\alpha}{m^{3/2}}.$$

It follows that (2.57) is bounded from above by

$$\left(1 + c_\alpha^{1/4} m^{-1/2}\right) \int_0^T \mathbb{P}_0^{Y^{(j)}|U=u}\left(V_\pi^+ \geqslant t\right) dt + T\delta + T\frac{c_\alpha}{m^{3/2}} \leqslant \left(1 + c_\alpha^{1/4} m^{-1/2}\right) \mathbb{E}_0^{Y^{(j)}|U=u} V_\pi^+ + T\delta + T\frac{c_\alpha}{m^{3/2}}.$$

Similarly, we have

$$\begin{aligned}\mathbb{E}_\pi^{Y^{(j)}|U=u} V_\pi^- &= \int_0^\infty \mathbb{P}_\pi^{Y^{(j)}|U=u}\left(V_\pi^- \geqslant t\right) dt \\ &= \int_0^T \mathbb{P}_\pi^{Y^{(j)}|U=u}\left(V_\pi^- \geqslant t\right) dt + \int_T^\infty \mathbb{P}_\pi^{Y^{(j)}|U=u}\left(V_\pi^- \geqslant t\right) dt. \qquad (2.58)\end{aligned}$$

Choosing $T \geqslant 1$ here results in the second term being zero, as $L^j_\pi \geqslant 0$. Applying Lemma 2.13, the right-hand side of the above display is further bounded from below by

$$\left(1 - c_\alpha^{1/4} m^{-1/2}\right) \int_0^T \mathbb{P}_0^{Y^{(j)}|U=u}\left(V_\pi^- \geqslant t\right) dt - T\delta - T\frac{c_\alpha}{m^{3/2}} \geqslant \left(1 - c_\alpha^{1/4} m^{-1/2}\right) \mathbb{E}_0^{Y^{(j)}|U=u} V_\pi^- - T\delta - T\frac{c_\alpha}{m^{3/2}},$$

where the inequality uses $V_\pi^- \leqslant 1$.

Combining the above bounds with the fact that $V_\pi^+ + V_\pi^- = |V_\pi|$ and $\mathbb{E}_0^{Y^{(j)}|U=u} V_\pi = 0$ yields that

$$\mathbb{E}_\pi^{Y^{(j)}|U=u} V_\pi = \mathbb{E}_\pi^{Y^{(j)}|U=u} V_\pi^+ - \mathbb{E}_\pi^{Y^{(j)}|U=u} V_\pi^- \leqslant \frac{c_\alpha^{1/4}\mathbb{E}_0^{Y^{(j)}|U=u}|V_\pi|}{\sqrt{m}} + 2T\delta + 2T\frac{c_\alpha}{m^{3/2}}.$$

Plugging in the choice of $T = 5m^{1/2}/\alpha$ and using that $\delta \leqslant c_\alpha m^{-3/2}$, we obtain

$$\mathbb{E}_\pi^{Y^{(j)}|U=u} V_\pi \leqslant \frac{c_\alpha^{1/4}\mathbb{E}_0^{Y^{(j)}|U=u}|V_\pi|}{\sqrt{m}} + \frac{20c_\alpha}{m\alpha}.$$

If $\mathbb{E}_0^{Y^{(j)}|U=u}|V_\pi| \lesssim m^{-1/2}$, we obtain $\mathbb{E}_\pi^{Y^{(j)}|U=u}V_\pi \lesssim m^{-1}(c_\alpha^{1/4} + c_\alpha/\alpha)$. Assume next that $\mathbb{E}_0^{Y^{(j)}|U=u}|V_\pi| \gtrsim m^{-1/2}$. Then,

$$\mathbb{E}_\pi^{Y^{(j)}|U=u}V_\pi \lesssim \frac{c_\alpha^{1/4}\mathbb{E}_0^{Y^{(j)}|U=u}|V_\pi|}{\sqrt{m}}.$$

Since $\mathbb{E}_\pi^{Y^{(j)}|U=u}V_\pi = \mathbb{E}_0^{Y^{(j)}|U=u}V_\pi^2$ and using that by Cauchy-Schwarz $\mathbb{E}_0^{Y^{(j)}|U=u}|V_\pi|$ is bounded above by $\sqrt{\mathbb{E}_0^{Y^{(j)}|U=u}V_\pi^2}$, we obtain that

$$\sqrt{\mathbb{E}_0^{Y^{(j)}|U=u}V_\pi^2} \lesssim C' c_\alpha^{1/4} m^{-1/2}$$

for a universal constant $C' > 0$ depending only on α. In both cases, we obtain that

$$\mathrm{B}_u^\pi = \prod_{j=1}^m \left(1 + \mathbb{E}_\pi^{Y^{(j)}|U=u}V_\pi\right) = \prod_{j=1}^m \left(1 + \mathbb{E}_0^{Y^{(j)}|U=u}V_\pi^2\right) \leqslant e^{C\sqrt{c_\alpha}}$$

for universal constant $C > 0$, finishing the proof of the lemma. □

2.4.2 Proof of Lemma 2.13

We start by proving the following general lemma, which is essentially Lemma 6.1 in [129], but for which we provide a proof that is perhaps easier to verify.

Lemma 2.18. *Consider random variables $X_1, \ldots, X_n \overset{i.i.d.}{\sim} P_1$ and $\tilde{X}_1, \ldots, \tilde{X}_n \overset{i.i.d.}{\sim} P_2$ defined on the same space. Write $X = (X_1, \ldots, X_n)$, $\tilde{X} = (\tilde{X}_1, \ldots, \tilde{X}_n)$ and let K be a Markov kernel between the sample space of X (equivalently $\tilde{X}$) and an arbitrary target space, satisfying a (ϵ, δ)-differential privacy constraint (i.e. (1.5)) with $\epsilon \leqslant 1$.*

Suppose that there exists a coupling $\mathbb{P}$ of $(\tilde{X}, X)$ such that $\mathbb{P}^{\tilde{X}} = P_1^n$, $\mathbb{P}^X = P_2^n$ and

$$D_i := \mathbb{1}\left\{\tilde{X}_i \neq X_i\right\} \sim Ber(p)\,, \ \textit{i.i.d. for } i = 1, \ldots, n, \ p \in [0, 1]$$

under $\mathbb{P}$.

Then, it holds that

$$P_1^n K\left(A|\tilde{X}\right) \leqslant e^{4\epsilon np} P_2^n K\left(A|X\right) + 2\delta np e^{\epsilon + 2np\epsilon}. \tag{2.59}$$

Proof. Let $\mathbb{E}$ denote expectation with respect to $\mathbb{P}$ and write $D = (D_i)_{i\in[n]}$, $S := \sum_{i=1}^n D_i$. We start by noting that

$$\mathbb{E}\left[K(A|\tilde{X})|S = 0\right] = \mathbb{E}\left[K(A|X)|S = 0\right]. \tag{2.60}$$

Next, we show that for all $k \in [n]$,

$$e^{-\epsilon}\mathbb{E}\left[K(A|\tilde{X})|S=k-1\right]-\delta \leqslant \mathbb{E}\left[K(A|\tilde{X})|S=k\right] \leqslant e^{\epsilon}\mathbb{E}\left[K(A|\tilde{X})|S=k-1\right]+\delta. \tag{2.61}$$

Write $v_{(-i)} = (v_i)_{[n]\backslash\{i\}}$ for a vector $v \in \mathbb{R}^n$. Let $k \in [n]$ be given and let $\mathcal{V}_k$ denote the set of $v \in \{0,1\}^n$'s such that $\sum_{i=1}^n v_i = k$. Using the definition of differential privacy, the integrand in the conditional expectation satisfies

$$e^{-\epsilon}K(A|\tilde{X}_1,\ldots,\breve{X}_i,\ldots,\tilde{X}_n)-\delta \leqslant K(A|\tilde{X}) \leqslant e^{\epsilon}K(A|\tilde{X}_1,\ldots,\breve{X}_i,\ldots,\tilde{X}_n)+\delta, \tag{2.62}$$

for any random variable $\breve{X}_i$ taking values in the sample space of $\tilde{X}_i$. In particular, if $v_i = 1$ it holds that

$$\begin{aligned}\mathbb{E}\left[K(A|x_1,\ldots,X_i,\ldots,x_n)\middle|D_i=v_i,\tilde{X}_{(-i)}=x_{(-i)},D_{(-i)}=v_{(-i)}\right] \leqslant \\ e^{\epsilon}\mathbb{E}\left[K(A|x_1,\ldots,X_i,\ldots,x_n)\middle|D_i=0,\tilde{X}_{(-i)}=x_{(-i)},D_{(-i)}=v_{(-i)}\right]+\delta,\end{aligned}$$

for all x in the sample space of $\tilde{X}$. It follows by the law of total probability that

$$\mathbb{E}\left[K(A|\tilde{X})|D=v\right] \leqslant e^{\epsilon}\mathbb{E}\left[K(A|\tilde{X})|D_i=0,\, D_k=v_k \text{ for } k\in[n]\backslash\{i\}\right]+\delta,$$

for all $i \in [n]$. For $v \in \mathcal{V}_k$, the event $\{D=v\}$ is equal to the event $\{D=v,\, S=k\}$ and similarly it holds that

$$\{D_k=v_k \text{ for } k\in[n]\backslash\{i\},\, D_i=0\} = \{D_k=v_k \text{ for } k\in[n]\backslash\{i\},\, D_i=0,\, S=k-1\}.$$

Consider now the sets

$$\begin{aligned}\mathcal{V}_{k-1}(v) &:= \left\{v'\in\mathcal{V}_{k-1} : v_l=v'_l \text{ except for one } l\in[n]\right\} \text{ for } v\in\mathcal{V}_k,\\ \mathcal{V}_k(v') &:= \left\{v\in\mathcal{V}_k : v_l=v'_l \text{ except for one } l\in[n]\right\} \text{ for } v'\in\mathcal{V}_{k-1}.\end{aligned}$$

By what we have derived so far, it holds that any $v \in \mathcal{V}_k$ and $v' \in \mathcal{V}_{k-1}(v)$,

$$\mathbb{E}\left[K(A|\tilde{X})|D=v,\, S=k\right] \leqslant e^{\epsilon}\mathbb{E}\left[K(A|\tilde{X})|D=v',\, S=k-1\right]+\delta.$$

Consider $\{I_k(v) : v \in \mathcal{V}_k\}$ independent random variables (on a possibly enlarged probability space) taking values in $[n]$ such that $\mathbb{P}(I_k(v)=i) = 1/k$ whenever $v_i = 1$.

Combining the above with the total law of probability we find that

$$\begin{aligned}
\mathbb{E}\left[K(A|\tilde{X})|S=k\right] = \\
\frac{1}{\binom{n}{k}}\sum_{v\in\mathcal{V}_k}\mathbb{E}\left[K(A|\tilde{X})|D=v,\, S=k\right] \leqslant \\
e^{\epsilon}\frac{1}{\binom{n}{k}}\sum_{v\in\mathcal{V}_k}\mathbb{E}\left[K(A|\tilde{X})|D_{I(v)}=0,\, D_{-I(v)}=v_{-I(v)},\, S=k-1\right]+\delta = \\
e^{\epsilon}\frac{1}{\binom{n}{k}}\frac{1}{k}\sum_{v\in\mathcal{V}_k}\sum_{v'\in\mathcal{V}_{k-1}(v)}\mathbb{E}\left[K(A|\tilde{X})|D=v',\, S=k-1\right]+\delta = \\
e^{\epsilon}\frac{1}{\binom{n}{k}}\frac{1}{k}\sum_{v'\in\mathcal{V}_{k-1}}\sum_{v\in\mathcal{V}_k(v')}\mathbb{E}\left[K(A|\tilde{X})|D=v',\, S=k-1\right]+\delta = \\
e^{\epsilon}\frac{1}{\binom{n}{k-1}}\sum_{v'\in\mathcal{V}_{k-1}}\mathbb{E}\left[K(A|\tilde{X})|D=v',\, S=k-1\right]+\delta = \\
e^{\epsilon}\mathbb{E}\left[K(A|\tilde{X})|S=k-1\right]+\delta,
\end{aligned}$$

where it is used that $|\mathcal{V}_k| = \binom{n}{k}$,

$$\mathbb{P}(D_1=v_1,\ldots,D_n=v_n|S=k) = \mathbb{P}(D_1=\tilde{v}_1,\ldots,D_n=\tilde{v}_n|S=k)$$

for all $v=(v_i)_{i\in[n]}, \tilde{v}=(\tilde{v}_i)_{i\in[n]} \in \mathcal{V}_k$ and for any $v'\in\mathcal{V}_{k-1}$ there are $n-k+1$ ways to obtain $v\in\mathcal{V}_k$ such that $v_k = v'_k$ except for one $i\in[n]$.

By applying the privacy lower bound of (2.62) and repeating the same steps, we also find that

$$e^{-\epsilon}\mathbb{E}\left[K(A|\tilde{X})|S=k-1\right]-\delta \leqslant \mathbb{E}\left[K(A|\tilde{X})|S=k\right].$$

This proves (2.61), which, applying iteratively, results in the bound

$$e^{-\epsilon k}\mathbb{E}\left[K(A|\tilde{X})|S=0\right]-\delta k \leqslant \mathbb{E}\left[K(A|\tilde{X})|S=k\right] \leqslant e^{\epsilon k}\mathbb{E}\left[K(A|\tilde{X})|S=0\right]+\delta k e^{\epsilon k}, \tag{2.63}$$

for $k=0,1,\ldots,n$. By symmetry of the argument, the same inequalities hold for X in place of $\tilde{X}$. Using the above inequalities, we can bound

$$P_1K(A|\tilde{X}) = \mathbb{E}K(A|\tilde{X}) = \mathbb{E}^S\mathbb{E}\left[K(A|\tilde{X})|S\right],$$

by

$$\mathbb{E}^S e^{S\epsilon}\mathbb{E}\left[K(A|\tilde{X})|S=0\right]+\delta\mathbb{E}Se^{S\epsilon}.$$

Similarly, applying (2.61) with X in place of $\tilde{X}$, we find

$$\begin{aligned}
P_2K(A|X) &= \mathbb{E}^S\mathbb{E}\left[K(A|X)|S\right] \\
&\geqslant \mathbb{E}^S e^{-S\epsilon}\mathbb{E}\left[K(A|X)|S=0\right]-\mathbb{E}\delta S.
\end{aligned} \tag{2.64}$$

Combining the two inequalities with (2.60), we obtain that

$$P_1 K(A|\tilde{X}) \leqslant \frac{\mathbb{E}^S e^{S\epsilon}}{\mathbb{E}^S e^{-S\epsilon}} \left(P_2 K(A|X) + \mathbb{E}^S \delta S \right) + \delta \mathbb{E} S e^{S\epsilon}. \tag{2.65}$$

In view of the moment generating function of the binomial distribution,

$$\frac{\mathbb{E}^S e^{S\epsilon}}{\mathbb{E}^S e^{-S\epsilon}} = \left(\frac{1 + p(e^{\epsilon} - 1)}{1 + p(e^{-\epsilon} - 1)} \right)^n \leqslant e^{4np\epsilon},$$

where the inequality follows from $0 \leqslant \epsilon, p \leqslant 1$, the inequality $e^x - e^{-x} \leqslant 3x$ for $0 \leqslant x \leqslant 1$ and Taylor expanding $\log(1+x) = x - x^2/2 + \ldots$. By Chebyshev's association inequality (e.g. Theorem 2.14 in [36]), $\mathbb{E}^S S \mathbb{E}^S e^{S\epsilon} \leqslant \mathbb{E}^S S e^{S\epsilon}$. Consequently, using the nonnegativity of S,

$$\delta \left(\frac{\mathbb{E}^S e^{S\epsilon}}{\mathbb{E}^S e^{-S\epsilon}} \mathbb{E}^S S + \mathbb{E} S e^{S\epsilon} \right) \leqslant 2\delta \mathbb{E} S e^{S\epsilon}.$$

Lemma 2.37 (a straightforward calculation) now finishes the proof. □

We are now ready to prove the desired result.

Proof of Lemma 2.13. Consider $\tilde{X}^{(j)} \sim P_\pi$ and $X^{(j)} \sim P_0$. We shall construct two couplings for $(\tilde{X}^{(j)}, X^{(j)})$, one for two different regimes of ϵ:

$$1/\sqrt{n} \geqslant \epsilon > 1/\sqrt{mnd} \quad \text{and} \quad \epsilon \leqslant 1/\sqrt{mnd}.$$

That is, for each of the regimes, we derive a joint distribution of $(\tilde{X}^{(j)}, X^{(j)})$ called $\mathbb{P}_{\pi,0}$ such that $\tilde{X}^{(j)} \sim \mathbb{P}_{\pi,0}^{\tilde{X}^{(j)}} = P_\pi$ and $X^{(j)} \sim \mathbb{P}_{\pi,0}^{X^{(j)}} = P_0$. The specific couplings that we construct aim at assuring that $d_H(\tilde{X}^{(j)}, X^{(j)})$ is small with high probability. After the construction of both of the couplings, the result follows by an application of Lemma 2.18.

Case 1: Consider $1/\sqrt{n} \geqslant \epsilon > 1/\sqrt{mnd}$. In this case, follow a construction similar to that of Theorem D.6 in [157].

If $n = 1$, Pinsker's inequality (see e.g. Lemma 2.5 in [204]) followed Lemma 2.40 and Lemma 2.10 applied with $m = 1$ yield that

$$\|P_0 - P_\pi\|_{\mathrm{TV}} \leqslant \sqrt{\frac{1}{2} D_{\chi^2}(P_0; P_\pi)} \leqslant C \frac{\sqrt{c_\alpha} \rho^2}{\sqrt{d}}$$

for a universal constant $C > 0$ (which we let vary from line to line). By Lemma 2.41, there exists a coupling $\mathbb{P}_{\pi,0}$ such that $\tilde{X}^{(j)} \sim \mathbb{P}_{\pi,0}^{\tilde{X}^{(j)}} = P_\pi$ and $X^{(j)} \sim \mathbb{P}_{\pi,0}^{X^{(j)}} = P_0$ and

$$p := \mathbb{P}\left(\tilde{X}^{(j)} \neq X^{(j)} \right) \leqslant \left(C \frac{\sqrt{c_\alpha} \rho^2}{\sqrt{d}} \right) \wedge 1. \tag{2.66}$$

Applying Lemma 2.18, it follows that

$$\begin{aligned} P_\pi^n K^j(A|\tilde{X}^{(j)}) &= \mathbb{E}_{\pi,0} K^j(A|\tilde{X}^{(j)}) \\ &\leqslant e^{4\epsilon np} P_0 K^j(A|X^{(j)}) + 2\delta np e^{\epsilon + 2\epsilon np}. \end{aligned}$$

By applying condition (2.44) or (2.45) and the bound on p of (2.66), we obtain that

$$e^{4\epsilon np} \leqslant e^{C\frac{\sqrt{c_\alpha}\epsilon\rho^2}{d^{1/2}}} \leqslant 1 + C\sqrt{c_\alpha}/\sqrt{m}$$

Similarly, using that $\delta \leqslant \epsilon/\sqrt{m}$,

$$\delta e^{4\epsilon np} \leqslant \delta + C' c_\alpha / m^{3/2}.$$

Hence, the first identity (2.48) follows for $n = 1$ and a sufficiently small enough choice of $c_\alpha > 0$.

In what follows, consider $n > 1$. Consider V a uniform draw from the unit sphere in $\mathbb{R}^d$ and $Z \sim N(0, I_d)$, both independent of the other random variables considered. We have

$$\overline{X^{(j)}} := \frac{1}{n}\sum_{i=1}^n X_i^{(j)} \stackrel{d}{=} \frac{\|Z\|_2}{\sqrt{n}} V$$

for $X^{(j)} \sim \mathbb{P}_0^{X^{(j)}}$ (see e.g. [210] Exercise 3.3.7). Similarly,

$$\overline{\tilde{X}^{(j)}} \stackrel{d}{=} \frac{\left\| (I_d + \frac{n c_\alpha^{1/2}\rho^2}{d}\overline{\Gamma})^{1/2} Z \right\|_2}{\sqrt{n}} V.$$

Next, we note that for $\eta_1, \ldots, \eta_n \sim N(0, I_d)$ independent of $X^{(j)} = (X_1^{(j)}, \ldots, X_n^{(j)})$, we have

$$X^{(j)} \stackrel{d}{=} \left(\overline{X^{(j)}} + \eta_i - \frac{1}{n}\sum_{i=1}^n \eta_i \right)_{1 \leqslant i \leqslant n}. \tag{2.67}$$

To see this, note that both the left- and right-hand side are mean zero Gaussians and

$$\begin{aligned} \mathbb{E}\left(\overline{X^{(j)}} + \eta_i - \frac{1}{n}\sum_{i=1}^n \eta_i \right)\left(\overline{X^{(j)}} + \eta_k - \frac{1}{n}\sum_{i=1}^n \eta_i \right)^\top = \\ \frac{1}{n} I_d + \mathbb{1}_{i=k} I_d - \frac{2}{n} I_d + \frac{1}{n} I_d = \mathbb{1}_{i=k} I_d, \end{aligned}$$

which means that the covariances of the left-hand side and right-hand side of (2.67) are equal too. Noting that $\tilde{X}^{(j)} \stackrel{d}{=} (F + X_i^{(j)})_{i \in [n]}$ and $\overline{\tilde{X}^{(j)}} \stackrel{d}{=} F + \overline{X^{(j)}}$, where $F \sim N(0, \sqrt{c_\alpha} d^{-1} \rho^2 \overline{\Gamma})$ is independent of $\overline{X^{(j)}}$, it follows that

$$\tilde{X}^{(j)} \stackrel{d}{=} \left(\overline{\tilde{X}^{(j)}} + \eta_i - \frac{1}{n}\sum_{i=1}^n \eta_i \right)_{1 \leqslant i \leqslant n}$$

by similar reasoning. Since the matrix $(I - VV^\top)$ is idempotent, we have that

$$\eta_i = VV^\top \eta_i + (I - VV^\top)\eta_i$$

where $VV^\top \eta_i$ is independent of $(I - VV^\top)\eta_i$ and $V^\top \eta_i$ is standard normally distributed, both conditionally and unconditionally on V. We can write

$$\eta_i - \frac{1}{n}\sum_{i=1}^{n}\eta_i = VV^\top \eta_i - \frac{1}{n}\sum_{i=1}^{n} VV^\top \eta_i + G_i,$$

where

$$G_i := (I - VV^\top)\eta_i - \frac{1}{n}\sum_{i=1}^{n}(I - VV^\top)\eta_i$$

and G_i is independent of $VV^\top \eta_i - \frac{1}{n}\sum_{i=1}^{n} VV^\top \eta_i$. Let $\tilde{\eta}_i$ be identically distributed to η_i for $i = 1, \ldots, n$. Combining the above assertions, we have that

$$X^{(j)} \stackrel{d}{=} \left\{ V\left(\frac{\|Z\|_2}{\sqrt{n}} + V^\top \eta_i - \frac{1}{n}\sum_{i=1}^{n} V^\top \eta_i \right) + G_i \right\}_{i\in[n]} =: (C_i)_{i\in[n]}, \tag{2.68}$$

$$\tilde{X}^{(j)} \stackrel{d}{=} \left\{ V\left(\frac{\left\|(I_d + \frac{nc_\alpha^{1/2}\rho^2}{d}\overline{\Gamma})^{1/2} Z\right\|_2}{\sqrt{n}} + V^\top \tilde{\eta}_i - \frac{1}{n}\sum_{i=1}^{n} V^\top \tilde{\eta}_i \right) + G_i \right\}_{i\in[n]} =: (\tilde{C}_i)_{i\in[n]}. \tag{2.69}$$

As further notations, we introduce

$$\zeta_i := \|Z\|_2/\sqrt{n} + V^\top \eta_i - \frac{1}{n}\sum_{i=1}^{n} V^\top \eta_i,$$

$$\tilde{\zeta}_i := \|(I_d + nc_\alpha^{1/2}\rho^2 d\overline{\Gamma})^{1/2} Z\|_2/\sqrt{n} + V^\top \tilde{\eta}_i - \frac{1}{n}\sum_{i=1}^{n} V^\top \tilde{\eta}_i.$$

We have that $\zeta_i | Z \sim N\left(\frac{\|Z\|_2}{\sqrt{n}}, \left(1 - \frac{1}{n}\right)\right)$ and

$$\tilde{\zeta}_i | Z \sim N\left(\frac{\left\|(I_d + \frac{nc_\alpha^{1/2}\rho^2}{d}\overline{\Gamma})^{1/2} Z\right\|_2}{\sqrt{n}}, \left(1 - \frac{1}{n}\right)\right).$$

By e.g. Lemma 6.5, we find that their respective push forward measures $\mathbb{P}^{\zeta_i|Z}$ and $\mathbb{P}^{\tilde{\zeta}_i|Z}$ satisfy

$$\begin{aligned}\|\mathbb{P}^{\zeta_i|Z} - \mathbb{P}^{\tilde{\zeta}_i|Z}\|_{\mathrm{TV}} &\leqslant \frac{1}{2\sqrt{1-1/n}\sqrt{n}} \left| \|Z\|_2 - \left\| (I_d + \frac{nc_\alpha^{1/2}\rho^2}{d}\overline{\Gamma})^{1/2} Z \right\|_2 \right| \\ &\leqslant \frac{\sqrt{n}c_\alpha^{1/2}\rho^2}{d} \left| \frac{Z^\top \overline{\Gamma} Z}{\sqrt{Z^\top I_d Z} + \sqrt{Z^\top (I_d + \frac{nc_\alpha^{1/2}\rho^2}{d}\overline{\Gamma}) Z}} \right| \\ &\leqslant \|\overline{\Gamma}\| \frac{\sqrt{n}c_\alpha^{1/2}\rho^2}{d} \|Z\|_2,\end{aligned}$$

where the second inequality follows from $n > 1$ in addition to the identity $(\sqrt{a} - \sqrt{b})(\sqrt{a} + \sqrt{b}) = a - b$ and the final inequality follows from the fact that $\overline{\Gamma}$ is positive semidefinite and $\overline{\Gamma} \leqslant \|\overline{\Gamma}\| I_d$. By Lemma 2.41, there exists a coupling of $\zeta_i|Z$ and $\tilde{\zeta}_i|Z$ such that

$$\mathbb{P}\left(\zeta_i \neq \tilde{\zeta}_i | Z\right) \leqslant \frac{\sqrt{n}c_\alpha^{1/2}\rho^2\|\overline{\Gamma}\|\|Z\|_2}{2d} \wedge 1. \tag{2.70}$$

By the independence structure, it holds for any joint distribution $\mathbb{P}$ of $V, Z, C, \tilde{C}$, $\zeta = (\zeta_1, \ldots, \zeta_n)$, $\tilde{\zeta} = \tilde{\zeta}_1, \ldots, \tilde{\zeta}_n$, and $G = (G_i)_{i\in[n]}$ that

$$d\mathbb{P}^{C,\tilde{C}} = i = 1 \bigotimes^n d\mathbb{P}^{C_i,\tilde{C}_i|Z}.$$

Take $\mathbb{P}^{\zeta_i,\tilde{\zeta}_i|Z}$ satisfying (2.70) and set $(X^{(j)}, \tilde{X}^{(j)}) = (C, \tilde{C})$ under $\mathbb{P}_{\pi,0}$. We have that

$$C_i^{(j)}|Z = \tilde{C}_i^{(j)}|Z \iff \zeta_i|Z = \tilde{\zeta}_i|Z$$

whilst the random variables $\mathbb{1}\{C_i \neq \tilde{C}_i\}$ are independent Bernoulli distributed for $i = 1, \ldots, n$.

To summarize, we have now obtained that there exists a joint distribution $\mathbb{P}_{\pi,0}$ of $(Z, X^{(j)}, \tilde{X}^{(j)})$ such that $\left(Z, X^{(j)}, \tilde{X}^{(j)}\right)$ satisfy

$$p := \mathbb{P}\left(X_i^{(j)} \neq \tilde{X}_i^{(j)}\right) = \mathbb{P}^Z \mathbb{P}\left(\zeta_i \neq \tilde{\zeta}_i | Z\right) \leqslant \mathbb{E}^Z \frac{\sqrt{n}c_\alpha^{1/2}\rho^2\|\overline{\Gamma}\|\|Z\|_2}{2d} \wedge 1,$$

and

$$S := \sum_{i=1}^n \mathbb{1}\{\tilde{X}_i^{(j)} \neq X_i^{(j)}\} \sim \mathrm{Bin}(n, p).$$

Let $\mathbb{E}_{\pi,0}$ denote the corresponding expectation. Consequently, by applying Lemma 2.18, we have for any measurable A that

$$\begin{aligned}P_\pi^n K^j(A|\tilde{X}^{(j)}) &= \mathbb{E}_{\pi,0} K^j(A|\tilde{X}^{(j)}) = \mathbb{E}_{\pi,0}^{\tilde{X}^{(j)}, X^{(j)}} K^j(A|\tilde{X}^{(j)}) \\ &\leqslant e^{4\epsilon np} P_0 K^j(A|X^{(j)}) + 2\delta np e^{\epsilon + 2\epsilon np}.\end{aligned}$$

By (2.70), $\|\overline{\Gamma}\| \asymp 1$ and the fact that $\|Z\|_2$ is $\sqrt{d}$-sub-exponential (using e.g. Proposition 2.7.1 in [210]), we obtain that

$$\mathbb{E}^Z p^k \leqslant \frac{\tilde{C}^k n^{k/2} (c_\alpha^{1/4}\rho)^{2k} k^k}{d^{k/2}}$$

for a universal constant $\tilde{C} > 0$. It follows that

$$\mathbb{E}^Z e^{4\epsilon n p_Z} \leqslant 1 + \sum_{k=1}^{\infty} \frac{4^k \tilde{C}^k \epsilon^k n^{3k/2} (c_\alpha^{1/4}\rho)^{2k} k^k}{d^{k/2} k!} \leqslant 1 + C' \frac{\epsilon n^{3/2} \rho^2}{\sqrt{c_\alpha} d^{1/2}},$$

for a universal constant $C' > 0$, where the second inequality follows from Stirling's approximation, the fact that under the assumptions on ρ^2 (i.e. condition (2.44) or (2.45)) that

$$\frac{\epsilon n^{3/2} c_\alpha^{1/2} \rho^2}{d^{1/2}} \leqslant \sqrt{c_\alpha}/\sqrt{m}$$

and a sufficiently small enough choice of $c_\alpha > 0$, such that the series is dominated by its first term. Similarly, using that $\delta \leqslant c_\alpha/m$,

$$\begin{aligned} \delta \mathbb{E}^Z e^{4\epsilon n p_Z} &= \delta + \delta \sum_{k=1}^{\infty} \frac{2^{2k} \epsilon^k n^k \mathbb{E}^Z p_Z^k}{k!} \\ &\leqslant \delta + C' c_\alpha / m^{3/2}. \end{aligned}$$

The first identity we wish to show, i.e. (2.48), now follows. Using the same coupling, the lower bound of (2.49) readily follows by a similar analysis, which closes the first case.

Case 2: Consider $\epsilon \leqslant 1/\sqrt{mnd}$. We will make use of the total variation coupling between $\tilde{X}_i^{(j)} \sim N(f, I_d)$ and $X_i^{(j)} \sim N(0, I)$, as given by Lemma 2.41. Since

$$\|N(0, I_d) - N(f, I_d)\|_{\mathrm{TV}} \leqslant \left(\frac{1}{2}\|f\|_2\right) \wedge 2$$

(see e.g. Lemma 6.5), we can couple the two data sets observation wise independently (simply taking the product space) such that

$$\sum_{i=1}^{n} \mathbb{1}\{\tilde{X}_i^{(j)} \neq X_i^{(j)}\} \sim \mathrm{Bin}(n, p_f)$$

where $p_f = (\|f\|_2/4) \wedge 1$. Given $k \in \mathbb{N}$, $\|f\|_2 \overset{d}{=} d^{-1/2} c_\alpha^{1/4} \rho \|N(0, I_d)\|_2$ and $\|N(0, I_d)\|_2$ is $\sqrt{d}$-sub-exponential we obtain (using e.g. Proposition 2.7.1 in [210])

$$\int p_f^k d\pi(f) \leqslant \int (\|f\|_2/4)^k d\pi(f) \leqslant \tilde{C}^k k^k (c_\alpha^{1/4}\rho)^k,$$

for a universal constant $\tilde{C} > 0$. The assumed condition on ρ (i.e. (2.44) or (2.45)) yields

$$\epsilon n c_\alpha^{1/4} \rho \leqslant c_\alpha^{1/4}/\sqrt{m},$$

which by similar arguments as before implies

$$\begin{aligned}\mathbb{E}^f e^{4\epsilon n p_f} &\leqslant 1 + C' c_\alpha^{1/4}/\sqrt{m},\\ \delta \mathbb{E}^f e^{4\epsilon n p_f} &\leqslant \delta + C' c_\alpha^{1/2}/m^{3/2},\end{aligned}$$

for a universal constant $C' > 0$. By applying Lemma 2.18 and using the assumptions on ρ, we obtain that

$$\begin{aligned}P_\pi^n K^j(A|\tilde{X}^{(j)}) = \mathbb{E}_{\pi,0} K^j(A|\tilde{X}^{(j)}) &= \int \mathbb{E}_{f,0} K^j(A|\tilde{X}^{(j)}) d\pi(f)\\ &\leqslant (1 + C c_\alpha^{1/4}/\sqrt{m}) P_0 K^j(A|X^{(j)}) + 2\delta + C c_\alpha^{1/2}/m^{3/2}\end{aligned}$$

as desired. Again, (2.49) follows by similar steps. □

2.4.3 Proof of Lemma 2.15

Proof. The bound $\mathrm{Tr}(\Xi_u^j) \leqslant nd$ follows by the fact that conditional expectation contracts the L_2-norm, i.e. the same arguments as in the proof of Lemma 2.11. For the second statement, we start introducing the notations $\overline{X^{(j)}} = n^{-1}\sum_{i=1}^n X_i^{(j)}$ and

$$G_i = \left\langle \mathbb{E}_0\left[n\overline{X^{(j)}}|Y^{(j)}, U = u\right], X_i^{(j)} \right\rangle.$$

For the remainder of the proof, consider versions of $X^{(j)}$ and $Y^{(j)}$ defined on the same probability given $U = u$, and we shall write as a shorthand

$$\mathbb{P}^j \equiv \mathbb{P}_0^{(X^{(j)}, Y^{(j)})|U=u} \text{ and } \mathbb{E}^j \equiv \mathbb{E}_0^{(X^{(j)}, Y^{(j)})|U=u}.$$

For random variables V, W defined on the same probability space, it holds that

$$\mathbb{E}W\mathbb{E}[W|V] = \mathbb{E}\mathbb{E}[W|V]\mathbb{E}[W|V],$$

since $W - \mathbb{E}[W|V]$ is orthogonal to $\mathbb{E}[W|V]$. Combining this fact with the linearity of the inner product and conditional expectation, we see that

$$\mathrm{Tr}(\Xi_u^j) = \mathbb{E}_0^{Y^{(j)}|U=u}\left\|\mathbb{E}_0[n\overline{X^{(j)}}|Y^{(j)}, U = u]\right\|_2^2 = \sum_{i=1}^n \mathbb{E}^j G_i. \tag{2.71}$$

Define also

$$\breve{G}_i = \left\langle \mathbb{E}_0[n\overline{X^{(j)}}|Y^{(j)}, U = u], \breve{X}_i^{(j)} \right\rangle,$$

where $\breve{X}_i^{(j)}$ is an independent copy of $X_i^{(j)}$ (defined on the same, possibly enlarged probability space) and note that $\mathbb{E}^j \breve{G}_i = 0$. Write $G_i^+ := 0 \vee G_i$ and $G_i^- = -(0 \wedge G_i)$. We have

$$\begin{aligned}
\mathbb{E}^j G_i^+ &= \int_0^\infty \mathbb{P}^j \left(G_i^+ \geqslant t\right) dt = \int_0^T \mathbb{P}^j \left(G_i^+ \geqslant t\right) + \int_T^\infty \mathbb{P}^j \left(G_i^+ \geqslant t\right) \\
&\leqslant e^\epsilon \int_0^T \mathbb{P}^j \left(\breve{G}_i^+ \geqslant t\right) dt + T\delta + \int_T^\infty \mathbb{P}^j \left(G_i^+ \geqslant t\right) \\
&\leqslant \int_0^T \mathbb{P}^j \left(\breve{G}_i^+ \geqslant t\right) dt + 2\epsilon \int_0^T \mathbb{P}^j \left(\breve{G}_i^+ \geqslant t\right) dt + T\delta + \int_T^\infty \mathbb{P}^j \left(G_i^+ \geqslant t\right) \\
&\leqslant \int_0^\infty \mathbb{P}_0^j \left(\breve{G}_i^+ \geqslant t\right) dt + 2\epsilon \int_0^\infty \mathbb{P}^j \left(\breve{G}_i^+ \geqslant t\right) dt + T\delta + \int_T^\infty \mathbb{P}^j \left(G_i^+ \geqslant t\right),
\end{aligned}$$

where in the second to last inequality follows by Taylor expansion and the fact that $\epsilon \leqslant 1$. Similarly, we obtain

$$\begin{aligned}
\mathbb{E}^j G_i^- &\geqslant \int_0^T \mathbb{P}^j \left(G_i^- \geqslant t\right) \\
&\geqslant e^{-\epsilon} \int_0^T \mathbb{P}^j \left(\breve{G}_i^- \geqslant t\right) dt - T\delta \\
&\geqslant \int_0^T \mathbb{P}^j \left(\breve{G}_i^- \geqslant t\right) dt - 2\epsilon \int_0^\infty \mathbb{P}^j \left(\breve{G}_i^- \geqslant t\right) dt - T\delta \\
&\geqslant \int_0^\infty \mathbb{P}^j \left(\breve{G}_i^- \geqslant t\right) dt - 2\epsilon \int_0^\infty \mathbb{P}^j \left(\breve{G}_i^- \geqslant t\right) dt - T\delta - \int_T^\infty \mathbb{P}^j \left(\breve{G}_i^- \geqslant t\right) dt.
\end{aligned}$$

Putting these together with $G_i = G_i^+ - G_i^-$, we get

$$\begin{aligned}
\mathbb{E}^j G_i &\leqslant \int_0^\infty \mathbb{P}^j \left(\breve{G}_i^+ \geqslant t\right) dt - \int_0^\infty \mathbb{P}^j \left(\breve{G}_i^- \geqslant t\right) dt + 2\epsilon \int_0^\infty \mathbb{P}^j \left(|\breve{G}_i| \geqslant t\right) dt \\
&\qquad + 2T\delta + \int_T^\infty \mathbb{P}^j \left(G_i^+ \geqslant t\right) dt + \int_T^\infty \mathbb{P}^j \left(\breve{G}_i^- \geqslant t\right) dt \\
&= \mathbb{E}^j \breve{G}_i + 2\epsilon \mathbb{E}^j |\breve{G}_i| + 2T\delta + \int_T^\infty \mathbb{P}^j \left(G_i^+ \geqslant t\right) dt + \int_T^\infty \mathbb{P}^j \left(\breve{G}_i^- \geqslant t\right) dt.
\end{aligned}$$

The first term in the last display equals 0. For the second term, observe that

$$\breve{G}_i \Big| \left[Y^{(j)}, X^{(j)}, U = u\right] \sim N(0, \|\mathbb{E}_0[n\overline{X^{(j)}}|Y^{(j)}, U = u]\|_2^2),$$

so

$$\mathbb{E}^j |\breve{G}_i| = \mathbb{E}^{X^{(j)}, Y^{(j)}} \mathbb{E}^{\breve{X}^{(j)}} |\breve{G}_i| = \mathbb{E}\|\mathbb{E}[n\overline{X^{(j)}}|Y^{(j)}, U = u]\|_2 \leqslant \sqrt{\mathrm{Tr}(\Xi_u^j)}$$

where the last inequality is Cauchy-Schwarz. To bound the terms

$$\int_T^\infty \mathbb{P}^j \left(G_i^+ \geqslant t\right) dt + \int_T^\infty \mathbb{P}^j \left(\breve{G}_i^- \geqslant t\right) dt$$

we shall employ tail bounds, which follow after showing that G_i is $\sqrt{dn}$-sub-exponential. To see this, note that by applying Cauchy-Schwarz and Jensen's inequality followed by the law of total probability, we have that

$$\begin{aligned}\mathbb{E}^j e^{t|G_i|} &= \mathbb{E}^j e^{t|\langle \mathbb{E}_0[n\overline{X^{(j)}}|Y^{(j)},U=u],X_i^{(j)}\rangle|} \\ &\leqslant \mathbb{E}^j e^{\frac{t}{2}\left(\left\|\mathbb{E}_0[n\overline{X^{(j)}}|Y^{(j)},U=u]\right\|_2^2+\left\|X_i^{(j)}\right\|_2^2\right)} \\ &\leqslant \sqrt{\mathbb{E}_0 e^{t\left\|\mathbb{E}_0[n\overline{X^{(j)}}|Y^{(j)},U=u]\right\|_2^2}}\sqrt{\mathbb{E}_0 e^{t\left\|X_i^{(j)}\right\|_2^2}} \\ &\leqslant \mathbb{E}_0 e^{t|\langle n\overline{X^{(j)}},X_i^{(j)}\rangle|},\end{aligned}$$

where the last equality follows from the fact that conditional expectation contracts the L_1-norm and the fact that U is independent of $X^{(j)}$.

Next, we bound $\mathbb{E}_0^{X^{(j)}} e^{t|\langle n\overline{X^{(j)}},X_i^{(j)}\rangle|}$. By the triangle inequality and Cauchy-Schwarz,

$$\mathbb{E}_0^{X^{(j)}} e^{t|\langle n\overline{X^{(j)}},X_i^{(j)}\rangle|} \leqslant \sqrt{\mathbb{E}_0^{X^{(j)}} e^{2t|\langle\sum_{k\neq i}^n X^{(j)},X_i^{(j)}\rangle|}}\sqrt{\mathbb{E}_0^{X_i^{(j)}} e^{2t|\langle X_i^{(j)},X_i^{(j)}\rangle|}}.$$

The random variable $\langle X_i^{(j)}, X_i^{(j)}\rangle$ is χ_d^2-distributed, so by Lemma 2.36 we obtain that

$$\mathbb{E}_0^{X_i^{(j)}} e^{2t|\langle X_i^{(j)},X_i^{(j)}\rangle|} = \left(\mathbb{E}e^{2tN(0,1)^2}\right)^d \leqslant e^{2td+8t^2d},$$

whenever $t \leqslant 1/8$. By Lemma 2.38,

$$\mathbb{E}_0^{X^{(j)}} e^{2t|\langle\sum_{k\neq i}^n X^{(j)},X_i^{(j)}\rangle|} \leqslant e^{\frac{4}{2}(t^2(n-1)d+t^4(n-1)^2d/4)},$$

where the inequality follows by Lemma 2.36 if $t^2(n-1)^2 \leqslant 1/8$. By the fact that $G_i^+ \leqslant |G_i|$ and Markov's inequality,

$$\mathbb{P}^j(G_i^+ > T) \leqslant \mathbb{P}^j(|G_i| > T) \leqslant e^{-tT}\mathbb{E}^j e^{t|G_i|}, \text{ for all } T, t > 0.$$

Combining this with the bound for the moment generating function derived above means that for $\delta = 0$, the result follows from letting $T \to \infty$. If $\delta > 0$, take $T = 32(d \vee \sqrt{nd})\log(1/\delta)$ to obtain that

$$\int_T^\infty \mathbb{P}^j\left(G_i^+ \geqslant t\right) dt \leqslant e^{-\log(1/\delta)}.$$

It is easy to see that the same bound applies to $\int_T^\infty \mathbb{P}_0^j\left(\breve{G}_i^- \geqslant t\right) dt$. We obtain that

$$\sum_{i=1}^n \mathbb{E}^j G_i \leqslant 2n\epsilon\sqrt{\mathrm{Tr}(\Xi_u^j)} + 64\delta(d \vee \sqrt{nd})\log(1/\delta) + 2n\delta.$$

If $\sqrt{\mathrm{Tr}(\Xi_u^j)} \leqslant n\epsilon$, the lemma holds (there is nothing to prove). So assume instead that $\sqrt{\mathrm{Tr}(\Xi_u^j)} \geqslant n\epsilon$. Combining the above display with (2.71), we get

$$\sqrt{\mathrm{Tr}(\Xi_u^j)} \leqslant 2n\epsilon + 64\delta \frac{d \vee \sqrt{nd}}{n\epsilon} \log(1/\delta) + \frac{2}{\epsilon}\delta.$$

Since $x^p \log(1/x)$ tends to 0 as $x \to 0$ for any $p > 0$, the result follows for $\delta \leqslant \left(\left(\frac{n}{d} \wedge \frac{n^{1/2}}{\sqrt{d}}\right)\epsilon^2\right)^{1+p}$ for some $p > 0$ as this implies that the last two terms are $O(n\epsilon)$. □

Chapter acknowledgements: We would like to thank Elliot H. Lieb for a helpful comment regarding the proof of Lemma 2.9.

2.5 Appendix

2.5.1 Mutual information, entropy and data processing

For a discrete random variable X and arbitrary random variable Y, define the *entropy of* X as

$$H(X) = -\sum_x \mathbb{P}(X = x) \log \mathbb{P}(X = x)$$

and the *conditional entropy of* X *given* Y as

$$H(X|Y) := -\int \sum_x \mathbb{P}(X = x|Y = y) \log \mathbb{P}(X = x|Y = y) d\mathbb{P}^Y(y).$$

The function $x \mapsto -x \log x$ is concave, so by Jensen's inequality we have $H(X) \geqslant 0$ and $H(X|Y) \geqslant 0$. Similarly, we have $H(X) \geqslant H(X|Y)$, i.e. *conditioning reduces entropy*. Following from this conditioning, for an arbitrary random vector Z, we similarly can conclude that conditioning also reduces conditional entropy:

$$H(X|Y) = \int H(X|Y = y) d\mathbb{P}^Y(y) \geqslant \int H(X|Y = y, Z) d\mathbb{P}^Y(y) = H(X|Y, Z).$$

If X and Y are independent, it is easy to see that $H(X|Y) = H(X)$. Furthermore, if $X \to Y \to Z$ form a Markov chain, $H(X|Y, Z) = H(X|Y)$. For random variables X, Y, Z we define the mutual information between X and Y and conditional mutual information between X and Y given Z as

$$\begin{aligned} I(X;Y) &= D_{\mathrm{KL}}(\mathbb{P}^{(X,Y)} \| \mathbb{P}^X \times \mathbb{P}^Y), \\ I(X;Y|Z = z) &= D_{\mathrm{KL}}(\mathbb{P}^{(X,Y)|Z=z} \| \mathbb{P}^{X|Z=z} \times \mathbb{P}^{Y|Z=z}), \\ I(X;Y|Z) &= \int I(X;Y|Z = z) d\mathbb{P}^Z(z). \end{aligned}$$

Next we recall some properties of the mutual information. First we note that $I(X,Y) = 0$ if and only if X is independent from Y. The chain rule for the mutual information between the random vector $Y = (Y^{(1)}, \ldots, Y^{(m)})$ and V is

$$I(V;Y) = \sum_{j=1}^{m} I(V;Y^{(j)}|Y^{(1)}, \ldots, Y^{(j-1)}), \tag{2.72}$$

which follows from straightforward algebra. For a discrete random variable X and an arbitrary random variable Y, we obtain the following relationship with entropy:

$$\begin{aligned} I(X;Y) &= \mathbb{E}^{(X,Y)} \log \frac{dP^{(X,Y)}}{dP^X dP^Y} \\ &= E^{(X,Y)} \log \frac{1}{dP^X} - \mathbb{E}^{(X,Y)} \log \frac{1}{d\mathbb{P}^{(X|Y=y)}} \\ &= H(X) - H(X|Y). \end{aligned} \tag{2.73}$$

By similar arguments, for an arbitrary random variable Z, we have

$$I(X;Y|Z) = H(X|Z) - H(X|Y,Z). \tag{2.74}$$

Next, we prove three lemmas that that reveal themselves as valuable for proving the data processing bounds of Sections 2.1 and 2.1.2. The first is a well known result that states says that mutual information is necessarily decreasing as we move further along a Markov chain, making it a type of data processing inequality.

Lemma 2.19 (Mutual information data processing inequality)**.** *Let and X, Z be discrete random variables and let Y an arbitrary random variable such that $X \to Y \to Z$ forms a Markov chain. It holds that*

$$I(X;Z) \leqslant I(X;Y) \quad \text{and } I(X;Z) \leqslant I(Y;Z).$$

Proof. This is a straightforward consequence of (2.74) combined with the fact that conditioning reduces entropy, which yields

$$I(X;Z) = H(X) - H(X|Z) \leqslant H(X) - H(X|Y,Z) = H(X) - H(X|Y) = I(X;Y),$$

where the second equality follows from the fact that $H(X|Y,Z) = H(X|Y)$ by fact that $X \to Y \to Z$ forms a Markov chain. Similarly,

$$I(X;Z) = H(Z) - H(Z|X) \leqslant H(Z) - H(Z|Y,X) = H(X) - H(X|Y) = I(Y;Z).$$

□

The next lemma is well known: it shows that mutual information cannot exceed the logarithm of the cardinality of the sample space.

Lemma 2.20. *Let X be a discrete random variable taking values in $\mathcal{X}$ and let Z be an arbitrary random variable. It holds that*

$$I(X;Z) \leqslant H(X) \leqslant \log|\mathcal{X}|.$$

Proof. The first inequality follows by (2.74) and the fact that $H(X|Y) \geqslant 0$. For the second inequality, by concavity of the log we have

$$H(X) = -\sum_x \mathbb{P}(X=x)\log\mathbb{P}(X=x) \leqslant \log\left(\sum_x \frac{\mathbb{P}(X=x)}{\mathbb{P}(X=x)}\right) = \log|\mathcal{X}|.$$

□

The next lemma is especially useful in the distributed setup. It allows us to exploit the independent nature of the machines in order to obtain an additive bound over the "local" mutual information for each machine. In our setup, we take into account the possible presence of a shared source of randomness and therefore it merits its own proof. A variation of the lemma excluding shared randomness is given for instance in [169].

Lemma 2.21 (Tensorization of the mutual information)**.** *Let us assume that the discrete random variable V and the discrete random vector W are such that the pair (V,W) is independent from the random variable U and the discrete random vector $Y = (Y_1,\dots,Y_m)$ satisfies that Y_j is conditionally independent from $Y_{1:j-1} := (Y_1,\dots,Y_{j-1})$ given U and (V,W), then*

$$I(V;Y) \leqslant \sum_{j=1}^m I(V;Y_j|U) + \sum_{j=1}^m I(W;Y_j|U,V).$$

Proof. First note that in view of (2.73) and since conditioning reduces entropy

$$I\big((Y,U);V\big) = H(V) - H(V|Y,U) \geqslant H(V) - H(V|Y) = I(Y;V).$$

Furthermore, by the chain rule (2.72) and the independence of U and V,

$$I\big((Y,U);V\big) = I\big(Y;V|U\big) + I(U;V) = I(Y;V|U).$$

Similarly, by the chain rule and nonnegativity of mutual information,

$$I\big(V;Y|U\big) = I\big((V,W);Y|U\big) - I\big(W;Y|U,V\big) \leqslant I\big((V,W);Y|U\big).$$

By the identity (2.74) and the chain rule (2.72),

$$\begin{aligned} I\big((V,W);Y|U\big) &= H(Y|U) - H(Y|V,W,U) \\ &= \sum_{j=1}^m H(Y_j|Y_{1:j-1},U) - H(Y_j|V,W,Y_{1:j-1},U). \end{aligned}$$

Since conditioning reduces entropy we have $H(Y_j|Y_{1:j-1},U) \leqslant H(Y_j|U)$. Furthermore, by the conditional independence of $Y_{1:j-1}$ and Y_j given (U,V,W) results in $H(Y_j|V,W,Y_{1:j-1},U) = H(Y_j|V,W,U)$. Using these two facts, we obtain that

$$\begin{aligned} I\big((V,W);Y|U\big) &\leqslant \sum_{j=1}^{m} H(Y_j|,U) - H(Y_j|V,W,U) \\ &= \sum_{j=1}^{m} I\big((V,W);Y_j|U\big). \end{aligned}$$

Combining the above displays and again applying the chain rule we now obtain that

$$I(Y;V) \leqslant \sum_{j=1}^{m} I\big((V,W);Y_j|U\big) = \sum_{j=1}^{m} \big[I\big(V;Y_j|U\big) + I\big(W;Y_j|U,V\big)\big].$$

□

The following lemma is Proposition 1 in the technical note [76]. It can be seen as a "distance-based" version of the original Fano's inequality. We provide a proof based on [76] for completeness.

Lemma 2.22. *Let $V,\hat{V},W$ be random variables forming a Markov chain $V \to W \to \hat{V}$, where V and $\hat{V}$ take values in a metric space $(\mathcal{V},d)$ with $|\mathcal{V}| < \infty$ and V is uniformly distributed on $\mathcal{V}$. Let*

$$N^*(t) := \max_{v\in\mathcal{V}} |\left\{v' \in \mathcal{V} : d(v,v') \leqslant t\right\}|, \; N_*(t) := \min_{v\in\mathcal{V}} |\left\{v' \in \mathcal{V} : d(v,v') \leqslant t\right\}|.$$

If $|\mathcal{V}| - N_*(t) > N^*(t)$, it holds that

$$Pr\Big(d(\hat{V},V) \geqslant t\Big) \geqslant 1 - \frac{I(V;W) + \log 2}{\log\left(|\mathcal{V}|/N^*(t)\right)}. \tag{2.75}$$

Remark 8. For the Hamming distance, the above reduces to the classical Fano's inequality of e.g. [97]. The advantage of employing this particular expression of Fano's inequality resides in its applicability without the necessity of delineating the packing set. Rather, one may choose to designate a prior distribution over a subset of finite cardinality and subsequently selecting a distribution for V that minimizes the mutual information.

Proof. Define the random variable $S = \mathbb{1}\{d(V,\hat{V}) < t\}$. By the chain rule for entropy,

$$H(S,V|\hat{V}) = H(V|\hat{V}) + H(S|V,\hat{V}).$$

The last term equals 0 as S is $\sigma(V,\hat{V})$-measurable. Conversely, since conditioning reduces entropy

$$H(S,V|\hat{V}) = H(S|\hat{V}) + H(V|S,\hat{V}) \leqslant H(S) + H(V|S,\hat{V}).$$

The second term equals

$$\mathbb{P}(S=1)H(V|S=1,\hat{V})+\mathbb{P}(S=0)H(V|S=0,\hat{V}).$$

Since conditionally on $S=1$, V is with probability 1 in a set of cardinality at most $N^*(t)$, it follows from the fact that conditioning reduces entropy and Lemma 2.20 that $H(V|S=1,\hat{V}) \leqslant H(V|S=1) \leqslant \log N^*(t)$. Similarly, $H(V|S=0,\hat{V}) \leqslant \log(|\mathcal{V}|-N_*(t))$. We now have that

$$H(V|\hat{V}) \leqslant H(S)+(1-\mathbb{P}(S=0))\log N^*(t)+\mathbb{P}(S=0)\log(|\mathcal{V}|-N_*(t)).$$

For a Markov chain $V \to W \to \hat{V}$, $H(V|W)=H(V|W,\hat{V}) \leqslant H(V|\hat{V})$ since conditioning reduces the entropy. Furthermore, since S equals either 0 or 1, $H(S) \leqslant \log 2$ by Lemma 2.20. We obtain that

$$H(V|W) \leqslant \log 2+(1-\mathbb{P}(S=0))\log N^*(t)+\mathbb{P}(S=0)\log(|\mathcal{V}|-N_*(t)),$$

which after rearranging yields

$$\Pr\left(d(\hat{V},V) \geqslant t\right)=\mathbb{P}(S=0) \geqslant \frac{H(V|W)-\log N^*(t)-\log 2}{\log\left(\frac{|\mathcal{V}|-N_*(t)}{N^*(t)}\right)}.$$

Since V is assumed to be uniform on $\mathcal{V}$, $H(V)=\log \mathcal{V}$. By (2.73), $I(V;W)=\log \mathcal{V}-H(V|W)$, which yields

$$\Pr\left(d(\hat{V},V) \geqslant t\right) \geqslant \frac{\log\left(\frac{|\mathcal{V}|}{N^*(t)}\right)}{\log\left(\frac{|\mathcal{V}|-N_*(t)}{N^*(t)}\right)}-\frac{I(V;W)+\log 2}{\log\left(\frac{|\mathcal{V}|-N_*(t)}{N^*(t)}\right)}.$$

Since it is assumed that $|\mathcal{V}|-N_*(t)>N^*(t)$, the result now follows by monotonicity of the logarithm. □

The following lemma is included for completeness, it can be seen as continuous version of Theorem 3.7 in [170] which concerns discrete sample spaces.

Lemma 2.23. *Consider random variables $V,W,\hat{V}$ forming a Markov chain $V \to W \to \hat{V}$ taking values in a Radon space. Suppose that $\mathbb{P}^{W|V=v} \ll \mathbb{P}^W$ and that the random variables*

$$\frac{d\mathbb{P}^{W|V=v}}{d\mathbb{P}^W}(W)$$

are $\sqrt{\gamma/2}$-sub-Gaussian for $0<\gamma<1$, $\mathbb{P}^V$-almost surely. Then, the Markov chain $V \to W \to \hat{V}$ satisfies the γ-strong data-processing inequality,

$$I(V;\hat{V}) \leqslant \gamma I(W;\hat{V}).$$

Proof. By Lemma 2.40 below,

$$I(V;\hat{V}) = D_{\mathrm{KL}}\left(\mathbb{P}^{V|\hat{V}} \| \mathbb{P}^{V}\right) \leqslant \mathbb{E}^{\hat{V}}\mathbb{E}^{(V|\hat{V})}\left(\frac{d\mathbb{P}^{V|\hat{V}}}{d\mathbb{P}^{V}}(V,\hat{V}) - 1\right)^2. \tag{2.76}$$

By subsequently using the Markov chain structure $V \to W \to \hat{V}$ and Bayes rule (using that $V, W, \hat{V}$ possess regular conditional probability distributions),

$$\begin{aligned}\frac{d\mathbb{P}^{V|\hat{V}}}{d\mathbb{P}^{V}}(v,\hat{v}) &= \mathbb{E}^{W|\hat{V}=\hat{v}}\left[\frac{d\mathbb{P}^{V|W}}{d\mathbb{P}^{V}}(v,W)\right] \\ &= \mathbb{E}^{W|\hat{V}=\hat{v}}\left[\frac{d\mathbb{P}^{W|V=v}}{d\mathbb{P}^{W}}(W)\right] \\ &= \mathbb{E}^{W}\left[\frac{d\mathbb{P}^{W|V=v}}{d\mathbb{P}^{W}}(W)\frac{d\mathbb{P}^{W|\hat{V}=\hat{v}}}{d\mathbb{P}^{W}}(W)\right].\end{aligned}$$

Define for $s \in \mathbb{R}$,

$$G_{s,v}(W) = s\left(\frac{d\mathbb{P}^{W|V=v}}{d\mathbb{P}^{W}}(W) - 1\right),\ H_{\hat{v}}(W) = \frac{d\mathbb{P}^{W|\hat{V}=\hat{v}}}{d\mathbb{P}^{W}}(W).$$

By Lemma 2.39, we have that

$$\mathbb{E}GH \leqslant \mathbb{E}H\log H + \log \mathbb{E}e^{G}$$

for any random variables G, H with $\mathbb{E}H = 1$ and $\mathbb{E}e^{G} < \infty$. Therefore, using the sub-Gaussianity of $G_{s,v}(W)$,

$$s\left(\frac{d\mathbb{P}^{V|\hat{V}}}{d\mathbb{P}^{V}}(v,\hat{v}) - 1\right) = \mathbb{E}^{W}\left[G_{s,v}(W)H_{\hat{v}}(W)\right] \leqslant \mathbb{E}^{W}H_{\hat{v}}(W)\log H_{\hat{v}}(W) + \frac{s^2\gamma}{2},$$

for all $s \in \mathbb{R}$. Choosing

$$s = \gamma^{-1}\left(\frac{d\mathbb{P}^{V|\hat{V}}}{d\mathbb{P}^{V}}(v,\hat{v}) - 1\right),$$

we obtain

$$\frac{1}{2}\left(\frac{d\mathbb{P}^{V|\hat{V}}}{d\mathbb{P}^{V}}(v,\hat{v}) - 1\right)^2 \leqslant \gamma\mathbb{E}^{W}H_{\hat{v}}(W)\log H_{\hat{v}}(W).$$

Putting things together, we obtain that

$$\begin{aligned}\mathbb{E}^{\hat{V}}\mathbb{E}^{(V|\hat{V})}\left[\mathbb{E}^{W}\left(\frac{d\mathbb{P}^{V|\hat{V}}}{d\mathbb{P}^{V}}(V,\hat{V}) - 1\right)\right]^2 &\leqslant 2\gamma\mathbb{E}^{\hat{V}}\mathbb{E}^{W}H_{\hat{V}}(W)\log H_{\hat{V}}(W) \\ &= 2\gamma D_{\mathrm{KL}}(\mathbb{P}^{W|\hat{V}}\|\mathbb{P}^{W}) = 2\gamma I(W;\hat{V}).\end{aligned}$$

□

2.5.2 Sub-Gaussianity of likelihoods

The following lemma is the key technical lemma enabling the data processing argument in the mutual information lower bound for testing in Section 2.1.2. First we recall some notations from Section 2.1.2. Let us denote by π the distribution of the random vector ϱR, where $R = (R_1, \dots, R_d)$ has independent Rademacher marginals and $\varrho > 0$ is small (it is taken to be $\varrho = \rho/\sqrt{d}$). We take $V \sim Ber(1/2)$ and set $X|(V=0) \sim N(0, \sigma^2 I_d)$ and $X|(V=1) \sim P_\pi$, where $P_\pi = \int P_f d\pi(f)$ and P_f is a multivariate Gaussian distribution with mean f and σ^2 times the identity variance. Let $\mathbb{P}^X$ and $\mathbb{P}^{X|V}$ denote the corresponding distributions of X and $X|V$.

The lemma below shows that the likelihood ratios $\frac{d\mathbb{P}^{X|V=0}}{d\mathbb{P}^X}(X)$ and $\frac{d\mathbb{P}^{X|V=1}}{d\mathbb{P}^X}(X)$ are sub-Gaussian.

Lemma 2.24. *The likelihood ratios*

$$\frac{d\mathbb{P}^{X|V=0}}{d\mathbb{P}^X}(X) \quad \textit{and} \quad \frac{d\mathbb{P}^{X|V=1}}{d\mathbb{P}^X}(X)$$

are $\sqrt{C\beta}$-sub-Gaussian with

$$\beta = \begin{cases} d\varrho^4/\sigma^4, & \text{if } \sigma^2/\varrho^2 < d/2, \\ 2\varrho^2/\sigma^2, & \text{if } \sigma^2/\varrho^2 \geqslant d/2 \end{cases} \tag{2.77}$$

and $C > 0$ a universal constant.

Proof. Using the notation

$$\mathscr{L}_v(X) := \frac{d\mathbb{P}^{X|V=v}}{d\mathbb{P}^X}(X), \quad v \in \{0,1\},$$

we show below that for all $t \in \mathbb{R}$, for some constant $C > 0$,

$$\mathbb{E}_X e^{t(\mathscr{L}_v(X) - \mathbb{E}_X \mathscr{L}_v(X))} \leqslant e^{C\beta t^2/2}.$$

This is implied by

$$\mathbb{P}^X\left(|\mathscr{L}_v - \mathbb{E}_X \mathscr{L}_v| \geqslant s\right) \leqslant 32 \exp\left(-\frac{s^2}{2\beta}\right) \quad \text{for all } s > 0, \tag{2.78}$$

where the equivalence is well known, but a proof can be found in Lemma 2.35. Since $|\mathscr{L}_v(X) - \mathbb{E}_X \mathscr{L}_v(X)| = |\mathscr{L}_v(X) - 1| \leqslant 1$, it is enough to consider $0 < s < 1$. Since the bound in the display above is vacuous for $\beta > 1/4$, consider $\beta \leqslant 1/4$.

To prove (2.78), let us first introduce the notation $L := \frac{dP_\pi}{dP_0}$, and note that

$$\mathscr{L}_0 = \frac{2}{1+L} \quad \text{and} \quad \mathscr{L}_1 = \frac{2}{1+L^{-1}}.$$

Then for $x \in \{\mathscr{L}_0 - 1 \geqslant s\}$ we have

$$\frac{2}{1+L}(x) = \mathscr{L}_0(x) \geqslant s+1 \quad \text{and} \quad 0 \leqslant \frac{2L}{1+L}(x) = 1 - \frac{1-L}{1+L}(x) \leqslant 1-s,$$

where the last inequality follows from $\mathscr{L}_0 - 1 = \frac{1-L}{1+L}$. Consequently, $L^{-1}(x) \geqslant \frac{s+1}{1-s}$. Similarly, for $x \in \{\mathscr{L}_0 - 1 \leqslant -s\}$,

$$0 \leqslant \frac{2}{1+L}(x) \leqslant 1-s \text{ and } \frac{2L}{1+L}(x) \geqslant 1+s$$

and thus $L(x) \geqslant \frac{s+1}{1-s}$. Combining the above bounds results in for $x \in \{|\mathscr{L}_0 - 1| \geqslant s\}$ that

$$|\log L(x)| \geqslant \log\left(\frac{1+s}{1-s}\right) \geqslant \frac{2s}{1+s} \geqslant s,$$

where the last two inequalities follow from $\log x \geqslant 1 - \frac{1}{x}$ and $0 < s < 1$.

Through the same computation, the above display is also true for $x \in \{|\mathscr{L}_1 - 1| \geqslant s\}$. Consequently, for $v = 0, 1$,

$$\begin{aligned}\mathbb{P}^X\left(|\mathscr{L}_v - \mathbb{E}\mathscr{L}_v| \geqslant s\right) &\leqslant \mathbb{P}^X\left(|\log L| \geqslant s\right) \\ &= \frac{1}{2}P_0\left(|\log(L)| \geqslant s\right) + \frac{1}{2}P_\pi\left(|\log(L)| \geqslant s\right).\end{aligned}$$

Using Markov's inequality the terms on the right-hand side can be further bounded as

$$\begin{aligned}&P_0\left(|\log(L)| \geqslant s\right) \leqslant e^{-\nu s}(\mathbb{E}^{X|V=0}L^\nu + \mathbb{E}^{X|V=0}L^{-\nu}) && \nu > 0 \text{ and} \\ &P_\pi\left(|\log(L)| \geqslant s\right) \leqslant e^{-\lambda_1 s}\mathbb{E}^{X|V=1}L^{\lambda_1} + e^{-\lambda_2 s}\mathbb{E}^{X|V=1}L^{-\lambda_2} && \text{for } \lambda_1, \lambda_2 > 0.\end{aligned}$$

Noting that $\mathbb{E}^{X|V=1}L^\lambda = \mathbb{E}^{X|V=0}L^{\lambda+1}$, we obtain that

$$\begin{aligned}\mathbb{P}^X\left(|\mathscr{L}_v - \mathbb{E}\mathscr{L}_v| \geqslant s\right) \leqslant{}& \frac{1}{2}e^{-\nu s}(\mathbb{E}^{X|V=0}L^\nu + \mathbb{E}^{X|V=0}L^{-\nu}) \\ &+ \frac{1}{2}e^{-\lambda_1 s}\mathbb{E}^{X|V=0}L^{\lambda_1+1} + \frac{1}{2}e^{-\lambda_2 s}\mathbb{E}^{X|V=0}L^{-(\lambda_2-1)}.\end{aligned}$$

We proceed by bounding the expectations in the above display after which minimizing in ν gives us the result of the lemma. Recall that $X|(V=0) \sim \mathcal{N}(0, \sigma^2 I_d)$ and $X_i|(V=1) \overset{i.i.d.}{\sim} \frac{1}{2}\mathcal{N}(\varrho, \sigma^2) + \frac{1}{2}\mathcal{N}(-\varrho, \sigma^2)$, $i = 1, \ldots, d$. Consequently,

$$\begin{aligned}L(X) &= \prod_{i=1}^d \left[\frac{\exp\left(-\frac{1}{2\sigma^2}(X_i - \varrho)^2\right) + \exp\left(-\frac{1}{2\sigma^2}(X_i + \varrho)^2\right)}{2\exp(-\frac{1}{2\sigma^2}X_i^2)}\right] \\ &= \prod_{i=1}^d \exp(-\frac{1}{2}\varrho^2/\sigma^2)\cosh(X_i\varrho/\sigma^2). \qquad (2.79)\end{aligned}$$

Then by independence of X_i, $i = 1, \dots, d$

$$\mathbb{E}_{X|V=0} L^{\nu} = \left(e^{-\frac{\nu}{2}\varrho^2/\sigma^2} \mathbb{E} \cosh^{\nu} \left(\frac{\varrho}{\sigma} Z \right) \right)^d,$$

where $Z \sim N(0,1)$.

In view of Lemma 2.26, whenever $|\nu|\varrho^2/\sigma^2 < 1/2$,

$$e^{-s\nu}(\mathbb{E}_0 L^{\nu} + \mathbb{E}_0 L^{-\nu}) \leqslant \exp\left(\nu^2 \frac{d\varrho^4}{2\sigma^4} - s\nu \right) \left(1 + e^{(3/2)\nu d\varrho^4/\sigma^4} \right).$$

By the same lemma, if $|\lambda_1 + 1|\varrho^2/\sigma^2 < 1/2$,

$$\mathbb{E}_{X|V=0} L^{\lambda_1+1} \leqslant \exp\left(3/8 + \lambda_1^2 \frac{d\varrho^4}{2\sigma^4} + 2\lambda_1 \frac{d\varrho^4}{2\sigma^4} \right),$$

where it is used that $\beta \leqslant 1/4$. Similarly,

$$\mathbb{E}_0 L^{-\lambda_2+1} \leqslant \exp\left(3/8 + \lambda_2^2 \frac{d\varrho^4}{2\sigma^4} + \frac{7}{2}\lambda_2 \frac{d\varrho^4}{2\sigma^4} \right).$$

Next we distinguish two cases. Suppose first that $2/d \leqslant \varrho^2/\sigma^2$. Let us take $\nu = \lambda_1 = \lambda_2 = s\sigma^4/(d\varrho^4)$. Then $\nu\varrho^2/\sigma^2 < 1/2$, as $0 < s < 1$, which in turn gives that

$$e^{-s\nu}(\mathbb{E}_0 L^{\nu} + \mathbb{E}_0 L^{-\nu}) \leqslant \exp\left(-\frac{s^2}{2} \frac{\sigma^4}{d\varrho^4} \right) \left(1 + e^{(3/2)s} \right).$$

Similarly,

$$e^{-s\lambda_1} \mathbb{E}_0 L^{\lambda_1+1} \leqslant \exp\left(3/8 - \frac{s^2}{2} \frac{\sigma^4}{d\varrho^4} + \frac{7}{2}s \right),$$

and

$$e^{-s\lambda_2} \mathbb{E}_0 L^{-\lambda_2+1} \leqslant \exp\left(3/8 - \frac{s^2}{2} \frac{\sigma^4}{d\varrho^4} + 2s \right).$$

The remaining case is when $2/d > \varrho^2/\sigma^2$. Choosing $\nu = s\sigma^2/(2\varrho^2)$ results in $\nu\varrho^2/\sigma^2 < 1/2$, which in turn implies

$$\begin{aligned} e^{-s\nu}(\mathbb{E}_0 L^{\nu} + \mathbb{E}_0 L^{-\nu}) &\leqslant \exp\left(\nu^2 \frac{d\varrho^4}{2\sigma^4} - s\nu \right) \left(1 + e^{(3/2)\nu d\varrho^4/\sigma^4} \right) \\ &\leqslant \exp\left(-\frac{s^2}{2} \frac{\sigma^2}{2\varrho^2} \right) \left(1 + e^{(3/2)s} \right). \end{aligned}$$

The bounds on $e^{-s\lambda_1} \mathbb{E}_{X|V=0} L^{\lambda_1+1}$ and $e^{-s\lambda_2} \mathbb{E}_{X|V=0} L^{-\lambda_2+1}$ follow similarly. Hence, by combining the above bounds and noting that for $0 < s < 1$ we have

$$\frac{1}{2}\left(1 + e^{(3/2)s} + e^{(3/8)+(7/2)s} + e^{(3/8)+2s} \right) \leqslant 32,$$

we arrive at (2.78), for β given in (2.77), concluding the proof of the lemma. □

The following has been established in previous literature (see e.g. [39] or [46]) and proves useful for obtaining estimation rates in distributed setting through mutual information based data processing.

Lemma 2.25. *Let R denote a Rademacher random variable, let for $\sigma > 0$, $X|R \sim N(\varrho R, \sigma^2)$ distributed. Then,*

$$\frac{d\mathbb{P}^{X|R=r}}{d\mathbb{P}^X}$$

is $\sqrt{C}\varrho/\sigma$-sub-Gaussian for $r \in \{-1, 1\}$ and a universal constant $C > 0$.

Proof. This follows by Lemma 2.24 (taking $d = 1$), applying it to $X' = X + \varrho$ which follows $X'|R \sim N(2\varrho R, \sigma^2)$. □

Lemma 2.26. *Let $Z \sim N(0, 1)$ and let $\nu \in \mathbb{R}$ such that $|\nu|\varrho^2/\sigma^2 < 1/2$. It holds that*

$$\mathbb{E}\cosh^\nu\left(\frac{\varrho}{\sigma}Z\right) \leqslant \exp\left(\nu\frac{\varrho^2}{2\sigma^2} + \nu^2\frac{3\varrho^4}{2\sigma^4} - \mathbb{1}_{\{\nu<0\}}\frac{3}{2}\nu\frac{\varrho^4}{\sigma^4}\right). \tag{2.80}$$

Proof. First assume that $\nu \geqslant 0$. Using $\cosh(x) \leqslant e^{x^2/2}$ we find

$$\mathbb{E}\cosh^\nu\left(\frac{\varrho}{\sigma}Z\right) \leqslant \mathbb{E}e^{\nu\frac{\varrho^2}{2\sigma^2}Z^2}.$$

In view of Lemma 2.36,

$$\mathbb{E}e^{\lambda(Z^2-1)} \leqslant e^{2\lambda^2} \text{ for all } 0 \leqslant \lambda \leqslant 1/4.$$

Applying this to the second last display yields (2.80).

Consider now the case that $\nu < 0$. We have

$$\begin{aligned}
\frac{d}{dx}\cosh^\nu\left(\frac{\varrho}{\sigma}x\right) &= \nu\frac{\varrho}{\sigma}\cosh^\nu\left(\frac{\varrho}{\sigma}x\right)\tanh\left(\frac{\varrho}{\sigma}x\right),\\
\frac{d^2}{dx^2}\cosh^\nu\left(\frac{\varrho}{\sigma}x\right) &= \nu\frac{\varrho^2}{\sigma^2}\cosh^\nu\left(\frac{\varrho}{\sigma}x\right)\left[(\nu-1)\tanh^2\left(\frac{\varrho}{\sigma}x\right)+1\right] =: \tau(x)
\end{aligned}$$

Since $\cosh(0) = 1$ and $\tanh(0) = 0$, a second order Taylor expansion of $x \mapsto \cosh^\nu\left(\frac{\varrho}{\sigma}x\right)$ about 0 yields

$$\mathbb{E}\cosh^\nu\left(\frac{\varrho}{\sigma}Z\right) = \mathbb{E}\left[1 + \frac{Z^2}{2!}\tau(r_Z Z)\right], \text{ for some } \quad r_Z \in [0, 1].$$

Since $\tanh^2(x) \leqslant x^2$ and $\cosh(x) \geqslant 1$ for all $x \in \mathbb{R}$,

$$\mathbb{E}\frac{Z^2}{2!}\tau(r_Z Z) \leqslant \nu\frac{\varrho^2}{2\sigma^2}\left[(\nu-1)\frac{\varrho^2}{\sigma^2}\mathbb{E}r_Z^2 Z^4 + 1\right] \leqslant \nu\frac{\varrho^2}{2\sigma^2}\left[(\nu-1)\frac{3\varrho^2}{\sigma^2} + 1\right].$$

Then by combining the above two displays

$$\mathbb{E}\cosh^{\nu}\left(\frac{\varrho}{\sigma}Z\right) \leqslant \exp\left(\nu\frac{\varrho^2}{2\sigma^2} + \nu^2\frac{3\varrho^4}{2\sigma^4} - \frac{3}{2}\nu\frac{\varrho^4}{\sigma^4}\right),$$

which concludes the proof of the lemma. □

2.5.3 Auxiliary lemmas for Section 2.2

For the following lemmas, assume the setting of Section 2.2.

Lemma 2.27. *Consider a sample space $\mathcal{X}$ and a distributed protocol with Markov kernels $K^j : \mathscr{Y}^{(j)} \times \mathcal{X} \times \mathcal{U} \to [0,1]$ for $j = 1, \ldots, m$ and shared randomness distribution $\mathbb{P}^U$. Writing $\mathscr{Y} = \bigotimes_{j=1}^m \mathscr{Y}^{(j)}$ for the product sigma-algebra, consider $K = \bigotimes_{j=1}^m K^j : \mathscr{Y} \times \mathcal{X}^m \times \mathcal{U}^m \to [0,1]$. It holds that*

$$K(\cdot|x,u) \ll \mathbb{P}_f^{Y|U=u}(\cdot), \;\; \mathbb{P}_f^{(X,U)} - \textit{almost surely,}$$

for all $f \in \mathbb{R}^d$.

Proof. Let $A \in \mathscr{Y}^{(j)}$. We have that

$$B := \left\{(x,u) \,:\, K^j(A|x,u) > 0\right\} = \bigcup_{l\in\mathbb{N}} \left\{(x,u) \,:\, K^j(A|x,u) > \frac{1}{l}\right\},$$

so if $\mathbb{P}_f^{(X,U)}(B) > 0$ for some $L \in \mathbb{N}$ it holds that $\mathbb{P}_f^{(X,U)}\left((x,u) \,:\, K^j(A|x,u) > \frac{1}{L}\right) > 0$. Since K^j is nonnegative, by Markov's inequality,

$$\begin{aligned}
\mathbb{P}_f^{Y^{(j)}|U=u}(A) &= \int K^j(A|x,u) d\mathbb{P}_f^{X^{(j)}} \times \mathbb{P}^U(x,u) \\
&\geqslant \int_B K^j(A|x,u) d\mathbb{P}_f^{X^{(j)}} \times \mathbb{P}^U(x,u) \\
&\geqslant \frac{1}{L}\mathbb{P}_f^{(X,U)}\left((x,u) \,:\, K^j(A|x,u) > \frac{1}{L}\right) > 0.
\end{aligned}$$

Since given U, $Y^{(1)}, \ldots, Y^{(m)}$, the statement for K follows as $K(\cdot|x^1, \ldots, x^m, u) := \bigotimes_{j=1}^m K^j(\cdot|x^j, u)$ and $\mathbb{P}_f^{Y|U=u} = \bigotimes_{j=1}^m \mathbb{P}_f^{Y^{(j)}|U=u}$. □

Lemma 2.28. *[Distributed Le Cam / chi-square divergence bound] Let $\mathcal{T}$ be a set consisting of distributed testing protocols. It holds that*

$$\inf_{T\in\mathcal{T}}\left(\mathbb{P}_0^Y T + \sup_{f\in H_\rho} \mathbb{P}_f^Y(1-T)\right) \geqslant \inf\left(\sup_{\pi}(1 - \|\mathbb{P}_0^Y - \mathbb{P}_\pi^Y\|_{\mathrm{TV}}) - \pi(H_\rho^c)\right), \quad (2.81)$$

where the supremum on the right-hand side is over all probability distributions π on $\mathbb{R}^d$ with $\mathbb{P}_\pi^Y := \int \mathbb{P}_f^Y d\pi(f)$ and the infimum on the right-hand side is over all Markov

kernels corresponding to a distributed testing protocol in $\mathcal{T}$. Furthermore, (2.81) *is further lower bounded by*

$$1 - \sup_{T\in\mathcal{T}} \inf_{\pi} \left(\sqrt{(1/2)\int \mathbb{E}_0^{Y|U=u} \left(L_\pi^{Y|U=u}(Y) - 1 \right)^2 d\mathbb{P}^U(u)} + \pi(H_\rho^c) \right),$$

the infimum on the right-hand side is over all probability distributions π on $\mathbb{R}^d$ and

$$L_\pi^{Y|U=u}(Y) = \frac{d\mathbb{P}_\pi^{Y|U=u}}{d\mathbb{P}_0^{Y|U=u}}(Y).$$

Proof. It trivially holds that for any distributed protocol $T' \equiv \{T', \{K^j\}_{j=1}^m, (\mathcal{U}, \mathscr{U}, \mathbb{P}^U)\} \in \mathscr{T}$ that

$$\left(\mathbb{P}_0^Y T'(Y) + \sup_{f\in H_\rho} \mathbb{E}_f^Y (1 - T'(Y)) \right) \geqslant \inf_{T\in\mathcal{T}} \left(\mathbb{P}_0^Y T(Y) + \sup_{f\in H_\rho} \mathbb{P}_f^Y (1 - T(Y)) \right).$$

Furthermore, for any prior distribution π on $\mathbb{R}^d$ it holds that

$$\begin{aligned} \sup_{f\in H_\rho} \mathbb{P}_f^Y (1 - T(Y)) &\geqslant \int_{\{f\in H_\rho\}} \mathbb{P}_f^Y (1 - T(Y)) d\pi(f) \\ &\geqslant \int \mathbb{P}_f^Y (1 - T(Y)) d\pi(f) - \pi(H_\rho^c). \end{aligned} \tag{2.82}$$

Hence the right-hand side of the second last display is further bounded from below by

$$\inf_T \left(\mathbb{P}_0^Y T(Y) + \mathbb{P}_\pi^Y (1 - T(Y)) - \pi(H_\rho^c) \right)$$

for all prior distributions π on $\mathbb{R}^d$. For any T, write $A_T = T^{-1}(\{0\})$ and note that

$$\mathbb{P}_0^Y T(Y) + \mathbb{P}_\pi^Y (1 - T(Y)) = 1 - \left(\mathbb{P}_0^Y (Y \in A_T) - \mathbb{P}_\pi^Y (Y \in A_T) \right).$$

By combining the above two displays we get that

$$\inf_{T\in\mathcal{T}} \left(\mathbb{P}_0^Y T(Y) + \sup_{f\in H_\rho} \mathbb{P}_f^Y (1 - T(Y)) \right) \geqslant 1 - \sup_A |\mathbb{P}_0^Y (A) - \mathbb{P}_\pi^Y (A)| - \pi(H_\rho^c).$$

Since the above is true for any distribution π on $\mathbb{R}^d$, the statement is true after taking the supremum over π also. This proves the first statement of the lemma.

Using that the measure $d\mathbb{P}_f^Y$ disintegrates as $d\mathbb{P}_f^{Y|U=u} d\mathbb{P}_f^U(u)$, and the fact that U is independent of the prior π, we find by Jensen's inequality that

$$\|\mathbb{P}_0^Y - \mathbb{P}_\pi^Y\|_{\mathrm{TV}} \leqslant \int \|\mathbb{P}_0^{Y|U=u} - \mathbb{P}_\pi^{Y|U=u}\|_{\mathrm{TV}} d\mathbb{P}^U(u).$$

Combining the first statement of the lemma with Pinsker's second inequality and the above inequality gives

$$\inf_{T\in\mathcal{T}}\mathcal{R}(H_\rho,T)\geqslant 1-\sup_{T\in\mathcal{T}}\inf_{\pi}\left(\int\sqrt{(1/2)D_{\mathrm{KL}}\left(\mathbb{P}_0^{Y|U=u};\mathbb{P}_\pi^{Y|U=u}\right)}d\mathbb{P}^U(u)+\pi(H_\rho^c)\right).$$

By applying Jensen's inequality once more and using Lemma 2.40, we can further bound the above display from below by

$$1-\sup_{T\in\mathcal{T}}\inf_{\pi}\left(\int\sqrt{(1/2)\int\mathbb{E}_0^{Y|U=u}\left(L_\pi^{Y|U=u}(Y)-1\right)^2d\mathbb{P}^U(u)}d\mathbb{P}^U(u)+\pi(H_\rho^c)\right),$$

where

$$D_{\chi^2}(\mathbb{P}_{0,K}^{Y|U=u};\mathbb{P}_{\pi,K}^{Y|U=u})=\mathbb{E}_{0,K}^{Y|U=u}\left(\frac{d\mathbb{P}_\pi^{Y|U=u}}{d\mathbb{P}_0^{Y|U=u}}(Y)\right)^2-1.$$

□

2.5.4 Auxiliary lemmas for Section 2.3

Let Ξ_u^j denote the matrix

$$\Xi_u^j=\mathbb{E}_0^{Y^{(j)}}\mathbb{E}_0^{Y^{(j)}|U=u}\left[\sum_{i=1}^n X_i^{(j)}\middle|Y^{(j)},U=u\right]\mathbb{E}_0^{Y^{(j)}|U=u}\left[\sum_{i=1}^n X_i^{(j)}\middle|Y^{(j)},U=u\right]^\top,$$

as in Section 2.2, (2.24). The following lemma is a (strong) data processing inequality; the covariance matrix of $X|Y$ is dominated by the covariance of the original process X, strongly so for the trace of the matrix if $b/d=o(1)$.

Lemma 2.29. *It holds that* $\Xi_u^j\leqslant nI_d$ *and*

$$Tr(\Xi_u^j)\leqslant 2\log(2)n(\log_2|\mathcal{Y}^{(j)}|).$$

In particular, for $\log_2|\mathcal{Y}^{(j)}|\leqslant b$,

$$Tr(\Xi_u^j)\leqslant\left(2\log(2)\frac{b}{d}\bigwedge 1\right)nd.$$

Proof. Let $v\in\mathbb{R}^d$, then

$$\begin{aligned}v^\top\Xi_u^jv&=\mathbb{E}_0^{Y^{(j)}}\mathbb{E}_0^{Y|U=u}\left[v^\top\sum_{i=1}^n X_i^{(j)}\middle|Y^{(j)},U=u\right]\mathbb{E}_0^{Y|U=u}\left[\left(\sum_{i=1}^n X_i^{(j)}\right)^\top v\middle|Y^{(j)},U=u\right]\\&=\mathbb{E}_0^{Y^{(j)}}\mathbb{E}_0^{Y|U=u}\left[v^\top\left(\sum_{i=1}^n X_i^{(j)}\right)\middle|Y^{(j)},U=u\right]^2.\end{aligned}$$

Since the conditional expectation contracts the L_2-norm, we obtain that the latter is bounded by

$$\mathbb{E}_0 v^\top \left(\sum_{i=1}^n X_i^{(j)}\right)\left(\sum_{i=1}^n X_i^{(j)}\right)^\top v = n\|v\|_2^2,$$

which completes the proof of the statement "$\Xi_u^j \leqslant nI_d$".

The second and third statement of the lemma, we start by noting that under $\mathbb{P}_0$, $\sum_{i=1}^n X_i^{(j)}$ follows a $N(0, nI_d)$ distribution. For any unit vector $v \in \mathbb{R}^d$ and $s \in \mathbb{R}$ this means that

$$\mathbb{E}_0 e^{s\langle \sum_{i=1}^n X_i^{(j)}, v\rangle} \leqslant e^{\frac{s^2 n}{2}}.$$

Furthermore, for arbitrary $y \in \mathcal{Y}$,

$$\begin{aligned}\sum_y \mathbb{P}^{Y^{(j)}|U=u}(y)\mathbb{E}_0\left[e^{s\langle \sum_{i=1}^n X_i^{(j)}, v\rangle}\middle|Y^{(j)} = y, U = u\right] \geqslant \\ \mathbb{P}^{Y^{(j)}|U=u}(y)\mathbb{E}_0\left[e^{s\langle \sum_{i=1}^n X_i^{(j)}, v\rangle}\middle|Y^{(j)} = y, U = u\right] \geqslant \\ \mathbb{P}^{Y^{(j)}|U=u}(y) e^{s\mathbb{E}_0\left[\langle \sum_{i=1}^n X_i^{(j)}, v\rangle\middle|Y^{(j)}=y, U=u\right]},\end{aligned}$$

where the last line follows by Jensen's inequality. By combining the above displays we obtain that

$$s\mathbb{E}_0\left[\langle \sum_{i=1}^n X_i^{(j)}, v\rangle\middle|Y^{(j)} = y, U = u\right] \leqslant \frac{s^2 n}{2} - \log \mathbb{P}^{Y^{(j)}|U=u}(y)$$

for all $s \in \mathbb{R}$. Choosing $s = n\mathbb{E}_0\left[\langle \sum_{i=1}^n X_i^{(j)}, v\rangle\middle|Y^{(j)} = y, U = u\right]$, we have for any unit vector $v \in \mathbb{R}^d$,

$$\mathbb{E}_0\left[\langle \sum_{i=1}^n X_i^{(j)}, v\rangle\middle|Y^{(j)} = y, U = u\right]^2 \leqslant -2n \log \mathbb{P}^{Y^{(j)}|U=u}(y).$$

Next, define for $y \in \mathcal{Y}^{(j)}$,

$$w_{1,y} = \frac{1}{\|\mathbb{E}_0(\sum_{i=1}^n X_i^{(j)}|Y^{(j)} = y, U = u)\|_2}\mathbb{E}_0\left[\sum_{i=1}^n X_i^{(j)}|Y^{(j)} = y, U = u\right]. \quad (2.83)$$

Choose now $w_{2,y}, \ldots, w_{d,y}$ such that together with $w_{1,y}$ the vectors form an orthonormal basis for $\mathbb{R}^d$. We then have

$$\begin{aligned}\mathrm{Tr}(\Xi_u^j) &= \sum_{y\in\mathcal{Y}^{(j)}} \mathbb{P}^{Y^{(j)}|U=u}(y)\sum_{i=1}^{d}\mathbb{E}_0\left[\langle w_{i,y}, \sum_{i=1}^{n} X_i^{(j)}\rangle|Y^{(j)}=y, U=u\right]^2 \\ &= \sum_{y\in\mathcal{Y}^{(j)}} \mathbb{P}^{Y^{(j)}|U=u}(y)\mathbb{E}_0\left[\langle w_{1,y}, \sum_{i=1}^{n} X_i^{(j)}\rangle|Y^{(j)}=y, U=u\right]^2 \\ &\leqslant -2n \sum_{y\in\mathcal{Y}^{(j)}} \mathbb{P}^{Y^{(j)}|U=u}(y)\log \mathbb{P}^{Y^{(j)}|U=u}(y) \leqslant 2n\log|\mathcal{Y}^{(j)}|,\end{aligned}$$

where the last inequality follows from the fact that the uniform distribution on $\mathcal{Y}^{(j)}$ maximizes the entropy on the left-hand side (see Lemma 2.20). In view of $\Xi_u^j \leqslant nI_d$, $\mathrm{Tr}(\Xi_u^j) \leqslant dn$. Combining the above upper bound with the one for $\log|\mathcal{Y}^{(j)}|$ leads to the final statement of the lemma. □

2.5.5 Auxiliary lemmas for Section 2.4

This section contains some results applying to the setting of Section 2.4, but some also apply to general Markov kernels K^j satisfying (ϵ,δ)-differential privacy constraints. For simplicity of the presentation, we simply assume the setting of the aforementioned section, except for suppressing the dependence on the shared randomness conditioning $U=u$ in certain places, whenever it bares no relevance to the results in this section.

The first lemma shows that, if we can approximate the Markov kernels of a distributed protocol sufficiently in terms of total variation by other Markov kernels, the testing risk corresponding to the distributed protocol can be considered in terms of the former.

Lemma 2.30. *Let $\alpha\in(0,1)$ be given. Let $(T,\{K^j\}_{j=1}^m,\mathbb{P}^U)$ be a distributed protocol for the testing problem (2.1) and suppose that there exist kernels $\{\tilde{K}^j\}_{j=1}^m$ such that for $j=1,\ldots,m$,*

$$\|P_0(K^j(\cdot|X^{(j)},u)-\tilde{K}^j(\cdot|X^{(j)},u))\|_{\mathrm{TV}} \leqslant \frac{\alpha}{2m} \quad \mathbb{P}^U\text{-a.s}$$

and

$$\|P_\pi(K^j(\cdot|X^{(j)},u)-\tilde{K}^j(\cdot|X^{(j)},u))\|_{\mathrm{TV}} \leqslant \frac{\alpha}{2m}, \quad \mathbb{P}^U\text{-a.s}$$

for a collection of distributions π on $\mathbb{R}^d$. Then,

$$\begin{aligned}&\mathbb{P}^U P_0^m K(T(Y)|X,U) + \mathbb{P}^U\int P_f^m K(1-T(Y)|X,U)d\pi(f) \geqslant \\ &\qquad \mathbb{P}^U P_0^m \tilde{K}(T(Y)|X,U) + \mathbb{P}^U\int P_f^m \tilde{K}(1-T(Y)|X,U)d\pi(f) - \alpha,\end{aligned}$$

for the same collection of distributions.

Proof. We omit the dependence of u in the proof, as it is of no consequence to the arguments below. Using standard arguments,

$$P_0^m K(T(Y)=1|X) + \int P_f^m K(T(Y)=0|X)d\pi(f) \geqslant$$
$$P_0^m \tilde{K}(T(Y)=1|X) + \int P_f^m \tilde{K}(T(Y)=0|X)d\pi(f) - \|P_0^m(K(\cdot|X)-\tilde{K}(\cdot|X))\|_{\mathrm{TV}}$$
$$- \|P_\pi^m(K(\cdot|X)-\tilde{K}(\cdot|X))\|_{\mathrm{TV}}.$$

By Lemma 6.8,

$$\|P_\pi^m(K(\cdot|X)-\tilde{K}(\cdot|X))\|_{\mathrm{TV}} \leqslant \sum_{j=1}^m \|P_\pi(K^j(\cdot|X^{(j)})-\tilde{K}^j(\cdot|X^{(j)}))\|_{\mathrm{TV}}.$$

By applying the same lemma to $\|P_0^m(K(\cdot|X)-\tilde{K}(\cdot|X))\|_{\mathrm{TV}}$, combined with what is assumed in this lemma, we obtain the result. □

The next lemma gives a construction that allows for a (ϵ,δ)-DP Markov kernel to be restricted to a set and "rebalanced" in order to result in $(\epsilon,2\delta)$-DP Markov kernel.

Lemma 2.31. *Let K be a Markov kernel from $(\mathcal{X},\mathscr{X})^n$ to $(\mathcal{Y},\mathscr{Y})$ satisfying an (ϵ,δ)-DP constraint (i.e. (1.5)) and define for a $A\in\mathscr{Y}$ and a probability measure μ on $\mathscr{Y}$*

$$\tilde{K}(B|x) := K(B\cap A|x) + K(A^c|x)\mu(B), \text{ for } x\in\mathcal{X},\ B\in\mathscr{Y}.$$

Then, $\tilde{K}$ is a Markov kernel $(\mathcal{X},\mathscr{X})$ to $(\mathcal{Y},\mathscr{Y})$ satisfying an $(\epsilon,2\delta)$-DP constraint.

Proof. First of, $\tilde{K}$ can be seen to be a Markov kernel, as the necessary measurability assumptions hold by construction and

$$\tilde{K}(\mathcal{Y}|x) = K(\mathcal{Y}\cap A|x) + K(A^c|x) = 1,$$

where it is used that μ is a probability measure. Furthermore, for arbitrary B and $x,x'\in\mathcal{X}^n$ such that $d_H(x,x')\leqslant 1$, it holds that

$$\begin{aligned}\tilde{K}(B|x) &\leqslant e^\epsilon K(B\cap A|x') + \delta + e^\epsilon K(A^c|x')\mu(B) + \mu(B)\delta\\ &\leqslant e^\epsilon \tilde{K}(B|x') + 2\delta.\end{aligned}$$

□

The following lemma allows approximation of a (ϵ,δ)-DP collection of kernels, which may have unbounded densities, with a $(\epsilon,3\delta)$-DP collection of kernels that have bounded densities. The construction of the approximating kernel is similar to that of Lemma 2.31. The approximation is in terms of total variation distance, which allows the comparison of the testing risks corresponding to both collections of kernels by using Lemma 2.30.

Lemma 2.32. *For any (ϵ, δ)-DP collection of kernels $\{K^j\}_{j=1}^m$, there exists a collection of $(\epsilon, 3\delta)$-DP kernels $\{\tilde{K}^j\}_{j=1}^m$ such that for a fixed constant $C > 0$,*

$$\sup_{x \in \mathbb{R}^{n\times d}} \frac{d\tilde{K}^j(\cdot|x)}{dP_0\tilde{K}^j(\cdot|X^{(j)})}(y) < C, \ P_0\tilde{K}^j(\cdot|X^{(j)})\text{-almost surely,}$$

whilst

$$\|P_f(K^j(\cdot|X^{(j)}) - \tilde{K}^j(\cdot, X^{(j)}))\|_{\mathrm{TV}} \leqslant \frac{\alpha}{2m}.$$

Proof. For any $x \in \mathbb{R}^{n\times d}$ and set $A \in \mathscr{Y}^{(j)}$, we have that

$$K^j(A|x) = \int_A \frac{dK^j(\cdot|x)}{dP_0K^j(\cdot|X^{(j)})}(y) dP_0K^j(y|X^{(j)}) \leqslant 1.$$

So, by Markov's inequality, there exists a set $A_x^M \in \mathscr{Y}^{(j)}$ such that

$$\frac{dK^j(\cdot|x)}{dP_0K^j(\cdot|X^{(j)})}(y) \leqslant M \ \text{ on } A_x^M,$$

whilst

$$K^j\left((A_x^M)^c|x\right) \leqslant 1/M. \tag{2.84}$$

Define for all $x \in \mathbb{R}^{n\times d}$,

$$\tilde{K}^j(B|x) := K^j\left(B \cap A_x^M|x\right) + K^j\left((A_x^M)^c|x\right) \frac{K^j(B \cap A_x^M|x)}{K^j(A_x^M|x)}. \tag{2.85}$$

Then, $\tilde{K}^j$ is $(\epsilon, 3\delta)$-DP whenever $M > 4\delta^{-1}$; for any $x, x' \in (\mathbb{R}^d)^n$ that are Hamming distance 1-apart and $B \in \mathscr{Y}^{(j)}$,

$$\begin{aligned}
\tilde{K}^j(B|x) &\leqslant K^j\left(B|x\right) + K^j\left((A_x^M)^c|x\right) \frac{K^j\left(B \cap A_x^M|x\right)}{K^j\left(A_x^M|x\right)} \\
&= K^j\left(B \cap A_{x'}^M|x\right) + K^j\left(B \cap (A_{x'}^M)^c|x\right) + K^j\left((A_x^M)^c|x\right) \frac{K^j\left(B \cap A_x^M|x\right)}{K^j\left(A_x^M|x\right)} \\
&\leqslant e^\epsilon K^j\left(B \cap A_{x'}^M|x'\right) + e^\epsilon K^j\left(B \cap (A_{x'}^M)^c|x'\right) + 2\delta + \frac{1}{M} \\
&\leqslant e^\epsilon \tilde{K}^j\left(B|x'\right) + (1 + e^\epsilon)M^{-1} + 2\delta,
\end{aligned}$$

where the second to last inequality follows by (2.84) and the last inequality follows by simply adding the nonnegative second term in (2.85). Its Radon-Nikodym derivative satisfies

$$\frac{d\tilde{K}^j(\cdot|x)}{dP_0\tilde{K}^j(\cdot|X^{(j)})}(y) \leqslant 2\mathbb{1}_{A_x^M} \frac{dK^j(\cdot|x)}{dP_0K^j(\cdot|X^{(j)})}(y) \leqslant 2M$$

$P_0\tilde{K}^j(\cdot|X^{(j)})$-almost surely. Furthermore, it holds for any $f \in \mathbb{R}^d$ that

$$\begin{aligned}\|P_f^n(K^j(\cdot|X^{(j)}) - \tilde{K}^j(\cdot, X^{(j)}))\|_{\mathrm{TV}} &\leqslant \int \|K^j(\cdot|x) - \tilde{K}^j(\cdot|x)\|_{\mathrm{TV}} dP_f^n(x) \\ &\leqslant 2\int |K^j\left((A_x^M)^c|x\right)| dP_f^n(x) \leqslant \frac{2}{M}.\end{aligned}$$

Since a choice $M > \delta^{-1} \vee 2m/\alpha$ yields the bound uniformly in $x \in \mathbb{R}^{n\times d}$, the result follows. □

Lemma 2.33 below is very well known, but included for completeness.

Lemma 2.33. *Let K^j be a (ϵ, δ)-DP Markov kernel, $k \in \mathbb{N}$, $0 \leqslant k \leqslant n$, $x_i, \tilde{x}_i \in \mathcal{X}$, $i = 1, \ldots, n$ such that $x_i = \tilde{x}_i$ for all but k $i \in [n]$ and let $x = (x_1, \ldots, x_n)$, $\tilde{x} = (\tilde{x}_1, \ldots, \tilde{x}_n)$. It holds that*

$$K^j(A|x) \leqslant e^{\epsilon k} K^j(A|\tilde{x}) + \delta k e^{\epsilon k}$$

for all measurable A.

Proof. If $k = 1$, the inequality follows by the definition of differential privacy. By applying the definition iteratively,

$$K^j(A|x) \leqslant e^{\epsilon k} K^j(A|\tilde{x}) + \sum_{l=1}^{k} \delta e^{\epsilon(k-l)}.$$

The statement now follows by a trivial inequality for the second term. □

The following lemma translates $(\epsilon, 0)$-differential privacy in the sense of (1.5) to the corresponding densities. In particular, densities corresponding to such kernels are bounded. It is well known and only included for completeness.

Lemma 2.34. *Let $\epsilon \geqslant$, $k \in \mathbb{N}$, $0 \leqslant k \leqslant n$, $x_i, \tilde{x}_i \in \mathcal{X}$, $i = 1, \ldots, n$ such that $x_i = \tilde{x}_i$ for all but k $i \in [n]$ and let $x = (x_1, \ldots, x_n)$, $\tilde{x} = (\tilde{x}_1, \ldots, \tilde{x}_n)$. Suppose that K^j satisfies an $(\epsilon, 0)$-differential privacy constraint in the sense of (1.5) and that it is dominated by some probability measure μ for all $x \in \mathcal{X}$. It holds μ-a.s. that*

$$\frac{dK^j(\cdot|x,u)}{d\mu}(y) \leqslant e^{\epsilon k} \frac{dK^j(\cdot|\tilde{x},u)}{d\mu}(y). \tag{2.86}$$

Furthermore, if $\mu(y) = \int K^j(y|x,u)d\mathbb{P}(x)$ for some probability measure $\mathbb{P}$, it holds μ-a.s. that $\sup_{x\in\mathcal{X}^n} \frac{dK^j(\cdot|x,u)}{d\mu}(y) \leqslant e^{n\epsilon}$.

Proof. Let $x, \tilde{x}$ be at most distance 1-apart in Hamming distance. By the definition of $(\epsilon, 0)$-differential privacy that, for any $A \in \mathscr{Y}$, we have that

$$\int_A \frac{dK^j(\cdot|x,u)}{d\mu}(y)d\mu(y) = \int_A dK^j(y|x,u) \\ \leqslant e^\epsilon \int_A dK^j(y|\tilde{x},u) = e^\epsilon \int_A \frac{dK^j(\cdot|\tilde{x},u)}{d\mu}(y)d\mu(y).$$

Applying this step k-times leads to the first statement of the lemma. For the second statement, $k = n$ yields

$$\frac{dK^j(\cdot|x,u)}{d\mu}(y) = \int \frac{dK^j(\cdot|x,u)}{d\mu}(y)d\mathbb{P}(\tilde{x}) \leqslant e^{\epsilon n}, \ \mu - \text{a.s.}$$

□

2.5.6 Distributed estimation under privacy constraints

The data processing results for differential privacy derived earlier in the chapter, yield (up to logarithmic factors) optimal rates for estimation in the distributed setting under differential privacy constraints as well. In particular, the data processing bound for the trace of the Fisher information of the distributed protocol derived in Lemma 2.15. We describe the resulting rates and provide a proof here.

The estimation results come in the form of Theorems 2.5 and 2.6. Together, they imply the minimax distributed estimation rate under (ϵ, δ)-differential privacy constraints (see Section 1.2 and Definition 3) is (up to logarithmic factors)

$$\left(\frac{d}{mn} + \frac{d^2}{mn^2\epsilon^2}\right) \bigwedge d, \tag{2.87}$$

whenever $\log(1/\delta) \asymp \log(mn)$.

Theorem 2.5. *Let $M \geqslant 1$ be given. Let $Y = (Y^{(1)}, \ldots, Y^{(m)})$ be generated by a (ϵ, δ)-differential privacy constrained distributed estimation protocol. It holds that*

$$\sup_{f \in \mathbb{R}^d : \|f\|_\infty \leqslant M} \mathbb{E}_f \|\hat{f}(Y) - f\|_2^2 \gtrsim \left(\frac{d}{mn} + \frac{d^2}{mn^2\epsilon^2}\right) \bigwedge d \ \text{ for all } d, n, m \in \mathbb{N}$$

for all $0 < \epsilon < 1$ and $\delta \leqslant \min\left(\frac{n}{d}, \frac{\sqrt{n}}{\sqrt{d}}\right)^p \epsilon^2$ for any constant $p > 1$.

Proof. Consider a differentiable prior for the parameter f with associated prior density π with respect to the Lebesgue measure that is of the form $\pi(f) = \prod_{k=1}^d \pi_k(f_k)$ and let $J(\pi)$ denote the "Fisher information" associated with π,

$$J(\pi) := \sum_{k=1}^d \int \frac{\pi_k'(f_k)^2}{\pi_k(f_k)} df_k.$$

Furthermore, let $I_{Y^{(1)},\dots,Y^{(m)}}(f)$ be the Fisher information of the model at f. A multivariate version of the van Trees inequality due to [105] (Theorem 1), bounds the Bayes-risk corresponding to π as follows;

$$\int_{\|f\|_\infty \leqslant M} \mathbb{E}_f \|\hat{f} - f\|_2^2 \pi(f) df \geqslant \frac{d^2}{\int_{\mathbb{R}^d} \mathrm{Tr}(I_{Y^{(1)},\dots,Y^{(m)}}(f))\pi(f) df + J(\pi)}. \tag{2.88}$$

Taking $\pi_k(t) = \cos^2(\pi t/2)\mathbb{1}\{|t| \leqslant 1\}$ for $k = 1, \dots, d$, $J(\pi)$ equals $d\pi^2$ (see e.g. [204]) and π has support contained in the sup-norm ball of radius $1 \leqslant M$ around zero. The Fisher information of the model is equal to the matrix Ξ, where we recall the notation of Section 2.2;

$$\Xi := \sum_{j=1}^m \Xi^j := \sum_{j=1}^m \mathbb{E}_0^{Y^{(j)}} \mathbb{E}_0 \left[\sum_{i=1}^n X_i^{(j)} \middle| Y^{(j)} \right] \mathbb{E}_0 \left[\sum_{i=1}^n X_i^{(j)} \middle| Y^{(j)} \right]^\top. \tag{2.89}$$

Hence, we have that the L_2-risk is lower bounded as follows

$$\sup_{f \in \mathbb{R}^d : \|f\|_\infty \leqslant M} \mathbb{E}_f \|\hat{f}(Y) - f\|_2^2 \geqslant \frac{d^2}{\mathrm{Tr}(\Xi) + d\pi^2}.$$

By employing the bound of Lemma 2.15 and the standard data processing bound $\mathrm{Tr}(\Xi) \leqslant dmn$, we obtain that

$$\sup_{f \in \mathbb{R}^d} \mathbb{E}_f \|\hat{f}(Y) - f\|_2^2 \geqslant \frac{d^2}{mnd \bigwedge mn^2\varepsilon^2 + d\pi^2},$$

which gives the rate of the theorem. □

The upper bound on the estimation risk $\mathbb{E}_f \|\hat{f}(Y) - f\|_2^2 \lesssim d$ follows from the fact that $\|f\|_\infty \leqslant M$ implies that $\|f\|_2^2 \leqslant dM^2$ and the estimator $\hat{f} \equiv 0$ does not require any sharing of information on the data.

Next, we provide a (ϵ, δ)-differentially private procedure which attains the rate of the previous theorem up to additional logarithmic factors whenever $\epsilon \gtrsim \sqrt{d}/\sqrt{mn^2}$. The resulting estimator can be seen as the average of private means of the m-data sets. Define for $x \in \mathbb{R}$ its *clipping between a and b* as

$$[x]_a^b := \begin{cases} b & \text{if } x > b, \\ x & \text{if } a \leqslant x \leqslant b, \\ a & \text{otherwise.} \end{cases}$$

As transcripts, we let machine j release

$$Y^{(j)} = \frac{1}{n} \sum_{j=1}^n [X_i^{(j)}]_{-\tau}^{\tau} + W^{(j)}, \quad \text{where} \quad W^{(j)} \sim N\left(0, 4\tau^2 d \log\left(\frac{1}{\delta}\right) \frac{1}{n^2\varepsilon^2} I_d\right). \tag{2.90}$$

By Lemma 3.27, this results in a (ϵ, δ)-differentially private distributed protocol, with the central machine being able to compute the estimator

$$\hat{f}(Y) = \frac{1}{m}\sum_{j=1}^{m} Y^{(j)}. \tag{2.91}$$

Theorem 2.6. *Let $M > 0$ be given and let $f \in \mathbb{R}^d$ satisfy $\|f\|_\infty \leqslant M$. Then, the (ϵ, δ)-differentially private distributed protocol generated by (2.90) with estimator as given in (2.91) with $\tau = \sqrt{2\log mn}$ satisfies*

$$\mathbb{E}_f \|\hat{f}(Y) - f\|_2^2 \lesssim \frac{d}{mn} + \frac{d^2}{mn^2\varepsilon^2} \cdot \log(mn) \log\left(\frac{1}{\delta}\right) \tag{2.92}$$

for all $m, n, d \in \mathbb{N}$, $0 < \epsilon \leqslant 1$ such that $\tau \geqslant M$ and $\epsilon \gtrsim \sqrt{d}/\sqrt{mn^2}$.

Proof. Using Cauchy-Schwarz and the inequality $2ab \leqslant a^2 + b^2$, we have that

$$\begin{aligned}\mathbb{E}_f \left\| \frac{1}{m}\sum_{j=1}^{m} Y^{(j)} - f \right\|_2^2 &= \mathbb{E}_f \left\| \frac{1}{m}\sum_{j=1}^{m} \frac{1}{n}\sum_{i=1}^{n} [X_i^{(j)}]_{-\tau}^{\tau} - f + \frac{1}{m}\sum_{i=1}^{m} W^{(j)} \right\|_2^2 \\ &\leqslant 2\mathbb{E}_f \left\| \frac{1}{m}\sum_{j=1}^{m} \frac{1}{n}\sum_{i=1}^{n} [X_i^{(j)}]_{-\tau}^{\tau} - f \right\|_2^2 + 2\mathbb{E} \left\| \frac{1}{m}\sum_{j=1}^{m} W^{(j)} \right\|_2^2.\end{aligned}$$

Next, observe that as $W^{(j)}$'s are centered independent random variables with variance given by (2.90), we have

$$\mathbb{E} \left\| \frac{1}{m}\sum_{j=1}^{m} W^{(j)} \right\|_2^2 = \frac{1}{m^2}\sum_{j=1}^{m} \mathbb{E} \left\| W^{(j)} \right\|_2^2 = 8\frac{d^2}{mn^2\varepsilon^2} \cdot \log(mn) \log\left(\frac{1}{\delta}\right).$$

Furthermore, it holds that

$$\begin{aligned}\mathbb{E}_f \left\| \frac{1}{m}\sum_{j=1}^{m} \frac{1}{n}\sum_{i=1}^{n} [X_i^{(j)}]_{-\tau}^{\tau} - f \right\|_2^2 &= \sum_{k=1}^{d} \mathbb{E}_f \left(\frac{1}{m}\sum_{j=1}^{m} \frac{1}{n}\sum_{i=1}^{n} \left([X_i^{(j)}]_{-\tau}^{\tau} \right) - f_k \right)^2 \\ &= \sum_{k=1}^{d} \mathbb{E} \left(\frac{1}{m}\sum_{j=1}^{m} \frac{1}{n}\sum_{i=1}^{n} [\left(Z_i^{(j)} \right)_k]_{-\tau+f_k}^{\tau-f_k} \right)^2,\end{aligned}$$

where we use that $\tau \geqslant M \geqslant \|f\|_\infty$ for mn large enough. Using that $\mathbb{E}V^2 = \mathrm{Var}(V) + (\mathbb{E}V)^2$ for any random variable V, the above display is further bounded by

$$\sum_{k=1}^{d} \mathrm{Var} \left(\frac{1}{m}\sum_{j=1}^{m} \frac{1}{n}\sum_{j=1}^{n} [\left(Z_i^{(j)} \right)_k]_{-\tau+f_k}^{\tau-f_k} \right) + \sum_{k=1}^{d} \left(\frac{1}{m}\sum_{j=1}^{m} \frac{1}{n}\sum_{j=1}^{n} \mathbb{E}[\left(Z_i^{(j)} \right)_k]_{-\tau+f_k}^{\tau-f_k} \right)^2.$$

By Lemma 3.22,

$$\sum_{k=1}^{d} \mathrm{Var}\left(\frac{1}{m}\sum_{j=1}^{m}\frac{1}{n}\sum_{j=1}^{n}[\left(Z_i^{(j)}\right)_k]_{-\tau+f_k}^{\tau-f_k}\right) \leqslant d\mathrm{Var}\left(\frac{1}{m}\sum_{j=1}^{m}\frac{1}{n}\sum_{j=1}^{n}\left(Z_i^{(j)}\right)_k\right) = \frac{d}{mn}.$$

Furthermore, by a standard tail bound for the normal distribution,

$$\left|\mathbb{E}[\left(Z_i^{(j)}\right)_k]_{-\tau+f_k}^{\tau-f_k}\right| \leqslant 2(\tau+M)\Pr\left(|Z_i^{(j)}| \geqslant \tau - M\right) \leqslant \sqrt{2}\frac{\tau+M}{\sqrt{\pi}\tau mn}.$$

Putting everything together, the result of the theorem follows. □

2.5.7 Folklore

The next lemma gives a well-known sufficient (and necessary) condition for the sub-Gaussian distribution. In the literature we did not find the present, required form of the lemma, hence for completeness we also provide its proof.

Lemma 2.35. *Let X a mean-zero random variable satisfying*

$$\mathbb{P}\left(|X| \geqslant s\right) \leqslant C\exp\left(-\frac{s^2}{2\beta}\right)$$

for some $C \geqslant 2$, $\beta > 0$ and for all $s \in [0,\infty)$. Then,

$$\mathbb{E}e^{sX} \leqslant e^{2\beta C s^2/2}.$$

Proof. For $k \in \mathbb{N}$, we have

$$\mathbb{E}|X|^k = \int_0^\infty \mathbb{P}\left(|X|^k > t\right)dt \leqslant C\int_0^\infty \exp\left(-\frac{t^{2/k}}{2\beta}\right)dt.$$

Changing coordinates to $u = t^{2/k}/(2\beta)$ yields that the right-hand side display equals

$$\frac{C}{2}(2\beta)^{k/2}k\int_0^\infty e^{-u}u^{k/2-1}du = \frac{C}{2}(2\beta)^{k/2}k\Gamma(k/2).$$

By the dominated convergence theorem, $\mathbb{E}X = 0$, and $C \geqslant 2$,

$$\begin{aligned}\mathbb{E}e^{sX} &= 1 + \sum_{k=2}^{\infty}\frac{s^k\mathbb{E}X^k}{k!} \leqslant 1 + \frac{C}{2}\sum_{k=2}^{\infty}\frac{(2\beta s^2)^{k/2}\Gamma(k/2)}{(k-1)!}\\ &\leqslant 1 + \sum_{k=1}^{\infty}\left[\frac{(C\beta s^2)^k\Gamma(k)}{(2k-1)!} + \frac{(C\beta s^2)^{k+1/2}\Gamma(k+1/2)}{(2k)!}\right].\end{aligned}$$

Since $\Gamma(k+1/2) \leqslant \Gamma(k+1) = k\Gamma(k) = k!$ and $(2k)! \geqslant 2^k (k!)^2$, the latter is further bounded by

$$1 + \left(1 + \sqrt{C\beta s^2}\right) \sum_{k=1}^{\infty} \frac{(C\beta s^2/2)^k}{k!} = e^{C\beta s^2/2} + \sqrt{C\beta s^2}(e^{\beta C s^2/2} - 1).$$

Since $(e^x - 1)(e^x - \sqrt{x}) \geqslant 0$, we obtain that

$$\mathbb{E}e^{sX} \leqslant e^{\frac{2C\beta s^2}{2}}.$$

□

The following lemma is a well known result and follows from standard calculus, but we included it as we did not find a stand-alone proof.

Lemma 2.36. *Let Z be $N(0,1)$, $0 \leqslant \lambda \leqslant 1/4$. Then,*

$$\mathbb{E}e^{\lambda(Z^2-1)} \leqslant e^{2\lambda^2}.$$

Proof. Using the change of variables $u = z\sqrt{1-2\lambda}$,

$$\begin{aligned}\mathbb{E}e^{\lambda(Z^2-1)} &= \frac{1}{\sqrt{2\pi}} \int e^{\lambda(z^2-1)} e^{-\frac{1}{2}z^2} dz \\ &= \frac{e^{-\lambda}}{\sqrt{2\pi(1-2\lambda)}} \int e^{-\frac{1}{2}z^2} dz = \frac{e^{-\lambda}}{\sqrt{(1-2\lambda)}}.\end{aligned}$$

The MacLaurin series of $-\frac{1}{2}\log(1-2\lambda)$ reads

$$\frac{1}{2} \sum_{k=1}^{\infty} \frac{(2\lambda)^k}{k},$$

which yields that the second last display equals

$$\exp\left(\frac{3}{2}\lambda^2 + \frac{1}{2}\sum_{k=3}^{\infty} \frac{(2\lambda)^k}{k}\right).$$

If $\lambda \leqslant 1/4$,

$$\sum_{k=3}^{\infty} \frac{(2\lambda)^k}{k} \leqslant \frac{(2\lambda)^3}{1-2\lambda} \leqslant \lambda^2,$$

from which the result follows. □

The following lemmas are straightforward calculations used multiple times in Section 2.4.2.

Lemma 2.37. *Let $S \sim Bin(p,n)$ for $p \in [0,1]$ and let $0 \leqslant \epsilon \leqslant 1$. It holds that*

$$\mathbb{E}Se^{\epsilon S} \leqslant npe^{\epsilon+2\epsilon np}.$$

Proof. Write $S = \sum_{i=1}^n B_i$, with $B_1, \dots, B_n \overset{\text{i.i.d.}}{\sim} \text{Ber}(p)$.

$$\begin{aligned}
\mathbb{E}Se^{\epsilon S} &= \sum_{i=1}^n \mathbb{E}B_i e^{\epsilon S} \\
&= \sum_{i=1}^n pe^{\epsilon}\mathbb{E}e^{\epsilon \sum_{k\neq i} B_k} \\
&= npe^{\epsilon}\left(\mathbb{E}e^{\epsilon B_1}\right)^{n-1} \\
&= npe^{\epsilon}\left(1+p(e^{\epsilon}-1)\right)^{n-1} \\
&\leqslant npe^{\epsilon+2\epsilon np},
\end{aligned}$$

where the inequality follows from the fact that $e^x - 1 \leqslant 2x$ for $0 \leqslant x \leqslant 1$. □

Lemma 2.38. *Let $a \in \mathbb{R}$ and let $Z, Z' \overset{i.i.d.}{\sim} N(0, I_d)$ for $d \in \mathbb{N}$.*
Then, $a\langle Z, Z'\rangle$ is $Ca\sqrt{d}$-sub-exponential for a universal constant $C > 0$ and

$$\mathbb{E}e^{t|a\langle Z,Z'\rangle|} \leqslant 2e^{t^2a^2d},$$

whenever $|t| \leqslant (2a^2)^{-1}$.

Proof. Since $\langle Z, Z'\rangle | Z' \sim N(0, \|Z'\|_2)$,

$$\mathbb{E}e^{ta\langle Z,Z'\rangle} = \mathbb{E}^{Z'}\mathbb{E}^{Z|Z'}e^{ta\langle Z,Z'\rangle} = \mathbb{E}^{Z'}e^{\frac{t^2a^2}{2}\|Z'\|_2^2}.$$

By Lemma 2.36, the latter is further bounded by

$$e^{\frac{t^2a^2d}{2}+\frac{t^4a^4d}{2}} \leqslant e^{t^2a^2d},$$

whenever $t^2a^2 \leqslant 1/2$. The conclusion then follows by e.g. Proposition 2.7.1 in [210], since $\langle Z, Z'\rangle$ is mean zero. For the last statement,

$$\langle Z, Z'\rangle | Z' \overset{d}{=} -\langle Z, Z'\rangle | Z'.$$

Consequently,

$$\begin{aligned}
\mathbb{E}^{Z|Z'}e^{t|a\langle Z,Z'\rangle|} &= \mathbb{E}^{Z|Z'}\mathbb{1}_{\{\langle Z,Z'\rangle>0\}}e^{ta\langle Z,Z'\rangle} + \mathbb{E}^{Z|Z'}\mathbb{1}_{\{\langle Z,Z'\rangle\leqslant 0\}}e^{-ta\langle Z,Z'\rangle} \\
&\leqslant 2\mathbb{E}^{Z|Z'}e^{ta\langle Z,Z'\rangle},
\end{aligned}$$

and the proof follows by what was shown above. □

The following lemma is a well known result, essentially one side of the Donsker-Varadhan duality (see e.g. Theorem 4.13 in [36]).

Lemma 2.39. *Consider a nonnegative random variable H with $\mathbb{E}H = 1$ and a random variable Z satisfying $\log \mathbb{E}e^Z < \infty$. It holds that*

$$\mathbb{E}HZ - \mathbb{E}H \log H \leqslant \log \mathbb{E}e^Z.$$

Proof. We have

$$\mathbb{E}HZ - \mathbb{E}H \log H = \mathbb{E}H \log(\frac{e^Z}{H}).$$

The result now follows by Jensen's inequality, using that $A \mapsto \mathbb{E}\mathbb{1}_A H$ defines a probability measure. □

The next lemma is a standard bound for the KL-divergence, see for instance Lemma 2.7 of [204].

Lemma 2.40. *Let P, Q probability measures on some measure space such that $Q \ll P$. Then,*

$$D_{\mathrm{KL}}(P\|Q) \leqslant \int \left(\frac{dP}{dQ} - 1\right)^2 dQ.$$

Definition 5. Consider probability measures P and Q on a measurable space $(\mathcal{X}, \mathscr{X})$. A *coupling* of P and Q is any probability measure $\mathbb{P}$ on $(\mathcal{X} \times \mathcal{X}, \mathscr{X} \otimes \mathscr{X})$ such that $\mathbb{P}$ has marginals P and Q:

$$P = \mathbb{P} \circ \pi_1^{-1},\ Q = \mathbb{P} \circ \pi_2^{-1}$$

where $\pi_i : \mathcal{X} \times \mathcal{X} \to \mathcal{X}$ is the projection onto the i-th coordinate (i.e. $\pi_i(x_1, x_2) = x_i$ for $i = 1, 2$).

Lemma 2.41 below is a well known result showing that, for random variables X and $\tilde{X}$ defined on a Polish space, a small total variation distance between their corresponding laws guarantees the existence of a coupling such that they are equal with high probability.

Lemma 2.41. *For any two probability measures P and Q on a measurable space $(\mathcal{X}, \mathscr{X})$ with $\mathcal{X}$ a Polish space and $\mathscr{X}$ its Borel sigma-algebra. There exists a coupling $\mathbb{P}^{X,\tilde{X}}$ such that*

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

For a proof, see e.g. Section 8.3 in [201].

Chapter 3

Optimal distributed testing protocols under bandwidth and privacy constraints

"*An algorithm must be seen to be believed.*" - Donald E. Knuth

In this chapter, we exhibit algorithms / methods attaining the lower bounds as described by the impossibility results of the previous chapter. Specifically the exhibited methods are optimal in the sense that they attain the lower bound rates of Theorem 3.1 and Theorem 3.2. Section 3.1 is concerned with constructing methods for the b-bit bandwidth constrained signal detection problem. In Section 3.2, methods are constructed that are optimal under differential privacy constraints.

There are similarities between the flavor of the testing strategies. The most important commonality is the contention between combining the locally optimal tests versus sharing information that allows to "reconstruct" the underlying full data. This second approach is more similar to the optimal approach typically followed in estimation problems. What is also similar, is that the "phase transitions" exhibited in the lower bound theorems typically correspond to different testing strategies (but not in all cases). Another parallel is the importance of shared randomness in cases where the "reconstruction" strategy is followed. The chapter closes with an in-depth discussion of this phenomenon in Section 3.3.

3.1 Testing protocols under bandwidth-constraints

"*There is a whole book of readymade, long and convincing, lavishly composed telegrams for all occasions. Sending such a telegram costs only*

> *twenty-five cents. You see, what gets transmitted over the telegraph is not the text of the telegram, but simply the number under which it is listed in the book, and the signature of the sender. This is quite a funny thing, reminiscent of Drugstore Breakfast #2. Everything is served up in a ready form, and the customer is totally freed from the unpleasant necessity to think, and to spend money on top of it.*" – Ilya Ilf & Yevgeny Petrov

In this section, we exhibit three distributed testing procedures that attain the rates posed by the lower bounds of Theorem 2.3. Together, they yield Theorem 3.1, which shows that the lower bounds in Theorem 2.3 are attainable and therefore tight. The first distributed testing procedure T_{I} communicates only a single bit per machine and can detect signals with a squared Euclidean norm of larger or equal order than $\sqrt{d}/(\sqrt{m}n)$ and does not need a shared randomness. As a second procedure, we consider a test using the shared randomness protocol T_{II} that achieves the rate $\frac{d}{nm\sqrt{b\wedge d}}$. The third procedure is a local randomness protocol and achieves the corresponding slower rate $d/(nm(b\wedge d))$. The existence of such distributed testing protocols proves the theorem below.

Theorem 3.1. *For each $\alpha \in (0,1)$ there exists a constant $C_\alpha > 0$ (depending only on α) such that if*

$$\rho^2 \geqslant C_\alpha \frac{\sqrt{d}}{mn}\left(\sqrt{\frac{d}{b\wedge d}} \bigwedge \sqrt{m}\right),$$

there exists $T \in \mathscr{T}_{SR}^{(b)}$ such that

$$\mathcal{R}(H_\rho, T) \leqslant \alpha \quad \text{for all} \quad n, m, d, b \in \mathbb{N}.$$

Similarly, for

$$\rho^2 \geqslant C_\alpha \frac{\sqrt{d}}{mn}\left(\frac{d}{b\wedge d} \bigwedge \sqrt{m}\right)$$

there exists $T \in \mathscr{T}_{LR}^{(b)}$ such that

$$\mathcal{R}(H_\rho, T) \leqslant \alpha \quad \text{for all} \quad n, m, d, b \in \mathbb{N}.$$

Sections 3.1.1, 3.1.2 and 3.1.3 describe distributed testing protocols that attain the rates in the above theorem. Combining Lemmas 3.2, 3.3 and 3.4, the proof of the theorem follows as an immediate consequence of these lemmas.

A common denominator in the construction of the three protocols is that the transcripts $Y^{(j)}$ are generated as vector of p_f^j-Bernoulli random variables taking values in $\{0,1\}^b$ where $p_f^j \in [0,1]^b$ depends on the underlying signal f, in a manner that ensures that $p_f^j = (1/2, \ldots, 1/2)$ under the null hypothesis (i.e. when $f = 0$). The concentration inequality for groups of Bernoulli random variables given in Lemma 3.1 provides a recipe for the construction of a central test for each of the three regimes.

The Type I error can be controlled since the distribution under the null hypothesis is known. The Type II error is small whenever the vectors of probabilities $p_f^1, \dots, p_f^m$ are sufficiently separated from $(1/2, \dots, 1/2)$ in Euclidean norm.

Below we design a testing procedure for l observations on the binary hypercube $\{0,1\}^k$. That is, given independent observations B_i for $i = 1, \dots, l$ taking values in $\{0,1\}^k$ with probability distribution $p = (p_1, \dots, p_k) \in [0,1]^k$ and $\sum_{i=1}^k p_i = 1$. The test aims to distinguish the hypothesis

$$H_0 : p = \frac{1}{2}\iota_k \text{ versus } H_1 : p \neq \frac{1}{2}\iota_k,$$

where we use $\iota_k = (1, 1, \dots, 1) \in \mathbb{R}^k$. Multiple algorithms exist that achieve this, we propose the following test.

Lemma 3.1. *For $k, l \in \mathbb{N}$, consider for $i = 1, \dots, k$ $j = 1, \dots, l$ independent draws B_i^j taking values in $\{0,1\}$ with mean p_i. The test*

$$T := \mathbb{1}\left\{ \left| \frac{1}{\sqrt{k}l} \sum_{i=1}^{k} \left(\sum_{j=1}^{l} (B_i^j - \frac{1}{2}) \right)^2 - \sqrt{k}/4 \right| \geqslant \kappa_\alpha \right\}$$

has at most level α for the null hypothesis $p_i = 1/2$ for $i = 1 \dots, k$. Furthermore, if for $p = (p_1, \dots, p_k) \in [0,1]^k$ it holds that

$$\eta_{p,k,l} := \frac{l-1}{2\sqrt{k}} \|p - 1/2\iota_k\|_2^2 \geqslant \kappa_\alpha, \tag{3.1}$$

then it also holds that

$$\mathbb{E}(1-T) \leqslant \frac{1/2 + 16k^{-1/2}\eta_{p,k,l}}{\eta_{p,k,l}^2}. \tag{3.2}$$

The lemma gives us a test that distinguishes between "strings" of bits generated by the machines depending on their stochastic behavior under the null hypothesis versus the anticipated behavior under the alternative. Bits under the null hypothesis are "fair coin flips". When they sufficiently deviate from fair flips in the sense that $\|p - 1/2\iota_k\|_2$ is large under the alternative hypothesis, the underlying signal that causes such a deviation can be detected with large power.

The proof of the lemma can be found in Section 3.4.1 of the chapter appendix where it is restated as Lemma 3.16.

3.1.1 Low communication budget: construction of T_{I}

We first compute the local test statistic $S_{\mathrm{I}}^{(j)} = n\|\overline{X^{(j)}}\|_2^2$ at every machine $j = 1, \dots, m$. Under the null hypothesis, $S_{\mathrm{I}}^{(j)}$ follows a chi-square distribution with d degrees of freedom, i.e. $S_{\mathrm{I}}^{(j)} \sim \chi_d^2$. Letting $F_{\chi_d^2}$ denote χ_d^2-cdf, the quantity $F_{\chi_d^2}\left(S_{\mathrm{I}}^{(j)}\right)$

can be seen as the p-value for the local test statistic $S_{\mathrm{I}}^{(j)}$. Based on these "local p-values", we then generate the randomized transcript $Y_{\mathrm{I}}^{(j)}$ for every j using Bernoulli random variables:

$$Y_{\mathrm{I}}^{(j)} | S_{\mathrm{I}}^{(j)} \sim \mathrm{Ber}\left(F_{\chi^2_d}\left(S_{\mathrm{I}}^{(j)}\right)\right).$$

For a given $\alpha \in (0,1)$, we can construct the test

$$T_{\mathrm{I}} = \mathbb{1}\Big\{\Big|\frac{1}{m}\Big(\sum_{j=1}^{m}(Y_{\mathrm{I}}^{(j)} - 1/2)\Big)^2 - 1/4\Big| \geqslant \kappa_\alpha\Big\} \tag{3.3}$$

at the central machine. In applications, one could set for instance κ_α such that $\mathbb{P}_0 T_{\mathrm{I}} \approx \alpha$ by considering that $\sum_{j=1}^{m} Y_I^{(j)}$ is $(m, 1/2)$-binomially distributed under the null. Lemma 3.2 below yields that for each $\alpha \in (0,1)$, there exist constants $\kappa_\alpha, C_\alpha, M_\alpha, D_0 > 0$ such that for $m \geqslant M_\alpha$ and $d \geqslant D_0$ it holds that $\mathcal{R}(H_\rho, T_{\mathrm{I}}) \leqslant \alpha$, whenever $\rho^2 \geqslant C_\alpha \frac{\sqrt{d}}{\sqrt{mn}}$.

The case $m \leqslant M_\alpha$ corresponds essentially to the non-distributed setting and is treated separately for technical reasons. In practice, one would simply use the test given in (3.3) also for $m \leqslant M_\alpha$. Furthermore, if one allows for a slightly larger amount of bits (e.g. $\log_2(mn)$ bits), one could opt to transmit an (approximation of) the test statistics $S_{\mathrm{I}}^{(j)}$ themselves, see e.g. Lemma 2.3 in [188], for which it is easy to prove that the rate of $\frac{\sqrt{d}}{\sqrt{mn}}$ is achieved without requiring any assumptions on m. For the sake completeness: by considering $\rho^2 \geqslant C_\alpha \sqrt{M_\alpha} \frac{\sqrt{d}}{mn}$, we see that the optimal rate of $\frac{\sqrt{d}}{\sqrt{mn}}$ can be achieved in the $m \leqslant M_\alpha$ case by simply taking

$$T'_{\mathrm{I}} := Y_{\mathrm{I}}^{(1)} := \mathbb{1}\left\{\frac{1}{\sqrt{d}}\left(S_{\mathrm{I}}^{(1)} - d\right) \geqslant \kappa_\alpha\right\} \tag{3.4}$$

for an appropriately large choice of the constant κ_α. Similarly, the requirement that d is larger than some constant D_0 (which is independent of α) appears for technical reasons. The case where $d \leqslant D_0$ is covered by the local randomness protocol T_{III} in Section 3.1.3.

Lemma 3.2. *For each $\alpha \in (0,1)$, there exist constants $\kappa_\alpha, C_\alpha, M_\alpha, D_0 > 0$ such that for $m \geqslant M_\alpha$ and $d \geqslant D_0$ it holds that*

$$\mathcal{R}(H_\rho, T_I) \leqslant \alpha,$$

whenever $\rho^2 \geqslant C_\alpha \frac{\sqrt{d}}{\sqrt{mn}}$.

Proof. Under the null hypothesis the random variables $Y_{\mathrm{I}}^{(j)} \sim^{i.i.d.} \mathrm{Bern}(1/2)$. Next we shall apply Lemma 3.1 with $k = 1$, and $l = m$. By the first statement of the lemma, we obtain that there exists $\kappa_\alpha > 0$ such that $\mathbb{P}_0 T_{\mathrm{I}} \leqslant \alpha/2$.

We give an upper bound for the Type II error by using the second statement of the lemma, but before that we show that condition (3.1) holds. Note that the law of total expectation yields

$$\mathbb{E}_f Y_I^{(j)} = \mathbb{E}_f \mathbb{E}_f \left[Y_I^{(j)} | S_{\mathrm{I}}^{(j)} \right] = \mathbb{E}_f F_{\chi^2_d} \left(S_{\mathrm{I}}^{(j)} \right) = \Pr(S_{\mathrm{I}}^{(j)} \geqslant W_d),$$

where $S_{\mathrm{I}}^{(j)}$ is noncentral chi-square distributed under $\mathbb{P}_f$ with d-degrees of freedom and noncentrality parameter $n\|f\|_2^2$ and W_d is an independent chi-square distributed random variable with d-degrees of freedom. Then Lemma 3.19 in the chapter appendix yields that

$$\eta_{p,m,1} = \frac{m-1}{2} \left(\mathbb{E}_f Y_I^{(j)} - \frac{1}{2} \right)^2 \geqslant \frac{m-1}{3200} \left(\frac{n\|f\|_2^2}{\sqrt{d}} \bigwedge \frac{1}{2} \right)^2. \tag{3.5}$$

whenever $d \geqslant D_0$ for some universal constant $D_0 > 0$. Consequently, as $\|f\|_2^2 \geqslant \rho^2 \geqslant C_\alpha \frac{\sqrt{d}}{\sqrt{mn}}$, we obtain that condition (3.1) is satisfied whenever $m \geqslant M_\alpha$ for some large enough $C_\alpha > 0$ and $M_\alpha > 0$. Hence, the Type II error is bounded by the right-hand side of (3.2), which is monotone decreasing in $\eta_{p,m,1}$ hence also in C_α. Therefore, by large enough choice of C_α the Type II error is bounded from above by $\alpha/2$. □

3.1.2 Public coin, high communication budget: construction of T_{II}

We now switch our attention to exhibiting a testing procedure that is optimal when $bm \gtrsim d$ and $b \lesssim d$, in the shared randomness case. The rate to attain in this case is $\rho^2 \gtrsim d/(nm\sqrt{b})$. That a shared source of randomness in distributed settings can be strictly better than private ones in terms of "communication complexity", is an idea that goes back to [222]. Essentially, the use of shared randomness allows for the machines coordinate their efforts in "covering" each of the d dimensions of the data even though all communication happens in just one round. We explore this phenomenon in Section 3.3, giving various explanations on top of the proof of Lemma 3.3 below. We adopt ideas proposed by [12], who consider the setting where $n = 1$ with asymptotics in m. This testing protocol is exhibited below and we provide a full proof covering also the case where $m \neq n$. To that extent, let U be a random rotation, i.e. U is drawn from the Haar measure (see e.g. Theorem F.13 in [24]) on the set of orthonormal matrices in $\mathbb{R}^{d\times d}$. At each machine, for $b \leqslant d$, we can compute the b-bit transcript $Y_{\mathrm{II}}^{(j)} \in \{0,1\}^b$ conditionally on the shared randomness draw U, where each of the $1 \leqslant i \leqslant b$ components is defined through

$$(Y_{\mathrm{II}}^{(j)})_i | U, X^{(j)} = \mathbb{1}\left\{ \left(\sqrt{n} U \overline{X^{(j)}} \right)_i > 0 \right\},$$

where $(v)_i$ denotes the projection onto the i-th coordinate of the vector $v \in \mathbb{R}^d$. The random rotation fulfills a similar purpose as the random reweighting algorithm

proposed in [192], but leads to an easier proof in the d-dimensional case because of rotational invariance of the Gaussian distribution.

Centrally, after transmitting $(Y_{\mathrm{II}}^{(1)}, \dots, Y_{\mathrm{II}}^{(m)})$, we compute the aggregated test statistics $S_{\mathrm{II}} = \sum_{j=1}^{m} Y_{\mathrm{II}}^{(j)}$ and define the corresponding test as

$$T_{\mathrm{II}} = \mathbb{1}\Big\{\Big|\frac{1}{\sqrt{b}m}\sum_{i=1}^{b}\Big((S_{\mathrm{II}})_i - \frac{m}{2}\Big)^2 - \sqrt{b}/4\Big| > \kappa_\alpha\Big\}. \tag{3.6}$$

Lemma 3.3 below shows that this test achieves the shared randomness lower bound when $mb \gtrsim d$ and $m \geqslant M_\alpha$.

Lemma 3.3. *For each $\alpha \in (0,1)$, there exist constants $\kappa_\alpha, C_\alpha, M_\alpha > 0$ such that for $m \geqslant M_\alpha$*

$$\mathcal{R}(H_\rho, T_{II}) \leqslant \alpha,$$

whenever $\rho^2 \geqslant C_\alpha \frac{d}{mn\sqrt{d \wedge b}}$.

Proof of Lemma 3.3. First note that it is sufficient to consider the case $b \leqslant d$ as one can simply take $b = b \wedge d$. Then note that under $\mathbb{P}_f$, $\sqrt{n}U\overline{X^{(j)}}|U \sim N_d(\sqrt{n}Uf, I_d)$ by the rotational invariance of the Gaussian distribution. By linearity of the coordinate projection, conditionally on U,

$$\mathbb{1}\left\{\left(\sqrt{n}U\overline{X^{(j)}}\right)_i > 0\right\} \overset{d}{=} \mathbb{1}\left\{\sqrt{n}(Uf)_i + Z > 0\right\},$$

where $Z \sim N(0,1)$. As a consequence, the vector S_{II} is conditionally on U coordinate wise independent binomially distributed with parameters m and $p_{f,U} \in [0,1]^b$ under $\mathbb{P}_f^{Y|U}$, where

$$(p_{f,U})_i = \Phi(\sqrt{n}(Uf)_i),$$

with Φ the standard normal CDF. Under the null hypothesis, $(S_{\mathrm{II}})_i$ is $\mathrm{Bin}(m, 1/2)$ distributed since $p_{0,U} = (1/2, \dots, 1/2) \in [0,1]^b$. Next we apply Lemma 3.1 with $k = b$ and $l = m$. By the first statement of the lemma, it follows that for κ_α large enough, $\mathbb{P}_0 T_{\mathrm{II}} \leqslant \alpha/2$.

In order apply the second statement of the lemma, which yields that the Type II error is bounded by $\alpha/2$, it suffices to show that the event

$$A = \Big\{\frac{m-1}{2\sqrt{b}}\sum_{i=1}^{b}\Big((p_{f,U})_i - \frac{1}{2}\Big)^2 \geqslant N_\alpha\Big\},$$

where $N_\alpha := \kappa_\alpha \vee \frac{16}{\alpha}$, occurs with $\mathbb{P}^U$-probability greater than $1 - \alpha/4$. Note that for this choice of N_α, (3.1) is satisfied on the event A and the right-hand side of (3.2) is smaller than $\alpha/4$. The Type II error is then bound by $\mathbb{P}_f T_{\mathrm{II}} \mathbb{1}_A + \mathbb{P}_f \mathbb{1}_{A^c} \leqslant \alpha/2$.

We proceed to show that $\mathbb{P}_f \mathbb{1}_{A^c} \leqslant \alpha/4$. By a standard bound on the Gaussian error function $x \mapsto 2\Phi(x) - 1$ (see Lemma 3.26),

$$\left(\Phi(\sqrt{n}(Uf)_i) - \frac{1}{2}\right)^2 \geqslant \frac{1}{12} \min\left\{n(Uf)_i^2, 1\right\},$$

which in turn implies that

$$\mathbb{P}^U\left(\frac{m-1}{2\sqrt{b}} \sum_{i=1}^{b} \left((p_{f,U})_i - \frac{1}{2}\right)^2 \leqslant N_\alpha\right) \leqslant \mathbb{P}^U\left(\frac{m-1}{24\sqrt{b}} \sum_{i=1}^{b} \min\left\{n(Uf)_i^2, 1\right\} \leqslant N_\alpha\right).$$

Note that $Uf \stackrel{d}{=} \|f\|_2 (Z_1, \ldots, Z_d)/\|Z\|_2$, where $Z = (Z_1, \ldots, Z_d) \sim N(0, I_d)$ (see e.g. Section 3.4 of [210]). Using that $\|f\|_2 \geqslant \rho$ and $\rho^2 \geqslant C_\alpha \frac{d}{mn\sqrt{b}}$, the previous display is further bounded by

$$\Pr\left(\frac{m-1}{24\sqrt{b}} \sum_{i=1}^{b} \min\left\{C_\alpha \frac{dZ_i^2}{m\sqrt{b}\|Z\|_2^2}, 1\right\} \leqslant N_\alpha\right).$$

Considering the intersection with the event $\{\|Z\|_2^2 \leqslant kd\}$ for some $k > 0$, the above display can be bounded by

$$\Pr\left(\sum_{i=1}^{b} \min\{Z_i^2, C_\alpha^{-1} m\sqrt{b} k\} \leqslant \frac{24bmk}{C_\alpha(m-1)} N_\alpha\right) + \Pr\left(\|Z\|_2^2 \geqslant kd\right).$$

For k large enough (independent of d), the second term is less than $\alpha/8$. By Lemma 3.27,

$$\Pr\left(\max_{1 \leqslant i \leqslant b} Z_i^2 \geqslant C_\alpha^{-1} m\sqrt{b} k\right) \leqslant \frac{2b}{e^{C_\alpha^{-1} m\sqrt{b} k/4}}.$$

For large enough $M_\alpha \geqslant C_\alpha$, the condition $m \geqslant M_\alpha$ implies that the right-hand side is less than $\alpha/8$. The first term in the second to last display is consequently bounded by

$$\begin{aligned}\Pr\Big(\sum_{i=1}^{b} Z_i^2 \leqslant \frac{24bmk}{C_\alpha(m-1)} N_\alpha\Big) + \Pr\left(\max_{1 \leqslant i \leqslant b} Z_i^2 \geqslant C_\alpha^{-1} m\sqrt{b} k\right) \\ \leqslant \Pr\left(\sum_{i=1}^{b} Z_i^2 \leqslant \frac{24bmk}{C_\alpha(m-1)} N_\alpha\right) + \alpha/8.\end{aligned}$$

For $m \geqslant M_\alpha \geqslant 25$ and by choosing C_α large enough such that the Chernoff–Hoeffding bound on the left tail of the chi-square distribution (see Lemma 3.28) can be applied to the first term of the preceding display we get that

$$\Pr\left(\sum_{i=1}^{b} Z_i^2 \leqslant \frac{25kN_\alpha}{C_\alpha} b\right) \leqslant \exp\left(-b \frac{\frac{25kN_\alpha}{C_\alpha} - 1 - \log\left(\frac{25kN_\alpha}{C_\alpha}\right)}{2}\right) \leqslant \alpha/8, \tag{3.7}$$

finishing the proof of the lemma. □

3.1.3 Private coin, high total communication budget: constructing T_{III}

Finally, we consider the case of not having access to shared randomness, but having a relatively large communication budget ($b^2 m \gtrsim d^2$). Note that we can assume without loss of generality that $m \geqslant M_\alpha d^2/b^2$ for a constant $M_\alpha > 0$, as otherwise the optimal rate is $\sqrt{d}/(\sqrt{m}n)$, obtained by the 1-bit local randomness test described by (3.3) (see Section 3.1.1). This case is the most involved one, and we construct a test consisting two sub-tests optimal in different sub-regimes.

The most obvious approach in this case is to divide the communication budget of each machine over the d coordinates as uniformly as possible. That is to say, to partition the coordinates $\{1, \dots, d\}$ into approximately d/b sets of size b (we assume without loss of generality that $b \leqslant d$, as we can always throw away excess budget and $b = d$ bits suffices for achieving the minimax rate). The machines are then equally divided over each of these partitions and communicate the coefficients corresponding to their partition. More formally, such a strategy entails taking sets $\mathcal{I}_i \subset \{1, \dots, m\}$ such that $|\mathcal{I}_i| = \lfloor \frac{mb}{d} \rfloor$ and each $j \in \{1, \dots, m\}$ is in $\mathcal{I}_i$ for b different indexes $i \in \{1, \dots, d\}$. For $i = 1, \dots, d$ and $j \in \mathcal{I}_i$, generate the transcripts according to

$$Y_i^{(j)} | X_i^{(j)} = \mathbb{1}\{X_i^{(j)} > 0\}. \tag{3.8}$$

Centrally, a natural test based on these transcripts is

$$T^1_{\mathrm{III}} := \mathbb{1}\Big\{ \Big| \frac{1}{|\mathcal{I}_1|\sqrt{d}} \sum_{i=1}^{d} \Big(\sum_{j \in \mathcal{I}_i} (Y_i^{(j)} - 1/2) \Big)^2 - \sqrt{d}/4 \Big| > \kappa_\alpha \Big\}. \tag{3.9}$$

It turns out that such a test does not cover all regimes where $m \gtrsim d^2/b^2$, because, there is a certain amount of information loss due to the nonlinearity of the quantization step (3.8), i.e. the test induces soft thresholding for the signal components which is suboptimal for (relatively) large signal components. For the exact statement on the testing error of this test, see Lemma 3.17 below.

For detecting signals including large coordinates we propose an adaptation of test T^1_{III}. We start by assuming that $b \geqslant 2\log(d+1)$ otherwise we do not construct the test. Then for $i = 1, \dots, d$ and $j = 1, \dots, m$, let us generate

$$B_{li}^{(j)} \overset{i.i.d.}{\sim} \mathrm{Ber}\Big(F_{\chi^2_1}\Big(\big(\sqrt{n} X_i^{(j)}\big)^2 \Big) \Big), \qquad l \in \{1, \dots, C_{b,d} = \lfloor 2^b/(d+1) \rfloor\}.$$

Note that $C_{b,d} \geqslant 1$ by assumption. Then machine j communicate the transcripts

$$Y^{(j)}_{count} = \sum_{l=1}^{C_{b,d}} \sum_{i=1}^{d} B_{li}^{(j)} \in \{0, 1, \dots, C_{b,d} d\}, \tag{3.10}$$

which can be done using $\log_2(C_{b,d}d+1) \leqslant b$ bits in total. Based on these transcripts, we compute the test

$$T^2_{\mathrm{III}} = \mathbb{1}\left\{\left|\frac{1}{dmC_{b,d}}\left(\sum_{j=1}^{m}(Y^{(j)}_{count} - Ld/2)\right)^2 - \frac{1}{4}\right| \geqslant \kappa_\alpha\right\} \tag{3.11}$$

centrally. The testing risk bound for the above test is given in Lemma 3.18 in the appendix.

Finally, we construct our test by combining the above ones. We construct both partial tests T^1_{III} and T^2_{III} if $b \geqslant 2\log(d+1)$ by transmitting $b' = \lfloor b/2 \rfloor$ bits per machine for each, otherwise we just construct T^1_{III}. Then we merge them by taking

$$T_{\mathrm{III}} = T^1_{\mathrm{III}} \vee T^2_{\mathrm{III}}\mathbb{1}_{\{b \geqslant 2\log(d+1)\}}, \tag{3.12}$$

where the indicator should be understood to rule out cases in which the transcripts for T^2_{III} cannot necessarily be communicated. This case, as shown below, is covered by the first test T^1_{III}. Lemma 3.4 below shows that T_{III} has sufficiently small testing risk in all cases where $m \geqslant M_\alpha d^2/b^2$.

Lemma 3.4. *For $\alpha \in (0,1)$, there exist constants $M_\alpha, C_\alpha > 0$ such that when $m \geqslant M_\alpha d^2/b^2$, the b-bit distributed private testing protocol T_{III} given in* (3.12) *satisfies*

$$\mathcal{R}(H_\rho, T_{III}) \leqslant \alpha,$$

whenever $\rho^2 \geqslant C_\alpha \frac{d\sqrt{d}}{mn(b \wedge d)}$.

Proof. Fix an arbitrary $f \in H_\rho$ and define

$$\mathcal{J} = \{i : 1 \leqslant i \leqslant d,\ \frac{n}{m}f_i^2 \geqslant 1\}. \tag{3.13}$$

By Lemma 3.17 in the appendix, the test T^1_{III} given in (3.9) with $\kappa_\alpha, C_\alpha, M_\alpha > 0$ large enough satisfies

$$\mathbb{E}_0 T^1_{\mathrm{III}} \leqslant \alpha/6, \quad \text{and} \quad \mathbb{E}_f(1 - T^1_{\mathrm{III}}) \leqslant \alpha/6,$$

whenever

$$\sum_{i \notin \mathcal{J}} f_i^2 \geqslant \rho^2/2 \ \text{ or } \ \frac{mb}{d\sqrt{d}} > M_\alpha. \tag{3.14}$$

Next we consider the case where (3.14) does not hold. Then $M_\alpha \geqslant \frac{mb}{d\sqrt{d}} \geqslant M_\alpha \frac{\sqrt{d}}{b}$, where the second inequality follows from the assumption of the lemma. This implies that $b \geqslant \sqrt{d}$. Since $\frac{mb}{d\sqrt{d}} \leqslant M_\alpha$ and m can be taken to be larger than arbitrary constant (otherwise we are in the non-distributed regime in which the minimax rate can be achieved locally), we can without loss of generality assume d is larger than

an arbitrary constant (depending only on α), hence $b \geqslant \sqrt{d} \geqslant 2\log(d+1)$ and the test T^2_{III} and the corresponding transcripts can be constructed. Furthermore, $\sum_{i\in\mathcal{J}^c} f_i^2 < \rho^2/2$ implies $\mathcal{J} \neq \varnothing$ in view of $\sum_i f_i^2 \geqslant \rho^2$. Consequently, the conditions of Lemma 3.18 are satisfied, yielding that there exists a test T^2_{III} such that $\mathbb{E}_0 T^2_{\mathrm{III}} \leqslant \alpha/6$ and $\mathbb{E}_f(1 - T^2_{\mathrm{III}}) \leqslant \alpha/6$. We note that in case $\frac{mb}{d\sqrt{d}} > M_\alpha$, the test T^2_{III} cannot necessarily be computed (not enough communication budget), but this is not required as this case is covered by T^1_{III}.

We now have that for any $f \in H_\rho$, whenever $\frac{mb}{d\sqrt{d}} \leqslant M_\alpha$, the test T_{III} can be computed and using that for nonnegative $x, y \geqslant 0$, $x \vee y \leqslant x + y$ and $x \vee y \geqslant x$, we obtain that

$$\begin{aligned}\mathcal{R}(H_\rho, T_{\mathrm{III}}) &\leqslant \mathbb{E}_0 T^1_{\mathrm{III}} + \mathbb{E}_0 T^2_{\mathrm{III}} \mathbb{1}_{\{b \geqslant 2\log(d+1)\}} \\ &\quad + \sup_{f\in H_\rho} \min\left\{\mathbb{E}_f(1 - T^1_{\mathrm{III}}), \mathbb{E}_f(1 - T^2_{\mathrm{III}}\mathbb{1}_{\{b\geqslant 2\log(d+1)\}})\right\} \\ &\leqslant 2\alpha/6 + \alpha/6 = \alpha/2.\end{aligned}$$

□

3.2 Testing protocols under privacy constraints

> "*An interesting thing about differential privacy is that it needs theorems even in practice. You can implement heuristic algorithms that are fast on your data, but 'heuristic privacy' does not exist. A mechanism isn't private without a theorem.*" – Jelani Nelson

In this section, we exhibit distributed differentially private testing procedures achieving (up to log factors) the rates posed by the lower bounds of Theorem 2.4 in Chapter 2. Together, they yield Theorem 1.2. We consider the test of hypotheses

$$H_0 : f = 0 \text{ versus the alternative hypothesis } f \in H_\rho = \left\{f \in \mathbb{R}^d : M \geqslant \|f\|_2 \geqslant \rho\right\}. \tag{3.15}$$

The restriction to signals of bounded norm is standard in privacy and does not change the conclusion of the lower bound, Theorem 2.4, see Remark 6. The rates attained by the procedures in this section are summarized by the theorem below.

Theorem 3.2. *Consider for some constant $M > 0$ the test of hypotheses in* (3.15)*. For all $\alpha \in (0,1)$, there exists a constant $C_\alpha > 0$ such that for all $n, m, d \in \mathbb{N}$ and $(mn)^{-1} < \epsilon \leqslant 1$, $\delta \geqslant \mathbb{1}_{\{\epsilon \geqslant n^{-1/2}\}}(mnd)^{-2}$, there exists a (ϵ, δ)-differentially private distributed testing protocol T using shared randomness such that $\mathcal{R}(H_\rho, T) \leqslant \alpha$ whenever*

$$\rho^2 > C_\alpha \log^6(1 + mnd)\left(\frac{d}{mn\sqrt{n\epsilon^2 \wedge 1}\sqrt{n\epsilon^2 \wedge d}} \bigwedge \left(\frac{\sqrt{d}}{\sqrt{m}n\sqrt{n\epsilon^2 \wedge 1}} \bigvee \frac{1}{mn^2\epsilon^2}\right)\right), \tag{3.16}$$

Similarly, for all $\alpha \in (0,1)$, there exists a constant $C_\alpha > 0$ such that for all $n, m, d \in \mathbb{N}$, $(mn)^{-1} < \epsilon \leqslant 1$ and $\delta \gtrsim \mathbb{1}_{\{\epsilon \geqslant n^{-1/2}\}}(mnd)^{-2}$, there exists a (ϵ, δ)-differentially private local randomness distributed testing protocol T such that $\mathcal{R}(H_\rho, T) \leqslant \alpha$ whenever

$$\rho^2 > C_\alpha \log^6(1+mnd)\left(\frac{d\sqrt{d}}{mn(n\epsilon^2 \wedge d)} \bigwedge \left(\frac{\sqrt{d}}{\sqrt{m}n\sqrt{n\epsilon^2 \wedge 1}} \bigvee \frac{1}{mn^2\epsilon^2}\right)\right). \quad (3.17)$$

Just as it was the case in the bandwidth constraint setting, the different rates, depending on ϵ comparatively to n, d and m, correspond to different testing "regimes". We shall coin these regimes using similar terminology. We start in Section 3.2.1 by designing a differentially private testing protocol that uses only local randomness and that is optimal in the "small ϵ regime"; $\epsilon \lesssim \sqrt{d}/\sqrt{nm}$ or $\epsilon \lesssim d/\sqrt{nm}$ for shared- and local randomness, respectively. Then, in Section 3.2.2, we design two local randomness protocols attaining the rates for the "in-between" and "large" ϵ regimes, where $\epsilon \gtrsim \sqrt{d}/\sqrt{nm}$ or $\epsilon \gtrsim d/\sqrt{nm}$ for shared- and local randomness, respectively. As remarked in Section 1.3.2, there are certain values of d, m, n, where some of these regimes do not occur.

The first distributed differentially private testing procedure T_{I} is $(\epsilon, 0)$-differentially private for any $\epsilon > 0$ and can detect signals with a squared Euclidean norm of a (poly-logarithmic) factor larger than $\sqrt{d}/(\sqrt{m}n\sqrt{n\epsilon^2 \wedge 1})$, whenever $(mn)^{-1} < \epsilon \leqslant 1$. This first procedure does not need a shared randomness.

As a second procedure, we consider a distributed differentially private testing protocol using shared randomness, that achieves the rate $\frac{d}{mn^{3/2}\epsilon\sqrt{n\epsilon^2 \wedge 1}}$ (again up to log-factors). Whenever $\epsilon \leqslant 1/\sqrt{n}$, this procedure can be implemented with a $(\epsilon, 0)$-differential privacy guarantee, in which case we shall denote it as $T^{\epsilon}_{\mathrm{II}}$. For the range of values $1/\sqrt{n} < \epsilon \leqslant 1$, we shall consider a version of this protocol that employs (ϵ, δ)-differential privacy, which we denote $T^{\epsilon,\delta}_{\mathrm{II}}$.

The third procedure, is a distributed differentially private testing protocol that uses only local randomness and achieves the rate $d\sqrt{d}/(mn^2\epsilon^2)$ (up to log-factors). Whenever $\epsilon \leqslant 1/\sqrt{n}$, the procedure satisfies $(\epsilon, 0)$-differential privacy constraints, and shall be denoted by $T^{\epsilon}_{\mathrm{III}}$. For the range of values $1/\sqrt{n} < \epsilon \leqslant 1$, we shall construct a (ϵ, δ)-differentially private version, $T^{\epsilon,\delta}_{\mathrm{III}}$.

The approximate differentially private tests $T^{\epsilon,\delta}_{\mathrm{II}}$ and $T^{\epsilon,\delta}_{\mathrm{III}}$ employed when $\epsilon > 1/\sqrt{n}$ attain the respective lower bound rates (up to logarithmic factors) for values of δ as small as $(mn)^{-C}$ for an arbitrary constant $C > 0$. The existence of such distributed testing protocols proves Theorem 3.2.

Before delving into the construction of these specific protocols for the different regime, we cover the general strategy for the design of these protocols. Similarly to how the bandwidth constraint protocols essentially boil down to testing uniformity of a sequence of bits, which are generated from the local data, the distributed privacy

protocols can be seen as combining noisy versions of statistics of the data. The "type" and "amount" of noise added depends on the *sensitivity* of the statistics.

Formally, consider a metric d on $\mathbb{R}^k$. Given n elements $x = (x_1, \ldots, x_n)$ in a sample space $\mathcal{X}$, the d-*sensitivity at* x of a map $S : \mathcal{X}^n \to \mathbb{R}^k$ is

$$\Delta_S(x) := \sup_{\breve{x} \in \mathcal{X}^n : \mathrm{d}_H(x, \breve{x}) \leqslant 1} \mathrm{d}\left(S(x), S(\breve{x})\right),$$

where d_H is the Hamming distance on $\mathcal{X}^n$,

$$\mathrm{d}_H(x, \breve{x}) := \sum_{i=1}^{n} \mathbb{1}\left\{x_i \neq \breve{x}_i\right\}. \tag{3.18}$$

The d-*sensitivity* of S is defined as $\Delta_S := \sup_x \Delta_S(x)$.

In our case, the sample space under consideration is $\mathbb{R}^d$. In this section, the noise mechanisms under consideration are *Laplace mechanism* and the *Gaussian mechanism*. These can be used to generate differentially private transcripts by adding either Laplacian or Gaussian noise to statistics under consideration. The Laplace mechanism yields ϵ-differentially private transcripts for statistics $x \mapsto S(x)$ that have bounded L_1-sensitivity, where the variance of the Laplace noise scales with the L_1-sensitivity. The Gaussian mechanism yields (ϵ, δ)-differentially private transcripts for statistics that have bounded L_2-sensitivity, with the noise variance scaling with the L_2-sensitivity.

The following lemma shows the way in which adding appropriately scaled Laplace noise can be used to guarantee ϵ-differential privacy. The result is well known (see e.g. [86]), but since the proof is short and instructive it is included below.

Lemma 3.5. *Suppose that the map $S : (\mathbb{R}^d)^n \to \mathbb{R}^k$ has $\|\cdot\|_1$-sensitivity $\Delta_S \in (0, \infty)$. Let $W = (W_1, \ldots, W_k)$ be a vector of i.i.d. centered Laplace random variables with scale parameter $\epsilon^{-1}\Delta_S$. Then, the transcript*

$$T(x) = S(x) + W$$

is ϵ-differentially private.

Proof. By the triangle inequality, the ratio of densities of the random variables $S(x) + W$ and $S(x') + W$ satisfies

$$e^{-\frac{\epsilon}{\Delta_S}\|S(x)+w\|_1 + \frac{\epsilon}{\Delta_S}\|S(x')+w\|_1} \leqslant e^{\frac{\epsilon}{\Delta_S}\|S(x)-S(x')\|_1} \leqslant e^{\epsilon}.$$

Consequently, Definition 3 can be seen to be satisfied:

$$\Pr\left(S(x) + W \in A\right) \leqslant e^{\epsilon} \Pr\left(S(x') + W \in A\right).$$

□

The next lemma shows how appropriately scaled Gaussian noise provides (ϵ, δ)-differential privacy for mappings with bounded L_2-sensitivity. This result is also well known, and a proof can be found in e.g. Appendix A of [86].

Lemma 3.6. *Suppose that the map $S : (\mathbb{R}^d)^n \to \mathbb{R}^k$ has $\|\cdot\|_2$-sensitivity $\Delta_S \in (0, \infty)$. Let $W = (W_1, \ldots, W_k)$ be a vector of i.i.d. standard Gaussian random variables. Then, the transcript*

$$T(x) = S(x) + \sqrt{3 \log(1/\delta)} \epsilon^{-1} \Delta_S W$$

is (ϵ, δ)-differentially private.

In order to obtain statistics with (uniformly) bounded sensitivity with respect to the L_1 and L_2-norms, it shall prove particularly useful to bound quantities between thresholds, which we shall refer to as *clipping*. Formally, for $a, b, x \in \mathbb{R}$ with $a < b$, let $[x]_a^b$ denote x clipped at a and b, that is

$$[x]_a^b := \begin{cases} b & \text{if } x > b, \\ x & \text{if } a \leqslant x \leqslant b, \\ a & \text{otherwise.} \end{cases}$$

For $x \in \mathbb{R}^d$, let $(x)_i$ denote the projection onto the i-th coordinate and let $[x]_a^b = \{[(x)_i]_a^b : i = 1, \ldots, d\}$.

In estimation, clipping and averaging (functionals of) the observations is a common strategy that enjoys good sensitivity. For $x_1, \ldots, x_n \in \mathbb{R}^d$, $x_i = (x_{i1}, \ldots, x_{id})$, the combination of clipping at τ and $-\tau$ and averaging over $i = 1, \ldots, n$ yields sensitivity of the order $2\tau d/n$ for the L_1-norm and $2\tau\sqrt{d}/n$ for the L_2-norm, uniformly over the sample space. That is, the map $S : \mathbb{R}^{n \times d} \to \mathbb{R}^d$ defined by $S(x_1, \ldots, x_n) := n^{-1} \sum_{i=1}^n [x_i]_{-\tau}^{\tau}$ satisfies

$$|S(x_1, \ldots, x_n)_k - S(\breve{x}_1, \ldots, \breve{x}_n)_k| = \left| \frac{1}{n} \sum_{i=1}^n [x_{ik}]_{-\tau}^{\tau} - \frac{1}{n} \sum_{i=1}^n [\breve{x}_{ik}]_{-\tau}^{\tau} \right| \leqslant \frac{2\tau}{n},$$

for $\breve{x}_1, \ldots, \breve{x}_n \in \mathbb{R}^d$ such that $\breve{x}_i = x_i$ for all but one $i \in [n]$. The larger L_1-sensitivity when $d > 1$ implies that the variance of the Laplace noise added will be larger than that of the Gaussian mechanism, which typically leads to a less powerful test.

Functions of the data that are superlinear on the entire sample space will typically have worse sensitivity than sublinear functions, such as the average. One technique that we will employ, that allows the use of e.g. a quadratic function is Lipschitz extension. The idea being that for $S : \mathcal{X} \to \mathbb{R}$, if $x \mapsto S(x)$ is D-Lipschitz on $\mathcal{C} \subset \mathcal{X}$ and we expect that most of our observations will be in $\mathcal{C}$, we can define S on $\mathcal{C}$ only and consider a Lipschitz extension $\tilde{S}$ of S to the whole space. This way, $\tilde{S}$ enjoys "sublinear sensitivity" on the whole space. In particular, if S is D-Lipschitz with respect to the Hamming distance, we have $|S(x) - S(\breve{x})| \leqslant D$ for all $x, \breve{x} \in \mathcal{X}^n$ such that $\mathrm{d}_H(x, \breve{x}) \leqslant 1$, so $\tilde{S}$ has sensitivity D.

The existence of such a Lipschitz extension is guaranteed by a version of the McShane–Whitney Extension Theorem. In particular, we use the construction of McShane. We provide our own proof which accounts for potential measurability issues stemming from the discrete topology of the Hamming distance in the construction.

Lemma 3.7. *Let $\mathcal{C} \subset \mathbb{R}^{n\times d}$ and $S : \mathcal{C} \to \mathbb{R}$ be a (Borel) measurable D-Lipschitz map with respect to the Hamming distance on $(\mathbb{R}^d)^n$ as defined in* (3.18)*. Then, there exists a map $\tilde{S} : \mathbb{R}^{n\times d} \to \mathbb{R}$ measurable with respect to the Borel sigma algebra such that it is D-Lipschitz with respect to the Hamming distance on its entire domain and $\tilde{S} = S$ on $\mathcal{C}$.*

We provide a proof of the lemma in Section 3.4.2.1 of the chapter appendix. In the next section, we leverage the Lipschitz-extension above to create a differentially private version of the chi-square test that would be locally optimal, but which has poor sensitivity on the entire sample space.

3.2.1 Testing using aggregated locally optimal private test statistics

In this section, we construct a test statistic that is locally optimal, in the sense that it reaches the local minimax rate under ϵ-differential privacy as established (up to poly-log factors) in [157]. That is, if $m = 1$, the ϵ-differentially private test statistic reaches the minimax rate of the local problem. When $m > 1$, in the central machine, combining m of these locally optimal tests results in a test which is optimal in the regime where $\epsilon \lesssim \sqrt{d/(mn)}$ in case of shared randomness or $\epsilon \lesssim d/\sqrt{mn}$ in case of local randomness.

Given $\tau > 0$ and $V^{(j)} \sim \chi^2_d$ independent of $X^{(j)}$, let

$$S^{(j)}_\tau(X^{(j)}) := \left[\frac{n}{\sqrt{d}}\left(\left\|\overline{X^{(j)}}\right\|^2 - \frac{V^{(j)}}{n}\right)\right]^\tau_{-\tau}. \tag{3.19}$$

For any τ, this test statistic can be seen to be mean zero under the null hypothesis, since $\|\sqrt{n}\overline{X^{(j)}}\|^2 \sim \chi^2_d$ under $\mathbb{P}_0$. Under the alternative hypothesis, the test statistic picks up a positive "bias" since $\|\sqrt{n}\overline{X^{(j)}}\|^2 \sim \chi^2_d(\|f\|_2^2)$ under $\mathbb{P}_f$.

We will use that for "typical" data, i.e. data that occurs with relatively high probability, $x \mapsto S^{(j)}_\tau(x)$ has relatively good sensitivity. However, because of the nonlinearity of the squared Euclidean norm, the sensitivity of the statistic $x \mapsto S^{(j)}_\tau(x)$ is large for certain data $x \in \mathbb{R}^{d\times n}$. To mitigate this, we follow a similar strategy to that proposed in [54] and improved upon by [157]. That is, we define $x \mapsto S^{(j)}_\tau(x)$ (as in (3.19)) only on a set on which the sensitivity is good. Lemma 3.10 below shows that on a set $\mathcal{C}_\tau \subset \mathbb{R}^{d\times n}$ depending on n, d, α, M and τ, the map $x \mapsto S^{(j)}_\tau(x)$ is D_τ-Lipschitz with respect to the Hamming distance. Specifically, for $x, \breve{x} \in \mathcal{C}_\tau$,

$$|S_\tau(x_1, \ldots, x_n) - S_\tau(\breve{x}_1, \ldots, \breve{x}_n)| \leqslant D_\tau \mathrm{d}_H(x, \breve{x})$$

with $D_{\tau,n,d,m,\gamma_\tau} \equiv D_\tau = \gamma_\tau / n\sqrt{d}$ and

$$\gamma_\tau \equiv \gamma_{\tau,n,d,m} = \tilde{\kappa}_\alpha \log(1+m) \left(\sqrt{\log^2(1+n) n\sqrt{d}\tau} \vee \sqrt{nd} \right), \tag{3.20}$$

for a constant $\tilde{\kappa}_\alpha > 0$. Then, in order to obtain a D_τ-Lipschitz statistic which is well-defined on the full sample space, we compute a Lipschitz extension of $x \mapsto S_\tau^{(j)}(x)$ to $\mathbb{R}^{d\times n}$. Such a Lipschitz extension exists by the McShane–Whitney Extension theorem, but the construction as applied in [54, 157] is not necessarily Borel measurable. A Borel-measurable Lipschitz extension of the map that is guaranteed to exist by Lemma 3.7, for which we provide a proof in the chapter appendix, Section 3.4.2.1. In addition, our construction differs from that of [54, 157], in order to allow easier combination of the test statistics and improving performance by several log-factors (which is due to a slightly sharper analysis). We denote the Lipschitz extension of $S_\tau^{(j)}$ by $\breve{S}_\tau^{(j)}$, which then satisfies $S_\tau^{(j)}(x) = \breve{S}_\tau^{(j)}(x)$ on $\mathcal{C}_\tau$ and is D_τ-Lipschitz on the entirety of $\mathbb{R}^{d\times n}$. The set $\mathcal{C}_\tau$ is constructed such that $X^{(j)}$ takes values in it with high probability. This results in $\breve{S}_\tau^{(j)}(X^{(j)})$ exhibiting similar probabilistic behavior as the original test $S_\tau^{(j)}(X^{(j)})$, whilst assuring it has a much smaller sensitivity is greatly improved over the whole parameter space. The explicit construction of $\mathcal{C}_\tau$ is given at the end of this section, as well as the proofs of the aforementioned lemmas. Consider for $\tau > 0$ $J \equiv J_{m,n} := \log(\lceil 1 + 2\log_2(mnM) \rceil)$ the (partial) transcript

$$Y_\tau^{(j)} = \frac{\epsilon}{D_\tau J} \breve{S}_\tau^{(j)}(X^{(j)}) + W_\tau^{(j)} \tag{3.21}$$

with $W_\tau^{(j)} \sim \mathrm{Lap}(1)$ independent for $j = 1, \dots, m$ and $\tau > 0$. Since $x \mapsto \frac{\epsilon}{JD_\tau} \breve{S}_\tau^{(j)}(x)$ can be seen to have sensitivity ϵ/J by the fact that $\breve{S}_\tau^{(j)}$ is D_τ-Lipschitz and consequently the partial transcript is ϵ/J-differentially private by Lemma 3.5. The lemma below shows that the test

$$\varphi_\tau^\epsilon := \mathbb{1}\left\{ \frac{1}{\sqrt{m}} \sum_{j=1}^m Y_\tau^{(j)} \geqslant \left(\frac{\epsilon}{D_\tau J} \vee 1 \right) \sqrt{J} \kappa_\alpha \right\} \tag{3.22}$$

has Type I error less than or equal to α/J for $\kappa_\alpha > 0$ large enough and detects signals that are "close" to the clipping τ.

Lemma 3.8. *The test φ_τ^ϵ defined in* (3.22) *satisfies $\mathbb{P}_0 \varphi_\tau^\epsilon \leqslant \alpha/J$. Furthermore, whenever*

$$\tau/4 \leqslant \frac{n\|f\|_2^2}{\sqrt{d}} \leqslant \tau/2,$$

and f satisfies (3.25) *for $C_\alpha > 0$ large enough, it holds that $\mathbb{P}_f(1 - \varphi_\tau) \leqslant \alpha$ for $J = \log(\lceil 1 + 2\log_2(mnM) \rceil)$.*

A proof of the lemma is given later on in the section. Essentially, the above test is calibrated for the detection of signals with signal size between $\tau/4$ and $\tau/2$. In order

to detect signals of any size larger than the right-hand side of (3.25), we follow what is essentially a multiple testing procedure. For large signals, we need a larger clipping to detect them, as well as a larger set $\mathcal{C}_\tau$ to assure that the data is in $\mathcal{C}_\tau$ with high probability, as larger signals increase the probability of "outliers" from the perspective of sensitivity of the Euclidean norm. The additional $J = \lceil 1 + 2\log_2(mnM) \rceil$ "blow up" can be seen as a Bonferroni correction.

Since $1/mn^2 \leqslant \|f\|_2 \leqslant M$, it suffices to compute tests for partial transcripts for

$$\tau \in \mathrm{T} := \left\{ 2^{-k+2} \frac{nM^2}{\sqrt{d}} : k = 1, \dots, J \right\}. \tag{3.23}$$

For each $\tau \in \mathrm{T}$, the machine transfers (3.21), yielding as a full transcript $Y^{(j)} = \{Y_\tau^{(j)} : \tau \in \mathrm{T}\}$. Since each partial transcript $Y_\tau^{(j)}$ is ϵ/J-differentially private with independent Laplacian noise, the full transcript $Y^{(j)}$ is ϵ-differentially private by Lemma 3.5. The test

$$T_{\mathrm{I}}^\epsilon := \max_{\tau \in \mathrm{T}} \varphi_\tau \tag{3.24}$$

then satisfies $\mathbb{P}_0 T_{\mathrm{I}}^\epsilon \leqslant \alpha$ via a union bound. Furthermore, for $f \in \mathbb{R}^d$ such that $M \geqslant \|f\|_2 \geqslant \rho$, we have a $\tau^* \in \mathrm{T}$ such that $\tau^*/4 \leqslant \sqrt{d} n \|f\|_2^2 \leqslant \tau^*/2$ and

$$\mathbb{P}_f(1 - T_{\mathrm{I}}^\epsilon) \leqslant \mathbb{P}_f(1 - \varphi_{\tau^*}) \leqslant \alpha/2.$$

Given what we have thus far, we obtain the following statement.

Lemma 3.9. *For all $M > 0$, $\alpha \in (0,1)$ there exists $\kappa_\alpha > 0$ and $C_\alpha > 0$ such that the test T_I^ϵ defined by* (3.24) *and* (3.22) *satisfies*

$$\mathbb{P}_0 T_I^\epsilon + \mathbb{P}_f(1 - T_I^\epsilon) \leqslant \alpha$$

for all $f \in \mathbb{R}^d$ such that

$$M^2 \geqslant \|f\|_2^2 \geqslant C_\alpha \log^6(1 + mn) \left(\frac{\sqrt{d}}{\sqrt{m} n (\sqrt{n}\epsilon \wedge 1)} \right) \bigvee \left(\frac{1}{mn^2\epsilon^2} \right). \tag{3.25}$$

Next, we discuss the construction of $\mathcal{C}_\tau$ and finish by proving the aforementioned lemmas. Define for $\tau > 0$ the sets

$$\begin{aligned} \mathcal{A}_\tau &= \left\{ (x_i) \in (\mathbb{R}^d)^n : \left| \left\| \textstyle\sum_{i \in \mathcal{J}} x_i \right\|_2^2 - kd \right| \leqslant k\gamma_\tau \ \ \forall \mathcal{J} \subset [n], |\mathcal{J}| = k \leqslant K \right\}, \\ \mathcal{B}_\tau &= \left\{ (x_i) \in (\mathbb{R}^d)^n : \left| \left\langle x_i, \textstyle\sum_{k \in [n] \setminus \{i\}} x_k \right\rangle \right| \leqslant \gamma_\tau, \ \ \forall i = 1, \dots, n \right\}, \end{aligned} \tag{3.26}$$

with $K = \lceil 2\tau D_\tau^{-1} \rceil$ and let $\mathcal{C}_\tau = \mathcal{A}_\tau \cap \mathcal{B}_\tau$. Lemma 3.10 below tells us that $x \mapsto S^{(j)}(x)$ is Lipschitz on $\mathcal{C}_\tau$ with Lipschitz constant $D_\tau = \frac{8\gamma}{n\sqrt{d}}$.

Lemma 3.10. *The map $x \mapsto S_\tau^{(j)}(x)$ defined in* (3.19) *is D_τ-Lipschitz with respect to $(\mathbb{R}^d)^n$-Hamming distance on $\mathcal{C}_\tau$.*

Proof. Consider $x = (x_i)_{i\in[n]}, \breve{x} = (\breve{x})_{i\in[n]} \in \mathcal{C}_\tau$ with $k := \mathrm{d}_H(x, \breve{x})$. If $k > \lceil 2\tau D_\tau^{-1}\rceil$, we have $|S_\tau^{(j)}(x) - S_\tau^{(j)}(\breve{x})| \leqslant 2\tau \leqslant D_\tau k$. If $k \leqslant \lceil 2\tau D_\tau^{-1}\rceil$, let $\mathcal{J} \subset [n]$ denote the indexes of columns in which x and $\breve{x}$ differ. Define the sum of the elements that x and $\breve{x}$ have in common as

$$v = \sum_{i\in[n]\setminus\mathcal{J}} x_i, \text{ such that } \sum_{i=1}^n x_i = v + w \text{ and } \sum_{i=1}^n \breve{x}_i = v + \breve{w}.$$

We have

$$\begin{aligned} S_\tau^{(j)}(x) - S_\tau^{(j)}(\breve{x}) &= \frac{n}{\sqrt{d}}\left(\left\|n^{-1}(v+w)\right\|^2 - \frac{V^{(j)}}{n}\right) - \frac{n}{\sqrt{d}}\left(\left\|n^{-1}(v+\breve{w})\right\|^2 - \frac{V^{(j)}}{n}\right) \\ &= \frac{1}{n\sqrt{d}}\left(2\langle w, v\rangle - 2\langle \breve{w}, v\rangle + \|w\|_2^2 - \|\breve{w}\|_2^2\right). \end{aligned}$$

The last two terms are bounded by $2k\gamma_\tau/(n\sqrt{d})$ since $x, \breve{x} \in \mathcal{A}_\tau$. The first two terms equal

$$\frac{2}{n\sqrt{d}}\left(\langle w, v+w\rangle - \langle \breve{w}, v+\breve{w}\rangle + \|\breve{w}\|_2^2 - \|w\|_2^2\right),$$

where the last two terms are bounded by $4k\gamma_\tau/(n\sqrt{d})$. It holds that

$$\langle w, v+w\rangle - \langle \breve{w}, v+\breve{w}\rangle = \sum_{i\in\mathcal{J}}\left(\left\langle x_i, \sum_{i\in[n]\setminus\mathcal{J}} x_i\right\rangle - \left\langle \breve{x}_i, \sum_{i\in[n]\setminus\mathcal{J}} \breve{x}_i\right\rangle + \|x_i\|_2^2 - \|\breve{x}_i\|_2^2\right),$$

which is bounded by $2k\gamma_\tau$ for $x \in \mathcal{A}_\tau \cap \mathcal{B}_\tau$. Putting it all together and by symmetry of the argument, we obtain that

$$\left|S_\tau^{(j)}(x) - S_\tau^{(j)}(\breve{x})\right| \leqslant \frac{8k\gamma_\tau}{n\sqrt{d}} = D_\tau k.$$

□

Under the null hypothesis, the observations the $X^{(j)}$'s are in $\mathcal{C}_\tau$ for every τ with high probability. For each element f from the alternative hypothesis, there exists a τ^* such that the $X^{(j)}$'s are in $\mathcal{C}_{\tau^*}$ with high probability. This is the content of the following lemma.

Lemma 3.11. *Whenever $n\|f\|_2^2 d^{-1/2} \leqslant \tau/2$, $\tau \leqslant nM^2/\sqrt{d}$ and $\tilde{\kappa}_\alpha$ in* (3.20) *is taken large enough, it holds that*

$$\mathbb{P}_f\left(X^{(j)} \notin \mathcal{C}_\tau\right) \leqslant \frac{\alpha}{2m}.$$

We now finish the proof of Lemma 3.9 by providing a proof for Lemma 3.8. Lemma 3.11 is proven in Section 3.4.2.3 in the chapter appendix.

Proof of Lemma 3.8. On the event that $X^{(j)} \in \mathcal{C}_\tau$ for all $j \in [m]$, we have that

$$\sum_{j=1}^m Y_\tau^{(j)} = \sum_{j=1}^m \left(\frac{\epsilon}{D_\tau J} \breve{S}_\tau(X^{(j)}) + W_\tau^{(j)} \right) = \sum_{j=1}^m \left(\frac{\epsilon}{D_\tau J} S_\tau(X^{(j)}) + W_\tau^{(j)} \right). \tag{3.27}$$

Consequently, $\mathbb{P}_0 \varphi_\tau$ is bounded above by

$$\mathbb{P}_0 \left(\frac{1}{\sqrt{m}} \sum_{j=1}^m \left(\frac{\epsilon}{D_\tau J} S_\tau(X^{(j)}) + W_\tau^{(j)} \right) \geqslant \left(\frac{\epsilon}{D_\tau J} \vee 1 \right) \sqrt{J} \kappa_\alpha \right) + \mathbb{P}_0 \left(\exists j : X^{(j)} \notin \mathcal{C}_\tau \right).$$

By Lemma 3.11 and a union bound, the second term is bounded above by $\alpha/2$. Under $\mathbb{P}_0$, the terms in (3.27) are independent mean zero. By Lemma 3.22 and a straightforward computation, the variance of each term in (3.27) is bounded above by $\frac{4\epsilon^2}{D_\tau^2 J^2} + 2$. By Chebyshev's inequality, the first term in the display above is therefore bounded by $\left(\sqrt{J}\kappa_\alpha(\epsilon/(D_\tau J) \vee 1)\right)^{-2} \left(\frac{4\epsilon^2}{D_\tau^2 J^2} + 2 \right)$, so choosing κ_α large enough yields the first statement of the lemma. For the second statement, note that the same union bound as above yields that $\mathbb{P}_f(1 - \varphi_\tau)$ is bounded above by

$$\mathbb{P}_f \left(\frac{1}{\sqrt{m}} \sum_{j=1}^m \left(\frac{\epsilon}{D_\tau J} S_\tau(X^{(j)}) + W_\tau^{(j)} \right) < \left(\frac{\epsilon}{D_\tau J} \vee 1 \right) \sqrt{J} \kappa_\alpha \right) + \alpha/2, \tag{3.28}$$

also using Lemma 3.11. Under the alternative hypothesis,

$$\frac{n}{\sqrt{d}} \left(\left\| \overline{X^{(j)}} \right\|_2^2 - \frac{V^{(j)}}{n} \right) \overset{d}{=} \frac{n\|f\|_2^2}{\sqrt{d}} + 2\frac{\sqrt{n}}{\sqrt{d}} \langle Z, f \rangle + \frac{\|Z\|_2^2 - V^{(j)}}{\sqrt{d}}. \tag{3.29}$$

By assumption, $\frac{n\|f\|_2^2}{\sqrt{d}} \leqslant \tau/2$, $\mathrm{Var}(\frac{\sqrt{n}}{\sqrt{d}} \langle Z, f \rangle) = n\|f\|_2^2/d \leqslant \tau/2$ and $(\|Z\|_2^2 - V^{(j)})/\sqrt{d}$ tends to a Gaussian with variance 4 for large d. The second and third term in (3.29) are symmetric in distribution about 0, have uniformly bounded densities (since the chi-square and normal densities are bounded, and the third term tends weakly to a Gaussian in d) and $d^{-1/2} n\|f\|_2^2 \leqslant \tau/2$, which means that the conditions of Lemma 3.21 are satisfied. Applying said lemma (with $\mu = d^{-1/2} n\|f\|_2^2$), we get that there exists a uniform constant $c > 0$ such that

$$\mathbb{E}_f \frac{1}{\sqrt{m}} \sum_{j=1}^m \left(\frac{\epsilon}{D_\tau J} S_\tau(X^{(j)}) + W_\tau^{(j)} \right) \geqslant c \frac{\sqrt{m} n \|f\|_2^2 \epsilon}{\sqrt{d} D_\tau J}.$$

Under $\mathbb{P}_f$, by independence of the data and the Laplacian noise,

$$\mathrm{Var}_f \left(\frac{1}{\sqrt{m}} \sum_{j=1}^m \frac{\epsilon}{D_\tau J} S_\tau(X^{(j)}) + W_\tau^{(j)} \right) = 1 + \mathrm{Var}_f \left(\frac{\epsilon}{D_\tau J} S_\tau(X^{(1)}) \right).$$

Since

$$\mathbb{E}_f \frac{n}{\sqrt{d}} \left(\left\| \overline{X^{(j)}} \right\|_2^2 - \frac{V^{(j)}}{n} \right) = \frac{n\|f\|_2^2}{\sqrt{d}} \leqslant \tau/2,$$

Lemma 3.22 yields

$$\begin{aligned} \mathrm{Var}_f \left(\frac{\epsilon}{D_\tau J} S_\tau(X^{(1)}) \right) &\leqslant \frac{\epsilon^2}{D_\tau^2 J^2} \mathrm{Var}_f \left(\frac{n}{\sqrt{d}} \left(\left\| \overline{X^{(j)}} \right\|_2^2 - \frac{V^{(j)}}{n} \right) \right) \\ &\leqslant \frac{\epsilon^2}{D_\tau^2 J^2} \left(\frac{4n\|f\|_2^2}{d} + 4 \right). \end{aligned}$$

Assume now that for all $C_\alpha > 0$ large enough,

$$\left(\frac{\epsilon}{D_\tau J} \vee 1 \right) \sqrt{J} \kappa_\alpha \leqslant c \frac{\sqrt{m} n \|f\|_2^2 \epsilon}{2\sqrt{d} D_\tau J}, \tag{3.30}$$

which is a claim we shall prove later on. Then, the first term in (3.28) is bounded above by

$$\mathbb{P}_f \left(\frac{1}{\sqrt{m}} \sum_{j=1}^{m} \left(\frac{\epsilon}{D_\tau J} S_\tau(X^{(j)}) + W_\tau^{(j)} - \mathbb{E}_f(S_\tau(X^{(j)}) + W_\tau^{(j)}) \right) < -c \frac{\sqrt{m} n \|f\|_2^2 \epsilon}{2\sqrt{d} D_\tau J} \right), \tag{3.31}$$

which, by Chebyshev's inequality is bounded by

$$\begin{aligned} &\left(c \frac{\sqrt{m} n \|f\|_2^2 \epsilon}{2\sqrt{d} D_\tau J} \right)^{-2} \left(1 + \frac{\epsilon^2}{D_\tau^2 J^2} \left(\frac{4n\|f\|_2^2}{d} + 4 \right) \right) \lesssim \\ &\left(\frac{\sqrt{m} n \|f\|_2^2 \epsilon}{\sqrt{d} D_\tau \log(nm)} \right)^{-2} + \left(mn\|f\|_2^2 \right)^{-1} + \left(\frac{\sqrt{m} n \|f\|_2^2}{\sqrt{d}} \right)^{-2}. \end{aligned}$$

For f satisfying (3.25), the last two terms are easily seen to be smaller than $\alpha/6$ for a large enough to choice for C_α. To see that this is also true for the first term, recall that $D_\tau = (8\gamma_\tau)/(n\sqrt{d})$ with γ_τ as defined in (3.20), which yields that the square root of the first term equals

$$\frac{\sqrt{m} n^2 \|f\|_2^2 \epsilon}{8\tilde{\kappa}_\alpha \log(1+m) \left(\sqrt{\log^2(1+n) n\sqrt{d}\tau} \vee \sqrt{nd} \right) \log(nm)},$$

which is larger than $C_\alpha \log(mn)$ when the maximum is taken in $\sqrt{nd}$. When the maximum is taken in $\sqrt{\log^2(1+n) n\sqrt{d}\tau}$, using that $4n\|f\|_2^2/\sqrt{d} \geqslant \tau$ yields that the above display is bounded by

$$\frac{\sqrt{m} n \|f\|_2 \epsilon}{16\tilde{\kappa}_\alpha \log(1+m) \log(1+n) \log(nm)} \geqslant C_\alpha.$$

In either case, it follows that the Type II error (i.e. (3.31)) can be made arbitrarily small per large enough choice of $C_\alpha > 0$.

We return to the claim of (3.30). First, we note that

$$\frac{\sqrt{mn}n\|f\|_2^2\epsilon}{\sqrt{d}D_\tau J} = \frac{\sqrt{mn}n^2\|f\|_2^2\epsilon}{\tilde{\kappa}_\alpha\sqrt{\log^2(1+n)n\sqrt{d}\tau} \vee \sqrt{nd}\log(1+nm)} \gtrsim 1$$

is what we have shown above. The inequality

$$\frac{\sqrt{mn}n\|f\|_2^2\epsilon}{\sqrt{d}D_\tau J} \geqslant \frac{\epsilon}{D_\tau J}$$

follows immediately for f satisfying (3.25). □

3.2.2 Tests using coordinate wise strategies under differential privacy

This section presents four different protocols, all aiming at the "high" and "medium" privacy budget regimes, which occur whenever $\epsilon \gtrsim \sqrt{d}/(\sqrt{mn})$ in the case of shared randomness strategies, or

$$\epsilon \gtrsim \frac{d}{\sqrt{mn}} \text{ if } \epsilon \gtrsim \frac{1}{\sqrt{n}} \text{ or } \frac{d}{\sqrt{mn}} \lesssim \epsilon \lesssim \frac{1}{\sqrt{n}},$$

in case of only local randomness.

A common element in these four strategies is that they try to reconstruct or approximate the aggregated data by combining a noisy and clipped version of the local data. This is in contrast to the aggregated statistics corresponding to the locally optimal private tests as considered in Section 3.2.1. These "coordinate wise reconstruction" strategies are similar to those typically employed in estimation [86], i.e. "clipping" statistics and adding appropriately scaled noise.

Where the strategies differ from estimation, is firstly in the dimensionality of the transcripts. There is interplay between the optimal number of coordinates that a transcript contains information on and the severity of the privacy constraint. When $\epsilon \lesssim 1/\sqrt{n}$, the cost of transmitting information on each of the observations is very costly. To mitigate this, the rate optimal strategies in this ϵ regime transmit no more than a single coordinate of (a linear transformation of) the data. Laplacian noise is added to the clipped coordinate in order to obtain $(\epsilon, 0)$-differentially privacy guarantee, resulting in the $(\epsilon, 0)$-differentially private protocols. Such a regime or optimal strategy is not observed in the equivalent estimation problem, where it seems always optimal to transmit information on all coordinates, as described by the results Section 2.5.6. The shared randomness test corresponding to this strategy shall be denoted by T^ϵ_{II} and its construction is given in Section 3.2.2.1, whereas the local randomness counterpart, $T^\epsilon_{\mathrm{III}}$, shall be described in Section 3.2.2.2.

When $\epsilon \gtrsim 1/\sqrt{n}$, strategies that communicate information on more than one coordinate (of a linear transformation) of the data become viable, in the sense that they perform equally well or better than the ones that communicate just one coordinate. For these strategies, we shall employ Gaussian noise, which scales better in dimension than the Laplacian noise. The Gaussian mechanism results in (ϵ, δ)-differentially private protocols. The corresponding shared randomness test shall be denoted by $T_{\mathrm{II}}^{\epsilon,\delta}$, its local randomness counterpart by $T_{\mathrm{III}}^{\epsilon,\delta}$, which are constructed in Sections 3.2.2.3 and 3.2.2.4, respectively. The optimal number of coordinates transmitted / dimensionality of the transcript depends on the privacy budget ϵ, n and d. The choice of δ could be as small as $(nm)^{-p}$, where $p \geqslant 1$ is a constant; we opt for the rather arbitrary choice of $\delta \asymp (nm)^{-2}$.

3.2.2.1 Pure differential privacy using shared randomness in the "in-between regime"

We now bring our attention to constructing a shared randomness distributed ϵ-DP testing protocol $T_{\mathrm{II}}^{\epsilon}$ that is rate-optimal up to a logarithmic factor whenever $\frac{\sqrt{d}}{\sqrt{mn}} \leqslant \epsilon \leqslant 1/\sqrt{n}$.

Let U denote a draw from the Haar measure on $\mathbb{R}^{d\times d}$. For machine $j = 1, \ldots, m$, generate

$$Y^{(j)} \equiv Y^{(j)}(X^{(j)}, U) = \frac{\epsilon}{2\tau} \sum_{i=1}^{n} [(UX_i^{(j)})_1]_{-\tau}^{\tau} + W^{(j)},$$

where $W^{(j)}$ is independent centered Laplace noise with scale parameter 1. We also recall the notation $(v)_k$, which denotes the projection of a vector $v \in \mathbb{R}^d$ onto the k-coordinate. The map $x \mapsto \frac{1}{2\tau}\sum_{i=1}^{n}[(ux_i^{(j)})_1]_{-\tau}^{\tau}$ has sensitivity 1 for any $u \in \mathbb{R}^{d\times d}$, which makes $Y^{(j)}$ ϵ-differentially private by Lemma 3.5.

In contrast to the multiple clippings used in the test of Section 3.2.1, we consider a single level of clipping:

$$\tau := \tilde{\kappa}_\alpha \sqrt{\log(1 + dmn)}. \tag{3.32}$$

Using these transcripts, the central machine computes the test

$$T_{\mathrm{II}}^{\epsilon} = \mathbb{1}\left\{ \left(\frac{1}{\sqrt{m}} \sum_{j=1}^{m} Y^{(j)} \right)^2 - 2 - \frac{n\epsilon^2}{4\tau^2} \geqslant \kappa_\alpha \sqrt{n\epsilon^2 \vee 1} \right\}.$$

Applying Lemma 3.23 with $\gamma = \epsilon/(2\tau)$ and $L = 1$, choosing $\kappa_\alpha > 0$ and $\tilde{\kappa}_\alpha > 0$ large enough yields that $\mathbb{P}_0 T_{\mathrm{II}}^{\epsilon} \leqslant \alpha/2$. Furthermore, this choice of γ reduces the condition of (3.67) to

$$\left(\frac{d}{mn\|f\|_2}\right) \vee \left(\frac{d\tilde{\kappa}_\alpha^2 \Lambda_{d,n,m}}{mn^2\epsilon^2\|f\|_2^2}\right) \vee \left(\frac{\kappa_\alpha^2 d^2}{m^2 n^2 \|f\|_2^4}\right) \vee \left(\frac{\kappa_\alpha^2 \tilde{\kappa}_\alpha^4 \Lambda_{d,n,m}^2}{m^2 n^4 \epsilon^4 \|f\|_2^4}\right) \leqslant c_\alpha, \tag{3.33}$$

where $\Lambda_{d,n,m} := \log(1 + dnm)$. Since $n\epsilon^2 \lesssim 1$, the condition

$$\|f\|_2^2 \geqslant C_\alpha \frac{d}{\log(1 + dmn)mn^2\epsilon^2} \tag{3.34}$$

for $C_\alpha > 0$ large enough yields that the maximum of (3.33) is taken in the last argument, which is in turn bounded by $\kappa_\alpha^2 \tilde{\kappa}_\alpha^4 / C_\alpha^2$. We obtain that the Type II error condition of Lemma 3.23 is satisfied and $\mathbb{P}_f(1 - T_{\mathrm{II}}^\epsilon) \leqslant \alpha/2$. In conclusion, we obtain the following result.

Lemma 3.12. *Let $\frac{\sqrt{d}}{\sqrt{mn}} \leqslant \epsilon \leqslant 1/\sqrt{n}$. There exists a distributed $(\epsilon, 0)$-differentially private testing protocol T_{II}^ϵ such that T_{II}^ϵ is of level α and has Type II error probability $\mathbb{P}_f(1 - T_{II}^\epsilon) \leqslant \alpha$ whenever $f \in H_\rho$ satisfies (3.67) for a constant $C_\alpha > 0$ depending only on α.*

3.2.2.2 Pure differential privacy strategy using only local randomness in the "in-between regime"

In this section, we shall prove the following lemma by constructing a distributed testing protocol $T_{\mathrm{III}}^\epsilon$ that is ϵ-differentially private, uses only private randomness and attains the lower bound rate of Theorem 2.4 up to log-factors whenever $\sqrt{n}\epsilon \leqslant 1$ and $\epsilon \geqslant \frac{d}{\sqrt{mn}}$.

For $\tau > 0$, $j \in [m]$ and $l = 1, \ldots, d$, consider transcripts of the form

$$Y_l^{(j)}(X^{(j)}) = \frac{\epsilon}{2L\tau} \sum_{i=1}^n [(X_i^{(j)})_l]_{-\tau}^\tau + W_l^{(j)}, \tag{3.35}$$

where the $W_l^{(j)}$'s are i.i.d. Laplace noise on the line with scale parameter 1 and $L \in \mathbb{N}$ such that $d \geqslant L$. Since $x \mapsto \frac{\epsilon}{2L\tau} \sum_{i=1}^n [(x_i^{(j)})_l]_{-\tau}^\tau$ has sensitivity ϵ/L for $l = 1, \ldots, L$, releasing

$$Y^{(j)}(X^{(j)}) = (Y_{i_1}^{(j)}(X^{(j)}), \ldots, Y_{i_L}^{(j)}(X^{(j)}))$$

for one $i_l \in [d]$ satisfies the ϵ-DP guarantee by Lemma 3.5 as that $Y^{(j)}$ has L_1-sensitivity 1.

The clipping $\tau := \tilde{\kappa}_\alpha \sqrt{\log(1 + dmn)}$ is taken similarly to that of T_{II}^ϵ in the previous section, where $\tilde{\kappa}_\alpha > 0$ is a constant depending on the desired level of the test only. This assures that "typical" observations under the null hypothesis are within the clipping, whilst only in rare cases outliers are required to be clipped. This clipping is the cause of the log-optimality of the testing procedure: with significantly more technical effort, we believe it can be shown that a large enough constant clipping attains the optimal rate.

The test statistic $x \mapsto Y^{(j)}(x)$ requires the $1/L$ rescaling to have sufficiently bounded L_1-sensitivity and choosing L too large means a possible loss of power. An approach in this case is to divide each machine over the d coordinates as uniformly as possible.

That is to say, to partition the coordinates $\{1,\dots,d\}$ into approximately d/L sets of size L. The machines are then equally divided over each of these partitions and communicate the sum of clipped $X^{(j)}$ coefficients corresponding to their partition. More formally, such a strategy entails taking sets $\mathcal{I}_l \subset \{1,\dots,m\}$ such that $|\mathcal{I}_l| = \lfloor \frac{mL}{d} \rfloor$ and each $j \in \{1,\dots,m\}$ is in $\mathcal{I}_l$ for L different indexes $l \in \{1,\dots,d\}$. For $l = 1,\dots,d$ and $j \in \mathcal{I}_l$, generate the transcripts according to (3.35). Interestingly, the optimal choice of L turns out to be $\lceil d/m \rceil$, which entails L being of constant order for the regime where $d/\sqrt{mn} \leqslant \sqrt{n}\epsilon \leqslant 1$ (as this implies that $m \geqslant d^2$). In other words, the optimal rate in this regime is achieved by each machine communicating information about just one (or an $O(1)$ selection) of the d coordinates. The information gained by communicating more than a constant number of coordinates is not worth the increased noise needed to guarantee differential privacy.

As a test, the central machine computes

$$T^{\epsilon}_{\mathrm{III}} = \mathbb{1}\left\{ \frac{1}{\sqrt{d}} \sum_{k=1}^{d} \left[\left(\frac{1}{\sqrt{|\mathcal{J}_k|}} \sum_{j \in \mathcal{J}_k} Y_k^{(j)} \right)^2 - \frac{n\epsilon^2}{4L^2\tau^2} - 2 \right] \geqslant \kappa_\alpha \tau \right\},$$

which, by applying Lemma 3.24 with $\gamma = \frac{\epsilon}{2L\tau}$, satisfies $\mathbb{P}_0 T^{\epsilon}_{\mathrm{III}} \leqslant \alpha/2$ for $\tilde{\kappa}_\alpha, \kappa_\alpha > 0$ large enough.

Since $\epsilon \leqslant n^{-1/2}$, we have $\gamma^2 \leqslant (\log(dnm)n)^{-1}$. Combining this with the fact that $L \asymp 1$, $\epsilon \lesssim 1/\sqrt{n}$ and the required condition (3.71) of the second statement of Lemma 3.24 reduces to showing that, for some constant $c_\alpha > 0$,

$$\frac{d}{mn\|f\|_2^2} \vee \frac{\log(1+dnm)d^2}{m^2n^2\epsilon^2\|f\|_2^2} \vee \frac{\log^2(1+dnm)d^3}{m^2n^4\epsilon^4\|f\|_2^4} \leqslant c_\alpha. \tag{3.36}$$

Whenever f satisfies (3.37) below, the first term is of the order $1/C_\alpha$ (again using $\epsilon \lesssim 1/\sqrt{n}$). Furthermore, using $m \gtrsim d^2$ yields that the second term in the maximum is of the order $1/(C_\alpha\sqrt{m})$ and for the third term we obtain

$$\frac{\log^2(1+dnm)d^3}{m^2n^4\epsilon^4\|f\|_2^4} \leqslant \frac{1}{C_\alpha^2},$$

which can be made arbitrarily small for per choice of $C_\alpha > 0$. The second statement of Lemma 3.24 consequently yields that $\mathbb{P}_f(1 - T^{\epsilon}_{\mathrm{III}}) \leqslant \alpha/2$. We consequently have proven Lemma 3.13 below.

Lemma 3.13. *Take $\alpha \in (0,1)$. Whenever $d/\sqrt{mn} \leqslant \sqrt{n}\epsilon \leqslant 1$, the distributed $(\epsilon,0)$-differentially private testing protocol T^{ϵ}_{III} of level α has Type II error $\mathbb{P}_f(1-T) \leqslant \alpha$ whenever*

$$\|f\|_2^2 \geqslant C_\alpha \frac{d^{3/2}}{\log(1+dmn)mn^2\epsilon^2}, \tag{3.37}$$

for a constant $C_\alpha > 0$ depending only on α.

3.2.2.3 Shared randomness in the "large" ϵ regime

When $\epsilon \gtrsim n^{-1/2}$, better rates can be achieved by communicating more than a constant number of clipped coordinates with added noise. In order to obtain a rate matching the lower bound in Theorem 2.4 (up to a log factor), the noise used for this strategy is Gaussian. The Gaussian noise requires the sensitivity to be small in L_2-norm, which means that the scaling of the Gaussian noise has a better dimensional dependency. This also means that the protocol is (ϵ, δ)-differentially private. For this strategy to attain the optimal rate, one can take $\delta \lesssim (nmd)^{-p}$ for any fixed $p > 1$ is deemed fit, so arguably the "impure" differentially private protocol is still "close" to being ϵ-differentially private, especially in terms a plausible deniability guarantee (see (1.6)). The protocol below does not require $\epsilon \geqslant n^{-1/2}$. The approach leads to a test attaining that attains the (log-optimal) rate and Type I and Type II guarantees laid out in the lemma below.

Lemma 3.14. *Let $(mn)^{-1} \leqslant \epsilon \leqslant 1$. There exists a distributed (ϵ, δ)-differentially private testing protocol $T_{II}^{\epsilon,\delta}$, with level α and has corresponding Type II error probability $\mathbb{P}_f(1 - T_{II}^{\epsilon,\delta}) \leqslant \alpha$ whenever*

$$\|f\|_2^2 \geqslant C_\alpha \frac{d \log(1/\delta) \log^2(1 + dnm)}{mn\sqrt{n\epsilon^2 \wedge d}\sqrt{n\epsilon^2 \wedge 1}} \tag{3.38}$$

for a constant $C_\alpha > 0$ depending only on α.

Consider for $L = \lceil n\epsilon^2 \wedge d \rceil$, $l = 1, \ldots, L$ and $j = 1, \ldots, m$ the transcripts

$$Y_l^{(j)} | (X^{(j)}, U) = \gamma_{\epsilon,\tau,n,m} \sum_{i=1}^{n} [(UX_i^{(j)})_l]_{-\tau}^{\tau} + W_l^{(j)}, \tag{3.39}$$

with $\gamma_{\epsilon,\tau,n,m} = \frac{\epsilon}{6\sqrt{2L \log(1/\delta)}\tau}$, $\tau = \tilde{\kappa}_\alpha \sqrt{\log(1 + dmn)}$, U a random rotation (drawn uniformly) and $(W_l^{(j)})_{j,l}$ i.i.d. centered standard Gaussian noise. For any rotation $u \in \mathbb{R}^{d \times d}$,

$$\sup_{\check{x} \in (\mathbb{R}^d)^n : \mathrm{d}_H(x,\check{x}) \leqslant 1} \left\| \gamma_{\epsilon,\tau,n,m} \left(\sum_{i=1}^{n} [(ux_i)_l]_{-\tau}^{\tau} - \sum_{i=1}^{n} [(u\check{x}_i)_l]_{-\tau}^{\tau} \right)_{l : j \in \mathcal{J}_l} \right\|_2 \leqslant$$
$$\frac{\epsilon}{2\sqrt{L}\tau} \sqrt{\sum_{l=1}^{L} \left(\sup_{\check{x} \in (\mathbb{R}^d)^n : \mathrm{d}_H(x,\check{x}) \leqslant 1} [(x_i)_l]_{-\tau}^{\tau} - [(\check{x}_i)_l]_{-\tau}^{\tau} \right)^2} \leqslant 1,$$

so by an application of Lemma 3.6, the transcript $Y^{(j)} := (Y_l^{(j)})_{l \in [L]}$ is (ϵ, δ)-differentially private. The test

$$T_{\mathrm{II}}^{\epsilon,\delta} = \mathbb{1} \left\{ \frac{1}{\sqrt{d}} \sum_{l=1}^{d} \left[\left(\frac{1}{\sqrt{m}} \sum_{j=1}^{m} Y_l^{(j)} \right)^2 - n\gamma_{\epsilon,\tau,n,m}^2 - 1 \right] \geqslant \kappa_\alpha \left(n\gamma_{\epsilon,\tau,n,m}^2 \vee 1 \right) \right\} \tag{3.40}$$

satisfies $\mathbb{P}_0\varphi \leqslant \alpha/2$ by Lemma 3.23 applied with $\gamma = \gamma_{\epsilon,\tau,n,m}$, for $\kappa_\alpha > 0$ large enough. Plugging in $\gamma_{\epsilon,\tau,n,m}$, we see that (3.67) is satisfied whenever the quantity

$$\left(\frac{d}{mnL\|f\|_2^2}\right) \vee \left(\frac{d\Lambda_{d,n,m}\tilde{\kappa}_\alpha^2}{mn^2\epsilon^2\|f\|_2^2}\right) \vee \left(\frac{\kappa_\alpha^2 d^2}{m^2n^2L\|f\|_2^4}\right) \vee \left(\frac{\kappa_\alpha^2\tilde{\kappa}_\alpha^4\Lambda_{d,n,m}^2 d^2 L}{m^2n^4\epsilon^4\|f\|_2^4}\right), \tag{3.41}$$

where $\Lambda_{d,n,m} := \log(1 + dnm)\log(1/\delta)$ can be made arbitrarily small per choice of $C_\alpha > 0$ in (3.38). When $\epsilon \lesssim 1/\sqrt{n}$, the same steps used to prove Lemma 3.12 (see (3.33)) show that the condition is satisfied. For $1/\sqrt{n} \lesssim \epsilon \lesssim d$, $L \asymp n\epsilon^2 \wedge d \gtrsim 1$ and (3.41) reduces further to

$$\frac{\kappa_\alpha^2\tilde{\kappa}_\alpha^4\Lambda_{d,n,m}^2 d^2}{m^2n^3\epsilon^2\|f\|_2^4} \lesssim c_\alpha.$$

This can also be seen to hold whenever f is such that (3.38). Lastly, when $n\epsilon^2 \gtrsim d$, (3.33) reduces to $\kappa_\alpha^2 d^2/(m^2n^2L\|f\|_2^4) \lesssim c_\alpha$, which also holds for $C_\alpha > 0$ large enough for f satisfying (3.38). Consequently, we obtain that $\mathbb{P}_f(1 - T_{\mathrm{II}}^{\epsilon,\delta}) \leqslant \alpha/2$ for C_α large enough, as desired.

3.2.2.4 Private randomness protocol in the "large" ϵ regime

Similarly to the case of shared randomness, we can combine the coordinate wise (local randomness) approach with Gaussian noise to allow for a larger amount of coordinates to be sent. Similarly to the shared randomness Gaussian mechanism, the protocol below does not require $\epsilon \geqslant n^{-1/2}$, but the rate that is attained is a factor $\log(1/\delta)$ larger. We shall take $\delta = (dnm)^{-p}$ with $p = 2$, but p can be taken any constant larger than one, which only affects the constant $C_\alpha > 0$ in (3.42).

Lemma 3.15. *Let $(mn)^{-1} \leqslant \epsilon \leqslant 1$ and $\delta = (dnm)^{-2}$. There exists a distributed (ϵ,δ)-differentially private testing protocol $T_{III}^{\epsilon,\delta}$ such that $T_{III}^{\epsilon,\delta}$ is of level α and has Type II error probability $\mathbb{P}_f(1 - T_{III}^{\epsilon,\delta}) \leqslant \alpha$ whenever*

$$\|f\|_2^2 \geqslant C_\alpha \frac{d^{3/2}\log(1/\delta)\log(1 + dmn)}{mn(n\epsilon^2 \wedge d)} \tag{3.42}$$

for a constant $C_\alpha > 0$ depending only on α.

Let $L = \lceil n\epsilon^2 \wedge d \rceil$ and take sets $\mathcal{I}_l \subset [m]$ such that $|\mathcal{I}_l| = \lfloor \frac{mL}{d} \rfloor$ and each $j \in \{1,\dots,m\}$ is in $\mathcal{I}_l$ for L different indexes $l \in \{1,\dots,d\}$. For $l \in [d]$, $j \in \mathcal{I}_l$, generate the transcripts according to

$$Y_l^{(j)}|X^{(j)} \equiv Y_l^{(j)}(X^{(j)}) = \frac{\epsilon}{6\sqrt{2L\log(dmn)}\tau}\sum_{i=1}^n [(X_i^{(j)})_l]_{-\tau}^{\tau} + W_l^{(j)} \tag{3.43}$$

with $\tau = \tilde{\kappa}_\alpha \sqrt{\log(1+dmn)}$ and $(W_l^{(j)})_{j,l}$ i.i.d. standard Gaussian noise. The clipped and $\frac{1}{2\sqrt{L}\tau}$-rescaled sums have at most L_2-sensitivity less than or equal to one:

$$\sup_{\breve{x}\in(\mathbb{R}^d)^n:\mathrm{d}_H(x,\breve{x})\leqslant 1} \frac{1}{2\sqrt{L}\tau} \left\| \left(\sum_{i=1}^n [(x_i)_l]_{-\tau}^{\tau} - \sum_{i=1}^n [(\breve{x}_i)_l]_{-\tau}^{\tau} \right)_{l:j\in\mathcal{J}_l} \right\|_2 \leqslant \tag{3.44}$$
$$\frac{1}{2\sqrt{L}\tau} \sqrt{\sum_{l=1}^L \left(\sup_{\breve{x}\in(\mathbb{R}^d)^n:\mathrm{d}_H(x,\breve{x})\leqslant 1} [(x_i)_l]_{-\tau}^{\tau} - [(\breve{x}_i)_l]_{-\tau}^{\tau} \right)^2} \leqslant 1.$$

Consequently, the transcript $Y^{(j)} := (Y_l^{(j)})_{l:j\in\mathcal{J}_l}$ is $(\epsilon, (nm)^{-2})$-differentially private as a result of Lemma 3.6. The test

$$T_{\mathrm{III}}^{\epsilon,\delta} = \mathbb{1}\left\{ \frac{1}{\sqrt{d}} \sum_{l=1}^d \left[\left(\frac{1}{\sqrt{|\mathcal{J}_l|}} \sum_{j\in\mathcal{J}_l} Y_l^{(j)} \right)^2 - \frac{n\epsilon^2}{4L\tau^2} - 1 \right] \geqslant \kappa_\alpha \left(\frac{n\epsilon^2}{4L\tau^2} \vee 1 \right) \right\} \tag{3.45}$$

satisfies $\mathbb{P}_0 T_{\mathrm{III}}^{\epsilon,\delta} \leqslant \alpha$ by Lemma 3.24 applied with $\gamma = \epsilon/(6\sqrt{2L\log(1/\delta)}\tau)$ whenever $\tilde{\kappa}_\alpha > 0$ and $\kappa_\alpha > 0$ are chosen large enough. In order to fulfill the condition of (3.71) in the same lemma, it suffices that

$$\frac{d}{mn(n\epsilon^2 \wedge d)\|f\|_2^2} \vee \frac{d\Lambda_{m,n,d}\tilde{\kappa}_\alpha^2}{mn^2\epsilon^2\|f\|_2^2} \vee \frac{\kappa_\alpha^2 d^3}{m^2n^2(n\epsilon^2\wedge d)^2\|f\|_2^4} \vee \frac{d^3\Lambda_{m,n,d}^2\tilde{\kappa}_\alpha^4\kappa_\alpha^2}{m^2n^4\epsilon^4\|f\|_2^4} \leqslant c_\alpha$$

where $\Lambda_{m,n,d} = \log(1/\delta)\log(1+dmn)$. When $n\epsilon^2 \lesssim d$, the maximum is taken in the fourth argument, which can be seen to satisfy the inequality for $C_\alpha > 0$ whenever f satisfies (3.42). Whenever $n\epsilon^2 \gtrsim d$, the maximum is taken in the third argument, which means the inequality is also satisfied in this case.

3.3 On the benefit of shared randomness in distributed decision problems

In this chapter and Chapter 2, we encounter the (in some cases strictly) better performance of shared randomness distributed testing protocols under privacy and bandwidth constraints. What drives this intriguing phenomenon? We will delve into this question in this section.

We approach the phenomenon from two different perspectives. In Section 3.3.1 below, we provide one approach, which is to study it abstractly in the framework of statistical decision theory as outlined in Section 1.2. Here, we shine a light on how it relates to the risk formulation of a statistical decision problem. In addition, we show that shared randomness offers no benefit in the distributed estimation settings considered in this thesis and offer some contemplation on what separates these distributed estimation problems from the distributed testing problem. As a second approach, in Section 3.3.2,

we explore an example of a simple minimax detection game which bears parallels with distributed hypothesis testing under communication constraints.

Both of these approaches provide insight into the potential benefit of access to shared randomness. For the reader who is further interested, we refer to Chapter 3 in [171] for a broad treatment of this phenomenon, in the context of how many bits two or more parties need to exchange to compute a specific function whose inputs are distributed among the parties.

3.3.1 Shared randomness for general decision problems

In this section, we will contemplate why access to shared randomness can have demonstrable benefit in distributed testing under bandwidth and privacy constraints, whilst it demonstrably does not in their minimax distributed estimation counterparts. That is, the distributed estimation problem under bandwidth constraints considered in Section 2.1.1 and the distributed estimation problem under privacy constraints studied in Section 2.4 exhibit the same minimax rate for shared randomness distributed protocols as for local randomness protocols. This is not explicitly established in those respective sections, so we aim to do so here in a general setting, in the form of Theorem 3.3 below. Whilst the result is straightforward and probably well known among experts, we did not find a proof of the following explicit statement in the literature. Related to the result, we then highlight how the specification of the minimax risk relates to the possibility of performance benefit from access to shared randomness.

Consider a sequence of models $\mathcal{P}_\nu = \{P_{f,\nu} : f \in \mathcal{F}\}$ defined on a measurable space $(\mathcal{X}, \mathscr{X})$, indexed by a measurable space $\mathcal{F}$, a decision space $(\mathcal{D}, \mathscr{D})$ and a sequence of (measurable) loss functions $\ell_\nu : \mathcal{D} \times \mathcal{F} \to [0, \infty)$. We recall the distributed setting of Section 1.2, in which $j = 1, \dots, m$ machines each observe an independent draw $X^{(j)}$ from $P_{f,\nu}$. Denote the full data as $X := (X^{(1)}, \dots, X^{(m)})$. Consider distributed decision protocols $\hat{f} \equiv \hat{f}_\nu \equiv \{\hat{f}, \{K^j\}_{j=1}^m, \mathbb{P}^U\}$ where $K^j : \mathscr{Y}^{(j)} \times \mathcal{X} \to [0,1]$ is a Markov kernel for $j = 1, \dots, m$ and $\hat{f} : \otimes_{j=1}^m \mathcal{Y}^{(j)} \to \mathcal{D}$ is a measurable function, for measurable spaces $(\mathcal{Y}^{(j)}, \mathscr{Y}^{(j)})$. Let $\mathscr{J} \equiv \mathscr{J}_\nu$ denote the class of all such distributed protocols such that $\{K^j\}_{j=1}^m$ satisfy either a b-bit bandwidth constraint or a (ϵ, δ)-differential privacy constraint and let $\mathscr{K} \subset \mathscr{J}$ denote the class of distributed protocols where $U \sim \mathbb{P}^U$ is degenerate (i.e. the subset of local randomness protocols). Let $\mathbb{E}_f \equiv \mathbb{E}_{f,\nu}$ denote expectation with respect to the joint distribution of Y with the data which is given by

$$P_f \mathbb{P}^U \otimes_{j=1}^m K^j(\cdot|X^{(j)}, U) = \mathbb{P}^U P_f \otimes_{j=1}^m K^j(\cdot|X^{(j)}, U),$$

where the interchange of $P_f \equiv P_{f,\nu}^m$ and $\mathbb{P}^U$ follows from the independence of the shared randomness and the data. We shall consider the risk for the loss function ℓ over the model $\mathcal{P}$;

$$\sup_{f \in \mathcal{F}} \mathbb{E}_{f,\nu}\, \ell_\nu(\hat{f}(Y), f). \tag{3.46}$$

This setup encapsulates the estimation and testing settings considered in the thesis. The following theorem shows that, if there exists a "sufficiently adversarial prior", there is no benefit to having access to shared randomness.

Theorem 3.3. *Suppose that, for some sequence of distributions $\pi \equiv \pi_\nu$ on $\mathcal{F}$ and sequences $\gamma_\nu, \varrho_\nu > 0$, it holds that*

$$\inf_{\hat{f}\in\mathscr{K}} \sup_{f\in\mathcal{F}} \mathbb{E}_f\, \ell_\nu(\hat{f}(Y), f) \leqslant \gamma_\nu \quad \textit{and} \quad \inf_{\hat{f}\in\mathscr{K}} \int \mathbb{E}_f\, \ell_\nu(\hat{f}(Y), f) d\pi(f) \geqslant \varrho_\nu.$$

Then,

$$\varrho_\nu \leqslant \inf_{\hat{f}\in\mathscr{J}} \sup_{f\in\mathcal{F}} \mathbb{E}_{f,\nu}\, \ell_\nu(\hat{f}(Y), f) \leqslant \gamma_\nu. \tag{3.47}$$

In particular, if $\gamma_\nu \asymp \rho_\nu$,

$$\inf_{\hat{f}\in\mathscr{J}} \sup_{f\in\mathcal{F}} \mathbb{E}_f\, \ell_\nu(\hat{f}(Y), f) \asymp \inf_{\hat{f}\in\mathscr{K}} \sup_{f\in\mathcal{F}} \mathbb{E}_f\, \ell_\nu(\hat{f}(Y), f).$$

Proof. The right-hand side inequality of (3.47) follows from the fact that $\mathscr{K} \subset \mathscr{J}$, which means

$$\inf_{\hat{f}\in\mathscr{J}} \sup_{f\in\mathcal{F}} \mathbb{E}_f\, \ell_\nu(\hat{f}(Y), f) \leqslant \inf_{\hat{f}\in\mathscr{K}} \sup_{f\in\mathcal{F}} \mathbb{E}_f\, \ell_\nu(\hat{f}(Y), f).$$

To obtain the other inequality, consider an arbitrary shared randomness distributed protocol $\{\hat{f}, \{K^j\}_{j=1}^m, \mathbb{P}^U\} \in \mathscr{J}$. By the independence of U with the data X, it holds for any $f \in \mathcal{F}$ that

$$\mathbb{E}_f\, \ell_\nu(\hat{f}(Y), f) = \int \mathbb{E}_f^{Y|U=u}\, \ell_\nu(\hat{f}(Y), f) d\mathbb{P}^U(u).$$

The triplet $\{\hat{f}, \{K^j\}_{j=1}^m, \delta_u\}$, where δ_u denotes the Dirac measure at u, is a local randomness distributed testing protocol, noting that $(A, x) \mapsto K^j(A|x, u)$ for any u in the sample space of U indeed defines a local randomness Markov kernel satisfying its original constraint (i.e. its bandwidth or differential privacy constraint). Consequently, we have that

$$\begin{aligned}
\sup_{f\in\mathcal{F}} \mathbb{E}_f\, \ell_\nu(\hat{f}(Y), f) &\geqslant \int \mathbb{E}_{f,\nu}^Y\, \ell_\nu(\hat{f}'(Y), f) d\pi(f) \\
&= \int\int \mathbb{E}_f^{Y|U=u}\, \ell_\nu(\hat{f}'(Y), f) d\pi(f) d\mathbb{P}^U(u) \\
&\geqslant \int \inf_{\hat{f}\in\mathscr{K}} \mathbb{E}_{f,\nu} \ell_\nu(\hat{f}'(Y), f) d\pi(f) \geqslant \varrho_\nu.
\end{aligned}$$

The conclusion of the theorem follows because $\{\hat{f}, \{K^j\}_{j=1}^m, \mathbb{P}^U\} \in \mathscr{J}$ was taken arbitrarily. □

Essentially, the theorem states that the "lack of benefit" of shared randomness in e.g. the estimation problems considered in Chapter 2 stems from the fact that the prior distribution used to prove the (Bayes risk) lower bounds does not depend on the Markov kernels $\{K^j\}_{j=1}^m$ constituting the distributed protocol. That is, in the language of the theorem above, π is not contingent on $\{K^j\}_{j=1}^m$. In the distributed testing problem, sharper lower bounds are obtained by taking π adversarial with respect to the local randomness protocol (i.e. $\{K^j\}_{j=1}^m$), see e.g. the distributed Le Cam bound of Lemma 2.28.

To illustrate this further, suppose that

$$\inf_{\hat{f}\in\mathscr{K}} \sup_{\pi} \int \mathbb{E}_f\, \ell_\nu(\hat{f}(Y), f) d\pi(f) \asymp \sup_{\pi} \inf_{\hat{f}\in\mathscr{K}} \int \mathbb{E}_f\, \ell_\nu(\hat{f}(Y), f) d\pi(f), \tag{3.48}$$

where the supremum is taken over all distributions π on $\mathcal{F}$. By an argument similar to that of the above theorem, it follows that

$$\inf_{\hat{f}\in\mathscr{J}} \sup_{f\in\mathcal{F}} \mathbb{E}_f\, \ell_\nu(\hat{f}(Y), f) \asymp \inf_{\hat{f}\in\mathscr{K}} \sup_{f\in\mathcal{F}} \mathbb{E}_f\, \ell_\nu(\hat{f}(Y), f).$$

That is, if the minimax problem in Bayes risk satisfies an (asymptotic) "minimax theorem" (i.e. (3.48) holds), shared randomness offers no improved minimax risk, up to a constant factor. From this observation, one can also conclude that only in problems where one considers "nature" to be possibly adversarial specifically to the choice of kernel one might have benefit in having access to shared randomness.

On a similar note, if one assumes "nature" to be adversarial to the shared randomness outcome, there is no benefit to having access to it. That is, if one defines the minimax risk as

$$\int \sup_{f\in\mathcal{F}} \mathbb{E}_f^{Y|U=u}\, \ell_\nu(\hat{f}(Y), f)\, d\mathbb{P}^U(u), \tag{3.49}$$

there is no benefit to having shared randomness. Indeed, through similar reasoning as in the proof of Theorem 3.3, the infimum over $\mathscr{J}$ of the above expression equals the infimum over $\mathscr{K}$.

Which risk formulation is appropriate to one's decision problem is contingent on what assumptions one is prepared to make about the "opposing forces" present. Fundamentally, the risk of formulation of (3.46) takes a viewpoint that assumes that, although nature may act in opposition to our selected protocol, it does not conspire against the source of shared randomness. This assumption appears sound; it is usually presumed that "nature" is not adversarial towards individual random occurrences. So at least in the scientific study of natural phenomena, this seems a reasonable formulation of minimax risk to work with. For example, in the context of testing a hypothesis concerning a natural phenomenon, the supremum is typically taken to *guarantee* power against an entire class of alternatives, not because nature is adversarial to one's testing protocol.

Yet, in a more contentious context, the utilization of shared randomness could be a genuine issue. For example, if servers coordinate to detect the presence of a hacker, using shared randomness might form a weakness in the sense that if the source of randomness is leaked, this could be exploited by the hacker. Similarly, in a military context, an enemy who gains access to these shared random elements could eavesdrop or even disrupt the coordinated efforts of the military units. Such contemplations are further exemplified in the detection game of the next section.

3.3.2 A simple minimax detection game

Consider the following detection game:

- It is the night of December 24th, and Santa Claus is about to visit the house of Alice and Bob.
- When Santa Claus arrives, he knocks on either the front door, a window on the side, or a window at the back of the house, leaving presents there.
- Alice and Bob are aware of this and position themselves at different windows on two floors of the house, trying to spot Santa Claus.
- The house has a first floor with a window next to the front door and a window to the side of the house, and a top floor with a side window and a window at the back of the house.

Santa Claus wants to remain undetected and somehow is aware of any strategy that Alice and Bob concoct beforehand. For example, if Alice and Bob decide to go to the front- and side window on the first floor, Santa Claus will always appear at the back of the house, remaining undetected. Since neither floor has windows covering all three sides of the house, Alice and Bob must split up to have a positive probability of detecting Santa Claus. Assume that Alice stays at the first floor, while Bob goes upstairs. When Alice and Bob are on different floors, they are no longer able to communicate.

The scenario fits into the framework of a minimax distributed detection game. Alice and Bob choose a strategy, knowing that Santa Claus will pick the best strategy against their chosen strategy. Let A the event in which Alice goes to the side window and let B be the even that Bob goes to the side window. The probability of detecting Santa Claus can be expressed as:

$$\min\{1 - \Pr(A), 1 - \Pr(B), \Pr(A \cup B)\}.$$

If Alice and Bob do not randomize their strategy (i.e., $\Pr(A), \Pr(B) \in \{0, 1\}$), Santa Claus will certainly not be detected. So, Alice and Bob will need to randomize their choice of window, since there are sides of the house to cover, with only two people.

Let us consider the case where Alice and Bob can choose A and B from sigma algebras Σ_a and Σ_b, respectively, to decide which window to guard. The minimax probability

of detection of Santa Claus for Alice and Bob is

$$\sup_{A\in\Sigma_a, B\in\Sigma_b} \min\{1-\Pr(A), 1-\Pr(B), \Pr(A\cup B)\}.$$

If Alice and Bob randomize their window choice using independent sources of randomness (with A and B being independent), the probability of the union of their choices can be expressed as:

$$\Pr(A\cup B) = \Pr(A) + \Pr(B) - \Pr(A\cap B) = \Pr(A) + \Pr(B) - \Pr(A)\Pr(B).$$

Thus, if Σ_a and Σ_b consist of independent events, the minimax problem for Alice and Bob is upper bounded by:

$$\max_{x,y\in[0,1]} \min\{1-x, 1-y, x+y-xy\},$$

which can be shown to equal $\frac{1}{2}(\sqrt{5}-1) = 0.61\ldots$.

Can a better outcome be achieved when Alice and Bob have access to a shared source of randomness? The answer is yes. An optimal strategy ensures that $P(A\cap B) = 0$, whilst also assuring $\Pr(A\cup B) = \Pr(A) + \Pr(B) = 2/3$. This is only possible if A and B are dependent, i.e. Alice and Bob use a shared source of randomness.

To exemplify such a strategy, suppose Alice and Bob possess two "entangled" coins, where the outcome (heads or tails) is always the same if the coins are flipped simultaneously. After flipping the shared randomness:

- If it lands on heads, Bob stands at the back window, and Alice rolls a die to determine whether she stands at the side window (if the number of eyes is less than 5) or the front window (otherwise).
- If it lands on tails, Alice stands at the front window, and Bob rolls a die to determine whether he stands at the side window (if the number of eyes is less than 5) or the back window (otherwise).

In this setup, $P(A\cap B) = 0$ and $P(A) = P(B) = 1/3$. Consequently, the probability of detecting Santa Claus under such a shared randomness strategy is $\min\{1-P(A), 1-P(B), P(A\cup B)\} = \frac{2}{3}$, which is strictly greater than the minimax detection probability with independent events in (3.3.2). So, a shared randomness strategy can achieve a higher detection probability of up to $\frac{2}{3}$ compared to a detection probability of approximately 0.61 when no shared randomness is available.

The intriguing aspect here is that there is no communication between Alice and Bob; what matters is the source of their randomness. The parallel between the scenario above and bandwidth constraint hypothesis testing is that in both cases, there are limited resources that need to "cover" multiple "locations". Under bandwidth constraints, we have a limited number of bits/machines that need to "cover" high

dimensional data adequately. Each of these dimensions are "locations" in which signal (Santa Claus in the scenario above) could "appear". If the number of bits is smaller than the dimension of the data (the number of "locations"), we can employ randomization to "divide" bits across the different "locations". When this randomness is shared, it is possible to improve coordination between the machines, allowing them to cover the "locations" more effectively.

In the privacy setting, the resource limitation originates from the fact that the more dimensions of the data are shared, the more noise must be injected. Typically, if we can get away with sharing something useful that is of lower dimensionality than the original data, this means less noise needs to be added. Using a shared source of randomness improves the coordination between the machines, effectively allowing the machines to cover more "locations" of the original data, whilst keeping the dimensionality of the object to which noise needs to be added limited.

Chapter acknowledgements: The quote from Section 3.1 was found in the blog of Maxim Raginsky. The quote originates from the travelogue [117].

3.4 Appendix

3.4.1 Lemmas for the upper bound theorems in the finite dimensional Gaussian mean model under bandwidth constraints

3.4.1.1 Proof of Lemma 3.1

We state a slightly extended version of Lemma 3.1.

Lemma 3.16. *Consider for $k, l \in \mathbb{N}$, $l \geqslant 2$, independent random variables $\{B_i^j : i = 1, \ldots, k,\ j = 1, \ldots, l\}$ with $B_i^j \sim Ber(p_i)$. If $p_i = 1/2$ for $i = 1, \ldots, k$, for each $\alpha \in (0, 1)$ there exists $\kappa_\alpha > 0$ such that*

$$Pr\left(\left|\frac{1}{\sqrt{k}l}\sum_{i=1}^{k}\left(\sum_{j=1}^{l}(B_i^j - \frac{1}{2})\right)^2 - \sqrt{k}/4\right| \geqslant \kappa_\alpha\right) \leqslant \alpha.$$

On the other hand, for arbitrary $c_{\alpha,n} > 0$,

$$\eta_{p,l,k} := \frac{l-1}{2\sqrt{k}}\sum_{i=1}^{k}\left(p_i - \frac{1}{2}\right)^2 \geqslant c_{\alpha,n}, \tag{3.50}$$

it holds that

$$Pr\left(\left|\frac{1}{\sqrt{k}l}\sum_{i=1}^{k}\left(\sum_{j=1}^{l}(B_i^j - \frac{1}{2})\right)^2 - \sqrt{k}/4\right| \leqslant c_{\alpha,n}\right) \leqslant \frac{1/2 + 16\eta_{p,l,k}/\sqrt{k}}{\eta_{p,l,k}^2}. \tag{3.51}$$

Proof. The left-hand side in the event having bounded variance: a straightforward computation (using that for $B_i^j \sim \mathrm{Bern}(p_i)$, the central fourth moment is $E(B_i^j - p_i)^4 = p_i(1-p_i)(1-3p_i(1-p_i)) \leqslant 1/16$ and $\mathrm{Var}(X) \leqslant EX^2$) yields

$$\begin{aligned}\mathbb{E}\Big[\frac{1}{\sqrt{kl}}\sum_{i=1}^{k}\Big(\sum_{j=1}^{l}(B_i^j - \frac{1}{2})\Big)^2 - \frac{\sqrt{k}}{4}\Big]^2 &= \frac{1}{kl^2}\sum_{i=1}^{k}\mathrm{Var}\Big[\Big(\sum_{j=1}^{l}(B_i^j - 1/2)\Big)^2\Big] \\ &\leqslant \frac{1}{l^2}\sum_{j=1}^{l}\mathbb{E}(B_i^j - 1/2)^4 + \frac{1}{l^2}\sum_{j=1}^{l}\big(\mathbb{E}(B_i^j - 1/2)^2\big)^2 \leqslant 1/8, \end{aligned} \tag{3.52}$$

after which Chebyshev's inequality yields the first statement.

We turn to the second statement. Adding and subtracting p_i and expanding the square, the left-hand side of the display in the lemma can be written as

$$\Pr\left(\left|\frac{1}{\sqrt{kl}}\sum_{i=1}^{k}\left(\sum_{j=1}^{l}B_i^j - lp_i\right)^2 - \mu_p + \frac{l-1}{\sqrt{k}}\sum_{i=1}^{k}\left(p_i - \frac{1}{2}\right)^2 + \zeta\right| \leqslant c_{\alpha,n}\right) \tag{3.53}$$

where

$$\mu_p := \frac{1}{\sqrt{k}}\sum_{i=1}^{k}p_i(1-p_i) \quad \text{and} \quad \zeta := \frac{2}{\sqrt{k}}\sum_{i=1}^{k}\left(p_i - \frac{1}{2}\right)\left(\sum_{j=1}^{l}B_i^j - lp_i\right).$$

The first term in the event of (3.53) has mean μ_p and variance (by the same computations as in (3.52))

$$\mathrm{Var}\Big[\frac{1}{\sqrt{kl}}\sum_{i=1}^{k}\Big(\sum_{j=1}^{l}B_i^j - lp_i\Big)^2\Big] = \frac{1}{kl^2}\sum_{i=1}^{k}\mathrm{Var}\Big[\Big(\sum_{j=1}^{l}B_i^j - lp_i\Big)^2\Big] \leqslant 1/8.$$

The term ζ has mean 0 and

$$\mathrm{Var}(\zeta) = \frac{4l}{k}\sum_{i=1}^{k}(p_i - \frac{1}{2})^2 p_i(1-p_i) \leqslant \frac{l}{k}\sum_{i=1}^{k}\left(p_i - \frac{1}{2}\right)^2.$$

Applying the reverse triangle inequality and condition (3.50), the probability in (3.53) is bounded from above by

$$\begin{aligned}&\Pr\Big[\Big|\frac{1}{\sqrt{kl}}\sum_{i=1}^{k}\Big(\sum_{j=1}^{l}B_i^j - lp_i\Big)^2 - \mu_p\Big| + |\zeta| \geqslant \frac{l-1}{2\sqrt{k}}\sum_{i=1}^{k}\Big(p_i - \frac{1}{2}\Big)^2\Big] \\ &\quad\leqslant \Pr\Big[\Big|\frac{1}{\sqrt{kl}}\sum_{i=1}^{k}\Big(\sum_{j=1}^{l}B_i^j - lp_i\Big)^2 - \mu_p\Big| \geqslant \eta_{p,l,k}/2\Big] + \Pr\Big[|\zeta| \geqslant \eta_{p,l,k}/2\Big] \\ &\quad\leqslant \frac{1/8}{(\eta_{p,l,k}/2)^2} + \frac{2lk^{-1/2}\eta_{p,l,k}/(l-1)}{(\eta_{p,l,k}/2)^2} \leqslant \frac{1/2 + 16\eta_{p,l,k}/\sqrt{k}}{\eta_{p,l,k}^2},\end{aligned}$$

where the last line follows by Chebyshev's inequality.

□

3.4.1.2 Proof of rate attainment of auxiliary local randomness tests T^1_{III} and T^2_{III}

Here, we provide the risk bounds for the partial tests T^1_{III} as defined in (3.9) and T^2_{III} as defined in (3.11), used in Section 3.1.3.

Lemma 3.17. *For any $\alpha \in (0,1)$ there exist constants $\kappa_\alpha, M_\alpha, C_\alpha > 0$ such that $\mathbb{E}_0 T^1_{III} \leqslant \alpha/2$. Furthermore, for $f \in H_\rho$ if $\rho^2 \geqslant C_\alpha \frac{d\sqrt{d}}{mn(d\wedge b)}$ and either $\frac{mb}{d\sqrt{d}} \geqslant M_\alpha$ or*

$$\sum\nolimits_{i\in\mathcal{J}^c} f_i^2 \geqslant \rho^2/2, \tag{3.54}$$

holds, where $\mathcal{J}$ was defined in (3.13)*, then*

$$\mathbb{E}_f(1 - T^1_{III}) \leqslant \alpha/2.$$

Proof. Under the null hypothesis, $Y_i^j \sim^{i.i.d.} \mathrm{Bern}(1/2)$. For each $\alpha \in (0,1)$ by applying Lemma 3.16 (with $k = d$ and $l = |\mathcal{I}_1|$) we get that $\mathbb{E}_0 T^1_{\mathrm{III}} \leqslant \alpha/2$ for large enough constant κ_α. For $f \in H_\rho$, we have

$$\mathbb{E}_f Y_i^j = \mathbb{E}_f \mathbb{E}_f \left[Y_i^{(j)} | X^{(j)} \right] = \Phi\left(\sqrt{n} f_i\right).$$

To bound the Type II error, we use the second statement of Lemma 3.16 (with $k = d$ and $l = |\mathcal{I}_1|$), but before that we show that condition (3.50) holds. Note that by Lemma 3.26,

$$\frac{|\mathcal{I}_1| - 1}{2\sqrt{d}} \sum_{i=1}^{d} \left(\mathbb{E}_f Y_i^{(j)} - \frac{1}{2} \right)^2 \geqslant \frac{|\mathcal{I}_1| - 1}{24\sqrt{d}} \sum_{i=1}^{d} \left(n f_i^2 \bigwedge 1 \right). \tag{3.55}$$

In case (3.54) holds, the preceding display is bounded from below by

$$\frac{|\mathcal{I}_1| - 1}{24\sqrt{d}} \sum_{i\in\mathcal{J}^c} n f_i^2 \geqslant \frac{n(|\mathcal{I}_1| - 1)\rho^2}{48\sqrt{d}}.$$

Note, that for large enough $C_\alpha > 0$, $\frac{n(|\mathcal{I}_1|-1)\rho^2}{48\sqrt{d}} \geqslant n(\frac{mb}{d}) C_\alpha \frac{d\sqrt{d}}{mnb} / (96\sqrt{d}) \geqslant \kappa_\alpha \vee \frac{16}{\alpha}$. If (3.54) does not hold, then there exists $i^* \in \{1, \ldots, d\}$ such that $f_{i*} \geqslant \sqrt{1/n}$, so (3.55) is lower bounded by

$$\frac{|\mathcal{I}_1| - 1}{24\sqrt{d}} \geqslant \frac{mb}{24d\sqrt{d}} - \frac{1}{12\sqrt{d}} \geqslant \frac{M_\alpha}{24} - \frac{1}{12}.$$

Then for large enough $M_\alpha > 0$, the condition (3.50) is satisfied. Consequently, the statement of the proof follows by the second statement of Lemma 3.16. □

Lemma 3.18. *For any $\alpha \in (0,1)$ there exists a $\kappa_\alpha > 0$ large enough such that $\mathbb{E}_0 T^2_{III} \leqslant \alpha/2$. Furthermore, if $\rho^2 \geqslant C_\alpha \frac{d\sqrt{d}}{mn(d\wedge b)}$, $m \geqslant M_\alpha$, for some large enough $C_\alpha, M_\alpha > 0$, the set $\mathcal{J}$ defined in* (3.13) *is non-empty and $b \geqslant 2\log(d+1)$, then $\mathbb{E}_f T^2_{III} \leqslant \alpha/2$.*

Proof of Lemma 3.18. We apply Lemma 3.16 (with $k = 1$ and $l = C_{b,d}dm$). Under the null hypothesis, $\left(\sqrt{n}X_i^{(j)}\right)^2$ follows a chi-square distribution with one degree of freedom. Consequently,

$$\mathbb{E}_0 B_{li}^{(j)} = \mathbb{E}_0 F_{\chi_1^2}\Big(\Big(\sqrt{n}X_i^{(j)}\Big)^2\Big) = 1/2$$

and

$$\sum_{j=1}^{m} Y_{count}^{(j)} \sim \mathrm{Bin}\left(1/2, mdC_{b,d}\right). \tag{3.56}$$

Then Lemma 3.16 yields that $\mathbb{E}_0 T_{\mathrm{III}}^2 \leqslant \alpha/2$.

Next we deal with the upper bound for the Type II error. Let $p_i := \mathbb{E}_f F_{\chi_1^2}\big((\sqrt{n}X_i^{(j)})^2\big)$ and note that $p_i \geqslant 1/2$. We apply again Lemma 3.16 (with $k = d$, $l = mC_{b,d}$). Hence, it is sufficient to show that the condition (3.50) of the lemma holds. For this first note that $\left(\sqrt{n}X_i^{(j)}\right)^2$ is a non-central chi-square distributed random variable with non-centrality parameter nf_i^2 and one degree of freedom. Consequently, for all $i \in \mathcal{J} \neq \varnothing$ we have

$$p_i = \mathbb{E}_f F_{\chi_1^2}\Big(\Big(\sqrt{n}X_i^{(j)}\Big)^2\Big) = \Pr\left(V \geqslant 1\right) > 3/5, \tag{3.57}$$

where it is used that V is noncentral F-distributed with noncentrality parameter $nf_i^2 \geqslant 1$ and $(1,1)$-degrees of freedom. Then by recalling that $\tilde{p}_i \geqslant 1/2$ we get that

$$\begin{aligned}\frac{mC_{b,d}-1}{2d}\Big(\sum_{i=1}^{d}(p_i - \frac{1}{2})\Big)^2 &\geqslant \frac{mC_{b,d}-1}{2d}\Big(\sum_{i\in\mathcal{J}}(\tilde{p}_i - \frac{1}{2})\Big)^2 \\ &\geqslant \frac{mC_{b,d}-1}{2d}(|\mathcal{J}|/10)^2 \geqslant \frac{m2^b}{400d^2} \geqslant M_\alpha/400,\end{aligned}$$

yielding (3.50) for large enough choice of M_α and hence concluding the proof of our statement. □

3.4.1.3 Auxiliary bandwidth constraint lemmas

Lemma 3.19. *Let U_d and $V_d^{\delta_d}$ be independent chi-square distributed random variables with d degree of freedom and non-centrality parameters zero and $\delta_d > 0$, respectively. Then for a universal $\overline{D} \in \mathbb{N}$, not depending on δ_d, we have for all $d \geqslant \overline{D}$ that*

$$Pr\left(V_d^{\delta_d} - U_d \geqslant 0\right) \geqslant \frac{1}{2} + \frac{1}{40}(\frac{\delta_d}{\sqrt{d}} \wedge \frac{1}{2}). \tag{3.58}$$

Proof. First note that the function $\delta \mapsto \Pr\left(V_d^{\delta} - U_d \geqslant 0\right)$ is monotone increasing. Then

$$\Pr\left(V_d^{\delta_d} - U_d \geqslant 0\right) \geqslant \Pr\left(V_d^{\delta_d \wedge \sqrt{d}/2} - U_d \geqslant 0\right),$$

so without loss of generality we can assume that $\delta_d \leqslant \sqrt{d}/2$.

The density of $V_d^{\delta_d}$ is

$$\sum_{k=0}^{\infty} \frac{e^{-\delta_d/2}\left(\delta_d/2\right)^k}{k!} p_{d+2k},$$

where p_k denotes the χ^2_k-density. By the independence of U_d and $V_d^{\delta_d}$,

$$\Pr\left(V_d^{\delta_d} - U_d \leqslant 0\right) = \sum_{k=0}^{\infty} \frac{e^{-\delta_d/2}\left(\delta_d/2\right)^k}{k!} \int_{\{v-u\geqslant 0\}} p_{d+2k}(v) p_d(u) d(v,u).$$

Let $U'_d \sim \chi^2_d$ and $U''_{2k} \sim \chi^2_{2k}$ be independent from each other and U_d. For any given $k \in \mathbb{N}$, we have

$$\int_{\{v-u\geqslant 0\}} p_{d+2k}(v) p_d(u) d(v,u) = \Pr\left(U_d - U'_d \leqslant U''_{2k}\right).$$

For convenience let us introduce the notation $W_d = (U_d - U'_d)/(2\sqrt{d})$. Conditioning and using independence once more, the latter equals

$$\int \Pr\left(W_d \leqslant \frac{u}{2\sqrt{d}}\right) d\mathbb{P}_{U''_{2k}}(u) = \frac{1}{2} + \int \Pr\left(0 \leqslant W_d \leqslant \frac{u}{2\sqrt{d}}\right) d\mathbb{P}_{U''_{2k}}(u).$$

Since U''_{2k} has a median larger than $2k/3$ and the map $u \mapsto \Pr\left(0 \leqslant W_d \leqslant \frac{u}{2\sqrt{d}}\right)$ is increasing, we have that the second term in the last display satisfies

$$\begin{aligned}\int \Pr\left(0 \leqslant W_d \leqslant \frac{u}{2\sqrt{d}}\right) p_{2k}(u) du &\geqslant \Pr\left(0 \leqslant W_d \leqslant \frac{k}{3\sqrt{d}}\right) \int_{[\frac{2k}{3},\infty)} p_{2k}(u) du \\ &\geqslant \frac{1}{2}\Pr\left(0 \leqslant W_d \leqslant \frac{k}{3\sqrt{d}}\right).\end{aligned}$$

By combining the above inequalities we obtain that

$$\Pr\left(V_d^{\delta_d} - U_d \leqslant 0\right) \geqslant \frac{1}{2} + \frac{1}{2}\sum_{k=0}^{\infty} \frac{e^{-\delta_d/2}\left(\delta_d/2\right)^k}{k!} \Pr\left(0 \leqslant W_d \leqslant \frac{k}{3\sqrt{d}}\right). \tag{3.59}$$

Assume now that $\delta_d \gtrsim 1$. Let k_d be the largest integer such that $k_d \leqslant 3\sqrt{d}$. We divide the sum on the right hand of the preceding display to two parts, i.e. $k < k_d$ and $k \geqslant k_d$. By applying Lemma 3.20 with $\varepsilon_d = k$, it holds that for $c_0 = e^{-9/8}/6$,

$$\begin{aligned}\sum_{k=0}^{k_d} \frac{e^{-\delta_d/2}\left(\delta_d/2\right)^k}{k!} \Pr\left(0 \leqslant W_d \leqslant \frac{k}{3\sqrt{d}}\right) &\geqslant \frac{c_0}{\sqrt{d}} \sum_{k=1}^{k_d} \frac{e^{-\delta_d/2}\left(\delta_d/2\right)^k}{(k-1)!} \\ &\geqslant \frac{c_0 \delta_d}{2\sqrt{d}} \sum_{k=0}^{k_d-1} \frac{e^{-\delta_d/2}\left(\delta_d/2\right)^k}{k!}.\end{aligned}$$

We have $\Pr(0 \leqslant W_d \leqslant 1) \xrightarrow{d} \Pr(0 \leqslant Z \leqslant 1) > 1/3$, hence there exists a $D_1 \in \mathbb{N}$, such that for all $d \geqslant D_1$ we have $\Pr(0 \leqslant W_d \leqslant 1) > 1/3$. For $k > k_d$ we have $k > 3\sqrt{d}$, hence for all $d \geqslant D_1$,

$$\sum_{k>k_d}^{\infty} \frac{e^{-\delta_d/2}(\delta_d/2)^k}{k!} \Pr\left(0 \leqslant W_d \leqslant \frac{k}{3\sqrt{d}}\right) \geqslant \frac{c_0}{2} \sum_{k>k_d}^{\infty} \frac{e^{-\delta_d/2}(\delta_d/2)^k}{k!}.$$

Since $\delta_d/\sqrt{d} \leqslant 1/2$, we have for $d \geqslant D_1$,

$$\frac{1}{2}\sum_{k=0}^{\infty} \frac{e^{-\delta_d/2}(\delta_d/2)^k}{k!} \Pr\left(0 \leqslant W_d \leqslant \frac{k}{3\sqrt{d}}\right) \geqslant \frac{c_0\delta_d}{2\sqrt{d}}\left(1 - \frac{e^{-\delta_d/2}(\delta_d/2)^{k_d}}{k_d!}\right).$$

The proof is finished by showing that for large enough d we have $c_0/2 - 1/40 > (\delta_d/2)^{k_d}/k_d! > 0$. Recalling that $2\sqrt{d} \leqslant 3\sqrt{d} - 1 \leqslant k_d \leqslant 3\sqrt{d}$ and hence $\delta_d \leqslant \sqrt{d}/2 \leqslant \sqrt{d}/4$ we get in view of Stirling's inequality, that for some universal constant $C > 0$,

$$\frac{(\delta_d/2)^{k_d}}{k_d!} \leqslant \frac{(k_d/4)^{k_d}}{k_d!} \lesssim e^{k_d(1-\log 4)} k_d^{-1/2}.$$

This is in turn bounded from above by $c_0/2 - 1/40$ for $d \geqslant D_1$, for some sufficiently large $D_1 > 0$.

□

Lemma 3.20. *Let $U_d, U_d' \overset{iid}{\sim} \chi^2_d$, and $0 < \varepsilon_d \leqslant C\sqrt{d}$. Then there exists a large enough $D_0 \in \mathbb{N}$, such that for all $d \geqslant D_0$*

$$Pr\left(0 \leqslant \frac{U_d - U_d'}{2\sqrt{d}} \leqslant \frac{\varepsilon_d}{\sqrt{d}}\right) \geqslant \frac{e^{-C^2/8}}{6}\frac{\varepsilon_d}{\sqrt{d}}.$$

Proof. The characteristic function of the random variable $W_d := (U_d - U_d')/(2\sqrt{d})$ is

$$\begin{aligned}\phi_d(t) &= \mathbb{E}e^{itW_d} = \mathbb{E}e^{i\frac{t}{2\sqrt{d}}U_d}\mathbb{E}e^{-i\frac{t}{2\sqrt{d}}U_d'} \\ &= (1 + it/\sqrt{d})^{-d/2}(1 - it/\sqrt{d})^{-d/2} \\ &= (1 + t^2/d)^{-d/2} \xrightarrow{d\to\infty} e^{-t^2/2}.\end{aligned}$$

Using the Fourier inversion formula, the density f_{W_d} of W_d satisfies

$$f_{W_d}(v) = \frac{1}{2\pi}\int_{\mathbb{R}} e^{itv}\phi_d(t)dt = \frac{1}{2\pi}\int_{\mathbb{R}} \cos(tv)\phi_d(t)dt,$$

where the second equality follows from the symmetry of ϕ_d. Let

$$g(v) := \frac{1}{2\pi}\int_{\mathbb{R}} \cos(tv)e^{-t^2/2}dt = \frac{1}{\sqrt{2\pi}}e^{-v^2/2},$$

where the last equation follows for instance by contour integration. Then by the dominated convergence theorem

$$|f_{W_d}(v) - g(v)| \leqslant \frac{1}{2\pi}\int_{\mathbb{R}} |e^{-t^2/2} - \phi_d(t)|dt \overset{d\to\infty}{\longrightarrow} 0.$$

By the earlier established uniform convergence we have that for every $d \geqslant D_0$, for some large enough D_0,

$$\int_0^{\frac{\varepsilon_d}{2\sqrt{d}}} f_{W_d}(v)dv \geqslant \frac{1}{3}e^{-\varepsilon_d^2/(8d)}\frac{\varepsilon_d}{2\sqrt{d}} = \frac{e^{-C^2/8}}{6}\frac{\varepsilon_d}{\sqrt{d}},$$

where the constant $1/3$ is arbitrary and could be taken anything smaller than $1/\sqrt{2\pi}$. □

3.4.2 Lemmas concerning the upper bounds in privacy constrained setting

3.4.2.1 Proof of Lemma 3.7

Proof. The case where $\mathcal{C}$ is empty is trivial, so we shall assume $\mathcal{C}$ to be nonempty. We follow the construction of McShane [153] whilst providing an additional argument to assure Borel-measurability of the extension. Consider the map $\breve{S} : \mathbb{R}^{n\times d} \to [-\infty, \infty)$ given by

$$\breve{S}(x) = \begin{cases} S(x) & \text{if } x \in \mathcal{C}, \\ \inf\{S(c) + D\mathrm{d}_H(c, x) \,:\, c \in \mathcal{C}\} & \text{otherwise.} \end{cases}$$

Fix any $c' \in \mathcal{C}$. Since S is D-Lipschitz with respect to the Hamming distance, we have for all $c \in \mathcal{C}$ that

$$S(c) + D\mathrm{d}_H(c, x) \geqslant S(c') - D\mathrm{d}_H(c', c) + D\mathrm{d}_H(c, x) \geqslant S(c') - D\mathrm{d}_H(c', x) > -\infty$$

where the last step follows from the triangle inequality. So, $\breve{S}$ is real valued. For all $x \in (\mathbb{R}^d)^n$ and $\gamma > 0$, there exists $c_\gamma \in \mathcal{C}$ such that

$$\breve{S}(x) \geqslant S(c_\gamma) + D\mathrm{d}_H(c_\gamma, x) - \gamma.$$

So for $x, x' \in (\mathbb{R}^d)^n$,

$$\breve{S}(x') - \breve{S}(x) \leqslant \breve{S}(c_\gamma) + D\mathrm{d}_H(c_\gamma, x') - \breve{S}(c_\gamma) - D\mathrm{d}_H(c_\gamma, x) + \gamma \leqslant D\mathrm{d}_H(x, x') + \gamma.$$

By symmetry of the argument and since $\gamma > 0$ is given arbitrarily, we conclude that $\breve{S}$ is D-Lipschitz with respect to the Hamming distance. Note, however, that this construction does not guarantee that $\breve{S}$ is measurable.

For any map $H : \mathbb{R}^{n\times d} \to [-\infty, \infty]$, let H^* denote its minimal Borel-measurable majorant. That is, a measurable map $H^* : \mathbb{R}^{n\times d} \to [-\infty, \infty]$ such that

1. $H \leqslant H^*$ and
2. $H^* \leqslant T$ $\mathbb{P}_0$-a.s. for every measurable $T : \mathbb{R}^{n\times d} \to [-\infty, \infty]$ with $T \geqslant H$.

Such a map exists by e.g. Lemma 1.2.1 in [208]. The map $\tilde{S} : (\mathbb{R}^d)^n \to \mathbb{R}$ defined by

$$\tilde{S}(x) = \breve{S}^*(x)\mathbb{1}_{x\notin\mathcal{C}} + S(x)\mathbb{1}_{x\in\mathcal{C}}$$

is measurable and can be seen to be a Borel-measurable majorant of $\breve{S}$; following from the fact that sums and products of measurable functions are measurable, $\breve{S} \leqslant \breve{S}^*$ and $\breve{S}(x) = S(x)$ for $x \in \mathcal{C}$.

Furthermore, by combining the fact that $\tilde{S}$ is measurable with e.g. Lemma 1.2.2 in [208], we get

$$|\tilde{S}(x) - \tilde{S}(x')| = |(\tilde{S}(x) - \tilde{S}(x'))^*| \leqslant |\breve{S}(x) - \breve{S}(x')|^*, \tag{3.60}$$

where $(x, x') \mapsto |\breve{S}(x) - \breve{S}(x')|^*$ is minimal Borel-measurable majorant of $(x, x') \mapsto |\breve{S}(x) - \breve{S}(x')|$. Since $\breve{S}$ is D-Lipschitz with respect to the Hamming distance $(x, x') \mapsto \mathrm{d}_H(x, x')$, which is a measurable map,

$$|\breve{S}(x) - \breve{S}(x')|^* \leqslant D\mathrm{d}_H(x, x').$$

From (3.60) it follows that for all $x, x' \in (\mathbb{R}^d)^n$,

$$|\tilde{S}(x) - \tilde{S}(x')| \leqslant D\mathrm{d}_H(x, x').$$

We have obtained a map $\tilde{S}$ that is D-Lipschitz with respect to the Hamming distance, measurable and $\tilde{S} = S$ on $\mathcal{C}$, concluding the proof. □

3.4.2.2 Lemmas concerning clipping

Lemma 3.21. *Let $\tau, \mu > 0$ satisfy $\tau/4 \leqslant \mu \leqslant \tau/2$, let V be a random variable symmetric about zero ($V \overset{d}{=} -V$) with Lebesgue density bounded by $M > 0$ and*

$$Pr\left(|V| \leqslant \frac{1}{12M} \vee (\tau/2)\right) \geqslant c$$

for some constant $c > 0$. It then holds that

$$\mathbb{E}\left[\mu + V\right]_{-\tau}^{\tau} \geqslant (c \wedge 1/2)\mu. \tag{3.61}$$

Proof. By definition of clipping,

$$\mathbb{E}\left[\mu + V\right]_{-\tau}^{\tau} = \mathbb{E}\left[V\right]_{-\tau-\mu}^{\tau-\mu} + \mu.$$

The first term equals

$$\begin{aligned}\mathbb{E}\left[V\right]_{-(\tau-\mu)}^{\tau-\mu} + \mathbb{E}\mathbb{1}_{\{V\in[-\tau-\mu,-\tau+\mu]\}}\left(\left[V\right]_{-\tau-\mu}^{-\tau+\mu} + (\tau-\mu)\right) \geqslant \\ \mathbb{E}\left[V\right]_{-(\tau-\mu)}^{\tau-\mu} - (\tau+\mu)\mathrm{Pr}\left(-\tau-\mu \leqslant V \leqslant -\tau+\mu\right) = \\ -(\tau+\mu)\mathrm{Pr}\left(-\tau-\mu \leqslant V \leqslant -\tau+\mu\right),\end{aligned}$$

where the last equality follows from the symmetry of V. By the condition on the Lebesgue density of V, the right-hand side of the above display can be further bounded from below by $-2M(\tau+\mu)\mu$. When $3M\tau < 1/2$, we obtain (3.61) with the constant $1/2$. Assume $3M\tau \geqslant 1/2$. Then, since $\mu > 0$ and V is symmetric about zero,

$$\begin{aligned}\mathbb{E}\left[\mu+V\right]_{-\tau}^{\tau} &= \mathbb{E}\left(\mu+V\right)\mathbb{1}\{|\mu+V| \leqslant \tau\} + \tau\left(\Pr\left(\mu+V>\tau\right) - \Pr\left(\mu+V<\tau\right)\right)\\ &\geqslant \mathbb{E}\left(\mu+V\right)\mathbb{1}\{|V| \leqslant \tau-\mu\}\\ &\geqslant \mathbb{E}\left(\mu+V\right)\mathbb{1}\{|V| \leqslant \tau/2\}\\ &= \mu\Pr\left(|V| \leqslant (1/12M) \vee (\tau/2)\right)\end{aligned}$$

where the last inequality follows from $\mu \leqslant \tau/2$ and the last equality follows from the symmetry of V about zero. □

Lemma 3.22. *For any $\tau > 0$ and random variable V with $|\mathbb{E}V| \leqslant \tau$,*

$$Var\left(\left[V\right]_{-\tau}^{\tau}\right) \leqslant Var\left(V\right).$$

If both V and $\left[V\right]_{-\tau}^{\tau}$ are mean 0, $\mathbb{E}\left(\left[V\right]_{-\tau}^{\tau}\right)^4 \leqslant \mathbb{E}V^4$.

Proof. Let $\mu = \mathbb{E}V$. Since the expectation of a random variable is the constant minimizing the L_2-distance to that random variable,

$$\mathrm{Var}\left(\left[V\right]_{-\tau}^{\tau}\right) \leqslant \mathbb{E}\left(\left[V\right]_{-\tau}^{\tau} - \mu\right)^2.$$

The latter expectation can be written as

$$\mathbb{E}\mathbb{1}_{V\in[-\tau,\tau]}\left(V-\mu\right)^2 + \mathbb{E}\mathbb{1}_{V>\tau}\left(\tau-\mu\right)^2 + \mathbb{E}\mathbb{1}_{V<-\tau}\left(-\tau-\mu\right)^2.$$

Assuming $\mu \geqslant 0$, $V < -\tau$ implies that $|-\tau-\mu| \leqslant |V-\mu|$. Since $\mu \leqslant \tau$, $V > \tau$ implies that $|\tau-\mu| \leqslant |V-\mu|$. Consequently, the above display is bounded from above by

$$\mathbb{E}\mathbb{1}_{V\in[-\tau,\tau]}\left(V-\mu\right)^2 + \mathbb{E}\mathbb{1}_{V>\tau}\left(V-\mu\right)^2 + \mathbb{E}\mathbb{1}_{V<-\tau}\left(V-\mu\right)^2 = \mathbb{E}\left(V-\mu\right)^2.$$

The case where $\mu < 0$ follows by the same reasoning. The last statement follows by a similar argument. □

3.4.2.3 Proof of Lemma 3.11

Proof. Since $\mathcal{C}_\tau = \mathcal{A}_\tau \cap \mathcal{B}_\tau$, it suffices to show that $\mathcal{A}_\tau^c$ and $\mathcal{B}_\tau^c$ as defined in (3.26) are small in $\mathbb{P}_f$-probability for a large enough choice of $\tilde{\kappa}_\alpha > 0$ in (3.20). For both sets, we proceed via a union bound:

$$\begin{aligned}\mathbb{P}_f\left(X^{(j)} \notin \mathcal{A}_\tau\right) &= \mathbb{P}_f\left(\exists \mathcal{J}\subset[n],\, |\mathcal{J}| \leqslant K : \left|\left\|\sum_{i\in\mathcal{J}} X_i^{(j)}\right\|_2^2 - |\mathcal{J}|d\right| > |\mathcal{J}|\gamma\right)\\ &\leqslant \sum_{k=1}^{K}\binom{n}{k}\Pr\left(\left|\|\sqrt{k}f - Z\|_2^2 - d\right| > \gamma\right) \qquad (3.62)\end{aligned}$$

where $Z \sim N(0, I_d)$. We have

$$\|\sqrt{k}f - Z\|_2^2 = k\|f\|_2^2 - 2\sqrt{k}f^\top Z + \|Z\|_2^2.$$

Recalling that $K = \lceil 2\tau D^{-1} \rceil$, $D = \gamma/n\sqrt{d}$ and

$$\gamma = \tilde{\kappa}_\alpha \log(1+m)\left(\sqrt{\log^2(1+n)n\sqrt{d}\tau} \vee \sqrt{nd}\right), \tag{3.63}$$

we obtain that

$$K \leqslant \frac{2n\sqrt{d}\tau}{\tilde{\kappa}_\alpha \log(1+m)\sqrt{\log^2(1+n)n\sqrt{d}\tau} \vee \sqrt{nd}} \lesssim \gamma/\log(1+n). \tag{3.64}$$

By the assumptions of the lemma ($n\|f\|_2^2 \leqslant \tau\sqrt{d}/2$ and $\tau \leqslant nM^2/\sqrt{d}$), we obtain that $k\|f\|_2^2 \leqslant KM^2$. Consequently, we have that for $\tilde{\kappa}_\alpha > 0$ large enough $K\|f\|_2^2 < \gamma/2$, so it holds that

$$\Pr\left(\|\sqrt{k}f - Z\|_2^2 - d > \gamma\right) \leqslant \Pr\left(\|Z\|_2^2 - d - 2\sqrt{k}f^\top Z > \gamma/2\right).$$

Using that $\Pr(A \cap B) + \Pr(A \cap B^c) \leqslant \Pr(A') + \Pr(A \cap B^c)$ for $A' \subset A \cap B$, it follows that the latter display is bounded above by

$$\Pr\left(\|Z\|_2^2 - d > \gamma/4\right) + \Pr\left(-2\sqrt{k}f^\top Z > \gamma/4\right).$$

By e.g. Lemma 3.28, the first probability is bounded by $e^{-d\gamma/8}$. Again using $K\|f\|_2^2 < \gamma/2$, the second term is bounded by $e^{-\gamma/32}$, where we note that the second term equals zero in the case that $f = 0$. The bound

$$\Pr\left(\|\sqrt{k}f - Z\|_2^2 - d < -\gamma\right) \leqslant e^{-\gamma/4} + e^{-\gamma/8}$$

follows by similar reasoning. Combining the above with the elementary bound $\sum_{k=1}^K \binom{n}{k} \leqslant e^{K\log(n)}$ means that

$$\mathbb{P}_f\left(X^{(j)} \notin \mathcal{A}_\tau\right) \leqslant 2\exp\left(K\log(n) - \frac{\gamma}{8}\right) \leqslant \alpha/(4mn).$$

Turning our attention to $\mathcal{B}_\tau$, we find that $\mathbb{P}_f\left(X^{(j)} \notin \mathcal{B}_\tau\right)$ is equal to

$$\mathbb{P}_f\left(\max_{i\in[n]}\left|\sum_{k\in[n]\setminus\{i\}}\langle X_i^{(j)}, X_k^{(j)}\rangle\right| > \gamma\right) \leqslant n\Pr\Big(\left|\langle f + Z, (n-1)f + \sqrt{n-1}Z'\rangle\right| > \gamma\Big),$$

where Z and Z' are independent $N(0, I_d)$ random vectors. Using another union bound, the above is further bounded by

$$\Pr\Big(\sqrt{n-1}\langle Z, Z'\rangle > \gamma/2 - (n-1)\|f\|_2^2\Big) + \Pr\Big((n-1)\langle f, Z'\rangle + \sqrt{n-1}\langle f, Z\rangle > \gamma/2\Big). \tag{3.65}$$

Using that $n\|f\|_2^2 \leqslant \tau\sqrt{d}/2$ and $\tau \leqslant nM^2/\sqrt{d}$ by assumption of the lemma and recalling (3.63), we see that

$$n\|f\|_2^2 \leqslant \tau\sqrt{d}/2 \leqslant \sqrt{n\sqrt{d}\tau}\frac{\sqrt{\sqrt{d}\tau}}{2\sqrt{n}} \leqslant \frac{M}{\tilde{\kappa}_\alpha}\gamma. \tag{3.66}$$

For $\tilde{\kappa}_\alpha > 0$, the latter can be seen to be larger than $\gamma/4$. Consequently, the first term in (3.65) can be seen to be bounded by

$$\Pr\left(\sqrt{n-1}\langle Z, Z'\rangle > \gamma/4\right) \leqslant e^{-\frac{\gamma}{4\sqrt{nd}}} \leqslant e^{-\kappa_\alpha \log(1+m)\log(1+n)},$$

where the inequality follows from Lemma 2.38. Since Z and Z' are independent standard Gaussian vectors, the second term in (3.65) is bounded above by

$$\Pr\left(\sqrt{2}(n-1)\langle f, Z\rangle > \gamma/2\right) \leqslant e^{-\frac{\gamma^2}{8n^2\|f\|_2^2}}.$$

By using that $n\|f\|_2^2 \leqslant \tau\sqrt{d}/2$,

$$\gamma^2 \geqslant 2\tilde{\kappa}_\alpha^2 \log^2(1+m)\log^2(1+n)n^2\|f\|_2^2,$$

so the exponent in the second last display is bounded from below by

$$-\tilde{\kappa}_\alpha^2 k \log^2(1+m)\log^2(1+n)/4.$$

Hence, we have obtained that

$$\mathbb{P}_f\left(X^{(j)} \notin \mathcal{B}_\tau\right) \leqslant \frac{\alpha}{4m}$$

for $\tilde{\kappa}_\alpha > 0$ large enough. This concludes the proof of the lemma. □

3.4.2.4 Lemmas concerning clipped averages coordinate wise strategies

Consider for $L \in \mathbb{N}$, $l = 1, \ldots, L$ and $j = 1, \ldots, m$ the transcripts

$$Y_l^{(j)}|(X^{(j)}, U) = \gamma \sum_{i=1}^{n} [(UX_i^{(j)})_l]_{-\tau}^{\tau} + W_l^{(j)},$$

where $\tau = \tilde{\kappa}_\alpha\sqrt{\log(1+dmn)}$, $(W_l^{(j)})_{j,l}$ is either i.i.d. centered Laplace with scale parameter 1 or standard Gaussian noise and U is an independent uniformly random rotation taking values in $\mathbb{R}^{d\times d}$.

Lemma 3.23. *Let $\gamma > 0$, $L \in \mathbb{N}$ be given. The test*

$$\varphi = \mathbb{1}\left\{\frac{1}{\sqrt{L}}\sum_{l=1}^{L}\left[\left(\frac{1}{\sqrt{m}}\sum_{j=1}^{m} Y_l^{(j)}\right)^2 - n\gamma^2 - \mathbb{E}(W_l^{(j)})^2\right] \geqslant \kappa_\alpha(\gamma^2 n \vee 1)\right\}$$

satisfies $\mathbb{P}_0\varphi \lesssim (1+mnd)^{2-\tilde\kappa_\alpha^2/4} + 1/\kappa_\alpha^2$. *In particular, for any* $\alpha \in (0,1)$, *there exist constants* $\kappa_\alpha, \tilde\kappa_\alpha > 0$ *such that the test is of level* $\mathbb{P}_0\varphi \leqslant \alpha/2$. *Furthermore,* $\mathbb{P}_f(1-\varphi) \leqslant \alpha/2$ *if in addition it holds that* $\|f\|_2 \leqslant M$ *and*

$$\left(\frac{d}{Lmn\|f\|_2^2}\right) \vee \left(\frac{d}{\gamma^2 mLn^2\|f\|_2^2}\right) \vee \left(\frac{\kappa_\alpha^2 d^2}{m^2Ln^2\|f\|_2^4}\right) \vee \left(\frac{\kappa_\alpha^2 d^2}{\gamma^4 m^2Ln^4\|f\|_2^4}\right) \leqslant c_\alpha \quad (3.67)$$

for some $c_\alpha > 0$ *depending only on* α.

Proof. Under $\mathbb{P}_f$, $(UX_i^{(j)})_l \overset{d}{=} (Uf)_l + (UZ_i)_l$, where $Z_i \sim N(0, I_d)$ independent of U and the centered i.i.d. $W_l^{(j)}$. Furthermore, $UZ_i \overset{d}{=} Z_i$, $(UZ_i)_l \sim N(0,1)$, still independent of $(W_l^{(j)})_{l,j}$. We obtain that

$$\gamma\sum_{i=1}^n (UX_i^{(j)})_l + W_l^{(j)} \overset{d}{=} \gamma n(Uf)_l + \gamma\sqrt{n}\eta + W_l^{(j)},$$

with $\eta \sim N(0,1)$ and all three terms independent. Therefore, a straightforward calculation shows that,

$$V_l := \left(\frac{1}{\sqrt{m}}\sum_{j=1}^m \left[\gamma\sum_{i=1}^n (UX_i^{(j)})_l + W_l^{(j)}\right]\right)^2$$

has expectation conditionally on U equal to $\gamma^2 mn^2(Uf)_l^2 + n\gamma^2 + \mathbb{E}(W_l^{(j)})^2$. Since $\mathbb{E}Z_{il}^4 = 3$ and $\mathbb{E}(W_l^{(j)})^4 \asymp 1$, its variance conditionally on U equals

$$\begin{aligned} &n^2\gamma^4 \mathrm{Var}\left(\eta^2|U\right) + \mathrm{Var}\left(\left(\frac{1}{\sqrt{m}}\sum_{j=1}^m W_l^{(j)}\right)^2 |U\right) + 2\gamma^4 n^3 m(Uf)_l^2\mathbb{E}\eta^2 \\ &\quad + 2\gamma^2 mn^2(Uf)_l^2\mathbb{E}(W_l^{(j)})^2 + 2\gamma^2 n\mathbb{E}(W_l^{(j)})^2\mathbb{E}\eta^2, \end{aligned}$$

which is of the order

$$(\gamma^4 mn^3(Uf)_l^2) \vee (\gamma^2 mn^2(Uf)_l^2) \vee \gamma^4 n^2 \vee 1.$$

If

$$\max_{1\leqslant i\leqslant d}|f_i| \leqslant \tau/2, \quad (3.68)$$

an application of the triangle inequality and Lemma 3.27 yield that, for $\tilde\kappa_\alpha > 0$ large enough, we have with probability at least

$$1 - 2mnde^{-\tau^2/4} \geqslant 1 - (1+mnd)^{2-\tilde\kappa_\alpha^2/4} \geqslant 1 - \alpha/4$$

that

$$\max_{i\in[n], j\in[m], l\in[d]} \left|(X_i^{(j)})_l\right| \leqslant \tau.$$

Consequently, under the null hypothesis ($f = 0$), $\mathbb{P}_0\varphi$ is bounded above by

$$\mathbb{P}_0\left(\frac{1}{\sqrt{L}}\sum_{l=1}^{L}[V_l - \mathbb{E}_0 V_l] \geqslant \kappa_\alpha(\gamma^2 n \vee 1)\right) + \frac{\alpha}{4}.$$

Chebyshev's inequality yields that the first term on the left-hand side is of the order $1/\kappa_\alpha^2$, so less than $\alpha/4$ for $\kappa_\alpha > 0$ large enough. Since that $\|f\|_2 \leqslant M$ under the alternative hypothesis, (3.68) is also satisfied with probability at least $1 - \alpha/4$ for $\tilde{\kappa}_\alpha$ large enough. In the case that (3.68) holds, we also have that $\mathbb{P}_f(1-\varphi)$ is bounded above by

$$\Pr\left(\frac{1}{\sqrt{L}}\sum_{l=1}^{L}\left[\left(\gamma\sqrt{m}(n(Uf)_l + \sqrt{n}Z) + \frac{1}{\sqrt{m}}\sum_{j=1}^{m}W_l^{(j)}\right)^2 - \mathbb{E}_0 V_l\right] < \kappa_\alpha(\gamma^2 n \vee 1)\right) + \frac{\alpha}{4}. \tag{3.69}$$

Under $\mathbb{P}^U$, we have that

$$(Uf)_k \stackrel{d}{=} \|f\|_2 \frac{Z_k}{\|Z\|_2},$$

for $Z = (Z_1, \ldots, Z_d) \sim N(0, I_d)$. As $\sum_{k=1}^{d}\mathbb{E}Z_k^2/\|Z\|_2^2 = 1$, $\mathbb{E}Z_k^2/\|Z\|_2^2 = 1/d$ by symmetry. Consequently,

$$\mathbb{E}_f V_l = \gamma^2 m n^2 \mathbb{E}^U (Uf)_l^2 + n\gamma^2 + \mathbb{E}(W_l^{(j)})^2 = \frac{\gamma^2 m n^2 \|f\|_2^2}{d} + n\gamma^2 + \mathbb{E}(W_l^{(j)})^2.$$

Subtracting $d^{-1}\gamma^2 m n^2 \sqrt{L}\|f\|_2^2$ on both sides, the first term in (3.69) is bounded above by

$$\Pr\left(\frac{1}{\sqrt{L}}\sum_{l=1}^{L}[V_l - \mathbb{E}_f V_l] < -\frac{\gamma^2 n^2 m\sqrt{L}\|f\|_2^2}{2d}\right)$$

whenever

$$\frac{\gamma^2 m n^2 \sqrt{L}\|f\|_2^2}{d} \geqslant 2\kappa_\alpha(\gamma^2 n \vee 1). \tag{3.70}$$

An application of Chebyshev's inequality and the variance bound computed in (3.4.2.4) now yields that the latter probability is of the order

$$\frac{(\gamma^4 m d^{-1} n^3 \|f\|_2^2) \vee (\gamma^2 m n^2 d^{-1}\|f\|_2^2) \vee \gamma^4 n^2 \vee 1}{\left(\frac{\gamma^2 m n^2\sqrt{L}}{d}\|f\|_2^2\right)^2}.$$

So, in order to obtain $\mathbb{P}_f(1-\varphi) \leqslant \alpha/2$ it suffices to have (3.67), noting that this also yields (3.70).

□

For the next lemma, consider for some $L \in \mathbb{N}$ sets $\mathcal{I}_l \subset \{1, \dots, m\}$ such that $|\mathcal{I}_l| = \lfloor \frac{mL}{d} \rfloor$ and each $j \in \{1, \dots, m\}$ is in $\mathcal{I}_l$ for L different indexes $l \in \{1, \dots, d\}$. For $\gamma > 0$, $l = 1, \dots, d$ and $j \in \mathcal{I}_l$, generate the transcripts according to

$$Y_l^{(j)} = \gamma \sum_{i=1}^{n} \left[(X_i^{(j)})_l \right]_{-\tau}^{\tau} + W_l^{(j)},$$

with $\tau = \tilde{\kappa}_\alpha \sqrt{\log(dmn)}$ and $(W_l^{(j)})_{j,l}$ either i.i.d. centered Laplace with scale parameter 1 or standard Gaussian noise. In essence, its content and proof are similar to the previous lemma, the key difference being the absence of the shared random rotation.

Lemma 3.24. *Let $\gamma > 0$ be given. The test*

$$\varphi = \mathbb{1}\left\{ \frac{1}{\sqrt{d}} \sum_{l=1}^{d} \left[\left(\frac{1}{\sqrt{|\mathcal{J}_l|}} \sum_{j \in \mathcal{J}_l} Y_l^{(j)} \right)^2 - n\gamma - \mathbb{E}(W_l^{(j)})^2 \right] \geqslant \kappa_\alpha(\gamma^2 n \vee 1) \right\}$$

satisfies $\mathbb{P}_0\varphi \lesssim (1 + mnd)^{2-\tilde{\kappa}_\alpha^2/4} + 1/\kappa_\alpha^2$. In particular, for any $\alpha \in (0,1)$, there exist constants $\kappa_\alpha, \tilde{\kappa}_\alpha > 0$ such that the test is of level $\mathbb{P}_0\varphi \leqslant \alpha/2$. Furthermore, $\mathbb{P}_f(1-\varphi) \leqslant \alpha/2$ if in addition it holds that $\|f\|_2 \leqslant M$ and

$$\frac{d}{mnL\|f\|_2^2} \vee \frac{d}{\gamma^2 mn^2 L\|f\|_2^2} \vee \frac{\kappa_\alpha^2 d^3}{m^2 n^2 L^2 \|f\|_2^4} \vee \frac{\kappa_\alpha^2 d^3}{\gamma^4 m^2 n^4 L^2 \|f\|_2^4} \leqslant c_\alpha \tag{3.71}$$

for some $c_\alpha > 0$ depending only on α.

Proof. Under $\mathbb{P}_f$, $(X_i^{(j)})_l \overset{d}{=} f_l + Z_{il}$ with i.i.d. $Z_{il} \sim N(0,1)$ and is independent of the centered i.i.d. $W_l^{(j)}$. The quantity

$$V_l := \left(\frac{1}{\sqrt{|\mathcal{J}_l|}} \sum_{j \in \mathcal{J}_l} \left[\gamma \sum_{i=1}^{n} (X_i^{(j)})_l + W_l^{(j)} \right] \right)^2$$

is in distribution equal to

$$\left(\gamma\sqrt{|\mathcal{J}_l|} n f_l + \gamma\sqrt{n}\eta + \frac{1}{\sqrt{|\mathcal{J}_l|}} \sum_{j \in \mathcal{J}_l} W_l^{(j)} \right)^2$$

under $\mathbb{P}_f$, with $\eta \sim N(0,1)$ independent. Therefore, a straightforward calculation shows that V_l has mean $\gamma^2 n^2 |\mathcal{J}_l| f_l^2 + n\gamma^2 + \mathbb{E}(W_l^{(j)})^2$ under $\mathbb{P}_f$. Since $\mathbb{E}Z_{il}^4 = 3$ and $\mathbb{E}(W_l^{(j)})^4 \asymp 1$, its variance equals

$$\begin{aligned} n^2\gamma^4 \mathrm{Var}\left(\eta^2\right) + \mathrm{Var}\left(\left(\frac{1}{\sqrt{|\mathcal{J}_l|}} \sum_{j \in \mathcal{J}_l} W_l^{(j)} \right)^2 \right) + \gamma^4 |\mathcal{J}_l| n^3 f_l^2 \mathbb{E}\eta^2 \\ + \gamma^2 |\mathcal{J}_l| n^2 f_l^2 \mathbb{E}(W_l^{(j)})^2 + n\gamma^2 \mathbb{E}(W_l^{(j)})^2 \mathbb{E}\eta^2, \end{aligned}$$

which is of the order

$$(\gamma^4|\mathcal{J}_l|n^3f_l^2) \vee (\gamma^2|\mathcal{J}_l|n^2f_l^2) \vee \gamma^4n^2 \vee 1.$$

If

$$\max_{1\leqslant i\leqslant d}|f_i| \leqslant \tau/2, \tag{3.72}$$

an application of the triangle inequality and Lemma 3.27 yield that, for $\tilde{\kappa}_\alpha > 0$ large enough, we have with probability at least $1 - 2mnde^{-\tau^2/4} \geqslant 1 - (1 + mnd)^{2-\tilde{\kappa}_\alpha^2/4} \geqslant 1 - \alpha/4$ that

$$\max_{i\in[n],j\in[m],l\in[d]}|(X_i^{(j)})_l| \leqslant \tau.$$

Consequently, under the null hypothesis ($f = 0$), using that $|\mathcal{J}_l| = |\mathcal{J}_1|$, $\mathbb{P}_0\varphi$ is bounded above by

$$\mathbb{P}_0\left(\frac{1}{\sqrt{d}|\mathcal{J}_1|}\sum_{l=1}^d\left[\left(\sum_{j\in\mathcal{J}_l}\left(\gamma(n\overline{X^{(j)}})_l + W_l^{(j)}\right)\right)^2 - \mathbb{E}_0V_l\right] \geqslant \kappa_\alpha(\gamma^2n \vee 1)\right) + \frac{\alpha}{4}.$$

Chebyshev's inequality yields that the first term on the left-hand side is bounded $\alpha/4$ for $\kappa_\alpha > 0$ large enough. Under the alternative hypothesis, note that $\|f\|_2 \leqslant M$, so also in this case (3.72) is satisfied for $\tilde{\kappa}_\alpha$ large enough (a more extensive but easy calculation shows that this test also works for signals larger than M). In the case that (3.72) holds, we also have that $\mathbb{P}_f(1-\varphi)$ is bounded above by

$$\Pr\left(\frac{1}{\sqrt{d}|\mathcal{J}_1|}\sum_{l=1}^d\left[\left(\sum_{j\in\mathcal{J}_l}\left(\gamma(nf_l + \sqrt{n}Z) + W_l^{(j)}\right)\right)^2 - \mathbb{E}_0V_l\right] < \kappa_\alpha(\gamma^2n \vee 1)\right) + \frac{\alpha}{4}.$$

Subtracting $d^{-1/2}\sum_{l=1}^d\gamma^2n^2|\mathcal{J}_l|f_l^2$ on both sides, the first term is bounded above by

$$\Pr\left(\frac{1}{\sqrt{d}|\mathcal{J}_1|}\sum_{l=1}^d\left[\left(\sum_{j\in\mathcal{J}_l}\left(\gamma(nf_l + \sqrt{n}Z) + W_l^{(j)}\right)\right)^2 - \mathbb{E}_fV_l\right] < -\frac{\gamma^2n^2|\mathcal{J}_1|\|f\|_2^2}{2\sqrt{d}}\right)$$

whenever

$$\frac{\gamma^2n^2|\mathcal{J}_l|}{\sqrt{d}}\|f\|_2^2 \geqslant 2\kappa_\alpha(\gamma^2n \vee 1) \iff \frac{2\kappa_\alpha(\gamma^2n \vee 1)d\sqrt{d}}{\gamma^2n^2mL\|f\|_2^2} \leqslant 1. \tag{3.73}$$

An application of Chebyshev's inequality and the variance bound computed in (3.4.2.4), now yields that the latter probability is of the order

$$\frac{(d^{-1}\gamma^4|\mathcal{J}_l|n^3\|f\|_2^2) \vee (d^{-1}\gamma^2|\mathcal{J}_l|n^2\|f\|_2^2) \vee \gamma^4n^2 \vee 1}{\left(\frac{\gamma^2n^2|\mathcal{J}_l|}{\sqrt{d}}\|f\|_2^2\right)^2}.$$

Since $|\mathcal{J}_l| \asymp mL/d$, for the above expression to be smaller than $\alpha/2$ and for (3.73) to hold, it suffices to have (3.71).

□

3.4.3 Auxiliary lemmas and folklore

The following lemma examines a well-known construction which demonstrates that, that for data that is independently and identically distributed, it is possible to create tests with different power and significance levels, starting from a test that has non-trivial power and level. This construction does not affect the minimax separation rate achieved by these tests: the minimax separations for the two tests are equal up to a constant, regardless of the power and level of the two tests. Essentially, the lemma warrants the study of the sum of Type I and Type II error when considering the performance of inference in terms of minimax rate for i.i.d. data. It applies generally to the setting discussed in Section 1.1, which we shall recall briefly.

Consider a statistical model $\mathcal{P}_{n,\nu}$ in which we observe $n \equiv n_\nu$ i.i.d. observations from $P_{f,n,\nu} \equiv P_f$, indexed by $f \in \mathcal{F}_\nu$ and $\nu \in \mathbb{N}$. Consider the simple null hypothesis $H_0 : f = f_{0,\nu}$ and an arbitrary collection of alternatives $H_\rho \subset \mathcal{F}_\nu$, $\rho > 0$, with $H_{\rho'} \subset H_\rho$ for $\rho \leqslant \rho'$. Given a level $\alpha \in (0,1)$, consider the minimax Type II error probability for tests of level α;

$$\beta_{\mathcal{P}_{n,\nu}}(\alpha, H_0, H_\rho) = \inf_T \sup_{f \in H_\rho} P_f(1-T),$$

where the infimum is over all tests of level at most α. The minimax separation at $\alpha, \beta \in (0,1)$ is given by

$$\rho^*_{\alpha,\beta,n,\nu} := \inf\left\{\rho > 0 : \beta_{\mathcal{P}_\nu}(\alpha, H_0, H_\rho) \leqslant \beta\right\}.$$

Lemma 3.25. *Assume that $n \equiv n_\nu \to \infty$ as $\nu \to \infty$. Suppose that for $\alpha, \beta \in (0,1)$ with $\alpha + \beta < 1$, it holds that*

$$\rho^*_{\alpha,\beta,n,\nu} \asymp \rho^*_{\alpha,\beta,\lfloor n/k \rfloor,\nu}$$

for any fixed $k > 0$ as $\nu \to \infty$.

Then, $\rho^*_{\alpha,\beta,n,\nu} \asymp \rho^*_{\alpha',\beta',n,\nu}$ for all $\alpha, \beta, \alpha', \beta' \in (0,1)$ such that $\alpha + \beta$ and $\alpha' + \beta'$ are strictly less than 1.

Proof. Let $\rho^*_{\alpha,\beta,n,\nu}$ and $\rho^*_{\alpha',\beta',n,\nu}$ be given with $\alpha, \beta, \alpha', \beta'$ satisfying the assumption of the lemma. Assume without loss of generality that $\rho^*_{\alpha,\beta,n,\nu} \geqslant \rho^*_{\alpha',\beta',n,\nu}$.

By assumption of the lemma,

$$c_k \rho^*_{\alpha,\beta,n,\nu} \leqslant \rho^*_{\alpha,\beta,\lfloor n/k \rfloor,\nu} \leqslant C_k \rho^*_{\alpha,\beta,n,\nu}$$

for constants $c_k, C_k > 0$ depending only on k.

For any $\gamma > 0$, $k > 0$, there exists a sequence of tests $\varphi_{\lfloor n/k \rfloor,\nu}$ of level α such that

$$\sup_{f \in H_{C_k \rho^*_{\alpha,\beta,n,\nu}}} P_f\left(1 - \varphi^i_{\lfloor n/k \rfloor,\nu}\right) \leqslant \sup_{f \in H_{\rho^*_{\alpha,\beta,\lfloor n/k \rfloor,\nu}}} P_{f,\lfloor n/k \rfloor,\nu}(1 - \varphi_{\lfloor n/k \rfloor,\nu}) \leqslant \beta + \gamma$$

for all n large enough. Split the n data points into k sets of $\lfloor n/k \rfloor$ observations each. Let the test $\varphi^i_{\lfloor n/k \rfloor,\nu}$ be equal to $\varphi_{\lfloor n/k \rfloor,\nu}$ applied to the i-th subset of observations. The test $\varphi^i_{\lfloor n/k \rfloor,\nu}$ has level less than or equal to α. Consider the test of level α' given by

$$\varphi_{\alpha',\beta',n,\nu} := \mathbb{1}\left\{ \sum_{i=1}^{k} \varphi^i_{\lfloor n/k \rfloor,\nu} \geqslant F^{-1}_{\mathrm{Bin}(k,\alpha)}(1-\alpha') \right\} \tag{3.74}$$

with $F^{-1}_{\mathrm{Bin}(k,\alpha)}$ the quantile function of a binomial distribution with parameters (k,α). The quantity

$$\frac{1}{\sqrt{k}}(F^{-1}_{\mathrm{Bin}(k,\alpha)}(1-\alpha') - k\alpha)$$

is bounded by k by the central limit theorem (see Lemma 4.9 in Chapter 4 for details) and is consequently a bounded sequence in k. Under any alternative hypothesis $f \in H_{C_k \rho^*_{\alpha,\beta,n,\nu}}$,

$$\alpha - P_f \varphi^i_{\lfloor n/k \rfloor,\nu} \leqslant \alpha - 1 + \sup_{f \in H_{\rho^*_{\alpha,\beta,\lfloor n/k \rfloor,\nu}}} P_f\left(1 - \varphi^i_{\lfloor n/k \rfloor,\nu}\right) \leqslant \alpha + \beta + \gamma - 1.$$

Since it is assumed that $\alpha + \beta < 1$, it follows that the right-hand side is strictly less than zero for $\gamma > 0$ small enough. Subtracting $kP_f\varphi^i_{\lfloor n/k \rfloor,\nu}$ on both sides and dividing by $\sqrt{k}$, we obtain that $P_f\left(1 - \varphi_{\alpha',\beta',n,\nu}\right)$ is bounded above by

$$P_f\left(\frac{1}{\sqrt{k}} \sum_{i=1}^{k} (\varphi^i_{\lfloor n/k \rfloor,\nu} - P_f\varphi^i_{\lfloor n/k \rfloor,\nu}) < \frac{1}{\sqrt{k}}\left(F^{-1}_{\mathrm{Bin}(k,\alpha)}(1-\alpha') - k\alpha\right) + \sqrt{k}\alpha - \sqrt{k}P_f\varphi^i_{n/k,\nu} \right)$$

which equals $\Pr\left(O_P(1) < O(1) - \sqrt{k}\right)$ by the arguments above. The latter quantity can be seen to be less than or equal to β' for k large enough depending only on $\alpha, \beta, \alpha', \beta'$. Since $f \in H_{C_k \rho^*_{\alpha,\beta,n,\nu}}$ was given arbitrarily, we can conclude that $C_k \rho^*_{\alpha,\beta,n,\nu} \geqslant \rho^*_{\alpha',\beta',n,\nu}$. By symmetry of the argument, we also obtain that $\rho^*_{\alpha,\beta,n,\nu} \lesssim \rho^*_{\alpha',\beta',n,\nu}$ and the conclusion of the lemma follows. □

The following three lemmas are standard, technical results, nevertheless we provided them for completeness.

Lemma 3.26. *Let Φ denote the CDF of a standard normal random variable. It holds that*

$$\left(\Phi(x) - \frac{1}{2}\right)^2 \geqslant \frac{1}{12}\min\left\{x^2, 1\right\}.$$

Proof. Since $\Phi(x) = 1 - \Phi(-x)$, it holds that $\left(\Phi(x) - \frac{1}{2}\right)^2 = \left(\Phi(-x) - \frac{1}{2}\right)^2$. Hence, one can consider $x \geqslant 0$ without loss of generality. We first lower bound $\Phi(x) - \frac{1}{2}$ for

$0 \leqslant x \leqslant \sqrt{2}$. We have

$$\Phi(x) - \frac{1}{2} = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-\frac{1}{2}z^2} dz = \frac{1}{\sqrt{2\pi}} \int_0^x \sum_{i=0}^{\infty} \frac{(-1)^i z^{2i}}{2^i i!} dz = \frac{x}{\sqrt{2\pi}} \Big(\sum_{i=0}^{\infty} \frac{(-1)^i (x/\sqrt{2})^{2i}}{(2i+1)i!} \Big), \tag{3.75}$$

where the last equation follows by Fubini's theorem. The series in the right-hand side is decreasing in $x \in [0, \sqrt{2}]$, as for each odd i it holds that

$$\frac{d}{d\epsilon} \left[\frac{(-1)^i \epsilon^{2i}}{(2i+1)i!} + \frac{(-1)^{i+1} \epsilon^{2i+2}}{(2i+3)(i+1)!} \right] = \frac{\epsilon^{2i-1} 2i}{i!(2i+1)} \left(\frac{\epsilon^2 (2i+1)(2i+2)}{(i+1)2i(2i+3)} - 1 \right) < 0$$

for $0 \leqslant \epsilon \leqslant 1$. Hence, for $0 \leqslant x/\sqrt{2} \leqslant c \leqslant 1$,

$$\frac{x}{\sqrt{2\pi}} \Big(\sum_{i=0}^{\infty} \frac{(-1)^i (x/\sqrt{2})^{2i}}{(2i+1)i!} \Big) \geqslant \frac{x}{\sqrt{2\pi}} \left(\sum_{i=0}^{\infty} \frac{(-1)^i c^{2i}}{(2i+1)i!} \right) = \frac{x}{\sqrt{2}c} \left(\Phi(\sqrt{2}c) - \frac{1}{2} \right),$$

where the last equality follows by (3.75). For $x > \sqrt{2}c$, it holds that

$$\Phi(x) - 1/2 \geqslant \Phi(\sqrt{2}c) - 1/2$$

as $x \mapsto \Phi(x) - 1/2$ is increasing. Taking $c = 1$ we obtain that

$$\Phi(x) - 1/2 \geqslant \min \Big\{ x\big(\Phi(\sqrt{2}) - 1/2\big)/\sqrt{2}, \Phi(\sqrt{2}) - 1 \Big\} > \min\{x, 1\}/\sqrt{12},$$

which finishes the proof. □

Lemma 3.27. *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 $\mathbb{E} \max_{1 \leqslant i \leqslant K} |G_i| \leqslant 3\|M\| \sqrt{\log(K) \vee \log(2)}$ and*

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

for all $x > 0$.

Proof. It holds that

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

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

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

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

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

Consequently,

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

Hence, it suffices to show that

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

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

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

see Lemma 2.36. Taking $t = 1/4$ and applying Markov's inequality yields the second statement of the lemma. Furthermore, in view of Jensen's inequality

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

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

Lemma 3.28. *Let X_d be chi-square random variable with d-degrees of freedom. For $0 < c < 1$ it holds that*

$$Pr(X_d \leqslant cd) \leqslant e^{-d\frac{c-1-\log(c)}{2}}.$$

Similarly, for $c > 1$ it holds that

$$Pr(X_d \geqslant cd) \leqslant e^{-d\frac{c-1-\log(c)}{2}}.$$

Proof. Let $t < 0$. We have

$$\Pr\left(X_d \leqslant cd\right) = \Pr\left(e^{tX_d} \geqslant e^{tcd}\right) \leqslant \frac{\mathbb{E} e^{tX_d}}{e^{tcd}}.$$

Using that $\mathbb{E} e^{tX_d} = (1-2t)^{-d/2}$, the latter display equals

$$\exp\Big(- d\big(tc + \frac{1}{2}\log(1-2t)\big)\Big).$$

The expression $tc + \frac{1}{2}\log(1-2t)$ is maximized when $t = \frac{1}{2}(1 - \frac{1}{c}) < 0$ which leads to the result. The second statement follows by similar steps. □

Lemma 3.29. *Let $X \sim Bin(n,p)$. For any $0 < \gamma < 1$. It holds that*

$$Pr((1-\gamma)\mathbb{E}X \leqslant X \leqslant (1+\gamma)\mathbb{E}X) \leqslant 2\exp\left(-\frac{\gamma^2\mathbb{E}X}{3}\right).$$

Proof. This follows by a Chernoff bound using the moment generating function of the binomial distribution. □

Chapter 4

Consequences for meta-analysis based on combined test statistics across independent studies

Combining test statistics from independent trials or experiments is a popular method of meta-analysis. However, there is very limited theoretical understanding of the power of the combined test, especially in high-dimensional models considering composite hypotheses tests. In this chapter, derive a mathematical framework to study standard meta-analysis testing approaches in the context of the many-normal-means model, which serves as the platform to investigate more complex models.

Given multiple data sets relating to the same hypothesis, one would like to combine the evidence. Sometimes, the full data sets are not available (e.g. due to privacy or proprietary reasons) or difficult to combine directly (e.g. due to the different experimental or observational setups). In such cases, the analysis must be carried out on the basis of the published results for each of the studies. Such "meta-analysis" can increase the statistical power by combining individually inconclusive or moderately significant tests, while keeping the false positive rate under control. Therefore, meta-analysis has received a lot of attention in various fields, for instance in genetics and system biology, when studying rare variants [19, 96] or in deep learning, for few shot image recognition and neural architecture search, see the review article [116].

The outcomes of the studies concerning hypothesis tests are, typically, summarized as real-valued test statistics and/or associated p-values. One expects the combination of m such p-values to result in an increase in power, but one also expects to pay

a price relative to computing a test on the basis of the full, pooled data of the m trials. The question of how to optimally combine independent real-valued test statistics concerning the same hypothesis into a single test has an extensive literature. A multitude of methods for combining independent tests of significance exist. For combining p-values, this starts with Fisher, Tippett and Pearson in the nineteen-thirties, see [202, 100, 164, 185, 108, 144, 209, 88, 155, 218, 212, 61, 214] and references therein. In Section 4.2, we collect and describe the most popular and frequently used p-value combination techniques.

We introduce a natural and mild restriction on the meta-level combination functions of the local trials. This allows us to mathematically quantify the cost of compressing m trials into real-valued test statistics and combining these. We then derive minimax lower and matching upper bounds for the separation rates of standard combination methods for e.g. p-values and e-values, quantifying the loss relative to using the full, pooled data. The results bare resemblance with the $b = 1$-bit bandwidth constraint setting of Chapters 2 and 3, where we reveal for example that in certain cases combining the locally optimal tests in each trial results in a suboptimal meta-analysis method and develop approaches to achieve the global optima. We also explore the possible gains of allowing limited coordination between the trial designs by using shared randomness. Our results connect meta-analysis with bandwidth constraint distributed inference.

As noted in [35], there does not exist a general uniformly most powerful p-value combination method for all alternative hypotheses. The distribution of a p-value or its underlying test statistic under the alternative hypothesis should be taken into consideration when selecting a method of combination. The performance of different p-value combination techniques was investigated extensively by empirical experiments in various synthetic and real world scenarios, see for instance [148, 224]. However, a unified, general theoretical description is lacking, especially in non-trivial, multidimensional composite testing problems, where the likelihood ratio test is not necessarily uniformly most powerful.

E-values are an increasingly popular and important notion of evidence, see [178, 111, 179]. E-values allow the combination of several tests in a straightforward manner while preserving the prescribed level of the tests (see Section 4.2.2). Formally, e-values are nonnegative random variables whose expected values under the null hypothesis are bounded by one. In contrast to p-values defined by probabilities, e-values are defined by expectation. This imposes significant differences in their interpretation, application and combination compared to the more standard p-values. However, as for p-values, very little is known about the power of these combination procedures. Theoretical results focus on specific optimality criteria, for instance the worst-case growth-rate (GROW), see [111]. However, these do not directly imply guarantees on the testing power, which is the main focus in practice.

We consider the signal detection problem in the many-normal-means model as considered in the earlier chapters. One possible interpretation of this testing problem

is to learn whether a treatment has an effect on any of the dimensions investigated. This model is directly applied in several fields where high-dimensional statistics and machine learning settings are concerned, such as detecting differentially expressed genes [167, 132, 149, 200, 92], bankruptcy prediction for publicly traded companies using Altman's Z-score in finance [21, 22], separation of the background and source in astronomical images [62, 112], and wavelet analysis [1, 125]. Furthermore, the model allows for tractable computations and it typically serves as the platform to investigate more difficult statistical and learning problems, including high- and infinite-dimensional models, see for instance [119, 204, 92, 106]. Results for the many-normal-means model in principle can translate to many other multidimensional models, where a certain loss in power is to be expected, since combining multidimensional data into a real-valued statistic (e.g. p-value or e-value) requires data compression.

In each experiment $j \in \{1, \dots, m\}$ the observations are summarized by an appropriate real-valued summary statistic $S^{(j)}$. These local test statistics (e.g. p- or e-values) are combined into $C_m(S^{(1)}, \dots, S^{(m)})$. We consider a general class of combination functions C_m, requiring only Hölder type continuity. Roughly speaking, such a restriction assures that C_m does not exploit the richness of the real numbers to encode the data in full. We aim to quantify the loss of summarizing, the gain of performing a meta-analysis and the best testing strategies in the individual experiments meta-analysis. This introduces only a mild restriction, and includes many standard meta-analysis techniques, for instance the standard p-value combination methods (see Section 4.2.1); e-value techniques (see Section 4.2.2); and other ad hoc and natural test statistic combination approaches, see the beginning of Section 4.2 for additional examples.

Our setting provides a principled and unified framework to study the power of standard meta-analysis testing methods. Within the framework of the many-normal-means model, we derive a minimax lower bound for the testing (separation) error and provide test statistics with associated combination methods that attain this theoretical limit (up to a logarithmic factor). Our results reveal that there is a certain unavoidable loss associated with compressing the data of each experiment to a real valued test statistic. We see that while it is always possible to obtain better testing rates using m trials instead of the best possible test based on a single trial, there is always a loss incurred when compared to the full, pooled data and optimal test in moderate- to large dimensional problems. Our theoretical results quantify these gains and losses in terms of the dimension d, sample size n and number of trials m.

Furthermore, we observe an elbow effect, which occurs when the number of trials is large compared to the dimension of the signal. In this regime, combinations of the (locally) optimal test in each individual trial performs suboptimally as a whole when aggregated and meta-analysis approaches based on directional test statistics are shown to perform better. Finally, we show that the performance of the meta-level tests can substantially improve (in certain regimes, depending on d, m, n) if a certain amount of coordination between the trials is allowed (e.g. by having access to the same random

seed). For the theoretical analysis of meta-analysis techniques we derive connections with the distributed statistical learning literature under communication constraints. Our paper builds on the recent information theoretical developments in distributed testing [8, 193, 195], allowing us to address several fundamental questions for the first time with mathematical rigor.

The chapter is organized as follows. In Section 4.1 we introduce the mathematical framework we consider in our investigation and present the corresponding minimax testing lower bound results. Next in Subsection 4.1.1 we show that the derived results are sharp by providing several meta-analysis approaches attaining the limits. Then we investigate the benefits of allowing a mild coordination between the trials in Subsection 4.1.2. We collect and discuss the standard p- and e-value combination methods in Section 4.2 and demonstrate our theoretical results numerically on synthetic data sets in Section 4.3. We discuss our results and derive conclusions in Section 4.4. The proofs of our results are deferred to the Appendix. In Section 4.5.1 we present the proof of our main results while the proofs of the technical lemmas are given in 4.5.2.

4.1 Main results

We recall the distributed version of the many-normal-means model as studied earlier in the thesis, where we tailor the language the meta-analysis setting: We assume that in each local trial $j \in \{1, \dots, m\}$ (in each machine) we observe a d-dimensional random variable $X^{(j)} \in \mathbb{R}^d$, subject to

$$X^{(j)} = f + \frac{1}{\sqrt{n}} Z^{(j)}, \qquad Z^{(j)} \overset{i.i.d.}{\sim} N(0, I_d), \quad j = 1, \dots, m, \tag{4.1}$$

for some unknown $f \in \mathbb{R}^d$. Denote by $\mathbb{P}_f$ the joint distribution of the observations and let $\mathbb{E}_f$ be the corresponding expectation. We note that this framework is equivalent to having n independent $N(f, I_d)$ observations within each local sample.

As in the earlier chapters, our goal is to test the presence or absence of the "signal component" $f \in \mathbb{R}^d$. More formally, we consider the simple null hypothesis $H_0 : f = 0$ versus composite alternative hypothesis $H_\rho : \|f\|_2 \geqslant \rho$, for some $\rho > 0$. This corresponds to testing for joint significance of variables, such as the presence of an effect of a treatment on any of the dimensions investigated. The difficulty in distinguishing the hypotheses depends on the effect size, the sample size and the dimension d. Here, ρ can be seen as the smallest effect size deemed important.

For a $\{0, 1\}$-valued test T, define the testing risk $\mathcal{R}(H_\rho, T)$ as the sum of the Type I error probability and worst case Type II error probability, i.e.

$$\mathcal{R}(H_\rho, T) := \mathbb{P}_0(T = 1) + \sup_{f \in H_\rho} \mathbb{P}_f (T = 0) . \tag{4.2}$$

In the case of a single trial (i.e. $m = 1$), this testing problem is known to have minimax separation rate or "detection boundary" $\rho^2 \asymp \sqrt{d}/n$.

This means that if $\rho^2 \gg \sqrt{d}/n$, there exist consistent[1] tests $T \equiv T_{d,n}$ in the sense that $\mathcal{R}(H_\rho, T) \to 0$, whilst no consistent tests exist when $\rho^2 \ll \sqrt{d}/n$. That is, for effect sizes of smaller order than $\sqrt{d}/n$, the null hypothesis cannot be consistently distinguished from the alternative hypothesis. Such a testing rate is attainable through a chi-square test based on $\|\sqrt{n}X^{(1)}\|_2^2$ (see e.g. [29]).

In case of m trials, if the full data were pooled (with aggregated sample size nm), the minimax separation rate would be $\sqrt{d}/(mn)$. However, pooling the data might not be possible or allowed in practice and often only real-valued test statistics are available that describe the significance in the local problems (e.g. a p- or an e-value). These m test statistics $S^{(j)}$, $j = 1, \dots, m$, then can be combined with some combination function $C_m : \mathbb{R}^m \to \mathbb{R}$, providing the test statistic in the meta-analysis. We now ask whether the above pooled testing rate is attainable with this meta-analysis procedure.

Without any restrictions on the test statistics $S = (S^{(1)}, \dots, S^{(m)})$ or the combination function C_m, any of the conventional optimal "full-data" tests can be reconstructed, since the real numbers and mappings between the real numbers form an overly rich class. We wish to restrict our analysis to S and C_m that are reasonable in practice and capture (most of) the relevant meta-analysis methods as listed in Section 4.2.

Based on each of the local observations $X^{(j)}$, a real-valued test statistic $S^{(j)}$ is computed, where each $S^{(j)}$ is a function of $X^{(j)}$ and possibly a source of randomness $U^{(j)}$ independent of $X := (X^{(1)}, \dots, X^{(m)})$.

Assumption 4.1. *For measurable functions $f_j : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}$ and independent random variables $U^{(1)}, \dots, U^{(m)}$ which are independent of the data X, the j-th test statistic $S^{(j)} = f_j(X^{(j)}, U^{(j)})$ satisfies $\mathbb{E}_0|S^{(j)}| \leqslant M$, for some $M > 0$, $j = 1, \dots, m$.*

We consider Hölder continuous combination functions $C_m : \mathbb{R}^m \to \mathbb{R}$. Arguably, this is the most important assumption in ruling out bijections between $\mathbb{R}^d$ and $\mathbb{R}$. This ensures that a small change in the underlying local test statistics cannot result in a large change in the combination of test statistics $C_m(S^{(1)}, \dots, S^{(m)})$.

Assumption 4.2. *There exist $L, p, q > 0$ such that for all $s, s' \in \mathbb{R}^m$*

$$\big|C_m(s) - C_m(s')\big| \leqslant L\Big(\sum_{j=1}^{m} |s_j - s'_j|^p\Big)^q. \tag{4.3}$$

The special case of $p = 2$ and $q = 1/2$ leads to Lipschitz continuous functions. Assumption 4.1 and Assumption 4.2 should be considered in conjunction. By rescaling and centering test statistics $S^{(j)}$, one can typically obtain test statistics satisfying Assumption 4.1. Rescaling and centering typically does affect how the test statistics need to be combined, which might "break" Assumption 4.2.

Finally, following the standard testing approach, we compare the aggregated test statistics $C_m(S^{(1)}, \dots, S^{(m)})$ to a threshold value. If the combined test statistics

[1] For any asymptotics in ρ, d and n such that $\rho^2 \gg \sqrt{d}/n$.

result in a large enough value, the null hypothesis of no effect is rejected. We note here that two-sided tests can be written as one-sided tests through straightforward transformations (e.g. centering and taking absolute value). More formally, we consider tests T_α of level α satisfying the following assumption.

Assumption 4.3. *There exists a strictly decreasing function $\alpha \mapsto \kappa_\alpha$ so that*

$$T_\alpha = \mathbb{1}\left\{C_m(S^{(1)}, \dots, S^{(m)}) \geqslant \kappa_\alpha\right\} \tag{4.4}$$

satisfies $\mathbb{P}_0 T_\alpha \leqslant \alpha$.

The map $\alpha \mapsto \kappa_\alpha$ could be taken as the quantile function of $C_m(S^{(1)}, \dots, S^{(m)})$ under its null distribution if it is appropriately standardized. If $\mathbb{E}_0 C_m(S^{(1)}, \dots, S^{(m)})$ is bounded in m, we can choose κ_α equal to $1/\alpha$ times the upper bound, in view of Markov's inequality.

Our first main result, Theorem 4.1 below, establishes a lower bound for tests of the form (4.4) and C_m and S satisfying the above assumptions. More concretely, under our assumptions, any test T_α (of level $\alpha \leqslant 0.1$) has large Type II-error under alternatives with ρ^2 of smaller order than $(\sqrt{m} \wedge \frac{d}{\log(m)})\sqrt{d}/(mn)$. When the number of trials is small compared to the dimension (i.e. $m\log^2(m) \leqslant d^2$), this means that the separation rate is at least $\sqrt{d}/(\sqrt{m}n)$. Thus even though there is a benefit in terms of separation rate compared to testing based on just a single trial, the gain is at best the square root of what one would gain based on testing on the pooled data. When $m\log^2(m) \geqslant d^2$, the rate in the lower bound changes to $d\sqrt{d}/(mn\log(m))$, resulting in an elbow effect.

Theorem 4.1. *Let $S^{(1)}, \dots, S^{(m)}$, C_m and T_α satisfy Assumptions 4.1–4.3 with T_α of level $\alpha \in (0, 0.1]$. Then there exists a constant $c > 0$ depending only on L, p, q and M, such that if*

$$\rho^2 \leqslant c\frac{(\sqrt{m} \wedge \frac{d}{\log(m)})\sqrt{d}}{mn}, \tag{4.5}$$

it holds for all $n, m, d \in \mathbb{N}$ that

$$\sup_{f \in H_\rho} \mathbb{P}_f\left(T_\alpha = 0\right) \geqslant 3/4. \tag{4.6}$$

Remark 9. The ranges of values $0 < \alpha \leqslant 0.1$ and $\beta = 3/4$ for the Type I and II errors, respectively, are arbitrary. Similar results hold for different choices as well. For instance, one can take arbitrary $\alpha \in (0, 1/5]$ and $\beta \in (0, 2/3]$, see the proof of the theorem for details. The result implies in particular that consistent testing is not possible for signals of a smaller order than the right-hand side of (4.5), where asymptotics can be considered in n, m and d simultaneously.

In the next section we show that the lower bounds in the theorems above are sharp (up to a logarithmic factor).

4.1.1 Rate optimal combination methods

To attain the lower bound rate derived in Theorem 4.1, different tests can be considered, for example the 1-bit testing strategies considered in Chapter 3. Here, we shall display tests that are based on single p-values, which attain the same rates (but probably outperform at the level of constants). The optimal rate in the setting described here displays a similar elbow effect around $m \asymp d^2$ as in the 1-bit bandwidth constrained case. When the dimension is large compared to the number of trials m (i.e. $m \lesssim d^2$), strategies that combine p-values for the optimal local tests (based on $\|\sqrt{n}X^{(j)}\|_2^2 \sim^{H_0} \chi^2_d$), turn out to achieve the optimal rate, as exhibited below. Such a test statistic is invariant to the directionality of $X^{(j)}$ and invariant under the model in the sense that the resulting power for the alternative $\mathbb{P}_f$ or $\mathbb{P}_g$ is the same as long as $\|f\|_2 = \|g\|_2$.

On the other hand, when the dimension is small compared to the number of trials (i.e. $m \gtrsim d^2$), optimal strategies exhibited below use information on the direction of $X^{(j)}$. In fact, we show in Theorem 4.4 in the Appendix that if no such information is available (i.e. the events defined by the signs of the $(X^{(j)})_{j=1,\dots,m}$ vector are not contained in the sigma algebra generated by the test statistics S), one cannot obtain a rate better than $\sqrt{d}/(\sqrt{m}n)$. This implies that by combining the locally optimal test statistics $S^{(j)} = \|\sqrt{n}X^{(j)}\|_2^2$ (or their arbitrary functions, e.g. the corresponding local p-values) would result in information loss and hence suboptimal rates in the meta-analysis.

Furthermore, it turns out, in accordance with the empirical literature discussed in the introduction, that there does not exist a uniquely best meta-analysis method. In fact, multiple standard meta-analysis techniques provide (up to a logarithmic factor) optimal rates, see below for some standard approaches attaining the lower bounds derived in Theorem 4.1.

First we consider the scenario when the dimension d of the model is large compared to the number of trials m, i.e. $m \lesssim d^2$. Locally the optimal test is based on the test statistic $\|\sqrt{n}X^{(j)}\|_2^2 \overset{H_0}{\sim} \chi^2_d$. A natural way to combine these statistics would be to sum these locally optimal test statistics to obtain

$$T_\alpha = \mathbb{1}\left\{ \sum_{j=1}^{m} \left\| \sqrt{n}X^{(j)} \right\|_2^2 \geqslant F^{-1}_{\chi^2_{dm}}(1-\alpha) \right\}, \tag{4.7}$$

which has level α. Alternatively, one could also apply p-value combination methods, such as Fisher's or Edgington's method based on the p-value $p^{(j)} = 1 - F_{\chi^2_d}(\|\sqrt{n}X^{(j)}\|_2^2)$, see Section 4.2. Lemma 4.6 in the appendix establishes that these tests are rate optimal.

Second, consider the case that the number of trials is large compared to the dimension, i.e. $m \gtrsim d^2$. Rate optimal tests can be constructed based on a variation of Edgington's or Stouffer's method, see Section 4.2 for their descriptions. Taking a partition of

$\{1,\dots,m\} = \cup_{i=1}^{d}\mathcal{J}_i$ where $|\mathcal{J}_i| \asymp m/d$ and setting $S^{(j)} = \sqrt{n}X_i^{(j)}$ if $j \in \mathcal{J}_i$, the meta-level test

$$T_\alpha = \mathbb{1}\left\{\frac{\sqrt{d}}{m}\sum_{i=1}^{d}\left(\sum_{j\in\mathcal{J}_i} S^{(j)}\right)^2 \geqslant d^{-1/2}F^{-1}_{\chi^2_d}(1-\alpha)\right\} \tag{4.8}$$

achieves the lower bounds. The above test is similar to employing Stouffer's method for each of the coordinates and averaging, i.e. computing approximately m/d i.i.d. p-values $p^{(j)} = \Phi(\sqrt{n}X_i^{(j)})$ for $j \in \mathcal{J}_i$ and applying the inverse Gaussian CDF $\Phi^{-1}(p^{(j)})$. Alternatively, the following variation of Edgington's method,

$$T_\alpha = \mathbb{1}\left\{\frac{\sqrt{d}}{m}\sum_{i=1}^{d}\left(\sum_{j\in\mathcal{J}_i}\left(p^{(j)}-\frac{1}{2}\right)\right)^2 \geqslant \kappa_\alpha\right\}, \tag{4.9}$$

is also rate optimal, as proven in Lemma 4.7 in the appendix. Essentially, these strategies divide the trials across the d different directions, and combines the evidence for each of the directions. Theorem 4.4 affirms that the information on the "direction" of the data is crucial to achieve the optimal rate in the $m \gtrsim d^2$ case, by showing that strategies that do not contain such information (rotationally invariant strategies such as norm-based test statistics) achieve the rate $\sqrt{d}/(\sqrt{mn})$ at best. We summarize the above testing upper bounds in the theorem below.

Theorem 4.2. *For all $\alpha, \beta \in (0,1)$ there exist S, $C_m : \mathbb{R}^m \to \mathbb{R}$ and tests T_α of level α satisfying Assumptions 4.1–4.3 such that if*

$$\rho^2 \geqslant C_{\alpha,\beta}\frac{(\sqrt{m}\wedge d)\sqrt{d}}{mn}, \tag{4.10}$$

we have

$$\sup_{f\in H_\rho}\mathbb{P}_f\left(T_\alpha = 0\right) \leqslant \beta$$

for a large enough constant $C_{\alpha,\beta} > 0$ depending only on $\alpha, \beta \in (0,1)$, for all $n, m, d \in \mathbb{N}$.

4.1.2 Benefits of coordination between the trials

When the dimension is small relative to the number of trials, as exhibited in the previous section, optimal strategies include information on the directionality of the observation vector. In this section we show that in this regime, there could be an additional benefit from allowing mild coordination between the trials through employing shared randomness, e.g. a shared random seed between the trials. Such a phenomenon has been observed before in the distributed testing literature [11, 8, 193, 195], which forms the basis of our analysis below.

We consider the following variation on Assumption 4.1, where the key difference is that the source of randomness is allowed to be shared between the m trials.

Assumption 4.4. *For functions $f_j : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}$ and a random variable U which is independent of the data X, the j-th test statistic $S^{(j)} = f_j(X^{(j)}, U)$ satisfies $\mathbb{E}_0|S^{(j)}| \leqslant M$ for some $M > 0$ and all $j = 1, \dots, m$.*

Test statistics satisfying this assumption shall be referred to as shared randomness (or public coin) protocols.

The theorem below establishes the optimal rate when coordination through shared randomness is allowed. When the number of trials is small compared to the dimension (i.e. $m \lesssim d/\log m$), there is no difference between protocols that coordinate using shared randomness or those without coordination. In fact, the optimal rate ($\rho^2 \asymp \sqrt{d}/(\sqrt{m}n)$) in this case is reached by the test (4.7) or the ones below it, which do not employ shared randomness. However, when the number of trials is large compared to the dimension (i.e. $m \gtrsim d$), the testing rate substantially improves in the shared randomness protocols.

Theorem 4.3. *Let $S^{(1)}, \dots, S^{(m)}$, C_m and T_α satisfy Assumptions 4.2–4.4. Then there exists a constant $c > 0$ depending only on L, p, q and M, such that if*

$$\rho^2 \leqslant c \frac{\left(\sqrt{m} \wedge \sqrt{\frac{d}{\log(m)}}\right)\sqrt{d}}{mn}, \tag{4.11}$$

it holds that $\sup_{f \in H_\rho} \mathbb{P}_f(T_\alpha = 0) > 2/3$ for all $n, m, d \in \mathbb{N}$ and any level $\alpha \in (0, 0.1]$.

At the same time, for all $\alpha, \beta \in (0,1)$ there exists a constant $C_{\alpha,\beta} > 0$ depending only on β, L, p, q, the function $\alpha \mapsto \kappa_\alpha$ and M, such that if

$$\rho^2 \geqslant C_{\alpha,\beta} \frac{\left(\sqrt{m} \wedge \sqrt{d}\right)\sqrt{d}}{mn} \tag{4.12}$$

it holds that $\sup_{f \in H_\rho} \mathbb{P}_f(T_\alpha = 0) \leqslant \beta$ for some test T_α of level α satisfying Assumptions 4.2–4.4.

Remark 10. Similarly to Theorem 4.1 the choice of ranges $0 < \alpha \leqslant 0.1$ and $\beta = 2/3$ in the lower bound result is arbitrary, other choices are also possible as presented in the proof.

A shared randomness method that attains the rate in (4.12) is given next. Consider drawing an orthonormal $d \times d$ matrix U taking values from the uniform measure on such matrices. As a test statistic, each trial computes $(\sqrt{n}UX^{(j)})_1$, which is a $N(0,1)$ random variable under the null hypothesis. A level $\alpha \in (0,1)$ meta-level test is then given by combining the local test statistics as

$$T_\alpha := \mathbb{1}\left\{ \left| \frac{1}{\sqrt{m}} \sum_{j=1}^{m} (\sqrt{n}UX^{(j)})_1 \right| \geqslant \Phi^{-1}(1 - \alpha/2) \right\}, \tag{4.13}$$

where Φ is the standard Gaussian CDF. The core idea here is that for each trial, the same 1-dimensional projection of the d -dimensional data is computed, where the projection is taken uniformly at random and the test is conducted along the projected direction. The above method corresponds to Stouffer's method for the p-values $p^{(j)} = \Phi(\sqrt{n}(UX^{(j)})_1)$ for $j = 1, \ldots, m$. Lemma 4.8 in the appendix shows that the above test attains a small Type II error probability whenever $\rho^2 \gtrsim d/(mn)$.

4.2 Examples for various meta-analysis methods

Combinations of independent test statistics that fall into the framework of Assumptions 4.1– 4.4 are subject to the rate optimality theory established by the main theorems in Section 4.1. In this section, we look into common methods for combining p-values, e-values and other test-statistics, as mentioned in the introduction.

When the distribution under the null hypothesis of the test statistics are known, certain combinations are natural. For example, the sum of normal or chi-square test statistics is again normal or chi-square distributed, respectively. Similarly, voting based mechanisms typically rely on summing Bernoulli random variables. It is easy to see that these and similar combinations methods fall into the framework of Assumptions 4.1–4.4.

For more specific test statistics, such as p-values or e-values, many general combination methods have been introduced in the literature. We cover some of the most prominent combination approaches for p-values and e-values in Section 4.2.1 and Section 4.2.2, respectively. The list of methods is certainly non-exhaustive and many more combination methods exist, but they serve as context for the range of techniques covered by our general theory. Our main results allow establishing lower bound rates for the ones listed below, whilst in Sections 4.1.1 and 4.1.2 attainability of these rates by some of the listed methods was exhibited.

4.2.1 Combinations of p-values

If $p^{(1)}, \ldots, p^{(m)}$ are p-values obtained from m independent test statistics concerning the same hypothesis, then under the null $p^{(j)} \sim^{i.i.d.} U(0,1)$. One can aim to combine the m p-values to form a test $T_\alpha \equiv T_\alpha(p^{(1)}, \ldots, p^{(m)})$ with Type I error probability α, which hopefully has higher power than a test based on one of the individual p-values. Below we list standard methods in the literature.

- Fisher's method [100]. Because the variables $-2\log p^{(j)}$'s are i.i.d. χ^2_2-distributed under the null hypothesis, their sum follows a χ^2_{2m}-distribution. Therefore, the combination method $\sum_{j=1}^m -2\log p^{(j)}$ results in a χ^2_{2m} distributed random variable, and the corresponding quantile function provides level-α one-sided tests at the meta-level.

- Similar flavor to Fisher's method are the combinations $\sum_{j=1}^{m} -\log(1-p^{(j)})$ (Pearson's method [164]), $\sum_{j=1}^{m} -\log p^{(j)}(1-p^{(j)})$ (the logit method / Mudholkar and George method [155]) and $m^{-1/2}\sum_{j=1}^{m}(p^{(j)}-1/2)$ (Edgington's method [88]).

- Order-based methods such as Tippett's method [202] based on $\min\{p^{(1)},\dots,p^{(m)}\} \overset{H_0}{\sim} \mathrm{Beta}(1,m)$.

- Methods based on inverse CDF's, such as by Stouffer et al. [185] based on $m^{-1/2}\sum_{j=1}^{m}\Phi^{-1}(p^{(j)}) \sim \mathrm{N}(0,1)$ under the null hypothesis.

- Generalized averages as considered in [212], $T_\alpha = \mathbb{1}\left\{a_{r,m}M_{r,m}(p^{(1)},\dots,p^{(m)}) \leqslant \alpha\right\}$, where $M_{r,m}(p^{(1)},\dots,p^{(m)})$ equals $\left(m^{-1}\sum_{j=1}^{m}(p^{(j)})^r\right)^{1/r}$ for $r \in \mathbb{R}\backslash\{0\}$, the geometric mean, minimum (i.e. Tippett's method) and maximum for $r = 0$, $r \to -\infty$, and $r \to \infty$, respectively. For $r \in \{-\infty\} \cup [1/(m-1),\infty]$, $a_{r,m}$ can be taken to obtain precisely level α tests (i.e. $\mathbb{P}_0 T_\alpha = \alpha$). We note that this means that canonical multiple testing methods (see e.g. [102]) such as Bonferroni's correction (which corresponds with taking as $M_{r,m}$ the minimum and $a_{r,m} = m$) also fall within our framework.

Lemma 4.1 below shows that all the methods mentioned above fall into the framework of Assumptions 4.1–4.4. This means that the error rate lower bounds of Theorem 4.1 and Theorem 4.3, respectively, apply to the p-value combination methods listed above. That is, one cannot attain a better separation rate when considering the worst case Type II error probability for the alternative hypothesis in (4.2), with any of the p-value combination methods listed above. Whether Assumption 4.1 or 4.4 applies depends on whether shared randomness is used in generating the p-values. To confirm that Assumptions 4.3 and 4.2 apply to tests based on the combined p-values, some algebra is needed. The proof of the lemma is deferred to the appendix.

Lemma 4.1. *Consider p-values $p^{(1)},\dots,p^{(m)}$, where each $p^{(j)}$ depends on the local data $X^{(j)}$ and possibly local randomness that is independent of the data. For each of the combination methods for p-values mentioned above and corresponding test T_α of level $\alpha \in (0,1)$, the conclusions of Theorem 4.1 holds.*

We remark that the p-values are obtained using shared randomness (i.e. in the sense of Assumption 4.1), the lower bound rate of Theorem 4.3 applies. Furthermore, as exhibited in Sections 4.1.1 and 4.1.2, for p-values corresponding to well-chosen test statistics, these combination methods can achieve the theoretical limits established in Theorems 4.1 and 4.3, respectively.

4.2.2 Combining e-values

An *e-value* is a nonnegative random variable E such that $\sup_{\mathbb{P}_0 \in H_0} \mathbb{P}_0 E \leqslant 1$. The *threshold test corresponding to E of level α* is $\mathbb{1}\{E \geqslant \alpha^{-1}\}$. This test yields a so called strict p-value; for $\mathbb{P}_0 \in H_0$ we have $\mathbb{P}_0(E \geqslant \alpha^{-1}) \leqslant \alpha$ by Markov's inequality.

E-values lend themselves for combining outcomes of independent studies for two main reasons. First, they are easy to combine, see Section 4 in [213] for an in-depth discussion of specific combination functions for independent e-values. Second, they are robust to misspecification and offer optional stopping/continuation guarantees [111]. Common examples of e-values are Bayes factors and likelihood ratios, which are non-negative and have expectation equal to 1 in the case of a simple null hypothesis such as considered in this article.

Several combination methods (e-merging functions) were proposed in the literature. For instance, the product of independent e-values is also again an e-value. This was shown to weakly dominate any other combination of independent e-values in the sense that $\Pi_{j=1}^m E^{(j)} \geqslant C_m(E)$, for any $E = (E^{(j)}) \in [1,\infty)^m$ and $E \mapsto C_m(E)$ such that $C_m(E^{(1)},\dots,E^{(m)})$ is an e-value for any independent e-values $E^{(1)},\dots,E^{(m)}$, see [213]. Similarly, the average of e-values is again an e-value. The product and the average are *admissible* in the sense that there is no e-merging function that strictly dominates them on $[0,\infty]^m$. The lemma below shows that these two, arguably most prominent e-value combination methods fulfill Assumptions 4.1– 4.4 and hence the lower bounds derived in Theorems 4.1 and 4.3 apply.

Lemma 4.2. *Consider e-values $E^{(1)},\dots,E^{(m)}$, where each $E^{(j)}$ depends on the local data $X^{(j)}$ and possibly local randomness that is independent of the data. Let $C_m : \mathbb{R}^m \to \mathbb{R}$ correspond to either the average or the product and let T_α be the corresponding threshold test of level $\alpha \in (0,1)$,*

$$T_\alpha = \mathbb{1}\left\{C_m(E^{(1)},\dots,E^{(m)}) \geqslant \alpha^{-1}\right\}.$$

If C_m is the product, assume in addition that $\mathbb{E}_0|\log E^{(j)}|$ is uniformly bounded. Then, the conclusion of Theorem 4.1 holds. In case the e-values are generated using shared randomness, then Theorem 4.3 applies.

4.3 Simulations

In this section, we investigate the numerical performance of the testing strategies outlined in Section 4.1.1 on synthetic data sets. We compare the tests based on their receiver operating characteristic (ROC) curve. For a range of significance levels we compute for each test the "true positive rate" (TPR) and "false positive rates" or (FPR), i.e. the fraction of the simulation runs in which the test correctly identifies the underlying signal, falsely rejects the null hypothesis, respectively. Plotting the TPR against the FPR (both given as a function of the significance level) provides us the ROC curve, visualizing the diagnostic ability of the test.

In our simulations we set $m = 20$, $n = 30$, let d range from 2 to 20 and take $\rho^2 = \sqrt{d}/(4n)$. This value of ρ^2 corresponds to a signal that is almost indistinguishable from noise using just a single trial, whilst consistently detectable if the data were to be pooled with $m \approx 20$ (which increases the signal size to noise ratio effectively by

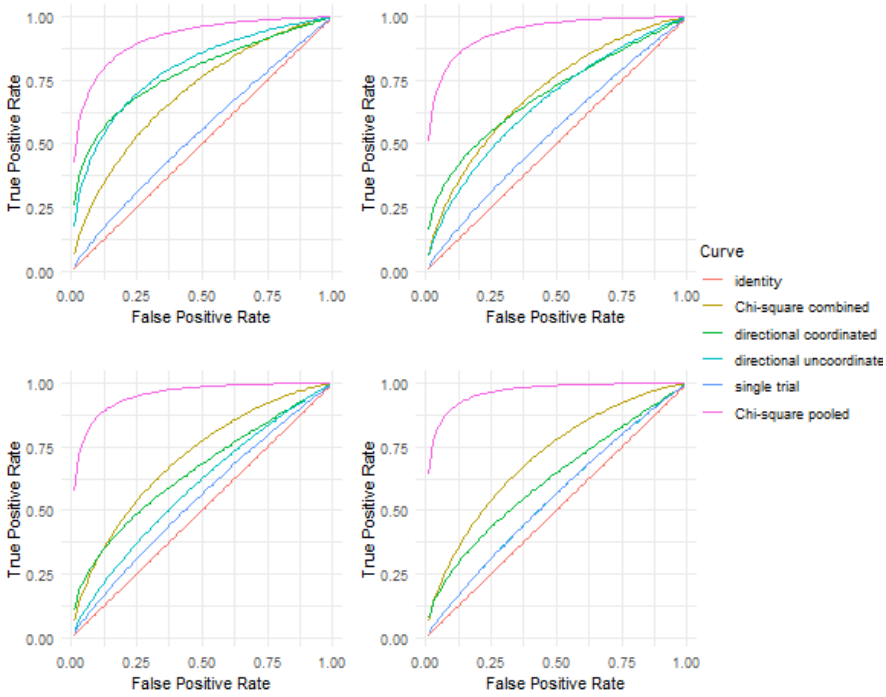


Figure 4.1: ROC curves for different values of d, whilst keeping $m = 20$, $n = 30$, $\rho^2 = \sqrt{d}/(4n)$. From left to right, top to bottom: $d = 2$, $d = 5$, $d = 10$, $d = 20$.

a factor $\sqrt{20} > 4$). For each level $\alpha \in \{0.01, 0.02, \ldots, 0.99\}$ we compute the power for different combination strategies 100 times, each time drawing a different $f \in \mathbb{R}^d$ with $\|f\|_2 = \rho$ according to $f_i = d^{-1/2}\rho R_i$ and R_i i.i.d. Rademacher random variables for $i = 1, \ldots, d$. As combination strategies, we compare the strategies (4.7), (4.13) and (4.8) from Section 4.1, which are called "chi-square combined", "coordinated directional" and "uncoordinated directional" in the legend of Figure 4.3. In addition, we display the ROC curves for the chi-square test based on pooled data ("chi-square pooled") and that of a single trial ("single trial").

We make the following observations, in line with our theoretical findings. The meta-analysis methods based on combining the locally optimal chi-squared test statistics (yellow curves) substantially out-performed the chi-squared test statistics based on a single trial (blue curve), but was substantially worse than the chi-square test based on the pooled data (pink curve). Second note that the large dimensional case ($d = 10$ and $d = 20$) the best strategy is indeed to combine the local chi-square statistics (yellow curve), while in the low dimensional setting ($d = 2$) it is more advantageous

to combine the directional test statistics $X_i^{(j)}$ (blue curve). Finally, note that allowing coordination between the trials by a shared randomness protocol can result in improved performance (green curve) compared to the independent experiments (blue curve). In fact this approach provides the best meta-analysis method in the small dimensional setting (e.g. $d = 2$ and $d = 5$ for small α, which is the most interesting case).

In the appendix, Section 4.5.6, we explore eight additional simulation settings, where we consider larger values of d and m. Whilst these simulations do not reveal additional phenomena to the ones observed in Figure 4.3, they do give insight into the relative performance of the testing methods for different values of d and m.

4.4 Discussion

We briefly summarize the main contributions of the chapter and discuss possible extensions and research directions. First, by establishing a connection between meta-analysis and distributed learning under communication constraints, we have provided a unified, theoretical framework for evaluating the behavior of standard meta-analysis techniques. In our analysis, we considered the many-normal-means model, but these results can be extended to other more complex models as well, building on the connection with distributed computation. For example, minimax estimation rates under communication constraints were derived for other parametric models [229], density estimation [30], signal-in-Gaussian-white-noise [231, 190, 47], nonparametric regression [189] and in abstract settings [226] including binary and Poisson regression, drift estimation, and more. The normal means model allows for a tractable analysis, but results in this model are known to extend to more complicated models, such as discrete density testing (see e.g. [57]). With the due technical work, our results are expected to translate to these settings as well, but we leave this for future endeavor.

In the normal means model we have shown that by combining the locally optimal chi-square statistics at a meta-level one can gain a factor of $\sqrt{m}$ compared to using a single trial. Nevertheless, regardless of the choice of the combination method, a factor of $\sqrt{m} \wedge \sqrt{d}$ is lost compared to the scenario when all data from all trials are at our disposal. This loss is clearly visible even in small sample sizes, dimensions and trial numbers, as demonstrated in our numerical analysis, as can be seen in the corresponding ROC curves. For more complex models, such a numerical study can be a first step to quantify the efficiency of the meta-analysis method. We have also shown that in the small dimension - large number of trials setting combining the locally optimal chi-square statistics (or any rotationally invariant statistics for that matter) results in information loss and suboptimal accuracy. In this case, better rates can be attained by test statistics based on the direction of the observations combined at the meta-level. In practice, one often cannot choose which test statistics can be obtained from independent trials. In such cases, the $\sqrt{m}$-factor loss in the case of e.g. rotationally invariant test statistics is of interest when considering power calcu-

lations. Meta-analysis approaches based on directional test statistics are designed for scenarios where individual datasets are not centrally collected, but there is some level of coordination among experimenters.

The assumption throughout the paper of homogeneity between the trials (i.e. each trial consisting of the same number of observations) simplifies the presentation, but the results can be extended to cases where the number of observations in each trial differ by constant factors. Situations where the number of observations differs greatly (e.g. $k \ll m$ trials have as many observations as the other $m-k$ trials combined) are certainly of interest, but beyond the scope of the thesis.

4.5 Appendix

The proofs of the main theorems (Theorem 4.1, 4.2 and 4.3) are divided over the subsections as follows. In Section 4.5.1, the lower bounds of Theorem 4.1 and 4.3 are proven. Auxiliary lemmas for the proof of the lower bounds are proven in 4.5.2. The attainability of the lower bound rates are given in Lemmas 4.6, 4.7 and 4.8 in Section 4.5.4. In Section 4.5.5, Lemmas 4.1 and 4.2 are proven.

4.5.1 Proof of the lower bounds (Theorems 4.1 and 4.3)

The proof is based around the following idea. If C_m satisfies the continuity condition of Assumption 4.2, it implies $C_m(S^{(1)}, \dots, S^{(m)})$ should not change too much if the statistics $S^{(1)}, \dots, S^{(m)}$ are replaced by finite bit approximations. If b is the number of bits used for the approximation of $S^{(j)}$, we should be able to get an approximation with accuracy of the order 2^{-b} through e.g. binary expansion. Since C_m and consequently the test based on C_m do not change (much) from passing to a finite bit approximation, tools and results from testing under bit-constrained communication apply, which finally yield the theorems.

Proof. We prove the statement for any $\alpha \in (0, 1/10]$. Since $\alpha \mapsto \kappa_\alpha$ is strictly decreasing, $\kappa_{1/8} < \kappa_{1/10} \leqslant \kappa_\alpha$ holds for any $\alpha \in (0, 1/10]$. Take $0 < \epsilon < \frac{1}{2}(\kappa_{1/10} - \kappa_{1/8})$. Then $|x - \kappa_\alpha| \leqslant \epsilon$ implies $x \geqslant \kappa_{1/8}$, which by the definition of the quantile function provides

$$\mathbb{P}_0 \left(|C_m(S) - \kappa_\alpha| \leqslant 2\epsilon \right) \leqslant 1/8. \tag{4.14}$$

By Lemma 4.3, there exist $B^{(j)}$-bit binary approximations $\tilde{S}^{(j)}$ such that

$$|S^{(j)} - \tilde{S}^{(j)}| \leqslant \left(\frac{\epsilon^{1/q}}{L^{1/q} m} \right)^{1/p} \tag{4.15}$$

and

$$\mathbb{E}_0 B^{(j)} \leqslant \mathbb{E}_0 \log_2(|S^{(j)}|) \vee 0 - \frac{1}{p} \log \left(\frac{\epsilon^{1/q}}{L^{1/q} m} \right) + 3. \tag{4.16}$$

Write $\tilde{S} = (\tilde{S}^{(1)}, \ldots, \tilde{S}^{(m)})$. By combining Assumption 4.2 with (4.15),

$$|C_m(S) - C_m(\tilde{S})| \leqslant \epsilon.$$

Consequently,

$$\begin{aligned}\mathcal{R}(T_\alpha, H_\rho) &\geqslant \mathbb{P}_0\left(C_m(\tilde{S}) - |C_m(S) - C_m(\tilde{S})| \geqslant \kappa_\alpha\right) \\ &\quad + \sup_{f \in H_\rho} \mathbb{P}_f\left(C_m(\tilde{S}) \leqslant \kappa_\alpha - |C_m(S) - C_m(\tilde{S})|\right) \\ &\geqslant \mathbb{P}_0\left(C_m(\tilde{S}) \geqslant \kappa_\alpha + \epsilon\right) + \sup_{f \in H_\rho} \mathbb{P}_f\left(C_m(\tilde{S}) \leqslant \kappa_\alpha - \epsilon\right).\end{aligned}$$

Define the test

$$T'_\alpha := \mathbb{1}\left\{C_m(\tilde{S}) > \kappa_\alpha - \epsilon\right\}.$$

Since

$$\mathbb{P}_0\left(C_m(\tilde{S}) \geqslant \kappa_\alpha + \epsilon\right) = \mathbb{P}_0\left(C_m(\tilde{S}) > \kappa_\alpha - \epsilon\right) - \mathbb{P}_0\left(-\epsilon \leqslant C_m(\tilde{S}) - \kappa_\alpha \leqslant \epsilon\right),$$

the second last display can now be written as

$$\mathcal{R}(T', H_\rho) - \mathbb{P}_0\left(|C_m(\tilde{S}) - \kappa_\alpha| \leqslant \epsilon\right).$$

Applying (4.3) again, using the reverse triangle inequality and (4.14), we obtain

$$\mathbb{P}_0\left(|C_m(\tilde{S}) - \kappa_\alpha| \leqslant \epsilon\right) \leqslant \mathbb{P}_0\left(|C_m(S) - \kappa_\alpha| \leqslant 2\epsilon\right) \leqslant 1/8.$$

It suffices to show that for ρ satisfying (4.5) in the case of Theorem 4.1 or ρ satisfying (4.11) in case of Theorem 4.3 for a small enough $c > 0$, we have

$$\mathcal{R}(T', H_\rho) \geqslant 7/8. \tag{4.17}$$

This follows from Lemma 4.4, where it is left to verify that

$$\sum_{j=1}^{m} d \wedge \mathbb{E}_0 B^{(j)} \lesssim m(d \wedge (1 \vee \log m)) \tag{4.18}$$

for a constant independent of d, n, m and $c > 0$. By (4.16) and $\mathbb{E}_0|S^{(j)}| \leqslant M$ for some constant $M > 0$ for $j = 1, \ldots, m$ (following from Assumption 4.1 or 4.4), we obtain that $\sum_{j=1}^{m} d \wedge \mathbb{E}_0 B^{(j)}$ is bounded by

$$m\left(d \bigwedge \left(\log_2(1+M) + 3 - \frac{1}{p}\log\left(\frac{\epsilon^{1/q}}{L^{1/q} m}\right)\right)\right),$$

from which (4.18) follows. Putting things together, we now have that for $c > 0$ small enough we obtain (4.17), from which we conclude that (4.17) holds and the proof of the theorems is concluded. □

4.5.2 Auxiliary lemmas to the lower bound theorems

As a first tool, we introduce finite bit approximations of real numbers through their binary expansion. Consider the binary expansion of $x \in \mathbb{R}$; i.e. there exist digits $a_k(x), \ldots, a_1(x), a_0(x) \in \{0,1\}$ for a $k_x \equiv k \in \mathbb{N} \cup \{0\}$ and $(b_i(x))_{i\in\mathbb{N}} \in \{0,1\}^{\mathbb{N}}$ such that

$$x = \operatorname{sign}(x) \left(\sum_{i=0}^{k} 2^i a_i(x) + \sum_{i=1}^{\infty} 2^{-i} b_i(x) \right) \tag{4.19}$$

with k the largest element in $\mathbb{N} \cup \{0\}$ such that $2^k - 1 \leqslant |x|$. We now define $\tilde{x}_B$ to be the B-bit binary expansion giving the smallest approximation error in absolute value, where the first bit encodes $\operatorname{sign}(x)$. That is, for $B \geqslant k + 2$, we have

$$|x - \tilde{x}_B| \leqslant \sum_{i=B-k-1}^{\infty} 2^{-i} b_i(x). \tag{4.20}$$

The following is well known, we exhibit its proof for completeness.

Lemma 4.3. *Let V be a random variable with a first moment. Given $1 > \epsilon > 0$, let $B_\epsilon \equiv B$ denote the number of bits required such that*

$$|V - \tilde{V}_{B_\epsilon}| \leqslant \epsilon. \tag{4.21}$$

It holds that

$$\mathbb{E}B \leqslant \mathbb{E}\log_2(|V|) \vee 0 + 1 + \log_2(1/\epsilon) + 2.$$

Proof. If $|V| < 1$, we have that

$$|V - \tilde{V}_B| \leqslant \sum_{i=B-1}^{\infty} 2^{-i} b_i(V).$$

So in the case that $|V| \leqslant 1$, since $b_i(V) \in \{0,1\}$, for (4.21) to hold it suffices that $B \geqslant \log_2(1/\epsilon) + 2$. Let B' denote the amount of bits required to obtain $|V - \tilde{V}_{B'}| \leqslant 1$. When $2^k \leqslant |V| < 2^{k+1}$, it holds that $B' \leqslant k + 1$. Using Markov's inequality,

$$\begin{aligned}
\mathbb{E}B' &= \mathbb{E}B' \sum_{k=0}^{\infty} \mathbb{1}\left\{2^k \leqslant |V| < 2^{k+1}\right\} \\
&\leqslant \mathbb{E} \sum_{k=0}^{\infty} (k+1) \mathbb{1}\left\{k \leqslant \log_2(|V|) < k+1\right\} \leqslant \mathbb{E}\log_2(|V|) \vee 0 + 1.
\end{aligned}$$

In conclusion, $\mathbb{E}B \leqslant \mathbb{E}\log_2(|V|) \vee 0 + 1 + \log_2(1/\epsilon) + 2$. □

For the lemmas below, we introduce the following notation. Let π be a probability distribution on $\mathbb{R}^d$. Write $\mathbb{P}_\pi := \int \mathbb{P}_f d\pi(f)$ for the mixture distribution, where $\mathbb{P}_f$ denotes the joint distribution on X, U and S. Let F denote the draw from π. Let

$\mathbb{P}_f^{\tilde{S}}$ denote the forward measure induced on the random variable $\tilde{S}$ and let $L_\pi^{\tilde{S}}$ denote the likelihood ratio of the mixture distribution and $\mathbb{P}_0$, i.e.

$$L_\pi^{\tilde{S}} = \int \frac{d\mathbb{P}_f^{\tilde{S}}}{d\mathbb{P}_0^{\tilde{S}}} d\pi(f). \tag{4.22}$$

Because of the Markov chain structure of $F \to (X, U) \to S$ and the independence between U and X, the joint distribution of (X, U, S) under the mixture disintegrates as

$$d\mathbb{P}_\pi^{X,U,S}(x, u, s) = \int d\mathbb{P}^{S|(X,U)}(s) d\mathbb{P}_f^X(x) d\mathbb{P}^U(u) d\pi(f) \tag{4.23}$$

where $\mathbb{P}^U$ is the marginal distribution of U. For the likelihood ratio conditionally on $U = u$, we shall write

$$L_\pi^{\tilde{S}|U=u} = \int \frac{d\mathbb{P}_f^{\tilde{S}|U=u}}{d\mathbb{P}_0^{\tilde{S}|U=u}} d\pi(f). \tag{4.24}$$

Furthermore, by the independence of the statistics given U,

$$d\mathbb{P}^{S|(X,U)} = \bigotimes_{j=1}^m d\mathbb{P}^{S^{(j)}|(X^{(j)},U)}. \tag{4.25}$$

Let $\tilde{S}^{(j)}$ denote the $B^{(j)}$-bit binary approximations to $S^{(j)}$ such that (4.15) holds. Note that the above displays are true for the random variable $\tilde{S} = (\tilde{S}^{(1)}, \ldots, \tilde{S}^{(m)})$ in place of S since $F \to (X, U) \to S \to \tilde{S}$ forms a Markov chain as well. The following lemma allows us to bound the chi-square divergence between the forward measure for $\tilde{S}$, which we will denote by $\mathbb{P}_\pi^{\tilde{S}}$ and $\mathbb{P}_0^{\tilde{S}}$.

The following lemma lower bounds the worst-case risk for any test T' depending only on $\tilde{S}$, the binary approximation of S as in (4.15).

Lemma 4.4. *Let T' be a test depending only on $\tilde{S}$ taking values in $\mathbb{R}^m$, satisfying* (4.23) *and where $\tilde{S}^{(j)}$ allows for an exact $B^{(j)}$-bit binary expansion as in* (4.19)*, with $\mathbb{E}_0 B^{(j)} < \infty$ for $j = 1, \ldots, m$.*

There exists $c > 0$ independent of n, m and d such that

$$\mathcal{R}(T', H_\rho) \geqslant 7/8$$

for all $n, m, d \in \mathbb{N}$ whenever

$$\sum_{j=1}^m d \wedge \mathbb{E}_0 B^{(j)} \lesssim m(d \wedge \log m) \tag{4.26}$$

in addition to

$$\rho^2 \leqslant c \frac{(\sqrt{m} \wedge \frac{d}{\log(m)})\sqrt{d}}{mn}, \tag{4.27}$$

if $\tilde{S}$ is generated using public randomness, or

$$\rho^2 \leqslant c\frac{(\sqrt{m} \wedge \sqrt{\frac{d}{\log(m)}})\sqrt{d}}{mn}, \tag{4.28}$$

in case $\tilde{S}$ is generated using only local randomness.

Proof. Consider a probability distribution π on $\mathbb{R}^d$ and $L_\pi^{\tilde{S}}$ as in (4.22). Consider the set

$$D := \left\{ u : \sum_{j=1}^m d \wedge \mathbb{E}_0[B^{(j)}|U=u] \leqslant 64 \sum_{j=1}^m d \wedge \mathbb{E}_0 B^{(j)} \right\},$$

whose complement, D^c, has $\mathbb{P}^U$-mass less than or equal to $1/64$ by Markov's inequality and $\mathbb{E}^U(d \wedge \mathbb{E}_0[B^{(j)}|U]) \leqslant d \wedge \mathbb{E}_0 B^{(j)}$. By conditioning on U (writing $\mathbb{P}_0^{|U=u} := \mathbb{P}_0(\cdot|U=u)$),

$$\begin{aligned}\mathcal{R}(T', H_\rho) &\geqslant \mathbb{P}_0 T' + \mathbb{P}_\pi(1-T') - \pi(f \notin H_\rho) \\ &\geqslant \int \left(\mathbb{P}_0^{|U=u}(T') + \mathbb{P}_\pi^{|U=u}(1-T') \right) \mathbb{1}_D(u) d\mathbb{P}^U(u) - \pi(f \notin H_\rho).\end{aligned}$$

Since $0 \leqslant T' \leqslant 1$ and $L_\pi^{\tilde{S}} \geqslant 0$, for all $0 < \gamma < 1$,

$$\begin{aligned}\mathbb{P}_0^{|U=u}(T') + \mathbb{P}_\pi^{|U=u}(1-T') &\geqslant \mathbb{P}_0^{|U=u}\left(\gamma T' + L_\pi^{\tilde{S}|U=u}(1-T') \mathbb{1}\left\{ L_\pi^{\tilde{S}|U=u} > \gamma \right\} \right) \\ &\geqslant \gamma \mathbb{P}_0^{|U=u}\left(L_\pi^{\tilde{S}|U=u} > \gamma \right) \\ &\geqslant \gamma\big(1 - \mathbb{P}_0^{|U=u}(|L_\pi^{\tilde{S}|U=u} - 1| \geqslant 1-\gamma)\big).\end{aligned}$$

The probability on the right-hand side of the above display can be bounded by applying Chebyshev's inequality and bounding the resulting chi-square divergence using the tools of [195], in particular using Lemma 10.1 from the aforementioned paper. This lemma applies if $\tilde{S}$ takes values in a space of finite, fixed cardinality.

Define $B^* = \sum_{j=1}^m 64\mathbb{E}_0|B^{(j)}|$ and the event

$$A := \left\{ \sum_{j=1}^m B^{(j)} \leqslant B^* \right\},$$

so that A^c by Markov's inequality occurs with $\mathbb{P}_0$-probability less than $1/64$.

Let $\breve{S}^{(j)}$ be the $\breve{B}^{(j)} := B^{(j)} \wedge B^*$ binary approximation of $\tilde{S}^{(j)}$ and note that on the event A, $\breve{S}^{(j)} = \tilde{S}^{(j)}$. We have

$$\int \mathbb{P}_0^{|U=u}\left(|L_\pi^{\tilde{S}|U=u} - 1| \geqslant 1-\gamma\right) \mathbb{1}_D(u) d\mathbb{P}^U(u) \leqslant$$
$$\int \mathbb{P}_0^{|U=u}\left(\left\{|L_\pi^{\tilde{S}|U=u} - 1| \geqslant 1-\gamma\right\} \cap A\right) \mathbb{1}_D(u) d\mathbb{P}^U(u) + \mathbb{P}_0(A^c) \leqslant$$
$$\int \mathbb{P}_0^{|U=u}\left(|L_\pi^{\breve{S}|U=u} - 1| \geqslant 1-\gamma\right) \mathbb{1}_D(u) d\mathbb{P}^U(u) + 1/64,$$

where $\breve{S} = (\breve{S}^{(1)}, \ldots, \breve{S}^{(m)})$. Using (4.23) and Chebyshev's inequality, it suffices to show that on the event D, $\mathbb{E}_0^{|U=u}|L_\pi^{\breve{S}|U=u} - 1|^2$ is smaller than $\frac{1}{32}(1-\gamma)^2$ for c small enough when ρ satisfies (4.27) or (4.28), some $\gamma \geqslant 5/6$ for a specific choice of π. By Lemma 4.5, such a distribution π exists, satisfying $\pi(f \notin H_\rho) \leqslant 1/32$, as long as $\mathrm{Tr}(\Xi_u)$ can be sufficiently bounded, which can be done in terms of (4.16), as we will show next.

Let $\mathcal{S}^{(j)}(b,u)$ be the space in which $\breve{S}^{(j)}|[B^{(j)} = b, U = u]$ takes values. Write

$$V_{s,u} = \mathbb{E}_0\left[X^{(j)} \middle| \breve{S}^{(j)} = s, U = u\right].$$

We have

$$\begin{aligned}\Xi_u^j &= \sum_s V_{s,u} V_{s,u}^\top \mathbb{P}_0(\breve{S}^{(j)} = s | U = u) \\ &= \sum_{b\in\mathbb{N}} \mathbb{P}_0(\breve{B}^{(j)} = b | U = u) \sum_{s=\mathcal{S}^j(b,u)} \mathbb{P}_0(\breve{S}^{(j)} = s | \breve{B}^{(j)} = b, U = u) V_{s,u} V_{s,u}^\top.\end{aligned}$$

By Lemma 2.11, the trace of the matrix

$$\sum_{s\in\mathcal{S}^j(b,u)} \mathbb{P}_0\left(\breve{S}^{(j)} = s | \breve{B}^{(j)} = b, U = u\right) V_{s,u} V_{s,u}^\top$$

is bounded by $\left(2\log(2)\frac{b}{d} \bigwedge 1\right)\frac{d}{n}$. By linearity of the trace operation,

$$\begin{aligned}\mathrm{Trace}(\Xi_u^j) &= \sum_{b\in\mathbb{N}} \mathbb{P}_0\left(\breve{B}^{(j)} = b | U = u\right)\left(2\log(2)\frac{b}{d} \bigwedge 1\right)\frac{d}{n} \\ &\leqslant 2\log(2)\frac{d \wedge \mathbb{E}_0[\breve{B}^{(j)}|U=u]}{n}\end{aligned}$$

and consequently, since $\breve{B}^{(j)} \leqslant B^{(j)}$ and $u \in D$,

$$\begin{aligned}\mathrm{Trace}(\sum_{j=1}^m \Xi_u^j) &\leqslant 2\log(2) n^{-1} \sum_{j=1}^m d \wedge \mathbb{E}_0\left[\breve{B}^{(j)} | U = u\right] \\ &\leqslant 128\log(2) n^{-1} \sum_{j=1}^m d \wedge \mathbb{E}_0\left[B^{(j)}\right].\end{aligned}$$

The result follows after using that ρ^2 satisfies (4.28) and (4.27) in the case of local or shared randomness protocols, respectively. □

Lemma 4.5. *Let $L_\pi^{\breve{S}}$ be as defined through (4.22), with $\breve{S} = (\breve{S}^{(1)}, \dots, \breve{S}^{(m)})$ taking values in a space of finite cardinality. Let $\Xi_u = \sum_{j=1}^m \Xi_u^j$ with*

$$\Xi_u^j := \mathbb{E}_0^{|U=u} \mathbb{E}_0 \left[X^{(j)} \middle| \breve{S}^{(j)}, U = u \right] \mathbb{E}_0 \left[X^{(j)} \middle| \breve{S}^{(j)}, U = u \right]^\top. \tag{4.29}$$

Let ρ^2 satisfy (4.27) or (4.28). For $c > 0$ small enough (in (4.27) or (4.28)) there exists a probability distribution π on $\mathbb{R}^d$ such that

$$\pi(f \notin H_\rho) \leqslant 1/32 \tag{4.30}$$

and

$$\mathbb{E}_0^{|U=u} |L_\pi^{\breve{S}|U=u} - 1|^2 \leqslant \exp\left(C(\frac{mn^2\rho^4}{cd} + \frac{mn^3\rho^4}{d^2 c} Tr(\Xi_u) \right) - 1, \tag{4.31}$$

for a constant $C > 0$ that does not depend on d, n, m or c. Furthermore, in case of private coin randomness (U is degenerate), there exists a probability distribution π on $\mathbb{R}^d$ such that (4.30) is satisfied and (the sharper bound)

$$\mathbb{E}_0 |L_\pi^{\breve{S}} - 1|^2 \leqslant \exp\left(C(\frac{mn^2\rho^4}{cd} + \frac{n^4\rho^4}{d^3 c} Tr(\Xi_u)^2 \right) - 1 \tag{4.32}$$

holds for $c > 0$ small enough.

Proof. The proof is an immediate consequence of Lemma 2.8. □

4.5.3 Theorem concerning necessity of signs

The theorem below tells us that in order to attain the rate of $\frac{d}{nm}$, the statistics $S^{(j)}$ need to contain at least *some* information on the signs of $X^{(j)}$, in the sense that $\sqrt{d}/(\sqrt{mn})$ is the rate that can be attained at best when $S^{(j)}$ is measurable with respect to the absolute values of the coordinates of $X^{(j)}$. This is in particular the case for statistics based on e.g. the norm $\|X^{(j)}\|_2$ or rotation invariant statistics such as the worst-case growth rate optimal e-values (see e.g. [111]), which consequently attain the rate $\frac{\sqrt{d}}{\sqrt{mn}}$ at best and are thus suboptimal when d is small compared to m.

Theorem 4.4. *Suppose that $S^{(j)} = f_j(X^{(j)}, U)$ is such that $S^{(j)}$ is measurable with respect to $\sigma(U, (|X_1^{(j)}|, \dots, |X_d^{(j)}|))$ for $j = 1, \dots, m$. Then, for any $\alpha \in (0, 0.1]$ there exists $c > 0$ such that*

$$\sup_{f \in H_\rho} \mathbb{P}_f (T_\alpha = 0) \geqslant 3/4, \tag{4.33}$$

whenever

$$\rho^2 \leqslant c \frac{\sqrt{d}}{\sqrt{mn}}. \tag{4.34}$$

Proof. In view of Lemma 4.4 and the proof of the main theorems in 4.5.1, it suffices to bound the trace of Ξ_u in (4.32) and (4.31) in Lemma 4.5 (the first term in the exponent is controlled by (4.34)). By assumption on $S^{(j)}$, we have

$$\sigma(S^{(j)}, U, (|X_1^{(j)}|, \ldots, |X_d^{(j)}|) = \sigma(U, (|X_1^{(j)}|, \ldots, |X_d^{(j)}|)), \tag{4.35}$$

which implies that the $\mathrm{sign}(X_i^{(j)})$ is independent of $\sigma(S^{(j)}, U, (|X_1^{(j)}|, \ldots, |X_d^{(j)}|)$. Writing $X_i^{(j)} = \mathrm{sign}(X_i^{(j)})|(X_i^{(j)})|$, we obtain that

$$\begin{aligned}\mathbb{E}_0\left[X^{(j)}\middle|S^{(j)}, U = u\right] &= \left(\mathbb{E}_0\left[\mathrm{sign}(X_i^{(j)})|(X_i^{(j)})|\middle|S^{(j)}, U = u\right]\right)_{1\leqslant i\leqslant d} \\ &= \left(\mathbb{E}_0\mathrm{sign}(X_i^{(j)})\mathbb{E}_0\left[|(X_i^{(j)})|\middle|S^{(j)}, U = u\right]\right)_{1\leqslant i\leqslant d} = 0,\end{aligned}$$

where the second last inequality follows from the fact that $\mathrm{sign}(X_i^{(j)})$ is independent of the sigma algebra in (4.35) and the final equality by the symmetry of the Gaussian distribution around the mean. Following the proof of Theorem 4.1 with $\Xi_u = 0$, we obtain that the testing risk is bounded from below whenever $\rho^2 \lesssim \frac{\sqrt{d}}{\sqrt{mn}}$. □

4.5.4 Lemmas related to rate attainability

Lemma 4.6. *Let T_α correspond to a test of level α based on Edgington's method based for p-values $p^{(j)} = \chi_d^2(\|\sqrt{n}X^{(j)}\|_2^2)$ or simply the sum of $\|\sqrt{n}X^{(j)}\|_2^2$. For all $\alpha, \beta \in (0,1)$ if*

$$\rho^2 \geqslant C_{\alpha,\beta}\frac{\sqrt{d}}{\sqrt{mn}} \tag{4.36}$$

we have

$$\sup_{f\in H_\rho} \mathbb{P}_f\left(T_\alpha = 0\right) \leqslant \beta$$

for $d \geqslant C_{\alpha,\beta}m$ a large enough constant $C_{\alpha,\beta}$ depending only on $\alpha, \beta \in (0,1)$. The above result holds for Fisher's method also, under the additional assumption that $\log(m) \lesssim \sqrt{d}$.

Proof. The test in (4.7) has level α under the null hypothesis. Under the alternative hypothesis,

$$\|\sqrt{n}X^{(j)}\|_2^2 \stackrel{d}{=} n\|f\|_2^2 + 2\sqrt{n}(Z^{(j)})^\top f + \|Z^{(j)}\|_2^2,$$

where $Z^{(j)} \sim N(0, I_d)$. Rearranging, the test T_α of (4.7) can be seen to equal

$$\mathbb{1}\left\{2\frac{\sqrt{n}}{\sqrt{d}}\left(m^{-1/2}\sum_{j=1}^m Z^{(j)}\right)^\top f + \frac{1}{\sqrt{md}}\sum_{j=1}^m\left(\|Z^{(j)}\|_2^2 - d\right) \geqslant \eta_{d,m} - \frac{\sqrt{mn}}{\sqrt{d}}\|f\|_2^2\right\} \tag{4.37}$$

in distribution under $\mathbb{P}_f$, with

$$\eta_{d,m} := \frac{1}{\sqrt{dm}} \left(F^{-1}_{\chi^2_{dm}}(1-\alpha) - md \right).$$

By Lemma 4.9, $\eta_{d,m} \to \Phi^{-1}(1-\alpha)$ as both or either $d, m \to \infty$, so $\eta_{d,m}$ is bounded in d and m. Consequently, $\mathbb{P}_f(1-T_\alpha)$ equals

$$\Pr\left((1+\frac{\sqrt{n}}{\sqrt{d}}\|f\|_2) O_P(1) \leqslant \eta_{d,m} - \frac{\sqrt{mn}}{\sqrt{d}}\|f\|_2^2 \right)$$

as the left-hand side of the test in (4.37) is mean 0 and has constant variance. Since $\|f\|_2^2 \geqslant C_{\alpha,\beta}\frac{\sqrt{d}}{\sqrt{mn}}$, the latter display can be bounded from above by

$$\Pr\left((1+\frac{\sqrt{n}}{\sqrt{d}}\|f\|_2) O_P(1) \leqslant -\frac{\sqrt{mn}}{2\sqrt{d}}\|f\|_2^2 \right)$$

for a large enough $C_{\alpha,\beta}$. The latter display is smaller than β for $C_{\alpha,\beta} > 0$ large enough depending only on α and β.

For Edgington's method, one can take $p^{(j)} = 1 - F_{\chi^2_d}(\|\sqrt{n}X^{(j)}\|_2^2)$ and compute the test

$$T_\alpha := \mathbb{1}\left\{ m^{-1/2}\zeta_{\alpha,m} \sum_{j=1}^m (p^{(j)} - \frac{1}{2}) \geqslant 12^{-1/2}\Phi^{-1}(1-\alpha) \right\}, \tag{4.38}$$

where $\zeta_{\alpha,m} \to 1$ in m is such that $\mathbb{P}_0 T_\alpha = \alpha$, by Lemma 4.9.

Under the alternative, $\mathbb{E}_f p^{(j)} = \Pr(\|\sqrt{n}f + Z^{(j)}\|_2^2 \leqslant \chi^2_d)$. Therefore, by Lemma 4 in [193],

$$\mathbb{E}_f p^{(j)} \geqslant \frac{1}{2} + \frac{1}{40}\left(d^{-1/2} n \|f\|_2^2 \bigwedge \frac{1}{2} \right),$$

where we note that we can take d larger than an arbitrary constant as the rate $\sqrt{d}/(\sqrt{mn}n)$ being optimal ($\sqrt{d}/(\sqrt{mn}n) \lesssim d/(mn)$) implies $d \gtrsim m$ and for constant order m there is nothing to prove. We obtain that

$$\begin{aligned}
\mathbb{P}_f(1-T_\alpha) &= \mathbb{P}_f\left(\frac{\zeta_{m,\alpha}}{\sqrt{m}} \sum_{j=1}^m (p^{(j)} - 1/2) \leqslant 12^{-1/2}\Phi^{-1}(1-\alpha) \right) \\
&= \mathbb{P}_f\left(\frac{\zeta_{m,\alpha}}{\sqrt{m}} \sum_{j=1}^m [(p^{(j)} - \mathbb{E}_f p^{(j)}) + \mathbb{E}_f p^{(j)} - \frac{1}{2}] \leqslant 12^{-1/2}\Phi^{-1}(1-\alpha) \right) \\
&\leqslant \Pr\left(O_P(1) + \frac{\zeta_{m,\alpha}\sqrt{m}}{40}\left(d^{-1/2} n \|f\|_2^2 \bigwedge \frac{1}{2} \right) \leqslant 12^{-1/2}\Phi^{-1}(1-\alpha) \right),
\end{aligned}$$

where the $O_P(1)$ term in last equality follows from the fact that $\zeta_{m,\alpha}$ is bounded and the central limit theorem (the $p^{(j)}$'s are bounded and independent still under $\mathbb{P}_f$). If

the minimum is taken in $1/2$, the result follows for large enough m. If the minimum is taken in the first argument,

$$\frac{\zeta_{m,\alpha}\sqrt{m}}{40}\left(d^{-1/2}n\|f\|_2^2 \bigwedge \frac{1}{2}\right) \geqslant \frac{C_{\alpha,\beta}\zeta_{m,\alpha}}{40}$$

so for large enough $C_{\alpha,\beta}$, we obtain that $\mathbb{P}_f(1-T_\alpha) \leqslant \beta$.

For Fisher's method, the test of level α is given by

$$T_\alpha := \mathbb{1}\left\{\sum_{j=1}^m -2\log p^{(j)} \geqslant F^{-1}_{\chi^2_{2m}}(1-\alpha)\right\}, \tag{4.39}$$

for the p-value $p^{(j)} := 1 - F_{\chi^2_d}(\|\sqrt{n}X^{(j)}\|_2^2)$ (or equivalently Pearson's method for the p-value $F_{\chi^2_d}(\|\sqrt{n}X^{(j)}\|_2^2)$).

For the Type II error bound, assume first that $n\|f\|_2^2 \geqslant 20\sqrt{d}$. We have that $\|Z^{(j)}\|_2^2 \geqslant d - 5\sqrt{d}$ on an event of probability at least $1-e^{-5}$, via e.g. Theorem 3.1.1 in [210]. By using a union and a standard Gaussian concentration inequality, the event

$$\max_{1\leqslant j\leqslant m}\left|2\frac{\sqrt{n}}{\sqrt{d}}f^\top Z^{(j)}\right| \leqslant \frac{n}{2\sqrt{d}}\|f\|_2^2, \tag{4.40}$$

has mass at least $1 - me^{-n\|f\|_2^2/32} \geqslant 1 - me^{-\sqrt{d}/2}$. On the intersection of these two events,

$$\begin{aligned} F_{\chi^2_d}(\|\sqrt{n}f + Z^{(j)}\|_2^2) &= \Pr\left(\frac{\chi^2_d - \|Z^{(j)}\|_2^2}{\sqrt{d}} \leqslant 2\frac{\sqrt{n}}{\sqrt{d}}f^\top Z^{(j)} + \frac{n}{\sqrt{d}}\|f\|_2^2\right) \\ &\geqslant \Pr\left(\frac{\chi^2_d - d}{\sqrt{d}} \leqslant \frac{n}{2\sqrt{d}}\|f\|_2^2 - 5\right) \\ &\geqslant \Pr\left(\frac{\chi^2_d - d}{\sqrt{d}} \leqslant 5\right), \end{aligned}$$

where the right-hand side tends to $\Phi(5)$ in d by the central limit theorem. As $\Phi(5) > e^{-2}$, we obtain $-\log p^{(j)} \geqslant 2$. Since $Z^{(1)}, \ldots, Z^{(m)}$ are independent, by binomial concentration, there are at least $(3/4)m$ indexes $j = 1, \ldots, m$ such that $\|Z^{(j)}\|_2^2 \geqslant d - 5\sqrt{d}$ whilst also satisfying (4.40) with probability $1 - e^{-\tau m} - me^{-\sqrt{d}/2}$ for some constant $\tau > 0$. Using that we can without loss of generality take $m \geqslant M_{\alpha,\beta}$ for a constant $M_{\alpha,\beta} > 0$ (otherwise the separation rate is effectively the same the one for $m = 1$) and since we consider $d \gtrsim m$, we obtain that the event the joint event occurs has mass less than $1-\beta$. Furthermore, on this event, we have $1 - T_\alpha = 0$ for $M_{\alpha,\beta} > 0$ large enough, since

$$\sum_{j=1}^m -2\log p^{(j)} \geqslant 4m \cdot (3/4)$$

and by the fact that the chi-square quantile tends to $2m + C_\alpha\sqrt{2m}$ for some constant only depending on α, which is less than $4m \cdot (3/4) = 3m$ for $m \geqslant M_{\alpha,\beta}$.

Assume now that $n\|f\|_2^2 \leqslant 20\sqrt{d}$. Consider the following claim: for d large enough it holds that

$$-2\mathbb{E}_f \log\left(1 - F_{\chi^2_d}(\|\sqrt{n}f + Z^{(j)}\|_2^2)\right) \geqslant 2 - e^{-\sqrt{d}/4} + \frac{n\|f\|_2^2}{C\sqrt{d}} \tag{4.41}$$

for a fixed constant $C > 0$. If the claim holds,

$$\mathbb{P}_f(1-T_\alpha) \leqslant \mathbb{P}_f\left(\frac{\sqrt{m}n\|f\|_2^2}{C\sqrt{d}} - \sqrt{m}e^{-c\sqrt{d}} + \frac{1}{\sqrt{m}}\sum_{j=1}^{m} - 2(\log p^{(j)} - \mathbb{E}_f \log p^{(j)}) \leqslant \eta_{m,\alpha}\right)$$

with $\eta_{m,\alpha} := \frac{1}{\sqrt{2m}}(F^{-1}_{\chi^2_{2m}}(1-\alpha) - 2m)$. Since the method is rate optimal when $m \lesssim d$, the second term of the left-hand side in the above display may be assumed to be small. For the third term, note that

$$\begin{aligned}\mathbb{E}_f(-2\log p^{(j)})^2 &= 4\mathbb{E}_f \log(1 - F_{\chi^2_d}(\|\sqrt{n}f + Z^{(j)}\|_2^2)) \\ &\leqslant 4\log(1 - F_{\chi^2_d}(n\|f\|_2^2 + d)),\end{aligned}$$

where the last inequality follows from the log-concavity of $x \mapsto 1 - F_{\chi^2_d}(x)$ (see e.g. Theorem 3.4 in [98]). For $n\|f\|_2^2 \leqslant 20\sqrt{d}$, the latter quantity is uniformly bounded in n, m and d. Since the second moment bounds the variance, this implies that

$$\frac{1}{\sqrt{m}}\sum_{j=1}^{m} - 2(\log p^{(j)} - \mathbb{E}_f \log p^{(j)}) = O_P(1)$$

by the independence of $p^{(j)}$ and $p^{(k)}$ for $k \neq j$. Consequently, for some constant $\tau > 0$,

$$\mathbb{P}_f(1 - T_\alpha) \leqslant \Pr\left(\frac{\sqrt{m}n\|f\|_2^2}{C\sqrt{d}} - \sqrt{m}e^{-\tau\sqrt{d}} + O_P(1) \leqslant \eta_{m,\alpha}\right).$$

Since $\eta_{m,\alpha} \to \Phi^{-1}(1-\alpha)$ by Lemma 4.9, the fact that

$$\frac{\sqrt{m}n\|f\|_2^2}{C\sqrt{d}} \geqslant C_{\alpha,\beta}/C$$

for large enough $C_{\alpha,\beta} > 0$ depending only on α and β and the fact that $m \lesssim d$, we have that $\mathbb{P}_f(1 - T_\alpha) \leqslant \beta$.

It remains to prove the claim of (4.41). We start by writing $-2\mathbb{E}_f \log\left(p^{(j)}\right)$ as

$$-2\mathbb{E}_f \log(1 - F_{\chi^2_d}(\|Z^{(j)}\|_2^2) - 2\mathbb{E}_f \log\left(\frac{1 - F_{\chi^2_d}(\|\sqrt{n}f + Z^{(j)}\|_2^2)}{1 - F_{\chi^2_d}(\|Z^{(j)}\|_2^2)}\right).$$

The first term equals 2. Using $\log(x) \leqslant |x-1|$, the second term is bounded from below by

$$2\mathbb{E}_f \left| \frac{F_{\chi^2_d}(\|\sqrt{n}f + Z^{(j)}\|_2^2) - F_{\chi^2_d}(\|Z^{(j)}\|_2^2)}{1 - F_{\chi^2_d}(\|Z^{(j)}\|_2^2)} \right| \geqslant 2\mathbb{E}_f \left| F_{\chi^2_d}(\|\sqrt{n}f + Z^{(j)}\|_2^2) - F_{\chi^2_d}(\|Z^{(j)}\|_2^2) \right|.$$

By the same argument as used for (4.40),

$$\mathbb{E}_f \mathbb{1}_{\{f^\top Z^{(j)} < 0\}} F_{\chi^2_d}(\|\sqrt{n}f + Z^{(j)}\|_2^2) \geqslant \mathbb{E}_f F_{\chi^2_d}(\frac{1}{2}\|\sqrt{n}f\|_2^2 + \|Z^{(j)}\|_2^2) - e^{-\sqrt{d}/4},$$

which is larger than $\mathbb{E}_f F_{\chi^2_d}(\|Z^{(j)}\|_2^2)$ for all large enough d. Additionally, on the event that $f^\top Z^{(j)} \geqslant 0$, it holds that

$$F_{\chi^2_d}(\|\sqrt{n}f + Z^{(j)}\|_2^2) \geqslant \mathbb{E}_f F_{\chi^2_d}(\frac{1}{2}\|\sqrt{n}f\|_2^2 + \|Z^{(j)}\|_2^2) \geqslant \mathbb{E}_f F_{\chi^2_d}(\|Z^{(j)}\|_2^2).$$

Furthermore, we have

$$\mathbb{E}_f F_{\chi^2_d}(\frac{1}{2}\|\sqrt{n}f\|_2^2 + \|Z^{(j)}\|_2^2) - \mathbb{E}_f F_{\chi^2_d}(\|Z^{(j)}\|_2^2) = \Pr\left(0 \leqslant \frac{\chi^2_d - \tilde{\chi}^2_d}{\sqrt{d}} \leqslant \frac{n}{2\sqrt{d}}\|f\|_2^2\right),$$

where $\chi^2_d, \tilde{\chi}^2_d$ are independent chi square random variables with d degrees of freedom, which tends in d to

$$\Phi\left(\frac{n}{2\sqrt{d}}\|f\|_2^2\right) - \Phi(0) \geqslant \frac{n}{C\sqrt{d}}\|f\|_2^2,$$

where the inequality holds under the assumption $n\|f\|_2^2 \leqslant 20\sqrt{d}$ for a large enough constant $C > 0$. Putting the above lower bounds together, we obtain (4.41). □

Lemma 4.7. *Let T_α correspond to a test of level α considered in* (4.8) *or* (4.9)*. For all $\alpha, \beta \in (0,1)$ if*

$$\rho^2 \geqslant C_{\alpha,\beta}\frac{d^{3/2}}{mn} \tag{4.42}$$

we have

$$\sup_{f \in H_\rho} \mathbb{P}_f\left(T_\alpha = 0\right) \leqslant \beta$$

for a large enough constant $C_{\alpha,\beta}$ depending only on $\alpha, \beta \in (0,1)$.

Proof. The proof follows a similar line of reasoning as e.g. the proof of Lemma A.8 in [195]. Starting with (4.8), note that

$$\mathbb{P}_f(1 - T_\alpha) = \Pr\Big(\frac{1}{\sqrt{d}}\sum_{i=1}^d \Big((d^{-1/2}\sqrt{mn}f_i + Z_i)\Big)^2 \leqslant d^{-1/2}F^{-1}_{\chi^2_d}(1-\alpha)\Big)$$

for independent $Z_1, \dots, Z_d \sim N(0,1)$. The latter display equals

$$\Pr\Big(\frac{nm}{d\sqrt{d}}\|f\|_2^2 + 2\frac{\sqrt{mn}}{d}\sum_{i=1}^d f_i Z_i + \frac{1}{\sqrt{d}}\sum_{i=1}^d \big(Z_i^2 - 1\big) \leqslant d^{-1/2}(F^{-1}_{\chi^2_d}(1-\alpha) - d)\Big) =$$
$$\Pr\Big(\frac{nm}{d\sqrt{d}}\|f\|_2^2 + (1 + \sqrt{\frac{nm}{d^2}}\|f\|_2)O_P(1) \leqslant d^{-1/2}(F^{-1}_{\chi^2_d}(1-\alpha) - d)\Big) \leqslant$$
$$\Pr\Big((1 + \sqrt{\frac{nm}{d^2}\|f\|_2^2})O_P(1) \leqslant -\frac{1}{2}\frac{nm}{d\sqrt{d}}\|f\|_2^2\Big),$$

where the last inequality holds for large enough $C_{\alpha,\beta}$ since $\frac{nm}{d\sqrt{d}}\|f\|_2^2 \geqslant C_{\alpha,\beta}$ and $d^{-1/2}(F^{-1}_{\chi^2_d}(1-\alpha) - d)$ is bounded in d by Lemma 4.9. The resulting probability can be made arbitrarily small by taking large enough $C_{\alpha,\beta}$.

For a variation to Edgington's method, i.e. (4.9), similar reasoning applies. Under the null hypothesis, $\mathbb{E}_0\Phi(\sqrt{n}X^{(j)}) = 1/2$, so a conservative test (i.e. $\mathbb{P}_0 T_\alpha \leqslant \alpha$) based on Edgington's method is given by

$$T_\alpha = \mathbb{1}\left\{\left|\frac{1}{\sqrt{d}}\sum_{i=1}^d\left[\frac{d}{m}\Big(\sum_{j\in\mathcal{J}_i}(p^{(j)} - \frac{1}{2})\Big)^2 - \mathrm{Var}_0(p^{(j)})\right]\right| \geqslant c\alpha^{-1/2}\right\}$$

for a constant $c > 0$ by e.g. Chebyshev's inequality. Under the alternative hypothesis, we have $p^{(j)} = \Phi(\sqrt{n}X_i^{(j)}) = \Phi(\sqrt{n}f_i + Z_i^{(j)})$ whenever $j \in \mathcal{J}_i$. The Type II error $\mathbb{P}_f(1 - T_\alpha)$ equals

$$\mathbb{P}_f\left(\left|\frac{1}{\sqrt{d}}\sum_{i=1}^d\left[\frac{d}{m}\Big(\sum_{j\in\mathcal{J}_i}(p^{(j)} - \Phi(Z_i^{(j)}) + \Phi(Z_i^{(j)}) - \frac{1}{2})\Big)^2 - \mathrm{Var}_0(p^{(j)})\right]\right| \leqslant c\alpha^{-1/2}\right) =$$
$$\mathbb{P}_f\left(\left|\zeta + \xi + \frac{1}{\sqrt{d}}\sum_{i=1}^d\frac{d}{m}\Big(\sum_{j\in\mathcal{J}_i}(\Phi(\sqrt{n}f_i + Z_i^{(j)}) - \Phi(Z_i^{(j)}))\Big)^2\right| \leqslant c\alpha^{-1/2}\right) \tag{4.43}$$

where

$$\zeta = \frac{1}{\sqrt{d}}\sum_{i=1}^d\left[\frac{d}{m}\Big(\sum_{j\in\mathcal{J}_i}(\Phi(Z_i^{(j)}) - \frac{1}{2})\Big)^2 - \mathrm{Var}_0(p^{(j)})\right]$$

and

$$\xi = \frac{2\sqrt{d}}{m}\sum_{i=1}^d\Big(\sum_{j\in\mathcal{J}_i}(\Phi(\sqrt{n}f_i + Z_i^{(j)}) - \Phi(Z_i^{(j)}))\Big)\Big(\sum_{j\in\mathcal{J}_i}(\Phi(Z_i^{(j)}) - \frac{1}{2})\Big).$$

By independence between $Z_i^{(j)}$ and $Z_i^{(k)}$ when $j \neq k$, the random variable ζ is mean 0 under $\mathbb{E}_f$ with constant variance (i.e. not depending on d, m, n) and is thus $O_P(1)$. Similarly, ξ has constant order variance and expectation. By Jensen's inequality

$$\mathbb{E}_f(\Phi(\sqrt{n}f_i + Z_i^{(j)}) - \Phi(Z_i^{(j)}))^2 \geqslant (\Phi(2^{-1/2}\sqrt{n}f_i) - \Phi(0))^2$$

where it is used that

$$\mathbb{E}_f \Phi(\sqrt{n} f_i + Z_i^{(j)}) = \Pr\left(\sqrt{n} f_i + Z \geqslant Z'\right) = \Phi(2^{-1/2}\sqrt{n} f_i).$$

By Lemma A.11 in [195], the right-hand side of the second last display is lower bounded by $\frac{1}{12}\min\{\frac{1}{2} n f_i^2, 1\}$. By the independence of $Z_i^{(j)}$ and $Z_i^{(k)}$ when $j \neq k$, it also holds that

$$\mathbb{E}_f(\Phi(\sqrt{n} f_i + Z_i^{(j)}) - \Phi(Z_i^{(j)}))(\Phi(\sqrt{n} f_i + Z_i^{(k)}) - \Phi(Z_i^{(k)})) = (\Phi(2^{-1/2}\sqrt{n} f_i) - \Phi(0))^2.$$

Therefore,

$$\mathbb{E}_f \frac{d}{m}\Big(\sum_{j \in \mathcal{J}_i} (\Phi(\sqrt{n} f_i + Z_i^{(j)}) - \Phi(Z_i^{(j)}))\Big)^2 \geqslant \frac{m}{12d}\min\{\frac{1}{2} n f_i^2, 1\}.$$

Adding and subtracting the above expectation and noting that

$$\frac{1}{\sqrt{d}} \sum_{i=1}^{d} \frac{d}{m}\Big(\sum_{j \in \mathcal{J}_i} (\Phi(\sqrt{n} f_i + Z_i^{(j)}) - \Phi(Z_i^{(j)}))\Big)^2$$

has constant variance by the independence of $Z_i^{(j)}$ and $Z_i^{(k)}$ when $j \neq k$, we obtain that (4.43) is bounded above by

$$\mathbb{P}_f\left(O_P(1) + \frac{m}{12d\sqrt{d}} \sum_{i=1}^{d} \min\{\frac{1}{2} n f_i^2, 1\} \leqslant c\alpha^{-1/2}\right).$$

If the minimum is taken by 1 for any $i = 1, \ldots, d$, the proof is completed by noting that $m \gtrsim d^2$ by assumption whenever the rate $\frac{d\sqrt{d}}{nm}$ is the optimal rate and considering m large enough. Otherwise, the power is arbitrarily small for

$$\frac{mn}{d\sqrt{d}} \|f\|_2^2 \geqslant C_{\alpha,\beta}$$

and $C_{\alpha,\beta}$ large enough. □

Lemma 4.8. *Let T_α correspond to a test of level α considered in* (4.13)*. For all $\alpha, \beta \in (0,1)$ if*

$$\rho^2 \geqslant C_{\alpha,\beta} \frac{d}{mn} \tag{4.44}$$

we have

$$\sup_{f \in H_\rho} \mathbb{P}_f\left(T_\alpha = 0\right) \leqslant \beta$$

for a large enough constant $C_{\alpha,\beta}$ depending only on $\alpha, \beta \in (0,1)$.

Proof. The proof follows a similar line of reasoning as e.g. the proof of Lemma A.7 in [195]. For any $f \in \mathbb{R}^d$ such that $\|f\|_2 \geqslant \rho$, we have

$$U\sqrt{n}X^{(j)} \stackrel{d}{=} \sqrt{n}Uf + Z^{(j)}$$

under $\mathbb{P}_f$ by rotational invariance of the normal distribution. The probability of a Type II error of the test of level α given in (4.13) is then equal to

$$\Pr\left(\left|\sqrt{n}\sqrt{m}(Uf)_1 + Z\right| \leqslant \Phi^{-1}(1-\alpha/2)\right),$$

with $Z \sim N(0,1)$. The random variable $(Uf)_1$ is in distribution equal to $\|f\|_2 Z_1'/\|Z'\|_2$ for a d-dimensional standard Gaussian random vector Z'. For any $\beta \in (0,1)$, there exists $c' > 0$ such that $\|Z'\|_2 > c'\sqrt{d}$ occurs with probability $1-\beta/2$. Also, for $\frac{\sqrt{nm}\|f\|_2}{c'\sqrt{d}} \geqslant C_{\alpha,\beta}/c'$ large enough,

$$\Pr\left(\left|\frac{\sqrt{nm}\|f\|_2}{c'\sqrt{d}} + Z\right| \leqslant \Phi^{-1}(1-\alpha/2)\right) \leqslant \beta/2.$$

This concludes the proof of the lemma. □

The following fact is well known and included for completeness. For a random variable V, let F_V denote its CDF.

Lemma 4.9. *Let $W_1, \ldots, W_m$ be random variables and let $V_m = \sum_{j=1}^m W_j$. Suppose that*

$$m^{-1/2}\sum_{j=1}^m (W_j - \mathbb{E}W_j) \rightsquigarrow N(0,\sigma^2).$$

Then, for all $\alpha \in (0,1)$,

$$(\sigma^2 m)^{-1/2}\left(F_{V_m}^{-1}(\alpha) - \sum_{j=1}^m \mathbb{E}W_j\right) \to \Phi^{-1}(\alpha),$$

where Φ is the standard Gaussian CDF.

Proof. The quantile function

$$F_{V_m}^{-1}(\alpha) = \inf\{x \in \mathbb{R} : \Pr(V_m \leqslant x) \geqslant \alpha\}$$

satisfies $z(F_{V_m}^{-1}(\alpha) - y) = F_{z(V_m - y)}^{-1}(\alpha)$. The result now follows by e.g. Lemma 21.2 in [205]. □

4.5.5 Proof Lemma 4.1 and Lemma 4.2

Proof of Lemma 4.1. The lemma directly follows from Theorem 4.1 and Theorem 4.3 after verifying the corresponding conditions. Assumption 4.1 is satisfied if $p^{(j)}$ is generated using only local randomness, while in case of shared randomness, the same conclusion holds for Assumption 4.4. Below, we prove Assumptions 4.2 and 4.3 for the examples listed in the lemmas.

1. Fisher's method: let $S^{(j)} = -2\log p^{(j)} \sim^{H_0} \chi^2_2$ and consider the test of level α as
$$\mathbb{1}\Big\{\eta_{\alpha,m}\frac{1}{\sqrt{2m}}\sum_{j=1}^{m}(S^{(j)}-2) \geqslant \Phi^{-1}(1-\alpha)\Big\}$$
with Φ^{-1} the inverse standard normal CDF and
$$\eta_{\alpha,m} := \Phi^{-1}(1-\alpha)\left(\frac{1}{\sqrt{2m}}\Big(F^{-1}_{\chi^2_{2m}}(1-\alpha)-2m\Big)\right)^{-1}.$$
In view of the CLT, see Lemma 4.9, the sequence $\eta_{\alpha,m}$ converges to one, hence it is bounded. Furthermore, note that the corresponding combination function $C_m(s) := (\eta_{\alpha,m}/\sqrt{m})\sum_{j=1}^{m}(s_j-1)$ with $s=(s_j)\in\mathbb{R}^m$ satisfies Assumption 4.2 (e.g. with $p=q=1$). This in turn implies the moment condition for $S^{(j)}$, concluding the proof.

2. Mudholkar and George's method: The corresponding combination function $C_m(s) := |m^{-1/2}\sum_{j=1}^{m} s_j|$, by triangle inequality, satisfies Assumption 4.4. Since $S^{(j)} := -\log(p^{(j)}(1-p^{(j)}))$, the moment conditions are also satisfied.

3. Pearson's and Edgington's methods: the proofs follow the same reasoning as above with an additional application of the reverse triangle inequality in case of a two-sided test.

4. Tippett's method: when small p-values are expected under the alternative hypothesis, a test of level $\alpha\in(0,1)$ is given by
$$T_\alpha = \mathbb{1}\left\{1-\big(1-\min\{p^{(1)},\dots,p^{(m)}\}\big)^m \leqslant \alpha\right\},$$
where $1-(1-\min\{p^{(1)},\dots,p^{(m)}\})^m$ is uniformly distributed under the null (see e.g. [202]). Observe that it is equivalent to
$$\mathbb{1}\left\{-m\min\left\{-\log(1-p^{(j)})\right\} \geqslant \log(1-\alpha)\right\}.$$
For $j=1,\dots,m$, take $S^{(j)} = -\log(1-p^{(j)}) \sim^{H_0} \mathrm{Exp}(1)$. The threshold $\alpha\mapsto\log(1-\alpha)$ is strictly decreasing and the combination function $C_m(s) = -m\min s_j$ satisfies
$$|C_m(s)-C_m(s')| \leqslant m\min|s_j-s'_j| \leqslant \sum_{j=1}^{m}|s_j-s'_j|.$$

Consequently, Assumptions 4.3 and 4.2 are satisfied.

5. Generalized averages: The case where $r = -\infty$ corresponds to Tippett's method above. Similarly, $r = \infty$ corresponds to the maximum of p-values, for which the proof follows by similar steps. For $r \in [\frac{1}{m-1}, \infty)$, $a_{r,m}$ can be chosen such that the test T_α in defined in Section 4.2.1 has precise level: $\mathbb{P}_0 T_\alpha = \alpha$, see Proposition 2 and 3 in [212]. For such $a_{r,m}$, the set $\{a_{r,m} : r \in [\frac{1}{m-1}, \infty)\}$ is bounded (see Table 1 in the aforementioned paper). This test can easily be seen to be of the form (4.3) and for the generalized average, we have $(m^{-1}\sum_{j=1}^m (s_j)^r)^{1/r} = \|m^{-1/r}s\|_r$, which yields

$$m^{-1/r}\big|\|s\|_r - \|s'\|_r\big| \leqslant m^{-1/r}\|s - s'\|_r \leqslant \max_j |s_j - s'_j|,$$

so Assumption 4.2 is satisfied since $a_{r,m}$ is bounded.

□

Proof of Lemma 4.2. **Product of e-values:** The e-value test T_α for the combination function $(e_j) \mapsto \prod_{j=1}^m e_j$ can be written as

$$T_\alpha = \mathbb{1}\Big\{\sum_{j=1}^m \log E^{(j)} \geqslant \log(1/\alpha)\Big\}.$$

For $S^{(j)} := \log E^{(j)}$ and $C_m(s) = \sum_{j=1}^m s_j$ note that $E_0|\log E^{(j)}| < \infty$ and C_m satisfies (4.3). Since $\alpha \mapsto \log(1/\alpha)$ is strictly decreasing on $(0, 1)$, the assumptions of Theorems 4.1 and 4.3 are met.

Average of e-values: Since $E^{(j)}$ is nonnegative, the moment condition is satisfied. The map $(e_j) \mapsto m^{-1}\sum_{j=1}^m e_j$ satisfies (4.3), while the map $\alpha \mapsto \alpha^{-1}$ is strictly decreasing and independent of m. Hence, the conditions of Theorems 4.1 and 4.3 are satisfied. □

4.5.6 Additional simulations

Figure 4.5.6 shows the further improvement of the combined chi-square tests compared to the directional methods as d grows with respect to the number of trials, for signals that are around the detection threshold. Figure 4.5.6 shows the further worsening of performance of the combined chi-square tests compared method as m grows with respect to the dimension, for signals that are around the detection threshold. For each of these simulations, $10,000$ repetitions for every value $\alpha \in \{0.01, 0.02, \ldots, 0.99\}$ of the level of the tests are considered.

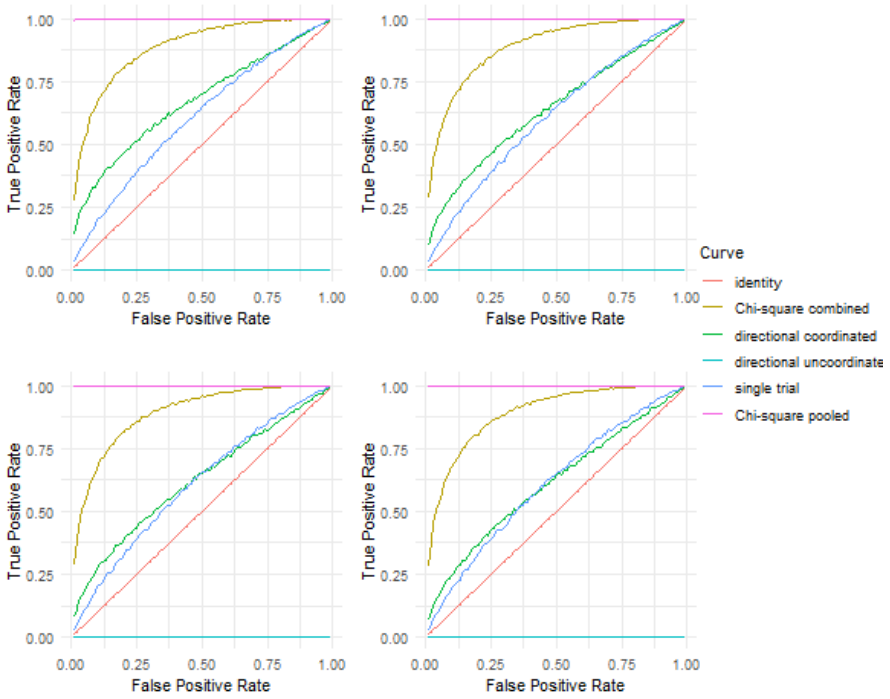


Figure 4.2: ROC curves for different values of d, whilst keeping $m = 20$, $n = 30$, $\rho^2 = 9\sqrt{d}/(16n)$. From left to right, top to bottom: $d = 30$, $d = 60$, $d = 90$, $d = 120$. The uncoordinated directional test requires $m \geqslant d$ and is therefore has TPR set to 0.

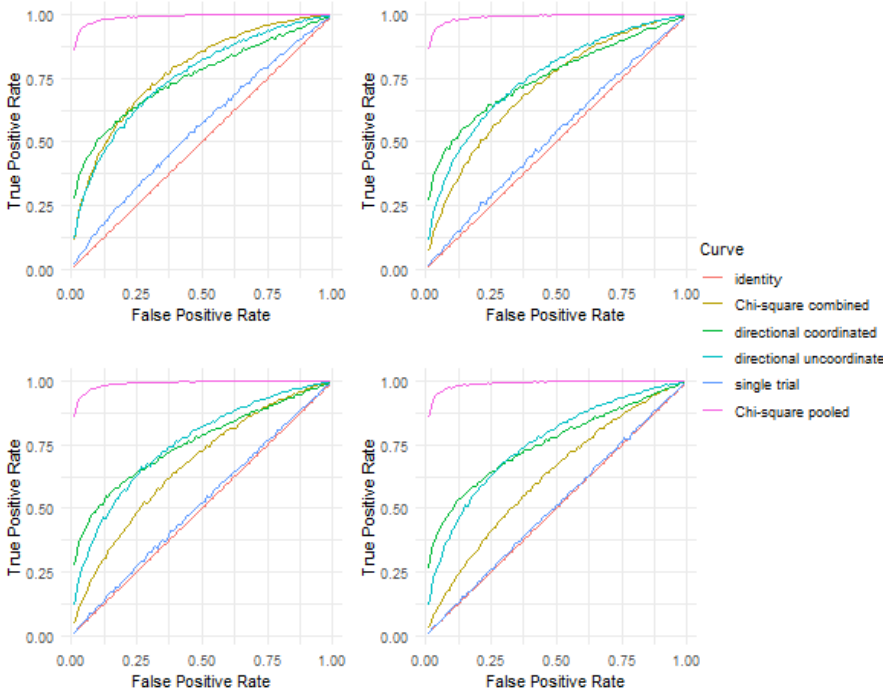


Figure 4.3: ROC curves for different values of m, whilst keeping $d = 5$, $n = 30$, $\rho^2 = 9d/(16nm)$. From left to right, top to bottom: $m = 30$, $m = 60$, $m = 100$, $m = 200$.

Chapter 5

Adaptation in nonparametric distributed testing with bandwidth and privacy constraints

> "*I wanted to see if it was really true that I had to give up all the beautiful features in order to do the applications. To answer this question, you must think about it differently. You need to find out what the constraints are imposed by the applications, whether those constraints can be added to the ones of the model, and whether the whole construction can be rebuilt.*"
> - Ingrid Daubechies

In this chapter, we shall describe the minimax rate for a nonparametric distributed problem for both bandwidth- and privacy constraints. A natural extension of the finite dimensional signal in Gaussian noise setting considered in the previous chapters is the infinite dimensional signal-in-white-noise model. The latter model serves as a benchmark model for nonparametric goodness-of-fit testing and has been extensively studied outside of the distributed setting, see [94, 121, 140, 184, 118]. In the distributed setting, the $j = 1, \dots, m$ machines each observe $i = 1, \dots, n$ i.i.d, $X_i^{(j)}$ taking values in $\mathcal{X} \subset L_2[0,1]$ and subject to the stochastic differential equation

$$dX_{t;i}^{(j)} = f(t)dt + dW_{t;i}^{(j)} \tag{5.1}$$

under P_f, with $W_i^{(1)}, \dots, W_i^{(m)}$ i.i.d. Brownian motions and $f \in L_2[0,1]$ for $i = 1, \dots, n$. Besides the difference in the local observations, the distributed setup considered for this model remains exactly the same. For notational convenience, we shall

use $N = mn$ throughout the chapter and will consider asymptotic regimes where $N \to \infty$. The results derived for the null hypothesis "$f = 0$" and alternative that "$\|f\|_2 \geqslant \rho$" in the many-normal-means model translate in the infinite dimensional model to testing $H_0 : f \equiv 0 \in L_2[0,1]$ against the alternative hypotheses that

$$f \in H_\rho^{s,R} := \{f \in \mathcal{H}^{s,R}[0,1] : \|f\|_{L_2} \geqslant \rho \text{ and } \|f\|_{\mathcal{H}^s} \leqslant R\}. \tag{5.2}$$

Here, $\mathcal{H}^{s,R} = \mathcal{H}^{s,R}([0,1])$ denotes the Sobolev ball of radius R in the space of s-smooth Sobolev functions and $\|\cdot\|_{\mathcal{H}^s}$ the Sobolev norm, see Section 5.5.3 in the chapter appendix for the definitions. We furthermore remark here that the results (minimax rates) extend to general Besov-(s,p,q) spaces (e.g. alternatives bounded in Besov norm), for any $p \in [2,\infty]$ and $q \in [1,\infty]$, up to possibly an additional logarithmic factor due to the use of a Gaussian prior in the Bayes risk.

The smoothness parameter $s > 0$ determines the difficulty of the classical (non-distributed, $m = 1$) nonparametric testing problem as considered in e.g. [123]. The asymptotic minimax rate for the non-distributed case is $\rho^2 \asymp (n)^{-\frac{2s}{2s+1/2}}$ for the s-smooth Sobolev alternative class.

This problem is closely related to "classical" nonparametric goodness-of-fit testing in the sense of [23, 182, 67, 211], in which we aim to distinguish the null hypothesis that an i.i.d. sample is generated from a cumulative distribution function $F = F_0$ versus the alternative hypothesis that $F \neq F_0$, see the Section 1.4 in the introduction of the thesis, and Section 1.4 in [123] and references therein for further discussion on this comparison.

If we consider the total variation distance between probability distributions admitting densities, which in the above example reduces to $(1/2)\|f_0 - f\|_1$ in case $F_0(t) = \int_0^t f_0(t)dt$, another motivation for the Gaussian shift experiment can be given through Le Cam theory and the equivalence of experiments, which we shall further explore in Chapter 6.

We consider the separation rate ρ in the nonparametric problem to be a sequence of positive numbers in both N, $m \equiv m_N$, $n := N/m$ and depending on the type of constraint, b or (ϵ, δ). A distributed test T in the nonparametric setting is called α-consistent for $\alpha \in (0,1)$ if $\mathcal{R}(H_\rho^{s,R}, T) \leqslant \alpha$ for all N large enough. The distributed setting for the nonparametric model remains unchanged in comparison with the finite dimensional model introduced in earlier chapters, except of course for the sample space in which the observations $X_i^{(j)}$ take values and the parameter space. These become (subsets of) $L_2[0,1]$ instead of $\mathbb{R}^d$.

The minimax rates when s is known, follow more or less straightforwardly from Chapter 2 and Chapter 3. Roughly speaking, after taking e.g. a Fourier or wavelet transform of the observations, the resulting problem in the infinite dimensional sequence space is well approximated by one in a finite dimensional subspace, by optimally truncating the sequence based on the knowledge of s. Such an approximation, combined with the results from the earlier chapter yield tight minimax lower- and upper

bounds for the minimax distributed testing problem in the nonparametric signal-in-white-noise model.

In practice, however, the true smoothness of the underlying parameter is unknown. When the smoothness s is unknown, the results and methods of the earlier chapters do not transfer as straightforwardly to the nonparametric problem. In this case, optimal methods need to be able to *adapt* to the unknown regularity of the underlying function.

In this chapter, we consider both the adaptive settings, both under bandwidth and differential privacy constraints. First, in Section 5.1, we derive tight minimax rates under bandwidth constraints when the smoothness is known. After this, in Section 5.2, we consider the setting where the smoothness is unknown, and derive minimax testing lower bounds in the adaptive setting under bandwidth constraints. In Section 5.2.1, where we exhibit adaptive testing methods attaining the lower bounds up to a $\log - \log$ factor. Then, we switch to the differential privacy setting, deriving (up to logarithmic factors) the known smoothness minimax optimal rates in Section 5.3 and exhibiting an adaptive method that attains these rates (up to logarithmic factors) whilst also adhering to the desired differential privacy constraints in Section 5.4.

5.1 Optimal nonparametric testing under bandwidth constraints with known regularity

The following theorem is a straightforward extension of Theorems 2.3 and 3.1. It describes the minimax testing rate under bandwidth constraints whenever the regularity is known. Optimal inference in this case boils down approximating the data using a sufficiently regular orthonormal basis that provides a good approximation of the (possibly) underlying signal in Sobolev space (such as a wavelet basis). Truncating the basis expansion of the data then effectively puts us in the setting of Chapters 2 and 3. A full proof of the theorem is given in Section 5.1.1.

Theorem 5.1. *Let $s, R > 0$ and let $b \equiv b_N$, $m \equiv m_N$ and $n \equiv N/m$ be sequences of natural numbers, take $\rho \equiv \rho_{n,b,m,s}$ be a sequence of positive numbers satisfying*

$$\rho^2 = \begin{cases} N^{-\frac{2s}{2s+1/2}}, & \text{if } b \geqslant N^{\frac{1}{2s+1/2}}, \\ \left(\sqrt{b}N\right)^{-\frac{2s}{2s+1}}, & \text{if } n^{\frac{1}{2s+1/2}} m^{\frac{-2s}{2s+1/2}} \leqslant b < N^{\frac{1}{2s+1/2}}, \\ (\sqrt{mn})^{-\frac{2s}{2s+1/2}}, & \text{if } b < n^{\frac{1}{2s+1/2}} m^{\frac{-2s}{2s+1/2}}. \end{cases} \tag{5.3}$$

For all $\alpha \in (0,1)$ there exist constants $C_\alpha, c_\alpha > 0$ depending only on α, s and R such that for all N large enough,

$$\inf_{T \in \mathscr{T}_{SR}^{(b)}} \mathcal{R}(H_{c_\alpha \rho}^{s,R}, T) > 1 - \alpha \quad \text{and} \quad \inf_{T \in \mathscr{T}_{SR}^{(b)}} \mathcal{R}(H_{C_\alpha \rho}^{s,R}, T) \leqslant \alpha.$$

Similarly, in the case of local randomness protocols, $\rho \equiv \rho_{n,b,m}$ satisfying

$$\rho^2 = \begin{cases} N^{-\frac{2s}{2s+1/2}} & \text{if } b \geqslant N^{\frac{1}{2s+1/2}}, \\ (bN)^{-\frac{2s}{2s+3/2}} & \text{if } n^{\frac{1}{2s+1/2}} m^{\frac{-s+1/4}{2s+1/2}} \leqslant b < N^{\frac{1}{2s+1/2}}, \\ (\sqrt{m}n)^{-\frac{2s}{2s+1/2}} & \text{if } b < n^{\frac{1}{2s+1/2}} m^{\frac{-s+1/4}{2s+1/2}}, \end{cases} \tag{5.4}$$

provides the minimax testing rate, i.e. for all $\alpha \in (0,1)$ there exist constants $C_\alpha, c_\alpha > 0$ depending only on α, s and R such that for all N large enough,

$$\inf_{T \in \mathscr{T}_{LR}^{(b)}} \mathcal{R}(H_{c_\alpha \rho}^{s,R}, T) > 1 - \alpha \quad \text{and} \quad \inf_{T \in \mathscr{T}_{LR}^{(b)}} \mathcal{R}(H_{C_\alpha \rho}^{s,R}, T) \leqslant \alpha.$$

The proof of the theorem is given in Section 5.1.1. The theorem reveals the relationship between the (local) signal-to-noise-ratio n, communication budget per machine b, the number of machines m and the smoothness of the signal s. Before providing the proof we briefly discuss the connection with distributed minimax estimation rates.

The distributed minimax estimation rates under bandwidth constraints were established in Corollary 2.2 of [187] or Theorem 3.1 in [230]. A slight reformulation of the latter yields that

$$\inf_{(\hat{f}, \mathcal{L}(Y)) \in \mathcal{E}(b)} \sup_{f \in \mathcal{H}^{s,R}} \mathbb{E}_f^Y \|\hat{f}(Y) - f\|_{L_2}^2 \asymp \begin{cases} N^{-\frac{2s}{2s+1}}, & \text{if } b \geqslant N^{\frac{1}{2s+1}}, \\ (bN)^{-\frac{2s}{2s+2}}, & \text{if } (nm^{-1-2s})^{\frac{1}{2s+1}} \leqslant b \leqslant N^{\frac{1}{2s+1}}, \\ (bm)^{-2s}, & \text{if } b \leqslant (nm^{-1-2s})^{\frac{1}{2s+1}}, \end{cases} \tag{5.5}$$

where $\mathcal{E}(b)$ is the class of all distributed estimators based on b-bit transcripts $Y = \big(Y^{(1)}, \ldots, Y^{(m)}\big)$.

A first observation is that consistent testing is possible in any regime of $b \geqslant 1$ and m, whereas this is not the case in estimation. Consider for instance the regime where m and b are fixed. In nonparametric distributed estimation, the L_2-risk does not improve once the sample size is large enough. In fact, even when allowing for asymptotics in b and m (but assuming that $(nm^{-1-2s})^{\frac{1}{2s+1}} \geqslant b$) one is better off performing the estimation locally using just one of the machines with local signal-to-noise-ratio n/m, attaining the locally optimal rate $n^{-\frac{2s}{2s+1}}$.

In the case of nonparametric testing, not only can we consistently test for any fixed m and b, the distributed testing rate is bounded from above by $(\sqrt{m}n)^{-2s/(2s+1/2)}$ (regardless of the communication budget b), which is significantly smaller (for large m) than the minimax testing rate based on the local signal-to-noise-ratio $n^{-2s/(2s+1/2)}$, which can be achieved by using only a single local machine. One possible explanation for this discrepancy is that in nonparametric estimation, the output of the inference is a high-dimensional object, which requires a large total communication budget to be reconstructed with sufficient granularity. In testing, the output of our inference is binary.

A perhaps less surprising difference is that a larger budget is needed for testing at the non-distributed minimax testing rate compared to estimation. That is, in order to obtain the non-distributed minimax rate of $\rho^2 \asymp N^{-\frac{2s}{2s+1/2}}$, the communication budget needs to satisfy $b \gtrsim N^{\frac{1}{2s+1/2}}$. On the other hand, the non-distributed minimax estimation rate $N^{-\frac{2s}{2s+1}}$ requires only $b \gtrsim N^{\frac{1}{2s+1}}$. This follows from the fact that the L_2-disk testing rate is faster than the L_2 estimation rate and hence to achieve this faster rate one has to collect information about the signal at higher frequency level as well (up to the $O(N^{\frac{1}{2s+1/2}})$ coefficients in the spectral decomposition).

Increasing m decreases the local signal-to-noise-ratio when the total number of observations $nm = N$ is kept fixed. When the total budget bm grows at a similar or faster rate than the "effective dimension" of the model, the rate that can be achieved no longer depends on m in both estimation and testing settings. In this regime, this effect is offset by the total number of bits being received by the central machine. What is different in testing problem, however, is that having access to shared randomness strictly improves the performance (until the local communication budget b reaches the effective dimension $N^{\frac{1}{2s+1/2}}$ as after that both methods reach the minimax non-distributed testing rate $N^{-\frac{2s}{2s+1/2}}$). One might wonder whether having access to shared randomness improves the rate in the estimation setting also. It turns out that this is not the case. See also Section 3.3, Theorem 3.3, which shows that under the shared randomness protocol the distributed minimax estimation rate does not improve compared to the local randomness protocol.

5.1.1 Proof of Theorem 5.1

For convenience, we consider a sufficiently smooth orthonormal wavelet basis $\{\psi_{lk} : l \in \mathbb{N}_0,\ k = 0, 1, \dots, 2^l - 1\}$ for $L_2[0,1]$, see Section 5.5.3 for a brief introduction of wavelets and collection of properties used during the proof. Nevertheless, we note, that other basis (e.g. Fourier) could be used equivalently.

For $L = L \in \mathbb{N}$, let $V_L = \{\psi_{lk} : l \leqslant L,\ k = 0, 1, \dots, 2^l - 1\}$ and define $\nu_L = \sum_{l=0}^{L} 2^l$. For $f \in L_2[0,1]$, let f^L denote the projection of f onto V_L, i.e.

$$f^L = \sum_{l=0}^{L} \sum_{k=0}^{2^l-1} \tilde{f}_{lk} \psi_{lk} \tag{5.6}$$

with $\tilde{f}_{lk} := \int f \psi_{lk}$. Let $X^{(j)} = (X_1^{(j)}, \dots, X_n^{(j)})$ and let $\overline{X^{(j)}}$ denote the average of the observed paths in machine $j = 1, \dots, m$, for which it holds that

$$d\overline{X^{(j)}}_t \overset{d}{=} f(t)dt + \frac{1}{\sqrt{n}} dW_t^{(j)},$$

where $W_\cdot^{(j)}$ is a Brownian motion. We denote the wavelet coefficients of $\overline{X^{(j)}}$ by $\tilde{X}_{lk}^{(j)} := \int_0^1 \psi_{lk} d\overline{X^{(j)}}_t$. For the coefficients at resolution level L, write

$$\tilde{X}_L^{(j)} = (\tilde{X}_{L0}^{(j)}, \dots, \tilde{X}_{L(2^L-1)}^{(j)})$$

for the $\mathbb{R}^{2^L}$ valued vector and let $\tilde{X}^{(j)}_{L':L}$ denote the concatenated coefficients from resolution level $L' < L$ up to resolution level L, i.e. $\tilde{X}^{(j)}_{L':L} = (\tilde{X}^{(j)}_{L'}, \dots, \tilde{X}^{(j)}_{L})$ taking values in $\mathbb{R}^{\nu_L - \nu_{L'-1}}$. The vector $\tilde{X}^{(j)}_{0:L} := (\tilde{X}^{(j)}_0, \tilde{X}^{(j)}_1, \dots, \tilde{X}^{(j)}_L)$ follows the dynamics

$$\tilde{X}^{(j)}_{0:L} = \tilde{f}^L + \frac{1}{\sqrt{n}} Z^{(j)}, \tag{5.7}$$

where $Z^{(j)} \sim^{i.i.d.} N(0, I_{\nu_L})$, $j = 1, \dots, m$, and $\tilde{f^L} := (\tilde{f}_{lk})_{l=0,\dots,L;\, k=0,\dots,2^l-1}$.

The existence of $C_\alpha > 0$ such that $f \in H^{s,R}_{C_\alpha \rho}$ can be detected.

In view of Theorem 3.1, there exists a constant $C'_\alpha > 0$ and a b-bit shared randomness distributed testing protocol T with transcripts generated according to $Y^{(j)} | (\tilde{X}^{(j)}_{0:L}, U) \sim K^j(\cdot | \tilde{X}^{(j)}_{0:L}, U)$ such that if $\|\tilde{f^L}\|_2^2 \geqslant (C'_\alpha)^2 \frac{\sqrt{2^L}}{mn} \left(\sqrt{\frac{2^L}{b \wedge 2^L}} \bigwedge \sqrt{m} \right)$, we have

$$\mathbb{E}_0 T + \mathbb{E}_{\tilde{f}^L}(1 - T) \leqslant \alpha.$$

Similarly, there exists a constant $C'_\alpha > 0$ and a b-bit local randomness distributed testing protocol T such that the above display holds if $\|\tilde{f^L}\|_2^2 \geqslant (C'_\alpha)^2 \frac{\sqrt{2^L}}{mn} \left(\frac{2^L}{b \wedge 2^L} \bigwedge \sqrt{m} \right)$. See Section 3.1 for the construction of such testing protocols.

Consequently, it suffices to show that for $f \in H^{s,R}_{C_\alpha \rho}$, $\|\tilde{f^L}\|_2^2$ satisfies the above lower bounds for some $L \in \mathbb{N}$ and $c > 0$. In view of $(a+b)^2/2 - b^2 \leqslant a^2$,

$$\|f^L\|_{L_2}^2 \geqslant \frac{\|f\|_{L_2}^2}{2} - \|f - f^L\|_{L_2}^2.$$

Furthermore, $f \in H^{s,R}_{C_\alpha \rho}$ implies that

$$\|f - f^L\|_{L_2}^2 = \sum_{l>L} \sum_{k=0}^{2^l-1} \tilde{f}_{lk}^2 \leqslant 2^{-2Ls} \sum_{l>L} \sum_{k=0}^{2^l-1} \tilde{f}_{lk}^2 2^{2ls} \leqslant \frac{\|f\|_{\mathcal{H}^s}^2}{2^{2Ls}} \leqslant \frac{R^2}{2^{2Ls}} \quad \text{and} \quad \|f\|_{L_2}^2 \geqslant C_\alpha^2 \rho^2.$$

Consequently, in view of Plancharel's theorem and taking $L = 1 \vee \lceil -\frac{1}{s} \log \rho \rceil$,

$$\|\tilde{f^L}\|_2^2 = \|f^L\|_{L_2}^2 \geqslant \rho^2 C_\alpha^2/2 - R^2 2^{-2Ls} \geqslant \rho^2 (C_\alpha^2/2 - R^2).$$

Consequently, there exists a b-bit shared randomness distributed testing protocol such that

$$\mathbb{E}_0 T + \mathbb{E}_f (1 - T) \leqslant \alpha$$

whenever

$$\rho^2 \gtrsim \frac{\sqrt{2^L}}{mn} \left(\sqrt{\frac{2^L}{b \wedge 2^L}} \bigwedge \sqrt{m} \right) \asymp \frac{\sqrt{1 \vee \rho^{-1/s}}}{mn} \left(\sqrt{\frac{1 \vee \rho^{-1/s}}{b \wedge (1 \vee \rho^{-1/s})}} \bigwedge \sqrt{m} \right), \tag{5.8}$$

since the constant $(\frac{C_\alpha^2}{2} - R^2)$ can be made arbitrary large by large enough choice of $C_\alpha > 0$. In the case that $b \geqslant (1 \vee \rho^{-1/s})$, the above display is satisfied whenever $\rho^{2+\frac{1}{2s}} \gtrsim (mn)^{-1}$, which provides the first case in (5.3). Similarly, if $b \leqslant \rho^{-1/s}$, the above display boils down to $\rho^{2+\frac{1}{s}} \gtrsim (\sqrt{b}mn)^{-1}$ whenever $bm \geqslant \rho^{-1/s}$, which leads to the second case in (5.3). If $bm \leqslant \rho^{-1/s}$, the inequality (5.8) reduces to $\rho^{2+\frac{1}{2s}} \gtrsim 1/\sqrt{mn}$ and consequently provides the third case in (5.3).

By similar argument as for the shared randomness protocol above, there exists a b-bit local randomness distributed testing protocol with testing risk less than α whenever

$$\rho^2 \gtrsim \frac{\sqrt{1 \vee \rho^{-1/s}}}{mn}\Big(\frac{1 \vee \rho^{-1/s}}{b \wedge (1 \vee \rho^{-1/s})} \bigwedge \sqrt{m}\Big) \tag{5.9}$$

and $C_\alpha > 0$ large enough. Then a similar computation as in the shared randomness case above leads to the three cases in (5.4).

The existence of c_α for which the risk is lower bounded. Furthermore, let $\Psi_L : \mathbb{R}^{2^L} \to L_2[0,1]$ be the measurable map defined by

$$\Psi_L \tilde{f}^L = \sum_{i=0}^{2^L-1} \tilde{f}_i \psi_{Li},$$

for $\tilde{f} = (\tilde{f}_0, \ldots, \tilde{f}_{2^L-1}) \in \mathbb{R}^{2^L}$. For any distribution π_L on $\mathbb{R}^{2^L}$, $\pi_L \circ \Psi_L^{-1}$ defines a probability measure on the Borel sigma algebra of $L_2[0,1]$. The testing risk is lower bounded as follows

$$\mathcal{R}(H^{s,R}_{c_\alpha\rho}, T) \geqslant \mathbb{P}_0(T=1) + \int \mathbb{P}_f(T=0) d\pi_L \circ \Psi^{-1}(f) - \pi_L\left(\tilde{f} \in \mathbb{R}^{2^L} : \Psi_L \tilde{f} \notin H^{s,R}_{c_\alpha\rho}\right).$$

The likelihood ratio $\frac{dP_{\tilde{f}}}{dP_0}(X^{(j)})$ with $f = \Psi_L \tilde{f}$ equals

$$\exp\left(n \int_0^1 f d\overline{X_t^{(j)}} - \frac{n}{2}\|f\|_2^2\right) = \exp\left(n(\tilde{f})^\top \tilde{X}_L^{(j)} - \frac{n}{2}\|\tilde{f}^L\|_2^2\right) =: \mathscr{L}_{\tilde{f}}(\tilde{X}_L^{(j)}),$$

where $\tilde{X}_L^{(j)} = (\int_0^1 \psi_{L0}(t) dX_t^{(j)}, \ldots, \int_0^1 \psi_{L(2^L-1)}(t) dX_t^{(j)}) \in \mathbb{R}^{2^L}$. That is, under $\mathbb{P}_0$, $\mathscr{L}_{\tilde{f}}(\tilde{X}_L^{(j)})$ is equal in distribution to the likelihood ratio

$$\frac{dN\left(\tilde{f}, \frac{1}{n} I_{2^L}\right)}{dN\left(0, \frac{1}{n} I_{2^L}\right)}.$$

This effectively puts us in the setting of Section 2.3. By Lemma 2.12, there exists a symmetric, idempotent matrix $\overline{\Gamma} \in \mathbb{R}^{2^L \times 2^L}$ such that for $\pi_L = N(0, \Gamma)$ with $\Gamma = \frac{\sqrt{c_\alpha}\rho^2}{2^L}\overline{\Gamma} \in \mathbb{R}^{2^L \times 2^L}$, it holds that

$$\mathcal{R}(H^{s,R}_{c_\alpha\rho}, T) \geqslant \alpha - \pi_L\left(\tilde{f} \in \mathbb{R}^{2^L} : \Psi_L \tilde{f} \notin H^{s,R}_{c_\alpha\rho}\right), \tag{5.10}$$

as long as ρ satisfies

$$\rho^2 \leqslant c_\alpha \frac{\sqrt{2^L}}{mn} \left(\frac{2^{L/2}}{\sqrt{b \wedge 2^L}} \bigwedge \sqrt{m} \right)$$

in the case of shared randomness protocols or

$$\rho^2 \leqslant c_\alpha \frac{\sqrt{2^L}}{mn} \left(\frac{2^{L}}{b \wedge 2^L} \bigwedge \sqrt{m} \right)$$

in the case of local randomness protocols, and $c_\alpha > 0$ small enough in both cases. Taking again $L = 2 \vee \lceil \log \rho^{-1/s} \rceil$, by similar argument as given below display (5.8) the above upper bounds for ρ^2 result in (5.3) and (5.4).

It remained to bound the prior mass term in (5.10) for $L = 2 \vee \lceil \log \rho^{-1/s} \rceil$. That is, we will show that

$$\pi_L \left(\tilde{f} \in \mathbb{R}^{2^L} : \|\Psi_L \tilde{f}\|_{L_2}^2 \geqslant c_\alpha \rho^2,\ \|\Psi_L \tilde{f}\|_{\mathcal{H}^s}^2 \leqslant R^2 \right) \geqslant 1 - \alpha/2, \tag{5.11}$$

for all n large enough. Note that for all $L \in \mathbb{N}$, $\|\Psi_L \tilde{f}\|_{\mathcal{H}^s}^2 \leqslant 2^{2Ls} \|\Psi_L \tilde{f}\|_{L_2}$. Consequently, using Plancharel's theorem, we obtain that the left-hand side of (5.11) is bounded from below by

$$\begin{aligned} \pi_L \left(\tilde{f} \in \mathbb{R}^{2^L} : c_\alpha \rho^2 \leqslant \|\tilde{f}\|_2^2 \leqslant 2^{-2Ls} R^2 \right) &\geqslant \Pr\left(c_\alpha \rho^2 \leqslant Z^\top \Gamma Z \leqslant R^2 \rho^2 \right) \\ &= \Pr\left(\sqrt{c_\alpha} 2^L \leqslant Z^\top \overline{\Gamma} Z \leqslant \frac{R^2}{\sqrt{c_\alpha}} 2^L \right), \end{aligned} \tag{5.12}$$

where Z is a 2^L-dimensional standard normal vector. Since the matrix $\overline{\Gamma}$ is symmetric, idempotent and has rank proportional to 2^L, Lemma 3.28 yields that the right-hand side of the above display is bounded from below by

$$1 - \exp\Big(- C 2^L \frac{\sqrt{c_\alpha} - 1 - 0.5 \log c_\alpha}{4} \Big) - \exp\Big(- C 2^L \frac{R^2/\sqrt{c_\alpha} - 1 - 0.5 \log\left(R^4/c_\alpha\right)}{4} \Big),$$

for a universal constant $C > 0$. The above expression can be set arbitrarily close to 1 per small enough choice of $c_\alpha > 0$, verifying the prior mass condition.

5.2 Adaptation under bandwidth constraints

In the previous section we have derived minimax lower and matching upper bounds for the nonparametric distributed testing problem in context of the Gaussian white noise model. The proposed tests, however, depend on the regularity hyperparameter s of the functional parameter of interest f. Typically, the regularity of the function is not known in practice and one has to use data driven methods to find the best testing strategies. In this section we derive distributed tests adapting to this unknown regularity. We derive both lower and upper bounds and observe surprising, additional

phase transition in the small budget regime which was not present in the non-adaptive setting.

First, we note that even in the non-distributed setting, we have to pay an additional $\log\log N$ factor as a price for adaptation (see e.g. Theorem 2.3 in [184] or Section 7 in [123]). More concretely, if $\rho_s \asymp N^{-s/(2s+1/2)}$, it holds that for any $s_{\min} < s_{\max}$,

$$\sup_{s\in[s_{\min},s_{\max}]} \mathcal{R}\left(H^{s,R}_{c_N M_{N,s}\rho_s}, T\right) \to 1,$$

for all tests T, $M_{N,s} = (\log\log N)^{\frac{s/4}{2s+1/2}}$ and any $c_N = o(1)$ whilst there exists a test T satisfying

$$\sup_{s\in[s_{\min},s_{\max}]} \mathcal{R}\left(H^{s,R}_{C M_{N,s}\rho_s}, T\right) \to 0,$$

for large enough constant $C > 0$.

The distributed testing problem is more complicated as we have to consider different regimes based on the number of transmitted bits, see Theorem 5.1. These regimes, however, depend on the unknown regularity hyperparameter and require different testing procedures to achieve consistent testing. The transcripts transmitted require a larger communication budget to attain the same performance as in Theorem 5.1. Theorem 5.2 and 5.3 below capture this increased difficulty in terms of lower- and upper bounds on the detection rate (tight up to a log-log factor). In the proof of the theorem, we derive such an adaptive distributed testing method which adapts to the smoothness. These methods are in principle based on taking a $1/\log mn$ grid of the regularity interval $[s_{\min}, s_{\max}]$, constructing optimal tests for each of the grid points and combining them using Bonferroni's correction. This results in loosing a logarithmic factor in the intermediate case as the budget has to be divided over $O(\log N)$ tests, each capturing a different possible level of smoothness.

The additional incurred cost in the distributed setting due to additional communication budget required is fundamental, as our accompanying lower bound shows. This additional difficulty translates to a $\sqrt{\log N}$ and $\log N$ factor more observations required in the intermediate budget regimes for the shared- and local randomness settings, respectively. In the small budget regime, such a loss is incurred when the local communication budget b is of smaller order than $\log N$. When $b \gtrsim \log N$ in the small budget regime, the same rate as in Theorem 5.1 can be obtained, up to the $\log\log N$ factor incurred by the Bonferroni correction.

The above described results are split over two theorems. The first, Theorem 5.2, concerns the case where $b \gtrsim \log N$. In the second, Theorem 5.3, the case where $b \lesssim \log N$ (both theorems coincide when $b \asymp \log N$). The case where $b = O(1)$ is of special interest, as $b = 1$ means each machine's local transcript forms a test itself and the global test can be seen as a "meta-analysis" on the basis of these m tests. The proofs of the upper bounds in both theorems are given in Section 5.2.1, while the proofs of the lower bound are deferred to Section 5.5.1 in the supplement.

Theorem 5.2. *Let us consider some $0 < s_{\min} < s_{\max} < \infty$, $R > 0$, let $m \equiv m_N$, $n \equiv N/m$ and $b \equiv b_N$ such that $b \gg \log N$ be sequences of natural numbers and take a sequence of positive numbers $\rho_s \equiv \rho_{n,b,m,s}$ satisfying*

$$\rho_s^2 \asymp \begin{cases} N^{-\frac{2s}{2s+1/2}}, & \text{if } b \geqslant \log(N) N^{\frac{1}{2s+1/2}}, \\ \left(\frac{\sqrt{b}N}{\sqrt{\log(N)}}\right)^{-\frac{2s}{2s+1}}, & \text{if } \log(N)\left(\frac{N^{\frac{1}{2s+1/2}}}{m^{\frac{2s+1}{2s+1/2}}} \bigvee 1\right) \leqslant b < \log(N) N^{\frac{1}{2s+1/2}}, \\ (\sqrt{mn})^{-\frac{2s}{2s+1/2}}, & \text{if } \log(N) \leqslant b < \log(N)\left(\frac{N^{\frac{1}{2s+1/2}}}{m^{\frac{2s+1}{2s+1/2}}} \bigvee 1\right), \end{cases} \tag{5.13}$$

in the shared randomness case, and

$$\rho_s^2 \asymp \begin{cases} N^{-\frac{2s}{2s+1/2}} & \text{if } b \geqslant \log(N) N^{\frac{1}{2s+1/2}}, \\ \left(\frac{bN}{\log(N)}\right)^{-\frac{2s}{2s+3/2}} & \text{if } \log(N)\left(\frac{N^{\frac{1}{2s+1/2}}}{m^{\frac{s+3/4}{2s+1/2}}} \bigvee 1\right) \leqslant b < \log(N) N^{\frac{1}{2s+1/2}}, \\ (\sqrt{mn})^{-\frac{2s}{2s+1/2}} & \text{if } \log(N) \leqslant b < \log(N)\left(\frac{N^{\frac{1}{2s+1/2}}}{m^{\frac{s+3/4}{2s+1/2}}} \bigvee 1\right), \end{cases} \tag{5.14}$$

in the case of only local randomness. Then, there exists a sequence of b-bit bandwidth constrained distributed testing procedures $T \equiv T_N$ in the respective setups such that

$$\sup_{s \in [s_{\min}, s_{\max}]} \mathcal{R}\left(H^{s,R}_{M_N \rho_s}, T\right) \to 0,$$

for any $M_N \gg \left(\log\log(N)\right)^{1/4}$. Similarly, for all distributed testing procedures in the respective setups, we have that for all $\alpha \in (0,1)$ there exists $c_\alpha > 0$ such that

$$\sup_{s \in [s_{\min}, s_{\max}]} \mathcal{R}\left(H^{s,R}_{c_\alpha \rho_s}, T\right) > \alpha.$$

The above theorem recovers (up to log-factors) the three rates corresponding to the three regimes also found in Theorem 5.1, the different regimes corresponding to different testing strategies. Since the true smoothness is unknown, these different distributed testing strategies are to be conducted simultaneously.

We note that for $m \geqslant N^{\frac{1}{2s_{\min}+1}}$ or $m \geqslant N^{\frac{1}{s_{\min}+3/4}}$ in the shared- and local randomness cases, respectively, the small budget regime no longer occurs. The reason for this is that, even though b could be relatively small, the total communication budget bm is large enough to warrant the strategy for the intermediate and high budget regimes. Furthermore, whenever $b > \log(N) N^{\frac{1}{2s+1/2}}$, the budget is large enough to recover the non-distributed regime rate.

For $b \lesssim \log N$ the separation rate is different from the non-adaptive low budget regime. Depending on the interplay between n and m either the minimax rate corresponding to the intermediate case applies or an additional poly-$(\log(mn)/b)$ factor is present

compared to the non-adaptive low budget regime, both in the local- and shared randomness settings. This results in an additional phase transition at $b = \log N$. The reason for this, is that in order to cover approximately $\log N$ different levels of smoothness using less than $\log N$ bits, each of the machines can no longer send an adequate amount of information on all the relevant smoothness levels. Instead, an optimal strategy is to divide the different machines over each of the smoothness levels, where each machines foregoes sending information regarding certain smoothness levels all together.

Theorem 5.3. *Assume the conditions of Theorem 5.2 with $b \lesssim \log N$ and assume $bm \gg \log N$. Let us consider*

$$\rho_s^2 \asymp \begin{cases} \left(\frac{\sqrt{b}N}{\sqrt{\log(N)}}\right)^{-\frac{2s}{2s+1}}, & \text{if } m \geqslant N^{\frac{1}{2s+1}}, \\ \left(\frac{\sqrt{b}\sqrt{m}n}{\log(N)}\right)^{-\frac{2s}{2s+1/2}}, & \text{if } m < N^{\frac{1}{2s+1}}, \end{cases} \tag{5.15}$$

in the shared randomness case and

$$\rho_s^2 \asymp \begin{cases} \left(\frac{bN}{\log(N)}\right)^{-\frac{2s}{2s+3/2}} & if\ m \geqslant N^{\frac{2}{2s+3/2}} \left(\frac{b}{\log(N)}\right)^{\frac{s-1/4}{2s+3/2}}, \\ \left(\frac{\sqrt{m}n\sqrt{b}}{\log(N)}\right)^{-\frac{2s}{2s+1/2}} & if\ \ m < N^{\frac{2}{2s+3/2}} \left(\frac{b}{\log(N)}\right)^{\frac{s-1/4}{2s+3/2}}, \end{cases} \tag{5.16}$$

in the local randomness case. Then, there exists a sequence of b-bit bandwidth constrained distributed testing procedures in the respective setups such that

$$\sup_{s\in[s_{\min},s_{\max}]} \mathcal{R}\left(H^{s,R}_{M_N\rho_s}, T\right) \to 0,$$

for arbitrary $M_N \gg \left(\log\log(N)\right)^{1/4}$. Similarly, for all b-bit bandwidth constrained distributed testing procedures in the respective setups, we have that for all $\alpha \in (0,1)$ there exists $c_\alpha > 0$ such that

$$\sup_{s\in[s_{\min},s_{\max}]} \mathcal{R}\left(H^{s,R}_{c_\alpha\rho_s}, T\right) > \alpha.$$

Remark 11. Both theorems together cover all cases where $mb \gg \log N$. The cases where $mb \lesssim \log N$ are excluded for technical reasons, as well as the fact that when $mb \lesssim \log N$, the optimal rate in (5.15)-(5.16) (up to at most a $\sqrt{\log\log N}$ factor) is attained by using a standard non-distributed method using just the data of one machine (see e.g. [184]). Similarly, in order to contain the level of technicality, we have foregone the $(\log\log N)^{1/4}$ additional factor in the lower bound which we esteem also to be present in the distributed setting. We refer the reader to the argument of Theorem 2.3 in [184] for how to obtain the $(\log\log N)^{1/4}$ factor in the lower bound in addition to the $\sqrt{\log N}$ and $\log N$ factors in the shared- and local randomness cases, respectively.

5.2.1 Adaptive tests attaining the bounds Theorem 5.2 and 5.3

Underlying the adaptive methods lies the wavelet transform of the observations, as introduced in Section 5.1.1. Let $\nu_L = \sum_{l=0}^{L} 2^l$ and let us introduce the notations $L_s = \lfloor s^{-1}\log(1/\rho_s)\rfloor \vee 1$, and for shorthand write $L_{\min} = L_{s_{\max}}$ and $L_{\max} = L_{s_{\min}}$ and note that $L_s \in \mathcal{C} := \{L_{\min}, \dots, L_{\max}\}$ for all $s \in [s_{\min}, s_{\max}]$. Note that $|\mathcal{C}| \leqslant \log N$.

For each regularity hyperparameter s, we distinguish low-budget ($2^{L_s} \gtrsim mb$ in the shared randomness case, and $2^{\frac{3}{2}L_s} \gtrsim mb$ in the local randomness setting) and high-budget (corresponding to $2^{L_s} \lesssim mb$ in the case of shared randomness and $2^{\frac{3}{2}L_s} \lesssim mb$ in the local randomness setting) cases. Since m and b are known for any given regularity s we know which regime it falls and is sufficient to construct that test. For notational convenience, without loss of generality, for each s we construct both the high-budget and the low-budget optimal tests using all the m machines (and do not split them between these two cases).

5.2.2 Proof of the upper bound in the low-budget regime

First we deal with the low-budget case (where the total budget is small compared to the effective dimension), which coincides in both setups. For each $L \in \mathcal{C}$ we take a subset of machines $M_L \subset \{1, \dots, m\}$ such that $|M_L| = m' := \frac{m(\log(N) \wedge b)}{\log(N)}$ and each machine appears in at most b such subsets. We note that this is possible since $m'|\mathcal{C}| \leqslant mb$. Then for each $j \in M_L$, $L \in \mathcal{C}$ we communicate

$$Y_I^{(j)}(L)|X^{(j)} \sim \mathrm{Ber}\left(\chi^2_{\nu_L}\left(\sqrt{n}\|\tilde{X}^{(j)}_{0:L}\|_2^2\right)\right) \tag{5.17}$$

and at the central machine, we can compute

$$S_I(L) = \frac{1}{\sqrt{m'}} \sum_{j \in M_L} (2Y_I^{(j)}(L) - 1).$$

Then we consider the following adaptive test based on Bonferroni's correction

$$T_I^{adapt} = \mathbb{1}\Big\{\max_{L \in \mathcal{C}} S_I(L) \geqslant 2\sqrt{\log\log N}\Big\}.$$

Since for $L \in \mathcal{C}$, it holds that $L \asymp \log(N)$, the above $\sqrt{\log\log N}$ blow up suffices to guarantee that the test has asymptotically vanishing Type I error control, i.e. $\mathbb{E}_0 T_I^{adapt} = o(1)$ by Lemma 5.1 in the Supplementary Material (as the random variables $2Y_I^{(j)}(L) - 1$ are i.i.d. Rademacher under $\mathbb{P}_0$).

For the Type II error note that

$$\mathbb{E}_f(1 - T_I^{adapt}) \leqslant \mathbb{P}_f\Big(S_I(L_s) < 2\sqrt{\log\log N}\Big)$$

and aim to apply Lemma 3.16. In view of Lemma 3.2, (with $\|f\|_2$ replaced by $\|\tilde{f}^{L_s}\|_2$ and $d = \nu_{L_s}$), noting that by triangle inequality $\|\tilde{f}^{L_s}\|_2^2 \geqslant \|f\|_2^2/2 - 2^{-2L_s s}R^2$ (see also

Section 5.1.1), we get for $\|f\|_2^2 \geqslant C_0^2\sqrt{\log\log(N)}\rho_s^2 \geqslant C_0^2\sqrt{\log\log(N)}\frac{\sqrt{2^{L_s} m\log(N)}}{n\sqrt{b\wedge\log(N)}}$, that for m large enough

$$\eta_{p,m',1} \gtrsim (m'-1)\Big(\frac{n\|\tilde{f}^{L_s}\|_2^2}{m2^{L_s/2}} \wedge \frac{1}{2}\Big)^2 \gtrsim m'\Big((\tilde{C}\frac{\log\log N}{m'}) \wedge (1/4)\Big),$$

with $\tilde{C} = C_0^2/2 - R^2$. By the assumption that $bm \gg \log(N)$, m' can be taken larger than arbitrary constant $M_0 > 0$. This means that, in view of Lemma 3.16 with $c_{\alpha,N} = 4\log\log N$ and large enough constant C_0 (depending on R), the Type II error is bounded by α.

5.2.3 Proof of the upper bound in the shared randomness, high budget regime

We use similar arguments as before, applying a Bonferroni-type of correction. First let us consider the shared randomness setting and take a one-to-one mapping ξ_L from $\{1,\dots,\nu_L\}$ to $\{(l,i) : l = 0,\dots,L,\ i = 0,1,\dots,2^l-1\}$. Let us define the test

$$(Y_{\mathrm{II}}^{(j)}(L))_i | U_L = \mathbb{1}\left\{\left(\sqrt{n}U_L\tilde{X}^{(j)}_{\xi_L(i)}\right)_i > 0\right\}, \tag{5.18}$$

where the random variable $U_L \in \mathbb{R}^{\nu_L\times\nu_L}$ is drawn from the Haar measure on the rotation group on $\mathbb{R}^{\nu_L}$. Similarly to before for each L we take a subset of machines $M_L \subseteq \{1,\dots,m\}$ such that $|M_L| = m' := \frac{m(b\wedge\log(N))}{\log(N)}$, and each machine appears at most in b such sets.

Then machine $j \in M_L$, $L \in \mathcal{C}$, transmits the bits $(Y_{\mathrm{II}}^{(j)}(L))_i$, $i = 1,\dots,b' := \frac{mb}{m'|\mathcal{C}|} \wedge \nu_L$ to the central machine, where these local test statistics are aggregated, similarly to (3.6), as

$$S_{\mathrm{II}}(L) = \frac{1}{\sqrt{b'}m'}\sum_{i=1}^{b'}\left[\left(\sum_{j\in M_L}\left[(Y_{\mathrm{II}}^{(j)}(L))_i - 1/2\right]\right)^2 - \frac{m'}{4}\right]. \tag{5.19}$$

In view of Lemma 5.1 the Type I error of the test

$$T_{\mathrm{II}}^{pub,adapt} := \mathbb{1}\Big\{\max_{L\in\mathcal{C}} S_{\mathrm{II}}(L) \geqslant 2\sqrt{\log\log N}\Big\}$$

is $o(1)$. For the Type II error note that

$$\mathbb{E}_f(1 - T_{\mathrm{II}}^{pub,adapt}) \leqslant \mathbb{E}_f\mathbb{1}\left\{S_{\mathrm{II}}(L_s) < 2\sqrt{\log\log N}\right\}.$$

By Lemma 5.2, the above display is $o(1)$ whenever $\rho^2 \gtrsim M_N\frac{2^{L_s}}{N\sqrt{\frac{b}{\log(N)}\wedge 2^{L_s}}}$, which, for the choice of $L_s = \lfloor s^{-1}\log(1/\rho_s)\rfloor \vee 1$ yields the rates of Theorem 5.2 and 5.3.

5.2.4 Proof of the upper bound in the local randomness case, high-budget regime

We proceed by adapting the test T_{III} provided in Section 3.1.3 to the nonparametric setting with unknown regularity using again a Bonferroni type correction to achieve adaptation. For simplicity, we again apply the map ξ_L introduced previously to move between the single and double index notations of the sequence model.

For all $L \in \mathcal{C}$, similarly to the previous cases we consider a collection of machines M_L with $|M_L| = m' = \frac{m(b \wedge \log(N))}{\log(N)}$ and similarly to Section 3.1.3 let us use the notation $\mathcal{I}_i(L) \subset M_L$ for the collection of machines corresponding the ith coordinate. We note that without loss of generality we can assume that $m' \geqslant M_\alpha \sqrt{\log\log N} 2^{2L_s}/(b')^2$, for some large enough constant M_α, otherwise the test T_I^{adapt} above covers the corresponding range. Then we modify the test given in (3.9) by increasing the threshold with the Bonferroni correction, i.e.

$$
\begin{aligned}
T_{\mathrm{III}}^{priv,adapt,1} &= \mathbb{1}\Big\{\max_{L\in\mathcal{C}} S^{\mathrm{III},1}(L) \geqslant 2\sqrt{\log\log N}\Big\}, \quad \text{where} \\
S^{\mathrm{III},1}(L) &= \Big|\frac{1}{|\mathcal{I}_1(L)|2^{L/2}} \sum_{i=1}^{\nu_L} \Big(\sum_{j\in\mathcal{I}_i(L)} (Y_i^{(j)} - 1/2)\Big)^2 - 2^{L/2}/4\Big|, \\
Y_i^{(j)} | \tilde{X}^{(j)}_{\xi_L(i)} &= \mathbb{1}\left\{\tilde{X}^{(j)}_{\xi_L(i)} > 0\right\}.
\end{aligned}
$$

To deal with large signal components, similarly to (3.9) (with $d = \nu_L$ and including the Bonferroni correction in the threshold), we propose the test,

$$
\begin{aligned}
T_{\mathrm{III}}^{priv,adapt,2} &= \mathbb{1}\Big\{\max_{L\in\mathcal{C}, 2\log(L)\leqslant b} S^{\mathrm{III},2}(L) \geqslant \kappa_\alpha \sqrt{\log\log N}\Big\}, \quad \text{where} \\
S^{\mathrm{III},2}(L) &= \left|\frac{1}{dm' C_{b,L}} \left(\sum_{j=1}^{m'} (Y_{count}^{(j)} - C_{b,L} 2^{L-1})\right)^2 - \frac{1}{4}\right|,
\end{aligned}
$$

with $C_{b,L} = 2^{b-L}$ and $Y_{count}^{(j)}$ given by

$$
Y_{count}^{(j)} = \sum_{l=1}^{C_{b,d}} \sum_{i=1}^{d} B_{li}^{(j)} \in \{0, 1, \ldots, C_{b,d} d\},
$$

with for $i = 1, \ldots, d$ and $j = 1, \ldots, m$, let us generate

$$
B_{li}^{(j)} \overset{i.i.d.}{\sim} \mathrm{Ber}\Big(F_{\chi_1^2}\Big(\big(\sqrt{n} X_i^{(j)}\big)^2\Big)\Big), \qquad l \in \{1, \ldots, C_{b,d} = \lfloor 2^b/(d+1)\rfloor\}.
$$

Note that $C_{b,d} \geqslant 1$ by assumption. Then machine j communicates the transcript $Y_{count}^{(j)}$ to the central machine, which can be done using $\log_2(C_{b,d} d + 1) \leqslant b$ bits in total. Finally, we aggregate these tests by taking

$$
T_{\mathrm{III}}^{priv,adapt} = T_{\mathrm{III}}^{priv,adapt,1} \vee T_{\mathrm{III}}^{priv,adapt,2}.
$$

In view of the law of Lemma 5.1 the Type I error tends to zero for both tests. Therefore, it remains to show that the Type II error is bounded by α. Similarly to the previous cases, note that

$$E_f(1-T_{\mathrm{III}}^{priv,adapt}) \leqslant \mathbb{E}_f\Big(\mathbb{1}\Big\{S^{\mathrm{III},1}(L_s) < 2\sqrt{\log\log N}\Big\} \wedge \mathbb{1}\Big\{S^{\mathrm{III},2}(L_s) < 2\sqrt{\log\log N}\Big\}\Big).$$

Following the proofs of Lemmas 3.4, 3.17 and 3.18 (with $d = \nu_{L_s}$, f taken to be the ν_{L_s} dimensional vector $\tilde{f}^{L_s}$, b replaced by b', and M_α replaced by $M_0\sqrt{\log\log n}$, for some large enough $M_0 > 0$), noting that for $C_0^2 > 4R^2$

$$\begin{aligned}\|\tilde{f}^{L_s}\|_2^2 &\geqslant \|f\|_2^2/2 - R^2 2^{-2L_s s} \gtrsim C_0\sqrt{\log\log N}\rho_s^2 \\ &= \frac{C_0 2^{3L_s/2}\sqrt{\log\log N}}{2N(\frac{b}{\log(N)} \wedge 2^{L_s})} \gtrsim \frac{C_0 2^{L_s}\sqrt{\log\log N}}{Nb'\frac{m'}{m}},\end{aligned}$$

and applying Lemmas 3.26 and 3.16 with $c_{n,\alpha} = 2\sqrt{\log\log N}$, we get that the Type II error of $T_{\mathrm{III}}^{priv,adapt}$ is bounded from above by $\alpha/2$.

Finally, we combine the above tests by taking

$$T^{priv,adapt} = T_{\mathrm{III}}^{priv,adapt} \vee T_{\mathrm{I}}^{priv,adapt} \quad \text{and} \quad T^{pub,adapt} = T_{\mathrm{II}}^{pub,adapt} \vee T_{\mathrm{I}}^{pub,adapt}.$$

Note that both of the above tests still have vanishing Type I error, while the Type II errors are bounded by the prescribed level α in view of taking the union of the above optimal tests.

5.3 Optimal nonparametric testing under differential privacy constraints

In this section, we study goodness-of-fit testing in the distributed nonparametric signal-in-white-noise model as described in the start of this chapter (i.e. in (5.1)) under differential privacy constraints, as laid out in Definition 3. The specific goodness-of-fit test we shall consider is that of testing $H_0 : f \equiv 0 \in L_2[0,1]$ against the alternative hypotheses that

$$f \in H_\rho^{s,R} := \{f \in \mathcal{H}^{s,R}[0,1] : \|f\|_{L_2} \geqslant \rho \text{ and } \|f\|_{\mathcal{H}^s} \leqslant R\}.$$

As is the case under bandwidth constraints, the nonparametric testing problem under privacy constraints closely resembles the goodness-of-fit testing problem in the many-normal-means model under privacy constraints, as studied in Chapters 2 and 3. Loosely speaking, this is a consequence of the fact that, when the model its parameter space restricted to the above Sobolev ball, it is well approximated by the many-normal-means model. That is, after applying e.g. a wavelet transform and considering the wavelet coefficients only up until a certain resolution determined by the model

characteristics, n, m, ϵ and s (sometimes referred to as its "effective dimension"), we end up in the many-normal-means model with the dimension as a function of the other model characteristics.

In this section, we shall consider the smoothness level s to be known and derive the minimax separation rate ρ for the nonparametric problem under (ϵ, δ)-differential privacy constraints. That is, for sufficiently δ, finding ρ as function of n, m, ϵ for which the minimax nonparametric testing risk

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

tends to either 0 or 1 depending on the sequence $M_N > 0$. Here, as in earlier chapters, $\mathscr{T}_{SR}^{(\epsilon,\delta)}$ consists of all distributed protocols satisfying the (ϵ, δ)-differential privacy constraint (see Definition 3). Likewise, we consider the class $\mathscr{T}_{LR}^{(\epsilon,\delta)}$ as consisting of distributed protocols using local randomness only and satisfying the same differential privacy constraint.

As is observed in Theorem 1.2, the separation rate many-normal-means model under privacy constraints is subject to many phase transitions, depending on the values of n, m, ϵ and d. These same phase transitions are observed in the nonparametric signal-in-white-noise models too, depending on n, m, ϵ and s.

In the case of shared randomness, the minimax rate in the nonparametric model is (up to logarithmic factors) given by

$$\rho^2 \asymp \begin{cases} (mn)^{-\frac{2s}{2s+1/2}} & \text{if } \epsilon \geqslant m^{\frac{1}{4s+1}} n^{\frac{1/2-2s}{4s+1}}, \\ (mn^{3/2}\epsilon)^{-\frac{2s}{2s+1}} & \text{if } m^{-\frac{2s}{4s+1}} n^{\frac{1/2-2s}{4s+1}} \leqslant \epsilon < m^{\frac{1}{4s+1}} n^{\frac{1/2-2s}{4s+1}}, \text{ and } \epsilon \geqslant n^{-1/2}, \\ (mn^2\epsilon^2)^{-\frac{2s}{2s+1}} & \text{if } m^{-\frac{1}{2}} n^{\frac{1-2s}{4s}} \leqslant \epsilon < n^{-1/2}, \\ (\sqrt{m}n)^{-\frac{2s}{2s+1/2}} & \text{if } n^{-1/2} \leqslant \epsilon < m^{-\frac{2s}{4s+1}} n^{\frac{1/2-2s}{4s+1}}, \\ (\sqrt{m}n^{3/2}\epsilon)^{-\frac{2s}{2s+1/2}} & \text{if } m^{-\frac{1}{2}} n^{-\frac{1+s}{2s+1}} \leqslant \epsilon < m^{-\frac{1}{2}} n^{\frac{1-2s}{4s}} \text{ and } \epsilon < n^{-1/2}, \\ (mn^2\epsilon^2)^{-1} & \text{if } \epsilon < m^{-\frac{1}{2}} n^{-\frac{1+s}{2s+1}}. \end{cases} \tag{5.20}$$

For different values of ϵ ranging between 0 and 1, the minimax rate changes, which we shall refer to as different regimes. We note also that, depending on the particular values of m, n and s, some of the above regimes do not occur for any value of $\epsilon \in (0, 1]$. When considering only local randomness protocols, the minimax rate (up to

logarithmic factors) for $s > 1/4$ satisfies

$$\rho^2 \asymp \begin{cases} (mn)^{-\frac{2s}{2s+1/2}} & \text{if } \epsilon \geqslant m^{\frac{1}{4s+1}} n^{\frac{1/2-2s}{4s+1}}, \\ (mn^2\epsilon^2)^{-\frac{2s}{2s+3/2}} & \text{if } m^{\frac{1/4-s}{4s+1}} n^{\frac{1/2-2s}{4s+1}} \leqslant \epsilon < m^{\frac{1}{4s+1}} n^{\frac{1/2-2s}{4s+1}} \text{ and } \epsilon \geqslant n^{-1/2}, \\ (mn^2\epsilon^2)^{-\frac{2s}{2s+3/2}} & \text{if } m^{-\frac{1}{2}} n^{\frac{5/2-2s}{4s-1}} \leqslant \epsilon < n^{-1/2}, \\ (\sqrt{m}n)^{-\frac{2s}{2s+1/2}} & \text{if } n^{-1/2} \leqslant \epsilon < m^{\frac{1/4-s}{4s+1}} n^{\frac{1/2-2s}{4s+1}}, \\ (\sqrt{m}n^{3/2}\epsilon)^{-\frac{2s}{2s+1/2}} & \text{if } m^{-\frac{1}{2}} n^{-\frac{1+s}{2s+1}} \leqslant \epsilon < m^{-\frac{1}{2}} n^{\frac{5/2-2s}{4s-1}} \text{ and } \epsilon < n^{-1/2}, \\ (mn^2\epsilon^2)^{-1} & \text{if } \epsilon < m^{-\frac{1}{2}} n^{-\frac{1+s}{2s+1}}. \end{cases} \tag{5.21}$$

We note that here, the minimax rate is subject to five different rates, where the rate $(mn^2\epsilon^2)^{-\frac{2s}{2s+3/2}}$ is split into two different cases. Even though the case where $m^{-\frac{1}{2}} n^{\frac{5/2-2s}{4s-1}} \leqslant \epsilon < n^{-1/2}$ does not change the minimax rate in (5.21), we do highlight it separately as it creates for an easier comparison with the shared randomness rate of (5.20).

Whenever $s \leqslant 1/4$, the conditions for local randomness minimax rate in (5.21) change to

$$\rho^2 \asymp \begin{cases} (\sqrt{m}n)^{-\frac{2s}{2s+1/2}} & \text{if } \epsilon \geqslant n^{-1/2}, \\ (\sqrt{m}n^{3/2}\epsilon)^{-\frac{2s}{2s+1/2}} & \text{if } m^{-\frac{1}{2}} n^{-\frac{1+s}{2s+1}} \leqslant \epsilon < n^{-1/2}, \\ (mn^2\epsilon^2)^{-1} & \text{if } \epsilon < m^{-\frac{1}{2}} n^{-\frac{1+s}{2s+1}}. \end{cases} \tag{5.22}$$

We further comment on the derived rates after the full statement on the minimax rate, which is given by the following theorem.

Theorem 5.4. *Let $s, R > 0$ be given and consider any sequences of natural numbers $m \equiv m_N$ and $n := N/m$ such that $N = mn \to \infty$, $\epsilon \equiv \epsilon_N$ in $(N^{-1}, 1]$ and $\delta \equiv \delta_N \asymp (mn)^{-p}$ for some constant $p \geqslant 2$. Let $\rho \equiv \rho_{n,m,\epsilon,\delta}$ be a sequence of positive numbers satisfying* (5.20).

Then,

$$\inf_{T \in \mathscr{T}_{SR}^{(\epsilon,\delta)}} \mathcal{R}(H^{s,R}_{M_N\rho}, T) \to \begin{cases} 0 & \text{for any } M_N \gg \log^3(N), \\ 1 & \text{for any } M_N \to 0. \end{cases}$$

Similarly, for ρ satisfying (5.21) *for $s > 1/4$ or* (5.21) *for $0 < s \leqslant 1/4$ we have that*

$$\inf_{T \in \mathscr{T}_{LR}^{(\epsilon,\delta)}} \mathcal{R}(H^{s,R}_{M_N\rho}, T) \to \begin{cases} 0 & \text{for any } M_N \gg \log^3(N), \\ 1 & \text{for any } M_N \to 0. \end{cases}$$

The theorem shows that the minimax rate under (ϵ, δ)-differentially privacy is indeed captured by (5.20) and (5.21) in the case of shared- and local randomness, respectively, up to a poly-logarithmic factor. The rate is asymptotic in the sense that the total

number of observations $N = nm$ is assumed to tend to infinity. We note here that, for the cases where $\epsilon \lesssim n^{-1/2}$, the rates are in fact attained by $(\epsilon, 0)$-differentially private protocols.

A first observation is that consistent testing against $H_\rho^{s,R}$ alternatives is possible for any value of $\epsilon \gg m^{-1/2}n^{-1}\log^6(N)$. That is, whenever $\epsilon \gg m^{-1/2}n^{-1}\log^6(N)$, the ρ tends to zero as $N \to \infty$, meaning that a signal in $\mathcal{H}^{s,R}[0,1]$ of arbitrary size can be consistently distinguished from 0 as long as the total sample size is large enough. Whenever $\epsilon \gg N^{\frac{1/2}{2s+1/2}}/\sqrt{n}$, the minimax testing rate of the unconstrained problem can be attained, up to poly-logarithmic factors. This means that, for the distributed nonparametric testing problem under privacy constraints with $\epsilon \leqslant 1$ exhibit similar performance as in the unconstrained problem, m is required to be small in comparison to n, where the smoothness plays a part in the threshold. To be precise, attaining the optimal, unconstrained rate for $\epsilon \leqslant 1$ requires $m \lesssim n^{2s-1/2}$, which in turn means that the unconstrained rate can only be attained whenever $s > 1/4$. This is true for both the shared- and local randomness distributed protocols. The drastic change in terms of achievable regimes around $s = 1/4$ in the case of local randomness protocols as prescribed by (5.21) and (5.22) is due to the "effective dimension" becoming too large (i.e. larger than $\sqrt{mn}$) for the "high-privacy-budget" regime to occur (see also Section 1.3.2).

The estimation rate in this model can be derived from Theorems 2.5 and 2.6 in Section 2.5.6. We will not provide the technical steps here in the thesis, but refer the reader to [51], where nonparametric regression is treated, which is subject to the same minimax estimation rate. This rate amounts to

$$\inf_{\hat{f}\in\mathscr{E}(\epsilon,\delta)} \sup_{f\in\mathcal{H}^{s,R}[0,1]} \mathbb{E}_f \left\| \hat{f}(Y) - f \right\|_{L_2}^2 \asymp M_N \left(N^{-\frac{2s}{2s+1}} + (mn^2\epsilon^2)^{-\frac{2s}{2s+1}} \right), \tag{5.23}$$

where $\mathscr{E}(\epsilon, \delta)$ denotes the class of all (ϵ, δ)-differentially private estimation protocols, $\log(1/\delta) \asymp \log(nm)$, $\epsilon \in ((\sqrt{mn})^{-1}, 1]$ and M_N is at most of the order $\log^2(N)$. The estimation rate reveals that consistent estimation uniformly over the Sobolev ball is possible whenever $\epsilon \gg m^{-1/2}n^{-1}$, the same threshold as for the testing problem. To attain the same rate as in the unconstrained problem (up to possibly a poly-logarithmic factor), the estimation problem requires $\epsilon \gg N^{\frac{1/2}{2s+1}}/\sqrt{n}$, which means that the unconstrained minimax rate can be attained in estimation for smaller privacy budgets than in testing. However, as can be observed in (5.20) and (5.21) the relative cost of privacy can be seen to be much higher in the estimation problem, in the sense that a change in ϵ has a larger effect on the estimation minimax rates.

How the privacy constraint affects the minimax rate can be seen to heavily depend on the regularity parameter s. In order to understand this impact, it helps to visualize the ρ-ϵ relationship as governed by (5.20) and (5.21). The relationship is depicted in Figure 5.1 below, which shows the minimax rate ρ as function of ϵ, for different smoothness levels. The slope of the curve captures the cost of increasing privacy

in terms of its effect on the minimax rate. The six regimes are summarized in the accompanying Table 5.1.

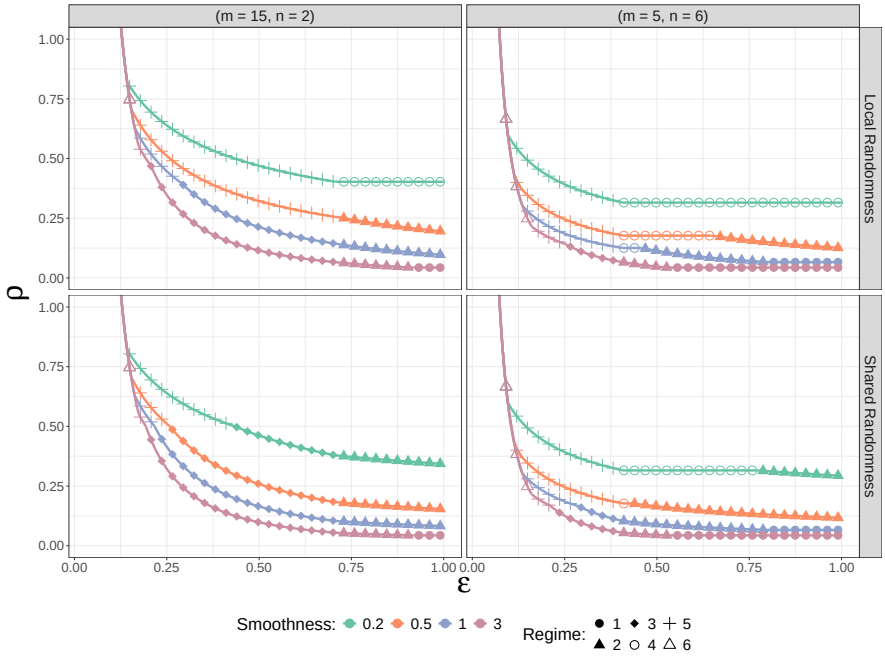


Figure 5.1: The relationship of the minimax testing rate ρ and ϵ, given by (5.20) and (5.21), for $(n, m) = (6, 5)$ in the left column and $(n, m) = (2, 15)$ in the right column, $\sigma = 1$ and smoothness levels $s = 1/5$, $s = 1/2$, $s = 1$ and $s = 3$. The panels on the first row correspond to distributed (ϵ, δ)-DP (local randomness only) protocols (i.e. (5.20)), the bottom row corresponds to distributed (ϵ, δ)-DP protocols with shared randomness (i.e. (5.21)). The regimes correspond to the six regimes (e.g. different rates) in Table 5.1.

	Regime 1	**Regime 2**	**Regime 3**	**Regime 4**	**Regime 5**	**Regime 6**
Shared U	$\left(\frac{1}{mn}\right)^{\frac{2s}{2s+1/2}}$	$\left(\frac{1}{mn^{3/2}\epsilon}\right)^{\frac{2s}{2s+1}}$	$\left(\frac{1}{mn^2\epsilon^2}\right)^{\frac{2s}{2s+1}}$	$\left(\frac{1}{\sqrt{mn}}\right)^{\frac{2s}{2s+1/2}}$	$\left(\frac{1}{\sqrt{mn^{3/2}}\epsilon}\right)^{\frac{2s}{2s+1/2}}$	$\frac{1}{mn^2\epsilon^2}$
Local only	$\left(\frac{1}{mn}\right)^{\frac{2s}{2s+1/2}}$	$\left(\frac{1}{mn^2\epsilon^2}\right)^{\frac{2s}{2s+3/2}}$	$\left(\frac{1}{mn^2\epsilon^2}\right)^{\frac{2s}{2s+3/2}}$	$\left(\frac{1}{\sqrt{mn}}\right)^{\frac{2s}{2s+1/2}}$	$\left(\frac{1}{\sqrt{mn^{3/2}}\epsilon}\right)^{\frac{2s}{2s+1/2}}$	$\frac{1}{mn^2\epsilon^2}$

Table 5.1: The minimax separation rates for the testing problem under privacy constraints, for both the local randomness and shared randomness settings. The rates are given up to logarithmic factors. The regimes are defined by the values of ϵ and the model characteristics m, n, s.

Figure 5.1 shows the minimax testing rates under privacy constraints with various

values for m and n and various levels of smoothness. In particular, the curves give insights into the cost of decreasing the privacy parameter ϵ depending on the regime. When the curve is flat, it means that ϵ can be decreased without incurring a cost in terms of having a larger detection boundary. A steep curve mean that, when ϵ decreases, it causes a large increase in the detection boundary. It can be observed that s has a "flattening" effect on the impact of ϵ on the detection boundary for "moderate" to "large" values of ϵ.

What constitute "moderate" or "large" values, depends on the size of m relative to n, as can be seen when comparing the $n = 6$ and $m = 5$ the setting with $n = 2$ and $m = 15$. It can be seen that, as the local sample n is larger compared to the number of times the total number of data points N is divided m, the cost of privacy is less. This underlines the idea that, in large samples, it is easier to retain privacy.

When ϵ becomes "very small" (smaller than a threshold depending on s, m and n), the smoothness starts to matter less and less, up to the point where the difficulty of the problem is no different for (very) different regularity levels. These scenarios correspond to settings where the privacy requirement underlying the problem is so stringent, that it effectively becomes the bottleneck of the testing problem. In such scenarios, the estimation problem locally becomes easier than the global testing problem under privacy constraints, meaning that the signal can locally be estimated at a smaller error than that the global testing problem can be solved, solely due to the presence of the differential privacy demands.

The estimation rate of (5.23) only exhibits one phase transition. This phase transition occurs between the optimal unconstrained rate and values of ϵ small enough such that the privacy constraint causes a worse rate. Comparing to the testing rates of Figures 5.1, we see that the cost of privacy for estimation is larger for "intermediate" to "small" values of ϵ, where the slope is much steeper for estimation, up until the "very small" values of ϵ, where both the testing and estimation minimax rates no longer tend to zero, which occurs for $\epsilon \lesssim 1/(\sqrt{mn})$.

Lastly, as in the case of bandwidth constraints, there is a benefit to shared randomness, strictly for moderate to large values of ϵ. [75, 44] study interactive versus non-interactive protocols and finds a difference in terms of minimax performance between the two in the local differential privacy setting. Interestingly, for $n = 1$ (which yields the local differential privacy setting), we find the similar minimax rates for nonparametric goodness-of-fit testing in the large privacy-budget regimes, for the shared randomness and local randomness protocols, as they do for interactive and non-interactive protocols, whenever ϵ is in the high-budget regime. Although they study a different model, observations from smooth densities; it is interesting to see that the same rates seem to be attainable without sequential interaction, by using shared randomness instead. We note here that, when sequential- or interactive protocols are allowed, shared randomness can be employed in particular. In real applications without interaction, one should always use shared randomness if at all possible.

5.3.1 Proof of Theorem 5.4

In a similar manenr to the proof the nonparametric testing rate under bandwidth constraints, we will make extensive use of the wavelet transform, which allows the tools of Chapter 2 and Chapter 3 to apply. We note that a wavelet basis is taken for convenience and other orthonormal bases for $L_2[0,1]$, such as Fourier- or spline bases would work as well. We separate the proof of the theorem into proving upper- and lower bounds. Before delving into the proof of the upper- and lower bounds, we introduce some notation.

We consider a smooth orthonormal wavelet basis $\{\psi_{li} : l \in \mathbb{N}_0,\, i = 0, 1, \ldots, 2^l - 1\}$. See Section 5.5.3 for a brief introduction of wavelets and collection of properties used in this proof.

For $L \in \mathbb{N}$, let $V_L = \{\psi_{li} : l \leqslant L,\ i = 0, 1, \ldots, 2^l - 1\}$ and define $\nu_L = \sum_{l=0}^{L} 2^l$. For $f \in L_2[0,1]$, let f^L denote the projection of f onto V_L, i.e.

$$f^L = \sum_{l=0}^{L} \sum_{i=0}^{2^l-1} \tilde{f}_{li} \psi_{li},$$

with $\tilde{f}_{li} := \int f \psi_{li}$. A slight difference with the approach taken in the case of bandwidth constraints, is that the wavelet transform is to be applied at the level of each observation. That is, consider

$$\tilde{X}^{(j)}_{lk;i} := \int_0^1 \psi_{lk} dX^{(j)}_{t;i}, \quad \text{for } i = 1, \ldots, n.$$

For the coefficients at resolution level L, write $\tilde{X}^{(j)}_{L;i} = (\tilde{X}^{(j)}_{L0;i}, \ldots, \tilde{X}^{(j)}_{L(2^L-1);i}) \in \mathbb{R}^{2^L}$ and let $\tilde{X}^{(j)}_{L':L;i}$ denote the concatenated coefficients from resolution level $L' < L$ up to resolution level L, i.e. $\tilde{X}^{(j)}_{L':L;i} = (\tilde{X}^{(j)}_{L';i}, \ldots, \tilde{X}^{(j)}_{L;i}) \in \mathbb{R}^{2^{L+1}-2^{L'+1}}$. The vector $\tilde{X}^{(j)}_{0:L;i} := (\tilde{X}^{(j)}_{0;i}, \tilde{X}^{(j)}_{1;i}, \ldots, \tilde{X}^{(j)}_{L;i})$ follows the dynamics

$$\tilde{X}^{(j)}_{0:L;i} = \tilde{f}^L + Z^{(j)}_i, \tag{5.24}$$

where $Z^{(j)}_i \sim^{i.i.d.} N(0, I_{\nu_L})$, $j = 1, \ldots, m$, and $\tilde{f^L} := (\tilde{f}_{li})_{l=0,\ldots,L;\, i=0,\ldots,2^l-1}$. Furthermore, let $\tilde{X}^{(j)}_{lk} = (\tilde{X}^{(j)}_{lk;i})_{i=1,\ldots,n}$ and $\tilde{X}^{(j)}_{L':L} = (\tilde{X}^{(j)}_{L':L;i})_{i=1,\ldots,n}$.

The existence of a sequence of (ϵ, δ)-DP consistent tests: The wavelet coefficients $\tilde{X}^{(j)}_{0:L;i}$ corresponding to the observation $X^{(j)}_i$, effectively place us in the many-normal-means setting of Chapters 2 and 3, with $d = 2^L$. For transcripts $Y^{(j)}_L$ generated according to

$$Y^{(j)}_L | (X^{(j)}, U) \sim K^j \left(\cdot | \tilde{X}^{(j)}_{0:L}, U \right),$$

a change in one datum $X^{(j)}_i$ translates to a change in $\tilde{X}^{(j)}_{0:L;i}$ only, which means that the privacy preserving mechanisms of Chapter 3 apply after the wavelet transform

and truncation up to resolution level L. Theorem 3.2 yields that, for $L \in \mathbb{N}$ and $\alpha \in (0,1)$, there exists a distributed (ϵ,δ)-differentially private, shared randomness testing protocol

$$T_{\alpha;L} \equiv \{T_{\alpha;L}, \{K^j\}_{j=1}^m, (\mathcal{U}, \mathscr{U}, \mathbb{P}^U)\} \tag{5.25}$$

such that $\mathbb{P}_0 T_{\alpha;L} \leqslant \alpha$ and furthermore the condition

$$\left\|\tilde{f}^L\right\|_2^2 \geqslant C_\alpha \log^6(2^L N)\left(\frac{2^L}{mn\sqrt{n\epsilon^2 \wedge 1}\sqrt{n\epsilon^2 \wedge 2^L}} \bigwedge \left(\frac{\sqrt{2^L}}{\sqrt{m}n\sqrt{n\epsilon^2 \wedge 1}} \bigvee \frac{1}{mn^2\epsilon^2}\right)\right),$$

implies that $\mathbb{P}_f(1 - T_{\alpha;L}) \leqslant \alpha$. For local randomness protocols, the same is true whenever

$$\left\|\tilde{f}^L\right\|_2^2 \geqslant C_\alpha \log^6(2^L N)\left(\frac{2^{(3/2)L}}{mn(n\epsilon^2 \wedge 2^L)} \bigwedge \left(\frac{\sqrt{2^L}}{\sqrt{m}n\sqrt{n\epsilon^2 \wedge 1}} \bigvee \frac{1}{mn^2\epsilon^2}\right)\right).$$

Next, we show that for $f \in H^{s,R}_{C_\alpha \rho}$, $\|\tilde{f^L}\|_2^2$ satisfies the above lower bounds for some $L \in \mathbb{N}$.

In view of $(a+b)^2/2 - b^2 \leqslant a^2$,

$$\|f^L\|_{L_2}^2 \geqslant \frac{\|f\|_{L_2}^2}{2} - \|f - f^L\|_{L_2}^2.$$

Furthermore, $f \in H^{s,R}_{C_\alpha \rho}$ implies that

$$\|f - f^L\|_{L_2}^2 = \sum_{l>L}\sum_{k=0}^{2^l-1} \tilde{f}_{lk}^2 \leqslant 2^{-2Ls}\sum_{l>L}\sum_{k=0}^{2^l-1} \tilde{f}_{lk}^2 2^{2ls} \leqslant \frac{\|f\|_{\mathcal{H}^s}^2}{2^{2Ls}} \leqslant \frac{R^2}{2^{2Ls}} \quad \text{and} \quad \|f\|_{L_2}^2 \geqslant C_\alpha^2 \rho^2.$$

Consequently, in view of Plancharel's theorem and taking $L = 1 \vee \lceil -\frac{1}{s}\log_2 \rho \rceil$,

$$\|\tilde{f^L}\|_2^2 = \|f^L\|_{L_2}^2 \geqslant \rho^2 C_\alpha^2/2 - R^2 2^{-2Ls} \geqslant \rho^2(C_\alpha^2/2 - R^2).$$

Consequently, whenever ρ satisfies either

$$\rho^2 \gtrsim \left(\frac{1 \vee \rho^{-1/s}}{mn\sqrt{n\epsilon^2 \wedge 1}\sqrt{n\epsilon^2 \wedge (1 \vee \rho^{-1/s})}} \bigwedge \left(\frac{\sqrt{1 \vee \rho^{-1/s}}}{\sqrt{m}n\sqrt{n\epsilon^2 \wedge 1}} \bigvee \frac{1}{mn^2\epsilon^2}\right)\right), \tag{5.26}$$

in the case of shared randomness, or

$$\rho^2 \gtrsim \left(\frac{(1 \vee \rho^{-1/s})^{3/2}}{mn(n\epsilon^2 \wedge (1 \vee \rho^{-1/s}))} \bigwedge \left(\frac{\sqrt{1 \vee \rho^{-1/s}}}{\sqrt{m}n\sqrt{n\epsilon^2 \wedge 1}} \bigvee \frac{1}{mn^2\epsilon^2}\right)\right), \tag{5.27}$$

in the case of local randomness, we have that for all $M_N \gg \log^6(2^L N)$,

$$\sup_{f \in H^{s,R}_{C_\alpha \rho}} (\mathbb{E}_0 T_{\alpha;L} + \mathbb{E}_f(1 - T_{\alpha;L})) \leqslant 2\alpha$$

for N large enough since $(\frac{C_\alpha^2}{2} - R^2)$ tends to zero as $C_\alpha \to \infty$. Since α is arbitrary, it follows that

$$\inf_{T \in \mathscr{T}_\cdot(\epsilon,\delta)} \mathcal{R}(H^{s,R}_{M_N\rho}, T) \to 0$$

for both classes of shared randomness protocols and local randomness protocols when $M_N \gg \log^6(2^L N)$, whenever ρ satisfies (5.26) or (5.27), respectively.

The testing risk lower bound for (ϵ, δ)-DP tests:

Consider for $L \in \mathbb{N}$ the linear operator $\Psi_L : \mathbb{R}^{2^L} \to L_2[0,1]$ defined by

$$\Psi_L \tilde{f}^L = \sum_{i=0}^{2^L-1} \tilde{f}_i \psi_{Li}, \tag{5.28}$$

for $\tilde{f}^L = (\tilde{f}_0, \dots, \tilde{f}_{2^L-1}) \in \mathbb{R}^{2^L}$. Since Ψ_L is measurable, any probability distribution π_L on $\mathbb{R}^{2^L}$, $\pi_L \circ \Psi_L^{-1}$ defines a probability measure on the Borel sigma algebra of $L_2[0,1]$. This means that the testing risk is lower bounded as follows

$$\mathcal{R}(H^{s,R}_{c_\alpha\rho}, T) \geqslant \mathbb{P}_0(T=1) + \int \mathbb{P}_f(T=0) d\pi_L \circ \Psi^{-1}(f) - \pi_L\left(\tilde{f} \in \mathbb{R}^{2^L} : \Psi_L \tilde{f} \notin H^{s,R}_{c_\alpha\rho}\right).$$

The likelihood ratio $\frac{dP_{\tilde{f}}}{dP_0}(X_i^{(j)})$ with $f = \Psi_L \tilde{f}$ equals

$$\exp\left(\int_0^1 f dX^{(j)}_{t;i} - \frac{1}{2}\|f\|_2^2\right) = \exp\left((\tilde{f})^\top \tilde{X}^{(j)}_{L;i} - \frac{1}{2}\|\tilde{f}^L\|_2^2\right) =: \mathscr{L}_{\tilde{f}}(\tilde{X}^{(j)}_{L;i}),$$

where $\tilde{X}^{(j)}_{L;i} = (\int_0^1 \psi_{L0}(t) dX_t^{(j)}, \dots, \int_0^1 \psi_{L(2^L-1)}(t) dX_t^{(j)}) \in \mathbb{R}^{2^L}$. That is, under $\mathbb{P}_0$, $\mathscr{L}_{\tilde{f}}(\tilde{X}^{(j)}_{L;i})$ is equal in distribution to the likelihood ratio

$$\frac{dN\left(\tilde{f}^L, I_{2^L}\right)}{dN\left(0, I_{2^L}\right)}.$$

Since the observations given $\tilde{f}^L$ are i.i.d., restricting to the above Bayes risk effectively puts us in the setting of Section 2.3 with $d = 2^L$. By Lemma 2.17, if $L \lesssim \log(N)$, there exists a symmetric, idempotent matrix $\overline{\Gamma} \in \mathbb{R}^{2^L \times 2^L}$ such that for $\pi_L = N(0, \Gamma)$ with $\Gamma = \frac{\sqrt{c_\alpha}\rho^2}{2^L}\overline{\Gamma} \in \mathbb{R}^{2^L \times 2^L}$, it holds that

$$\mathcal{R}(H^{s,R}_{c_\alpha\rho}, T) \geqslant \alpha - \pi_L\left(\tilde{f} \in \mathbb{R}^{2^L} : \Psi_L \tilde{f} \notin H^{s,R}_{c_\alpha\rho}\right), \tag{5.29}$$

as long as ρ satisfies

$$\rho^2 \leqslant c_\alpha \left(\frac{2^L}{mn\sqrt{n\epsilon^2 \wedge 1}\sqrt{n\epsilon^2 \wedge 2^L}} \bigwedge \left(\frac{\sqrt{2^L}}{\sqrt{m}n\sqrt{n\epsilon^2 \wedge 1}} \bigvee \frac{1}{mn^2\epsilon^2} \right) \right)$$

in the case of shared randomness protocols or

$$\rho^2 \leqslant c_\alpha \left(\frac{2^{(3/2)L}}{mn(n\epsilon^2 \wedge 2^L)} \bigwedge \left(\frac{\sqrt{2^L}}{\sqrt{m}n\sqrt{n\epsilon^2 \wedge 1}} \bigvee \frac{1}{mn^2\epsilon^2} \right) \right)$$

in the case of local randomness protocols, and $c_\alpha > 0$ small enough in both cases. The choice minimizing the left-hand side whilst also satisfying the prior mass requirement and $L \lesssim \log(N)$, is shown below to be $L = 2 \vee \lceil \log_2 \rho^{-1/s} \rceil$. The condition $L \lesssim \log(N)$ follows if ρ satisfies (5.20) or (5.21). We verify the prior mass requirement further down below. The choice of $L = 2 \vee \lceil \log_2 \rho^{-1/s} \rceil$ yields that if

$$\rho \lesssim \left(\frac{1 \vee \rho^{-1/s}}{mn\sqrt{n\epsilon^2 \wedge 1}\sqrt{n\epsilon^2 \wedge (1 \vee \rho^{-1/s})}} \bigwedge \left(\frac{\sqrt{1 \vee \rho^{-1/s}}}{\sqrt{m}n\sqrt{n\epsilon^2 \wedge 1}} \bigvee \frac{1}{mn^2\epsilon^2} \right) \right), \quad (5.30)$$

in the case of shared randomness, or

$$\rho \lesssim \left(\frac{(1 \vee \rho^{-1/s})^{3/2}}{mn(n\epsilon^2 \wedge (1 \vee \rho^{-1/s}))} \bigwedge \left(\frac{\sqrt{1 \vee \rho^{-1/s}}}{\sqrt{m}n\sqrt{n\epsilon^2 \wedge 1}} \bigvee \frac{1}{mn^2\epsilon^2} \right) \right), \quad (5.31)$$

in case of local randomness, the corresponding class $\mathscr{T}.(\epsilon, \delta)$ is such that

$$\inf_{T \in \mathscr{T}.(\epsilon,\delta)} \mathcal{R}(H^{s,R}_{M_N \rho}, T) \to 0$$

for any sequence $M_N \to 0$.

It remained to bound the prior mass term in (5.29) for $L = 2 \vee \lceil \log_2 \rho^{-1/s} \rceil$. That is, we will show that

$$\pi_L \left(\tilde{f} \in \mathbb{R}^{2^L} : \|\Psi_L \tilde{f}\|^2_{L_2} \geqslant c_\alpha \rho^2,\ \|\Psi_L \tilde{f}\|^2_{\mathcal{H}^s} \leqslant R^2 \right) \geqslant 1 - \alpha/2, \quad (5.32)$$

for all n large enough. Note that for all $L \in \mathbb{N}$, $\|\Psi_L \tilde{f}\|^2_{\mathcal{H}^s} \leqslant 2^{2Ls} \|\Psi_L \tilde{f}\|_{L_2}$. Consequently, using Plancharel's theorem, we obtain that the left-hand side of (5.32) is bounded from below by

$$\begin{aligned} \pi_L \left(\tilde{f} \in \mathbb{R}^{2^L} : c_\alpha \rho^2 \leqslant \|\tilde{f}\|^2_2 \leqslant 2^{-2Ls} R^2 \right) &\geqslant \Pr\left(c_\alpha \rho^2 \leqslant Z^\top \Gamma Z \leqslant R^2 \rho^2 \right) \\ &= \Pr\left(\sqrt{c_\alpha} 2^L \leqslant Z^\top \overline{\Gamma} Z \leqslant \frac{R^2}{\sqrt{c_\alpha}} 2^L \right), \quad (5.33) \end{aligned}$$

where Z is a 2^L-dimensional standard normal vector. Since the matrix $\overline{\Gamma}$ is symmetric, idempotent and has rank proportional to 2^L, Lemma 3.28 yields that the right-hand side of the above display is bounded from below by

$$1 - \exp\Big(- C 2^L \frac{\sqrt{c_\alpha} - 1 - 0.5 \log c_\alpha}{4} \Big) - \exp\Big(- C 2^L \frac{R^2/\sqrt{c_\alpha} - 1 - 0.5 \log\left(R^4/c_\alpha\right)}{4} \Big),$$

for a universal constant $C > 0$. The above expression can be set arbitrarily close to 1 per small enough choice of $c_\alpha > 0$, verifying the prior mass condition.

In summary:

Putting everything together, we have obtained the result of the theorem whenever (5.26) and (5.30) hold in the case of shared randomness protocols, or (5.27) and (5.31) in case of local randomness protocols. That is, when

$$\rho \asymp \left(\frac{1 \vee \rho^{-1/s}}{mn\sqrt{n\epsilon^2 \wedge 1}\sqrt{n\epsilon^2 \wedge (1 \vee \rho^{-1/s})}} \bigwedge \left(\frac{\sqrt{1 \vee \rho^{-1/s}}}{\sqrt{m}n\sqrt{n\epsilon^2 \wedge 1}} \bigvee \frac{1}{mn^2\epsilon^2} \right) \right),$$

in the case of shared randomness, or

$$\rho \asymp \left(\frac{(1 \vee \rho^{-1/s})^{3/2}}{mn(n\epsilon^2 \wedge (1 \vee \rho^{-1/s}))} \bigwedge \left(\frac{\sqrt{1 \vee \rho^{-1/s}}}{\sqrt{m}n\sqrt{n\epsilon^2 \wedge 1}} \bigvee \frac{1}{mn^2\epsilon^2} \right) \right),$$

in the case of local randomness protocols, we obtain the statement of the theorem. A straightforward calculation yields the corresponding expressions of (5.20), (5.21) and (5.22) for ρ^2.

5.4 Adaptive nonparametric methods under privacy constraints

In the previous section, minimax (up to a poly-log factor) optimal (ϵ, δ)-DP distributed testing protocols $T_{\alpha;L}$ were derived, where the choice of L yielded the optimal performance. This optimal choice of L was contingent on the hyperparamter $s > 0$, the true smoothness of the signals in the alternative class.

In many settings, the true underlying smoothness of a signals is not known in advance and it is desirable in these cases to consider testing risk for various levels of smoothness. In such cases, it makes sense to consider the minimax testing risk

$$\sup_{s \in [s_{\min}, s_{\max}]} \mathcal{R}\left(H^{s,R}_{M_{N,s}\rho_s}, T \right),$$

for certain predetermined values $0 < s_{\min} < s_{\max} < \infty$. Here, we consider separation rates ρ_s depending on the underlying smoothness. In the case that $s = s_{\min}$, which results in a relatively larger separation rate than when (for example) $s = s_{\max}$. The results of the previous section indicate that the rate $\rho_{s_{\min}}$ can be attained (up to a poly-logarithmic factor) by a (ϵ, δ)-DP distributed protocol when $\rho_{s_{\min}}$ satisfies (5.20), (5.21) and (5.22) with $s = s_{\min}$. However, in the case that the true smoothness s is larger than $s_{\min}$ we would like to attain the smaller of the two rates ρ_s.

In this section, it is shown that adaptation under (ϵ, δ)-differential privacy constraints, *adaptive* testing is possible at the cost of at most a logarithmic factor in N. Specifically, by deriving tests that adapt to the optimal rate (up to a poly-logarithmic factor) established in the previous section, we prove the following theorem.

Theorem 5.5. *Let $0 < s_{\min} < s_{\max} < \infty$, $R > 0$ be given and consider sequences of natural numbers $m \equiv m_N$, $n = N/m$, positive numbers $\epsilon \equiv \epsilon_N$ in $((mn)^{-1}, 1]$ and $\delta \equiv \delta_N \asymp (mn)^{-p}$ for some constant $p \geqslant 2$. Consider for every $s \in [s_{\min}, s_{\max}]$ a sequence of nonnegative ρ_s satisfying (5.20). Then, there exists a shared randomness distributed (ϵ, δ)-differentially private protocol T_{SR} such that*

$$\sup_{s \in [s_{\min}, s_{\max}]} \mathcal{R}\left(H^{s,R}_{CM_{N,s}\rho_s}, T_{SR}\right) \to 0$$

whenever $M_{N,s} \gg \log^{p_s}(N)$, where $p_s \leqslant 11/4$ is a constant depending only on $s > 0$. Furthermore, whenever each ρ_s satisfies (5.21) when $s > 1/4$ or (5.22) when $s \leqslant 1/4$, there exists a local randomness distributed (ϵ, δ)-differentially private protocol T_{LR}

$$\sup_{s \in [s_{\min}, s_{\max}]} \mathcal{R}\left(H^{s,R}_{CM_{N,s}\rho_s}, T_{LR}\right) \to 0,$$

whenever $M_{N,s} \gg \log^{p_s}(N)$ where p_s is a constant depending only on $s > 0$.

Proof. We start by introducing the notation $L_s = \lfloor s^{-1} \log_2(1/\rho_s) \rfloor \vee 1$. Define furthermore $\mathcal{C} := \{L_{s_{\min}}, \ldots, L_{s_{\max}}\}$ and note that $L_s \in \mathcal{C}$ for all $s \in [s_{\min}, s_{\max}]$ and $|\mathcal{C}| \leqslant C_{s_{\max}} \log N$ for some constant $C_{s_{\max}} > 0$ depending only on $s_{\max}$, whenever ρ_s satisfies the conditions of the theorem.

The adaptive test we construct can be seen as the maximum of the tests $T_{\alpha;L_s}$ considered in (5.25) in addition to a Bonferroni correction to compensate for the increasing Type I error resulting from taking the maximum of tests. To be precise, let $\epsilon' = \epsilon/|\mathcal{C}|$ and $\delta' = \delta/|\mathcal{C}|$. For every $s \in [s_{\min}, s_{\max}]$ and $L_s \in \mathcal{C}$, we release the (ϵ', δ')-DP transcripts $(Y^{(j)}_{L_s})_{j \in [m]}$ corresponding to the rate-optimal test T_{L_s} of the previous section (i.e. as in (5.25)). The full collection of transcripts received is

$$\left\{ (Y^{(j)}_{L_s})_{j \in [m]} \,:\, L_s \in \mathcal{C} \right\},$$

which can be generated through independent noise mechanisms, is (ϵ, δ)-DP (see e.g. Theorem 3.16 in [82]).

Next, we discuss the construction of the test T_{L_s} for each $L_s \in \mathcal{C}$. What is the optimal test depends on whether there is access to shared randomness or not, so we consider these cases separately. We recall the notation $\nu_L := \sum_{l=0}^{L} 2^l$.

An adaptive shared randomness protocol: For $\nu_{L_s}/\sqrt{mn} \leqslant \epsilon'$, let $(Y^{(j)}_{L_s})_{j \in [m]}$ be generated by (3.21) with as the underlying observations the wavelet coefficients

$\tilde{X}^{(j)}_{0:L_s}$, i.e. setting $d = \nu_{L_s}$ and compute the test $\varphi^{\epsilon'}_\tau$ defined in (3.22) for all $\tau \in \mathrm{T}$, where the latter collection is given (3.23) with $M = R$. For the critical value of $\varphi^{\epsilon'}_\tau$, setting $J = 1/(|\mathcal{C}||\mathrm{T}|)$ results in the test

$$T_{L_s} := \max_{\tau \in \mathrm{T}} \varphi^{\epsilon'}_\tau \tag{5.34}$$

having Type I error less than $\alpha/|\mathcal{C}|$ by Lemma 3.8 for arbitrary $\alpha \in (0,1)$. By the proof of the same Lemma, the above test has Type II error of the order α whenever

$$\left\| \tilde{f}^{L_s} \right\|_2^2 \geqslant C_\alpha \log^6(1+mn) \left(\frac{\sqrt{2^{L_s}}}{\sqrt{m}n(\sqrt{n}\epsilon' \wedge 1)} \right) \bigvee \left(\frac{1}{mn^2(\epsilon')^2} \right),$$

with $C_\alpha \geqslant C'_\alpha |\mathcal{C}|$ with $C'_\alpha > 0$ large enough.

Whenever $\nu_{L_s}/\sqrt{mn} > \epsilon'$, let $(Y^{(j)}_{L_s})_{j\in[m]}$ be generated by (3.39) ($d = \nu_{L_s}$) with as the underlying observations the wavelet coefficients $\tilde{X}^{(j)}_{0:L_s}$. Using these transcripts, the test $T^{\epsilon',\delta'}_{II}$ as defined in (3.40) with $\kappa_\alpha = \kappa'_\alpha \sqrt{|\mathcal{C}|}$, $\kappa'_\alpha > 0$ has Type I error less than $\alpha/|\mathcal{C}|$ by Lemma 3.23. By Lemma 3.14 combined with Lemma 3.23, this test has Type II error of the order $\kappa^2_\alpha / C^2_\alpha \leqslant \alpha$ whenever

$$\left\| \tilde{f}^{L_s} \right\|_2^2 \geqslant C_\alpha \frac{2^{L_s} \log(1+2^{L_s}nm) \log(1+nm)}{mn\sqrt{n(\epsilon')^2 \wedge 2^{L_s}} \sqrt{n(\epsilon')^2 \wedge 1}}$$

and $C_\alpha \geqslant C'_\alpha \kappa_\alpha \gtrsim C'_\alpha |\mathcal{C}|$ with $C'_\alpha > 0$ large enough.

For the distributed protocol described above, the test

$$T := \max_{L_s \in \mathcal{C}} T_{L_s}$$

satisfies

$$\mathbb{E}_0 T + \mathbb{E}_f(1-T) \leqslant \sum_{L_s \in \mathcal{C}} \mathbb{E}_0 T_{L_s} + \mathbb{E}_f(1 - T_{L^*_s}) \leqslant \alpha + \mathbb{E}_f(1 - T_{L_{s^*}})$$

for any $s^* \in [s_{\min}, s_{\max}]$. This means that the test T has its Type I error bounded by α, and its Type II error is also less than α whenever

$$\left\| \tilde{f}^{L_{s^*}} \right\|_2^2 \geqslant M_N^2 \left(\frac{2^{L_{s^*}}}{mn\sqrt{n\epsilon^2 \wedge 1}\sqrt{n\epsilon^2 \wedge 2^{L_{s^*}}}} \bigwedge \left(\frac{\sqrt{2^{L_{s^*}}}}{\sqrt{m}n\sqrt{n\epsilon^2 \wedge 1}} \bigvee \frac{1}{mn^2\epsilon^2} \right) \right),$$

for some nonnegative sequence M_N^2 that is at most of the order $\log^{9/2}(N)$. By the same computation as in the proof of Theorem 5.4, the above display is satisfied when $f \in H^{s^*,R}_{M_{N,s^*}\rho_{s^*}}$ with $s^* \in [s_{\min}, s_{\max}]$, $s \mapsto M_{N,s}$ as in the assumptions of the theorem and $s \mapsto \rho_s$ satisfying

$$\rho_s \gtrsim \left(\frac{1 \vee \rho_s^{-1/s}}{mn\sqrt{n\epsilon^2 \wedge 1}\sqrt{n\epsilon^2 \wedge (1 \vee \rho_s^{-1/s})}} \bigwedge \left(\frac{\sqrt{1 \vee \rho_s^{-1/s}}}{\sqrt{m}n\sqrt{n\epsilon^2 \wedge 1}} \bigvee \frac{1}{mn^2\epsilon^2} \right) \right),$$

where the latter follows from the choice $L_s = \lfloor s^{-1}\log_2(1/\rho_s)\rfloor \vee 1$. Solving for ρ_s yields the rates described by the theorem. Taking a positive sequence $\alpha \equiv \alpha_N$ converging zero slow enough (depending on $M_{N,s}$) completes the proof for shared randomness protocols.

An adaptive local randomness protocol: The procedure in this case is essentially the same as in the case of having access to shared randomness, except for using a different testing procedure when ϵ' is "large" relative to 2^{L_s}, m and n. Whenever $\epsilon' \leqslant 1/\sqrt{n}$ and $\epsilon' \leqslant 2^{L_s}/\sqrt{mn}$, or whenever $\epsilon' > 1/\sqrt{n}$ and $\epsilon' \leqslant 2^{L_s}/(\sqrt{m}n)$, set the test T_{L_s} equal to the one computed in (5.34) and let $(Y^{(j)}_{L_s})_{j\in[m]}$ be generated by (3.21) with as the underlying observations the wavelet coefficients $\tilde{X}^{(j)}_{0:L_s}$, i.e. setting $d = \nu_{L_s}$. We note here that the kernel generating the transcripts underlying this test require no shared randomness.

Otherwise, whenever $\epsilon' \leqslant 1/\sqrt{n}$ and $\epsilon' > 2^{L_s}/\sqrt{mn}$, or whenever $\epsilon' > 1/\sqrt{n}$ and $\epsilon' > 2^{L_s}/(\sqrt{m}n)$, let $(Y^{(j)}_{L_s})_{j\in[m]}$ be generated by (3.43) with as the underlying observations the wavelet coefficients $\tilde{X}^{(j)}_{0:L_s}$. Using these transcripts, let $T_{L_s} = T^{\epsilon',\delta'}_{III}$ (with $d = \nu_{L_s}$) as computed in (3.45), with $\kappa_\alpha = \kappa'_\alpha\sqrt{|\mathcal{C}|}$. By Lemma 3.24, this test has level $\alpha/|\mathcal{C}|$ for $\kappa'_\alpha > 0$ large enough. Furthermore, by the same lemma combined with Lemma 3.15, gives that T_{L_s} as such has Type II error less than α whenever

$$\left\|\tilde{f}^{L_s}\right\|_2^2 \geqslant C_\alpha|\mathcal{C}|\frac{2^{(3/2)L_s}\log(1+2^{L_s}nm)\log(1+nm)}{mn(n(\epsilon')^2\wedge 2^{L_s})},$$

which is in particular true whenever

$$\left\|\tilde{f}^{L_s}\right\|_2^2 \gtrsim \log^3(N)\log(1+2^{L_s}N)\log(1+N)\frac{2^{(3/2)L_s}}{mn(n\epsilon^2\wedge 2^{L_s})}.$$

Since $L_s = \lfloor s^{-1}\log_2(1/\rho_s)\rfloor \vee 1$,

$$\rho_s \gtrsim \left(\frac{(1\vee\rho_s^{-1/s})^{3/2}}{mn(n\epsilon^2\wedge(1\vee\rho_s^{-1/s}))}\bigwedge\left(\frac{\sqrt{1\vee\rho_s^{-1/s}}}{\sqrt{m}n\sqrt{n\epsilon^2\wedge 1}}\bigvee\frac{1}{mn^2\epsilon^2}\right)\right),$$

ensures that the test $T := \max_{L_s\in\mathcal{C}} T_{L_s}$ has Type I and $\mathbb{P}_f$-Type II error less than α whenever $f \in H^{s^*,R}_{M_{N,s^*}\rho_{s^*}}$ for any $s^* \in [s_{\min}, s_{\max}]$, $s \mapsto M_{N,s}$ as in the assumptions of the theorem. □

Chapter acknowledgements: The quote at the start of the chapter is taken from [60].

5.5 Appendix

5.5.1 Proof of the adaptation lower bounds Theorems 5.2 and 5.3

Let f^L and $\tilde{X}^{(j)}_{L':L}$ as defined in (5.6) and (5.7), respectively. Let $T = (T, K, \mathbb{P}^U)$ be a distributed testing protocol (with U degenerate in the case it is a local randomness protocol) and fix $\alpha \in (0,1)$. For given $s_{\min} < s_{\max}$, consider for $s \in [s_{\min}, s_{\max}]$ the map $s \mapsto \rho_s$.

Recall that for Ψ_L as defined in (5.28) and any distribution π_L on $\mathbb{R}^{\nu(L)}$, $\pi_L \circ \Psi_L^{-1}$ defines a probability measure on the Borel sigma algebra of $L_2[0,1]$. Define the mixture of the above probability measures by

$$\Pi = \frac{1}{|\mathcal{C}_0|} \sum_{L \in \mathcal{C}_0} \pi_L \circ \Psi_L^{-1}, \tag{5.35}$$

where $\mathcal{C}_0 \subseteq \mathcal{C}$. There exists a grid of points $\mathcal{S} \subset [s_{\min}, s_{\max}]$ such that the map $s \mapsto L_s$ is a one-to-one map from $\mathcal{S}$ to $\mathcal{C}$. Let $L \mapsto s_L$ denote its inverse.

By the same steps as in (2.82),

$$\sup_{f \in H^{s_L,R}_{c_\alpha \rho_{s_L}}} \mathbb{P}^Y_f(T=0) \geqslant \mathbb{P}^Y_{\pi_L}(T=0) - \pi_L \circ \Psi_L^{-1}\left(f \notin H^{s_L,R}_{c_\alpha \rho_{s_L}}\right), \tag{5.36}$$

for all $L \in \mathcal{C}$. Using the above display, we can bound the risk in the adaptive setting from below:

$$\begin{aligned}\sup_{s \in [s_{\min}, s_{\max}]} \mathcal{R}(H^{s,R}_{c_\alpha \rho_s}, T) &\geqslant \frac{1}{|\mathcal{C}|} \sum_{L \in \mathcal{C}} \mathcal{R}(H^{s_L,R}_{c_\alpha \rho_{s_L}}, T) \\ &\geqslant \mathbb{P}^Y_0(T=1) + \mathbb{P}^Y_\Pi(T=0) - \frac{1}{|\mathcal{C}_0|} \sum_{L \in \mathcal{C}_0} \pi_L \circ \Psi_L^{-1}\left(f \notin H^{s_L,R}_{c_\alpha \rho_{s_L}}\right).\end{aligned} \tag{5.37}$$

Taking π_L as in the proof of Theorem 5.1, then by the same reasoning as in proof the proof of Theorem 5.1 that the third term in the above display can be made arbitrarily small per choice of c_α for ρ_s satisfying (5.13)-(5.14). For the first two terms, define

$$\mathcal{L}^{Y|u}_{\pi_L} := \int \frac{d\mathbb{P}^{Y|U=u}_f}{d\mathbb{P}^{Y|U=u}_0} d\pi_L(f)$$

and note that

$$\begin{aligned}\mathbb{P}_0^Y(T=1)+\mathbb{P}_\Pi^Y(T=0) &= \frac{1}{|\mathcal{C}_0|}\sum_{L\in\mathcal{C}_0}\int \mathbb{P}_0^{Y|U=u}\left(T+\mathcal{L}_{\pi_L}^{Y|u}(1-T)\right)d\mathbb{P}^U(u)\\ &\geqslant \frac{1}{|\mathcal{C}_0|}\sum_{L\in\mathcal{C}_0}\int \mathbb{E}_0^{Y|U=u}\left(\gamma T+\mathcal{L}_{\pi_L}^{Y|u}(1-T)\right)\mathbb{1}\left\{\mathcal{L}_{\pi_L}^{Y|u}>\gamma\right\}d\mathbb{P}^U(u)\\ &\geqslant \gamma\frac{1}{|\mathcal{C}_0|}\sum_{L\in\mathcal{C}_0}\int \mathbb{P}_0^{Y|U=u}\left(\mathcal{L}_{\pi_L}^{Y|u}>\gamma\right)d\mathbb{P}^U(u),\end{aligned}$$

where the conditioning follows from the Markov chain structure and the inequality holds for $0<\gamma<1$. We can conclude that it suffices to show that for all $\varepsilon>0$,

$$\frac{1}{|\mathcal{C}_0|}\sum_{L\in\mathcal{C}_0}\mathbb{P}_0^{(Y,U)}\left(\left|\mathcal{L}_{\pi_L}^{Y|U}-1\right|>\varepsilon\right) \tag{5.38}$$

can be made arbitrarily small per small enough choice of c_α in order obtain the required lower bound in (5.37). Using $\mathbb{P}_0^{(Y,U)}=d\mathbb{P}^U d\mathbb{P}_0^{Y|U}$, conditioning on the $\mathbb{P}_0^{Y|U}$-variance of $\mathcal{L}_\Pi^{Y|u}$ with Chebyshev's inequality and $\mathbb{E}_0^{Y|U=u}\mathcal{L}_\Pi^{Y|u}=1$ lead to

$$\frac{1}{|\mathcal{C}_0|}\sum_{L\in\mathcal{C}_0}\mathbb{P}_0^{(Y,U)}\left(\left(\mathcal{L}_{\pi_L}^{Y|U}-1\right)^2>\varepsilon^2\right)\leqslant\frac{1}{|\mathcal{C}_0|}\sum_{L\in\mathcal{C}_0}\mathbb{P}^U\left(\mathbb{E}^{Y|U}(\mathcal{L}_{\pi_L}^{Y|U})^2>1+\zeta\right)+\frac{\zeta}{\varepsilon^2}$$

for all $\varepsilon>0$ and $\zeta>0$. Noting that $\mathbb{E}^{Y|U=u}(\mathcal{L}_{\pi_L}^{Y|U=u})^2\geqslant 1$, sufficiently bounding (5.38) follows from Markov's inequality and showing

$$\frac{1}{|\mathcal{C}_0|}\sum_{L\in\mathcal{C}_0}\int\log\left(\mathbb{E}^{Y|U=u}(\mathcal{L}_{\pi_L}^{Y|U=u})^2\right)d\mathbb{P}^U(u)\lesssim c_\alpha. \tag{5.39}$$

Noting that $\mathbb{E}^{Y|U=u}(\mathcal{L}_{\pi_L}^{Y|U=u})^2=D_{\chi^2}(\mathbb{P}_{0,K}^{Y|U=u};\mathbb{P}_{\pi_L,K}^{Y|U=u})+1$, we can apply the argument of the proof of Theorem 2.3 for bounding the chi-square divergence, and we obtain that for some fixed $C>0$,

$$\log\left(\mathbb{E}^{Y|U=u}(\mathcal{L}_{\pi_L}^{Y|U=u})^2\right)\leqslant\begin{cases}Cc_\alpha\frac{n^4\rho_{s_L}^4}{2^{3L}}\mathrm{Tr}\left(\Xi_{L,u}\right)^2+A_{L,u}, & \text{if } U \text{ is degenerate,}\\ Cc_\alpha\frac{mn^3\rho_{s_L}^4}{2^{2L}}\mathrm{Tr}\left(\Xi_{L,u}\right)+A_{L,u}, & \text{otherwise,}\end{cases} \tag{5.40}$$

where

$$A_{L,u}=\sum_{j=1}^m\log\left(\mathbb{E}_0^{Y^{(j)}|U=u}\left(\mathbb{E}_0\left[\int\frac{d\mathbb{P}_f^{\tilde{X}^{(j)}}}{d\mathbb{P}_0^{\tilde{X}^{(j)}}}(\tilde{X}_L^{(j)})d\pi_L(f)\middle|Y^{(j)},U=u\right]^2\right)\right)$$

and $\Xi_{L,u}=\sum_{j=1}^m\Xi_{L,u}^j$ with $\Xi_{L,u}^j=\mathbb{E}_0\mathbb{E}_0\left[\tilde{X}_L^{(j)}\middle|Y^{(j)},U=u\right]\mathbb{E}_0\left[\tilde{X}_L^{(j)}\middle|Y^{(j)},U=u\right]^\top$. Via a data processing argument (Lemma 5.5),

$$\frac{1}{|\mathcal{C}_0|}\sum_{L\in\mathcal{C}_0}\int A_{L,u}d\mathbb{P}^U(u)\lesssim\max_{L\in\mathcal{C}_0}\frac{c_\alpha mn^2\rho_{s_L}^4(b\wedge|\mathcal{C}_0|)}{2^L|\mathcal{C}_0|}.$$

When U is degenerate, Lemma 5.3 implies that there exists a choice for $\mathcal{C}_0 \subset \mathcal{C}_0$ such that for all $L \in \mathcal{C}_0$,

$$\mathrm{Tr}\left(\Xi_{L,u}\right)^2 \lesssim \left(\frac{b}{|\mathcal{C}|} \wedge 2^L\right)^2 \frac{m^2}{n^2}.$$

When U is not degenerate, Lemma 5.4 implies that taking $\mathcal{C}_0 = \mathcal{C}$,

$$\frac{1}{|\mathcal{C}|}\sum_{L\in\mathcal{C}} \frac{mn^3\rho_{s_L}^4}{2^{2L}}\mathrm{Tr}\left(\Xi_{L,u}\right) \lesssim \max_{L\in\mathcal{C}} \frac{N^2\rho_{s_L}^4}{2^{2L}}\left(\frac{b}{|\mathcal{C}|} \wedge 2^L\right).$$

Combining the above with the fact that $s \mapsto L_s = \lfloor s^{-1}\log(1/\rho_s)\rfloor \vee 1$ maps a grid $\mathcal{S} \subset [s_{\min}, s_{\max}]$ one-to-one to $\mathcal{C}_0$ with inverse map $L \mapsto s_L$ on $\mathcal{C}_0$, we obtain

$$\frac{1}{|\mathcal{C}_0|}\sum_{L\in\mathcal{C}_0}\int \log\left(\mathbb{E}^{Y|U=u}(\mathcal{L}_{\pi_L}^{Y|U=u})^2\right) d\mathbb{P}^U(u) \lesssim c_\alpha \cdot \begin{cases} \max\limits_{L\in\mathcal{C}} \frac{N^2\rho_{s_L}^4\left(\frac{b}{\log(n)}\wedge 2^L\right)^2}{2^{3L}} \bigvee \frac{N^2\rho_{s_L}^4(b\wedge\log(N))}{m2^L\log(N)}, \\ \max\limits_{L\in\mathcal{C}} \frac{N^2\rho_{s_L}^4\left(\frac{b}{\log(n)}\wedge 2^L\right)}{2^{2L}} \bigvee \frac{N^2\rho_{s_L}^4(b\wedge\log(N))}{m2^L\log(N)}, \end{cases}$$

where the first case corresponds to a degenerate U, the latter to the general (shared randomness) case. The conditions (5.13)-(5.14) for ρ_{s_L} yield (5.39), which in turn finishes the proof.

5.5.2 Auxiliary lemmas concerning adaptation under bandwidth constraints

The following lemma controls the Type I error of the adaptive tests defined in Section 5.2.

Lemma 5.1. *Consider for $L \in \mathbb{N}$ and a nonnegative positive integer sequence K_n,*

$$S_n(L) := \frac{1}{\sqrt{K_n}}\sum_{i=1}^{K_n}\zeta_{i,L}$$

where $(\zeta_{1,L}, \ldots, \zeta_{K_n,L})$ independent random variables with mean 0 and unit variance.

Assume that the random variables satisfy Cramér's condition, i.e. for some $\epsilon > 0$ and all $t \in (-\epsilon, \epsilon)$, $i = 1, \ldots, K_n$ and $L \in \mathcal{C}$, for some set $\mathcal{C} \subset \mathbb{N}$ satisfying $|\mathcal{C}| \asymp \log(n)$,

$$\mathbb{E}e^{t\zeta_{i,L}} < \infty.$$

Then for $K_n \gg (\log\log n)^6$, it holds that

$$Pr\left(\max_{L\in\mathcal{C}} |S_n(L)| \geqslant c\sqrt{\log\log(n)}\right) \to 0$$

for all $c > \sqrt{2}$ as $n \to \infty$.

If the random variables are i.i.d. Rademacher or are of the form

$$\zeta_{i,L} = \frac{1}{4Q}\left[\left(\sum_{q=1}^{Q} R_{qL}\right)^2 - Q\right]$$

with $R = (R_{1L}, \ldots, R_{QL})$ *independent Rademacher random variables and* $Q \in \mathbb{N}$*, the statement holds for any sequence* K_n *as* $n \to \infty$*.*

Proof. By using union bounds,

$$\Pr\left(\max_{L\in\mathcal{C}} S_n(L) \geqslant c\sqrt{\log\log(n)}\right) \leqslant \sum_{L\in\mathcal{C}} \Pr\left(|S_n(L)| \geqslant c\sqrt{\log\log(n)}\right) \leqslant$$
$$\sum_{L\in\mathcal{C}}\left[\Pr\left(S_n(L) \geqslant c\sqrt{\log\log(n)}\right) + \Pr\left(-S_n(L) \geqslant c\sqrt{\log\log(n)}\right)\right].$$

The proof follows by showing that $S_n(L)$ and $-S_n(L)$ are or tend to sub-Gaussian variables with sub-Gaussianity constant less than or equal to 1, since this allows for bounding the above display by

$$2\sum_{L\in\mathcal{C}} e^{-\frac{c^2}{2}\log\log(n)} \lesssim \frac{1}{(\log(n))^{c^2/2-1}}$$

and the result follows.

For the first statement, by Cramér's theorem (see e.g. Theorem 7 in Section 8.2 of [165]),

$$\frac{\Pr\left(S_n(L) \geqslant c\sqrt{\log\log(n)}\right)}{1-\Phi(c\sqrt{\log\log(n)})} = \exp\left(O(1)\cdot\frac{(\log\log n)^3}{\sqrt{K_n}}\right)\left(1+O\left(\frac{\log\log n}{\sqrt{K_n}}\right)\right) \to 1.$$

Note that the above statement holds for $-S_n(L)$ also. The statement now follows by using $1-\Phi(x) \leqslant e^{-x^2/2}$.

For the second statement, note that by symmetry of the Rademacher distribution, it suffices to consider only $S_n(L)$. In case the $\zeta_{i,L}$'s are i.i.d. Rademacher, note that a Chernoff bound yields

$$\Pr\left(S_n(L) \geqslant c\sqrt{\log\log(n)}\right) \leqslant \inf_{t>0} e^{\frac{t^2}{2} - ct\sqrt{\log\log(n)}} = e^{-\frac{c^2}{2}\log\log(n)}. \tag{5.41}$$

Similarly, for the sum of Rademacher random variables, we have

$$\mathbb{E}\exp\left(\frac{t}{\sqrt{K_n}}\zeta_{i,L}\right) = \mathbb{E}\exp\left(\frac{t}{4Q\sqrt{K_n}}\left[\sum_{q\neq q'}^{Q} R_{qL}R_{q'L}\right]\right)$$
$$\leqslant \mathbb{E}\exp\left(\frac{t}{Q\sqrt{K_n}}\left[\sum_{q\neq q'}^{Q} R_{qL}R'_{q'L}\right]\right),$$

where the inequality follows from e.g. Theorem 6.1.1 in [210] with $R' = (R'_{1L}, \dots, R'_{QL})$ independent of R. The latter implies that $(R_{qL}R'_{q'L})_{(q,q')\in\{1,\dots,Q\}^2}$ itself is a vector of independent Rademacher random variables, and consequently the above display is further bounded by

$$\exp\left(\frac{t^2Q(Q-1)}{2K_nQ^2}\right) \leqslant \exp\left(\frac{t^2}{2K_n}\right).$$

The proof of the last statement now follows via Chernoff bound as in (5.41). □

The next lemma controls the Type 2 error of the adaptive test in the high-budget case for the shared randomness protocol.

Lemma 5.2. *Consider $S_{II}(L_s)$ as in (5.19) in the paper. It holds that*

$$\mathbb{E}_f \mathbb{1}\left\{S_{II}(L_s) < 2\sqrt{\log\log n}\right\} \leqslant \alpha/2$$

whenever $f \in H^{s,R}_{C_\alpha\rho_s}$ with $\rho^2 \geqslant C_0\sqrt{\log\log(n)}\frac{2^{L_s}}{n\sqrt{\frac{b}{\log(n)}\wedge 2^{L_s}}}$ for C_0 large enough, depending only on R.

Proof. The proof is similar in spirit to that of the risk bound in the finite dimensional, non-adaptive, shared randomness setting given in Lemma 3.3.

We show below that the event

$$A = \left\{\frac{m'-1}{2\sqrt{b'}}\sum_{i=1}^{b'}(Y^{(j)}_{\mathrm{II}}(L))_i - 1/2)^2 \geqslant 2\sqrt{\log\log N}\right\},$$

occurs with $\mathbb{P}_f$-probability greater than $1-\alpha/4$. Since on A the condition of Lemma 3.16 is satisfied with $c_{\alpha,n} = 2\sqrt{\log\log N}$ and consequently, by the conclusion of Lemma 3.16, $\mathbb{E}_f \mathbb{1}\left\{S_{\mathrm{II}}(L_s) < 2\sqrt{\log\log N}\right\}$ is bounded by $\alpha/2$.

Following the proof of Lemma 3.3 (with $d = \nu_{L_s}$, considering the ν_{L_s} dimensional vector $f^{\nu_{L_s}}$, and taking $N_\alpha = 2\sqrt{\log\log N}$), and noting that for $C_0^2 > 4R^2$

$$\|\tilde{f}^{L_s}\|_2^2 \geqslant \|f\|_2^2/2 - R^2 2^{-2L_s s} \gtrsim \frac{C_0 2^{L_s}\sqrt{\log\log(N)}}{2N\sqrt{\frac{b}{\log(N)}\wedge 2^{L_s}}} \gtrsim \frac{C_0 2^{L_s}\sqrt{\log\log(N)}}{N\sqrt{b'\frac{m'}{m}}},$$

we get that

$$\mathbb{E}_f \mathbb{1}_{A^c} \leqslant \Pr\left(\frac{m'-1}{24\sqrt{b'}}\sum_{i=1}^{b'}\min\left\{\frac{C_0\sqrt{\log\log N}2^{L_s}Z_i^2}{2m'\sqrt{b'}\|Z\|_2^2}, 1\right\} \leqslant 2\sqrt{\log\log N}\right). \quad (5.42)$$

Considering the intersection with the event $\{\|Z\|_2^2 \leqslant k2^{L_s}\}$ for some large enough $k > 0$, and noting that by Lemma 3.27,

$$\Pr\left(\max_{1\leqslant i\leqslant b'} Z_i^2 \geqslant \frac{2m'\sqrt{b'}k}{C_0\sqrt{\log\log N}}\right) \leqslant 2b'\exp\left(-\frac{m'\sqrt{b'}k}{2C_0\sqrt{\log\log N}}\right) = o(1),$$

the right-hand side of (5.42) is further bounded by

$$\Pr\left(\sum_{i=1}^{b'} Z_i^2 \leqslant \frac{96b'm'k}{C_0(m'-1)}\right) + o(1) + \alpha/8 \leqslant \alpha/4,$$

where the last inequality holds for large enough choices $m' := \frac{m(b\wedge\log(n))}{\log(n)}$, $b' := \frac{mb}{m'|\mathcal{C}|} \wedge \nu_L$ and large enough choice of C_0 (depending on k), see e.g. (3.7) in the proof of Lemma 3.3, which finishes the proof of our statement. □

Next we provide the lemmas for the lower bound. From now on in this section, we consider the setting of Section 5.2. That is, let $\tilde{X}_L^{(j)}$, $\tilde{X}_{1:L}^{(j)}$ denote the wavelet coefficients of $X^{(j)}$ as in (5.7). Define in addition the matrices

$$\Xi_{L,u}^j = \mathbb{E}_0\mathbb{E}_0\left[\tilde{X}_L^{(j)}|Y^{(j)}, U=u\right]\mathbb{E}_0\left[\tilde{X}_L^{(j)}|Y^{(j)}, U=u\right]^\top,$$
$$\Xi_{L':L,u}^j = \mathbb{E}_0\mathbb{E}_0\left[\tilde{X}_{L':L}^{(j)}|Y^{(j)}, U=u\right]\mathbb{E}_0\left[\tilde{X}_{L':L}^{(j)}|Y^{(j)}, U=u\right]^\top,$$

$\Xi_{L,u} := \sum_{j=1}^m \Xi_{L,u}^j$ and $\Xi_u = \sum_{j=1}^m \Xi_{L_{\min}:L_{\max},u}^j$. The lemma below allows for extending the data processing inequality of Lemma 2.11 to the adaptive local randomness case, in which extra demands are placed on the communication budget in terms of the budget needing to cover the coordinates corresponding to each resolution level.

Lemma 5.3. *Suppose $Y^{(j)}$ takes values in a space with cardinality at most $2^b \in \mathbb{N}$, for $j = 1, \ldots, m$ and let $\mathcal{C} = \{L_{\min}, \ldots, \mathbb{L}_{\max}\}$, for some $L_{\min} < L_{\max} \in \mathbb{N}$. There exists $\mathcal{C}_0 \subset \mathcal{C}$ such that*

$$Tr(\Xi_{L,u}) \lesssim \left(\frac{b}{|\mathcal{C}|} \wedge 2^L\right)\frac{m}{n}$$

for all $L \in \mathcal{C}_0$.

Proof. Define $\Delta_L = \mathrm{Tr}\,(\Xi_{L,u})$ and let $\ell : \{1, \ldots, L_{\max} - L_{\min} + 1\} \to \mathcal{C}$ a map that respects the ordering of the Δ_L's in the sense that

$$\Delta_{\ell(i)} \leqslant \Delta_{\ell(k)} \text{ if } i \leqslant k.$$

Let $\mathcal{C}_0$ denote the first $\lfloor \frac{L_{\max}-L_{\min}+1}{2} \rfloor$ elements of the collection $\{\Delta_{\ell 1}, \Delta_{\ell(2)}, \ldots, \Delta_{\ell(L_{\max}-L_{\min}+1)}\}$. For all $L^\circ \in \mathcal{C}$,

$$\mathrm{Tr}\,(\Xi_{L^\circ,u}) \leqslant \frac{2}{|\mathcal{C}|}\sum_{L\in\mathcal{C}\setminus\mathcal{C}_0} \mathrm{Tr}\,(\Xi_{L,u}).$$

By definition of the trace of a matrix, $\sum_L \mathrm{Tr}(\Xi_{L,u}) = \mathrm{Tr}(\Xi_{L_{\min}:L_{\max},u})$. By Lemma 2.11,

$$\mathrm{Tr}\,(\Xi_{L_{\min}:L_{\max},u}) = \sum_{j=1}^m \mathrm{Tr}\left(\Xi_{L_{\min}:L_{\max},u}^j\right) \leqslant \frac{2\log(2)mb}{n}.$$

Combining the above two displays, we obtain that

$$\mathrm{Tr}\left(\Xi_{L^\circ,u}\right) \lesssim \frac{mb}{n(|\mathcal{C}|}.$$

By an application of Lemma 2.11 and a straightforward computation as in the proof of Lemma 2.11,

$$\mathrm{Tr}\left(\Xi_{L^\circ,u}\right) \leqslant \frac{m}{n} 2^{L^\circ}. \tag{5.43}$$

Combining the two bounds for $\mathrm{Tr}\left(\Xi_{L^\circ,u}\right)$ gives the result. □

The next lemma applies to the adaptive shared randomness setting. The bound below is slightly more relaxed than the previous one, which relates to the local randomness setting. The reason for this is the fact that in the shared randomness setting, the hyperprior cannot be chosen in an adversarial way because of the shared randomness draw essentially allowing multiple (coordinated) protocols across the m machines.

Lemma 5.4. *With the notation as in the proof of Theorem 5.2, it holds that*

$$\frac{1}{|\mathcal{C}|}\sum_{L\in\mathcal{C}} \frac{N^3\rho_{s_L}^4}{m2^{2L}} Tr\left(\Xi_{L,u}\right) \lesssim \max_{L\in\mathcal{C}} \frac{N^2\rho_{s_L}^4}{2^{2L}} \left(\frac{b}{|\mathcal{C}|} \wedge 2^L\right).$$

Proof. Similarly to the proof of Lemma 5.3, we note that by the linearity of the trace,

$$\sum_{L\in\mathcal{C}} \mathrm{Tr}\left(\Xi_{L,u}\right) = \mathrm{Tr}\left(\Xi_u\right),$$

where $\Xi_u = \sum_{j=1}^m \Xi^j_{L_{\min}:L_{\max},u}$. Lemma 2.11 yields $\mathrm{Tr}\left(\Xi_u\right) \leqslant 2\log(2)\frac{bm}{n}$. Otherwise, applying Lemma 2.11 yields $\mathrm{Tr}\left(\Xi_{L,u}\right) \leqslant \frac{2^L m}{n}$. Combining these two inequalities yields the result:

$$\begin{aligned}
\frac{1}{|\mathcal{C}|}\sum_{L\in\mathcal{C}} \frac{N^3\rho_{s_L}^4}{m2^{2L}} \mathrm{Tr}\left(\Xi_{L,u}\right) &\leqslant \frac{1}{|\mathcal{C}|}\sum_{L\in\mathcal{C}} \frac{N^2\rho_{s_L}^4}{2^{2L}} \left(\frac{n}{m}\mathrm{Tr}\left(\Xi_{L,u}\right) \bigwedge 2^L\right) \\
&\leqslant \max_{L^*} \frac{N^2\rho_{s_L^*}^4}{2^{2L^*}} \left(\frac{n}{m}\frac{1}{|\mathcal{C}|}\sum_{L\in\mathcal{C}} \mathrm{Tr}\left(\Xi_{L,u}\right) \bigwedge 2^{L^*}\right) \\
&\lesssim \max_{L^*} \frac{N^2\rho_{s_L^*}^4}{2^{2L^*}} \left(\frac{b}{|\mathcal{C}|} \bigwedge 2^{L^*}\right).
\end{aligned}$$

□

Whereas in the nonadaptive setting of Theorem 2.3 and Theorem 5.1 the local "chi-square" based terms need no special data processing treatment, it does in the adaptive case. For each of the $\log(N)$ resolution levels L, information on the norm of $\tilde{X}_L^{(j)}$ is

communicated. Using $b \asymp \log(N)$ to this without loss (compared to Theorem 5.1) turns out to be fundamental, as is the content of the lemma below. The proof of the lemma is based on exploiting the fact that even though $2^{-L/2}(\|\sqrt{n/m}\tilde{X}_L^{(j)}\|_2^2 - 2^L)$ is sub-exponential, the fact that it tends to a sub-Gaussian random variable can be exploited whenever the communication budget is small enough.

Lemma 5.5. *Let π_L as in the proof of Theorem 5.2, with $\rho_s = \rho_{s_L}$ satisfying* (5.14) *or* (5.13)*. Furthermore, let*

$$A_{L,u} = \sum_{j=1}^{m} \log\left(\mathbb{E}_0^{Y^{(j)}|U=u}\left(\mathbb{E}_0\left[\int \frac{d\mathbb{P}_f^{\tilde{X}^{(j)}}}{d\mathbb{P}_0^{\tilde{X}^{(j)}}}(\tilde{X}_L^{(j)})d\pi_L(f)\middle| Y^{(j)}, U=u\right]^2\right)\right).$$

Then for arbitrary $\mathcal{C} \subset \mathbb{N}$,

$$\frac{1}{|\mathcal{C}|}\sum_{L\in\mathcal{C}}\int A_{L,u}d\mathbb{P}^U(u) \lesssim \max_{L\in\mathcal{C}} \frac{c_\alpha m n^2 \rho_{s_L}^4 (b \wedge |\mathcal{C}|)}{2^L|\mathcal{C}|}.$$

Proof. Recalling the notation from Section 5.1.1, we shall write

$$\mathscr{L}_{\pi_L}(\tilde{X}_L^{(j)}) = \int \mathscr{L}_f(\tilde{X}_L^{(j)})d\pi_L(f)$$

with

$$\mathscr{L}_f(\tilde{X}_L^{(j)}) := \frac{d\mathbb{P}_f^{\tilde{X}^{(j)}}}{d\mathbb{P}_0^{\tilde{X}^{(j)}}}(\tilde{X}_L^{(j)}) = e^{nf^\top \tilde{X}_L^{(j)} - \frac{n}{2}\|f\|_2^2}.$$

Note that, using $\log(x) \leqslant x-1$, $\mathbb{E}_0\mathscr{L}_{\pi_L}(\tilde{X}_L^{(j)}) = 1$ and the fact that by the law of total probability

$$\mathbb{E}_0^{Y^{(j)}|U=u}\mathbb{E}_0\left[\mathscr{L}_{\pi_L}(\tilde{X}_L^{(j)})\middle| Y^{(j)}, U=u\right] = 1,$$

we obtain that

$$A_{L,u} \leqslant \sum_{j=1}^{m}\mathbb{E}_0^{Y^{(j)}|U=u}\left(\mathbb{E}_0\left[\mathscr{L}_{\pi_L}(\tilde{X}_L^{(j)}) - 1\middle| Y^{(j)}, U=u\right]^2\right). \tag{5.44}$$

We work out the case where $\pi = N(0, \varrho_s^2 I_{2^L})$, the case where $\pi = N(0, \varrho_s^2\Gamma)$ with $\|\Gamma\| \asymp 1$ follows similarly with additional bookkeeping. Since $f \sim N(0, \varrho_s^2 I_{2^L})$ with $\varrho_s = c_\alpha^{1/4}\rho_s/2^{L/2}$,

$$\mathscr{L}_{\pi_L}(\tilde{X}_L^{(j)}) = \prod_{i=0}^{2^L-1}\int \frac{e^{nf_i\tilde{X}_{Li}^{(j)} - \frac{1}{2}(n+\varrho_s^{-2})f_i^2}}{\sqrt{2\pi\varrho_s^2}}df_i = \frac{e^{n\varrho_s^2\frac{\|\sqrt{n}\tilde{X}_L^{(j)}\|_2^2}{2(1+n\varrho_s^2)}}}{(1+n\varrho_s^2)^{2^L/2}} \tag{5.45}$$

where the last equality follows by the substitution $u = f_i\sqrt{1+n\varrho_s^2}$ and completing the square. Taking the logarithm and using that $\frac{(1+x)\log(1+x)}{x} > 1$ for $x > 0$, we find

$$V_L^{(j)} := n\varrho_s^2 \frac{\|\sqrt{n}\tilde{X}_L^{(j)}\|_2^2}{2(1+n\varrho_s^2)} - 2^{L-1}\log(1+n\varrho_s^2) \leqslant \frac{n\varrho_s^2}{2}\left(\|\sqrt{n}\tilde{X}_L^{(j)}\|_2^2 - 2^L\right) \tag{5.46}$$

Therefore, using (5.45), Taylor expanding, $(a+b)^2 \leqslant 2a^2 + 2b^2$ and (5.46), we can upper bound (5.44) by

$$2\sum_{j=1}^m \mathbb{E}_0^{Y^{(j)}|U=u}\left(\mathbb{E}_0\left[V_L^{(j)}\middle|Y^{(j)}, U=u\right]^2\right) + 2\sum_{j=1}^m \mathbb{E}_0^{Y^{(j)}|U=u}(D^j)^2, \tag{5.47}$$

with

$$D^j = \mathbb{E}_0\left[\sum_{k=2}^\infty \frac{n^k\varrho_s^{2k}}{2^k k!}\left|\|\sqrt{n}\tilde{X}_L^{(j)}\|_2^2 - 2^L\right|^k \middle| Y^{(j)}, U=u\right].$$

We deal with the two terms in (5.47) separately. Since conditional expectation contracts the L_2-norm,

$$\sum_{j=1}^m \mathbb{E}_0^{Y^{(j)}|U=u}(D^j)^2 \lesssim m\cdot\sum_{k=2}^\infty\sum_{i=2}^\infty \frac{n^k c_\alpha^{k/2}\rho_s^{2k}}{2^k 2^{kL_s/2}k!}\frac{n^i c_\alpha^{i/2}\rho_s^{2i}}{2^i 2^{iL_s/2}i!}\mathbb{E}W^{i+k}$$

where $W \stackrel{d}{=} 2^{-L/2}\left(\|\sqrt{n}\tilde{X}_L^{(j)}\|_2^2 - 2^L\right)$. Furthermore, since $\|\sqrt{n}X_L^{(j)}\|_2^2 \sim \chi^2_{2^L}$ is $2^{L/2}$-sub-exponential, $\mathbb{E}W^{i+k} \leqslant C^{k+i}(i+k)^{i+k}$, where $C > 0$ is a constant (see e.g. Proposition 2.7.1 in [210]). Then in view of $(i+k)^{i+k} \leqslant 2^{i+k}i!k!$, we the above display is $O(\frac{c_\alpha^2 n^4\rho_s^8}{2^{2L_s}})$ whenever $\frac{c_\alpha^2 n^4\rho_s^8}{C^2 2^{2L_s}} < 1$. This is certainly the case when $\rho_s^2 \lesssim \left(\frac{\sqrt{m\log(mn)}}{mn\sqrt{b\wedge\log(nm)}}\right)^{\frac{2s}{2s+1/2}}$ and $mb \gtrsim \log(nm)$, which yields that

$$\sum_{j=1}^m \mathbb{E}_0^{Y^{(j)}|U=u}(D^j)^2 \lesssim \frac{c_\alpha^2 n^2\rho_s^4}{m2^{L_s/2}}\cdot O\left(\frac{\log(mn)}{m(b\wedge\log(n))}\right).$$

It remained to deal with the first term in (5.47), where we proceed by a data processing argument. When $b \geqslant \log(n)$,

$$2\sum_{j=1}^m \mathbb{E}_0^{Y^{(j)}|U=u}\left(\mathbb{E}_0\left[V_L^{(j)}\middle|Y^{(j)}, U=u\right]^2\right) \leqslant 2\sum_{j=1}^m \mathbb{E}_0^{\tilde{X}^{(j)}}\left(V_L^{(j)}\right)^2 \leqslant \frac{c_\alpha mn^2\rho_s^4}{2^{L_s}},$$

in which case the result follows.

We continue with the case where $b < \log(n)$, which implies $|\mathcal{Y}^{(j)}| \leqslant 2^{\log(n)}$. We bound the average of the first terms in (5.47) over $\mathcal{C}$, by

$$\frac{1}{|\mathcal{C}|}\sum_{L\in\mathcal{C}}\sum_{j=1}^{m}\frac{n^2\rho_s^4}{2^{L_s}}\mathbb{E}_0^{Y^{(j)}|U=u}\left(\mathbb{E}_0\left[G_L^{(j)}\middle|Y^{(j)},U=u\right]^2\right)\leqslant \max_{L\in\mathcal{C}}\frac{n^2\rho_{s_L}^4}{2^L|\mathcal{C}|}\sum_{j=1}^{m}\mathbb{E}_0^{Y^{(j)}|U=u}\mathrm{Tr}(M^{(j)}(Y^{(j)})), \tag{5.48}$$

where $M^{(j)}(y) = \mathbb{E}_0\left[G_{\mathcal{C}}^{(j)}\middle|Y^{(j)}=y,U=u\right]\mathbb{E}_0\left[G_{\mathcal{C}}^{(j)}\middle|Y^{(j)}=y,U=u\right]^\top$, $G_{\mathcal{C}}^{(j)} = (G_L^{(j)})_{L\in\mathcal{C}}$, and $G_L^{(j)} = \left(\frac{n\rho_s^2}{2^{L_s/2}}\right)^{-1}V_L^{(j)}$. We show below that for all $v = (v_L)_{L\in\mathcal{C}}$ of unit norm

$$\mathbb{E}_0^{Y^{(j)}|U=u}\langle v_{\mathcal{C}}, G_{\mathcal{C}}^{(j)}\rangle^2 \leqslant b, \tag{5.49}$$

which by taking $v = G_{\mathcal{C}}^{(j)}/\|G_{\mathcal{C}}^{(j)}\|_2$ yields that (5.48) is $O(\max_s \frac{mn^2\rho_s^4}{2^{L_s}|\mathcal{C}|}b)$ as required. Therefore, it remains to verify (5.49). For any $\lambda\in\mathbb{R}$, independence and (5.46) yield

$$\mathbb{E}_0^{X^{(j)}}e^{\lambda v^\top G_{\mathcal{C}}^{(j)}} \leqslant \prod_{L\in\mathcal{C}}\mathbb{E}_0^{X^{(j)}}e^{\frac{\lambda}{2^{L_s/2}}v_L\sum_{i=0}^{2^L-1}(\tilde{X}_{Li}^2-1)}.$$

When $\frac{|\lambda|}{2\cdot 2^{L_s/2}}v_L \leqslant \frac{1}{4}$, the latter can be further bounded by

$$\prod_{L\in\mathcal{C}}\exp\left(\lambda^2 v_L^2\right) = \exp\left(\lambda^2\right),$$

see e.g. Lemma 12 in [192]. In view of $0\leqslant K(y|X^{(j)},u)\leqslant 1$ and the previously shown sub-exponential behavior of $\langle v_{\mathcal{C}}, G_{\mathcal{C}}^{(j)}\rangle$, we get that

$$\begin{aligned}
&\mathbb{P}^{Y^{(j)}|U=u}(y)\mathbb{E}_0\left[\langle v_{\mathcal{C}}, G_{\mathcal{C}}^{(j)}\rangle\middle|Y^{(j)}=y,U=u\right]\\
&\quad=\mathbb{E}_0^{X^{(j)}}\langle v_{\mathcal{C}}, G_{\mathcal{C}}^{(j)}\rangle K(y|X^{(j)},u)\leqslant\mathbb{E}_0^{X^{(j)}}\int_0^\infty \mathbb{1}\left\{|\langle v_{\mathcal{C}}, G_{\mathcal{C}}^{(j)}\rangle|>t\right\}K(y|X^{(j)},u)dt\\
&\quad\leqslant\int_0^\infty\min\left\{\mathbb{P}_0^{X^{(j)}}\left(|\langle v_{\mathcal{C}}, G_{\mathcal{C}}^{(j)}\rangle|>t\right),\mathbb{P}^{Y^{(j)}|U=u}(y)\right\}dt\leqslant e^{-t_0}+t_0\mathbb{P}^{Y^{(j)}|U=u}(y).
\end{aligned}$$

Taking $t_0 = -\log(\mathbb{P}^{Y^{(j)}|U=u}(y))$ yields

$$\mathbb{E}_0\left[\langle v_{\mathcal{C}}, G_{\mathcal{C}}^{(j)}\rangle\middle|Y^{(j)}=y,U=u\right]\leqslant -2\log(\mathbb{P}^{Y^{(j)}|U=u}(y))\vee(1-\log(\mathbb{P}^{Y^{(j)}|U=u}(y))). \tag{5.50}$$

Furthermore, for $\lambda_y \in \mathbb{R}$ and y satisfying

$$-2^{L_s/2+2} \leqslant \lambda_y = \mathbb{E}_0\left[\langle v_{\mathcal{C}}, G_{\mathcal{C}}^{(j)}\rangle \middle| Y^{(j)} = y, U = u\right] \leqslant 2^{L_s/2+2}, \tag{5.51}$$

the argument of Lemma 2.11 yields

$$\mathbb{E}_0\left[\langle v_{\mathcal{C}}, G_{\mathcal{C}}^{(j)}\rangle \middle| Y^{(j)} = y, U = u\right]^2 \leqslant -\log\left(\mathbb{P}^{Y^{(j)}|U=u}(y)\right). \tag{5.52}$$

Note, that if (5.51) does not hold, then in view of (5.50), $-\log(\mathbb{P}^{Y^{(j)}|U=u}(y)) \geqslant 2^{L_s/2+1}$.

Let us write $p_y = \mathbb{P}^{Y^{(j)}|U=u}(y)$ and define $\mathcal{Y}_*^{(j)} = \{y \in \mathcal{Y}^{(j)} : \log(1/p_y) \leqslant 2^{L_s/2+2}\}$. Since $x \mapsto x\log^2(1/x)$ is increasing on $(0, e^{-2})$, it holds that

$$p_y \log^2(1/p_y) \leqslant e^{-2^{L_s/2+2}+(L_s+4)\log(2)}$$

for $y \in (\mathcal{Y}_*^{(j)})^c$. Then, in view of (5.50) and (5.52) we get that

$$\begin{aligned}\sum_{y\in\mathcal{Y}^{(j)}} p_y \mathbb{E}_0\left[\langle v_{\mathcal{C}}, G_{\mathcal{C}}^{(j)}\rangle \middle| Y^{(j)} = y, U = u\right]^2 &\leqslant \sum_{y\in\mathcal{Y}_*^{(j)}} p_y \log(1/p_y) + 4\sum_{y\in(\mathcal{Y}_*^{(j)})^c} p_y \log^2(1/p_y) \\ &\lesssim \log|\mathcal{Y}^{(j)}| + 2^b e^{-2^{L_s/2+2}+(L_s+4)\log(2)} \lesssim b,\end{aligned}$$

concluding the proof of (5.49) and hence the lemma. □

5.5.3 Definitions and notations for wavelets

In this section we briefly introduce wavelets and collect some properties used in the article. For a more detailed and elaborate introduction of wavelets we refer to [115, 106].

In our work we consider the Cohen, Daubechies and Vial construction of compactly supported, orthonormal, N-regular wavelet basis of $L_2[0,1]$, see for instance [64]. First for any $N \in \mathbb{N}$ one can follow Daubechies' construction of the father $\phi(.)$ and mother $\psi(.)$ wavelets with N vanishing moments and bounded support on $[0, 2N-1]$ and $[-N+1, N]$, respectively, see for instance [70]. The basis functions are then obtained as

$$\left\{\phi_{j_0 m}, \psi_{jk} : m \in \{0, \ldots, 2^{j_0}-1\}, \quad j > j_0, \quad k \in \{0, \ldots, 2^j - 1\}\right\},$$

with $\psi_{jk}(x) = 2^{j/2}\psi(2^j x - k)$, for $k \in [N-1, 2^j - N]$, and $\phi_{j_0 k}(x) = 2^{j_0}\phi(2^{j_0}x - m)$, for $m \in [0, 2^{j_0} - 2N]$, while for other values of k and m, the basis functions are specially constructed, to form a basis with the required smoothness property. For notational

convenience we take $j_0 = 0$ and denote the father wavelet by ψ_{00}. Then the function $f \in L_2[0,1]$ can be represented in the form

$$f = \sum_{j=j_0}^{\infty} \sum_{k=0}^{2^j-1} f_{jk}\psi_{jk},$$

with $f_{jk} = \langle f, \psi_{jk} \rangle$. Note that in view of the orthonormality of the wavelet basis the L_2-norm of the function f is equal to

$$\|f\|_2^2 = \sum_{j=j_0}^{\infty} \sum_{k=0}^{2^j-1} f_{jk}^2.$$

Next we give an equivalent definition of Sobolev spaces using wavelets. Let us define the norm for $s \in (0, N)$ as

$$\|f\|_{\mathcal{H}^s}^2 = \sum_{j \geqslant j_0} 2^{2js} \sum_{k=0}^{2^j-1} f_{jk}^2.$$

Then the Sobolev space $\mathcal{H}^s([0,1])$ and Sobolev ball $\mathcal{H}^{s,R}([0,1])$ of radius $R > 0$ are defined as

$$\mathcal{H}^s = \{f \in L_2[0,1] : \|f\|_{\mathcal{H}^s} < \infty\}, \quad \text{and} \quad \mathcal{H}^{s,R}([0,1]) = \{f \in L_2[0,1] : \|f\|_{\mathcal{H}^s} \leqslant R\},$$

respectively. The above definition of the Sobolev space and norm is equivalent to the classical one based on the weak derivatives of the function (see e.g. Chapter 4 in [106]). Similarly, we can define s-smooth Hölder function spaces using wavelets. Consider the norm

$$\|f\|_{C^s} := \|f\|_\infty + \sup_{j \geqslant 0} 2^{js+1/2} \max_{0 \leqslant k \leqslant 2^j-1} |f_{jk}|,$$

which is equivalent to the s-smooth Hölder norm defined through the modulus of smoothness (e.g. Chapter 4 in [106]). The Hölder space $C^s([0,1])$ and Hölder ball $C^{s,R}([0,1])$ of radius $R > 0$ are defined as

$$C^s = \{f \in L_2[0,1] : \|f\|_{C^s} < \infty\}, \quad \text{and} \quad C^{s,R}([0,1]) = \{f \in L_2[0,1] : \|f\|_{C^s} \leqslant R\},$$

respectively.

Chapter 6

Statistical equivalence under communication constraints

> "*Oh, my distances are very impossible to calculate; you know that. But bounds are feasible. And for the Bayes risk, I know that just the metric structure does not catch everything, but I don't know what else to look at, except, as you said, calculations.*" - Lucien Le Cam

In this final chapter of the thesis, we explore the degree to which the results derived in this thesis extend beyond the many-normal-means model and the infinite dimensional signal-in-white-noise model studied in the earlier chapters. To do so, we shall leverage existing results concerning the comparison of models, called Le Cam theory. This allows us to obtain minimax rates for goodness-of-fit testing in other models, such as the multinomial model, nonparametric regression and nonparametric density testing.

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[1], 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. [137, 186]. For a brief introduction; [138, 151].

It has been a long-standing and persistent finding that models that describe seemingly very different data and dynamics, can still be subject to the same phenomena,

[1] Or their corresponding *statistical experiments*, see Section 6.1.

such as the asymptotic minimax risk coinciding as the number of samples grows. This finds mathematical substantiation using the Le Cam distance: if the Le Cam distance between models tends to zero as e.g. the size of the data grows, they are called *asymptotically equivalent*. For parametric models, asymptotic equivalence has been established for a huge variety of models, in particular models that are "locally asymptotically normal", see for instance [138]. Starting in 1996, asymptotic equivalence between the nonparametric signal-in-white-noise model studied in Chapter 5 to observing i.i.d. draws from a density function or a nonparametric regression model has been established by [40, 161]. Since then, asymptotic equivalence to the signal-in-white-noise model has been established for many models, such as nonparametric generalized linear models [109], nonparametric regression with non-identically distributed data [124], nonparametric regression with non-Gaussian errors [110], nonparametric regression with random design [42], nonparametric drift diffusion models [71, 104, 68, 69], the spectral density of a Gaussian process [107], densities with known discontinuities [152] and jump-diffusion models [150].

A phenomenon which is also of interest in this chapter, is that of asymptotic nonequivalence. 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. Whilst a Le Cam distance lower bound only indicates that in *some* loss functions concerning certain specific decision problems two models behave differently, these results are still of interest. Firstly, they give fundamental insight into specific statistical models. Secondly, they serve as a warning to tread carefully in such cases, by indicating that the usual statistical phenomena one might expect in the well studied simple model might not necessarily occur in the model of interest. Asymptotic nonequivalence has been studied for the signal-in-white-noise model, showing nonequivalence to e.g. nonparametric regression and i.i.d. draws from a density whenever the underlying space of functions is not of sufficient regularity [91, 41] or nonequivalence with i.i.d. sampling from densities when the class of densities are not sufficiently bounded from below [173]. In [217], nonequivalence is shown between the drift diffusion model and a stochastic volatility model.

There are many reasons to study the Le Cam distance of models, not the least of which scientific interest. The main concern in this chapter is the ability to obtain distributed inference performance bounds in complex models which are known to be asymptotically equivalent to the many-normal-means model and the infinite dimensional signal-in-white-noise model. This allows us to obtain distributed bandwidth and differential privacy constraint minimax testing rates for models for which these have, up until now, not been established.

Whilst classically minimax goodness-of-fit testing rates are perhaps more easily derived by studying the different models directly, this does not seem to be the case for the bandwidth and differential privacy constraint distributed equivalent of these testing problems. From the results of the earlier chapters of the thesis, it is clear that to obtain the minimax rates in simple, stylized Gaussian models already requires

substantial effort. By leveraging asymptotic equivalence in the distributed setting, we can establish minimax distributed testing rates for the d-dimensional multinomial model under bandwidth and local differential privacy constraints. Prior to this work, such rates had only been available in the literature for the case of having just one observation per machine ($n = 1$) in [9, 10, 15]. To obtain the rates for the multinomial model, we use the Le Cam deficiency bound between the Gaussian and multinomial model of [57]. The multinomial model shall be our main illustrative example to elucidate the role of the sample space in the distributed communication constraint setting. These results give insight into whether models that are asymptotically equivalent, have equivalent rates in the distributed setting as well. The answer here turns out to be partly yes, but not always.

Furthermore, we extend our results of Chapter 5 to nonparametric models more commonly encountered in practice, such as nonparametric regression and goodness-of-fit testing for nonparametric densities based on i.i.d. observations. The Le Cam distance bound of [172] allows us to establish bandwidth and local minimax distributed testing rates for nonparametric density testing, which have been established in the literature only for the case of just one observation per machine under privacy constraints in [75, 136], where in [136], the authors consider adaptation as well. The work of this chapter is also the first to establish minimax distributed testing rates for nonparametric regression under bandwidth and local differential privacy constraints. The latter results leverage the Le Cam distance bound of [174] between the signal-in-white-noise model and nonparametric regression. The results for both of these models apply to their respective adaptive settings.

Besides deriving distributed testing rates, we shall also study asymptotic nonequivalence. The route with which we shall study asymptotic nonequivalence is novel and perhaps surprising. By leveraging our result concerning asymptotic equivalence in the distributed setting, we exhibit a proof method for obtaining lower bounds on the Le Cam distance between models which, even though without communication constraints they behave similarly, display drastically different behavior when distributed communication constraints are in place. We illustrate this principle using the multinomial model and the many-normal-means-model. For the multinomial model and the many-normal-means model, although the unconstrained minimax testing rate is $\sqrt{d}/(mn)$ for both models, we exploit distributed settings in which these models have different minimax rates to obtain lower bounds on the Le Cam distance of the models that apply generally, that is to say; *these lower bounds apply outside of the distributed setting as well.*

The chapter is structured as follows. First, in Section 6.1, we recall the formal notions surrounding the Le Cam distance and prove results for general distributed settings with communication constraints. In Section 6.2, we study the consequences of the general theory for the multinomial model, obtaining minimax distributed testing rates for the multinomial model under bandwidth and privacy constraints, as well as a lower bound on the Le Cam distance between the many-normal-means model and the

multinomial model. Finally, in Section 6.3, we derive minimax distributed testing rates for nonparametric regression and density testing using the machinery developed in Section 6.1.

6.1 Le Cam theory in distributed setting

We introduce some formal notions of Le Cam theory first in Section 6.1.1. Then, in Section 6.1.2, we study the equivalence of models in the distributed setting.

6.1.1 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

$$\mathfrak{d}(\mathcal{P}; \mathcal{Q}) = \inf_{C} \sup_{f \in \mathcal{F}} \|P_f C - Q_f\|_{\mathrm{TV}}. \tag{6.1}$$

Here, we use the total variation norm as defined in earlier chapters (see e.g. (1.3)), 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

$$P_f C(A) := \int_{x \in \mathcal{X}} C(A|x) dP_f(x). \tag{6.2}$$

This is equivalent to the more general notion of deficiency of [53] for dominated models on Polish spaces (see Proposition 9.2 in [161]).

The deficiency $\mathfrak{d}(\mathcal{P}; \mathcal{Q})$ quantifies the degree to which $\mathcal{Q}$ can be approximated by an experiment $\mathcal{P}$. If $\mathfrak{d}(\mathcal{P}; \mathcal{Q}) \leqslant \varrho$, 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 $\mathfrak{d}(\mathcal{P}; \mathcal{Q}) \leqslant \varrho$, 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

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

Here, the Markov kernel $Q_f D$ is to be understood in the sense of (6.2) 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

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

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 6.7 in the appendix, since $x \mapsto \int \ell(f,\varphi) dD(\varphi|x)$ is measurable. In the more extensive framework considered in e.g. [53], 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

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

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 6.1. *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*

$$\frac{dP_f}{d\mu}(x) = g_f(S(x))h(x) \text{ for almost every } x \in \mathcal{X} \text{ and every } f \in \mathcal{F}. \tag{6.3}$$

A proof for both the lemma and the last statement of the previous paragraph can be found in Chapter 5 of [137].

6.1.2 Equivalence of distributed decision problems

We now turn to the distributed setting considered in this thesis, where $j = 1, \dots, m$ machines each receive data $X^{(j)}$ drawn from a distribution P_f and sample space $(\mathcal{X}, \mathscr{X})$. Each of the machines 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.

We start by stating the distributed setting as given in Section 1.2 in the current context. A *distributed protocol for the experiment $\mathcal{P}$ with decision space $(\mathcal{D}, \mathscr{D})$* consists of a triplet $\{D, \{K^j\}_{j=1,\dots,m}, (\mathcal{U}, \mathscr{U}, \mathbb{P}^U)\}$, where $\{K^j\}_{j=1,\dots,m}$ is a collection of Markov kernels $K^j : \mathscr{Y}^{(j)} \times (\mathcal{X} \times \mathcal{U}) \to [0,1]$ defined on a measurable space $(\mathcal{Y}^{(j)}, \mathscr{Y}^{(j)})$, a Markov kernel $D : \mathscr{D} \times \bigotimes_{j=1}^m \mathcal{Y}^{(j)} \to [0,1]$ and a probability space $(\mathcal{U}, \mathscr{U}, \mathbb{P}^U)$.

To unpack all this notation: the Markov kernel D takes the role of the decision procedure, where the decision is to be made on the basis of the transcripts generated by $\{K^j\}_{j=1,\dots,m}$. 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 machines. 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 terms of random variables, we have $X^{(j)} \sim P_f$, $U \sim \mathbb{P}^U$, $Y^{(j)}|(X^{(j)}, U) \sim K^j(\cdot|X^{(j)}, U)$ for $j = 1, \dots, m$ and $\varphi \sim D(\cdot|Y)$ with $Y = (Y^{(1)}, \dots, Y^{(m)})$. This gives rise to a Markov chain

$$\begin{array}{ccccc} X^{(1)} & \longrightarrow & Y^{(1)}|U & \searrow & \\ \vdots & \longrightarrow & \vdots & \longrightarrow & \varphi. \\ X^{(m)} & \longrightarrow & Y^{(m)}|U & \nearrow & \end{array} \tag{6.4}$$

For $x = (x^{(1)}, \dots, x^{(m)}) \in \mathcal{X}^m$, $u \in \mathcal{U}$ and $\{K^j\}_{j=1,\dots,m}$, let $x \mapsto K(A|x,u)$ be the Markov kernel product distribution $\bigotimes_{j=1}^m K^j(\cdot|x^{(j)}, u)$. Following the notation in the earlier chapters, given a distributed protocol and i.i.d. data from P_f we shall use $\mathbb{P}_f$ to denote the joint distribution of the data $X \sim P_f^m$, the shared randomness $U \sim \mathbb{P}^U$ and $Y = (Y^{(1)}, \dots, Y^{(m)})$ 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

$$\mathbb{P}_f^Y(A) = P_f^m \mathbb{P}^U K(A) = \mathbb{P}^U P_f^m K(A) = \int d \bigotimes_{j=1}^m P_f K^j(\cdot|X^{(j)}, u)(A) d\mathbb{P}^U(u), \tag{6.5}$$

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

We shall consider two types of communication constraints in this chapter: bandwidth constraints and differential privacy constraints. For the first of these constraints, the definition of a bandwidth constraint protocol is straightforward and fully overlaps with the one considered in the rest of the thesis (i.e. Definition 2). A distributed protocol is said to satisfy a *b-bit bandwidth constraint* if its kernels $\{K^j\}_{j=1,\dots,m}$ are defined on spaces satisfying $|\mathcal{Y}^{(j)}| \leqslant 2^b$.

Given a Markov Kernel $C : \mathscr{X} \times \tilde{\mathcal{X}} \to [0,1]$, a distributed protocol

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

for the model $\mathcal{P}$, yields a distributed protocol for the model $\mathcal{Q}$:

$$\{D, \{CK^j\}_{j=1,\dots,m}, (\mathcal{U}, \mathscr{U}, \mathbb{P}^U)\}.$$

If $\{K^j\}_{j=1,\dots,m}$ is a b-bit bandwidth constraint, the collection of kernels $\{CK^j\}_{j=1,\dots,m}$ do so too, as each CK^j is defined on $\mathscr{Y}^{(j)} \times \tilde{\mathcal{X}}$.

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

$$K(A|x) \leqslant e^{\epsilon} K(A|x') + \delta \quad \text{for all } A \in \mathscr{Y} \text{ and } x, x' \in \mathcal{X}. \tag{6.6}$$

A distributed protocol shall be called locally (ϵ, δ)-differentially private if (6.6) holds for each K^j; $j = 1, \dots, m$. The difference with Definition 3 is that we essentially wish to retain privacy for the entire "local sample", instead of for each observation in the sample. Llocal differential privacy is a more demanding notion of differential privacy than what was considered in earlier chapters and it is less general, as it cannot accomodate for the fact that the datums in a server belong to e.g. different individuals. The restriction to local differential privacy arises naturally, due to the fact that sample spaces (and thus datums) can differ between different experiments. The following lemma shows that local (ϵ, δ)-differential privacy, just like bandwidth constraints, carry over from one model to the other.

Lemma 6.2. *Let $(\mathcal{X}, \mathscr{X})$ and $(\tilde{\mathcal{X}}, \tilde{\mathscr{X}})$ be measurable spaces and consider Markov kernels $C : \mathscr{X} \times \tilde{\mathcal{X}} \to [0,1]$ and $K : \mathscr{Y} \times \mathcal{X} \to [0,1]$. If K is b-bit bandwidth constraint, so is the Markov kernel $CK : \mathscr{Y} \times \tilde{\mathcal{X}} \to [0,1]$. If K is locally (ϵ, δ)-differentially private, so is CK.*

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

and applying (6.6) to K yields

$$\begin{aligned} CK(A|\tilde{x}) &= \int K(A|x)dC(x|\tilde{x}) = \int\int K(A|x)dC(x|\tilde{x})dC(x'|\tilde{x}') \\ &\leqslant e^{\epsilon}\int K(A|x')dC(x'|\tilde{x}') + \delta = e^{\epsilon}CK(A|\tilde{x}') + \delta, \end{aligned}$$

which shows CK is (ϵ,δ)-differentially private. □

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

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

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 6.3, shall serve as one of the main tools for deriving the main results of this chapter. 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 machines, 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 6.3. *Let $m\in\mathbb{N}$. Consider two experiments $\mathcal{P}$ and $\mathcal{Q}$ with indexing set $\mathcal{F}$, satisfying $m\mathfrak{d}(\mathcal{Q};\mathcal{P}) \leqslant \varrho$ 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 $\ell : \mathcal{F}\times\mathcal{D} \to [0,1]$,

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

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

The same statement holds for $\mathscr{J}_{\mathcal{P}}$ and $\mathscr{J}_{\mathcal{Q}}$ denoting the classes of b-bit bandwidth constraint local randomness protocols, distributed protocols satisfying (shared or local randomness) local (ϵ,δ)-differential privacy constraints.

Remark 12. The result, might seem rudimentary as not much more than the triangle inequality seems to be going into the proof. However, the statement is sufficient to derive minimax rates for distributed goodness-of-fit testing in models other than the many-normal-means model and the signal-in-white-noise model considered in the previous chapters. What is more, in Section 6.2, the lemma is leveraged to obtain lower bounds on the deficiency between two models whenever two models have (substantially) different distributed risks for the same decision problem under communication constraints. 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 [137], $m\mathfrak{d}(\mathcal{Q};\mathcal{P}) \leqslant \varrho$ implies that there exists a kernel $C : \mathscr{X} \times \tilde{\mathcal{X}} \to [0,1]$ such that

$$\sup_{f\in\mathcal{F}} \|P_f - Q_f C\|_{\mathrm{TV}} \leqslant \varrho/m. \tag{6.7}$$

By Lemma 6.2, the kernels $\tilde{K}^j := CK^j$, $j = 1,\dots,m$ all satisfy a b-bit bandwidth constraint or local (ϵ,δ)-differential privacy constraint if the collection $\{K^j\}_{j=1,\dots,m}$ does. That is, given a 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 an element of $\mathscr{J}_{\mathcal{Q}}$.

Using the fact that ℓ is bounded by one and Lemma 6.8 in the appendix, it follows that

$$\begin{aligned}\mathcal{R}_{\mathcal{Q}}(\tilde{D},\ell) - \mathcal{R}_{\mathcal{P}}(D,\ell) &\leqslant \|\mathbb{P}^U \bigotimes_{j=1}^m P_f K^j - \mathbb{P}^U \bigotimes_{j=1}^m Q_f C K^j\|_{\mathrm{TV}} \\ &\leqslant \sum_{j=1}^m \|\mathbb{P}^U P_f K^j - \mathbb{P}^U Q_f C K^j\|_{\mathrm{TV}}.\end{aligned}$$

By Lemma 6.9 in the appendix,

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

which combined with (6.7) finishes the proof. □

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 testing a simple null versus a composite alternative hypothesis in the distributed setting. To that extent, consider a test of the hypotheses

$$H_0 : f = f_0 \ \text{ versus the alternative hypothesis } \ f \in H_1 \tag{6.8}$$

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 (6.5), define the distributed testing risk for the hypotheses in (6.8) and the model $\mathcal{P}$ as

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

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 $\mathscr{T}_{\mathrm{SR}}^b(\mathcal{P})$ (resp. $\mathscr{T}_{\mathrm{LR}}^b(\mathcal{P})$) denote the set of shared randomness (resp. local randomness) distributed testing protocols for $\mathcal{P}$ satisfying a b-bit bandwidth constraint. Similarly, let $\mathscr{T}_{\mathrm{SR}}^{(\epsilon,\delta)}(\mathcal{P})$ (resp. $\mathscr{T}_{\mathrm{LR}}^{(\epsilon,\delta)}(\mathcal{P})$) 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 6.3, we obtain the following result.

Corollary 6.1. *Consider experiments $\mathcal{P}, \mathcal{Q}$ such that $m\mathfrak{d}(\mathcal{Q}; \mathcal{P}) \leqslant \varrho$ for $\varrho > 0$. It holds that*

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

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 6.3 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

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

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 result above yields that for experiments with matching indexing sets, matching hypotheses and that are close in Le Cam distance, the minimax separation rates (see Section 1.1 for a definition) for the hypotheses is the same in distributed settings, as long as m is not too large compared to the Le Cam distance between the models. We remark that a similar result can also be obtained for the meta-analysis "combination of real-valued test-statistics" framework considered in Chapter 4.

The implications of Lemma 6.3 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

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

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

In the next sections, we will explore the consequences of Corollary 6.1 for minimax distributed testing rates for both bandwidth- and privacy constraints.

6.2 Distributed multinomial observations under communication constraints

The multinomial distribution describes n discrete random variables that take one of d mutually exclusive states, applying to any setting in which you sample independently from a probability distribution on a discrete set.

Recently, there have been numerous applications in areas that handle large samples of multinomial data over extensive domains, such as population genetics [166, 196] and computer science; where it is used for e.g. information retrieval [228, 177], speech and text and classification [126], text mining [49] and large language models [168].

This has sparked recent interest in studying the statistical decision theoretic properties of the multinomial model, see [27] for an overview. It has been extensively studied in the context distributed inference under differential privacy- and bandwidth constraints, see e.g. [101, 78, 30, 113, 58, 59, 17, 16, 13]. For distributing testing under privacy- and bandwidth constraints specifically, much is still to be uncovered, with minimax rates only having been obtained for the case of having just one draw from a discrete distribution per machine [9, 10, 15] at the time of writing. For some investigations into the multiple observations case, see [73, 99].

The multinomial model describes sampling independently from a probability distribution on a discrete set. We start by giving a formal description of the model in the distributed setting. Let $\mathbb{S}^d$ denote the $d-1$-dimensional probability simplex

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

In the distributed multinomial model, each machine $j = 1, \ldots, m$ observes data $\tilde{X}^{(j)}$ taking values in $\{1, \ldots, d\}^n$

$$\tilde{X}^{(j)} = (\tilde{X}_1^{(j)}, \ldots, \tilde{X}_n^{(j)}) \sim Q \equiv Q_{n,q}, \quad X_i^{(j)} \overset{i.i.d.}{\sim} \text{Multinomial}(1, q) \ \text{ for } q \in \mathcal{F} \quad (6.9)$$

where

$$\mathcal{F} = \left\{ q \in \mathbb{S}^d \,:\, \frac{\max_i q_i}{\min_i q_i} \leqslant R \right\} \tag{6.10}$$

for some fixed constant $R > 4$. The statistical decision problem of interest shall be that of uniformity testing, i.e. distinguishing the hypotheses

$$H_0 : q = q_0 \;\text{ versus }\; H_1 : q \in \{q \in \mathcal{F} : \|q - q_0\|_1 \geqslant \rho\} =: H_\rho, \tag{6.11}$$

with $q_0 = (q_{01}, \ldots, q_{0d}) = (1/d, \ldots, 1/d) \in \mathbb{S}^d$. We note that the results can easily be extended to the case where $q_0 = (q_{01}, \ldots, q_{0d}) \in \mathcal{F}$, with e.g.

$$\frac{\max_i q_{0i}}{\min_i q_{0i}} \leqslant R/4.$$

The latter assumption allows for (slightly) more flexibility than uniformity. However, some sort of uniformity assumption is critical, as the minimax rates for the multinomial model depend on the "degree of uniformity" of null hypothesis (see [28]). Outside of the (approximately) uniform case, goodness-of-fit testing exhibits different phenomena compared to the many-normal-means model.

The minimax rate for the hypothesis above in the $m = 1$ case is $\rho^2 \asymp \frac{\sqrt{d}}{n}$, as was established in [163] and [206]. For distributed data, minimax rates have been derived in [9, 10] for the hypothesis test above under both privacy- and bandwidth constraints for the case where $n = 1$. This case corresponds with receiving only one observation per machine. For their lower bounds, the authors use clever series expansions of a combinatorial nature that are difficult to generalize to the large n case.

Theorems 6.1 and 6.3 below will partly extend these results by giving minimax rates in a regime where n is large. We will do so by comparing the statistical experiment of the multinomial observations to that of the Gaussian many-normal-means model considered in Chapters 2 and 3 of the thesis.

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

$$X_i^{(j)} = \sqrt{q_i} + \frac{1}{\sqrt{2n}} Z_i^{(j)} \tag{6.12}$$

$Z^{(j)} = (Z_1^{(j)}, \ldots, Z_d^{(j)}) \sim N(0, I_d)$. 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 [57] 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 [57] that

$$\Delta(\mathcal{P}, \mathcal{Q}) \leqslant C_R \frac{d \log d}{\sqrt{n}}, \tag{6.13}$$

where $C_R > 0$ is a constant depending only on R. Combining this with Corollary 6.1 and the lower bound results from Chapter 2, we obtain the following result.

Theorem 6.1. *For any sequences $m \equiv m_\nu$, $b \equiv b_\nu$, $d \equiv d_\nu$ and $n \equiv n_\nu$ such that $md \to \infty$ whilst*

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

the minimax separation rate in the distributed multinomial model $\mathcal{Q}$ for testing the hypotheses (6.11) *under a b-bandwidth constraint is given by*

$$\rho^2 \asymp \left(\frac{d}{\sqrt{d \wedge bmn}}\right) \bigwedge \left(\frac{\sqrt{d}}{\sqrt{mn}}\right) \tag{6.14}$$

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

$$\rho^2 \asymp \left(\frac{d^{3/2}}{(d \wedge b)mn}\right) \bigwedge \left(\frac{\sqrt{d}}{\sqrt{mn}}\right). \tag{6.15}$$

Remark 13. The rates obtained for the L_1-norm separated alternatives in the multinomial model, can be seen to correspond to L_2-separated alternative hypothesis rates in the many-normal-means model. While this might seem odd up on first reading, these are natural 'equivalent hypotheses' to consider. To see this, consider that the total variation distance in the many-normal-means model ($n = 1$) satisfies (see e.g. [72])

$$\|P_{\sqrt{q_1}} - P_{\sqrt{q_2}}\|_{\mathrm{TV}} \asymp \|\sqrt{q_1} - \sqrt{q_2}\|_2.$$

On the other hand, we have that $\|q_1 - q_2\|_{\mathrm{TV}} = \|q_1 - q_2\|_1/2$ (see e.g. Lemma 6.6). An explicit comparison of these hypotheses shows up in the proof of the theorem.

We discuss a similar result for distributed testing under local differential privacy constraints later on in the section, in the form Theorem 6.3. We provide a proof for both the aforementioned theorem and the one above at the very end of the section, but before doing so, let us consider the ramifications of Theorem 6.1.

The distributed b-bit bandwidth constraint minimax rate for the hypotheses (6.11) in the multinomial model with $n = 1$ is established in [9, 10]. Specifically, they find that

$$\rho^2 \asymp \begin{cases} \frac{d}{m\sqrt{2^b \wedge d}} & \text{in case of access to shared randomness,} \\ \frac{d\sqrt{d}}{m(2^b \wedge d)} & \text{without access to shared randomness.} \end{cases} \tag{6.16}$$

Several aspects of this minimax rate are intriguing. Firstly, there is no elbow effect, as is observed in the "large n case" for the same model and hypothesis (see (6.14) and (6.15)). Secondly, the benefit (i.e. efficiency gain) from an increase in bandwidth is exponential, compared to the polynomial factor observed in the Gaussian model. We shall delve into this "communication super-efficiency" phenomenon further below.

The multinomial model considered thus far, where we draw n independent and indentically distributed draws taking values in $\{1, \ldots, d\}$, is equivalent to the Multinomial(n, d)

model in which one observes $N^{(j)} = (N_1^{(j)}, \dots, N_d^{(j)})$ taking values in $\{1, \dots, n\}^d$, where

$$N_k^{(j)} \equiv N_k^{(j)}\left(\tilde{X}^{(j)}\right) = \left|\left\{i : (\tilde{X}^{(j)})_i = k\right\}\right|. \tag{6.17}$$

Let $\mathcal{Q}'$ denote the model generated by the observations $N^{(j)}$ in (6.17). It easily follows from Neyman-Fisher factorization (e.g. Lemma 6.1), that model is equivalent to $\mathcal{Q}$, meaning $\Delta(\mathcal{Q}, \mathcal{Q}') = 0$. 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 $\{1, \dots, n\}^d$, the full data can be transmitted whenever there are at least $d \log_2 n$-bits are available per machine. However, recalling that the observation $\tilde{X}^{(j)}$ takes values in the space $\{1, \dots, d\}^n$, the cardinality which is 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 (6.11) 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 2.3. This indicates a kind of "tipping point" occurring 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. More generally, it means that, even though models $\mathcal{P}^m = \{P_f^m : f \in \mathcal{F}\}$ and $\mathcal{Q}^m = \{Q_f^m : f \in \mathcal{F}\}$ could be 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.

Interestingly, this does not imply that the multinomial model is "easier" from a distributed testing under bandwidth constraints perspective, as there are 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, combined with Corollary 6.1, 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 6.2. *There exists constants $C, c > 0$ such that for any $n, d \in \mathbb{N}$ with*

$$\frac{d}{n \log(d)} \geqslant C \quad \text{and} \quad n \geqslant \sqrt{d} \log(d) \tag{6.18}$$

it holds that

$$\mathfrak{d}(\mathcal{Q}, \mathcal{P}) \geqslant c, \tag{6.19}$$

where $\mathcal{P}$ is the experiment generated by the observations in (6.12)*, $\mathcal{Q}$ is generated according to* (6.9)*, both indexed by $\mathcal{F}$ as given in* (6.10)*.*

The conclusion of the theorem, that for such large d compared to n, the multinomial model is asymptotically nonequivalent to the Gaussian model, is unsuprising for uniform-like distributions q. The proof of the theorem, however, uses the distributed bandwidth constraint results derived earlier in perhaps an interesting way: 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 by Corollary 6.1. The fact that the separation in the respective (distributed) testing risks occurs for a constant number of machines, yields that the two models are asymptotically nonequivalent whenever (6.18) holds. 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 (6.11) would have similar minimax performance for both of the models. Whilst it is unlikely that the condition $d/n \log(d) \gtrsim 1$ is "tight" for the above non-equivalence result, the proof technique used in the theorem could be of interest for other settings where non-equivalence is suspected due to differences in the models' sample spaces.

We now turn to the case of local differential privacy constraints. Before stating the theorem, let us avoid possible confusion by stressing that the privacy constraints considered in the earlier chapters of the thesis study a setting more general than local differential privacy. Here, we shall require that a distributed protocol transcript generating Markov kernels $\{K^j\}_{j\in[m]}$ satisfy (6.6). We stress that this type of differential privacy guarantee concerns the local data $\tilde{X}^{(j)}$ as the "unit of privacy" or as an "individual", even $\tilde{X}^{(j)}$ consists of multiple (i.e. n) observations.

For the Gaussian models studied in Chapters 2, 3 and 5, local differential privacy is a special case corresponding to "$n = 1$" in the results found in these chapters. This translates to guaranteeing differential privacy for a single observation

$$X^{(j)} = \sqrt{q} + \frac{1}{\sqrt{2n}} Z^{(j)} \tag{6.20}$$

per machine $j = 1, \ldots, m$ in the Gaussian model considered in this section. Whilst the above mode is equivalent to observing $i = 1, \ldots, 2n$ i.i.d. $\sqrt{q} + Z_i^{(j)}$ observations classically, under privacy constraints there is a pronounced difference.

The rates under local differential privacy for the Gaussian signal detection problem follow from Theorems 2.4 and 3.2 by considering "$n = 1$" in the setting of those chapters and considering $\sqrt{n}$-rescaled signal in the (single) observation received at each machine: $X^{(j)} = \sqrt{n} f + Z^{(j)}$. The hypotheses considered are $H_0 : f = 0$ versus the alternative

$$f \in H_\rho := \left\{ f \in \mathbb{R}^d \,:\, \|\sqrt{n} f\|_2 \geqslant \rho',\ \|\sqrt{n} f\|_2 \leqslant \sqrt{n} M \right\},$$

where ρ' is the minimax rate for the single observation model under differential privacy, and setting $\rho = \rho'/\sqrt{n}$ we obtain the minimax rate corresponding to (6.20). For details, we defer the reader [52]. In this "rescaled" version of the problem, the minimax rate for $(mn)^{-1} \leqslant \epsilon \leqslant 1$ and $\log(1/\delta) \lesssim \log(mnd)$ is given by

$$\rho^2 \asymp \text{poly-log}(d, m, n, 1/\delta) \begin{cases} \frac{d}{mn\epsilon^2} & \text{if } \epsilon \geqslant \frac{\sqrt{d}}{\sqrt{m}}, \\ \frac{\sqrt{d}}{\sqrt{mn}\epsilon} & \text{if } \frac{1}{\sqrt{md}} \leqslant \epsilon < \frac{\sqrt{d}}{\sqrt{m}}, \\ \frac{1}{mn\epsilon^2} & \text{if } \epsilon < \frac{1}{\sqrt{md}}, \end{cases} \tag{6.21}$$

in case of locally (ϵ, δ)-differentially private shared randomness protocols and

$$\rho^2 \asymp \text{poly-log}(d, m, n, 1/\delta) \begin{cases} \frac{d\sqrt{d}}{mn\epsilon^2} & \text{if } \epsilon \geqslant \frac{d}{\sqrt{m}}, \\ \frac{\sqrt{d}}{\sqrt{mn}\epsilon} & \text{if } \frac{1}{\sqrt{md}} \leqslant \epsilon < \frac{d}{\sqrt{m}}, \\ \frac{1}{mn\epsilon^2} & \text{if } \epsilon < \frac{1}{\sqrt{md}}, \end{cases} \tag{6.22}$$

in case of non-shared randomness protocols. In the above, the poly-$\log(d, m, n)$ factor should be understood as a factor at most of poly-logarithmic rate in d, m and n. Because only the local case is considered, a lot of the phase transitions observed in the earlier chapters are not observed (the ones that occur when $\epsilon \geqslant 1/\sqrt{n}$). Leveraging asymptotic equivalence between the two models, we obtain the following theorem.

Theorem 6.3. *Consider sequences $m \equiv m_\nu$, $d \equiv d_\nu$ and $n \equiv n_\nu$ such that $md \to \infty$,*

$$\frac{md \log d}{\sqrt{n}} \overset{\nu \to \infty}{\to} 0,$$

$n^{-1/4} \ll \epsilon \equiv \epsilon_\nu \leqslant 1$ and $\delta \equiv \delta_\nu \lesssim (md)^{-p}$ for some $p \geqslant 2$. The minimax separation rate in the distributed multinomial model $\mathcal{Q}$ for testing the hypotheses (6.11) *using locally (ϵ, δ)-differentially private protocols is* (6.21) *in the case of having access to shared randomness. In the case of having only access to shared randomness, it is given by* (6.22).

Remark 14. Also in the case of privacy, there is a difference between the one observation per machine case minimax rate ($n = 1$) and the multiple observations per machine with local differential privacy case. The minimax rate in the multinomial for $n = 1$ is worked out to be

$$\rho^2 \asymp \begin{cases} \frac{d}{m\epsilon^2} & \text{in case of access to shared randomness,} \\ \frac{d\sqrt{d}}{m\epsilon^2} & \text{without access to shared randomness.} \end{cases} \tag{6.23}$$

(see [9, 15]). Comparing this rate to the rate obtained in Theorem 6.3, we observe phase transitions in the distributed testing problem for multinomial model under local differential privacy constraints which are not observed if the number of observations locally is small compared to the cardinality of the sample space.

6.2.1 Proofs of Theorems 6.1, 6.3 and 6.2

Proof of Theorems 6.1 and 6.3. 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 Corollary 6.1 and the bound (6.13) that the testing risks satisfy

$$\inf_{T \in \mathscr{T}_\mathcal{Q}} \mathcal{R}_{\mathcal{Q}_\nu}(H_{\rho_\nu}, T) = \inf_{T \in \mathscr{T}_\mathcal{P}} \mathcal{R}_{\mathcal{P}_\nu}(H_{\rho_\nu}, T) + o(1). \tag{6.24}$$

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 (6.14), (6.15), (6.21) or (6.22). 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 (6.12) is equivalent in the Le Cam sense the one generated by

$$X^{(j)} = \sqrt{q} - \sqrt{q_0} + \frac{1}{\sqrt{2n}} Z^{(j)} \text{ with } Z^{(j)} \sim N(0, I_d), \tag{6.25}$$

for $q \in \mathcal{F}$, which we shall denote by $\tilde{\mathcal{P}}$. Consequently, by another application of Corollary 6.1, it suffices to show

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

If $\|q - q_0\|_1 \geqslant \rho$, Lemma 6.10 implies that $\|\sqrt{q} - \sqrt{q_0}\|_2 \geqslant \rho/2$. 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 $H_\rho : \|\sqrt{q} - \sqrt{q_0}\|_2 \geqslant \rho/2$. By the established equivalence of the minimax risks (6.24), 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 2.12 and 2.17, 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 $i = 1,\dots,d/2$ and $q_i^f = 1/d - f_{i-d/2}/\sqrt{d}$ for $i = d/2+1,\dots,d$. To see that $q^f \in \mathcal{F}$, note that $\sum_{i=1}^d q_i^f = 1$, $q^f \geqslant 0$ and

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

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

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

and

$$H_\rho' := \left\{q : q\in\mathcal{F}', \|q-q_0\|_1 \geqslant \rho\right\}.$$

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

$$\inf_{T\in\mathscr{T}_\mathcal{P}} \mathcal{R}_\mathcal{P}(T,H_\rho) \geqslant \inf_{T\in\mathscr{T}_\mathcal{P}} \mathcal{R}_\mathcal{P}(T,H_\rho'). \tag{6.26}$$

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

$$\begin{aligned}\left\|P^{nm}_{\sqrt{q^f}-\sqrt{q_0}} - P^{nm}_{g_f}\right\|_{\mathrm{TV}} &\leqslant 1\wedge\sqrt{\frac{mn}{4}D_{\mathrm{KL}}(P_{\sqrt{q}-\sqrt{q_0}};P_{g_f})}\\ &= 1\wedge\frac{\sqrt{mn}}{2}\left\|\sqrt{q_0+2g_f/\sqrt{d}}-\sqrt{q_0}-g_f\right\|_2 =: D_f,\end{aligned}$$

where $P^n_{\sqrt{q}-\sqrt{q_0}}$ denotes the distribution of (6.25) and the square root is to be understood as applied coordinate wise.

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

We have that

$$\begin{aligned}\inf_{T\in\mathscr{T}_\mathcal{P}} \mathcal{R}_\mathcal{P}(T,H_\rho') \geqslant \inf_{T\in\mathscr{J}_\mathcal{P}} &\left[\mathbb{P}_0 T(Y) + \int \mathbb{P}_{g_f}(1-T(Y))d\pi(f)\right] - 2\int D_f d\pi(f)\\ &- \pi\left(f : f\notin (I_R/\sqrt{d})^{d/2} \text{ or } \left\|(q_i^f)_{i\in[d]}-q_0\right\|_1 < \rho\right).\end{aligned}$$

By Lemma 6.4, the model $\{P_{g_f} : f\in I_R^{d/2}/\sqrt{d}\}$ is equivalent to the model generated by the observations

$$S_i^{(j)} := f_i + \frac{1}{\sqrt{n}}Z_i^{(j)} \tag{6.27}$$

for $i = 1, \dots, d/2$. Since $\mathcal{F}'$ is bijective with $(I_R/\sqrt{d})^{d/2}$, the aforementioned equivalence and Corollary 6.1 implies that

$$\inf_{T \in \mathscr{J}_{\mathcal{P}}} \left[\mathbb{P}_0 T(Y) + \int \mathbb{P}_{g_f}(1 - T(Y)) d\pi(f) \right] = \inf_{T \in \mathscr{J}_{\tilde{\mathcal{P}}}} \left[\mathbb{P}'_0 T(Y) + \int \mathbb{P}'_f (1 - T(Y)) d\pi(f) \right] \tag{6.28}$$

where $\tilde{\mathcal{P}}$ is the model generated by the observations in display (6.27) for $i = 1, \dots, d/2$ and $\mathbb{P}'_f$ denotes the distribution of the distributed protocol with data generated from $f \in \tilde{\mathcal{P}}$.

It follows from Lemma 2.12 in the case of bandwidth constraints or Lemma 2.17 in the case of privacy constraints (using that $\rho \ll \rho^*$ in both cases) that the latter distributed testing risk is lower bounded by

$$1 - o(1) - \pi\left(f \in \mathbb{R}^{d/2} \,:\, f \notin (I_R/\sqrt{d})^{d/2} \text{ or } \left\| (q_i^f)_{i \in [d]} - q_0 \right\|_1 < \rho \right) - 2 \int D_f d\pi(f). \tag{6.29}$$

Addressing the third term in the display above; the theorem(s) assume that $md \log d/\sqrt{n}$ tends to zero as $\nu \to \infty$, $b \geqslant 1$ and $\epsilon \gg n^{-1/4}$, we have that $\rho^* \ll 1/\sqrt{\log(d)}$, which implies that

$$f_i \in I_R/\sqrt{d} \quad \text{ for all } i = 1, \dots, d/2, \tag{6.30}$$

as $\|\sqrt{d} f_i\|_\infty \to 0$ with π-probability tending to one (see e.g. Lemma 3.27).

Next, we show that $\|(q_i^f)_{i \in [d]} - q_0\|_1 \geqslant \rho$ 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 $c, c' > 0$,

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

where in the expression on the right-hand side, $Z \sim N(0, I_{d/2})$. Since $\rho \ll \rho^*$ and $\overline{\Gamma}$ is idempotent with rank of the order d, we can conclude that the expression vanishes. This takes care of the third term in (6.29).

For the last term in (6.29), 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

$$\left| \frac{1}{\sqrt{d}} \left(\sqrt{1 + \sqrt{d} f_i} - 1 - f_i/2 \right) \right| \leqslant \frac{\sqrt{d} f_i^2}{4}$$

on a set of π-probability tending to one. This yields that

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

Since the theorem(s) assume that $md \log d/\sqrt{n} \overset{\nu \to \infty}{\to} 0$, $b \geqslant 1$ and $\epsilon \gg n^{-1/4}$, the right-hand side of the above display vanishes when ρ^* satisfies either of the bounds (6.14), (6.15), (6.21) or (6.22).

□

Proof of Theorem 6.2. Let $\mathscr{T}_{\mathcal{Q}}, \mathscr{T}_{\mathcal{P}}$ 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 $\mathcal{Q}$ and $\mathcal{P}$, 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 (6.18), 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 (6.11), with $H_0 : q_0 = (1/d, \dots, 1/d) \in \mathbb{S}^d$ and H_ρ as in the display.

Set $b = \lceil n \log_2(d) \rceil$. When $b \geqslant n \log_2(d)$, the observations $\tilde{X}^{(j)}$ in the multinomial model as given in (6.9) are valid b-bit transcripts, since $|\{1, \dots, d\}^n| \leqslant n \log_2(d)$. These transcripts are therefore sufficient for the nondistributed / unconstrained model $\mathcal{Q}^m$, i.e. corresponding to observations

$$\tilde{X} = (\tilde{X}^{(1)}, \dots, \tilde{X}^{(m)}) \sim Q_{q,nm} \text{ for } q \in \mathcal{F}.$$

Consequently, the distributed, b-bit bandwidth constraint testing risk for $\mathcal{Q}$ is equal to the testing risk $\mathcal{Q}^m$;

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

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 $\mathcal{Q}$ such that

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

where H_ρ as defined in (6.11), as the minimax rate for the unconstrained problem with mn observations is $\rho^2_{\mathcal{Q}^m} := \sqrt{d}/(mn)$ (see e.g. Theorem 3 in [163]).

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

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

is bounded from below by $\rho^2_{\mathcal{P}} \asymp \sqrt{d}/(\sqrt{mn})$, as a consequence of Theorem 2.3. Specifically, following the proof of Theorem 6.1 above, we have that

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

where

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

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

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

for a symmetric, idempotent matrix $\overline{\Gamma} \in \mathbb{R}^{d/2 \times d/2}$ with $d/4 \leqslant \operatorname{rank}(\overline{\Gamma}) \leqslant d/2$ $\alpha \in (0,1)$, whenever $\rho^2 \leqslant c_\alpha \frac{\sqrt{d}}{\sqrt{m}n}$ for some small enough constant $c_\alpha > 0$. By the same analysis as conducted in the proof of Theorem 6.1 above (using that $n \leqslant \log(d)$), 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 $\rho^2 \leqslant c\sqrt{d}/(\sqrt{m}n)$ and m, n, b, d such that $mb \leqslant d$, where the number $1/3$ is chosen without particular significance.

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

$$\inf_{T \in \mathscr{T}_{\mathcal{Q}}} \mathcal{R}_{\mathcal{Q}}(H_\rho, T) < 1/6 < 1/3 < \inf_{T \in \mathscr{T}_{\mathcal{P}}} \mathcal{R}_{\mathcal{P}}(H_\rho, T). \tag{6.31}$$

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

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

Now suppose that $C\mathfrak{d}(\mathcal{Q}, \mathcal{P}) \leqslant 1/6$. Corollary 6.1 then implies in that

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

This contradicts (6.31). We conclude that

$$C\mathfrak{d}(\mathcal{Q}, \mathcal{P}) > 1/6, \tag{6.32}$$

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

6.3 Distributed testing rates for nonparametric models

In this section, we revisit the results for goodness-of-fit testing in the nonparametric signal-in-white-noise model under communication constraints studied in Chapter 5 and exhibit how these results extend to goodness-of-fit testing in other nonparametric models.

Specifically, we revisit the distributed setting in which $j = 1, \dots, m$ machines each observe

$$dX_t^{(j)} = f(t)dt + \frac{1}{\sqrt{n}} dW_t \tag{6.33}$$

where $f \in L_2[0,1]$. We shall denote the experiment generated by the observed sample path $X^{(j)}$, indexed by $\mathcal{F} \subset \mathcal{H}^{s,R}[0,1]$, for $s > 0$ and $R > 0$ as $\mathcal{P}_{s,R}$.

We discuss the extension of the distributed testing rates for goodness-of-fit testing with two other nonparametric models, namely nonparametric regression (Section 6.3.1) and nonparametric densities (Section 6.3.2), for both the adaptive and nonadaptive settings.

Before deriving results for the other models, we briefly recall the results derived in Chapter 5 as they apply to our setting here.

In the case of bandwidth constraints, a tight minimax rate for the model (6.33) is derived in Theorem 5.1 when the smoothness of the underlying alternative is known. Theorems 5.2 and 5.3 provide tight rates (up to a log-log(mn) factor) whenever s is in a given range $[s_{\min}, s_{\max}]$.

For privacy constraints, the theory derived in this section concerns *local* differential privacy constraints for nonparametric regression and nonparametric density testing. The appropriate comparison in terms of the minimax rates for these respective testing problems under local differential privacy constraints is to a distributed testing problem corresponding to (6.33), where the "individual" for which the privacy guarantee is to be satisfied is $X^{(j)}$. We stress that "n" plays a different role here than what is considered in Chapter 5, where n is the number of "individuals for which privacy is to be guaranteed", per machine. In (6.33), n takes the role of the "noise level", but bares no relationship to the privacy guarantee.

The minimax rates for the testing problem under local DP of $H_0 : f = 0$ versus $f \in \mathcal{H}^{s,R}[0,1]$ with $\|f\|_2 \geqslant \rho$, $s > 1/2$, for data generated according to (6.33) can be obtained by an easy adjustment to the proof of Theorem 5.4 in Section 5.3.1 (i.e. considering the single observation case for the rescaled model given by the SDE $dX_t^{(j)} = \sqrt{n} f(t)dt + dW_t$, we refer the reader to [52] for details) yields that the

minimax rate ρ satisfies

$$\rho^2 \asymp \begin{cases} (mn\epsilon^2)^{-\frac{2s}{2s+1}} & \text{if } m^{-\frac{1}{2}} n^{\frac{1}{4s}} \leqslant \epsilon \leqslant 1, \\ (\sqrt{m}n\epsilon)^{-\frac{2s}{2s+1/2}} & \text{if } m^{-\frac{1}{2}} n^{-\frac{1}{4s+2}} \leqslant \epsilon < m^{-\frac{1}{2}} n^{\frac{1}{4s}}, \\ (mn\epsilon^2)^{-1} & \text{if } \epsilon < m^{-\frac{1}{2}} n^{-\frac{1}{4s+2}}, \end{cases} \tag{6.34}$$

for locally (ϵ, δ)-DP shared randomness protocols. For local randomness protocols, we have

$$\rho^2 \asymp \begin{cases} (mn\epsilon^2)^{-\frac{2s}{2s+3/2}} & \text{if } m^{-\frac{1}{2}} n^{\frac{1}{2s-1/2}} \leqslant \epsilon \leqslant 1 \\ (\sqrt{m}n\epsilon)^{-\frac{2s}{2s+1/2}} & \text{if } m^{-\frac{1}{2}} n^{-\frac{1}{4s+2}} \leqslant \epsilon < m^{-\frac{1}{2}} n^{\frac{1}{2s-1/2}}, \\ (mn\epsilon^2)^{-1} & \text{if } \epsilon < m^{-\frac{1}{2}} n^{-\frac{1}{4s+2}}. \end{cases} \tag{6.35}$$

We show that, under local differential privacy constraints, these rates for the signal-in-white-noise model of (6.33) extend to goodness-of-fit testing in the nonparametric regression model and nonparametric density testing, in Sections 6.3.1 and 6.3.2, respectively.

6.3.1 Nonparametric regression

We consider the following version of the fixed design version of the nonparametric regression model, where machines $j = 1, \ldots, m$ each observe random variables $X_1^{(j)}, \ldots, X_n^{(j)}$ satisfying

$$X_i^{(j)} = f(i/n) + Z_i^{(j)}, \tag{6.36}$$

under the probability distribution Q_f for $f \in L_2[0,1]$ and $Z_1^{(j)}, \ldots, Z_n^{(j)}$ i.i.d. standard Gaussian random variables. The above model is sometimes thought of as a discretized version of (6.33). Observations are in practice often discrete, although it is tempting to replace it with a continuous version of (6.33), which is more convenient to work with as it avoids discretization effects.

In the random design nonparametric regression model we consider, machines $j = 1, \ldots, m$ each observe random variables $(X_1^{(j)}, \zeta_1^{(j)}), \ldots, (X_n^{(j)}, \zeta_n^{(j)})$ satisfying

$$X_i^{(j)} = f(\zeta_i^{(j)}) + Z_i^{(j)} \quad \text{under } Q_f, \tag{6.37}$$

with $Z_1^{(j)}, \ldots, Z_n^{(j)}$ i.i.d. standard Gaussian random variables and $\zeta_1^{(j)}, \ldots, \zeta_n^{(j)}$ i.i.d. uniformly distributed on $[0,1]$, independent of $Z_1^{(j)}, \ldots, Z_n^{(j)}$.

Variations of the above model include non-equispaced or non-uniform random design, or non-Gaussian errors, for which much of the theory that follows can also be extended to, as long as the required asymptotic equivalence with (6.33) can be established. See for example [40] for results on more general fixed- and random design and [110] for non-Gaussian errors.

For $f \in \mathcal{H}^{s,R}[0,1]$, let $\mathcal{Q}^{\mathrm{fxd}}$ be the experiment generated by the observations in (6.36) and let $\mathcal{Q}^{\mathrm{rdm}}$ be the experiment generated by the observations in (6.37). Outside of the distributed setting, this model is well-studied, with both minimax estimation and testing rates known, see e.g. [90, 122]. In the distributed, communication constrained setting, only the estimation rates have been derived [187, 231, 47], with the testing minimax rates unknown until now. Leveraging Corollary 6.1, we are able to (partly) derive these rates.

Specifically, consider the goodness-of-fit testing problem $H_0 : f \equiv 0 \in L_2[0,1]$ against the alternative hypotheses that

$$f \in H^{s,R}_{\rho_s} := \{f \in \mathcal{H}^{s,R}[0,1] : \|f\|_{L_2} \geqslant \rho_s \text{ and } \|f\|_{\mathcal{H}^s} \leqslant R\},$$

for $\rho_s > 0$. The minimax distributed testing risk for the above hypotheses and a distributed testing procedure T for the model $\mathcal{Q} \equiv \mathcal{Q}_{s,R} \in \{\mathcal{Q}^{\mathrm{fxd}}_{s,R}, \mathcal{Q}^{\mathrm{rdm}}_{s,R}\}$ is given by

$$\mathcal{R}_{\mathcal{Q}_{s,R}}(H^{s,R}_{\rho_s}, T) = \mathbb{Q}^Y_0 T(Y) + \sup_{f \in H^{s,R}_{\rho_s}} \mathbb{Q}^Y_f \left(1 - T(Y)\right),$$

where $\mathbb{Q}^Y_f$ denotes the marginal distribution of the transcripts when the data is generated from Q_f. For some class of distributed protocols $\mathscr{J}_{\mathcal{Q}}$ for the model $\mathcal{Q}_{s,R}$, we shall compare the distributed testing risk with that of the model $\mathcal{P}_{s,R}$ over a class of distributed protocols $\mathscr{J}_{\mathcal{P}}$. In the nonadaptive setting, this means we compare

$$\inf_{T \in \mathscr{J}_{\mathcal{Q}_{s,R}}} \mathcal{R}_{\mathcal{Q}_{s,R}}(H^{s,R}_{\rho_s}, T) \text{ to the quantity } \inf_{T \in \mathscr{J}_{\mathcal{P}}} \mathcal{R}_{\mathcal{P}_{s,R}}(H^{s,R}_{\rho_s}, T).$$

When the regularity of the true underlying signal is unknown (but assumed to lie in a given range $[s_{\min}, s_{\max}]$), it is desirable for a method to adapt to the true underlying smoothness. For a given range $1/2 < s_{\min} < s_{\max} < \infty$, the adaptive testing risk

$$\inf_{T \in \mathscr{J}_{\mathcal{Q}_{s,R}}} \sup_{s \in [s_{\min}, s_{\max}]} \mathcal{R}_{\mathcal{Q}_{s,R}}(H^{s,R}_{\rho_s}, T)$$

is to be compared to

$$\inf_{T \in \mathscr{J}_{\mathcal{P}_{s,R}}} \sup_{s \in [s_{\min}, s_{\max}]} \mathcal{R}_{\mathcal{P}_{s,R}}(H^{s,R}_{\rho_s}, T).$$

Bounds on the Le Cam distance between nonparametric regression and the signal-in-white-noise model were initially derived in [40], here we use the ones derived in [174]. For fixed design points, we shall take the Le Cam distance bound of Theorem 2.8 of the aforementioned paper, which gives

$$\Delta(\mathcal{P}_{s,R}, \mathcal{Q}^{\mathrm{fxd}}_{s,R}) \leqslant C_s R n^{1/2-s} \tag{6.38}$$

for a constant $C_s > 0$ depending only on $s > 1/2$. The assumption $s > 1/2$ is strictly necessary here for asymptotic equivalence, see Remark 4.6 in [40] for a counter

example. For the i.i.d. uniform design case, the same paper (Theorem 4.8) offers the Le Cam distance bound

$$\Delta(\mathcal{P}_{s,R}, \mathcal{Q}_{s,R}^{\mathrm{rdm}}) \leqslant C_s R n^{\frac{1-2s}{2+4s}} \tag{6.39}$$

for a constant $C_s > 0$ depending only on $s > 1/2$.

We present the minimax testing rates for distributed nonparametric regression under both bandwidth and privacy constraints, known and unknown s, across three theorems. We defer the proofs of these three theorems until the end of this section.

The first theorem gives the minimax rates for distributed nonparametric regression under bandwidth constraints, for both fixed and random design, when s is known. The rates are the same as those derived for the signal-in-white-noise model, where all observed regimes occur depending on the values of s, m, n and b.

Theorem 6.4. *Let $R > 0$, $s > 1/2$, $\mathcal{Q}_{s,R} \in \{\mathcal{Q}_{s,R}^{fxd}, \mathcal{Q}_{s,R}^{rdm}\}$ and let $b \equiv b_N$, $m \equiv m_N$ and $n \equiv N/m$ be sequences of natural numbers such that*

$$\begin{aligned} mn^{1/2-s} &\to 0 \text{ in case } \mathcal{Q}_{s,R} = \mathcal{Q}_{s,R}^{fxd}, \\ mn^{\frac{1-2s}{2+4s}} &\to 0 \text{ in case } \mathcal{Q}_{s,R} = \mathcal{Q}_{s,R}^{rdm}. \end{aligned} \tag{6.40}$$

Take $\mathscr{T}^{(b)} \in \{\mathscr{T}_{SR}^{(b)}, \mathscr{T}_{LR}^{(b)}\}$ and let $\rho \equiv \rho_{n,b,m,s}$ be a sequence of positive numbers satisfying

$$\rho^2 \asymp \begin{cases} N^{-\frac{2s}{2s+1/2}}, & \text{if } b \geqslant N^{\frac{1}{2s+1/2}}, \\ \left(\sqrt{b}N\right)^{-\frac{2s}{2s+1}}, & \text{if } n^{\frac{1}{2s+1/2}} m^{\frac{-2s}{2s+1/2}} \leqslant b < N^{\frac{1}{2s+1/2}}, \\ (\sqrt{mn})^{-\frac{2s}{2s+1/2}}, & \text{if } b < n^{\frac{1}{2s+1/2}} m^{\frac{-2s}{2s+1/2}}, \end{cases} \tag{6.41}$$

if $\mathscr{T}^{(b)} = \mathscr{T}_{SR}^{(b)}$, or

$$\rho^2 \asymp \begin{cases} N^{-\frac{2s}{2s+1/2}} & \text{if } b \geqslant N^{\frac{1}{2s+1/2}}, \\ (bN)^{-\frac{2s}{2s+3/2}} & \text{if } n^{\frac{1}{2s+1/2}} m^{\frac{-s+1/4}{2s+1/2}} \leqslant b < N^{\frac{1}{2s+1/2}}, \\ (\sqrt{mn})^{-\frac{2s}{2s+1/2}} & \text{if } b < n^{\frac{1}{2s+1/2}} m^{\frac{-s+1/4}{2s+1/2}}, \end{cases} \tag{6.42}$$

if $\mathscr{T}^{(b)} = \mathscr{T}_{LR}^{(b)}$. It holds that

$$\inf_{T \in \mathscr{T}^{(b)}} \mathcal{R}_{\mathcal{Q}}(H_{\rho'}^{s,R}, T) \to \begin{cases} 1 & \text{if } \rho' \ll \rho, \\ 0 & \text{if } \rho' \gg \rho. \end{cases}$$

The next theorem shows that, when s is unknown, but in some fixed range $1/2 < s_{\min} \leqslant s \leqslant s_{\max} < \infty$, the adaptive minimax rate for distributed nonparametric regression under bandwidth constraints matches that of the distributed signal-in-white-noise model, as derived in Theorems 5.2 and 5.3.

Theorem 6.5. *Let $R > 0$, $1/2 < s_{\min} < s_{\max} < \infty$ be given and consider for $s \in [s_{\min}, s_{\max}]$, $\mathcal{Q}_{s,R} \in \{\mathcal{Q}_{s,R}^{fxd}, \mathcal{Q}_{s,R}^{rdm}\}$. Let $b \equiv b_N$, $m \equiv m_N$ and $n \equiv N/m$ be sequences of natural numbers such that*

$$\begin{aligned} mn^{1/2-s_{\min}} &\to 0 \text{ in case } \mathcal{Q}_{s,R} = \mathcal{Q}_{s,R}^{fxd}, \\ mn^{\frac{1-2s_{\min}}{2+4s_{\min}}} &\to 0 \text{ in case } \mathcal{Q}_{s,R} = \mathcal{Q}_{s,R}^{rdm}. \end{aligned} \tag{6.43}$$

Take $\mathscr{T}^{(b)} \in \{\mathscr{T}_{SR}^{(b)}, \mathscr{T}_{LR}^{(b)}\}$. Consider for $s \in [s_{\min}, s_{\max}]$ a sequence of positive numbers satisfying $\rho_s \equiv \rho_{n,b,m,s}$ satisfying the minimax rate conditions of Theorems 5.2 and 5.3, i.e. (5.13)-(5.15) *in case $\mathscr{T}^{(b)} = \mathscr{T}_{SR}^{(b)}$ or* (5.14)-(5.16) *in case $\mathscr{T}^{(b)} = \mathscr{T}_{LR}^{(b)}$. It holds that*

$$\inf_{T \in \mathscr{T}^{(b)}} \sup_{s \in [s_{\min}, s_{\max}]} \mathcal{R}_{\mathcal{Q}}(H_{\rho_s'}^{s,R}, T) \to \begin{cases} 1 & \text{if } \rho_s' \ll \rho, \\ 0 & \text{if } \rho_s' \gg (\log\log(N))^{1/4}\rho_s. \end{cases}$$

Theorem 6.5 is a direct consequence of Corollary 6.1 and the Le Cam distance bounds of (6.38) and (6.39). It essentially says that adaptation is equally difficult in distributed nonparametric regression under bandwidth constraints as in the signal-in-white-noise model.

The next theorem shows that, under the local differential privacy constraints, the rates of the distributed signal-in-white-noise model transfer to the distributed nonparametric regression setting as well, both under fixed and random design. As in the bandwidth constraint setting, this even holds in the adaptive setting.

Theorem 6.6. *Let $R > 0$, $1/2 < s_{\min} < s_{\max} < \infty$ be given and for $s \in [s_{\min}, s_{\max}]$, take $\mathcal{Q}_{s,R} \in \{\mathcal{Q}_{s,R}^{fxd}, \mathcal{Q}_{s,R}^{rdm}\}$. Consider sequences of natural numbers $m \equiv m_N$ and $n := N/m$ such that $N = mn \to \infty$, $\epsilon \equiv \epsilon_N$ in $(N^{-1}, 1]$ and $\delta \equiv \delta_N \lesssim (mn)^{-p}$ for some constant $p \geqslant 2$.*

Take $\mathscr{T}(\epsilon, \delta) \in \{\mathscr{T}_{SR}(\epsilon, \delta), \mathscr{T}_{LR}(\epsilon, \delta)\}$ (i.e. either the class of shared or local randomness distributed locally (ϵ, δ)-DP distributed protocols), let $\rho_s \equiv \rho_{n,m,\epsilon,\delta}$ be a sequence of positive numbers satisfying (6.34) *in case $\mathscr{T}(\epsilon, \delta) = \mathscr{T}_{SR}(\epsilon, \delta)$ or* (6.35) *in case $\mathscr{T}(\epsilon, \delta) = \mathscr{T}_{LR}(\epsilon, \delta)$.*

If in addition, m, n and $s > 1/2$ satisfy (6.40)*, it holds that*

$$\inf_{T \in \mathscr{T}(\epsilon,\delta)} \mathcal{R}_{\mathcal{Q}}(H_{\rho'}^{s,R}, T) \to \begin{cases} 1 & \text{if } \rho' \ll \rho_s, \\ 0 & \text{if } \rho' \gg \log(1/\delta)\log^3(N)\rho_s. \end{cases}$$

If (6.43) *holds, we furthermore have that*

$$\sup_{s \in [s_{\min}, s_{\max}]} \inf_{T \in \mathscr{T}(\epsilon,\delta)} \mathcal{R}_{\mathcal{Q}}(H_{\rho_s'}^{s,R}, T) \to \begin{cases} 1 & \text{if } \rho_s' \ll \rho_s, \\ 0 & \text{if } \rho_s' \gg \log(1/\delta)\log^5(N)\rho_s. \end{cases}$$

Proof of Theorems 6.4, 6.5 and 6.6. Corollary 6.1 implies that, for any sequences of distributed protocols $T \equiv T_N$ in $\mathscr{T}_{\mathcal{Q}}$ there exists a sequence of distributed protocols $\tilde{T} \equiv \tilde{T}_N$ in $\mathscr{T}_{\mathcal{P}}$

$$|\mathcal{R}_{\mathcal{Q}_{s,R}}(H^{s,R}_{\rho_s}, T) - \mathcal{R}_{\mathcal{P}_{s,R}}(H^{s,R}_{\rho_s}, \tilde{T})| \leqslant m\Delta(\mathcal{P}_{s,R}, \mathcal{Q}_{s,R}). \tag{6.44}$$

The same statement holds when the roles of $\mathscr{T}_{\mathcal{Q}}$ and $\mathscr{T}_{\mathcal{P}}$ are reversed. Combining (6.40) with the bounds of (6.38) and (6.39), the right-hand side tends to zero. Consequently, the statement of Theorem 6.4 follows now by Theorem 6.4.

If

$$\sup_{s\in[s_{\min},s_{\max}]} m\Delta(\mathcal{P}_{s,R}, \mathcal{Q}_{s,R}) \to 0,$$

(6.44) implies that

$$\inf_{T\in\mathscr{T}_{\mathcal{Q}}} \sup_{s\in[s_{\min},s_{\max}]} \mathcal{R}_{\mathcal{Q}_{s,R}}(H^{s,R}_{\rho_s}, T) \geqslant \inf_{T\in\mathscr{T}_{\mathcal{P}}} \sup_{s\in[s_{\min},s_{\max}]} \mathcal{R}_{\mathcal{P}_{s,R}}(H^{s,R}_{\rho_s}, T) + o(1).$$

Since (6.44) also holds with the roles of $\mathscr{T}_{\mathcal{Q}}$ and $\mathscr{T}_{\mathcal{P}}$ are reversed, the above statement holds with the reverse inequality also, from which we can conclude that

$$\inf_{T\in\mathscr{T}_{\mathcal{Q}}} \sup_{s\in[s_{\min},s_{\max}]} \mathcal{R}_{\mathcal{Q}_{s,R}}(H^{s,R}_{\rho_s}, T) = \inf_{T\in\mathscr{T}_{\mathcal{P}}} \sup_{s\in[s_{\min},s_{\max}]} \mathcal{R}_{\mathcal{P}_{s,R}}(H^{s,R}_{\rho_s}, T) + o(1).$$

Consequently, the Le Cam bounds (6.38)-(6.39) combined with (6.43) and Theorems 5.2 and 5.3 yield the statement of Theorem 6.5.

Through the same steps, Theorem 6.6 follows from the results of Chapter 5, in particular the local differential privacy minimax rates as given in (6.34) and (6.35). □

6.3.2 Nonparametric density testing

Through a similar strategy, rates can be obtained for nonparametric density testing, which is closely related to the multinomial model discussed in Section 6.2. Consider a distribution on $[0,1]$ with Lebesgue density q. Let Q_q^n denote the product distribution. Each machine $j = 1, \ldots, m$ observes

$$X^{(j)} = (X_1^{(j)}, \ldots, X_n^{(j)}), \text{ with } X_i^{(j)} \overset{\text{i.i.d.}}{\sim} q. \tag{6.45}$$

Consider for some $\kappa \geqslant 2$ the set

$$\mathcal{Q}_{s,R} = \left\{ q \in C^{s,R}[0,1] \,:\, q \geqslant 0,\ \int_0^1 q(t)dt = 1,\ \|q\|_\infty \leqslant \kappa,\ \|1/q\|_\infty \leqslant \kappa \right\},$$

where $C^{s,R}[0,1]$ is the set of s-smooth Hölder functions with Hölder-norm bounded by $R > 0$, see Section 5.5.3 for a definition. In a slight abuse of notation, we let $\mathcal{Q}_{s,R}$ also denote the model generated by the observations of (6.45) with the Lebesgue

densities $q \in \mathcal{Q}_{s,R}$, i.e. the set consisting of the probability measures on $[0,1]^n$ given by $A \mapsto \int_A (\otimes_{i=1}^n q)(s)ds$ with $q \in \mathcal{Q}_{s,R}$.

For the above model, the minimax rates in the nondistributed case are known for both estimation and testing, see e.g. [89, 95]. In the distributed case, the rates are derived under bandwidth constraints in the case of a large number of observations locally in [30, 226] or for the case of each machine having just one observation locally ($n = 1$) in [14]. Under differential privacy constraints, only the case where each machine has just one observation ($n = 1$) has been studied in the context of estimation in [79, 176, 134, 43] and testing in [136]. Below, we derive the local differential privacy rate for multiple draws of the density per machine (i.e. $n \gg 1$).

Let q_0 denote the Lebesgue density of the uniform distribution on $[0,1]$ and consider the testing problem of the hypotheses

$$H_0 : q = q_0 \text{ versus } q \in H_{\rho_s}^{s,R} := \left\{ q \in \mathcal{Q}_{s,R} \cap \mathcal{H}^{s,R}[0,1] \, : \, \|q - q_0\|_1 \geqslant \rho \right\}. \tag{6.46}$$

Let $\mathcal{P}_{\mathcal{Q}}$ denote the model corresponding to observations generated by the observation

$$dX_t^{(j)} = \sqrt{q(t)}dt + \frac{1}{\sqrt{2n}} dW_t^{(j)}. \tag{6.47}$$

Theorem 2 in [172] gives the following bound on the Le Cam distance between this signal-in-white-noise model and the model generated by (6.45);

$$\Delta(\mathcal{P}_{s,R}, \mathcal{Q}_{s,R}) \leqslant C_{s,\kappa} R n^{\frac{1-2s}{2+4s}}, \tag{6.48}$$

for a constant $C_{s,\kappa}$ depending only on $\kappa > 0$ and s, with $s \mapsto C_{s,\kappa}$ bounded on $1/2 < s \leqslant 1$. Leveraging Lemma 6.3, we can use the above bound to obtain that the rates governing goodness-of-fit testing in the signal-in-white-noise model as studied in Chapter 5 also govern density testing. We summarize the minimax rate results for distributed density testing under either bandwidth or local differential privacy constraints, known and unknown regularity, in a single theorem below.

Theorem 6.7. *Let $1/2 < s_{\min} \leqslant s_{\max} \leqslant 1$, $\kappa, R > 0$ be given and consider a sequence $m \equiv m_N$, $n := N/m$ such that*

$$m n^{\frac{1-2s_{\min}}{2+4s_{\min}}} \to 0. \tag{6.49}$$

Let $\mathscr{J}_{\mathcal{Q}}$ denote either shared or local randomness distributed protocols under either a $b \equiv b_N$-bandwidth constraint or (ϵ, δ)-differential privacy constraint and let ρ_s the minimax rate of Theorem 6.4 or (6.34)-(6.35) for each type of protocol, respectively. Assume in addition that $\epsilon \equiv \epsilon_N$ and $\delta \equiv \delta_N$ satisfy $m^{1/4} n^{-1/4} \ll \epsilon \leqslant 1$ and $\log(1/\delta) \asymp \log N$.

For each of the choices of $\mathscr{J}_{\mathcal{Q}}$ described above there exists a positive sequence M_N, at most of poly-logarithmic order in N, such that

$$\sup_{s \in [s_{\min}, s_{\max}]} \inf_{T \in \mathscr{J}_{\mathcal{Q}}} \mathcal{R}_{\mathcal{Q}}(H_{\rho_s'}^{s,R}, T) \to \begin{cases} 1 & \text{if } \rho_s' \ll M_N^{-1} \rho_s, \\ 0 & \text{if } \rho_s' \gg M_N \rho_s. \end{cases}$$

Remark 15. The restriction to $C^{s,R}[0,1] \cap \mathcal{H}^{s,R}[0,1]$ instead of just $C^{s,R}[0,1]$ in the alternative hypothesis of (6.46) is made for simplicity, aligning the alternative hypothesis with the results in Chapter 5 for the signal-in-white-noise model that are only derived for Sobolev spaces. As remarked in the aforementioned chapter, the minimax rates derived for the signal-in-white-noise model are the same for $C^{s,R}[0,1]$ alternatives as for $H^{s,R}[0,1]$ alternatives, up to a logarithmic factor.

In the unconstrained case, the theorem recovers the minimax goodness-of-fit density testing rate for the hypothesis (6.46) as derived in [120] (up to logarithmic factors). Furthermore, under the restricted setting of the theorem, distributed density testing is shown to be approximately equally difficult as the signal-in-white-noise detection problem considered in Chapter 5, for both bandwidth and local differentially privacy constraints.

The proof is less straightforward than that of the theorems concerning nonparametric regression, as the null- and alternative hypotheses of (6.46) do not immediately translate to the hypotheses studied for the signal-in-white-noise model. To that extent, the proof has similarities with the proof of Theorem 6.1.

Proof. The Le Cam distance bound of (6.48) and Corollary 6.1 together imply that for every sequence of distributed protocols $T \equiv T_N$ in $\mathscr{J}_{\mathcal{Q}}$ there exists $T' \equiv T'_N$ in $\mathscr{J}_{\mathcal{P}}$ such that

$$\left|\mathcal{R}_{\mathcal{Q}}(H^{s,R}_{\rho'_s}, T) - \mathcal{R}_{\mathcal{P}}(H^{s,R}_{\rho'_s}, T')\right| \leqslant m\Delta(\mathcal{P}_{s,R}, \mathcal{Q}_{s,R}). \tag{6.50}$$

The same statement is true with the roles of $\mathscr{J}_{\mathcal{Q}}$ and $\mathscr{J}_{\mathcal{P}}$ reversed. The conditions (6.49) and (6.48) give an upper bound on the right-hand side of the above display, uniformly in $s \in [s_{\min}, s_{\max}]$. Consequently,

$$\inf_{T \in \mathscr{J}_{\mathcal{Q}}} \sup_{s \in [s_{\min}, s_{\max}]} \mathcal{R}_{\mathcal{Q}}(H^{s,R}_{\rho'_s}, T) = \inf_{T \in \mathscr{J}_{\mathcal{P}}} \sup_{s \in [s_{\min}, s_{\max}]} \mathcal{R}_{\mathcal{P}}(H^{s,R}_{\rho'_s}, T) + o(1). \tag{6.51}$$

Since $X^{(j)} - \sqrt{q_0}$ is a sufficient statistic in the model $\mathcal{P}_{s,R}$, we can (by another application of Corollary 6.1), consider the model $\mathcal{P}_{s,R}$ to be generated by

$$dX_t^{(j)} = \left(\sqrt{q(t)} - 1\right) dt + \frac{1}{\sqrt{2n}} dW_t^{(j)}. \tag{6.52}$$

We split the remainder of the proof into showing "ρ_s is an upper bound" and "ρ_s is a lower bound" for the minimax rate.

The rate ρ_s is an upper bound for the minimax rate (up to logarithmic factors): We start by noting that the L_1-norm is bounded above by a multiple of the Hellinger distance (Lemma 6.10 in the appendix). That is,

$$\|q_0 - q\|_1 \leqslant 2\sqrt{\int_0^1 \left(\sqrt{q_0}(s) - \sqrt{q}(s)\right)^2 ds},$$

which implies that the function $\sqrt{q_0} - \sqrt{q}$ has L_2-norm bounded from below by $\rho_s/2$ whenever $q \in H^{s,R}_{\rho_s}$. The function $x \mapsto \sqrt{x}$ has a bounded derivative on the domain $(1/L, \infty)$ and can be smoothly extended to a function on $\mathbb{R}$ that vanishes at 0 and has a bounded derivative on $\mathbb{R}$. Since q is assumed to be in $C^{s,R}[0,1] \cap \mathcal{H}^{s,R}[0,1]$ and $1/\kappa \leqslant |q| \leqslant \kappa$, by e.g. Theorem 2.87 in [26] we obtain that

$$\|\sqrt{q}\|_{\mathcal{H}^s[0,1]} \leqslant C_\kappa R$$

for a constant $C_\kappa > 0$ depending only on κ and clearly a similar bound holds for $\sqrt{q} - \sqrt{q_0}$. Consequently, it holds that

$$\inf_{T \in \mathscr{J}_\mathcal{P}} \sup_{s \in [s_{\min}, s_{\max}]} \mathcal{R}_\mathcal{P}(H^{s,R}_{M_N \rho'_s}, T) \to 0$$

by Theorems 5.2, 5.3 and 5.5 for each of the classes considered for $\mathscr{J}_\mathcal{P}$, corresponding minimax rates ρ_s and appropriate (at most) poly-logarithmic in N factor M_N. As a consequence of (6.51), the adaptive testing risk vanishes for the adaptive density testing risk.

The rate ρ_s is a lower bound for the minimax rate (up to logarithmic factors): Fix an arbitrary $s \in [s_{\min}, s_{\max}]$. We consider a similar prior on $L_2[0,1]$ as the one used in the proofs of Theorem 5.1 and Theorem 5.4. That is, for $L \in \mathbb{N}$ let the linear operator $\Psi_L : \mathbb{R}^{2^L} \to L_2[0,1]$ be defined by

$$\Psi_L \tilde{f}^L = \sum_{i=0}^{2^L - 1} \tilde{f}_i \psi_{Li},$$

for $\tilde{f}^L = (\tilde{f}_0, \ldots, \tilde{f}_{2^L-1}) \in \mathbb{R}^{2^L}$ and $\{\psi_{li} : l \geqslant l_0,\ i = 0, \ldots, 2^l - 1\}$ forming an orthonormal S-smooth wavelet basis such that $\int_0^1 \psi_{li}(s) ds = 0$ for all $l > L_0$ for some fixed $L_0 > l_0$, $S > s_{\max}$ and compactly supported (see Section 5.5.3 for a definition).

As Ψ_L is measurable, $\pi_L \circ \Psi_L^{-1}$ defines a probability measure on the Borel sigma algebra of $L_2[0,1]$ for any probability distribution π_L on the Borel sigma-algebra of $\mathbb{R}^{2^L}$. To that extent, let $\pi_L = N(0, \Gamma)$, with $\Gamma = C 2^{-L} \rho_s^2 M_N^{-2} \overline{\Gamma} \in \mathbb{R}^{2^L \times 2^L}$ for a symmetric, idempotent matrix $\overline{\Gamma} \in \mathbb{R}^{2^L \times 2^L}$ with $\operatorname{rank}(\overline{\Gamma}) \asymp 2^L$ and $C > 0$ a constant. Taking $L \equiv L_s = \lceil 1 \vee \frac{1}{s} \log(1/\rho_s) \rceil$, it follows by the proof of Theorem 5.1 (or Theorem 5.4) that

$$\pi_L \circ \Psi_L^{-1} \left(\mathcal{H}^{s,R}[0,1] \right) = 1 - o(1).$$

Furthermore, for a fixed constant $R' > 0$,

$$\pi_L \circ \Psi_L^{-1} \left(C^{s,R'}[0,1] \right) \geqslant 1 - \Pr \left(M_N^{-1} \rho^{1+1/(2s)} 2^{Ls + 1/2} \max_{1 \leqslant i \leqslant 2^L} |Z_i| \geqslant \sqrt{C} R' \right),$$

for $Z_1, \ldots, Z_{2^L}$ i.i.d. standard Gaussian. By Lemma 3.27, the probability on the right-hand side tends to zero for $M_N \gg \sqrt{\log(N)}$, meaning that f is in the s-smooth

Hölder ball of radius R' with probability tending to one. Similarly, by Lemma 3.27,

$$\pi_L \circ \Psi_L^{-1}\left(\|f\|_\infty \leqslant \rho_s\right) = 1 - o(1).$$

Setting $q_f = 1+f$ we can conclude that for $f \sim \pi_L \circ \Psi_L^{-1}$, q_f is in $C^{s,R}[0,1] \cap \mathcal{H}^{s,R}[0,1]$ with probability tending to 1. Furthermore, q_f is a probability density, since $|f| \leqslant 1$ and $\int_0^1 f(s)ds = 0$. To see the latter, note that $L \to \infty$ as $\rho \to 0$, for which we have

$$\int_0^1 f(s)ds = \sum_{k=0}^{2^L-1} f_{Lk} \int_0^1 \psi_{Lk}(s)ds = 0 \quad \forall L > L_0. \tag{6.53}$$

So, $\|f\|_1 \geqslant \rho_s/M_N$ implies that $q_f \in H^{s,R}_{M_N^{-1}\rho_s}$. The latter condition holds with $\pi_L \circ \Psi_L^{-1}$-probability $1 + O(1/C)$. To see this, note that the wavelets are compactly supported, $\int_0^1 |\psi_{Lk}(s)|ds \gtrsim 2^{-L/2}$, so it follows that

$$\int \|f\|_1 d\pi_L \circ \Psi_L^{-1}(f) \gtrsim \rho_s M_N^{-1} 2^{-L} \mathbb{E}\|\overline{\Gamma} Z\|_1 \gtrsim \rho_s M_N^{-1},$$

where $Z \sim N(0, I_{2^L})$ and the last step follows from the fact that $\overline{\Gamma}$ is idempotent and has rank of the order 2^L. By the fact that

$$\int \|f\|_1^2 d\pi_L \circ \Psi_L^{-1}(f) \leqslant \int \|f\|_2^2 d\pi_L \circ \Psi_L^{-1}(f) \lesssim C\rho_s^2/M_N^2,$$

we obtain that, for a large enough choice of $C > 0$,

$$\pi_L \circ \Psi_L^{-1}\left(\|f\|_1 \leqslant \rho_s/M_N\right) \leqslant \pi_L \circ \Psi_L^{-1}\left(\frac{M_N}{\rho_s}\left|\|f\|_1 - \int \|f\|_1 d\pi_L \circ \Psi_L^{-1}(f)\right| > C/2\right)$$

which by Chebyshev's inequality and the abound on the second moment is of the order $1/C$.

The rest of the proof follows along a similar argument as the one used in the proof of Theorem 6.1. Consider the model $\mathcal{P}'_{s,R}$ to be generated by

$$dX_t^{(j)} = \frac{f(t)}{2}dt + \frac{1}{\sqrt{2n}} dW_t^{(j)}. \tag{6.54}$$

By Pinsker's inequality,

$$\|P_f^{nm} - P_g^{nm}\|_{\mathrm{TV}} \leqslant \sqrt{\frac{mn}{2} D_{\mathrm{KL}}(P_f; P_g)} = \frac{\sqrt{mn}}{\sqrt{2}} \|f - g\|_2$$

for any $f, g \in L_2[0,1]$ and with $P_{f/2}^{2n}$ denoting the distribution of (6.54). Consequently, the probability distribution corresponding to $j = 1, \ldots, m$ i.i.d. draws of (6.54) is at most

$$D_f := \sqrt{mn}\, \|\sqrt{q} - 1 - f/2\|_2$$

far away in total variation distance from the probability distribution of the $j = 1, \dots, m$ i.i.d. draws of (6.52). Via the Taylor approximation $\sqrt{1+y}-1 = y/2-y^2/8+\frac{y^3}{16(1+\eta_y^{5/2})}$ for some $\eta \in [0,y]$, the display above with $q = 1+f$ is further bounded by $\sqrt{mn}\left\|f^2\right\|_2/2$, where it is used that $|f| \leqslant 1/2$ for all N large enough. The latter quantity is less than $\sqrt{mn}\rho_s^2/4$ with probability tending to one under $\pi_L \circ \Psi_L^{-1}$. Let $\mathbb{P}'_f$ denote the joint distribution of the transcripts $Y = (Y^{(1)}, \dots, Y^{(m)})$ corresponding to the distributed protocol T and the data $X = (X^{(1)}, \dots, X^{(m)})$ with $X^{(j)}$ governed by (6.54). Combining the above with Lemma 6.9, it follows that

$$\inf_{T \in \mathscr{J}_{\mathcal{P}}} \sup_{s \in [s_{\min}, s_{\max}]} \mathcal{R}_{\mathcal{P}}(H^{s,R}_{\rho'_s}, T) \geqslant$$

$$\inf_{T \in \mathscr{J}_{\mathcal{P}}} \sup_{s \in [s_{\min}, s_{\max}]} \Bigg[\mathbb{P}'_0 T(Y) + \int \mathbb{P}'_{f/2}(1 - T(Y)) d\pi_{L_s} \circ \Psi_{L_s}^{-1}(f)$$

$$- \pi_{L_s} \circ \Psi_{L_s}^{-1}\left(f : q_f \notin H^{s,R}_{\rho_s}, \|f\|_\infty \leqslant \rho_s\right) - \sqrt{mn}\rho_s^2/4 \Bigg].$$

It was established earlier in the proof that for $C > 0$ large enough, the second term on the right-hand side can be made arbitrarily small. In case of ρ_s corresponding to the minimax rate under bandwidth constraints, we have that

$$\sqrt{mn}\rho_s^2 \lesssim \sqrt{mn}\left(\frac{1}{\sqrt{mn}}\right)^{\frac{2}{3}} \to 0 \quad \text{as } mn \to \infty,$$

where it is used that $s > 1/2$ and $m/n \to 0$ by (6.49). Under differential privacy constraints, we similarly have that

$$\sqrt{mn}\rho_s^2 \lesssim \sqrt{mn}\left(\frac{1}{\sqrt{mn}\epsilon}\right)^{\frac{2s}{2s+1/2}} \lesssim \frac{m^{1/6}}{n^{1/6}\epsilon^{2/3}} \to 0$$

whenever $\epsilon \gg m^{1/4}n^{-1/4}$. By combining the above with (6.51), we conclude that

$$\inf_{T \in \mathscr{J}_{\mathcal{Q}}} \sup_{s \in [s_{\min}, s_{\max}]} \mathcal{R}_{\mathcal{Q}}(H^{s,R}_{M_N^{-1}\rho_s}, T) = 1 - o(1),$$

by (the proofs of) Theorems 5.2, 5.3, 5.5 for bandwidth and local differential privacy constraints for appropriately large but at most poly-logarithmic factors M_N, finishing the proof. □

Chapter acknowledgements: The quote at the start of the chapter is from [220].

6.4 Appendix

The lemmas in this section are well known in the literature, but we provide proofs for completeness. The first lemma below is used in the comparison of the multinomial model to the many-normal-means model.

Lemma 6.4. *Let $d \in 2\mathbb{N}$, $\mathcal{F} \subset \mathbb{R}^{d/2}$, and consider for $i = 1, \dots, d$ independent random variables $X_i = h_i + \sigma Z_i$ with $\sigma > 0$ and $Z_i \sim N(0,1)$ satisfying*

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

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 $X := (X_1, \dots, X_d) \sim P_f$, $f \in \mathcal{F}$ and let $\mathcal{Q}$ denote the model generated by

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

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

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

Proof. We shall show that 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 6.1). We have

$$\begin{aligned} \frac{dP_f}{dP_0}(X) &= \prod_{i=1}^{d} \exp\left(\sigma^{-1} X_i h_i - \frac{1}{2\sigma^2} h_i^2\right) \\ &= \prod_{i=1}^{d/2} \exp\left(\sigma^{-1}(a_i X_i - a_i X_{d/2+1}) f_i - \frac{1}{\sigma^2} f_i^2\right) \qquad = e^{\sigma^{-1} S^\top f - \frac{1}{\sigma^2}\|f\|_2^2}. \end{aligned}$$

In distribution, $\tilde{X} = (\tilde{X}_i)_{i\in[d]}$ is equal to S, which implies $\Delta(\mathcal{P}, \mathcal{Q}) = 0$ per Lemma 6.1. □

The following lemmas are well known but included for completeness.

Lemma 6.5. *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

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

Proof. By Pinsker's inequality,

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

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 6.6. *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 [204]. □

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

Lemma 6.7. *Let P be a signed, bounded measure defined on measurable space $(\mathcal{X}, \mathscr{X})$ and suppose that $P \ll \nu$ for a sigma-finite measure ν. It holds that*

$$\|P\|_{\mathrm{TV}} = \frac{1}{2} \sup \left\{ \int f dP : |f| \leqslant 1 \text{ and } f : \mathcal{X} \to \mathbb{R} \text{ is measurable} \right\}. \tag{6.55}$$

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, $\{f \geqslant 0\}, \{f \leqslant 0\} \in \mathscr{X}$, so $|f| \leqslant 1$ means that

$$\begin{aligned} \int f dP &\leqslant \int f \mathbb{1}_{\{f \geqslant 0\}} dP^+ - \int f \mathbb{1}_{\{f \leqslant 0\}} dP^- \\ &\leqslant \int \mathbb{1}_{\{f \geqslant 0\}} dP^+ + \int \mathbb{1}_{\{f \leqslant 0\}} dP^- \\ &\leqslant \|P^+\|_{\mathrm{TV}} + \|P^-\|_{\mathrm{TV}} \leqslant 2\|P\|_{\mathrm{TV}}. \end{aligned}$$

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

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

where the last equality follows from Lemma 6.6.

□

Lemma 6.8. *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 $(\mathcal{X}, \mathscr{X})$, with probability densities p_j, q_j for $j = 1, \dots m$. It holds that*

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

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 6.6 we obtain

$$\|P - Q\|_{\mathrm{TV}} = \frac{1}{2} \int |\prod_{j=1}^{m} p_j(x_j) - \prod_{j=1}^{m} q_j(x_j)| d\mu(x_1, \ldots, x_m). \tag{6.56}$$

By the telescoping product identity

$$a_1 \cdot a_2 \cdots a_m - b_1 \cdot b_2 \cdots b_m = \sum_{j=1}^{m} (a_j - b_j) \prod_{k=1}^{j-1} a_k \prod_{k=j+1}^{m} b_k \tag{6.57}$$

and Fubini's Theorem, the right-hand side of (6.56) is bounded by

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

□

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

Lemma 6.9. *Let $(\mathcal{X}, \mathscr{X})$ and $(\mathcal{Y}, \mathscr{Y})$ be two measurable spaces and let $K : \mathscr{Y} \times \mathcal{X} \to [0, 1]$ be a Markov kernel. For any probability measures P, Q defined on $\mathscr{X}$ it holds that*

$$\|PK - QK\|_{\mathrm{TV}} \leqslant \|P - Q\|_{\mathrm{TV}}.$$

Proof. This follows immediately from the representation in Lemma 6.7 combined with the fact that, for $|f| \leqslant 1$, $x \mapsto \int f(y) dK(y|x)$ is a measurable function bounded by 1, since K is Markov kernel. Hence,

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

□

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

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

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. [204] for details. □

Discussion

The results of this thesis mathematically characterize and quantify the impact of various communication constraints in distributed hypothesis testing, where we have investigated bandwidth constraints, differential privacy constraints, and a meta-analysis setting.

The specific statistical task that is central to the thesis is that of "signal detection" or "goodness-of-fit" testing, where we wish to decide between a null hypothesis that the data is generated by a particular specified "null" probability distribution, versus the alternative hypothesis that the data is generated by another probability distribution belonging to a family of alternatives.

The results provide insight into how the statistical problem of testing gets more difficult depending on the severity of the bandwidth or privacy constraint. The theory is derived in an abstract distributed setting, in which we have m machines (e.g. locations; hospitals, sensors, servers, etc.). Each of the $j = 1, \ldots, m$ machines communicates a transcript on the basis of a local independent sample of n data points drawn from an unknown distribution. Each transcript has to satisfy a certain communication constraint. In case of a bandwidth constraint, the transcript is considered to contain at most b-bits of information. In case of privacy constraint, the transcript $Y^{(j)}$ must satisfy a differential privacy constraint governed by two parameters ϵ and δ, where smaller values for ϵ and δ give stronger privacy guarantees. Chapter 4 concerns a meta-analysis setting, where the constraint comes in the form of restricting the type of test statistics communicated by each of the machines to those that are "typical" when only the outcome of studies are published (e.g. a test outcome or p-value).

Within the minimax paradigm, the theory in this thesis captures the difficulty of the various distributed and constraint testing problems, in terms of the characteristics of the underlying model and statistical setting. That is to say, it describes the difficulty in terms of the minimax separation rate as a function of b in the bandwidth constraint setting and ϵ and δ in the differential privacy constraint setting, as well as m, n and d (or in terms of the regularity hyperparameter "s" in the case of a nonparametric model formulation).

Chapter 2 and 3 together establish the minimax rate for the canonical many-normal-means model for both bandwidth and privacy constraints. The first of these two chapters focusses on proving impossibility theorems for these settings. Where typical proof techniques in the literature used for e.g. estimation minimax rates fall short, the chapter exhibits a novel framework which proves fruitful in deriving distributed testing minimax separation rates. Chapter 3 establishes that the rates of the impossibility theorems are indeed sharp, by exhibiting methods that attain these rates. Hence, the methods derived in the former chapter are optimal in terms of the minimax separation rate.

The results of Chapter 2 and 3 fully establish the bandwidth constraint rates in a testing setting, where previously rigorous study of performance of distributed bandwidth constraint problems has been mostly conducted for estimation. In terms of the privacy results, the findings of these two chapters contribute to the existing literature by establishing the minimax testing (and estimation) rates under differential privacy a fully general distributed setting, where earlier literature derived optimality results only in the case of local differential privacy (i.e. $n = 1$) or central differential privacy (i.e. $m = 1$).

The results of Chapter 2 and 3 testing are contrasted with known results for estimation under bandwidth constraints and differential privacy constraints, where the latter optimality results are novel and derived in Chapter 2 also. Here, multiple fundamental differences between estimation and testing which occur under the presence of communication constraints are uncovered. Where classically the high-dimensional testing problem is already fundamentally different from estimation, it is revealed that these differences persist and are in many ways exacerbated under bandwidth and privacy constraints.

In the presence of bandwidth or differential privacy constraints, it turns out that there are more possibilities in terms of testing than for its natural estimation counterparts. In the presence of these constraints, consistent testing is possible in regimes where consistent estimation is not. Furthermore, testing is subject to many phase transitions, in which different testing strategies need to be adopted for optimal performance, whereas estimation under these constraints typically be performed by a single procedure. That consistent testing is 'easier' than estimation and has more options in terms of different consistent testing strategies, conversely means that showing impossibility results turns out to be much more involved, as is exhibited in Chapter 2.

In Chapter 4, the theory derived in Chapter 2 and 3 is extended to the setting of meta-analysis, by establishing a connection between meta-analysis and distributed learning under bandwidth constraints. This chapter provides a unified, theoretical framework for evaluating the behavior of standard meta-analysis techniques, such as Fisher's and Bonferroni's method. In the normal means model, it is shown that by combining the locally optimal chi-square statistics at a meta-level one can gain a factor of $\sqrt{m}$ compared to using just a single trial. Nevertheless, regardless of the choice of the combination method, a factor of $\sqrt{m} \wedge \sqrt{d}$ is lost compared to the scenario when all

data from all trials are at our disposal. This loss in efficiency, as captured by the minimax separation rate, is the same as the one suffered under a 1-bit bandwidth constraint in the many-normal-means model, as exemplified by the theory derived in the earlier Chapters 2 and 3.

In Chapter 5, the minimax rate for goodness-of-fit testing in the nonparametric distributed signal-in-white noise model is derived, for both bandwidth- and privacy constraints settings. The nonparametric signal-in-white-noise model is a natural extension of the finite dimensional many-normal-means model considered in the previous chapters. This nonparametric model serves as a benchmark for nonparametric goodness-of-fit testing, and the results here follow from the theory established in Chapter 2 and 3. As an added difficulty, we consider the setting where the true smoothness s of the underlying signal parameter is unknown. When the smoothness s is unknown, the results and methods of the earlier chapters do not transfer as straightforwardly to the nonparametric problem. It is shown that rate optimal methods can adapt to the unknown regularity s of the underlying function with a cost of at most additional logarithmic factors in both the bandwidth and differential privacy constraint settings, where we characterize the cost of adaptation exactly under the former constraints.

In Chapter 6, some bandwidth and privacy results of the earlier chapters are shown to extend to other models as well, such as the multinomial model, nonparametric densities and nonparametric regression. The focus of this chapter is distributed goodness-of-fit testing under communication constraints, and the minimax rates for the aforementioned models are derived by leveraging existing model comparisons from the literature in the distributed setting. The chapter also exemplifies a scenario in which the distributed bandwidth constraint testing problem with n observations from the d dimensional multinomial model behaves drastically different from its many-normal-means model counterpart. The latter fact is used to show that these models are asymptotically nonequivalent when d/n is large.

A remarkable finding that is consistent across each of the constraint types and models considered, is that there is fundamentally a benefit of having access to shared randomness in the distributed setting. For certain constraint budgets, the improvement over protocols that rely solely on local sources of randomness is strict. In real applications without interaction, one should always use shared randomness if at all possible.

The theory in this thesis can be extended in many directions. In Chapter 6, we learn that depending on a relation between d and n, the multinomial model and many-normal-means model do not always exhibit the same behavior under bandwidth or privacy constraints. Understanding these differences well can give insights into the impact of the model on the cost of communication constraints. Furthermore, this provides a lens to understand differences between models outside of the distributed context. To understand these problems better, one might require to adapt the Brascamp-Lieb inequality type of argument to a non-Gaussian setting or use a different technique altogether.

Other extensions apply to the many-normal-means and signal-in-white-noise settings as well. To list a few, one could consider different alternative hypotheses, alternatives that are sparse in an appropriate sense or a multiple testing setting. Another extension is to consider settings in which variances of the noise are unknown. In principle, since variances can be estimated at an $n^{-1/2}$-rate locally in each machine (see e.g. [183]), one could conjecture that this leaves the established rates unchanged (up to logarithmic factors as observed in [47]), but this has not been verified by the author beyond the back of an envelope. Furthermore, whilst our analysis supports differing budgets to the extent that $\epsilon_j \asymp \epsilon_k$, $\delta_j \asymp \delta_k$, $b_j \asymp b_k$ and $n_j \asymp n_k$, it would be interesting to consider settings in which the machines are heterogeneous in their constraints and number of observations, such as considered in the estimation results of [189, 51]. Since differential privacy is not the only formal notion of privacy, other notions of privacy could be considered. Lastly, the distributed setting considered here is a federated setting where data is observed and shared "at once" to a single "central server". Architectures in which machines share transcripts, for example, sequentially, do not fall under this scope and merit their own study.

Bibliography

[1] F. Abramovich, Y. Benjamini, D. L. Donoho, and I. M. Johnstone. Adapting to unknown sparsity by controlling the false discovery rate. *The Annals of Statistics*, pages 584–653, 2006.

[2] J. Acharya, Z. Sun, and H. Zhang. Differentially private testing of identity and closeness of discrete distributions. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[3] J. Acharya, Z. Sun, and H. Zhang. Differentially private testing of identity and closeness of discrete distributions. *Advances in Neural Information Processing Systems*, 31, 2018.

[4] J. Acharya, C. Canonne, C. Freitag, and H. Tyagi. Test without trust: Optimal locally private distribution testing. In K. Chaudhuri and M. Sugiyama, editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 2067–2076. PMLR, 16–18 Apr 2019.

[5] J. Acharya, K. Bonawitz, P. Kairouz, D. Ramage, and Z. Sun. Context aware local differential privacy. In H. D. III and A. Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 52–62. PMLR, 13–18 Jul 2020.

[6] J. Acharya, C. L. Canonne, and H. Tyagi. Inference under information constraints i: Lower bounds from chi-square contraction. *IEEE Transactions on Information Theory*, 66(12):7835–7855, 2020. doi: 10.1109/TIT.2020.3028440.

[7] J. Acharya, C. L. Canonne, and H. Tyagi. Inference under information constraints ii: Communication constraints and shared randomness. *IEEE Transactions on Information Theory*, 66(12):7856–7877, 2020. doi: 10.1109/TIT.2020.3028439.

[8] J. Acharya, C. L. Canonne, and H. Tyagi. Distributed signal detection under communication constraints. In *Conference on Learning Theory*, pages 41–63. PMLR, 2020.

[9] J. Acharya, C. L. Canonne, and H. Tyagi. Inference under information constraints i: Lower bounds from chi-square contraction. *IEEE Transactions on Information Theory*, 66(12):7835–7855, 2020. doi: 10.1109/TIT.2020.3028440.

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

[11] J. Acharya, C. L. Canonne, and H. Tyagi. Inference Under Information Constraints I: Lower Bounds From Chi-Square Contraction. *IEEE Transactions on Information Theory*, 66(12):7835–7855, Dec. 2020. ISSN 0018-9448, 1557-9654. doi: 10.1109/TIT.2020.3028440.

[12] J. Acharya, C. L. Canonne, and H. Tyagi. Distributed signal detection under communication constraints. In J. Abernethy and S. Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 41–63. PMLR, 09–12 Jul 2020.

[13] J. Acharya, C. Canonne, Y. Liu, Z. Sun, and H. Tyagi. Distributed estimation with multiple samples per user: Sharp rates and phase transition. *Advances in neural information processing systems*, 34:18920–18931, 2021.

[14] J. Acharya, C. Canonne, A. V. Singh, and H. Tyagi. Optimal rates for nonparametric density estimation under communication constraints. *Advances in Neural Information Processing Systems*, 34:26754–26766, 2021.

[15] J. Acharya, C. L. Canonne, C. Freitag, Z. Sun, and H. Tyagi. Inference under information constraints iii: Local privacy constraints. *IEEE Journal on Selected Areas in Information Theory*, 2(1):253–267, 2021. doi: 10.1109/JSAIT.2021.3053569.

[16] J. Acharya, P. Kairouz, Y. Liu, and Z. Sun. Estimating sparse discrete distributions under privacy and communication constraints. In V. Feldman, K. Ligett, and S. Sabato, editors, *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*, volume 132 of *Proceedings of Machine Learning Research*, pages 79–98. PMLR, 16–19 Mar 2021.

[17] J. Acharya, C. L. Canonne, Y. Liu, Z. Sun, and H. Tyagi. Interactive inference under information constraints. *IEEE Transactions on Information Theory*, 68 (1):502–516, 2022. doi: 10.1109/TIT.2021.3123905.

[18] J. Acharya, Y. Liu, and Z. Sun. Discrete distribution estimation under user-level local differential privacy. In F. Ruiz, J. Dy, and J.-W. van de Meent, editors, *Proceedings of The 26th International Conference on Artificial Intelligence and*

Statistics, volume 206 of *Proceedings of Machine Learning Research*, pages 8561–8585. PMLR, 25–27 Apr 2023.

[19] S. Aerts, D. Lambrechts, S. Maity, P. Van Loo, B. Coessens, F. De Smet, L.-C. Tranchevent, B. De Moor, P. Marynen, B. Hassan, et al. Gene prioritization through genomic data fusion. *Nature biotechnology*, 24(5):537–544, 2006.

[20] R. Ahlswede and I. Csiszar. Hypothesis testing with communication constraints. 32(4):533–542. ISSN 0018-9448. doi: 10.1109/TIT.1986.1057194. Number: 4.

[21] E. I. Altman. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The journal of finance*, 23(4):589–609, 1968.

[22] E. I. Altman, M. Iwanicz-Drozdowska, E. K. Laitinen, and A. Suvas. Distressed firm and bankruptcy prediction in an international context: A review and empirical analysis of altman's z-score model. *Available at SSRN 2536340*, 2014.

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

[24] G. W. Anderson, A. Guionnet, and O. Zeitouni. *An Introduction to Random Matrices*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 2009. doi: 10.1017/CBO9780511801334.

[25] Apple Differential Privacy Team. Learning with privacy at scale. 2017.

[26] H. Bahouri. *Fourier analysis and nonlinear partial differential equations*. Springer, 2011.

[27] S. Balakrishnan and L. Wasserman. Hypothesis testing for high-dimensional multinomials: A selective review. *The Annals of Applied Statistics*, 12(2):727 – 749, 2018. doi: 10.1214/18-AOAS1155SF.

[28] S. Balakrishnan and L. Wasserman. Hypothesis testing for densities and high-dimensional multinomials. *The Annals of Statistics*, 47(4):1893–1927, 2019.

[29] Y. Baraud. Non-asymptotic minimax rates of testing in signal detection. *Bernoulli*, pages 577–606, 2002.

[30] L. P. Barnes, Y. Han, and A. Özgür. Lower bounds for learning distributions under communication constraints via fisher information. *The Journal of Machine Learning Research*, 21(1):9583–9612, 2020.

[31] F. Beaufays, K. Rao, R. Mathews, and S. Ramaswamy. Federated learning for emoji prediction in a mobile keyboard. *arXiv preprint arXiv:1906.04329*, 2019.

[32] J. O. Berger. *Statistical decision theory and Bayesian analysis*. Springer Science & Business Media, 2013.

[33] T. Berger and Z. Zhang. On the ceo problem. In *Proceedings of 1994 IEEE International Symposium on Information Theory*, pages 201–, 1994. doi: 10.1109/ISIT.1994.394767.

[34] T. Berrett and C. Butucea. Locally private non-asymptotic testing of discrete distributions is faster using interactive mechanisms. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 3164–3173. Curran Associates, Inc., 2020.

[35] A. Birnbaum. Combining independent tests of significance. *Journal of the American Statistical Association*, 49(267):559–574, 1954.

[36] S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities: a nonasymptotic theory of independence*. Oxford University Press, Oxford, 1st ed edition, 2013. ISBN 978-0-19-953525-5. OCLC: ocn818449985.

[37] H. J. Brascamp and E. H. Lieb. Best constants in young's inequality, its converse, and its generalization to more than three functions. *Advances in Mathematics*, 20(2):151–173, 1976.

[38] M. Braverman and A. Rao. Towards coding for maximum errors in interactive communication. In *Proceedings of the forty-third annual ACM symposium on Theory of computing*, pages 159–166, 2011.

[39] M. Braverman, A. Garg, T. Ma, H. L. Nguyen, and D. P. Woodruff. Communication lower bounds for statistical estimation problems via a distributed data processing inequality. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 1011–1020, 2016.

[40] L. D. Brown and M. G. Low. Asymptotic equivalence of nonparametric regression and white noise. *The Annals of Statistics*, 24(6):2384 – 2398, 1996. doi: 10.1214/aos/1032181159.

[41] L. D. Brown and C.-H. Zhang. Asymptotic nonequivalence of nonparametric experiments when the smoothness index is 1/2. *The Annals of Statistics*, 26(1): 279–287, 1998. ISSN 00905364.

[42] L. D. Brown, T. T. Cai, M. G. Low, and C.-H. Zhang. Asymptotic equivalence theory for nonparametric regression with random design. *The Annals of Statistics*, 30(3):688 – 707, 2002. doi: 10.1214/aos/1028674838.

[43] C. Butucea, A. Dubois, M. Kroll, and A. Saumard. Local differential privacy: Elbow effect in optimal density estimation and adaptation over Besov ellipsoids. *Bernoulli*, 26(3):1727 – 1764, 2020. doi: 10.3150/19-BEJ1165.

[44] C. Butucea, A. Rohde, and L. Steinberger. Interactive versus noninteractive locally differentially private estimation: Two elbows for the quadratic functional. *Annals of Statistics*, 51(2), Apr. 2023. doi: 10.1214/22-AOS2254.

[45] T. T. Cai and H. Wei. Distributed adaptive gaussian mean estimation with unknown variance: interactive protocol helps adaptation. *arXiv preprint arXiv:2001.08877*, .

[46] T. T. Cai and H. Wei. Distributed gaussian mean estimation under communication constraints: Optimal rates and communication-efficient algorithms. .

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

[48] T. T. Cai, Y. Wang, and L. Zhang. The cost of privacy: Optimal rates of convergence for parameter estimation with differential privacy. *The Annals of Statistics*, 49(5):2825–2850, 2021.

[49] T. T. Cai, Z. T. Ke, and P. Turner. Testing high-dimensional multinomials with applications to text analysis. *arXiv preprint arXiv:2301.01381*, 2023.

[50] T. T. Cai, Y. Wang, and L. Zhang. Score attack: A lower bound technique for optimal differentially private learning. *arXiv preprint arXiv:2303.07152*, 2023.

[51] T. T. Cai, A. Chakraborty, and L. Vuursteen. Optimal federated learning for nonparametric regression with heterogeneous distributed differential privacy constraints. 2024.

[52] T. T. Cai, A. Chakraborty, and L. Vuursteen. Federated nonparametric hypothesis testing with differential privacy constraints: Optimal rates and adaptive tests. 2024.

[53] L. L. Cam. Sufficiency and Approximate Sufficiency. *The Annals of Mathematical Statistics*, 35(4):1419 – 1455, 1964. doi: 10.1214/aoms/1177700372.

[54] C. L. Canonne, G. Kamath, A. McMillan, J. Ullman, and L. Zakynthinou. Private identity testing for high-dimensional distributions. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 10099–10111. Curran Associates, Inc., 2020.

[55] E. A. Carlen. Superadditivity of Fisher's information and logarithmic Sobolev inequalities. *Journal of Functional Analysis*, 101(1):194–211, Oct. 1991. ISSN 00221236. doi: 10.1016/0022-1236(91)90155-X.

[56] E. A. Carlen and D. Cordero-Erausquin. Subadditivity of the entropy and its relation to Brascamp-Lieb type inequalities. *arXiv:0710.0870 [math]*, Jan. 2008. arXiv: 0710.0870.

[57] A. V. Carter. Deficiency distance between multinomial and multivariate normal experiments. *The Annals of Statistics*, 30(3):708 – 730, 2002. doi: 10.1214/aos/1028674839.

[58] W.-N. Chen, P. Kairouz, and A. Ozgur. Pointwise bounds for distribution estimation under communication constraints. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 24593–24603. Curran Associates, Inc., 2021.

[59] W.-N. Chen, P. Kairouz, and A. Ozgur. Breaking the dimension dependence in sparse distribution estimation under communication constraints. In M. Belkin and S. Kpotufe, editors, *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 1028–1059. PMLR, 15–19 Aug 2021.

[60] Y. Chen and F. Chen. An interview with professor ingrid daubechies [interview]. *IEEE Circuits and Systems Magazine*, 22(4):3–61, 2022. doi: 10.1109/MCAS.2022.3214401.

[61] Z. Chen. Optimal tests for combining p-values. *Applied Sciences*, 12(1):322, 2021.

[62] N. Clements, S. K. Sarkar, and W. Guo. Astronomical transient detection controlling the false discovery rate. In *Statistical challenges in modern astronomy V*, pages 383–396. Springer, 2012.

[63] E. S. Cochran, J. F. Lawrence, C. Christensen, and R. S. Jakka. The quake-catcher network: Citizen science expanding seismic horizons. *Seismological Research Letters*, 80(1):26–30, 2009.

[64] A. Cohen, I. Daubechies, and P. Vial. Wavelets on the interval and fast wavelet transforms. *Applied and computational harmonic analysis*, 1993.

[65] T. M. Cover and J. A. Thomas. ELEMENTS OF INFORMATION THEORY. page 774.

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

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

[68] A. Dalalyan and M. Reiß. Asymptotic statistical equivalence for scalar ergodic diffusions. *Probability theory and related fields*, 134:248–282, 2006.

[69] A. Dalalyan and M. Reiß. Asymptotic statistical equivalence for ergodic diffusions: the multidimensional case. *Probability theory and related fields*, 137: 25–47, 2007.

[70] I. Daubechies. *Ten lectures on wavelets*. SIAM, 1992.

[71] S. Delattre and M. Hoffmann. Asymptotic equivalence for a null recurrent diffusion. *Bernoulli*, 8(2):139–174, 2002. ISSN 13507265.

[72] L. Devroye, A. Mehrabian, and T. Reddad. The total variation distance between high-dimensional gaussians with the same mean. *arXiv preprint arXiv:1810.08693*, 2018.

[73] I. Diakonikolas, T. Gouleakis, D. M. Kane, and S. Rao. Communication and memory efficient testing of discrete distributions. In A. Beygelzimer and D. Hsu, editors, *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 1070–1106, Phoenix, USA, 25–28 Jun 2019. PMLR.

[74] B. Ding, J. Kulkarni, and S. Yekhanin. Collecting telemetry data privately. *Advances in Neural Information Processing Systems*, 30, 2017.

[75] A. Dubois, T. Berrett, and C. Butucea. Goodness-of-Fit Testing for Hölder Continuous Densities Under Local Differential Privacy. In *Foundations of Modern Statistics*, volume PROMS-425 of *Springer Proceedings in Mathematics & Statistics*, pages 53–119. Springer International Publishing, 2023. doi: 10.1007/978-3-031-30114-8_2.

[76] J. C. Duchi and M. J. Wainwright. Distance-based and continuum fano inequalities with applications to statistical estimation. 2013.

[77] J. C. Duchi, M. I. Jordan, M. J. Wainwright, and Y. Zhang. Optimality guarantees for distributed statistical estimation.

[78] J. C. Duchi, M. I. Jordan, and M. J. Wainwright. Local privacy and statistical minimax rates. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pages 429–438. IEEE, 2013.

[79] J. C. Duchi, M. I. Jordan, and M. J. Wainwright. Minimax optimal procedures for locally private estimation. *Journal of the American Statistical Association*, 113(521):182–201, 2018.

[80] G. Duncan and D. Lambert. The risk of disclosure for microdata. *Journal of Business & Economic Statistics*, 7(2):207–217, 1989.

[81] G. T. Duncan and R. W. Pearson. Enhancing access to microdata while protecting confidentiality: Prospects for the future. *Statistical Science*, 6(3):219–232, 1991.

[82] C. Dwork. Differential privacy. In *International colloquium on automata, languages, and programming*, pages 1–12. Springer, 2006.

[83] C. Dwork and M. Naor. On the difficulties of disclosure prevention in statistical databases or the case for differential privacy. *Journal of Privacy and Confidentiality*, 2(1), 2010.

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

[85] C. Dwork, F. McSherry, K. Nissim, and A. Smith. 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*, pages 265–284. Springer, 2006.

[86] C. Dwork, A. Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

[87] C. Dwork, A. Smith, T. Steinke, and J. Ullman. Exposed! a survey of attacks on private data. *Annual Review of Statistics and Its Application*, 4:61–84, 2017.

[88] E. S. Edgington. An additive method for combining probability values from independent experiments. *The Journal of Psychology*, 80(2):351–363, 1972.

[89] S. Y. Efroimovich. Nonparametric estimation of a density of unknown smoothness. *Theory of Probability & Its Applications*, 30(3):557–568, 1986.

[90] S. Efromovich. On nonparametric regression for iid observations in a general setting. *The Annals of Statistics*, 24(3):1126–1144, 1996. ISSN 00905364.

[91] S. Efromovich and A. Samarov. Asymptotic equivalence of nonparametric regression and white noise model has its limits. *Statistics & Probability Letters*, 28 (2):143–145, 1996. ISSN 0167-7152. doi: https://doi.org/10.1016/0167-7152(95)00109-3.

[92] B. Efron. *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*, volume 1. Cambridge University Press, 2012.

[93] U. Erlingsson, V. Pihur, and A. Korolova. Rappor: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, CCS '14, page 1054–1067, New York, NY, USA, 2014. Association for Computing Machinery. ISBN 9781450329576. doi: 10.1145/2660267.2660348.

[94] M. Ermakov. Asymptotically minimax tests for nonparametric hypotheses concerning the distribution density. *Journal of Soviet Mathematics*, 52:2891–2898, 1990.

[95] M. S. Ermakov. Minimax nonparametric testing of hypotheses on the distribution density. *Theory of Probability & Its Applications*, 39(3):396–416, 1995. doi: 10.1137/1139028.

[96] E. Evangelou and J. P. Ioannidis. Meta-analysis methods for genome-wide association studies and beyond. *Nature Reviews Genetics*, 14(6):379–389, 2013.

[97] R. M. Fano and D. Hawkins. Transmission of information: A statistical theory of communications. *American Journal of Physics*, 29(11):793–794, 1961.

[98] H. Finner and M. Roters. Log-concavity and inequalities for chi-square, f and beta distributions with applications in multiple comparisons. *Statistica Sinica*, pages 771–787, 1997.

[99] O. Fischer, U. Meir, and R. Oshman. Distributed uniformity testing. In *Proceedings of the 2018 ACM Symposium on Principles of Distributed Computing*, PODC '18, page 455–464, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450357951. doi: 10.1145/3212734.3212772.

[100] R. A. Fisher. Statistical methods for research workers. In *Breakthroughs in statistics*, pages 66–70. Springer, 1992.

[101] A. Friedman and A. Schuster. Data mining with differential privacy. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 493–502, 2010.

[102] M. Fromont, M. Lerasle, and P. Reynaud-Bouret. Family-Wise Separation Rates for multiple testing. *The Annals of Statistics*, 44(6):2533 – 2563, 2016. doi: 10.1214/15-AOS1418.

[103] M. Gaboardi, H. Lim, R. Rogers, and S. Vadhan. Differentially private chi-squared hypothesis testing: Goodness of fit and independence testing. In M. F. Balcan and K. Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 2111–2120, New York, New York, USA, 20–22 Jun 2016. PMLR.

[104] V. Genon-Catalot, C. Laredo, and M. Nussbaum. Asymptotic equivalence of estimating a Poisson intensity and a positive diffusion drift. *The Annals of Statistics*, 30(3):731 – 753, 2002. doi: 10.1214/aos/1028674840.

[105] R. D. Gill and B. Y. Levit. Applications of the van trees inequality: a bayesian cramér-rao bound. *Bernoulli*, pages 59–79, 1995.

[106] E. Gine and R. Nickl. *Mathematical Foundations of Infinite-Dimensional Statistical Models*. Cambridge University Press, Cambridge, 2016. ISBN 978-1-107-33786-2. doi: 10.1017/CBO9781107337862.

[107] G. K. Golubev, M. Nussbaum, and H. H. Zhou. Asymptotic equivalence of spectral density estimation and Gaussian white noise. *The Annals of Statistics*, 38(1):181 – 214, 2010. doi: 10.1214/09-AOS705.

[108] I. Good. On the weighted combination of significance tests. *Journal of the Royal Statistical Society: Series B (Methodological)*, 17(2):264–265, 1955.

[109] I. Grama and M. Nussbaum. Asymptotic equivalence for nonparametric generalized linear models. *Probability Theory and Related Fields*, 111:167–214, 1998.

[110] I. Grama and M. Nussbaum. Asymptotic equivalence for nonparametric regression. *Mathematical methods of statistics*, 11(1):1–36, 2002.

[111] P. Grünwald, R. de Heide, and W. M. Koolen. Safe testing. In *2020 Information Theory and Applications Workshop (ITA)*, pages 1–54, 2020. doi: 10.1109/ITA50056.2020.9244948.

[112] F. Guglielmetti, R. Fischer, and V. Dose. Background–source separation in astronomical images with bayesian probability theory–i. the method. *Monthly Notices of the Royal Astronomical Society*, 396(1):165–190, 2009.

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

[114] A. Hard, K. Rao, R. Mathews, S. Ramaswamy, F. Beaufays, S. Augenstein, H. Eichner, C. Kiddon, and D. Ramage. Federated learning for mobile keyboard prediction. *arXiv preprint arXiv:1811.03604*, 2018.

[115] W. Härdle, G. Kerkyacharian, D. Picard, and A. Tsybakov. *Wavelets, approximation, and statistical applications*, volume 129. Springer Science & Business Media, 2012.

[116] T. Hospedales, A. Antoniou, P. Micaelli, and A. Storkey. Meta-learning in neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(9):5149–5169, 2021.

[117] I. Ilf and Y. Petrov. *Little Golden America*. Ishi Press, 1937.

[118] Y. Ingster and I. A. Suslina. *Nonparametric goodness-of-fit testing under Gaussian models*, volume 169. Springer Science & Business Media, 2003.

[119] Y. I. Ingster. Minimax testing of nonparametric hypotheses on a distribution density in the l_p metrics. 31(2):333–337. ISSN 0040-585X, 1095-7219. doi: 10.1137/1131042. Number: 2.

[120] Y. I. Ingster. Minimax testing of nonparametric hypotheses on a distribution density in the l_p metrics. *Theory of Probability & Its Applications*, 31(2): 333–337, 1987.

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

[122] Y. I. Ingster and T. Sapatinas. Minimax goodness-of-fit testing in multivariate nonparametric regression. *Mathematical Methods of Statistics*, 18:241–269, 2009.

[123] Y. I. Ingster and I. A. Suslina. *Nonparametric Goodness-of-Fit Testing Under Gaussian Models*, volume 169 of *Lecture Notes in Statistics*. Springer New York, New York, NY, 2003. ISBN 978-0-387-95531-5 978-0-387-21580-8. doi: 10.1007/978-0-387-21580-8.

[124] M. Jähnisch and M. Nussbaum. Asymptotic equivalence for a model of independent non identically distributed observations. *Statistics & Decisions*, 21(3): 197–218, 2003.

[125] I. M. Johnstone and B. W. Silverman. Needles and straw in haystacks: Empirical Bayes estimates of possibly sparse sequences. *The Annals of Statistics*, 32 (4):1594 – 1649, 2004. doi: 10.1214/009053604000000030.

[126] D. Jurafsky and J. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, volume 2. Prentice Hall, 02 2008.

[127] G. Kamath, J. Li, V. Singhal, and J. Ullman. Privately learning high-dimensional distributions. In *Conference on Learning Theory*, pages 1853–1902. PMLR, 2019.

[128] G. Kamath, A. Mouzakis, M. Regehr, V. Singhal, T. Steinke, and J. Ullman. A bias-variance-privacy trilemma for statistical estimation. *arXiv preprint arXiv:2301.13334*, 2023.

[129] V. Karwa and S. Vadhan. Finite sample differentially private confidence intervals. *arXiv preprint arXiv:1711.03908*, 2017.

[130] J. Konečnỳ, H. B. McMahan, D. Ramage, and P. Richtárik. Federated optimization: Distributed machine learning for on-device intelligence. *arXiv preprint arXiv:1610.02527*, 2016.

[131] J. Konečnỳ, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, and D. Bacon. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*, 2016.

[132] A. Krämer, J. Green, J. Pollard Jr, and S. Tugendreich. Causal analysis approaches in ingenuity pathway analysis. *Bioinformatics*, 30(4):523–530, 2014.

[133] O. P. Kreidl, J. N. Tsitsiklis, and S. I. Zoumpoulis. On decentralized detection with partial information sharing among sensors. 59(4):1759–1765. ISSN 1053-587X, 1941-0476. doi: 10.1109/TSP.2010.2099223. Number: 4.

[134] M. Kroll. On density estimation at a fixed point under local differential privacy. *Electronic Journal of Statistics*, 15(1):1783 – 1813, 2021. doi: 10.1214/21-EJS1830.

[135] J. Kulynych and H. T. Greely. 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):94–132, 2017.

[136] J. Lam-Weil, B. Laurent, and J.-M. Loubes. Minimax optimal goodness-of-fit testing for densities and multinomials under a local differential privacy constraint. *Bernoulli*, 28(1):579–600, 2022.

[137] L. Le Cam. *Asymptotic methods in statistical decision theory*. Springer Science & Business Media, 2012.

[138] L. Le Cam and G. L. Yang. *Asymptotics in statistics: some basic concepts*. Springer Science & Business Media, 2000.

[139] E. L. Lehmann, J. P. Romano, and G. Casella. *Testing statistical hypotheses*, volume 3. Springer, 1986.

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

[141] D. Levy, Z. Sun, K. Amin, S. Kale, A. Kulesza, M. Mohri, and A. T. Suresh. Learning with user-level privacy. *Advances in Neural Information Processing Systems*, 34:12466–12479, 2021.

[142] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith. Federated learning: Challenges, methods, and future directions. *IEEE signal processing magazine*, 37(3):50–60, 2020.

[143] E. H. Lieb. Gaussian kernels have only Gaussian maximizers. *Inventiones Mathematicae*, 102(1):179–208, Dec. 1990. ISSN 0020-9910. doi: 10.1007/BF01233426. Publisher: Springer New York.

[144] T. Lipták. On the combination of independent tests. *Magyar Tud Akad Mat Kutato Int Kozl*, 3:171–197, 1958.

[145] J. Liu, T. A. Courtade, P. Cuff, and S. Verdu. Brascamp-Lieb inequality and its reverse: An information theoretic view. In *2016 IEEE International Symposium on Information Theory (ISIT)*, pages 1048–1052, Barcelona, Spain, July 2016. IEEE. ISBN 978-1-5090-1806-2. doi: 10.1109/ISIT.2016.7541459.

[146] J. Liu, T. A. Courtade, P. Cuff, and S. Verdu. Smoothing Brascamp-Lieb Inequalities and Strong Converses for Common Randomness Generation. *arXiv:1602.02216 [cs, math]*, Feb. 2016. arXiv: 1602.02216.

[147] Y. Liu, A. T. Suresh, F. X. X. Yu, S. Kumar, and M. Riley. Learning discrete distributions: user vs item-level privacy. *Advances in Neural Information Processing Systems*, 33:20965–20976, 2020.

[148] T. M. Loughin. A systematic comparison of methods for combining p-values from independent tests. *Computational statistics & data analysis*, 47(3):467–485, 2004.

[149] B. M. Malone, F. Tan, S. M. Bridges, and Z. Peng. Comparison of four chip-seq analytical algorithms using rice endosperm h3k27 trimethylation profiling data. *PloS one*, 6(9):e25260, 2011.

[150] E. Mariucci. Asymptotic equivalence for inhomogeneous jump diffusion processes and white noise. *ESAIM: Probability and Statistics*, 19:560–577, 2015.

[151] E. Mariucci. Le cam theory on the comparison of statistical models. *arXiv preprint arXiv:1605.03301*, 2016.

[152] E. Mariucci. Asymptotic equivalence for density estimation and Guassian white noise: an extension. *Annales de l'ISUP*, 60(1-2):23–34, 2016. 11 pages.

[153] E. J. McShane. Extension of range of functions. *Bulletin of the American Mathematical Society*, 1934.

[154] G. E. Moore. Cramming more components onto integrated circuits. *Proceedings of the IEEE*, 86(1):82–85, 1998.

[155] G. S. Mudholkar and E. George. The logit statistic for combining probabilities-an overview. 1977.

[156] A. Narayanan and V. Shmatikov. Robust de-anonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy (sp 2008)*, pages 111–125. IEEE, 2008.

[157] S. Narayanan. Private high-dimensional hypothesis testing. In P.-L. Loh and M. Raginsky, editors, *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 3979–4027. PMLR, 02–05 Jul 2022.

[158] S. Narayanan, V. Mirrokni, and H. Esfandiari. Tight and robust private mean estimation with few users. In *International Conference on Machine Learning*, pages 16383–16412. PMLR, 2022.

[159] J. Neyman and E. S. Pearson. Ix. on the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231(694-706):289–337, 1933.

[160] A. Nguyen, T. Do, M. Tran, B. X. Nguyen, C. Duong, T. Phan, E. Tjiputra, and Q. D. Tran. Deep federated learning for autonomous driving. In *2022 IEEE Intelligent Vehicles Symposium (IV)*, pages 1824–1830. IEEE, 2022.

[161] M. Nussbaum. Asymptotic equivalence of density estimation and Gaussian white noise. *The Annals of Statistics*, 24(6):2399 – 2430, 1996. doi: 10.1214/aos/1032181160.

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

[163] L. Paninski. A Coincidence-Based Test for Uniformity Given Very Sparsely Sampled Discrete Data. *IEEE Transactions on Information Theory*, 54(10): 4750–4755, Oct. 2008. ISSN 0018-9448. doi: 10.1109/TIT.2008.928987.

[164] K. Pearson. On a new method of determining" goodness of fit". *Biometrika*, 26 (4):425–442, 1934.

[165] V. V. Petrov. Sums of independent random variables. In *Sums of Independent Random Variables*. De Gruyter, 2022.

[166] J. K. Pritchard, M. Stephens, and P. Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959, 2000.

[167] J. Quackenbush. Microarray data normalization and transformation. *Nature genetics*, 32(4):496–501, 2002.

[168] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[169] M. Raginsky. Strong Data Processing Inequalities and $\Phi $ -Sobolev Inequalities for Discrete Channels. *IEEE Transactions on Information Theory*, 62(6):3355–3389, June 2016. ISSN 1557-9654. doi: 10.1109/TIT.2016.2549542. Conference Name: IEEE Transactions on Information Theory.

[170] M. Raginsky. Strong data processing inequalities and ϕ -sobolev inequalities for discrete channels. *IEEE Transactions on Information Theory*, 62(6):3355–3389, 2016. doi: 10.1109/TIT.2016.2549542.

[171] A. Rao and A. Yehudayoff. *Communication Complexity: and Applications*. Cambridge University Press, 2020. ISBN 9781108497985.

[172] K. Ray and J. Schmidt-Hieber. The le cam distance between density estimation, poisson processes and gaussian white noise. *Mathematical Statistics and Learning*, 1(2):101–170, 2018.

[173] K. Ray and J. Schmidt-Hieber. Asymptotic nonequivalence of density estimation and Gaussian white noise for small densities. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 55(4):2195 – 2208, 2019. doi: 10.1214/18-AIHP946.

[174] M. Reiß. Asymptotic equivalence for nonparametric regression with multivariate and random design. *The Annals of Statistics*, 2008.

[175] I. M. Rodriguez, W. N. Sexton12, P. E. Singer, and L. Vilhuber. The modernization of statistical disclosure limitation at the us census bureau. 2020.

[176] M. Sart. Density estimation under local differential privacy and Hellinger loss. *Bernoulli*, 29(3):2318 – 2341, 2023. doi: 10.3150/22-BEJ1543.

[177] H. Schütze, C. D. Manning, and P. Raghavan. *Introduction to information retrieval*, volume 39. Cambridge University Press Cambridge, 2008.

[178] G. Shafer, A. Shen, N. Vereshchagin, and V. Vovk. Test Martingales, Bayes Factors and p-Values. *Statistical Science*, 26(1):84–101, Feb. 2011. ISSN 0883-4237. doi: 10.1214/10-STS347. arXiv: 0912.4269.

[179] G. Shafer et al. Testing by betting: A strategy for statistical and scientific communication. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 184(2):407–431, 2021.

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

[181] O. Sheffet. Locally private hypothesis testing. In *International Conference on Machine Learning*, pages 4605–4614. PMLR, 2018.

[182] N. Smirnov. Table for Estimating the Goodness of Fit of Empirical Distributions. *The Annals of Mathematical Statistics*, 19(2):279 – 281, 1948. doi: 10.1214/aoms/1177730256.

[183] V. Spokoiny. Variance estimation for high-dimensional regression models. *Journal of Multivariate Analysis*, 82(1):111–133, 2002.

[184] V. G. Spokoiny. Adaptive hypothesis testing using wavelets. *The Annals of Statistics*, 24(6), Dec. 1996. ISSN 0090-5364. doi: 10.1214/aos/1032181163.

[185] S. A. Stouffer, E. A. Suchman, L. C. DeVinney, S. A. Star, and R. M. Williams Jr. The american soldier: Adjustment during army life.(studies in social psychology in world war ii), vol. 1. 1949.

[186] H. Strasser. *Mathematical theory of statistics: statistical experiments and asymptotic decision theory*. Number 7 in De Gruyter studies in mathematics. W. de Gruyter, Berlin ; New York, 1985. ISBN 978-0-89925-028-1.

[187] B. Szabo and H. van Zanten. Adaptive distributed methods under communication constraints. *The Annals of Statistics*, 48(4):2347–2380, 2020.

[188] B. Szabo and H. van Zanten. Distributed function estimation: adaptation using minimal communication. *arXiv preprint arXiv:2003.12838*, 2020.

[189] B. Szabó and H. van Zanten. Adaptive distributed methods under communication constraints. *The Annals of Statistics*, 48(4):2347 – 2380, 2020. doi: 10.1214/19-AOS1890.

[190] B. Szabó and H. van Zanten. Distributed function estimation: Adaptation using minimal communication. *Mathematical Statistics and Learning*, 5(3):159–199, 2022.

[191] B. Szabo and A. Zaman. Distributed nonparametric estimation under communication constraints. *arXiv preprint arXiv:2204.10373*, 2022.

[192] B. Szabó, L. Vuursteen, and H. van Zanten. Optimal high-dimensional and nonparametric distributed testing under communication constraints. *arXiv preprint arXiv:2202.00968*, 2022.

[193] B. Szabó, L. Vuursteen, and H. Van Zanten. Optimal distributed composite testing in high-dimensional gaussian models with 1-bit communication. *IEEE Transactions on Information Theory*, 68(6):4070–4084, 2022.

[194] B. Szabo, A. van der Vaart, L. Vuursteen, and H. van Zanten. Optimal testing using combined test statistics across independent studies. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[195] B. Szabó, L. Vuursteen, and H. Van Zanten. Optimal high-dimensional and nonparametric distributed testing under communication constraints. *The Annals of Statistics*, 51(3):909–934, 2023.

[196] Y. C. Tai and T. P. Speed. A multivariate empirical Bayes statistic for replicated microarray time course data. *The Annals of Statistics*, 34(5):2387 – 2412, 2006. doi: 10.1214/009053606000000759.

[197] A. Tarighati, J. Gross, and J. Jalden. Decentralized Hypothesis Testing in Energy Harvesting Wireless Sensor Networks. *IEEE Transactions on Signal Processing*, 65(18):4862–4873, Sept. 2017. ISSN 1053-587X, 1941-0476. doi: 10.1109/TSP.2017.2716909.

[198] Te Sun Han and S. Amari. Statistical inference under multiterminal data compression. 44(6):2300–2324. ISSN 00189448. doi: 10.1109/18.720540. Number: 6.

[199] R. R. Tenney and N. R. Sandell. Detection with distributed sensors. AES-17 (4):501–510. ISSN 2371-9877. doi: 10.1109/TAES.1981.309178. Number: 4.

[200] J. G. Thomas, J. M. Olson, S. J. Tapscott, and L. P. Zhao. An efficient and robust statistical modeling approach to discover differentially expressed genes using genomic expression profiles. *Genome Research*, 11(7):1227–1236, 2001.

[201] H. Thorisson. *Coupling, Stationarity, and Regeneration*. Probability and Its Applications. Springer New York, 2000. ISBN 9780387987798.

[202] L. H. C. Tippett et al. The methods of statistics. an introduction mainly for experimentalists. *The methods of statistics. An introduction mainly for experimentalists.*, 1941.

[203] J. N. Tsitsiklis. Decentralized detection by a large number of sensors. 1(2):167–182, 1988. ISSN 0932-4194, 1435-568X. doi: 10.1007/BF02551407. Number: 2.

[204] A. B. Tsybakov. *Introduction to nonparametric estimation*. Springer series in statistics. Springer, New York ; London, 2009. ISBN 978-0-387-79051-0 978-0-387-79052-7. OCLC: ocn300399286.

[205] A. W. v. d. Vaart. *Asymptotic statistics*. Cambridge series in statistical and probabilistic mathematics. Cambridge University Press. ISBN 978-0-521-49603-2.

[206] G. Valiant and P. Valiant. An automatic inequality prover and instance optimal identity testing. *SIAM Journal on Computing*, 46(1):429–455, 2017. doi: 10.1137/151002526.

[207] D. Van der Hoeven, H. Hadiji, and T. van Erven. Distributed online learning for joint regret with communication constraints. In S. Dasgupta and N. Haghtalab, editors, *Proceedings of The 33rd International Conference on Algorithmic Learning Theory*, volume 167 of *Proceedings of Machine Learning Research*, pages 1003–1042. PMLR, 29 Mar–01 Apr 2022.

[208] A. W. Van Der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes*. Springer Series in Statistics. Springer, 1996. ISBN 9780387946405; 0387946403.

[209] W. Van Zwet and J. Oosterhoff. On the combination of independent test statistics. *The Annals of Mathematical Statistics*, 38(3):659–680, 1967.

[210] R. Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 1 edition, Sept. 2018. ISBN 978-1-108-23159-6 978-1-108-41519-4. doi: 10.1017/9781108231596.

[211] R. Von Mises. Statistik und wahrheit. *Julius Springer*, 20, 1928.

[212] V. Vovk and R. Wang. Combining p-values via averaging. *Biometrika*, 107(4):791–808, 2020.

[213] V. Vovk and R. Wang. E-values: Calibration, combination and applications. *The Annals of Statistics*, 49(3):1736–1754, 2021.

[214] V. Vovk, B. Wang, and R. Wang. Admissible ways of merging p-values under arbitrary dependence. *The Annals of Statistics*, 50(1):351–375, 2022.

[215] L. Vuursteen. Optimal private and communication constraint distributed goodness-of-fit testing for discrete distributions in the large sample regime. *Preprint available upon request*, 2024.

[216] A. Wald. Statistical Decision Functions. *The Annals of Mathematical Statistics*, 20(2):165 – 205, 1949. doi: 10.1214/aoms/1177730030.

[217] Y. Wang. Asymptotic nonequivalence of GARCH models and diffusions. *The Annals of Statistics*, 30(3):754 – 783, 2002. doi: 10.1214/aos/1028674841.

[218] M. C. Whitlock. Combining probability from independent tests: the weighted z-method is superior to fisher's approach. *Journal of evolutionary biology*, 18 (5):1368–1373, 2005.

[219] A. Xu and M. Raginsky. Information-Theoretic Lower Bounds on Bayes Risk in Decentralized Estimation. *arXiv:1607.00550 [cs, math, stat]*, July 2016. arXiv: 1607.00550.

[220] G. L. Yang and L. L. Cam. A conversation with lucien le cam. *Statistical Science*, 14(2):223–241, 1999. ISSN 08834237.

[221] Y. Yang and A. Barron. Information-theoretic determination of minimax rates of convergence. *Annals of Statistics*, pages 1564–1599, 1999.

[222] A. C.-C. Yao. Some complexity questions related to distributive computing(preliminary report). In *Proceedings of the Eleventh Annual ACM Symposium on Theory of Computing*, STOC '79, page 209–213, New York, NY, USA, 1979. Association for Computing Machinery. ISBN 9781450374385. doi: 10.1145/800135.804414.

[223] M. Ye and A. Barg. Optimal schemes for discrete distribution estimation under locally differential privacy. *IEEE Transactions on Information Theory*, 64(8): 5662–5676, 2018. doi: 10.1109/TIT.2018.2809790.

[224] S. Yoon, B. Baik, T. Park, and D. Nam. Powerful p-value combination methods to detect incomplete association. *Scientific reports*, 11(1):6980, 2021.

[225] M. Yuan and D.-X. Zhou. Minimax optimal rates of estimation in high dimensional additive models. *The Annals of Statistics*, 44(6):2564 – 2593, 2016. doi: 10.1214/15-AOS1422.

[226] A. Zaman and B. Szabó. Distributed nonparametric estimation under communication constraints. *arXiv preprint arXiv:2204.10373*, 2022.

[227] R. Zamir. A proof of the Fisher information inequality via a data processing argument. *IEEE Transactions on Information Theory*, 44(3):1246–1250, May 1998. ISSN 00189448. doi: 10.1109/18.669301. Number: 3.

[228] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to information retrieval. *ACM Transactions on Information Systems (TOIS)*, 22(2):179–214, 2004.

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

[230] Y. Zhu and J. Lafferty. Distributed nonparametric regression under communication constraints. In J. Dy and A. Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 6009–6017. PMLR, 10–15 Jul 2018.

[231] Y. Zhu and J. Lafferty. Distributed nonparametric regression under communication constraints. In *International Conference on Machine Learning*, pages 6009–6017. PMLR, 2018.

Publications

Published

[193] B. Szabó, L. Vuursteen, and H. Van Zanten. Optimal distributed composite testing in high-dimensional gaussian models with 1-bit communication. *IEEE Transactions on Information Theory*, 68(6): 4070–4084, 2022

[195] B. Szabó, L. Vuursteen, and H. Van Zanten. Optimal high-dimensional and nonparametric distributed testing under communication constraints. *The Annals of Statistics*, 51(3):909–934, 2023

[194] B. Szabo, A. van der Vaart, L. Vuursteen, and H. van Zanten. Optimal testing using combined test statistics across independent studies. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023

Submitted

[51] T. T. Cai, A. Chakraborty, and L. Vuursteen. Optimal federated learning for nonparametric regression with heterogeneous distributed differential privacy constraints. 2024

[52] T. T. Cai, A. Chakraborty, and L. Vuursteen. Federated nonparametric hypothesis testing with differential privacy constraints: Optimal rates and adaptive tests. 2024

[215] L. Vuursteen. Optimal private and communication constraint distributed goodness-of-fit testing for discrete distributions in the large sample regime. *Preprint available upon request*, 2024