Complexity of High-Dimensional Identity Testing with Coordinate Conditional Sampling

ANTONIO BLANCA, Pennsylvania State University, University Park, PA, USA
ZONGCHEN CHEN, Georgia Institute of Technology, Atlanta, GA, USA
DANIEL ŠTEFANKOVIČ, University of Rochester, Rochester, NY, USA
ERIC VIGODA, University of California, Santa Barbara, CA, USA

We study the identity testing problem for high-dimensional distributions. Given as input an explicit distribution μ, an $\varepsilon > 0$, and access to sampling oracle(s) for a hidden distribution π, the goal in identity testing is to distinguish whether the two distributions μ and π are identical or are at least ε-far apart. When there is only access to full samples from the hidden distribution π, it is known that exponentially many samples (in the dimension) may be needed for identity testing, and hence previous works have studied identity testing with additional access to various "conditional" sampling oracles. We consider a significantly weaker conditional sampling oracle, which we call the Coordinate Oracle, and provide a computational and statistical characterization of the identity testing problem in this new model.

We prove that if an analytic property known as approximate tensorization of entropy holds for an n-dimensional visible distribution μ, then there is an efficient identity testing algorithm for any hidden distribution π using $\widetilde{O}(n/\varepsilon)$ queries to the Coordinate Oracle. Approximate tensorization of entropy is a pertinent condition as recent works have established it for a large class of high-dimensional distributions. We also prove a computational phase transition: For a well-studied class of n-dimensional distributions, specifically sparse antiferromagnetic Ising models over $\{+1, -1\}^n$, we show that in the regime where approximate tensorization of entropy fails, there is no efficient identity testing algorithm unless RP = NP. We complement our results with a matching $\Omega(n/\varepsilon)$ statistical lower bound for the sample complexity of identity testing in the Coordinate Oracle model.

CCS Concepts: • **Theory of computation** → *Design and analysis of algorithms*; *Machine learning theory*; • **Mathematics of computing** → *Probabilistic inference problems*;

Additional Key Words and Phrases: identity testing, conditional sampling, high-dimensional distributions, entropy

ACM Reference format:
Antonio Blanca, Zongchen Chen, Daniel Štefankovič, and Eric Vigoda. 2024. Complexity of High-Dimensional Identity Testing with Coordinate Conditional Sampling. *ACM Trans. Algor.* 21, 1, Article 7 (November 2024), 58 pages.
https://doi.org/10.1145/3686799

A. Blanca was supported by NSF grant CCF-2143762 and E. Vigoda was supported by NSF grant CCF-2147094.
Authors' Contact Information: Antonio Blanca (corresponding author), Pennsylvania State University, University Park, PA, USA; e-mail: ablanca@cse.psu.edu; Zongchen Chen, Georgia Institute of Technology, Atlanta, GA, USA; e-mail: chenzongchen@gatech.edu; Daniel Štefankovič, University of Rochester, Rochester, NY, USA; e-mail: stefanko@cs.rochester.edu; Eric Vigoda, University of California, Santa Barbara, CA, USA; e-mail: vigoda@ucsb.edu.

ACM 1549-6333/2024/11-ART7
https://doi.org/10.1145/3686799

1 Introduction

A fundamental problem in statistics and machine learning is the identity testing problem (also known as the goodness-of-fit problem). Roughly speaking, we are explicitly given a visible distribution μ and oracle access to samples from an unknown/hidden distribution π; the goal is to determine if these distributions are identical using as few samples from π as possible.

The complexity of identity testing for general distributions is now well-understood; this includes conditions on the visible and hidden distributions which enable efficient identity testing; see [15, 16] for a comprehensive survey. An intriguing line of work considers a different perspective: what additional assumptions *on the sampling oracle* for the hidden distribution are required to ensure efficient identity testing. We present tight results with more modest oracle assumptions than considered previously.

Let us begin with a formal definition of the classical identity testing framework. Let $\mathcal{X}$ be a finite state space of size $N = |\mathcal{X}|$, and let $d(\cdot, \cdot)$ denote a metric or divergence between distributions over $\mathcal{X}$; the standard choices for $d(\cdot, \cdot)$ are **total variation distance (TV distance)** or **Kullback–Leibler divergence (KL divergence)**. For a distribution μ over $\mathcal{X}$ and a parameter $\varepsilon > 0$, denote by ID-TEST$(d, \varepsilon; \mu)$ the identity testing problem for μ: Given as input the full description of the visible distribution μ, and given access to a sampling oracle for an unknown distribution π, our goal is to distinguish between the cases $\pi = \mu$ vs. $d(\pi, \mu) \geq \varepsilon$ with probability at least $2/3$.

For a distribution μ over $\mathcal{X}$, there are efficient identity testing algorithms with sample complexity $O(\sqrt{N}/\varepsilon^2)$ which matches, asymptotically, the information-theoretic lower bound; see [61, 67] for landmark results and [1, 24, 34, 35, 48, 67] for other relevant works. (We recall that the sample or query complexity of an identity testing algorithm is the number of queries it sends to the sampling oracle.)

In practice, data are often high-dimensional, which raises the question of whether identity testing can be solved more effectively for high-dimensional distributions; this will be our focus. To be more precise, let $\mathcal{K} = \{1, \ldots, k\}$ be a label (spin/color) set and let $\mathcal{X} = \mathcal{K}^n$ be a product space of dimension n. We study the identity testing problem ID-TEST$(d, \varepsilon; \mu)$ for n-dimensional distributions μ over $\mathcal{X}$.

Identity testing for high-dimensional distributions has recently attracted some attention; see, e.g., [5–7, 11, 18, 31, 33]. The focus is on visible distributions μ that have a poly(n) size description or parametrization; otherwise one could not hope to design efficient testing algorithms. Such distributions include product distributions (including the uniform distribution), Bayesian nets, and undirected graphical models (also known as spin systems) among others.

The goal is to design identity testing algorithms with poly(n) sample complexity and running times. It is known, however, that identity testing may require a super-polynomial (in n) number of samples [5, 11]. (The algorithms for the general identity testing problem have sample complexity $\Omega(k^{n/2}/\varepsilon^2)$ in the high-dimensional setting since $|\mathcal{X}| = k^n$.)

Consequently, in order to design efficient algorithms, there are two types of further conditions that one may attach to the identity testing problem. The first approach is to restrict the unknown distribution π to be in some particular class of distributions; a natural example is to require that π is from the same class as μ. For example, Bhattacharyya et al. [7] study the setting where both μ and π are product distributions, Canonne et al. and Daskalakis and Pan [18, 33] require μ and π to be Bayesian nets, and Daskalakis et al. [31] study the problem when μ and π are Ising models. More recently, Bhattacharyya et al. [6] consider the case where μ is a product distribution, and π is a Bayesian net. While such an approach leads to fruitful results for testing high-dimensional distributions, it is not ideal from a practical perspective, where π can be, for example, a "noisy" version of μ and may not necessarily belong to a nice class of distributions.

An alternative approach to overcome the apparent intractability of identity testing in the high-dimensional setting is to assume access to stronger sampling oracles from the hidden distribution π; specifically, access to conditional sampling oracles for π (in addition to the sampling oracle for π). This approach for high-dimensional distributions is the focus of this article.

There are several types of conditional sampling oracles, and here we mention the most popular choices. The first is the general conditional sampling oracle—see [19, 23, 37]—which given any subset $\mathcal{X}'$ of the space $\mathcal{X}$ generates a sample from the projection of π to $\mathcal{X}'$; that is, the oracle returns an element x from $\mathcal{X}'$ with probability $\pi(x)/\pi(\mathcal{X}')$. This oracle is not well-suited for the high-dimensional setting because the query subset $\mathcal{X}'$ could be exponentially large in n, and thus one could not hope to formulate the queries to the oracle efficiently (unless restricted to a special class of subsets $\mathcal{X}'$).

The second is the pairwise conditional sampling oracle (Pairwise Oracle) which takes a pair of configurations and generates a sample from the distribution restricted to these two choices: Given $x, y \in \mathcal{X}$ the oracle returns x with probability $\pi(x)/(\pi(x) + \pi(y))$ and y otherwise; see [19]. The queries for Pairwise Oracle can be easily formulated for high-dimensional distributions, and identity testing has been studied in this setting. Recently, [60] provided an identity testing algorithm for the Pairwise Oracle model with $\widetilde{O}(\sqrt{n}/\varepsilon^2)$ sample complexity and a matching statistical lower bound; the $\widetilde{O}$ notation hides poly-logarithmic factors in n and $1/\varepsilon$.

The other conditional oracle previously studied in the high-dimensional setting is the subcube conditional sampling oracle (Subcube Oracle) introduced by Bhattacharyya and Chakraborty [8] and also studied in [17, 25]. A query to the Subcube Oracle consists of a subset $\Lambda \subseteq [n] = \{1, \ldots, n\}$ of variables and a configuration $x \in \mathcal{K}^\Lambda$ on Λ. If $\pi(x) > 0$, the Subcube Oracle returns a sample $x' \in \mathcal{K}^{[n]\setminus\Lambda}$ from the conditional distribution $\pi(\cdot \mid x)$ (see Definition 3.1). For the Subcube Oracle, an identity testing algorithm using $\widetilde{O}(n^2/\varepsilon^2)$ queries was given in [8]; improved algorithms were presented for uniformity testing in [17] and for testing juntas in [25].

In this work, we study identity testing for high-dimensional distributions under a weaker conditional sampling oracle, which we call the Coordinate Oracle. The Coordinate Oracle corresponds to the Subcube Oracle restricted to query sets Λ where $|\Lambda| = n - 1$; that is, we fix the configuration at all but one coordinate and look at the conditional distribution at this particular coordinate given a fixed configuration on the remaining coordinates. Hence, access to the Coordinate Oracle is a much weaker assumption than access to the Subcube Oracle. We also note that the Subcube Oracle model can be significantly harder to simulate. For instance, for the classical ferromagnetic Ising model simulating the Coordinate Oracle is trivial, but sampling conditionally on arbitrary configurations, as required by the Subcube Oracle, is computationally hard [45].

Access to the Coordinate Oracle is also a weaker assumption than access to the Pairwise Oracle in the following sense. When $k = 2$ and $\mathcal{X} = \{0, 1\}^n$, Coordinate Oracle access corresponds to Pairwise Oracle access restricted to pairs of configurations that differ in *exactly one coordinate*. When $k \geq 3$, one can simulate an δ-approximate Coordinate Oracle with Pairwise Oracle access in $\mathrm{poly}(k, \log(1/\delta))$ time (or a perfect one with $\mathrm{poly}(k)$ expected time) using a Markov chain; see Remark 3 for the details. In addition, as in the case of the Subcube Oracle, simulating the Pairwise Oracle can be computationally more demanding than simulating the Coordinate Oracle. For example, in the context of the ferromagnetic Ising model on an n-vertex bounded degree graphs, a query to the Coordinate Oracle will require $O(1)$ random bits, but queries to the Pairwise Oracle may require $\Omega(n)$ random bits.

We provide a computational and statistical characterization of the identity testing problem in the Coordinate Oracle model. Our focus is on imposing no conditions on the hidden distribution π, other than access to Coordinate Oracle, and explore which conditions on the visible distribution μ

are necessary and sufficient for identity testing. We mention that the Coordinate Oracle oracle has already been implicitly used in [17] for uniformity testing (i.e., the special case of testing whether π is the uniform distribution).

Algorithmic Results. For our algorithmic work we consider the identity testing problem under KL divergence, which we denote by $D_{\mathrm{KL}}(\cdot \,\|\, \cdot)$ and is formally defined in Section 3. From an algorithmic perspective, the choice of KL divergence is a natural one since, by Pinsker's inequality, a testing algorithm for ID-TEST$(D_{\mathrm{KL}}(\cdot \,\|\, \cdot), 2\varepsilon^2; \mu)$ yields one for ID-TEST$(d_{\mathrm{TV}}(\cdot,\cdot), \varepsilon; \mu)$ (i.e., for identity testing under TV distance) albeit with potentially sub-optimal sample complexity and running time; the reverse is not true in general.

We start by introducing a key analytic property for the visible distribution, known as *approximate tensorization of entropy* [20], which we will show is a sufficient (and essentially also necessary) condition for efficient identity testing in the high-dimensional setting. Approximate tensorization of entropy roughly states that the entropy of a distribution is bounded by the sum of the average conditional entropies at each coordinate.

Definition 1.1 (Approximate Tensorization of Entropy). A distribution μ fully supported on $\mathcal{K}^n$ satisfies *approximate tensorization of entropy* with constant C if for any distribution π over $\mathcal{K}^n$:

$$D_{\mathrm{KL}}(\pi \,\|\, \mu) \leq C \sum_{i=1}^{n} \mathbb{E}_{x \sim \pi_{n\setminus i}} \Big[D_{\mathrm{KL}}(\pi_i(\cdot \mid x) \,\|\, \mu_i(\cdot \mid x)) \Big], \tag{1}$$

where $\pi_{n\setminus i}(\cdot)$ denotes the marginal distribution of π on $[n] \setminus \{i\}$, and $\pi_i(\cdot \mid x)$ and $\mu_i(\cdot \mid x)$ denote the marginals of π and μ, respectively, on the ith coordinate conditional on x.

The constant C achieves the minimum $C = 1$ when μ is a product distribution. More details about approximate tensorization and equivalent formulations are provided in Section 3.4.

Approximate tensorization of entropy is known to imply optimal mixing times of single-site update Markov chains, known as the Gibbs sampler or Glauber dynamics [22, 28]. It is also used to establish **modified log-Sobolev inequalities (MLSIs)** and the concentration of Lipschitz functions under the distribution [12, 59].

There are a plethora of recent results establishing approximate tensorization in a wide variety of settings. In particular, [27] showed that the spectral independence condition introduced by [3] implies approximate tensorization of entropy for sparse undirected graphical models (i.e., spin systems on bounded degree graphs). Furthermore, recent works showed that spectral independence (and hence approximate tensorization) is implied by certain forms of correlation decay [26, 27, 38], path coupling for local Markov chains [10, 55], and the stability of the partition function [29]. As such, approximate tensorization is now known to hold with constant $C = O(1)$ (independent of n) for a variety of high-dimensional distributions; see, e.g., [10, 29, 30, 40, 41, 55].

We show that approximate tensorization of the visible distribution μ yields an efficient identity testing algorithm, provided access to the Coordinate Oracle and the General Oracle for the hidden distribution π. Access to the General Oracle oracle (i.e., to independent full samples from π) is a standard assumption for testing under conditional sampling oracles. In particular, the General Oracle corresponds to Subcube Oracle restricted to $\Lambda = \emptyset$, so access to the Subcube Oracle implies access to the General Oracle, and previous work under the Pairwise Oracle assumes access to the General Oracle as well.

For our algorithmic result, we have four additional basic assumptions on the visible distribution μ. Specifically, we require that:

(1) μ has a description (parametrization) of poly(n) size;
(2) the Coordinate Oracle can be implemented efficiently for the *visible* distribution μ;

(3) μ is η-*b*alanced: there is a lower bound η so that the conditional probability of any label $a \in \mathcal{K}$ at any coordinate i, fixing any configuration on $[n] \setminus \{i\}$, is at least η (see Section 3.3);
(4) μ is fully supported on $\mathcal{K}^n$.

We discuss these assumptions in detail below (see Remark 1.3). Our algorithmic result for the Coordinate Oracle model follows.

Theorem 1.2. *Given a distribution μ over $\mathcal{X} = \mathcal{K}^n$ satisfying (i)–(iv) and Approximate Tensorization with constant C, there is a testing algorithm for ID-TEST$(D_{\mathrm{KL}}(\cdot \| \cdot), \varepsilon; \mu)$ with* Coordinate Oracle *and* General Oracle *access with $\widetilde{O}(n/\varepsilon)$ sample complexity and polynomial running time.*

We refer the reader to Theorem 4.1 for a more precise theorem statement indicating the explicit dependence on C and η in the sample complexity. See also Section 4.4 for applications of Theorem 1.2 to several well-studied high-dimensional distributions.

We shall see in what follows that our algorithmic result for the Coordinate Oracle model in Theorem 1.2 is tight, both statistically and computationally; that is, we establish a matching $\Omega(n/\varepsilon)$ sample complexity lower bound and show that there is a class of high-dimensional distributions where identity testing is computationally hard in exactly the same settings where approximate tensorization of entropy does not hold. (These results also hold for the easier problem of testing under TV; see Theorems 1.4 and 1.5.)

A surprising feature of our algorithm is that it bypasses sampling from visible distribution μ; it does not even require the concentration of any statistics under μ. As in some of the previous algorithms for high-dimensional testing—e.g., those in [17, 31]—our algorithm starts by "localizing" the testing problem (i.e., reducing it to a one-dimensional setting). For this, we crucially use the Approximate Tensorization of entropy of the visible distribution.

We then consider the problem of testing general (one-dimensional) distributions under KL divergence. It turns out that this problem has been largely overlooked in the literature (the aforementioned known results for identity testing are all under TV distance). This is likely because there are pairs of distributions with infinite KL divergence but arbitrarily small TV distance, and so testing under KL divergence is considered unsolvable in a worst-case sense; see [32]. However, we can aim for algorithms with sample complexities that depend on the visible distribution, i.e., instance-specific bounds instead of worst-case ones, as done in [9, 35, 67].

We provide here an algorithm for the classical identity testing problem (that is, only access to the General Oracle is assumed) for general distributions under KL divergence; the sample complexity of our algorithm depends on the visible distribution (see Lemma 2.1). This is a key technical development toward establishing Theorem 1.2, and one we believe could be of independent interest.

Remark 1.3. We pause now to discuss assumptions (i)–(iv) in Theorem 1.2. As mentioned, condition (i) is necessary as otherwise one can not hope to design testing algorithms with poly(n) running times. Condition (ii) formally states that for any coordinate i, and any fixed assignment σ for the other $n-1$ coordinates, we can compute the conditional distribution at i given σ in polynomial time. This is equivalent to requiring that a step of the Gibbs Sampler Markov chain for μ can be implemented efficiently; we believe (ii) is a mild assumption.

The notion of η-balancedness in condition (iii) is a byproduct of working with KL-divergence and is closely related to other coordinate marginal conditions that are required for efficient learning and sampling; specifically, under the assumption that μ has full support, it is equivalent to the notions δ-biased in [53] and of b-marginally bounded distributions from [10, 29]. Finally, we note that condition (iv) is also a byproduct of working with KL divergence but can be relaxed; we could require instead that the support of π is a subset of the support of μ. We emphasize that these conditions are all for the visible distribution μ, and that we impose no restrictions on the

hidden distribution π (other than oracle access). For example, the uniform distribution, product distributions, and undirected graphical models (e.g., the Ising and hard-core models) satisfy the conditions in Theorem 1.2.

Computational Hardness Results. We show next that the algorithmic result in Theorem 1.2 is computationally tight. In particular, for the anti-ferromagnetic Ising model (defined below), we establish the following computational phase transition for identity testing in the Coordinate Oracle model: (i) when approximate tensorization holds the problem can be solved efficiently, and (ii) when approximate tensorization does not hold, there is no polynomial-time testing algorithm unless RP = NP.

We do not directly prove that identity testing is hard when approximate tensorization fails. We show instead that the same strong correlations that cause approximate tensorization to fail, combined with the hardness of identifying the ground states of the model in the presence of strong correlations, imply the hardness of identity testing. (The ground states are the most likely configurations in the model, and for the anti-ferromagnetic Ising model correspond to the maximum cuts of the graph.)

We introduce the Ising model next, which is the simplest and most well-studied example of an undirected graphical model. Given a graph $G = (V, E)$, the set of configurations of the model is denoted by $\Omega = \{+1, -1\}^V$. For a real-valued parameter β, the probability of a configuration $\sigma \in \Omega$ is given by the Gibbs or Boltzmann distribution:

$$\mu_{G,\beta}(\sigma) = \frac{1}{Z_{G,\beta}} \cdot \exp\left(\beta \sum\nolimits_{\{v,w\} \in E} \sigma_v \sigma_w\right), \tag{2}$$

where the normalizing constant $Z_{G,\beta}$ is known as the partition function. When $\beta > 0$ the model is ferromagnetic/attractive and when $\beta < 0$ then the model is anti-ferromagnetic/repulsive; see Section 4.4.2 for a more general definition of the model.

The anti-ferromagnetic Ising model undergoes an intriguing computational phase transition at the threshold $\beta_c(d) = -\frac{1}{2}\ln(\frac{d}{d-2})$ for the parameter β. This threshold corresponds to the so-called uniqueness/non-uniqueness phase transition on the infinite d-regular tree defined as follows. Let p_ℓ^+ denote the marginal probability that the root of the complete d-regular tree of depth ℓ (i.e., the tree where all internal vertices have degree d and all leaves are on the same level) has label/spin +1 when one fixes the leaves to the all +1 configuration. Similarly, let p_ℓ^- denote the analogous marginal probability for the root to be +1 when the leaves are instead fixed to the all −1 configuration. When $\beta < \beta_c(d)$, then in the limit as $\ell \to \infty$ the two marginals are the same, i.e., $\lim_{\ell\to\infty} p_\ell^+ = \lim_{\ell\to\infty} p_\ell^-$; this is known as the *tree uniqueness region* since it implies that there is a unique Gibbs distribution for the infinite d-regular tree. On the other hand, when $\beta > \beta_c(d)$ then the limits are different; this is called the *tree non-uniqueness region* as there are multiple Gibbs distributions for the infinite d-regular tree. A key consequence for general graphs is the following rough statement: In graphs of maximum degree at most d, when $\beta < \beta_c(d)$ long-range correlations die off, whereas when $\beta > \beta_c(d)$ long-range correlations persist and mark the onset of hardness for several computational problems (e.g., counting and sampling) on graphs of degree at most d; see [63] for further details.

For constant $d \geq 3$ and all $0 > \beta > \beta_c(d)$, the approximate sampling and counting (i.e., approximating the partition function $Z_{G,\beta}$) problems can be solved efficiently on any graph of maximum degree d [28]. Moreover, approximate tensorization holds in this regime, and hence Theorem 1.2 applies for identity testing in the Coordinate Oracle model. In contrast, it is also known that when $\beta < \beta_c(d)$ there are no polynomial-time approximate sampling or counting algorithms unless RP = NP [42, 65].

We establish here the computational hardness of identity testing in the Coordinate Oracle model in the same parameter regime $\beta < \beta_c(d)$, which thereby exhibits a similar computational phase transition for identity testing for the class of anti-ferromagnetic Ising models.

Theorem 1.4. *For sufficiently large constant $d \geq 3$ and constant $\beta < 0$, consider identity testing for the family of anti-ferromagnetic Ising models on n-vertex graphs of max degree d with parameter β.*

(i) If $\beta > \beta_c(d)$, then there exists a polynomial-time algorithm for identity testing under KL divergence with access to the Coordinate Oracle *and the* General Oracle *with sample complexity $\widetilde{O}(n/\varepsilon)$;*

(ii) If $\beta < \beta_c(d)$, then there is no polynomial-time algorithm for identity testing under TV distance (and hence under KL divergence) with access to the Coordinate Oracle *and the* General Oracle *unless* RP = NP.

There are few analogous computational hardness results for identity testing; most lower-bound results in this setting are information-theoretic. The few examples appeared in [5, 11], and these earlier results apply to the identity testing problem with access only to General Oracle and require both the hidden and visible models to be Ising models. In our current setting, the visible model is an Ising model, but the hidden is an arbitrary high-dimensional distribution. This is a significant conceptual difference, and the techniques from [5, 11] do not easily extend (see Remark 2.2).

At a high level, as in [5], we prove the hardness result in Theorem 1.4 (ii) using a reduction from the maximum cut problem. That is, given a graph $G = (V, E)$, we construct a testing instance that if solved, would find the maximum cut of G. In this approach, constructing a testing instance of small degree is a key challenge, and the "degree reducing" gadgets from [5, 11] no longer work in our setting.

Instead, we use a gadget introduced in [64] to establish the computational hardness of approximate counting anti-ferromagnetic spin systems. An interesting technical aspect of our proof is that we are required to design polynomial-time sampling algorithms to simulate the hidden oracles. This is difficult for us because sampling anti-ferromagnetic Ising models throughout the non-uniqueness regime, i.e., for all $\beta < \beta_c(d)$, is a notoriously hard problem (the problem is NP-hard even for regular graphs). We manage to design efficient sampling algorithms for our testing instances using the recent algorithmic result of Koehler et al. [54] that give an approximate sampling algorithm for Ising models when the edge interaction matrix has low rank, in conjunction with the sampling methods from [50] that use polymer models. A detailed overview of our reduction is given in Section 2.2. We mention that in the reductions in [5, 11], sampling is trivial, since there it is assumed that $\beta \ll \beta_c(d)$ (specifically, $|\beta|d = \Omega(\log n)$) and the instance is bipartite, so the Gibbs distribution concentrates in the configurations that align with the bi-partition; see Remark 2.2 for a detailed account of the novelties in our reduction to establish Theorem 1.4 (ii).

Finally, we mention that the hardness result in Theorem 1.4 (ii) extends to *any* conditional sampling oracle that could be implemented in polynomial time for the anti-ferromagnetic Ising model, and thus applies to identity testing in the Pairwise Oracle model, complementing the algorithmic results from [19, 60]. On the other hand, they do not extend to the Subcube Oracle model since we do not know how to simulate this oracle efficiently.

Statistical Lower Bounds. We present next an information-theoretic lower bound for identity testing problem in the Coordinate Oracle model that matches the sample complexity of our testing algorithm for this model (Theorem 1.2). Our lower bound is for the special case of uniformity testing under TV distance when $k = 2$, i.e., the visible distribution is the uniform distribution over $\{0, 1\}^n$.

Theorem 1.5. *Let μ be the uniform distribution over $\{0,1\}^n$. Then, any algorithm for* ID-TEST $(d_{\mathrm{TV}}(\cdot,\cdot),\varepsilon;\mu)$ *with access to both the* Coordinate Oracle *and the* General Oracle *requires* $\Omega(n/\varepsilon^2)$ *samples.*

A direct corollary of this result is that solving the identity testing problem under KL divergence requires $\Omega(n/\varepsilon)$ samples in the Coordinate Oracle model, thus showing that the sample complexity of our algorithm in Theorem 1.2 is asymptotically tight (up to logarithmic in n and $1/\varepsilon$ factors).

Our proof of Theorem 1.5 follows a well-known strategy. We construct a family of "bad" distributions $\mathcal{B}$, each of which has TV distance at least ε from the uniform distribution μ over $\{0,1\}^n$. The lower bounds follow from the fact that, for this carefully constructed family $\mathcal{B}$, one can not distinguish between sequences of independent samples from μ or from a distribution π chosen uniformly at random from $\mathcal{B}$. However, since our setting is adaptive, i.e., the choice of conditional queries of the testing algorithm may depend on the output to previous ones, we need to consider *query histories*, as in [19, 60]. (Roughly speaking, a query history is a sequence of queries that the testing algorithm sends the oracle along with the outputs from the oracle.) To show that two query histories are indistinguishable (under μ or π), we use ideas from [19] and the so-called *hybrid argument* in cryptography; see [46].

New Results for the Subcube Oracle Model. While the main focus of this work is the study of identity testing under weaker oracle assumptions (i.e., the Coordinate Oracle model), we also provide new results for identity testing in the previously studied Subcube Oracle model. Our first results for this model is an improved identity testing algorithm.

Theorem 1.6. *Let μ be an η-balanced distribution fully supported on $\mathcal{K}^n$ that has a* $\mathrm{poly}(n)$ *size parameterization. If we can compute the marginal probability at any coordinate conditioned on any partial configuration on any subset of coordinates, then there is an identity testing algorithm for ID-TEST$(D_{\mathrm{KL}}(\cdot\,\|\,\cdot),\varepsilon;\mu)$ for the* Subcube Oracle *model with $\widetilde{O}(n/\varepsilon)$ sample complexity and running time that depends on the time it takes to compute the coordinate conditional marginals.*

This algorithm, compared to the best-known algorithm for the Subcube Oracle model in [8], additionally requires that μ is η-balanced, but improves the sample complexity significantly from $\widetilde{O}(n^2/\varepsilon^2)$ to $\widetilde{O}(n/\varepsilon)$. In addition, compared to Theorem 1.2, this result for the Subcube Oracle does *not* require approximate tensorization of entropy. In fact, we point out several relevant settings where Theorem 1.6 applies, but approximate tensorization fails (or we do not know if it holds) and hence Theorem 1.2 does not apply: undirected graphical models (e.g., Ising model) on trees, Bayesian networks, mixtures of product distributions, and high-temperature Ising models and monomer-dimer models (i.e., weighted matchings) on arbitrary graphs. We remark that, similar to Theorem 1.2, Theorem 1.6 also holds under the weaker assumption that the support of μ contains the support of π; see Theorems 7.2 and 7.5 for more details.

We also provide a matching lower bound for uniformity testing in the Subcube Oracle.

Theorem 1.7. *Let μ be the uniform distribution over $\{0,1\}^n$. Then, any algorithm for* ID-TEST $(D_{\mathrm{KL}}(\cdot\,\|\,\cdot),\varepsilon;\mu)$ *with access to the* Subcube Oracle *requires* $\Omega(n/\varepsilon)$ *samples.*

Note that when μ is the uniform distribution, then $\eta = \Theta(1)$, so Theorems 1.6 and 1.7 provide asymptotically matching sample complexity bounds for identity testing under KL divergence. Interestingly, if one considers uniformity testing under TV distance and Subcube Oracle access, then the recent work [17] shows that $\widetilde{O}(\sqrt{n}/\varepsilon^2)$ oracle queries suffice. As far as we know, it is unclear if the testing algorithm from [17] with sublinear sample complexity can be used for other high-dimensional distributions, e.g., general product distributions.

Our last result concerns *tolerant* identity testing in the Subcube Oracle model. In this problem, the goal is to distinguish between the cases $D_{\mathrm{KL}}(\pi\,\|\,\mu) \le \delta$ and $D_{\mathrm{KL}}(\pi\,\|\,\mu) \ge \delta+\varepsilon$ for $\delta,\varepsilon > 0$;

identity testing corresponds to $\delta = 0$. We show that, under the same assumptions as in Theorem 1.6, one can estimate $D_{\mathrm{KL}}(\pi \,\|\, \mu)$ within additive error ε using $\widetilde{O}(n^4/\varepsilon^4)$ queries to the Subcube Oracle.

Theorem 1.8. *Let μ be an η-balanced visible distribution fully supported on $\mathcal{K}^n$ that has a* $\mathrm{poly}(n)$ *size parametrization. Suppose we can compute the marginal probability for μ at any coordinate conditioned on any partial configuration on a subset of coordinates. Given access to the* Subcube Oracle *for a hidden distribution π, there is an algorithm that for any $\varepsilon > 0$ computes $\widehat{S}$ such that, with probability at least $2/3$, we have $|\widehat{S} - D_{\mathrm{KL}}(\pi \,\|\, \mu)| \le \varepsilon$. The sample complexity of the algorithm is $\widetilde{O}\left(n^4/\varepsilon^4\right)$. The running time of the algorithm depends on the time it takes to compute the coordinate conditional marginals.*

2 Overview of Techniques

We present proof overviews for our main results in the Coordinate Oracle model: our testing algorithm (Theorem 1.2), the computational hardness (Theorem 1.4 (ii)), and the lower bound (Theorem 1.5).

2.1 Algorithmic Result for Coordinate Oracle Model: Theorem 1.2

Suppose μ is the visible distribution and let π be an arbitrary distribution over $\mathcal{K}^n$. If approximate tensorization of entropy holds for μ with constant C, the following holds:

$$D_{\mathrm{KL}}(\pi \,\|\, \mu) \le Cn\, \mathbb{E}_{(i,x)}\left[D_{\mathrm{KL}}\left(p_i^x \,\|\, q_i^x\right)\right],$$

where $i \in [n]$ is a uniformly random coordinate, $x \in \mathcal{K}^{n\setminus i}$ is generated from the marginal distribution $\pi_{n\setminus i}$ of π on $[n] \setminus \{i\}$, $p_i^x = \pi_i(\cdot \mid x)$, and $q_i^x = \mu_i(\cdot \mid x)$ (see Definition 1.1). Therefore, to distinguish between the cases $\pi = \mu$ and $D_{\mathrm{KL}}(\pi \,\|\, \mu) \ge \varepsilon$, it suffices to distinguish between

$$p_i^x = q_i^x \text{ for all pairs } (i,x) \quad \text{vs.} \quad \mathbb{E}_{(i,x)}\left[D_{\mathrm{KL}}\left(p_i^x \,\|\, q_i^x\right)\right] \ge \frac{\varepsilon}{Cn}.$$

This is the first step toward localizing the testing problem to a single coordinate. Now, under the η-balanced assumption for μ, we have that $0 \le D_{\mathrm{KL}}\left(p_i^x \,\|\, q_i^x\right) \le \ln(1/\eta)$. Hence, if $\mathbb{E}_{(i,x)}\left[D_{\mathrm{KL}}\left(p_i^x \,\|\, q_i^x\right)\right] \ge \frac{\varepsilon}{Cn}$, one can show via a reverse Markov inequality that there exists an integer $\ell \ge 1$, such that $2^\ell = O(n)$ and

$$\Pr_{(i,x)}\left(D_{\mathrm{KL}}\left(p_i^x \,\|\, q_i^x\right) \ge 2^\ell \cdot \frac{\varepsilon}{2Cn}\right) \ge \frac{1}{2^\ell D}, \tag{3}$$

where $D = \Theta\big(\log(\frac{n\cdot\log(1/\eta)}{\varepsilon})\big)$; see Lemma 4.2 for a precise statement.

With (3), it is not difficult to find a pair (i,x) such that $D_{\mathrm{KL}}\left(p_i^x \,\|\, q_i^x\right) \ge 2^\ell \cdot \frac{\varepsilon}{2Cn}$. This can done by first generating $O(\mathrm{poly}(n) \cdot D)$ pairs (i_t, x_t), by choosing $i_t \in [n]$ uniformly at random and then using the General Oracle to sample the partial configuration x_t on $[n] \setminus \{i_t\}$. Then, we can exhaustively check for each ℓ (note that $\ell = O(\log n)$) whether among the generated pairs (i_t, x_t)'s there is one, say (i,x), satisfying that $D_{\mathrm{KL}}\left(p_i^x \,\|\, q_i^x\right) \ge 2^\ell \cdot \frac{\varepsilon}{2Cn}$. By (3), this will likely be the case.

In conclusion, we reduce identity testing for μ to solving identity testing for the one-dimensional distributions $p := p_i^x$ and $q := q_i^x$ on a domain of size k with respect to KL divergence. We assume q can be computed for the visible distribution μ (this is assumption (iii) in Theorem 1.2), and we have access to a sampling oracle for p_i^x using the Coordinate Oracle for π.

As mentioned earlier, in the distribution testing literature, testing under KL divergence has been overlooked. This is because there are pairs of distributions with infinite KL divergence but arbitrarily small TV distance, which entails that identity testing requires arbitrarily many samples even though the KL divergence is arbitrarily large. For example, this happens when q is the distribution on a single point 0 and p is the Bernoulli distribution with arbitrarily small mean [32]. As such, identity

testing under KL divergence has been considered unsolvable in the sense of worst-case sample complexity for arbitrary p and q.

However, the identity testing problem under KL divergence makes perfect sense for specific visible distributions q if we are interested in the instance-specific sample complexity instead of the worst-case one, as in [67] under TV distance. Namely, for a given distribution q, what is the number of samples required, potentially depending on q, for the identity testing problem for q under KL divergence? We give next a first attempt at solving this problem. The sample complexity of our testing algorithm depends on the minimum probability $\eta = \min_{a\in\mathcal{K}} q(a)$.

Lemma 2.1. *Let $k \in \mathbb{N}^+$ and let $\varepsilon > 0$, $\eta \in (0, 1/2]$. Given a visible distribution q over domain $\mathcal{K}$ of size k such that $q(a) \geq \eta$ for any $a \in \mathcal{K}$, and given sample access to an unknown distribution p over $\mathcal{K}$, there exists a polynomial-time identity testing algorithm that distinguishes with probability at least $2/3$ between the cases $p = q$ or $D_{\mathrm{KL}}(p \,\|\, q) \geq \varepsilon$. The sample complexity of the identity testing algorithm is $O\big(\min\big\{\frac{1}{\varepsilon\sqrt{\eta}}, \frac{\sqrt{k}\ln(1/\eta)}{\varepsilon^2}\big\}\big)$ for $k \geq 3$ and $O\big(\frac{\ln(1/\eta)}{\varepsilon}\big)$ for $k = 2$.*

We remark that the dependency on η in the sample complexity is inevitable; see Remark 4.4.

A natural first approach to identity testing under KL divergence to prove Lemma 2.1 is a reduction to testing under TV distance via the so-called reversed Pinsker's inequality: $D_{\mathrm{KL}}(p \,\|\, q) \leq (2/\eta) d_{\mathrm{TV}}(p, q)^2$ (see Lemma 4.5). The sample complexity of such algorithm is $O(\sqrt{k}/(\varepsilon\eta))$. This is not optimal, for example, if q is the uniform distribution over $\mathcal{K}$ one has $\eta = 1/k$ and so the sample complexity is $O(k^{3/2}/\varepsilon)$, but one would expect the sample complexity to be $O(\sqrt{k}/\varepsilon)$, by analogy to what happens for testing in TV distance. A better reduction is to testing under ℓ_2 distance via the inequality: $D_{\mathrm{KL}}(p \,\|\, q) \leq (1/\eta)\,\|p - q\|_2^2$ (see Lemma 4.5). We then need to distinguish between $p = q$ and $\|p - q\|_2 \geq \sqrt{\varepsilon\eta}$, which allows us to apply results from [35] and obtain an algorithm with sample complexity of $O(\|q\|_2/(\varepsilon\eta))$; this time we get the $O(\sqrt{k}/\varepsilon)$ sample complexity bound when q is the uniform distribution.

However, two major challenges ought to be solved for this approach to work. First, while $\|q\|_2$ can be bounded for certain specific distributions q (e.g., the uniform distribution), in general we do not have a bound for $\|q\|_2$. This can be solved via the *flattening* method from [35] which, roughly speaking, constructs a new testing instance (i.e., distributions p' and q' over $\mathcal{K}'$) that is equivalent to the initial one with the additional property that $\|q'\|_2$ is small. The idea is to divide "heavy" elements (those $a \in \mathcal{K}$ with large density $q(a)$) into many copies so that $q'(a') \approx 1/\ell$ for all $a' \in \mathcal{K}'$ where $\ell = |\mathcal{K}'|$, i.e., q' is close to uniform.

The second challenge is that, even if when $\|q\|_2$ small, the dependency on η could still be inverse-polynomial. This is particularly problematic when η decays sharply as k grows, e.g., $\eta = 2^{-k}$. To overcome this, we divide the domain $\mathcal{K}$ into two parts, those with small density $q(a) < \zeta$ and those with large density $q(a) \geq \zeta$ for some parameter ζ. We then deal with the two parts separately by running different testing algorithms on each. This can be viewed as a simple application of the *bucketing* technique from [4] using only two buckets.

While flattening and bucketing were previously known, the novelty of our approach is to combine them to get a stronger bound for the sample complexity, specifically by selecting the right scale ℓ for flattening and the right threshold ζ for bucketing. Our bound achieves $O(\sqrt{k}/\varepsilon)$ for q with $\eta = \Theta(1/k)$ such as the uniform distribution, and also maintains a $\sqrt{k}$ dependency even for biased q of tiny η, with only a logarithmic dependency on $1/\eta$. For details see Lemmas 4.3 and 4.10.

2.2 Computational Hardness in the Coordinate Oracle Model: Theorem 1.4 (ii)

We establish hardness of the identity testing problem as stated in Theorem 1.4 (ii) for the anti-ferromagnetic Ising model with Coordinate Oracle and General Oracle access via a reduction from

the maximum cut problem. Let $\{G = (V_G, E_G), k\}$ be an instance of the maximum cut problem. That is, we want to check whether $\textsf{max-cut}(G) < k$ or $\textsf{max-cut}(G) \geq k$. In our reduction, we construct an identity testing instance for the anti-ferromagnetic Ising model, feed it as input to a presumed testing algorithm, and claim that the output of the algorithm solves $\{G = (V_G, E_G), k\}$ with probability at least 2/3; this is not possible unless RP = NP.

We start by constructing the multi-graph $F = (V_F, E_F)$ by adding two special vertices s and t to G, i.e., $V_F = V_G \cup \{s, t\}$. These two vertices are connected with $N^2 - k$ edges, where $N = |V_G|$. We also add N edges between s and each vertex of V_G and do the same for t so that:

(1) When $\textsf{max-cut}(G) < k$, the cut $(\{s, t\}, V_G)$ of size $2N^2$ is the unique maximum cut of F;
(2) When $\textsf{max-cut}(G) \geq k$, there exists another cut in F, other than $(\{s, t\}, V_G)$, of size $\geq 2N^2$.

This is because for $S \subset V_G$, the cut $(S \cup \{s\}, V_G \setminus S \cup \{t\})$ of F will have size: $\textsf{max-cut}(G) + |S|N + |V_G \setminus S|N + N^2 - k = 2N^2 + \textsf{max-cut}(G) - k$, which is $\geq 2N^2$ only when $\textsf{max-cut}(G) \geq k$.

We consider the anti-ferromagnetic Ising model on F. There is a natural bijection between the cuts of F and the configurations of the Ising model. In particular, each cut $(S, V_F \setminus S)$ of F corresponds to exactly two Ising configurations: Vertices in S are assigned +1 and those in $V_F \setminus S$ are assigned -1 (and vice versa). From the definition of the model (see (2)) we also see that the "ground states" of the anti-ferromagnetic Ising model on F, that is the configurations of maximum probability in the Gibbs distribution, correspond precisely to the maximum cuts of F.

Let Ω be the set of all cuts of F and let Ω_0 be the set of all cuts $(S, V_F \setminus S)$ of F *except* those where $s \in S$, $t \in V_F \setminus S$, and the corresponding cut for G, i.e., $(S \setminus \{s\}, V_F \setminus \{S, t, s\})$, has size $\geq k$. This way, if $\textsf{max-cut}(G) < k$, then $\Omega_0 = \Omega$, and if $\textsf{max-cut}(G) \geq k$, then $\Omega \setminus \Omega_0$ contains the cuts of F corresponding to cuts of G of size $\geq k$.

We set the visible distribution of our testing instance to be the Gibbs distribution $\mu_{F,\beta}$ of the anti-ferromagnetic Ising model on F with $\beta < \beta_c(d) < 0$ in the tree non-uniqueness region. The hidden distribution will be $\mu_{F,\beta}(\cdot \mid \Omega_0)$, that is, $\mu_{F,\beta}$ conditioned on configurations that correspond to cuts in Ω_0. Our construction ensures that if $\textsf{max-cut}(G) < k$, then $\Omega = \Omega_0$ and so $\mu_{F,\beta}(\cdot \mid \Omega_0) = \mu_{F,\beta}$. Moreover, when $\textsf{max-cut}(G) \geq k$, we have $\Omega \neq \Omega_0$ and $\mu_{F,\beta}(\cdot \mid \Omega_0) \neq \mu_{F,\beta}$. In fact, it can be shown that the TV distance between $\mu_{F,\beta}(\cdot \mid \Omega_0)$ and $\mu_{F,\beta}$ is $1 - o(1)$; intuitively, this is because $\Omega \setminus \Omega_0$ contains large cuts of F that account for a non-trivial portion of the probability mass of $\mu_{F,\beta}$.

Our reduction is then completed by generating samples from $\mu_{F,\beta}(\cdot \mid \Omega_0)$ and giving these samples and $\mu_{F,\beta}$ to the identity testing algorithm as input. The testing algorithm is guaranteed to succeed with probability at 2/3. If the algorithm detects that the samples did not come from $\mu_{F,\beta}$, it means that $\textsf{max-cut}(G) \geq k$; otherwise, it means that $\textsf{max-cut}(G) < k$. Hence, we have a polynomial running time algorithm that solves the maximum cut problem with probability at least 2/3, which is not possible unless RP = NP.

There are two important complications in this approach. First, F is a multi-graph of unbounded degree, and our goal is to establish hardness for the class of anti-ferromagnetic Ising models graphs of maximum degree $d = O(1)$ when $\beta < \beta_c(d)$. Second, we do not know how to generate samples from $\mu_{F,\beta}(\cdot \mid \Omega_0)$ efficiently in polynomial time.

Let us address first how we solve the issue of F being a multi-graph with large maximum degree. For this, we use a "degree reducing" gadget; the one we use was introduced in [64] to establish the hardness of approximate counting and sampling anti-ferromagnetic spin systems. Specifically, each vertex of F is replaced by a gadget H which consists of a (nearly) d-regular random bipartite graph with a relatively small number of trees attached to it. Being more precise, the leaves of each tree will be identified with unique vertices on the same side of the bipartite graph; see Section 5.1

for the precise construction. The roots of these trees are called *ports* and are used to connect the gadgets as dictated by the edges of F. This results in a simple d-regular graph $\widehat{F}$.

A key feature of the gadget H is that in the tree non-uniqueness region $\beta < \beta_c(d)$, a sample from $\mu_{H,\beta}$ will have mostly +1's on one side of H and mostly -1's on the other, or vice versa. Hence there are two possible "phases" for the gadget which we use to simulate the spin of the corresponding vertex in F, i.e., the phase of the gadget is mapped to the spin of the corresponding vertex of F. Therefore, in a configuration in $\widehat{F}$, the phase of all the gadgets determines a cut for F, and thus one for G. Consequently, the reduction described above from the maximum cut problem to identity testing using F can be done using $\widehat{F}$ instead.

The second technical complication is that we are required to sample from $\mu_{\widehat{F},\beta}(\cdot \mid \Omega_0)$. For this, we observe first that sampling a phase assignment from $\mu_{\widehat{F},\beta}(\Omega_0)$ is straightforward (see Lemma 5.3). We then sample the port configuration given the phase vector from Ω_0. This is done via a rejection sampling procedure by noting that the marginal distribution on the ports is within $o(1)$ TV distance of a suitably defined product distribution. Once the port configuration is sampled within the desired accuracy, we sample the configuration on each gadget (independently) given the configuration of the ports. For this we use a hybrid approach: We use the recent algorithm from [54] for low-rank Ising models for one range of values of β (i.e., when $|\beta|\sqrt{d} = O(1)$) and polymer models—see [50]—for the other. To use these algorithms, we fleshed out the spectrum of the incidence matrix of the gadget. Note that simulating the Coordinate Oracle for the Ising model is straightforward as the spin probability is a function of the number of neighboring +1 and -1 spins.

Remark 2.2. In [5], hardness of identity testing was established when both the visible and hidden distributions are anti-ferromagnetic Ising models on graphs of bounded degree also via a reduction from the maximum cut problem. As such, we believe it is meaningful to detail the conceptual and technical differences, as well as some similarities, between the reduction described above and the one from [5]. Conceptually, in [5] the hidden distribution π is assumed to be from the same class as μ, so the testing problem in consideration is easier. In fact, this problem is not hard for all $\beta < \beta_c(d) < 0$ since when $|\beta|d = O(\log n)$ it can be solved using the learning algorithm from [53] to learn π. Only when $|\beta|d = \Omega(\log n)$, this variant of identity testing becomes computationally hard, and this is precisely what is established in [5]. Our goal here is to show hardness throughout the entire non-uniqueness regime $\beta < \beta_c(d)$ (not only for $|\beta|d = \Omega(\log n)$), so the hidden distribution in our reduction can not be an Ising model. Our hidden distribution $\mu_{\widehat{F},\beta}(\cdot \mid \Omega_0)$ is instead a conditional anti-ferromagnetic Ising distribution, and, as noted, sampling from it is challenging.

At a technical level, the necessary assumption in [5] that $|\beta|d = \Omega(\log n)$ simplifies matters significantly. In particular, the degree reducing gadgets there simply consist of random regular bipartite graphs; when $\beta d = \omega(\log n)$, sampling from the anti-ferromagnetic Ising model on these gadgets is trivial since $1 - o(1)$ of the probability mass is concentrated on two trivial configurations (+1 in one side of the bipartite graph, -1 in the other side and vice versa). When $\beta < \beta_c(d)$, the correlations in the model are super-polynomially weaker, i.e., there is no such strong concentration in the ground states. As such, we must use a more sophisticated degree-reducing gadget (the one from [64] as discussed earlier), and consider the phase of the gadget to simulate spin assignments to vertices. In terms of similarities, the construction of the multi-graph F from the max-cut instance detailed above is nearly identical to the construction in [5], i.e., F is essentially the same, but $\widehat{F}$ is not since we must use a different gadget.

3 Preliminaries

In this section we gather a number of standard definitions and results that we will refer to in our proofs. Let $k, n \in \mathbb{N}^+$ be integers. Let $\mathcal{K} = \{1, \ldots, k\}$ denote a finite alphabet set of size k, and

let π be an arbitrary distribution over $\mathcal{K}^n$. Throughout the article, we use n in the subscript and superscript to represent the set $[n] = \{1, \ldots, n\}$ and use $n \setminus i$ to represent the set $[n] \setminus \{i\}$ to ease the notation.

3.1 Coordinate Conditional Sampling Oracle

We recall next the formal definitions of the various sampling oracles discussed in the article.

Definition 3.1. The sampling oracles for the hidden distribution π are defined as follows:

- *General Sampling Oracle* (General Oracle): Generate a sample x from π.
- *Coordinate Conditional Sampling Oracle* (Coordinate Oracle): Given $i \in [n]$ and $x \in \mathcal{K}^{n\setminus i}$ as inputs to the oracle:
  - If $\pi(X_{n\setminus i} = x) > 0$, the oracle samples $a \in \mathcal{K}$ from the conditional marginal distribution $\pi(X_i = \cdot \mid X_{n\setminus i} = x)$;
  - If $\pi(X_{n\setminus i} = x) = 0$, the oracle outputs $a \in \mathcal{K}$ arbitrarily.
- *Subcube Conditional Sampling Oracle* (Subcube Oracle): Given $\Lambda \subseteq [n]$ and $x \in \mathcal{K}^{\Lambda}$ as inputs to the oracle:
  - If $\pi(X_\Lambda = x) > 0$, the oracle samples $x' \in \mathcal{K}^{[n]\setminus\Lambda}$ from the conditional distribution $\pi(X_{V\setminus\Lambda} = \cdot \mid X_\Lambda = x)$;
  - If $\pi(X_\Lambda = x) = 0$, the oracle outputs $x' \in \mathcal{K}^{V\setminus\Lambda}$ arbitrarily.
- *Pairwise Conditional Sampling Oracle* (Pairwise Oracle): Given $x, y \in \mathcal{K}^n$, the oracle returns x with probability $\pi(x)/(\pi(x) + \pi(y))$ and y otherwise.

We provide next two brief remarks noting that access to a Coordinate Oracle is a weaker assumption than access to a Pairwise Oracle or a Subcube Oracle.

Remark 3.2. Pairwise Oracle is generally a stronger oracle than Coordinate Oracle. When $k = 2$ and the state space is the binary hypercube $\{0, 1\}^n$, this is obvious since the Coordinate Oracle essentially generates samples conditioned in the set $\{x, y\}$ where x and y differ in exactly one coordinate, while Pairwise Oracle can handle any pair vectors $x, y \in \mathcal{K}^n$. If $k \geq 3$ is a constant (independent of n), then one can also simulate an ε-approximate Coordinate Oracle with Pairwise Oracle access in $\mathrm{poly}(k, \log(1/\varepsilon))$ time (or a perfect one with $\mathrm{poly}(k)$ expected time) using a Markov chain. Given a query (i, x) where $i \in [n]$ and $x \in \mathcal{K}^{n\setminus i}$, to generate a random value at the coordinate i conditional on x, one can simulate the Markov chain that in each step picks an element $a \in \mathcal{K}$ uniformly at random and lets $a_{t+1} = a$ with probability $\mu(x_{i,a})/(\mu(x_{i,a}) + \mu(x_{i,a_t}))$ and $a_{t+1} = a_t$ otherwise; $x_{i,a}$ denotes the vector with the ith coordinating being a and all other coordinates given by x. Every step of the Markov chain can be perfectly implemented with the Pairwise Oracle, and a simple coupling argument shows that the ε-mixing time is $\mathrm{poly}(k, \log(1/\varepsilon))$. For perfect sampling with $\mathrm{poly}(k)$ expected time, one can use the Coupling from the Past Method; see [62].

Remark 3.3. The Subcube Oracle subsumes the Coordinate Oracle + General Oracle combination implying that:

- Algorithms with both Coordinate Oracle + General Oracle access give algorithms with Subcube Oracle access
- Lower bounds for the Subcube Oracle model imply lower bounds the Coordinate Oracle + General Oracle.

3.2 Identity Testing

We provide next the formal definition of the identity testing problem for a distribution μ over $\mathcal{K}^n$. Let d be any metric or divergence for distributions over $\mathcal{K}^n$.

ID-TEST$(d, \varepsilon; \mu)$.
Input: Description of a distribution μ over $\mathcal{K}^n$.
Provided: Access to Coordinate Oracle + General Oracle for an unknown distribution π over $\mathcal{K}^n$.
Goal: Determine whether $\pi = \mu$ or $d(\pi, \mu) \geq \varepsilon$.

Let $\mathcal{F}$ denote a family of distributions (with varying dimensions), each of which is supported on $\mathcal{K}^n$ for some integer $n \in \mathbb{N}^+$ and can be represented with poly(n) parameters. We say an algorithm $\mathcal{A}$ is an identity testing algorithm for the family $\mathcal{F}$ if for every $\mu \in \mathcal{F}$ it solves ID-TEST$(d, \varepsilon; \mu)$ with probability at least $2/3$. Note that the unknown distribution π does not necessarily belong to the family $\mathcal{F}$.

3.3 Coordinate Balancedness and Marginal Boundedness

We say a distribution μ supported on $\mathcal{K}^n$ is η-*balanced*, if for every $i \in [n]$, every $x \in \mathcal{K}^{n\setminus i}$ with $\mu(X_{n\setminus i} = x) > 0$, and every $a \in \mathcal{K}$, one has

$$\text{either } \mu(X_i = a \mid X_{n\setminus i} = x) = 0, \quad \text{or } \mu(X_i = a \mid X_{n\setminus i} = x) \geq \eta.$$

On the other hand, we say the distribution μ is b-*marginally bounded* if for every $\Lambda \subseteq [n]$, every $x \in \mathcal{K}^\Lambda$ with $\mu(X_\Lambda = x) > 0$, every $i \in [n] \setminus \Lambda$, and every $a \in \mathcal{K}$, one has

$$\text{either } \mu(X_i = a \mid X_\Lambda = x) = 0, \quad \text{or } \mu(X_i = a \mid X_\Lambda = x) \geq b.$$

Note that marginal boundedness is a generalization of coordinate balance, and in particular, if in the definition of b-marginally bounded one restricts to Λ where $|\Lambda| = n - 1$ then we obtain b-balanced. Hence, any b-marginally bounded distribution is also b-balanced. Moreover, if μ has full support, then both notions are equivalent. See also Remark 7.3 for a weaker version of marginal boundedness. Related notions to marginal boundedness appeared in [28, 53].

3.4 Approximate Tensorization of Entropy

Let μ be a distribution over $\mathcal{K}^n$. For any non-negative function $f : \mathcal{K}^n \to \mathbb{R}_{\geq 0}$, the expectation of f is defined to be $\mu(f) = \sum_{x\in\mathcal{K}^n} \mu(x) f(x)$, and the (relative) entropy of f is defined as

$$\mathrm{Ent}_\mu(f) = \mu(f \ln f) - \mu(f) \ln(\mu(f)),$$

with the convention that $0 \ln 0 = 0$.

Given a coordinate i and a partial configuration $x \in \mathcal{K}^{n\setminus i}$ on all coordinates but i, one can define the entropy of the function f with respect to the conditional distribution $\mu_i(\cdot \mid x)$, which we denote by $\mathrm{Ent}_i^x(f)$. Furthermore, we regard $\mathrm{Ent}_i^x(f)$ as a function of x and its expectation, when x is generated from $\mu_{n\setminus i}$, is denoted as $\mu[\mathrm{Ent}_i(f)]$. We are now ready to give the formal definition of approximate tensorization of entropy in the functional inequality form, as in [20–22].

Definition 3.4 (Approximate Tensorization of Entropy: Functional Form). We say that a distribution μ over $\mathcal{K}^n$ satisfies approximate tensorization of entropy with constant C if for any non-negative function $f : \mathcal{K}^n \to \mathbb{R}_{\geq 0}$ one has

$$\mathrm{Ent}(f) \leq C \sum_{i=1}^{n} \mu[\mathrm{Ent}_i(f)]. \tag{4}$$

As mentioned in the introduction, approximate tensorization is an important tool for proving functional inequalities like the MLSI; it is also useful for deriving optimal mixing time bounds for the Glauber dynamics. Although it is most often stated in this functional inequality form, mainly because of several useful analytic properties, in this article we will consider its probabilistic version, as in [49, 57].

For two distributions μ and π over a discrete state space $\mathcal{K}^n$, we write $\pi \ll \mu$ if $\mu(x) = 0$ implies $\pi(x) = 0$ for any $x \in \mathcal{K}^n$, i.e., the support of π is contained in the support of μ. The KL divergence is defined as

$$D_{\mathrm{KL}}(\pi \,\|\, \mu) = \sum_{x \in \mathcal{K}^n} \pi(x) \ln\left(\frac{\pi(x)}{\mu(x)}\right).$$

The following definition of approximate tensorization is slightly more general than Definition 1.1 from the introduction.

Definition 3.5 (Approximate Tensorization of Entropy: Probabilistic Form). We say a distribution μ over $\mathcal{K}^n$ satisfies approximate tensorization of entropy with constant C if for any distribution π over $\mathcal{K}^n$ such that $\pi \ll \mu$ one has

$$D_{\mathrm{KL}}(\pi \,\|\, \mu) \le C \sum_{i=1}^{n} \mathbb{E}_{x \sim \pi_{n\setminus i}}\Big[D_{\mathrm{KL}}(\pi_i(\cdot \mid x) \,\|\, \mu_i(\cdot \mid x))\Big]. \tag{5}$$

Note that in Definition 1.1 we required that μ has full support, instead of the more general assumption $\pi \ll \mu$. We remark that in (5) the partial configuration $x \in \mathcal{K}^{n\setminus i}$ is drawn from π rather than μ. It is easy to check that the two definitions (Definitions 3.4 and 3.5) are equivalent to each other by letting $f = \pi/\mu$; see [57].

Remark 3.6. To the best of our knowledge, there is no known analog of the probabilistic form of approximate tensorization of entropy (5) for other f-divergences, even when the visible distribution μ is a product measure. An analog of (4), the functional form of the same condition, does exist for the variance functional, which can be translated into a more intricate, weighted version of (5) for χ^2-divergence, with the entropy replaced by variance; the weights depend on ratios of the two densities and can be exponentially large. Although it is possible to establish such an inequality for certain high-dimensional distributions, how to use it for algorithmic purposes remains unclear. This contributes to our rationale for selecting KL in our work.

4 Identity Testing via Approximate Tensorization

For integer $k \ge 2$ and real $C \ge 1, \eta > 0$, let $\mathcal{F}_k(C, \eta)$ denote the family of all distributions over $\mathcal{K}^n$ (for any $n \in \mathbb{N}^+$) with $\mathrm{poly}(n)$ many parameters that are η-balanced and satisfy approximate tensorization of entropy with constant C. The goal of this section is to give an identity testing algorithm for the family $\mathcal{F}_k(C, \eta)$ in terms of the KL divergence. We observe that this also implies a tester for the TV distance by the Pinsker's inequality.

For applications in Section 4.4 all the parameters k, C, η are constants independent of n. In the theorem below, however, we consider a more general setting where these parameters are functions of the dimension n with only mild assumptions on their growth rate. This allows us to have a clearer picture on the sample complexity and the dependency on all the parameters involved.

Theorem 4.1. *Let $k = k(n)$ be an integer and $C = C(n) \ge 1$, $\eta = \eta(n) \in (0, 1/2]$ be reals. Suppose that*

$$\max\{\log C, \log\log(1/\eta)\} = O(\log n).$$

Then, there is an identity testing algorithm for the family $\mathcal{F}_k(C, \eta)$ with query access to both the Coordinate Oracle *and the* General Oracle *and for KL divergence with distance parameter $\varepsilon > 0$. The query complexity of the identity testing algorithm is*

$$O\left(\min\left\{\frac{C}{\sqrt{\eta}} \cdot \frac{n}{\varepsilon}\log^3\left(\frac{n}{\varepsilon}\right),\ C^2\sqrt{k}\log\left(\frac{1}{\eta}\right) \cdot \frac{n^2}{\varepsilon^2}\log^2\left(\frac{n}{\varepsilon}\right)\right\}\right).$$

The running time of the algorithm is polynomial in all parameters ($1/\eta$ for the first bound, and $\log(1/\eta)$ for the second) and also proportional to the time of computing the conditional marginal distributions $\mu_i(\cdot \mid x)$ for any $i \in [n]$ and any feasible $x \in \mathcal{K}^{n\setminus i}$. Furthermore, if $k = 2$, i.e., we have a binary domain $\mathcal{K} = \{0, 1\}$, the query complexity can be improved to

$$O\left(C \log\left(\frac{1}{\eta}\right) \cdot \frac{n}{\varepsilon} \log^3\left(\frac{n}{\varepsilon}\right)\right).$$

4.1 Algorithm

Before presenting our algorithm, we first give a well-known fact, e.g., see [47, Section 8.2.4] and [60, Proposition 6.7].

Lemma 4.2. *Let $\varepsilon, M > 0$ be reals and let $L = \lceil \log_2(M/\varepsilon) \rceil$. If Y is a non-negative random variable such that $Y \le M$ always and $\mathbb{E}Y \ge \varepsilon$, then there exists a non-negative integer $\ell \le L$ such that*

$$\Pr(Y \ge 2^{\ell-1}\varepsilon) \ge \frac{1}{2^\ell (L+1)}.$$

Proof. Suppose for sake of contradiction that for all $0 \le \ell \le L$ it holds

$$\Pr(Y \ge 2^{\ell-1}\varepsilon) < \frac{1}{2^\ell (L+1)}.$$

Notice that $2^L \varepsilon \ge M$. Then we have

$$\begin{aligned}
\mathbb{E}Y &= \int_0^M \Pr(Y \ge y)dy = \int_0^{\varepsilon/2} \Pr(Y \ge y)dy + \sum_{\ell=0}^{L} \int_{2^{\ell-1}\varepsilon}^{2^\ell \varepsilon} \Pr(Y \ge y)dy \\
&\le \frac{\varepsilon}{2} + \sum_{\ell=0}^{L} (2^\ell \varepsilon - 2^{\ell-1}\varepsilon) \Pr(Y \ge 2^{\ell-1}\varepsilon) \\
&< \frac{\varepsilon}{2} + \sum_{\ell=0}^{L} 2^{\ell-1}\varepsilon \cdot \frac{1}{2^\ell (L+1)} = \varepsilon,
\end{aligned}$$

which is a contradiction. □

For $i \in [n]$ and $x \in \mathcal{K}^{n\setminus i}$, we define $q_i^x = \mu_i(\cdot \mid x)$ to be a distribution over $\mathcal{K}$ induced by the pair (i, x) from μ, where we think of i and x as the parameters. Similarly, we define $p_i^x = \pi_i(\cdot \mid x)$ with respect to π.

Recall that the approximate tensorization of entropy for μ can be written as

$$D_{\mathrm{KL}}(\pi \,\|\, \mu) \le C \sum_{i=1}^{n} \mathbb{E}_{x \sim \pi_{n\setminus i}} \left[D_{\mathrm{KL}}(\pi_i(\cdot \mid x) \,\|\, \mu_i(\cdot \mid x)) \right] = Cn\, \mathbb{E}_{(i,x)} \left[D_{\mathrm{KL}}\left(p_i^x \,\|\, q_i^x\right) \right],$$

where $i \in [n]$ is a uniformly random coordinate and x is generated from the marginal distribution $\pi_{n\setminus i}$. Therefore, the original identity testing problem boils down to the following testing problem:

$$\Pr_{(i,x)}(p_i^x = q_i^x) = 1 \quad \text{v.s.} \quad \mathbb{E}_{(i,x)} \left[D_{\mathrm{KL}}\left(p_i^x \,\|\, q_i^x\right) \right] \ge \varepsilon',$$

where $\varepsilon' = \varepsilon/(Cn)$. Notice that $D_{\mathrm{KL}}\left(p_i^x \,\|\, q_i^x\right) \le \ln(1/\eta)$ for all (i, x) assuming η-balancedness and $p \ll q$. By Lemma 4.2 it further boils down to the following sequence of testing problems: let $L = \lceil \log_2(\ln(1/\eta)/\varepsilon') \rceil$ and for each $\ell \le L$, for a random pair (i, x), distinguish between $p_i^x = q_i^x$ surely versus

$$\Pr_{(i,x)} \left(D_{\mathrm{KL}}\left(p_i^x \,\|\, q_i^x\right) \ge 2^{\ell-1}\varepsilon' \right) \ge \frac{1}{2^\ell (L+1)}.$$

Algorithm 1: Identity Testing for $\mathcal{F}_k(C,\eta)$ for KL Divergence

Input: Description (parametrization) of a given distribution $\mu \in \mathcal{F}_k(C,\eta)$, query access to both Coordinate Oracle and General Oracle for an unknown distribution π, and distance parameter $\varepsilon > 0$.

$\varepsilon' \leftarrow \varepsilon/(Cn)$;
$L \leftarrow \lceil \log_2(\ln(1/\eta)/\varepsilon') \rceil$;
for $0 \le \ell \le L$ **do**
  $\varepsilon_\ell \leftarrow 2^{\ell-1}\varepsilon'$; /* Distance parameter */
  $\delta \leftarrow 2^{-2L-6}$; /* Failure probability */
  $T_\ell \leftarrow 2^{\ell+2}(L+1)$; /* Need T_ℓ samples of (i,x) to see $D_{\mathrm{KL}}\left(p_i^x \,\|\, q_i^x\right) \ge \varepsilon_\ell$ */
  for $t = 1, 2, \ldots, T_\ell$ **do**
    Sample (i,x) from π' via General Oracle for π;
    Call $\mathcal{A}_{\text{KL-ID}}$ from Lemmas 4.3 and 4.10 to distinguish between p_i^x and q_i^x with distance parameter ε_ℓ and failure probability δ (samples from p_i^x are obtained via Coordinate Oracle for π) ; /* Check whether $D_{\mathrm{KL}}\left(p_i^x \,\|\, q_i^x\right) \ge \varepsilon_\ell$ */
    if *$\mathcal{A}_{\text{KL-ID}}$ returns No (i.e., $D_{\mathrm{KL}}\left(p_i^x \,\|\, q_i^x\right) \ge \varepsilon_\ell$)* **then**
      Output: No (i.e., $D_{\mathrm{KL}}\left(\pi \,\|\, \mu\right) \ge \varepsilon$), and the algorithm ends;
    end
  end
end
Output: Yes (i.e., $\pi = \mu$)

For this testing problem, we sample (i,x) for $O\left(2^\ell(L+1)\right)$ times so that we get to see the event $D_{\mathrm{KL}}\left(p_i^x \,\|\, q_i^x\right) \ge 2^{\ell-1}\varepsilon'$, and when it happens the problem is reduced to a classical identity testing setting on a finite state space where we can apply previously known identity testing algorithm. To accomplish this we also give an identity testing algorithm for the KL divergence, which is missing in the literature; see Lemmas 4.3 and 4.10.

We give a few more definitions before presenting our algorithm formally. For a distribution π over $\mathcal{X} = \mathcal{K}^n$, we define the set $\mathcal{X}'$ by

$$\mathcal{X}' = \left\{(i,x) : i \in [n],\, x \in \mathcal{K}^{n\setminus i}\right\}$$

to be the set of all pairs (i,x) where i is one coordinate and x contains the values of all coordinates other than i. We then define a distribution π' over $\mathcal{X}'$ by

$$\pi'(i,x) = \frac{1}{n}\,\pi_{n\setminus i}(x) = \frac{1}{n}\,\pi\left(X_{n\setminus i} = x\right),$$

so that a sample from π' can be obtained in the following way: first pick $i \in [n]$ uniformly at random, and then sample x from the marginal distribution $\pi_{n\setminus i}$.

Our algorithm is given in Algorithm 1, which also appeared in the previous work [17] for uniformity testing over the binary hypercube $\{0,1\}^n$. We now give our proof of Theorem 4.1.

Proof of Theorem 4.1. Suppose first that $\pi = \mu$. Then each time we call the KL tester in Line 1, it returns Yes with probability at least $1-\delta$ since $p_i^x = q_i^x$ for any (i,x). If every time the result is Yes then Algorithm 1 will return Yes (i.e., $\pi = \mu$). By a simple union bound, the probability that

Algorithm 1 mistakenly outputs No is at most

$$\sum_{\ell=0}^{L} T_\ell \cdot \delta = \sum_{\ell=0}^{L} 2^{\ell+2}(L+1) \cdot 2^{-2L-6} \le 2^{L+3}(L+1) \cdot 2^{-2L-6} \le \frac{1}{8},$$

where the last inequality is due to $L + 1 \le 2^L$.

Next assume that $D_{\mathrm{KL}}(\pi \,\|\, \mu) \ge \varepsilon$. Then by approximate tensorization of entropy we have

$$\mathbb{E}_{(i,x)} \left[D_{\mathrm{KL}} \left(p_i^x \,\|\, q_i^x \right) \right] \ge \varepsilon',$$

where $\varepsilon' = \varepsilon/(Cn)$. By Lemma 4.2, there exists a non-negative integer $\ell \le L$ such that

$$\Pr \left(D_{\mathrm{KL}} \left(p_i^x \,\|\, q_i^x \right) \ge 2^{\ell-1}\varepsilon' \right) \ge \frac{1}{2^\ell (L+1)}.$$

For this ℓ, the algorithm repeats for $T_\ell = 2^{\ell+2}(L+1)$ times to find such a pair (i, x) via the general sampling oracle; the probability that the algorithm fails to find such (i, x) is upper bounded by

$$\left(1 - \frac{1}{2^\ell (L+1)}\right)^{T_\ell} \le \exp\left(-\frac{T_\ell}{2^\ell (L+1)}\right) = e^{-4} \le \frac{1}{50}.$$

In the case that such a pair is successfully found, the KL tester in Line 1 will return No with probability at least $1 - \delta$, and hence then Algorithm 1 returns No. Therefore, if Algorithm 1 wrongly outputs Yes then either a good pair (i, x) is not found, or the KL tester in Line 1 makes a mistake on a good pair (i, x). The probability of outputting Yes is then upper bounded by $1/50 + \delta \le 1/8$.

Finally, Lemma 4.3, combined with the amplification technique for failure probability (e.g., see [16, Lemma 1.1.1]), implies that the number of samples required by Algorithm 1 is at most

$$\begin{aligned}
&\sum_{\ell=0}^{L} T_\ell \cdot O\left(\min\left\{ \frac{\ln(1/\delta)}{\varepsilon_\ell \sqrt{\eta}}, \frac{\sqrt{k}\ln(1/\eta)\ln(1/\delta)}{\varepsilon_\ell^2} \right\} \right) \\
&= \sum_{\ell=0}^{L} O\left(\min\left\{ \frac{CL^2 n}{\varepsilon\sqrt{\eta}}, \frac{C^2\sqrt{k}L^2 n^2 \ln(1/\eta)}{2^\ell \varepsilon^2} \right\} \right) \\
&= O\left(\min\left\{ \frac{CL^3 n}{\varepsilon\sqrt{\eta}}, \frac{C^2\sqrt{k}L^2 n^2 \ln(1/\eta)}{\varepsilon^2} \right\} \right).
\end{aligned}$$

Since $L = O(\log(n/\varepsilon))$ under our assumptions $\log(C) = O(\log n)$ and $\log\log(1/\eta) = O(\log n)$, we obtain the sample complexity upper bound from the theorem. For $k = 2$ the sample complexity is obtained in the same way, using Lemma 4.10 instead. □

4.2 Identity Testing for KL Divergence on General Domain

In this and next subsection, we prove Lemma 2.1 from Section 2.1, which is also a key sub-routine of Algorithm 1. We first consider general $k \ge 2$ in this subsection, and then give an improved sample complexity bound for $k = 2$ in the next subsection.

Lemma 4.3. *Let $k \in \mathbb{N}^+$ be an integer, and let $\varepsilon > 0$, $\eta \in (0, 1/2]$ be reals. Given a target distribution q over domain $\mathcal{K}$ of size k such that either $q(a) = 0$ or $q(a) \ge \eta$ for any $a \in \mathcal{K}$, and given sample access to an unknown distribution $p \ll q$ over $\mathcal{K}$, there exists a polynomial-time identity testing algorithm that distinguishes with probability at least $2/3$ between the two cases*

$$p = q \quad \text{and} \quad D_{\mathrm{KL}}(p \,\|\, q) \ge \varepsilon \tag{6}$$

with sample complexity $O\left(\min\left\{ \frac{1}{\varepsilon\sqrt{\eta}}, \frac{\sqrt{k}\ln(1/\eta)}{\varepsilon^2} \right\} \right)$.

The running time is polynomial in $1/\eta$ if we apply the first bound, and $\log(1/\eta)$ for the second.

Remark 4.4. We remark that the dependency on η in the sample complexity is inevitable. This is because, if $\eta \to 0$, $d_{\mathrm{TV}}(p,q)$ could tend to 0 but $D_{\mathrm{KL}}(p \,\|\, q)$ can be independent of η. For example, if p and q are Bernoulli random variables with means $\hat{p}, \hat{q} \in (0, 1/2)$, respectively, one can have $D_{\mathrm{KL}}(p \,\|\, q) = 0.1$, for some $\hat{p} = \hat{p}(\hat{q})$, with $d_{\mathrm{TV}}(p,q) = |\hat{p} - \hat{q}| \to 0$ as $\hat{q} \to 0$. In Lemma 2.1 we show that the sample complexity for identity testing with respect to KL divergence depends, in the worst case, logarithmically on $1/\eta$. In particular, for uniformity testing under KL divergence, Lemma 2.1 gives an $O(\sqrt{k}/\varepsilon)$ sample complexity which matches what one would expect given the $O(\sqrt{k}/\varepsilon^2)$ sample complexity for uniformity testing under TV distance and Pinsker's inequality.

We need the following standard inequalities between statistical divergences; see [16] for more.

Lemma 4.5. *Let q be a distribution fully supported on a finite set $\mathcal{K}$, and let $\eta = \min_{a \in \mathcal{K}} q(a)$. Then for any distribution p over $\mathcal{K}$ with $p \ll q$ it holds*

$$D_{\mathrm{KL}}(p \,\|\, q) \le \chi^2(p \,\|\, q) \le \frac{1}{\eta} \|p - q\|_2^2 \le \frac{2}{\eta} d_{\mathrm{TV}}(p,q)^2 .$$

Our KL tester uses the following identity testing algorithm from [35] for ℓ_2 distance, and also the flattening technique proposed there to reduce the ℓ_2 norm. See also [16, Theorem 2.2.2] for an exposition of the algorithms and techniques.

Lemma 4.6 *([35]). Given the distance parameter $\varepsilon > 0$, full description of the target distribution q with domain $\mathcal{K}$ of size k, and general sample access to an unknown distribution p over $\mathcal{K}$, there exists a polynomial-time identity testing algorithm $\mathcal{A}_{\ell_2\text{-ID}}$ that distinguishes with probability at least $2/3$ between the two cases*

$$\|p - q\|_2 \le \frac{\varepsilon}{2} \quad \text{and} \quad \|p - q\|_2 \ge \varepsilon \tag{7}$$

with sample complexity $O\Big(\max\big\{\frac{\|q\|_2}{\varepsilon^2}, \frac{1}{\varepsilon}\big\}\Big)$.

We are now ready to give our proof of Lemma 4.3.

Proof of Lemma 4.3. Without loss of generality we may assume that q is fully supported over $\mathcal{K}$, i.e., $q(a) \ge \eta$ for all $a \in \mathcal{K}$. We establish the two bounds in the lemma separately using two testing algorithms depending on the range of parameters, both based on the ℓ_2 tester in [35]. To clarify, our testing algorithm will check the two bounds $1/(\varepsilon\sqrt{\eta})$ and $(\sqrt{k}/\varepsilon^2)\ln(1/\eta)$, find the smaller one, and run the algorithm for that bound.

Algorithm A. We construct a new instance p', q' (including the oracle for p') of the identity testing problem from p, q such that $\eta/2 \le q'(a) \le \eta$ for all $a \in \mathcal{K}'$, where $\mathcal{K}'$ is a new domain of size $k' = \Theta(1/\eta)$; this is achieved using the flattening technique from [35]. We show that the new identity testing problem with p', q' is equivalent to the original one with p, q, requiring the same number of samples. And we can apply Lemma 4.6 to the new instance p', q' with nicer properties to obtain a better bound on the sample complexity instead of doing it directly to p, q. For each $a \in \mathcal{K}$, define

$$k_a = \left\lfloor \frac{q(a)}{\eta} \right\rfloor + 1.$$

We split each $a \in \mathcal{K}$ into k_a distinct copies, denoted by $a_1, \ldots, a_{k_a}$, which constitute the new domain $\mathcal{K}'$, namely,

$$\mathcal{K}' = \{a_i : a \in \mathcal{K},\ 1 \le i \le k_a\}.$$

Notice that the size $k' = |\mathcal{K}'|$ of the new domain is bounded by

$$k' = \sum_{a \in \mathcal{K}} k_a \leq \sum_{a \in \mathcal{K}} \left(\frac{q(a)}{\eta} + 1 \right) = \frac{1}{\eta} + k \leq \frac{2}{\eta},$$

where the last inequality follows from $\eta \leq 1/k$ since q is fully supported on $\mathcal{K}$. The new target distribution q' is given by for every $a \in \mathcal{K}$ and $i \in [k_a]$,

$$q'(a_i) = \frac{q(a)}{k_a},$$

and similarly p' is given by $p'(a_i) = p(a)/k_a$. We can easily transform a sample from p into a sample from p': If we receive a as a sample from p, then we can compute k_a (since the target distribution q is given in full description) and generate $i \in [k_a]$ uniformly at random, so that a_i is a sample from the distribution p'. The crucial fact here is that the KL divergence (more generally, any f-divergence) is preserved under flattening. Indeed, observe that

$$D_{\mathrm{KL}}\left(p' \,\|\, q'\right) = \sum_{a \in \mathcal{K}} \sum_{i \in [k_a]} p'(a_i) \ln \frac{p'(a_i)}{q'(a_i)} = \sum_{a \in \mathcal{K}} \sum_{i \in [k_a]} \frac{p(a)}{k_a} \ln \frac{p(a)}{q(a)} = D_{\mathrm{KL}}\left(p \,\|\, q\right).$$

Thus, we only need to solve the identity testing problem for the flattened distribution p' and q'. Moreover, we observe that for all $a \in \mathcal{K}$ and $i \in [k_a]$ it holds

$$\frac{\eta}{2} \leq q'(a_i) = \frac{q(a)}{k_a} \leq \eta,$$

since we have

$$\frac{2q(a)}{\eta} \geq k_a = \left\lfloor \frac{q(a)}{\eta} \right\rfloor + 1 \geq \frac{q(a)}{\eta},$$

where the first inequality follows from $q(a) \geq \eta$. Therefore, we observe that

$$\|q'\|_2^2 = \sum_{a_i \in \mathcal{K}'} q'(a_i)^2 \leq \eta \sum_{a_i \in \mathcal{K}'} q'(a_i) = \eta.$$

Note that $D_{\mathrm{KL}}\left(p' \,\|\, q'\right) \leq (2/\eta)\,\|p' - q'\|_2^2$ by Lemma 4.5. Applying Lemma 4.6, we are able to distinguish between $p' = q'$ versus $\|p' - q'\|_2^2 \geq \varepsilon\eta/2$, and hence between $p' = q'$ versus $D_{\mathrm{KL}}\left(p' \,\|\, q'\right) \geq \varepsilon$, using

$$O\left(\max\left\{\frac{\|q'\|_2}{\varepsilon\eta}, \frac{1}{\sqrt{\varepsilon\eta}}\right\}\right) = O\left(\frac{1}{\varepsilon\sqrt{\eta}}\right)$$

samples from the unknown distribution p'. This then gives an identity testing algorithm for p and q for KL divergence using the same number of samples from p.

Algorithm B. The previous algorithm works well when $1/\eta$ is not too large. To get a better dependency on $1/\eta$ as in the second bound, more work is required. Our first step is still flattening the distributions, but up to the scale $1/k$ instead of η. This is done exactly in [35] and [16, Theorem 2.2.2]. Let $k_a = \lfloor kq(a) \rfloor + 1$ for each $a \in \mathcal{K}$ and let $q'(a_i) = q(a)/k_a$ for each $a \in \mathcal{K}$ and $i \in [k_a]$. The flattened distributions p' and q' satisfy the following properties:

(a) Given an explicit description of q, one can efficiently give an explicit description of the flattened distribution q;
(b) Given access to the sampling oracle for p, one can efficiently generate samples from the flattened distribution p';
(c) The KL divergence is preserved, i.e., $D_{\mathrm{KL}}\left(p' \,\|\, q'\right) = D_{\mathrm{KL}}\left(p \,\|\, q\right)$;
(d) The size of the new domain is $k' \leq 2k$;

(e) For every $a_i \in \mathcal{K}'$, we have $\eta/2 \le q'(a_i) \le 2/k'$;
(f) We have $\|q'\|_2 \le \sqrt{2/k'}$.

The proofs of these properties are the same as before or as in [16, Theorem 2.2.2] so we omit here. We only mention the lower bound on $q'(a_i)$: since $k_a = \lfloor kq(a) \rfloor + 1 \le kq(a) + q(a)/\eta$ we have that

$$q'(a_i) = \frac{q(a)}{k_a} \ge \frac{q(a)}{kq(a) + \frac{q(a)}{\eta}} = \frac{\eta}{k\eta + 1} \ge \frac{\eta}{2}.$$

Therefore, it suffices to consider the identity testing problem with respect to distributions p' and q' satisfying properties (e) and (f). For ease of notation, in the rest of the proof we assume that our p, q are already flattened to satisfy (e) and (f), instead of writing p', q', and k'.

Our second step is to divide elements in $\mathcal{K}$ into two classes, those with larger probability mass and those with smaller one, and to upper bound the KL divergence by dealing with the two classes separately. Let

$$\zeta = \frac{\varepsilon}{10k \ln(2/\eta)},$$

and let $\mathcal{K}_1 = \{a \in \mathcal{K} : q(a) \ge \zeta\}$ and $\mathcal{K}_2 = \{a \in \mathcal{K} : \eta/2 \le q(a) < \zeta\}$. Hence, $\{\mathcal{K}_1, \mathcal{K}_2\}$ forms a partition of $\mathcal{K}$. We upper bound the KL divergence of p and q as follows. Observe that

$$D_{\mathrm{KL}}(p \,\|\, q) = \sum_{a \in \mathcal{K}_1} p(a) \ln \frac{p(a)}{q(a)} + \sum_{a \in \mathcal{K}_2} p(a) \ln \frac{p(a)}{q(a)}. \tag{8}$$

For the second term, we have

$$\sum_{a \in \mathcal{K}_2} p(a) \ln \left(\frac{p(a)}{q(a)}\right) \le \sum_{a \in \mathcal{K}_2} p(a) \ln(2/\eta) = (\ln(2/\eta))\, p(\mathcal{K}_2). \tag{9}$$

For the first term, we have

$$\begin{aligned}
\sum_{a \in \mathcal{K}_1} p(a) \ln \left(\frac{p(a)}{q(a)}\right) &\le \sum_{a \in \mathcal{K}_1} p(a) \left(\frac{p(a)}{q(a)} - 1\right) \\
&= \sum_{a \in \mathcal{K}_1} (p(a) - q(a)) \left(\frac{p(a)}{q(a)} - 1\right) + \sum_{a \in \mathcal{K}_1} q(a) \left(\frac{p(a)}{q(a)} - 1\right) \\
&= \sum_{a \in \mathcal{K}_1} \frac{(p(a) - q(a))^2}{q(a)} + p(\mathcal{K}_1) - q(\mathcal{K}_1) \\
&\le \frac{1}{\zeta} \|p - q\|_2^2 + q(\mathcal{K}_2) - p(\mathcal{K}_2),
\end{aligned} \tag{10}$$

where the last inequality is because $q(a) \ge \zeta$ for $a \in \mathcal{K}_1$. Therefore, combining (8)–(10) we obtain

$$D_{\mathrm{KL}}(p \,\|\, q) \le \frac{1}{\zeta} \|p - q\|_2^2 + (\ln(2/\eta) - 1)\, p(\mathcal{K}_2) + q(\mathcal{K}_2). \tag{11}$$

In particular, (11) directly implies the following fact.

Fact 4.7. *If* $D_{\mathrm{KL}}(p \,\|\, q) \ge \varepsilon$, *then*

$$\textit{either } p(\mathcal{K}_2) \ge \frac{\varepsilon}{5 \ln(2/\eta)} \textit{ or } \|p - q\|_2^2 \ge \frac{4}{5} \varepsilon \zeta.$$

To see this, suppose on contrary that $p(\mathcal{K}_2) < \varepsilon/(5\ln(2/\eta))$ and $\|p-q\|_2^2 < \frac{4}{5}\varepsilon\zeta$. Since we know

$$q(\mathcal{K}_2) \le \zeta k = \frac{\varepsilon}{10\ln(2/\eta)} < \frac{\varepsilon}{5\ln(2/\eta)},$$

we deduce from (11) that

$$\varepsilon \le D_{\mathrm{KL}}(p \,\|\, q) < \frac{4}{5}\varepsilon + \frac{1}{5}\varepsilon = \varepsilon,$$

which is a contradiction.

Our identity testing algorithm proceeds by conducting two tests independently. In the first test, we try to distinguish between $p(\mathcal{K}_2) = q(\mathcal{K}_2)$ and $p(\mathcal{K}_2) \ge \varepsilon/(5\ln(2/\eta))$ with failure probability $1/6$ and sample complexity $m_1 = O(\ln(1/\eta)/\varepsilon)$. (If $\mathcal{K}_2 = \emptyset$ then we do nothing in this first stage.) To be more precise, let X (respectively, Y) be the indicator of the event that a sample drawn from p (respectively q) is contained in $\mathcal{K}_2$. So both X and Y are Bernoulli random variables, where the expectation $q(\mathcal{K}_2) > 0$ of Y is known to the algorithm, while the expectation $p(\mathcal{K}_2)$ of X is unknown but we have sample access to X via samples from p. We would like to distinguish between the two cases $p(\mathcal{K}_2) = q(\mathcal{K}_2)$, i.e., X and Y are the same, and $p(\mathcal{K}_2) \ge \varepsilon/(5\ln(2/\eta))$, i.e., X and Y are far from each other since $q(\mathcal{K}_2) \le \varepsilon/(10\ln(2/\eta))$. This is a standard property testing problem for Bernoulli random variables. We use the testing algorithm from Lemma 4.9 for

$$\gamma = \frac{\varepsilon}{5q(\mathcal{K}_2)\ln(2/\eta)} - 1 \ge 1$$

with failure probability $1/6$, using $m_1 = O(\ln(1/\eta)/\varepsilon)$ samples from X.

In the second stage, we run the tester from Lemma 4.6 to distinguish between $p = q$ and $\|p-q\|_2^2 \ge \frac{4}{5}\varepsilon\zeta$ with failure probability $1/6$. Let m_2 be the number of samples that the ℓ_2 tester uses, and we obtain from Lemma 4.6 that

$$m_2 = O\left(\max\left\{\frac{\|q\|_2}{\varepsilon\zeta}, \frac{1}{\sqrt{\varepsilon\zeta}}\right\}\right) = O\left(\frac{\sqrt{k}\ln(1/\eta)}{\varepsilon^2}\right),$$

where we use the property (f) from flattening.

Suppose in both tests the outputs are Yes (i.e., $p(\mathcal{K}_2) = q(\mathcal{K}_2)$ in the first and $p = q$ in the second), then our identity testing algorithm will output Yes (i.e., $p = q$). If in at least one test the output is No, then our identity testing algorithm will output No (i.e., $D_{\mathrm{KL}}(p \,\|\, q) \ge \varepsilon$). To finish up the proof, we still need to bound the failure probability and the number of samples needed for our testing algorithm. Suppose first that $p = q$, and hence $p(\mathcal{K}_2) = q(\mathcal{K}_2)$. Our testing algorithm wrongly outputs No if at least one of the two tests makes a mistake and outputs No. By a simple union bound, the probability of this is at most $1/6 + 1/6 = 1/3$. On the other hand, if $D_{\mathrm{KL}}(p \,\|\, q) \ge \varepsilon$, then either $p(\mathcal{K}_2) \ge \varepsilon/(5\ln(2/\eta))$ or $\|p-q\|_2^2 \ge \frac{4}{5}\varepsilon\zeta$ by Fact 4.7, and so at least one of the two tests should output No if it does not make a mistake. Hence, the failure probability is at most $1/6$. Finally, the number of samples we need is

$$m_1 + m_2 = O\left(\frac{\ln(1/\eta)}{\varepsilon}\right) + O\left(\frac{\sqrt{k}\ln(1/\eta)}{\varepsilon^2}\right) = O\left(\frac{\sqrt{k}\ln(1/\eta)}{\varepsilon^2}\right).$$

This establishes the second bound of the lemma. □

4.3 Identity Testing for KL Divergence on Binary Domain

If $k = 2$, i.e., we have a binary domain $\mathcal{K} = \{0, 1\}$, then the sample complexity for the KL tester is better.

For $p \in [0,1]$, the Bernoulli distribution denoted by $\mathrm{Ber}(p)$ is the distribution over $\{0,1\}$ such that $\Pr(X=1)=p$. We record below the standard Chernoff bounds.

Lemma 4.8 (Chernoff bounds). *Suppose $X_1,\dots,X_m$ are independent Bernoulli random variables from $\mathrm{Ber}(p)$ where $p \in [0,1]$. Let $\hat{p} = \frac{1}{m}\sum_{i=1}^{m} X_i$ denote the sample mean. Then for all $\delta \geq 0$,*

$$\Pr\left(\hat{p} \leq (1-\delta)p\right) \leq e^{-\delta^2 pm/2};$$

$$\Pr\left(\hat{p} \geq (1+\delta)p\right) \leq e^{-\delta^2 pm/(2+\delta)} \leq \begin{cases} e^{-\delta^2 pm/3}, & 0 \leq \delta \leq 1; \\ e^{-\delta pm/3}, & \delta \geq 1. \end{cases}$$

The following is a folklore fact.

Lemma 4.9. *Let $\gamma > 0$ be a real number. Given $q \in (0, 1/(1+\gamma)]$ and sample access to $\mathrm{Ber}(p)$ with unknown $p \in [0,1]$, there exists a polynomial-time identity testing algorithm that distinguishes with probability at least $2/3$ between the two cases*

$$p = q \quad \text{and} \quad p \geq (1+\gamma)q \tag{12}$$

with sample complexity

$$O\left(\frac{1+\gamma}{\gamma^2 q}\right) = \begin{cases} O\left(\dfrac{1}{\gamma^2 q}\right), & 0 < \gamma \leq 1; \\ O\left(\dfrac{1}{(1+\gamma)q}\right), & \gamma \geq 1. \end{cases}$$

Proof. Let $\hat{p}$ denote the sample mean of m independent samples from $\mathrm{Ber}(p)$ where

$$m = \left\lceil \frac{10(1+\gamma)}{\gamma^2 q} \right\rceil.$$

If $\hat{p} \leq (1+\gamma/2)q$, then the tester concludes $p = q$; otherwise, it concludes $p \geq (1+\gamma)q$.

Suppose first $p = q$. Then by the Chernoff bound Lemma 4.8 we have

$$\Pr\left(\hat{p} \geq \left(1+\frac{\gamma}{2}\right)q\right) \leq \exp\left(-\frac{\gamma^2 qm}{2(4+\gamma)}\right) \leq \frac{1}{3}.$$

If $p \geq (1+\gamma)q$, then again by the Chernoff bound Lemma 4.8 we have

$$\begin{aligned} \Pr\left(\hat{p} \leq \left(1+\frac{\gamma}{2}\right)q\right) &\leq \Pr\left(\hat{p} \leq \left(\frac{1+\gamma/2}{1+\gamma}\right)p\right) = \Pr\left(\hat{p} \leq \left(1-\frac{\gamma}{2(1+\gamma)}\right)p\right) \\ &\leq \exp\left(-\frac{\gamma^2 pm}{8(1+\gamma)^2}\right) \leq \exp\left(-\frac{\gamma^2 qm}{8(1+\gamma)}\right) \leq \frac{1}{3}. \end{aligned}$$

Finally, for $0 < \gamma \leq 1$ one has $(1+\gamma)/\gamma^2 \leq 2/\gamma^2$, and for $\gamma \geq 1$ one has $(1+\gamma)/\gamma^2 \leq 4/(1+\gamma)$, which completes the proof of the lemma. □

We now give our testing algorithm for Bernoulli random variables.

Lemma 4.10. *Let $\varepsilon > 0$ be a real number. Given $q \in (0,1)$ and sample access to $\mathrm{Ber}(p)$ with unknown $p \in [0,1]$, there exists a polynomial-time identity testing algorithm that distinguishes with probability at least $2/3$ between the two cases*

$$p = q \quad \text{and} \quad D_{\mathrm{KL}}\left(\mathrm{Ber}(p) \,\|\, \mathrm{Ber}(q)\right) \geq \varepsilon \tag{13}$$

with sample complexity

$$O\left(\frac{\ln(1/\eta)}{\varepsilon}\right)$$

where $\eta = \min\{q, 1-q\}$.

Proof. We may assume without loss of generality that $q \le 1/2$ and $\eta = q$, as otherwise we can flip the Bernoulli. For $q \in (0,1)$ and $p \in [0,1]$, we define

$$\varphi_{\mathrm{kl}}(p,q) = D_{\mathrm{KL}}\left(\mathrm{Ber}(p) \,\|\, \mathrm{Ber}(q)\right) = p \ln \frac{p}{q} + (1-p) \ln \frac{1-p}{1-q}.$$

The testing algorithm is as follows. Let $S < T$ be parameters which, as will be clear soon, depend on the distance parameter ε and the mean q of the given Bernoulli (note that both ε and q are known to the algorithm). Compute the sample mean $\hat{p}$ for p using m independent samples from $\mathrm{Ber}(p)$. The testing algorithm determines $p = q$ or $\varphi_{\mathrm{kl}}(p,q) \ge \varepsilon$ by checking whether $\hat{p}$ belongs to the interval $[S,T]$ or not. More specifically, if $\hat{p} \in [S,T]$, then it outputs $p = q$. If $\hat{p} \notin [S,T]$, then it outputs $\varphi_{\mathrm{kl}}(p,q) \ge \varepsilon$. We need to choose suitable S and T so that the algorithm is accurate with high probability and the number of samples required is minimized. Given $q \in (0, 1/2]$ and $\varepsilon > 0$, we will consider three separate cases.

Case 1: $\varepsilon \le 2q$. We choose $S = q - \sqrt{\varepsilon q/8}$, $T = q + \sqrt{\varepsilon q/8}$, and let

$$m = \left\lceil \frac{64}{\varepsilon} \right\rceil.$$

be the number of samples. If $p = q$, then by the Chernoff bound (Lemma 4.8) we have

$$\Pr\left(\hat{p} \notin [S,T]\right) \le \Pr\left(|\hat{p} - q| \ge \sqrt{\frac{\varepsilon q}{8}}\right) \le 2\exp\left(-\frac{\varepsilon m}{24}\right) \le \frac{1}{3},$$

where we use $\sqrt{\varepsilon q/8} \le q/2$ by the assumption $\varepsilon \le 2q$.

Now suppose $\varphi_{\mathrm{kl}}(p,q) \ge \varepsilon$. By Lemma 4.5, we have

$$\varepsilon \le \varphi_{\mathrm{kl}}(p,q) \le \frac{2}{q}(p-q)^2,$$

and hence either $p \le q - \sqrt{\varepsilon q/2}$ or $p \ge q + \sqrt{\varepsilon q/2}$. Suppose $p \le q - \sqrt{\varepsilon q/2}$. If $p = 0$ then trivially $\Pr\left(\hat{p} \in [S,T]\right) = 0$ since $S = q - \sqrt{\varepsilon q/8} \ge q/2 > 0$. If $0 < p \le q - \sqrt{\varepsilon q/2}$, then again by the Chernoff bound (Lemma 4.8) we have

$$\Pr\left(\hat{p} \in [S,T]\right) \le \Pr\left(\hat{p} \ge S\right) \le \exp\left(-\frac{(S-p)^2 m}{2p + (S-p)}\right) \le \exp\left(-\frac{\varepsilon m}{16}\right) \le \frac{1}{3},$$

where the second to last inequality follows from $S + p \le 2q$ and $(S-p)^2 \ge \varepsilon q/8$. If $p \ge q + \sqrt{\varepsilon q/2}$, then Lemma 4.8 gives

$$\Pr\left(\hat{p} \in [S,T]\right) \le \Pr\left(\hat{p} \le T\right) \le \exp\left(-\frac{(p-T)^2 m}{2p}\right) \le \exp\left(-\frac{\varepsilon m}{32}\right) \le \frac{1}{3},$$

where the second to last inequality follows from the fact that $(p-T)^2/(2p)$ is minimized at $p = q + \sqrt{\varepsilon q/2} \le 2q$ and hence

$$\frac{(p-T)^2}{2p} \ge \frac{(q + \sqrt{\varepsilon q/2} - T)^2}{2(q + \sqrt{\varepsilon q/2})} \ge \frac{\varepsilon}{32}.$$

Case 2: $2q < \varepsilon \le 2q\ln(1/q)$. (This case is possible only for $q < 1/e$.) Again if $\varphi_{\mathrm{kl}}(p,q) \ge \varepsilon$ then either $p \le q - \sqrt{\varepsilon q/2}$ or $p \ge q + \sqrt{\varepsilon q/2}$. But since $\varepsilon > 2q$, we have $q - \sqrt{\varepsilon q/2} < 0$ and hence it must be $p \ge q + \sqrt{\varepsilon q/2}$. This means that, we need to distinguish between $p = q$ versus

$$p \ge q + \sqrt{\varepsilon q/2} \ge 2q$$

as $\varepsilon > 2q$. Therefore, we can apply the identity tester from Lemma 4.9 for $\gamma = 1$ with sample complexity

$$m = O\left(\frac{1}{q}\right) = O\left(\frac{\ln(1/q)}{\varepsilon}\right)$$

since $\varepsilon \le 2q\ln(1/q)$.

Case 3: $\varepsilon > \max\{2q,\, 2q\ln(1/q)\}$. Just as in Case 2, if $\varphi_{\mathrm{kl}}(p,q) \ge \varepsilon$ then one must have $p \ge q$, since $p < q$ implies

$$\varphi_{\mathrm{kl}}(p,q) \le \varphi_{\mathrm{kl}}(0,q) = \ln\left(\frac{1}{1-q}\right) \le \frac{q}{1-q} \le 2q.$$

Since $p \ge q$, we have $1 - p \le 1 - q$ and thus

$$\varepsilon \le \varphi_{\mathrm{kl}}(p,q) = p\ln\frac{p}{q} + (1-p)\ln\frac{1-p}{1-q} \le p\ln\frac{p}{q} \le p\ln\frac{1}{q}.$$

Therefore, it suffices to distinguish between $p = q$ and $p \ge \varepsilon/\ln(1/q) > 2q$. The identity tester from Lemma 4.9 for $\gamma = \varepsilon/(q\ln(1/q)) - 1 \ge 1$ can achieve $2/3$ success probability with sample complexity

$$m = O\left(\frac{\ln(1/q)}{\varepsilon}\right).$$

This completes the proof of the lemma. □

4.4 Applications

Here we give several applications of Theorem 4.1.

4.4.1 Product Distributions. For each $i \in [n]$ let μ_i be an arbitrary distribution over $\mathcal{K}$, and define a product distribution $\mu = \mu_1 \otimes \cdots \otimes \mu_n$ over $\mathcal{K}^n$. It is well-known that every product distribution satisfies approximate tensorization of entropy with an optimal constant $C = 1$.

Lemma 4.11 ([20, 22, 56]). *Let μ be any product distribution over $\mathcal{K}^n$. For any distribution π over $\mathcal{K}^n$ such that $\pi \ll \mu$, we have*

$$D_{\mathrm{KL}}(\pi \,\|\, \mu) \le \sum_{i=1}^{n} \mathbb{E}_{x\sim\pi_{n\setminus i}}\Big[D_{\mathrm{KL}}\left(\pi_i(\cdot \mid x) \,\|\, \mu_i(\cdot \mid x)\right)\Big].$$

Namely, every product distribution satisfies approximate tensorization of entropy with constant 1.

For a product distribution μ, define $\eta(\mu) = \min_{i\in[n]} \min_{a\in\mathcal{K}:\,\mu_i(a)>0} \mu_i(a)$. Observe that μ is $\eta(\mu)$-balanced. Let $\mathcal{P}(\eta)$ denote the collection of all product distributions μ such that $\eta(\mu) \ge \eta$. The following corollary follows immediately from Theorem 4.1 and Lemma 4.11.

Corollary 4.12. *Let $\eta \in (0, 1/2]$ be real. There is a polynomial-time identity testing algorithm for the family $\mathcal{P}(\eta)$ of η-balanced product distributions with query access to both* Coordinate Oracle *and* General Oracle *and for KL divergence with distance parameter $\varepsilon > 0$. The query complexity of the identity testing algorithm is $O((n/\varepsilon)\log^3(n/\varepsilon))$.*

4.4.2 Sparse Ising Models in the Uniqueness Region. An Ising model is a tuple (G, β, h) where

— $G = (V, E)$ is a finite simple graph;
— $\beta : E \to \mathbb{R}$ is a function of edge couplings;
— $h : V \to \mathbb{R}$ is a function of vertex external fields.

We may also view β and h as vectors; in particular, we write β_{uv} to represent the edge coupling of an edge $\{u, v\} \in E$, and write h_v to represent the external field of a vertex $v \in V$.

The Gibbs distribution of an Ising model (G, β, h) is given by

$$\mu_{(G,\beta,h)}(\sigma) = \frac{1}{Z_{(G,\beta,h)}} \exp\left(\sum_{\{u,v\} \in E} \beta_{uv}\sigma_u\sigma_v + \sum_{v \in V} h_v\sigma_v \right), \quad \forall \sigma \in \{+,-\}^V,$$

where

$$Z_{(G,\beta,h)} = \sum_{\sigma \in \{+,-\}^V} \exp\left(\sum_{\{u,v\} \in E} \beta_{uv}\sigma_u\sigma_v + \sum_{v \in V} h_v\sigma_v \right)$$

is the partition function.

Definition 4.13. (The Family $\mathcal{IS}(\Delta, \delta, h^*)$ of Ising Models in Tree-Uniqueness). For an integer $\Delta \geq 3$ and reals $\delta \in (0, 1), h^* > 0$, let $\mathcal{IS}(\Delta, \delta, h^*)$ be the family of Gibbs distributions of Ising models (G, β, h) satisfying:

(1) The maximum degree of G is at most Δ;
(2) We have $(\Delta - 1)\tanh(\beta^*) \leq 1 - \delta$, where $\beta^* = \max_{\{u,v\} \in E} |\beta_{uv}|$ denotes the maximum edge coupling in absolute value;
(3) For each $v \in V(G)$, we have $|h_v| \leq h^*$.

Recent works toward establishing optimal mixing of Glauber dynamics have shown approximate tensorization of entropy for the family $\mathcal{IS}(\Delta, \delta, h^*)$.

Lemma 4.14. ([27, 28]). *For any integer $\Delta \geq 3$ and reals $\delta \in (0, 1), h^* > 0$, there exists a constant $C = C(\Delta, \delta, h^*) \geq 1$, such that every Ising distribution μ from the family $\mathcal{IS}(\Delta, \delta, h^*)$ satisfies approximate tensorization of entropy with constant C.*

We then deduce the following corollary from Theorem 4.1 and Lemma 4.14.

Corollary 4.15. *Suppose $\Delta \geq 3$ is an integer and $\delta \in (0, 1), h^* > 0$ are reals. There is a polynomial-time identity testing algorithm for the family $\mathcal{IS}(\Delta, \delta, h^*)$ of Ising models with query access to both* Coordinate Oracle *and* General Oracle *and for KL divergence with distance parameter $\varepsilon > 0$. The query complexity of the identity testing algorithm is $O((n/\varepsilon)\log^3(n/\varepsilon))$.*

4.4.3 Distributions Satisfying Dobrushin Uniqueness Condition. Let μ be a distribution over $\mathcal{K}^n$. For $i, j \in [n]$, the Dobrushin influence of i on j is given by

$$a_{u,v} = \max_{(x,x') \in C_{i,j}} d_{\mathrm{TV}}\left(\mu_j(\cdot \mid X_{n\setminus j} = x),\, \mu_j(\cdot \mid X_{n\setminus j} = x')\right),$$

where $C_{i,j}$ denotes the collection of all pairs (x, x') of vectors in $\mathcal{K}^{n\setminus i}$ such that $\mu(X_{n\setminus j} = x) > 0$, $\mu(X_{n\setminus j} = x') > 0$, and x, x' either are the same or differ exactly at the coordinate i. The Dobrushin influence matrix A is an $n \times n$ matrix with entries given as earlier. Note that A is not symmetric in general.

For $b \in (0, 1/2]$, we say the distribution μ is b-marginally bounded if for every $\Lambda \subseteq [n]$, every $x \in \mathcal{K}^\Lambda$ with $\mu(X_\Lambda = x) > 0$, every $i \in [n] \setminus \Lambda$, and every $a \in \mathcal{K}$, one has

$$\text{either } \mu(X_i = a \mid X_\Lambda = x) \geq b \quad \text{or } \mu(X_i = a \mid X_\Lambda = x) = 0.$$

Note that though seemingly similar, the notion of marginal boundedness is not the same as the coordinate balancedness defined in Section 3.3. We observe that any b-marginally bounded distribution is also b-balanced.

For $\delta \in (0, 1)$ and $b \in (0, 1/2]$, let $\mathcal{D}(\delta, b)$ be the family of all distributions over $\mathcal{K}^n$ satisfying the following conditions:

(1) The Dobrushin influence matrix A of μ satisfies $\|A\|_2 \leq 1 - \delta$;
(2) μ is b-marginally bounded.

Marton proved that every distribution from the family $\mathcal{D}(\delta, b)$ satisfies approximate tensorization of entropy.

Lemma 4.16 ([57]). *Suppose $\delta \in (0, 1)$ and $b \in (0, 1/2]$ are reals. Every distribution μ from the family $\mathcal{D}(\delta, b)$ satisfies approximate tensorization of entropy with constant $C = 1/(b\delta^2)$.*

The following corollary follows from Theorem 4.1 and Lemma 4.16.

Corollary 4.17 *Suppose $\delta \in (0, 1)$ and $b \in (0, 1/2]$ are reals. There is a polynomial-time identity testing algorithm for the family $\mathcal{D}(\delta, b)$ with query access to both the* Coordinate Oracle *and the* General Oracle *and for KL divergence with distance parameter $\varepsilon > 0$. The query complexity of the identity testing algorithm is $O((n/\varepsilon)\log^3(n/\varepsilon))$.*

For Ising models, there is also a stronger version of Dobrushin uniqueness in literature.

Definition 4.18 (The Family $\mathcal{IS}_{\mathrm{D}}(\delta, h^*)$ of Ising Models in Dobrushin-Uniqueness). For $\delta \in (0, 1)$ and $h^* > 0$, let $\mathcal{IS}_{\mathrm{D}}(\delta, h^*)$ be the family of Gibbs distributions of Ising models (G, β, h) satisfying:

(1) For each $v \in V(G)$, we have $\sum_{u \in N(v)} |\beta_{uv}| \leq 1 - \delta$;
(2) For each $v \in V(G)$, we have $|h_v| \leq h^*$.

Notice that in the Ising model we have $a_{u,v} \leq \tanh(|\beta_{uv}|) \leq |\beta_{uv}|$ for $\{u, v\} \in E$ and $a_{u,v} = 0$ for non-edges. So we have $\mathcal{IS}_{\mathrm{D}}(\delta, h^*) \subseteq \mathcal{D}(\delta, b)$ for $b \geq 1/(e^{2(h^*+1)} + 1)$. Hence, the following corollary follows immediately from Corollary 4.17.

Corollary 4.19. *Suppose $\delta \in (0, 1)$ and $h^* > 0$ are reals. There is a polynomial-time identity testing algorithm for the family $\mathcal{IS}_{\mathrm{D}}(\delta, h^*)$ of Ising models with query access to both* Coordinate Oracle *and* General Oracle *and for KL divergence with distance parameter $\varepsilon > 0$. The query complexity of the identity testing algorithm is $O((n/\varepsilon)\log^3(n/\varepsilon))$.*

4.5 Identity Testing for TV Distance

One of the main goals of this article is to give efficient identity testing algorithms *without any restriction* on the noisy, unknown distribution π. However, since we work with KL divergence in most parts of our algorithmic results, one assumption we have to make is that the support of the hidden distribution π is contained in that of the visible μ, denoted by $\pi \ll \mu$. This is necessary for the KL divergence $D_{\mathrm{KL}}(\pi \,\|\, \mu)$ to be finite. However, we emphasize that this assumption is fairly mild and does not introduce any restriction in many settings for the following two reasons: (1) In many cases the visible distribution μ is already fully supported on $\mathcal{K}^n$ and hence the hidden one π can be arbitrary, e.g., μ is the uniform distribution or from an Ising model. (2) Testing algorithms for KL divergence can be easily applied as a black box to obtain identity testing algorithms for

TV distance, where in the latter we do not require $\pi \ll \mu$. Here we show how our identity testing algorithm (Algorithm 1) can be used to test for TV distance.

Lemma 4.20. *Suppose $\mathcal{A}_{KL\text{-}ID}$ is an identity testing algorithm for a family $\mathcal{F}$ of distributions with query access to both* Coordinate Oracle *and* General Oracle *and for KL divergence with distance parameter $\varepsilon > 0$. The query complexity of $\mathcal{A}_{KL\text{-}ID}$ is $m(n, 1/\varepsilon)$ and the running time of $\mathcal{A}_{KL\text{-}ID}$ is polynomial in n and $1/\varepsilon$. Then there exists a polynomial-time identity testing algorithm $\mathcal{A}_{TV\text{-}ID}$ for $\mathcal{F}$ with the same query access and for TV distance with distance parameter $\varepsilon > 0$. The query complexity of $\mathcal{A}_{TV\text{-}ID}$ is $O(m(n, 2/\varepsilon^2) + 1/\varepsilon)$.*

Proof. Let $\mathcal{X}_\mu \subseteq \mathcal{X}$ denote the support of μ. By the law of total probability we have $\pi(\cdot) = \pi(\mathcal{X}_\mu)\,\pi(\cdot \mid \mathcal{X}_\mu) + \pi(\mathcal{X}_\mu^{\mathsf{c}})\,\pi(\cdot \mid \mathcal{X}_\mu^{\mathsf{c}})$ where $\mathcal{X}_\mu^{\mathsf{c}} = \mathcal{X} \setminus \mathcal{X}_\mu$ is the complement. Therefore, we obtain from the triangle inequality that

$$d_{\mathrm{TV}}(\pi, \mu) \le \pi(\mathcal{X}_\mu^{\mathsf{c}}) + d_{\mathrm{TV}}\left(\pi(\cdot \mid \mathcal{X}_\mu), \mu\right).$$

In particular, if $d_{\mathrm{TV}}(\pi, \mu) \ge \varepsilon$, then either $\pi(\mathcal{X}_\mu^{\mathsf{c}}) \ge \varepsilon/2$ or $d_{\mathrm{TV}}\left(\pi(\cdot \mid \mathcal{X}_\mu), \mu\right) \ge \varepsilon/2$, where the latter implies $D_{\mathrm{KL}}\left(\pi(\cdot \mid \mathcal{X}_\mu) \,\|\, \mu\right) \ge \varepsilon^2/2$ via the Pinsker's inequality.

Our testing algorithm $\mathcal{A}_{\text{TV-ID}}$ runs in two stages. In the first stage, we distinguish between $\pi(\mathcal{X}_\mu^{\mathsf{c}}) = 0$ versus $\pi(\mathcal{X}_\mu^{\mathsf{c}}) \ge \varepsilon/2$ using $O(1/\varepsilon)$ samples from π, and we say π passes this stage if none of these samples is in $\mathcal{X}_\mu^{\mathsf{c}}$. In particular, by choosing suitable constants we can make the failure probability at most $1/3$, i.e., if $\pi(\mathcal{X}_\mu^{\mathsf{c}}) \ge \varepsilon/2$ then the probability that π passes is at most $1/3$. Observe that if $\pi(\mathcal{X}_\mu^{\mathsf{c}}) = 0$ then it always passes the first stage.

In the second stage, we test between $\pi(\cdot \mid \mathcal{X}_\mu) = \mu$ versus $D_{\mathrm{KL}}\left(\pi(\cdot \mid \mathcal{X}_\mu) \,\|\, \mu\right) \ge \varepsilon^2/2$, using $\mathcal{A}_{\text{KL-ID}}$ with failure probability $1/3$. Note that if we saw samples that belong to $\mathcal{X}_\mu^{\mathsf{c}}$ when running $\mathcal{A}_{\text{KL-ID}}$, either from calls of Coordinate Oracle or from calls of General Oracle, we can safely conclude that $\pi \ne \mu$ and hence $d_{\mathrm{TV}}(\pi, \mu) \ge \varepsilon$. Otherwise, these samples can be viewed as generated perfectly from the conditional distribution $\pi(\cdot \mid \mathcal{X}_\mu)$. We say π passes the second stage if $\mathcal{A}_{\text{KL-ID}}$ outputs Yes (i.e., $\pi(\cdot \mid \mathcal{X}_\mu) = \mu$).

If π passes both stages then $\mathcal{A}_{\text{TV-ID}}$ outputs Yes (i.e., $\pi = \mu$); otherwise it outputs No (i.e., $d_{\mathrm{TV}}(\pi, \mu) \ge \varepsilon$). Observe that, if $\pi = \mu$ then it passes the first stage always and passes the second stage with probability at least $2/3$. Meanwhile, if $d_{\mathrm{TV}}(\pi, \mu) \ge \varepsilon$ then either $\pi(\mathcal{X}_\mu^{\mathsf{c}}) \ge \varepsilon/2$ or $D_{\mathrm{KL}}\left(\pi(\cdot \mid \mathcal{X}_\mu) \,\|\, \mu\right) \ge \varepsilon^2/2$. If $\pi(\mathcal{X}_\mu^{\mathsf{c}}) \ge \varepsilon/2$ then it passes the first stage with probability at most $1/3$. And if $D_{\mathrm{KL}}\left(\pi(\cdot \mid \mathcal{X}_\mu) \,\|\, \mu\right) \ge \varepsilon^2/2$ it passes the second stage with probability at most $1/3$. Hence, the probability that π passes both stages is at most $1/3$. Therefore, $\mathcal{A}_{\text{TV-ID}}$ is a polynomial-time identity testing algorithm with sample complexity $O(m(n, 2/\varepsilon^2) + 1/\varepsilon)$. □

5 Hardness of Identity Testing When Approximate Tensorization Fails

In this section we show that approximate tensorization is essentially a necessary condition for efficient identity testing, in the sense that there are high-dimensional distributions, specifically the anti-ferromagnetic Ising model, for which either approximate tensorization holds with constant $C = O(1)$ (and thus there is an efficient identity algorithm from Theorem 1.2) or there is no polynomial-time identity testing algorithm with General Oracle and Coordinate Oracle access unless RP = NP.

We prove the hardness result in Theorem 1.4 from the introduction in the following sections. We use a reduction from the maximum cut problem to identity testing. In particular, given a hard maximum cut instance, we construct an identity testing instance whose outputs provide the maximum cut. Our reduction is inspired by the one in [5], but we use a different "degree reducing" gadget (namely, the one from [64]), and we are also required to design an algorithm to sample from

the hidden model we construct. This is challenging because sampling from the anti-ferromagnetic Ising model is NP-hard in general, but for our instance we manage to do it using a hybrid approach. Specifically, we use the recent algorithm from [54] for low-rank Ising model for one range of parameters and polymer models [50] for the other. Both algorithms rely on the fact that the graph in our testing is a random bipartite graph with trees attached to it that happens to be a good expander.

Our proof is organized as follows. First, we introduce our degree reducing gadget in Section 5.1. The testing instance construction and the reduction is then provided in Section 5.2. Finally, Sections 5.3 and 5.4 contain our sampling algorithm.

5.1 The Degree Reducing Gadget

The gadget construction has as parameters integers $n \geq 1$, $d \geq 3$ and real numbers $0 < \theta, \psi < 1/8$. Let $\ell = 2\lfloor \frac{\psi}{2} \log_{d-1} n \rfloor$, $t = (d-1)^{\lfloor \theta \log_{d-1} n \rfloor}$ and $m = t(d-1)^{\ell}$. The gadget is constructed as follows:

(1) Let $\hat{G} = (V_{\hat{G}}, E_{\hat{G}})$ be a random bipartite graph with $n + m$ vertices on each side.
(2) For $s \in \{+, -\}$, let the vertices on the s-side of $\hat{G}$ be $W_s \cup U_s$, where $|W_s| = n$ and $|U_s| = m$.
(3) Let $M_1, \ldots, M_{d-1}$ be $d-1$ random perfect matchings between $W_+ \cup U_+$ and $W_- \cup U_-$, that is, each M_i is drawn uniformly at random from the set of all perfect matching between $W_+ \cup U_+$ and $W_- \cup U_-$.
(4) Let M' be a random perfect matching between W_+ and W_-.
(5) Set $E_{\hat{G}} = M' \cup \left(\bigcup_{i=1}^{d-1} M_i \right)$.
(6) Construct collections $\mathcal{T}_+$ and $\mathcal{T}_-$ each of t disjoint $(d-1)$-ary trees of height ℓ.
(7) Adjoin $\mathcal{T}_+$ (resp., $\mathcal{T}_-$) to $\hat{G}$ by identifying each vertex of U_+ (resp., of U_-) with one of the leafs of the trees in $\mathcal{T}_+$ (resp., $\mathcal{T}_-$). We denote the set of roots of the trees in $\mathcal{T}_+$ (resp., $\mathcal{T}_-$) by R_+ (resp., R_-).

Let $G = (V_G, E_G)$ be the random multi-graph resulting from this construction.

5.2 The Reduction

Let $(K = (V_K, E_K), k)$ be an instance of the maximum cut problem. Namely, we want to distinguish between the cases $\mathsf{max\text{-}cut}(K) < k$ and $\mathsf{max\text{-}cut}(K) \geq k$, where $\mathsf{max\text{-}cut}(K)$ denotes the size of the maximum cut of the graph K.

Let $N = |V_K|$; we may assume that $N = n^{\theta/12}$, where n and θ are the parameters for the degree reducing gadget construction in the previous section. Form the multi-graph $F = (V_F, E_F)$ by adding two special vertices s and t to K (i.e., $V_F = V_K \cup \{s, t\}$), connecting s and t with $N^2 - k$ edges, and adding N edges between each s and t and each vertex in V_K; note that F has $|E_K| + 3N^2 - k$. This construction ensures that:

(1) When $\mathsf{max\text{-}cut}(K) < k$, then $(\{s, t\}, V_K)$ is the unique maximum cut of F and has size $2N^2$;
(2) When $\mathsf{max\text{-}cut}(K) \geq k$, then there exists another cut in F whose size is at least $2N^2$; this cut is obtained by taking the maximum cut for K and adding s and t to opposite sides of it.

Next, we generate an instance $G = (V_G, E_G)$ of the degree reducing gadget from Section 5.1. We then obtain the multi-graph $\widehat{F} = (V_{\widehat{F}}, E_{\widehat{F}})$ by replacing every vertex $v \in V_F$ with a copy G; we label each copy of G by G^v and let R_+^v and R_-^v denote R_+ and R_- for G_v. Moreover, for each edge $\{u, v\} \in E_F$, we add a matching of size $n^{3\theta/4}$ between R_+^v and R_+^u, and another matching of the same size between R_-^v and R_-^u. Note that $\widehat{F}$ is a d-regular multi-graph.

We will consider the anti-ferromagnetic Ising model on the multi-graph $\widehat{F}$. (See Section 4.4.2 for the definition of the Ising model on a simple graph. The definition extends to the multi-graph setting by simply considering multi-edges in the summation.) For a configuration $\sigma \in \{+1,-1\}^{V_{G_v}}$, we define its *phase* $Y_v(\sigma)$ as +1 if the number of vertices assigned -1 in W_+ is greater than the number of vertices assigned -1 in W_-; otherwise we set $Y(\sigma) = -1$. For a configuration $\sigma \in \{+1,-1\}^{V_{\widehat{F}}}$, we let $Y(\sigma)$ denote the *phase vector* of σ, which contains as coordinates the phase of σ in each gadget G_v.

Let $\Omega = \{+1,-1\}^{V_F}$ be the set of all phase vectors. Let $\xi^+_{st} \in \Omega$ (resp., $\xi^-_{st} \in \Omega$) be the phase vector that assigns +1 (resp., -1) to s,t and -1 (resp., +1) to every other gadget in $\widehat{F}$. Let $\Omega_{st} = \{\xi^+_{st}, \xi^-_{st}\}$. Observe that each phase vector $Y(\sigma)$ corresponds to a cut in the graph F, with the phase determining the side of the cut for each vertex.

Let $\Omega'_0 \subseteq \Omega$ be the collection of all phase vectors corresponding to cuts $(\{s\} \cup U, \{t\} \cup V_F \setminus U)$ of F, which in turn correspond to cuts $(U, V_K \setminus U)$ of K of size $< k$. Let Ω_0 be Ω'_0 together with the phase vectors for cuts $(\{s,t\} \cup U, V_F \setminus U)$ of F. Then:

(1) if $\mathsf{max\text{-}cut}(K) < k$, then $\Omega_0 = \Omega$;
(2) if $\mathsf{max\text{-}cut}(K) \geq k$, then $\Omega_0 \subsetneq \Omega$ and $\Omega \setminus \Omega_0$ contains at least one phase vector corresponding to a cut $(\{s\} \cup U, \{t\} \cup V_F \setminus U)$ of F, where $(U, V_K \setminus U)$ is a maximum cut for K.

We are now ready to describe our instance for the identity testing problem. Let $\beta < \beta_c(d) := -\frac{1}{2}\ln(\frac{d}{d-2})$; this parameter regime corresponds to the so-called tree uniqueness region for $(d-1)$-ary infinite trees. The visible distribution of our testing instance will be the Gibbs distribution $\mu_{\widehat{F},\beta}$ for the anti-ferromagnetic Ising model on $\widehat{F}$. The hidden distribution will be $\mu_{\widehat{F},\beta}(\cdot \mid Y(\sigma) \in \Omega_0)$, that is, $\mu_{\widehat{F},\beta}$ conditioned on the phase vector being in Ω_0. Our construction ensures that if $\mathsf{max\text{-}cut}(K) < k$, then $\mu_{\widehat{F},\beta}(\cdot \mid Y(\sigma) \in \Omega_0) = \mu_{\widehat{F},\beta}$. In addition, we have the following fact.

Lemma 5.1. *If* $\mathsf{max\text{-}cut}(K) \geq k$ *and* $\beta < \beta_c(d)$, *then* $d_{\mathrm{TV}}\left(\mu_{\widehat{F},\beta}(\cdot \mid Y(\sigma) \in \Omega_0), \mu_{\widehat{F},\beta}\right) = 1 - o(1)$.

Proof. Observe that

$$d_{\mathrm{TV}}\left(\mu_{\widehat{F},\beta}(\cdot \mid Y(\sigma) \in \Omega_0), \mu_{\widehat{F},\beta}\right) = \sum_{\sigma : Y(\sigma) \in \Omega \setminus \Omega_0} \mu_{\widehat{F},\beta}(\sigma).$$

Since $\mathsf{max\text{-}cut}(K) \geq k$, the set $\Omega \setminus \Omega_0$ contains (at least) the phase vector corresponding to a maximum cut of F. Hence, $\sum_{\sigma : Y(\sigma) \in \Omega \setminus \Omega_0} \mu_{\widehat{F},\beta}$ is at least the probability that a sample from $\mu_{\widehat{F},\beta}$ reveals a maximum cut for F. The results in [14, 42, 64] imply that this probability is indeed $1 - 1/2^{n^{\theta/4}}$, as desired. Specifically, the argument in the proof of Theorems 1 and 2 in [64] shows that this holds (under certain conditions) for the hard-core model; [42] extends the argument for any anti-ferromagnetic spin system (including the Ising model); and Lemma 22 from [14] shows that the required condition holds for all $\beta < -\frac{1}{2}\ln(\frac{d}{d-2})$ in the tree uniqueness region. □

The idea of our reduction is to provide this testing instance to a presumed polynomial-time identity testing algorithm and use its output to determine whether $\mu_{\widehat{F},\beta}(\cdot \mid Y(\sigma) \in \Omega_0) = \mu_{\widehat{F},\beta}$ or

$$d_{\mathrm{TV}}\left(\mu_{\widehat{F},\beta}(\cdot \mid Y(\sigma) \in \Omega_0), \mu_{\widehat{F},\beta}(\cdot)\right) = 1 - o(1).$$

This gives whether $\Omega_0 = \Omega$ or not, and thus whether the $\mathsf{max\text{-}cut}(K) < k$ or not, which would solve the maximum cut problem in randomized polynomial time and imply that there is no polynomial-time identity testing algorithm unless RP = NP.

All that remains to complete the reduction is that we show how to sample (in polynomial time) from the hidden distribution $\mu_{\widehat{F},\beta}(\cdot \mid Y(\sigma) \in \Omega_0)$ and how to simulate the Coordinate Oracle for it.

Simulating the conditional marginal oracle for $\mu_{\widehat{F},\beta}(\cdot \mid Y(\sigma) \in \Omega_0)$ is straightforward. Given a vertex $v \in V_{\widehat{F}}$ and a configuration $\sigma \in \{+1,-1\}^{V_{\widehat{F}}\setminus\{v\}}$, we can first check if $Y(\sigma) \notin \Omega_0$; if this is the case, we output $\{+1,-1\}$ arbitrarily. Otherwise, we sample from the vertex marginal $\mu_{\widehat{F},\beta}(\cdot \mid \sigma)$, which can be done in $O(d)$ time. Sampling from $\mu_{\widehat{F},\beta}(\cdot \mid Y(\sigma) \in \Omega_0)$ is much trickier, but it can be done relying heavily on the structure of the graph $\widehat{F}$; note that the problem of approximately sampling anti-ferromagnetic is computationally hard, even in the bounded degree case. We prove the following.

Lemma 5.2. *For any $\varepsilon \in (0,1)$ and any phase vector $\boldsymbol{\mathcal{Y}} \in \Omega_0$ there is an algorithm that generates a sample from a distribution μ_{ALG} such that $d_{\mathrm{TV}}\left(\mu_{ALG}, \mu_{\widehat{F},\beta}(\cdot \mid \boldsymbol{\mathcal{Y}})\right) \le \varepsilon + \ln(1/\varepsilon)e^{-\Omega(n^{\theta/4})}$ with running time $\mathrm{poly}(|V_{\widehat{F}}|, 1/\varepsilon)$.*

The proof of this lemma is provided in Section 5.3. We are now ready to prove Theorem 1.4 from the introduction.

Proof of Theorem 1.4. The first part of the theorem was proved in Section 4.4.2. For the second part, suppose there is an identity testing algorithm with polynomial running time and sample complexity.

Let $(K = (V_K, E_K), k)$ be the instance of the maximum cut problem with $|V_K| = n^{\theta/12}$. Set $\mu_{\widehat{F},\beta}$ to be the visible distribution and $\mu_{\widehat{F},\beta}(\cdot \mid Y(\sigma) \in \Omega_0)$ to be the hidden one. Suppose $L = \mathrm{poly}(n)$ is the sample complexity of the testing algorithm in this instance. Generate a set $\mathcal{S}$ of L samples from the distribution μ_{ALG} from Lemma 5.2 setting $\varepsilon = 1/(100L)$, so that

$$d_{\mathrm{TV}}\left(\mu_{\mathrm{ALG}}^{\otimes L}, \mu_{\widehat{F},\beta}^{\otimes L}(\cdot \mid Y(\sigma) \in \Omega_0)\right) \le L \cdot d_{\mathrm{TV}}\left(\mu_{\mathrm{ALG}}, \mu_{\widehat{F},\beta}(\cdot \mid Y(\sigma) \in \Omega_0)\right) \le \frac{1}{50},$$

where $\mu_{\mathrm{ALG}}^{\otimes L}$ and $\mu_{\widehat{F},\beta}^{\otimes L}(\cdot \mid Y(\sigma) \in \Omega_0)$ denote the product distributions corresponding to L independent samples from μ_{ALG} and $\mu_{\widehat{F},\beta}(\cdot \mid Y(\sigma) \in \Omega_0)$, respectively.

Our algorithm for solving $(K = (V_K, E_K), k)$ gives $\mathcal{S}$ to the testing algorithm. Recall that our construction ensures that if $\mathsf{max\text{-}cut}(K) < k$, then $\mu_{\widehat{F},\beta}(\cdot \mid Y(\sigma) \in \Omega_0) = \mu_{\widehat{F},\beta}$ and that if $\mathsf{max\text{-}cut}(K) \ge k$ then

$$d_{\mathrm{TV}}\left(\mu_{\widehat{F},\beta}(\cdot \mid Y(\sigma) \in \Omega_0), \mu_{\widehat{F},\beta}\right) = 1 - o(1); \tag{14}$$

see Lemma 5.1.

If $\pi^{\otimes L}$ is the optimal coupling of the distributions $\mu_{\mathrm{ALG}}^{\otimes L}$ and $\mu_{\widehat{F},\beta}^{\otimes L}(\cdot \mid Y(\sigma) \in \Omega_0)$, and $(\mathcal{S}, \mathcal{S}')$ is sampled from $\pi^{\otimes L}$, then $\mathcal{S}' = \mathcal{S}$ with probability at least $49/50$, $\mathcal{S} \sim \mu_{\mathrm{ALG}}^{\otimes L}$ and $\mathcal{S}' \sim \mu_{\widehat{F},\beta}^{\otimes L}(\cdot \mid Y(\sigma) \in \Omega_0)$. Therefore, if (14) holds (i.e., $\mathsf{max\text{-}cut}(K) \ge k$), then

$$\begin{aligned}
&\Pr[\textsc{Tester} \text{ outputs } \textsc{Yes} \text{ when given samples } \mathcal{S} \text{ where } \mathcal{S} \sim \mu_{\mathrm{ALG}}^{\otimes L}] \\
&= \Pr[\textsc{Tester} \text{ outputs } \textsc{Yes} \text{ when given samples } \mathcal{S} \text{ where } (\mathcal{S}, \mathcal{S}') \sim \pi^{\otimes L}] \\
&\le \Pr[\textsc{Tester} \text{ outputs } \textsc{Yes} \text{ when given samples } \mathcal{S}' \text{ where } (\mathcal{S}, \mathcal{S}') \sim \pi^{\otimes L}] + \pi^{\otimes L}(\mathcal{S} \ne \mathcal{S}') \\
&= \Pr[\textsc{Tester} \text{ outputs } \textsc{Yes} \text{ when given samples } \mathcal{S}' \text{where } \mathcal{S}' \sim \mu_{\widehat{F},\beta}^{\otimes L}(\cdot \mid Y(\sigma) \in \Omega_0)] + \pi^{\otimes L}(\mathcal{S} \ne \mathcal{S}') \\
&\le \frac{1}{3} + \frac{1}{50} = \frac{53}{150}.
\end{aligned} \tag{15}$$

Hence, the Tester returns No with probability at least $3/5$ in this case.

Now, when $\mathsf{max\text{-}cut}(K) < k$ and $\mu_{\widehat{F},\beta}(\cdot \mid Y(\sigma) \in \Omega_0) = \mu_{\widehat{F},\beta}$, we can analogously deduce that the Tester returns Yes with probability at least $2/3$. Therefore, our algorithm can solve any maximum

cut instance $(K = (V_K, E_K), k)$ in polynomial time with probability at least 3/5, and the result follows. □

5.3 Sampling Conditional on the Phase Vector: Proof of Lemma 5.2

We start with a number of definitions and facts required to describe and analyze our algorithm to establish Lemma 5.2. The proofs of these facts are provided in Section 5.4. The first lemma states that it essentially suffices to sample from the simpler conditional distribution $\mu_{\widehat{F},\beta}(\cdot \mid Y(\sigma) \in \Omega_0)$.

LEMMA 5.3. $d_{\mathrm{TV}}\left(\mu_{\widehat{F},\beta}(\cdot \mid Y(\sigma) \in \Omega_{st}), \mu_{\widehat{F},\beta}(\cdot \mid Y(\sigma) \in \Omega_0)\right) \leq \frac{1}{2^{n^{\theta/4}}}$.

We call the roots in $\bigcup_{v \in V_F}(R_+^v \cup R_-^v)$ used to connect the degree reducing gadgets *ports*. Let P denote the set of all ports of $\widehat{F}$; we also use $P^v \subset P$ to denote the set of ports of the gadget G_v.

For a configuration $\{+1,-1\}^{P_v}$, let $Z_{G_v,\beta}(\sigma_{P_v})$ denote the sum of the weights of all the configurations on G_v that agree with σ_{P_v}. We will need an approximation algorithm for this quantity and an approximate sampling algorithm for $\mu_{G_v,\beta}(\cdot \mid \sigma_{P_v})$. A **fully polynomial-time randomized approximation scheme (FPRAS)** for $Z_{G_v,\beta}(\sigma_{P_v})$ is an algorithm that for every $\varepsilon > 0$ and $\delta \in (0,1)$ outputs $\hat{Z}$ so that, with probability at least $1-\delta$, $e^{-\varepsilon}\hat{Z} \leq Z_{G_v,\beta}(\sigma_{P_v}) \leq e^{\varepsilon}\hat{Z}$ and runs in time polynomial in $|V_{G_v}|$, $1/\varepsilon$ and $\log(1/\delta)$. A polynomial-time sampling algorithm for $\mu_{G_v,\beta}(\cdot \mid \sigma_{P_v})$ is a randomized algorithm that for every $\varepsilon > 0$ runs in time polynomial in $|V_{G_v}|$ and $1/\varepsilon$ and outputs a sample from a distribution ε-close in TV distance to $\mu_{G_v,\beta}(\cdot \mid \sigma_{P_v})$.

LEMMA 5.4. *Let $\sigma_{P_v} \in \{+1,-1\}^{P_v}$ be an arbitrary spin configuration on P_v. For all sufficiently large $d = O(1)$, with probability $1 - o(1)$ over the choice of the random multi-graph G_v, for all $\beta < 0$ there is an FPRAS for $Z_{G_v,\beta}(\sigma_{P_v})$ and a polynomial-time sampling algorithm for $\mu_{G_v,\beta}(\cdot \mid \sigma_{P_v})$.*

For $\mathcal{Y} \in \Omega$, let $\mu_{P,\beta}(\cdot \mid \mathcal{Y})$ denote the marginal distribution of $\mu_{\widehat{F},\beta}(\cdot \mid \mathcal{Y})$ on P. When $\beta < \beta_c(d)$, in the non-uniqueness regime for the infinite $(d-1)$-ary tree, there are two semi-translation invariant measures, denoted μ^+ and μ^-. These measures can be obtained by conditioning on the leaves at level $2h$ (resp., $2h+1$) to have spin -1, and then taking the weak limits as $h \to \infty$. Let p^+ (resp., p^-) be the probability that the root of the tree is assigned -1 under μ^+ (resp., p^-).

Let $P_+^v = R_+^v \cap P_v$ and $P_-^v = R_-^v \cap P_v$. For $i \in \{+1,-1\}$, we define the following product distribution over configurations $\sigma \in \{+1,-1\}^{P_v}$ on P_v:

$$Q_v^i(\sigma) = (p^i)^{|\sigma^-(-1)\cap P_+^v|}(1-(p^i))^{|\sigma^-(+1)\cap P_+^v|}(p^{-i})^{|\sigma^-(-1)\cap P_-^v|}(1-(p^{-i}))^{|\sigma^-(+1)\cap P_-^v|},$$

where $\sigma^-(i)$ denotes the set of vertices from P_v assigned spin i in σ.

The product distribution Q_v^+ (resp., Q_v^-) is known to be a good approximation for $\mu_{G_v,\beta}(\cdot \mid \mathcal{Y}_v = +1)$ (resp., $\mu_{G_v,\beta}(\cdot \mid \mathcal{Y}_v = -1)$), as formalized in the following lemma. Here $\mathcal{Y}_v$ denote the phase of the gadget G_v.

LEMMA 5.5 (LEMMA 22 [14] AND LEMMA 19 [42]). *Let $\beta < \beta_c(d)$. Then, there exists θ and ψ such that for $s \in \{+1,-1\}$, with probability $1 - o(1)$ over the choice of the random n-vertex multi-graph G_v, for any $\sigma_{P_v} \in \{+1,-1\}^{P_v}$ we have*

$$1 - n^{-2\theta} \leq \frac{\mu_{G_v,\beta}(\sigma_{P_v} \mid \mathcal{Y}_v = s)}{Q_v^s(\sigma_{P_v})} \leq 1 + n^{-2\theta}. \tag{16}$$

Moreover, for $s \in \{+1,-1\}$ we have $\mu_{G_v,\beta}(Y_v = s) \geq \frac{1}{n}$.

Next, for a phase vector $\mathcal{Y} \in \Omega$ we define another product measure this time over configurations $\sigma \in \{+1,-1\}^P$ on P. Let

$$w_P^{\mathcal{Y}}(\sigma) = \prod_{v \in V_F} Q_v^{\mathcal{Y}_v}(\sigma_{P_v}) \prod_{\{u,v\} \in E(P)} e^{\beta \sigma_u \sigma_v},$$

where $E(P)$ is the set of edges with both endpoints in P. Let $Z_P^{\mathcal{Y}} = \sum_{\sigma \in \{+1,-1\}^P} w_P^{\mathcal{Y}}(\sigma)$ and define

$$Q_P^{\mathcal{Y}}(\sigma) = \frac{w_P^{\mathcal{Y}}(\sigma)}{Z_P^{\mathcal{Y}}}.$$

We have the following approximation for $\mu_{P,\beta}(\cdot \mid \mathcal{Y})$ in terms of $Q_P^{\mathcal{Y}}$.

Lemma 5.6 *Let $\beta < \beta_c(d)$. For every $\mathcal{Y} \in \Omega$ and $\sigma \in \{+1,-1\}^P$, we have*

$$\left| \frac{\mu_{P,\beta}(\sigma \mid \mathcal{Y})}{Q_P^{\mathcal{Y}}(\sigma)} - 1 \right| = o(1).$$

Finally, we will also use the following fact.

Lemma 5.7 *Let $\mathcal{Y} \in \Omega$. Suppose $\sigma_P \in \{+1,-1\}^P$ is sampled from $Q_P^{\mathcal{Y}}$ and that $\sigma \in \{+1,-1\}^{V_{\widehat{F}}}$ is then sampled from $\mu_{\widehat{F},\beta}(\cdot \mid \sigma_P)$. Then, the phase vector of σ is $\mathcal{Y}$ with probability at least $1 - e^{-\Omega(n^{3\theta/4})}$.*

We can now provide the proof of Lemma 5.2.

Proof of Lemma 5.2. Let σ^+ denote the all-plus configuration on P. For $\sigma \in \{+1,-1\}^P$, recall that we use $Z_{\widehat{F},\beta}(\sigma, \mathcal{Y})$ to denote the total weight of the configurations of $\widehat{F}$ that agree with σ on P and have phase vector $\mathcal{Y}$. The algorithm is as follows:

(1) Sample $\mathcal{Y} \in \{\xi_{st}^+, \xi_{st}^-\}$ uniformly at random. Note that by ignoring all other phase vectors in Ω_0, the error is at most $1/2^{n^{\theta/4}}$ by Lemma 5.3.
(2) Sample $\sigma_P \in \{+1,-1\}^P$ from a distribution $\varepsilon/3$-close in TV distance to $\mu_{P,\beta}(\cdot \mid \mathcal{Y})$ with the following rejection sampling algorithm:
  2.1 Generate $\sigma_P \in \{+1,-1\}^P$ from the product distribution $Q_P^{\mathcal{Y}}$.
  2.2 Compute the approximation $\hat{Z}(\sigma_P)$ for $Z_{\widehat{F},\beta}(\sigma_P, \mathcal{Y})$ such that

$$\left(1 - \frac{\varepsilon}{10}\right) Z_{\widehat{F},\beta}(\sigma_P, \mathcal{Y}) \le \hat{Z}(\sigma_P) \le \left(1 + \frac{\varepsilon}{10}\right) Z_{\widehat{F},\beta}(\sigma_P, \mathcal{Y});$$

  this can be done in time $\mathrm{poly}(|V_{\widehat{F}}|, 1/\varepsilon)$ with success probability at least $1 - \varepsilon/3$ by Lemma 5.4.
  2.3 Accept σ_P with probability:

$$r(\sigma_P) = \frac{1}{10} \cdot \frac{Q_P^{\mathcal{Y}}(\sigma^+)}{Q_P^{\mathcal{Y}}(\sigma_P)} \cdot \frac{\hat{Z}(\sigma_P)}{\hat{Z}(\sigma^+)}.$$

  2.4 Repeat until accept or exceed $T = c \ln(1/\varepsilon)$ rounds, for a suitable constant $c > 0$, in which case we let $\sigma_P = \sigma^+$.
(3) Sample the configuration of each gadget G_v conditional on the port configuration σ_{P_v} on P_v from a distribution $\frac{\varepsilon}{3|V_F|}$-close to $\mu_{G_v,\beta}(\cdot \mid \sigma_{P_v})$ with the algorithm from Lemma 5.4.
(4) Output the resulting configuration σ.

For the analysis of this algorithm, let us focus first on the rejection sampling process in Step 2. First, note that the process is well-defined since

$$r(\sigma_P) = \frac{1}{10} \cdot \frac{Q_P^{\mathcal{Y}}(\sigma^+)}{Q_P^{\mathcal{Y}}(\sigma_P)} \cdot \frac{\hat{Z}(\sigma_P)}{\hat{Z}(\sigma^+)} \leq \frac{1}{5} \cdot \frac{Q_P^{\mathcal{Y}}(\sigma^+)}{Q_P^{\mathcal{Y}}(\sigma_P)} \cdot \frac{Z_{\widehat{F},\beta}(\sigma_P, \mathcal{Y})}{Z_{\widehat{F},\beta}(\sigma^+, \mathcal{Y})}$$
$$= \frac{1}{5} \cdot \frac{\mu_{P,\beta}(\sigma_P \mid \mathcal{Y})}{Q_P^{\mathcal{Y}}(\sigma_P)} \cdot \frac{Q_P^{\mathcal{Y}}(\sigma^+)}{\mu_{P,\beta}(\sigma^+ \mid \mathcal{Y})} \leq 1,$$

where the last inequality follows from Lemma 5.6. Second, each iteration of the rejection sampling algorithm can be implemented in polynomial time; in particular, we can compute $r(\sigma_P)$ using the FPRAS from Lemma 5.4 to obtain $\hat{Z}(\sigma_P)$ and $\hat{Z}(\sigma^+)$. Finally, we claim that the output of this algorithm is at least $\varepsilon/3$-close to $\mu_{P,\beta}(\cdot \mid \mathcal{Y})$ in TV distance. To see this, note that for each $\sigma_P \in \{+1, -1\}^P$, the probability that process outputs σ_P in one round is

$$Q_P^{\mathcal{Y}}(\sigma_P) r(\sigma_P) = \frac{1}{10} \cdot Q_P^{\mathcal{Y}}(\sigma^+) \cdot \frac{\hat{Z}(\sigma_P)}{\hat{Z}(\sigma^+)} \propto \hat{Z}(\sigma_P).$$

Therefore, conditioned on the algorithm accepting on the first $T = c \ln(1/\varepsilon)$ rounds, the probability that σ_P is the output is

$$\frac{\hat{Z}(\sigma_P)}{\sum_{\sigma'} \hat{Z}(\sigma')} \leq \left(1 + \frac{\varepsilon}{5}\right) \frac{Z_{\widehat{F},\beta}(\sigma_P, \mathcal{Y})}{\sum_{\sigma'} Z_{\widehat{F},\beta}(\sigma', \mathcal{Y})} = \left(1 + \frac{\varepsilon}{5}\right) \mu_{P,\beta}(\sigma_P \mid \mathcal{Y}),$$

and similarly for the lower bound. Moreover, since by Lemma 5.6

$$r(\sigma_P) = \frac{1}{10} \cdot \frac{Q_P^{\mathcal{Y}}(\sigma^+)}{Q_P^{\mathcal{Y}}(\sigma_P)} \cdot \frac{\hat{Z}(\sigma_P)}{\hat{Z}(\sigma^+)} \geq \frac{1}{20} \cdot \frac{Q_P^{\mathcal{Y}}(\sigma^+)}{Q_P^{\mathcal{Y}}(\sigma_P)} \cdot \frac{Z_{\widehat{F},\beta}(\sigma_P, \mathcal{Y})}{Z_{\widehat{F},\beta}(\sigma^+, \mathcal{Y})}$$
$$= \frac{1}{20} \cdot \frac{\mu_{P,\beta}(\sigma_P \mid \mathcal{Y})}{Q_P^{\mathcal{Y}}(\sigma_P)} \cdot \frac{Q_P^{\mathcal{Y}}(\sigma^+)}{\mu_{P,\beta}(\sigma^+ \mid \mathcal{Y})} \geq \frac{1}{100},$$

the probability that the algorithm accepts in the first $T = c \ln(1/\varepsilon)$ rounds is at least

$$1 - \left(1 - \frac{1}{100}\right)^{c \ln(1/\varepsilon)} \geq 1 - \frac{\varepsilon}{10},$$

for a suitable constant $c > 0$.

Now, note that by Lemma 5.7 and union bound, the phase vector of σ agrees with $\mathcal{Y}$ with probability at least $1 - |T| e^{-\Omega(n^{3\theta/4})}$; hence, the output distribution of the algorithm satisfies

$$d_{\mathrm{TV}}\left(\mu_{\mathrm{ALG}}, \mu_{\widehat{F},\beta}(\cdot \mid \mathcal{Y})\right) \leq \frac{\varepsilon}{3} + \frac{\varepsilon}{3} + |V_F| \cdot \frac{\varepsilon}{3|V_F|} + |T| e^{-\Omega(n^{3\theta/4})} + 2^{-n^{\theta/4}} \leq \varepsilon + \ln(1/\varepsilon) e^{-\Omega(n^{\theta/4})},$$

as claimed. □

5.4 Sampling Conditional on the Phase Vector: Proof of Auxiliary Facts

We provide in this section the proofs of Lemmas 5.3, 5.4, 5.6, and 5.7.

Proof of Lemma 5.3. We have

$$d_{\mathrm{TV}}\left(\mu_{\widehat{F},\beta}(\cdot \mid Y(\sigma) \in \Omega_0), \mu_{\widehat{F},\beta}(\cdot \mid Y(\sigma) \in \Omega_{st})\right) = \sum_{\sigma : Y(\sigma) \in \Omega_0 \setminus \Omega_{st}} \mu_{\widehat{F},\beta}(\sigma \mid Y(\sigma) \in \Omega_0).$$

This is the probability of obtaining a phase vector in $\Omega_0 \setminus \Omega_{st}$ under $\mu_{\widehat{F},\beta}(\cdot \mid Y(\sigma) \in \Omega_0$. Observe that among all the cuts of F corresponding to phase vectors in Ω_0, the largest ones are those

corresponding to the phase vectors in Ω_{st}. Hence, following the argument in the proof of Lemma 5.1, we get from the results in [14, 42, 64] that this probability is at least $1 - 1/2^{n^{\theta/4}}$ as desired. □

Proof of Lemma 5.6. Let $\sigma \in \{+1,-1\}^P$. Let $Z_{\widehat{F},\beta}(\sigma, \mathcal{Y})$ be the sum of the weights of all the configurations of $\widehat{F}$ with phase vector $\mathcal{Y}$ that agree with σ in P and, similarly, define $Z_{\widehat{F},\beta}(\mathcal{Y})$ as the sum of the weights of all configurations with phase vector $\mathcal{Y}$, so that

$$\mu_{P,\beta}(\sigma \mid \mathcal{Y}) = \frac{Z_{\widehat{F},\beta}(\sigma,\mathcal{Y})}{Z_{\widehat{F},\beta}(\mathcal{Y})} = \frac{1}{Z_{\widehat{F},\beta}(\mathcal{Y})} \cdot \prod_{v\in V_F} Z_{G_v,\beta}(\sigma_{P_v}, \mathcal{Y}_v) \cdot \prod_{\{u,v\}\in E(P)} e^{\beta\sigma_u\sigma_v}.$$

By Lemma 5.5, we have

$$\begin{aligned}\mu_{P,\beta}(\sigma \mid \mathcal{Y}) &\le \frac{(1+n^{-2\theta})^{|V_F|}}{Z_{\widehat{F},\beta}(\mathcal{Y})} \prod_{v\in V_F} Z_{G_v,\beta}(\mathcal{Y}_v) Q_v^{\mathcal{Y}_v}(\sigma_{P_v}) \prod_{\{u,v\}\in E(P)} e^{\beta\sigma_u\sigma_v} \\ &= \frac{(1+n^{-2\theta})^{|V_F|}}{Z_{\widehat{F},\beta}} \cdot w_P^{\mathcal{Y}}(\sigma) \cdot \prod_{v\in V_F} Z_{G_v,\beta}(\mathcal{Y}_v).\end{aligned}$$

Then,

$$\frac{\mu_{P,\beta}(\sigma \mid \mathcal{Y})}{Q_P^{\mathcal{Y}}(\sigma)} \le (1+n^{-2\theta})^{|V_F|} \cdot \frac{Z_P^{\mathcal{Y}}}{Z_{\widehat{F},\beta}} \cdot \prod_{v\in V_F} Z_{G_v,\beta}(\mathcal{Y}_v).$$

Now,

$$Z_P^{\mathcal{Y}} \cdot \prod_{v\in V_F} Z_{G_v,\beta}(\mathcal{Y}_v) = \sum_{\sigma\in\{+1,-1\}^P} \prod_{\{u,v\}\in E(P)} e^{\beta\sigma_u\sigma_v} \cdot \prod_{v\in V_F} Q_v^{\mathcal{Y}_v}(\sigma_{P_v}) Z_{G_v,\beta}(\mathcal{Y}_v).$$

From (16), we have

$$\frac{1}{1+n^{-2\theta}} \mu_{P_v,\beta}(\sigma_{P_v} \mid \mathcal{Y}_v) \le Q_v^{\mathcal{Y}_v}(\sigma_{P_v}) \le \frac{1}{1-n^{-2\theta}} \mu_{P_v,\beta}(\sigma_{P_v} \mid \mathcal{Y}_v),$$

and so

$$\begin{aligned}&Z_P^{\mathcal{Y}} \cdot \prod_{v\in V_F} Z_{G_v,\beta}(\mathcal{Y}_v) \\ &\quad\le \frac{1}{(1-n^{-2\theta})^{|V_F|}} \sum_{\sigma\in\{+1,-1\}^P} \prod_{\{u,v\}\in E(P)} e^{\beta\sigma_u\sigma_v} \cdot \prod_{v\in V_F} \mu_{P_v,\beta}(\sigma_{P_v} \mid \mathcal{Y}_v) Z_{G_v,\beta}(\mathcal{Y}_v) \\ &\quad= \frac{1}{(1-n^{-2\theta})^{|V_F|}} Z_{\widehat{F},\beta}.\end{aligned}$$

Thus, we have obtained the upper bound

$$\frac{\mu_{P,\beta}(\sigma \mid \mathcal{Y})}{Q_P^{\mathcal{Y}}(\sigma)} \le \left(\frac{1+n^{-2\theta}}{1-n^{-2\theta}}\right)^{|V_F|},$$

and we can deduce analogously that

$$\frac{\mu_{P,\beta}(\sigma \mid \mathcal{Y})}{Q_P^{\mathcal{Y}}(\sigma)} \ge \left(\frac{1-n^{-2\theta}}{1+n^{-2\theta}}\right)^{|V_F|}.$$

Recall that $|V_F| = n^{\theta/12} + 2$, so that

$$1 - o(1) \le \frac{\mu_{P,\beta}(\sigma \mid \mathcal{Y})}{Q_P^{\mathcal{Y}}(\sigma)} \le 1 + o(1)$$

and the result follows. □

Proof of Lemma 5.7. Consider a gadget G_v of $\widehat{F}$ such that $\mathcal{Y}_v = +1$. Let $u \in P_v^+$. We claim that since $p^+ > 1/2$, then $Q_P^{\mathcal{Y}}(\sigma_u = -1) > 1/2$. To see this, let w be the neighbor of u in P and suppose that the phase of the gadget containing w is +1. Then,

$$\frac{Q_P^{\mathcal{Y}}(\sigma_u = -1)}{Q_P^{\mathcal{Y}}(\sigma_u = +1)} = \frac{(p^+)^2 e^{\beta} + p^+(1-p^+)e^{-\beta}}{(1-p^+)^2 e^{\beta} + p^+(1-p^+)e^{-\beta}} > 1,$$

which implies that $Q_P^{\mathcal{Y}}(\sigma_u = -1) > 1/2$. An analogous calculation shows that the same holds when the gadget containing w is in the -1 phase. With the same reasoning, we can similarly deduce that when $u \in P_v^-$, then $Q_P^{\mathcal{Y}}(\sigma_u = +1) > 1/2$. This implies by a Chernoff bound that if $\sigma_P \sim Q_P^{\mathcal{Y}}$ and $\mathcal{Y}_v = +1$, then there exists a constant $\delta > 0$ such that

$$|\sigma_{P_v}^-(-1) \cap P_v^+| - |\sigma_{P_v}^-(-1) \cap P_-^v| \geq \delta |P_v|,$$

and

$$|\sigma_{P_v}^-(+1) \cap P_-^v| - |\sigma_{P_v}^-(+1) \cap P_+^v| \geq \delta |P_v|,$$

with probability at least $1 - \exp(-\Omega(|P_v|))$.

Now,

$$\begin{aligned}\frac{\mu_{G_v,\beta}(\mathcal{Y}_v = +1 \mid \sigma_{P_v})}{\mu_{G_v,\beta}(\mathcal{Y}_v = -1 \mid \sigma_{P_v})} &= \frac{\mu_{G_v,\beta}(\sigma_{P_v} \mid \mathcal{Y}_v = +1)}{\mu_{G_v,\beta}(\sigma_{P_v} \mid \mathcal{Y}_v = -1)} \cdot \frac{\mu_{G_v,\beta}(\mathcal{Y}_v = +1)}{\mu_{G_v,\beta}(\mathcal{Y}_v = -1)} \\ &\geq \left(1 - \frac{c}{n^{2\theta}}\right) \cdot \frac{Q_v^+(\sigma_{P_v})}{Q_v^-(\sigma_{P_v})} \cdot \frac{\mu_{\widehat{F},\beta}(\mathcal{Y}_v = +1)}{\mu_{\widehat{F},\beta}(\mathcal{Y}_v = -1)}\end{aligned}$$

by Lemma 5.5. Lemma 5.5 also implies that $\frac{\mu_{G_v,\beta}(\mathcal{Y}_v=+1)}{\mu_{G_v,\beta}(\mathcal{Y}_v=-1)} \geq \frac{1}{n}$. Moreover, from the definition of Q_v^+ and Q_v^- we have

$$\begin{aligned}\frac{Q_v^+(\sigma_{P_v})}{Q_v^-(\sigma_{P_v})} &= \frac{(p^+)^{|\sigma_{P_v}^-(-1)\cap P_+^v|}(1-p^+)^{|\sigma_{P_v}^-(+1)\cap P_+^v|}(p^-)^{|\sigma_{P_v}^-(-1)\cap P_-^v|}(1-p^-)^{|\sigma_{P_v}^-(+1)\cap P_-^v|}}{(p^-)^{|\sigma_{P_v}^-(-1)\cap P_+^v|}(1-p^-)^{|\sigma_{P_v}^-(+1)\cap P_+^v|}(p^+)^{|\sigma_{P_v}^-(-1)\cap P_-^v|}(1-p^+)^{|\sigma_{P_v}^-(+1)\cap P_-^v|}} \\ &= \left(\frac{p^+}{p^-}\right)^{|\sigma_{P_v}^-(-1)\cap P_+^v| - |\sigma_{P_v}^-(-1)\cap P_-^v|} \left(\frac{1-p^-}{1-p^+}\right)^{|\sigma_{P_v}^-(+1)\cap P_-^v| - |\sigma_{P_v}^-(+1)\cap P_+^v|} \\ &\geq a^{\delta|P_v|},\end{aligned}$$

for a suitable constant $a > 1$ since $p^+ > 1/2$ and $p^- < 1/2$. This implies that for a suitable constant $c_0 > 0$, we have

$$\mu_{G_v,\beta}(\mathcal{Y}_v = +1 \mid \sigma_{P_v}) \geq 1 - \frac{c_0 n}{a^{\delta|P_v|}} \geq 1 - \frac{c_0 n}{a^{\delta n^{3\theta/4}}},$$

since $|P_v| \geq n^{3\theta/4}$. Finally, we note that

$$\mu_{\widehat{F},\beta}(\mathcal{Y} \mid \sigma_P) = \prod_{v \in V_F} \mu_{G_v,\beta}(\mathcal{Y}_v \mid \sigma_{P_v}) \geq \left(1 - \frac{c_0 n}{a^{\delta n^{3\theta/4}}}\right)^{|V_F|} \geq 1 - \frac{c_0 n^{1+\theta/12}}{a^{\delta n^{3\theta/4}}}$$

since $|V_F| = n^{\theta/12} + 2$, and the result follows. □

5.4.1 Sampling from the Degree Reducing Gadget. We focus now in proving Lemma 5.4. Let $\mu_{G,\beta}$ and $Z_{G,\beta}$ be the anti-ferromagnetic Ising distribution and its corresponding partition function on a degree reducing gadget G. We need to show to prove Lemma 5.4 how to approximately sample from $\mu_{G,\beta}$ and how to compute $Z_{G,\beta}$ when conditioning on an arbitrary configuration on the ports P of G. (Note that with a slight abuse of notation we are using P for the set of ports of a single gadget G throughout this section.) Let $\tau\{+1,-1\}^P$ be a configuration on the ports. Let $Z^{\tau}_{G,\beta}$ and $\mu^{\tau}_{G,\beta}$ denote the conditional Ising distribution and the corresponding partition function.

To establish Lemma 5.4 we provide two different algorithms: one based on the recent results from [54] that works when $\beta \geq -1/\sqrt{10d}$, and another based on polymer models that works when $\beta \leq -\frac{c\ln d}{d}$ (for a sufficiently large constant $c > 0$), so that each value of the regime $\beta < 0$ is covered by one of these algorithms provided d is large enough.

Both algorithms use facts about the spectrum of the multi-graph induced by $V_G \setminus P$. Hence, let $H = (V_H, E_H)$ be the multi-graph that results from removing P from V_G. Let A_H be the adjacency matrix for the multi-graph H, that is, $A_H(u,v)$ is the multiplicity of the edge $\{u,v\}$ in H. For $S \subseteq V_H$, let $\partial_e(S)$ be the set of edges from E_H with one endpoint in S and one $V_H \setminus S$. For any real symmetric matrix Q, let $\lambda_i(Q)$ denote its ith largest eigenvalue.

Fact 5.8. *Suppose $d = O(1)$. Then, with probability $1 - o(1)$:*

(1) $d - 2\sqrt{d} - 2 \leq \lambda_1(A_H) \leq d + 2\sqrt{d}$;
(2) $-d - 2\sqrt{d} \leq \lambda_{|V_H|}(A_H) \leq -d + 2\sqrt{d} + 2$;
(3) For $i \geq 2$, $-4\sqrt{d} - 2 \leq \lambda_2(A_H) \leq 4\sqrt{d} + 2$;
(4) For every $S \subseteq V_H$ such that $|S| \leq |V_H|/2$, we have $|\partial_e(S)| \geq \frac{d-4\sqrt{d}-2}{2}|S|$.

Proof. Consider the symmetric matrices A, B, and T, of dimension $|V_H| \times |V_H|$ defined by:

— $A(u,v) = \kappa$ if $u \in W_+ \cup U_+$ and $v \in W_- \cup U_-$ (or vice versa) and the edge $\{u,v\}$ appears κ times in $\bigcup_{i=1}^{d-1} M_i$; all other entries of A are 0.
— $B(u,v) = 1$ if $u \in W_+$ and $v \in W_-$ (or vice versa) and $\{u,v\} \in M$; all other entries of B are 0.
— $T(u,v) = 1$ if $\{u,v\} \in E_H$ and either u or v (or both) are vertices in $V_H \setminus (W_+ \cup U_+ \cup W_- \cup U_-)$; all other entries of T are 0.

Note that $A_H = A + B + T$, so it follows from Weyl's inequality (see [39]) that

$$\lambda_i(A) + \lambda_{|V_H|}(B) + \lambda_{|V_H|}(T) \leq \lambda_i(A_H) \leq \lambda_i(A) + \lambda_1(B) + \lambda_1(T).$$

From Theorem 4 in [13] and contiguity (see Theorem 4 and Corollary 1 in [58]), we know that A has real eigenvalues $\lambda_1(A) \geq \lambda_2(A) \geq \cdots \geq \lambda_{|V_H|}(A)$, where $\lambda_1(A) = d - 1$, $\lambda_i(A) = -\lambda_{|V_H|-i+1}(A)$, and $2\sqrt{d} - 1 \leq \lambda_2(A) \leq 2\sqrt{d} + 1$ with probability $1 - o(1)$. The matrix B has eigenvalues 1 and -1. Also, all the eigenvalues of the matrix T are real and belong to the interval $[-2\sqrt{d}, 2\sqrt{d}]$ (see Theorem 3 in [44]). Combining these facts, we obtain parts 1, 2, and 3; part 4 follows from Cheeger's inequality (for multi-graphs). □

Proof of Lemma 5.4. Let J be a $|V_H| \times |V_H|$ matrix indexed by the vertices of H with entries $J(u,v) = \beta \cdot A_H(u,v)$ for $u \neq v$ and $J(u,u) = \alpha$ where α is a real number we choose later. Let $\partial P \subset V_H$ be the set of vertices of H that were incident to P in G. Define a magnetic field h by letting $h_v = \beta$ (resp., $h_v = -\beta$) if $v \in \partial P$ and the vertex adjacent to v in P has +1 (resp., −1) spin in τ; we set $h_v = 0$ otherwise. The Ising model on H with edge interaction β and external field h assigns to

each configuration σ on H probability:

$$\mu_{H,\beta}(\sigma) = \frac{1}{Z_{H,\beta}} \exp\Big(\beta \sum_{\{u,v\}\in E_H} \sigma_u\sigma_v + \sum_{v\in V_H} h_v\sigma_v\Big) = \frac{1}{\hat{Z}_{H,\beta}} \exp\Big(\frac{1}{2}\langle\sigma, J\sigma\rangle + \langle h,\sigma\rangle\Big), \tag{17}$$

where $\hat{Z}_{H,\beta} = e^{\alpha|V_H|}Z_{H,\beta}$, and we interpret σ and h as vectors indexed by the vertices of H. By construction, $Z^{\tau}_{G,\beta} = Z_{H,\beta}$ and $\mu^{\tau}_{G,\beta}(\sigma) = \mu_{H,\beta}(\sigma)$ for every $\sigma \in \{+,-\}^{V_H}$.

The matrix J has real eigenvalues $\lambda_1(J) \geq \lambda_2(A) \geq \cdots \geq \lambda_{|V_H|}(J)$. To bound the spectrum of J, we note that $J = \beta A_H + \alpha I$ and so $\lambda_i(J) = \beta\lambda_i(A_H) + \alpha$. Hence, setting $\alpha = -\beta(4\sqrt{d}+2)$ and assuming that $0 > \beta \geq -1/(10\sqrt{d})$ and that d is sufficiently large, we obtain from Fact 2 that with probability $1 - o(1)$: $\lambda_1(J) = \Theta(\sqrt{d})$, $\lambda_{|V(H)|}(J) = -\Theta(\sqrt{d})$ and that every other eigenvalue of J is in the interval $[0,1]$. Then, Theorem 1.1 from [54] implies that:

(1) There is an algorithm that with probability $1 - e^{-|V_H|}$ produces an e^{ε}-multiplicative approximation for $\hat{Z}_{H,\beta} = Z^{\tau}_{G,\beta}$ with running time $\mathrm{poly}(|V_H|, 1/\varepsilon)$; and
(2) There is an algorithm to sample from a distribution within ε TV distance from $\mu_{H,\beta} = \mu^{\tau}_{G,\beta}$ with running time $\mathrm{poly}(|V_H|, \log(1/\varepsilon))$.

Hence, we have established the result for the case when $\beta \geq -\frac{1}{10\sqrt{d}}$. We consider next the case when $\beta \leq -\frac{c\ln d}{d}$, for a suitably large constant $c > 0$. For this, we introduce the notion of polymer models.

For a fixed configuration τ in P, let $P^+ \subset \partial P$ be the set of vertices of ∂P adjacent to a vertex assigned "+" in τ; define $P^- \subset \partial P$ similarly. For a configuration σ on H, let $p^+(\sigma)$ (resp., $p^-(\sigma)$) denote the number of vertices from P^+ (resp., P^-) that are assigned spin -1 (resp., $+1$) in σ. Let also $D(H,\sigma)$ denote the number of edges incident to two vertices with different spins in σ. Then, we can renormalize the Ising distribution (17) as

$$\mu_{H,\beta}(\sigma) = \frac{1}{\tilde{Z}_H} e^{-2\beta(D(H,\sigma)+p^+(\sigma)+p^-(\sigma))} =: \frac{w(\sigma)}{\tilde{Z}_{H,\beta}},$$

where $\tilde{Z}_{H,\beta} = e^{-\beta(|E_H|+|\partial P|)}Z_{H,\beta}$.

Let $\Omega = \{+1,-1\}^{V_H}$. Observe the graph H is bipartite with partition (L,R) where $L \cup R = V_H$ and $|L| = |R|$. Let $\Omega^{\pm} \subset \Omega$ be the subset of configurations where the number vertices that are assigned $+1$ in L and -1 in R are more than $|V_H|/2$. Define $\Omega^{\mp}$ analogously. Let $Z^{\pm}_H = \sum_{\sigma\in\Omega^{\pm}} w(\sigma)$ and $Z^{\mp}_H = \sum_{\sigma\in\Omega^{-=}} w(\sigma)$ so that $\tilde{Z}_{H,\beta} = Z^{\pm}_H + Z^{\mp}_H$. We define a polymer model whose partition function will serve as a good approximation for $Z^{\pm}_H$ and $Z^{\mp}_H$.

We say $\gamma \subset V_H$ is a polymer if the subgraph induced by γ is connected and $|\gamma| < |V_H|/2$. Two polymers are compatible if the graph distance between them is at least 2. Let $\mathcal{G}$ be the family of all sets of mutually compatible polymers. To each polymer γ we assign the weight

$$w_\gamma = e^{-2\beta(-|\partial_e(\gamma)|+|P^-\cap L\cap\gamma|+|P^+\cap R\cap\gamma|-|P^+\cap L\cap\gamma|-|P^-\cap R\cap\gamma|)}.$$

Define the polymer partition function

$$\Phi = \sum_{\Gamma\in\mathcal{G}} \prod_{\gamma\in\Gamma} w_\gamma.$$

We say $S \subset V_H$ is sparse if every connected component of S has size less than $|V_H|/2$. Note that there is a one-to-one correspondence between the sparse subsets of V_H and polymer configurations

from $\mathcal{G}$. Then:

$$\begin{aligned}\hat{\Phi} &:= e^{-2\beta(|E_H|+|P^-\cap R|+|P^+\cap L|)} \cdot \Phi \\ &= e^{-2\beta(|E_H|+|P^-\cap R|+|P^+\cap L|)} \sum_{\Gamma\in\mathcal{G}} \prod_{\gamma\in\Gamma} w_\gamma \\ &= \sum_{S \text{ sparse}} e^{-2\beta(|E_H|-|\partial_e(S)|+|P^-\cap R|+|P^+\cap L|+|P^-\cap L\cap S|+|P^+\cap R\cap S|-|P^+\cap L\cap S|-|P^-\cap R\cap S|)} \\ &= \sum_{S \text{ sparse}} e^{-2\beta(|E_H|-|\partial_e(S)|+|((R\setminus S)\cup(S\cap L))\cap P^-|+|((L\setminus S)\cup(S\cap R))\cap P^+|)}.\end{aligned}$$

Now, we say $S \subset V_H$ is small if $|S| < |V_H|/2$ (otherwise we say it is large), so that

$$Z_H^{\mp} = \sum_{S \text{ small}} e^{-2\beta(|E_H|-|\partial_e(S)|+|((R\setminus S)\cup(S\cap L))\cap P^-|+|((L\setminus S)\cup(S\cap R))\cap P^+|)}.$$

Hence,

$$0 \le \hat{\Phi} - Z_H^{\mp} \le \sum_{S \text{ sparse, large}} e^{-2\beta(|E_H|-|\partial_e(S)|+|((R\setminus S)\cup(S\cap L))\cap P^-|+|((L\setminus S)\cup(S\cap R))\cap P^+|)}.$$

If S is sparse, by part 4 of Fact 5.8, each connected component S_i of S satisfies $\partial_e(S_i) \ge \theta|S_i|$ with $\theta = \frac{d-4\sqrt{d}-2}{2}$. Summing over the components of S we get $\partial_e(S) \ge \theta|S| \ge \theta|V_H|/2$ when S is large. Then,

$$|\hat{\Phi} - Z_H^{\mp}| \le \sum_{S \text{ sparse, large}} e^{-2\beta(|E_H|-\theta|V_H|/2+|((R\setminus S)\cup(S\cap L))\cap P^-|+|((L\setminus S)\cup(S\cap R))\cap P^+|)}$$

and since $Z_H^{\mp} \ge e^{-2\beta(|E_H|+|L\cap P^+|+|R\cap P^-|)}$ and $|S| < |V_H|/2$, we have

$$\left|1 - \frac{\hat{\Phi}}{Z_H^{\mp}}\right| \le 2^{|V_H|} e^{-2\beta(-\theta|V_H|/2+|V_H|/2)} \le e^{-|V_H|}, \tag{18}$$

provided $\theta > 1$ and $-\beta \ge \frac{1+\ln 2}{\theta-1}$. An analogous argument yields the same bound for $Z_H^{\pm}$.

Our goal now is to use Theorem 8 from [50] to obtain an approximation for Φ and consequently for $\hat{\Phi}$, $Z_H^{\mp}$, $Z_H^{\pm}$ and ultimately for $Z_{H,\beta} = Z_H^{\pm} + Z_H^{\mp}$. For this, it suffices to check that our polymer model satisfies the so-called Kotecký–Preiss condition (see, e.g., Equation (3) from [50]). This condition requires that for every polymer γ:

$$\sum_{\gamma' : d(\gamma,\gamma')\le 1} w_{\gamma'} e^{2|\gamma'|} \le |\gamma|, \tag{19}$$

where $d(\cdot,\cdot)$ denotes graph distance. First note that

$$w_{\gamma'} = e^{-2\beta(-|\partial_e(\gamma')|+|P^-\cap L\cap\gamma'|+|P^+\cap R\cap\gamma'|-|P^+\cap L\cap\gamma'|-|P^-\cap R\cap\gamma'|)} \le e^{-2\beta(-\theta/2+1)|\gamma'|}.$$

Hence,

$$\sum_{\gamma' : d(\gamma,\gamma')\le 1} w_{\gamma'} e^{2|\gamma'|} \le \sum_{\gamma' : d(\gamma,\gamma')\le 1} e^{|\gamma'|(2-2\beta(-\theta/2+1))} \le \sum_{v\in\gamma\cup\partial_v(\gamma)} \sum_{\gamma' : v\in\gamma'} e^{|\gamma'|(2-2\beta(-\theta/2+1))}.$$

The number of polymers of size k that contain a given vertex is at most $(ed)^k$ (see Lemma 2.1 in [43]), so

$$\sum_{\gamma' : v\in\gamma'} e^{|\gamma'|(2-2\beta(-\theta/2+1))} \le \sum_{t\ge 1} \left(de^{(3-2\beta(-\theta/2+1))}\right)^t \le \frac{1}{d+1}$$

when $-\beta \geq \frac{3+\ln(d(d+2))}{\theta-2}$ and $\theta > 2$. (Note that the latter is true when d is large enough.) Since $|\gamma \cup \partial_v(\gamma)| \leq (d+1)|\gamma|$, (19) follows. Hence, for sufficiently large $d = O(1)$, for a suitable constant $c > 0$, Theorem 8 from [50] gives an FPTAS for Φ when $-\beta \geq \frac{c \ln d}{d}$. This yields the desired FPTAS for $Z_{H,\beta}$. Note that if the desired approximation factor is smaller than $e^{-|V_H|}$, which is the best approximation for $Z_{H,\beta}$ we could obtain using the polymer function Φ (see (18)), then we could instead use brute force for counting and sampling, since the running time would be allowed to be exponential in $|V_H|$. Finally, Theorem 9 from [50] gives the polynomial-time approximate sampling algorithm for the distribution

$$\nu(\Gamma) = \frac{\prod_{\gamma \in \Gamma} w_\gamma}{\Phi}.$$

Once a polymer configuration Γ is sampled from ν, it can be easily transformed into an Ising configuration by setting the vertices in $L \setminus \Gamma$ and $R \cap \Gamma$ to +1 with probability $\frac{Z_H^+}{Z_H^+ + Z_H^-}$ and all other vertices to -1, and doing the opposite with the remaining probability. □

6 Statistical Lower Bounds

In this section we establish lower bounds on the number of samples required to perform uniformity testing over the hypercube $\{0,1\}^n$ (i.e., $k = 2$), with a focus on comparisons between testing for KL divergence and TV distance, and between testing with Coordinate Oracle + General Oracle and Subcube Oracle. Throughout this section, we assume $k = 2$ and $\mathcal{K} = \{0,1\}$. Let u_n denote the uniform distribution over $\mathcal{X}_n = \{0,1\}^n$ for an integer $n \in \mathbb{N}^+$. We omit the subscript n when it is clear from context.

6.1 Statistical Lower Bounds for Coordinate Oracle Model: Proof Sketch

We provide next an overview of our proof approach for Theorem 1.5 in which we establish an information-theoretic lower bound for uniformity testing over the binary hypercube $\{0,1\}^n$ in the Coordinate Oracle model; our proof of Theorem 1.7 for the Subcube Oracle is similar and we comment on it below. Our proof follows a well-known strategy. We construct a family of "bad" distributions $\mathcal{B}$, each of which has TV distance (or KL divergence) at least ε from the uniform distribution over $\{0,1\}^n$. Then, the lower bounds follow from, roughly speaking, the fact that the joint distributions of L independent samples from the uniform distribution and of L independent samples from a distribution from $\mathcal{B}$ (chosen uniformly at random) are close to each other.

Such an argument works nicely for non-adaptive identity testing algorithms, where the queries are pre-determined before receiving any sample. In the presence of conditional sampling oracles, we are required to show the lower bounds for adaptive testing algorithms which is necessary with, and so we need to consider the whole *query history*, as in [19, 60]. Informally speaking, a query history is a sequence of queries that the testing algorithm asks the oracle along with the outputs from the oracle. Each step, the tester determines, possibly at random, a new query based on all previous queries that have been asked and the corresponding outputs from the oracle. The output of the testing algorithm can be viewed as a function (possibly randomized) of the query history.

Consequently, we need to show that the following two processes generate close query histories in TV distance. In the first process, in each step the algorithm computes a query and the oracle outputs a sample using the uniform distribution. In the second one, we first pick a bad distribution $\pi \in \mathcal{B}$ uniformly at random, and then the oracle outputs samples using π. To show that the two generated query histories are close, we use ideas from [19] and also the so-called *hybrid argument* in cryptography (see, e.g., [46]). For each $\ell \leq L$, we consider a hybrid query history where the first ℓ queries are answered by the oracle using the uniform distribution, while the other $L - \ell$ queries are answered by a single $\pi \in \mathcal{B}$ chosen uniformly at random. It then suffices to show that

every pair of "adjacent" hybrid query histories are close to each other. Since two adjacent hybrid query histories differ only at one step, this can be done for a carefully constructed family $\mathcal{B}$ of bad distributions.

Our family of bad distributions for the Coordinate Oracle model is the same as in earlier works [18, 31]. Each distribution in $\mathcal{B}$ is constructed by taking a perfect matching of all coordinates (we may assume n is even) and considering the distribution such that coordinates from different matched pairs are independent of each other while within each pair the two coordinates are correlated with covariance $\Theta(\varepsilon/\sqrt{n})$. Then, one can show that the joint distribution of $O(n/\varepsilon^2)$ samples from the uniform distribution and that from a bad distribution corresponding to a uniformly random perfect matching are close to each other. Furthermore, the Coordinate Oracle does not help in the following sense: For the uniform distribution, the Coordinate Oracle outputs uniform Bernoulli $\mathrm{Ber}(1/2)$ random variables, and for any distribution from $\mathcal{B}$ it outputs a sample from $\mathrm{Ber}(1/2+\xi)$ or $\mathrm{Ber}(1/2-\xi)$, where $\xi = \Theta(\varepsilon/\sqrt{n})$. We show that to distinguish between a sequence of $\mathrm{Ber}(1/2)$ and a sequence of adaptively chosen $\mathrm{Ber}(1/2 \pm \xi)$, one needs $\Omega(1/\xi^2) = \Omega(n/\varepsilon^2)$ samples in this specific setting; this is proved in Section 6.3.

Finally, we briefly describe our construction of the family $\mathcal{B}$ for establishing Theorem 1.7. It is inspired by studying approximate tensorization of entropy. In particular, our identity testing algorithm fails within $O(n/\varepsilon)$ steps if for most of pairs (i,x) it holds $\pi_i(\cdot \mid x) = \mathrm{Ber}(1/2)$ but only for an ε/n fraction of the pairs the KL divergence is large, which means we need $\Omega(n/\varepsilon)$ steps to be able to see it. For Subcube Oracle we would like to construct bad distributions with similar behavior. Namely, for most (random) conditionings on a (random) subset of coordinates, the conditional distribution is the same as what one gets from the uniform distribution, and with probability $O(\varepsilon/n)$ the KL divergence between the two conditional distributions is as large as $\Theta(n)$. We achieve this using the following type of construction. We pick a random subset A of size t such that $2^t = O(n/\varepsilon)$, and pick a vector $\sigma \in \mathcal{K}^n$. To generate a sample x from the bad distribution $\pi = \pi_{A,\sigma}$, we first sample x_A uniformly at random. If $x_A \neq \sigma_A$ then the other coordinates are sampled randomly, but if $x_A = \sigma_A$ then we take $x = \sigma$. One can check, with careful calculations, that such bad distributions satisfy our requirements. In particular, while the KL divergence for any such bad distribution to the uniform distribution is ε, the TV distance is ε/n instead, and so a $\Omega(n/\varepsilon)$ lower bound is not a surprise for this construction of family of bad distributions.

6.2 Uniformity Testing with Subcube Oracle for KL Divergence

In this subsection we consider uniformity testing over $\{0,1\}^n$ with access to Subcube Oracle for KL divergence, and give an information-theoretic lower bound of $\Omega(n/\varepsilon)$ on the number of samples needed.

Let Alg denote an arbitrary uniformity testing algorithm (possibly randomized and adaptive), and for simplicity let $\mathsf{Ora}[\pi]$ denote the Subcube Oracle with respect to a distribution π over $\{0,1\}^n$.

Definition 6.1. A *pinning* τ is a partial configuration on a subset of coordinates, namely $\tau \in \{0,1\}^\Lambda$ where $\Lambda \subseteq [n]$.

Observe that pinnings are exactly the inputs to the subcube oracle $\mathsf{Ora}[\pi]$.

Definition 6.2 (Query History for Subcube Oracle*).* Let $\mathcal{T}$ be the collection of all pinnings on all subsets of coordinates. For $L \in \mathbb{N}^+$, define the *(subcube) query history* with respect to Alg and $\mathsf{Ora}[\pi]$ of length L to be the random vector in $(\mathcal{T} \times \{0,1\}^n)^L$ generated as follows:

—For $i = 1, \ldots, L$:
  – Alg receives $((\tau_1, x_1), \ldots, (\tau_{i-1}, x_{i-1}))$ as input and generates $\tau_i \in \mathcal{T}$ (randomly) as output;

– $\mathsf{Ora}[\pi]$ receives τ_i as input and generates $x_i \in \{0,1\}^n$ as output.
—The (subcube) query history is $H = ((\tau_1, x_1), \ldots, (\tau_L, x_L))$.

Note that the output of Alg with sample complexity L is a (randomized) function of the query history H of length L.

Our main theorem is stated as below in terms of the query history, from which Theorem 1.6 follows immediately.

Theorem 6.3. *Let $n \in \mathbb{N}^+$ be a sufficiently large integer and $\varepsilon > 0$ be a real. Let $u = u_n$ denote the uniform distribution over $\{0,1\}^n$. There is no algorithm which can achieve the following properties using only $L \le n/(64\varepsilon)$ samples:*

—$\Pr_H(\mathsf{output} = \mathsf{Yes}) \ge 2/3$ *for a random query history H of length L with respect to* Alg *and* $\mathsf{Ora}[u]$*;*
—$\Pr_{H'}(\mathsf{output} = \mathsf{No}) \ge 2/3$ *for a random query history H' of length L with respect to* Alg *and* $\mathsf{Ora}[\pi]$ *where π is any distribution such that $D_{\mathrm{KL}}(\pi \,\|\, u) \ge \varepsilon$.*

Our plan, as in many previous works, is to construct a family $\mathcal{B}$ of bad distributions that are all ε far away from u in KL divergence, such that when picking a bad distribution from $\mathcal{B}$ uniformly at random and drawing limited number of samples, the joint distributions of these samples are close to that of samples drawn from u. We present now our construction of the bad family $\mathcal{B}$. Let $t = \lceil \log_2(n/\varepsilon) \rceil - 3$ for sufficiently large n. For any $A \subseteq [n]$ with $|A| = t$ and any $\sigma \in \{0,1\}^n$, define the distribution $\pi_{A,\sigma}$ in the following way. A sample from $\pi_{A,\sigma}$ is generated by:

—For each $i \in A$ independently sample $x_i \in \{0,1\}$ uniformly at random;
—If $x_A \ne \sigma_A$, then for each $j \in [n] \setminus A$ independently sample $x_j \in \{0,1\}$ uniformly at random and output x;
—If $X_A = \sigma_A$ then output $x = \sigma$.

We remark that all steps are independent. Finally, we define

$$\mathcal{B} = \left\{ \pi_{A,\sigma} : A \in \binom{[n]}{t}, \sigma \in \{0,1\}^n \right\}.$$

We first show that the distributions in $\mathcal{B}$ are all bad in the sense that their KL divergence to the uniform distribution is at least ε. A key intuition in our construction of $\pi_{A,\sigma}$ here is that while the KL divergence $D_{\mathrm{KL}}(\pi_{A,\sigma} \,\|\, u) = \Theta(\varepsilon)$, the TV distance is much smaller than ε and is $d_{\mathrm{TV}}(\pi_{A,\sigma}, u) = \Theta(2^{-t}) = \Theta(\varepsilon/n)$. Hence, intuitively, it will take $\Theta(1/d_{\mathrm{TV}}(\pi, u)) = \Theta(n/\varepsilon)$ samples to test between the family $\mathcal{B}$ and the uniform distribution u.

Claim 6.4. *For all $\pi \in \mathcal{B}$ one has*

$$D_{\mathrm{KL}}(\pi \,\|\, u) \ge \varepsilon.$$

Proof. Suppose $\pi = \pi_{A,\sigma} \in \mathcal{B}$ is a bad distribution. By definition we have $\pi(x) = u(x) = 2^{-n}$ if $x_A \ne \sigma_A$, and $\pi(\sigma) = 2^{-t}$. Hence, we get

$$D_{\mathrm{KL}}(\pi \,\|\, u) = \pi(\sigma) \ln\left(\frac{\pi(\sigma)}{u(\sigma)}\right) = \frac{\ln 2}{2^t}(n-t) \ge \frac{2\varepsilon}{n}(n-t) \ge \varepsilon,$$

for n sufficiently large. □

Define H to be the random query history of length L with respect to Alg and $\mathsf{Ora}[u]$, and let output denote the random output with respect to H and Alg. Define H' to be the random query history of length L generated by

—Pick $\pi \in \mathcal{B}$ uniformly at random;
—Let H' be the random query history of length L with respect to Alg and Ora$[\pi]$.

Further, let output′ denote the random output with respect to H' and Alg. Our goal is to show that the TV distance between the two query histories H and H' is small and therefore by the data processing inequality the TV distance between output and output′ is also small so the two properties in Theorem 6.3 cannot simultaneously hold.

Lemma 6.5. *For the family $\mathcal{B}$ of bad distributions, query histories H, H' of length $L \le n/(64\varepsilon)$, and* output, output′ *defined as above, we have*

$$d_{\mathrm{TV}}\left(\mathsf{output}, \mathsf{output}'\right) \le d_{\mathrm{TV}}\left(H, H'\right) \le \frac{1}{4}.$$

We present next the proof of Theorem 6.3 provided in Lemma 6.5. The proof of the latter is postponed to Section 6.2.1.

Proof of Theorem 6.3. Suppose for sake of contradiction that Alg satisfies both properties as in Theorem 6.3. Then for the family $\mathcal{B}$ of bad distributions, query histories H, H' of length $L \le n/(64\varepsilon)$, and output, output′ defined as earlier, we know from these two properties that

$$\Pr\left(\mathsf{output} = \mathsf{Yes}\right) \ge 2/3 \quad \text{and} \quad \Pr\left(\mathsf{output}' = \mathsf{Yes}\right) \le 1/3.$$

This implies $d_{\mathrm{TV}}\left(\mathsf{output}, \mathsf{output}'\right) \ge 1/3$ which contradicts Lemma 6.5. □

6.2.1 Proof of Lemma 6.5. Our proof is inspired by the hybrid argument from cryptography as in [19]; we flesh out the details of the proof in what follows.

For $0 \le \ell \le L$, define the *hybrid query history* $H^{(\ell)}$ with respect to Alg, Ora$[u]$, and Ora$[\pi]$ to be the random vector in $(\mathcal{T} \times \{0,1\}^n)^L$ generated as follows:

—For $i = 1, \ldots, \ell$:
  – Alg receives $((\tau_1, x_1), \ldots, (\tau_{i-1}, x_{i-1}))$ as input and generates $\tau_i \in \mathcal{T}$ (randomly) as output;
  – Ora$[u]$ receives τ_i as input and generates $x_i \in \{0,1\}^n$ as output.
—Pick $\pi \in \mathcal{B}$ uniformly at random.
—For $i = \ell + 1, \ldots, L$:
  – Alg receives $((\tau_1, x_1), \ldots, (\tau_{i-1}, x_{i-1}))$ as input and generates $\tau_i \in \mathcal{T}$ (randomly) as output;
  – Ora$[\pi]$ receives τ_i as input and generates $x_i \in \{0,1\}^n$ as output.
—The hybrid query history is $H^{(i)} = ((\tau_1, x_1), \ldots, (\tau_L, x_L))$.

Observe that $H^{(0)} = H'$ and $H^{(L)} = H$ in distribution. We will prove the following lemma regarding the distance between two adjacent hybrid query histories.

Lemma 6.6. *For every $1 \le \ell \le L$, we have*

$$d_{\mathrm{TV}}\left(H^{(\ell-1)}, H^{(\ell)}\right) \le \frac{16\varepsilon}{n}.$$

Note that Lemma 6.5 is an immediate consequence of Lemma 6.6.

Proof of Lemma 6.5. By the triangle inequality and Lemma 6.6, we have that

$$d_{\mathrm{TV}}\left(H, H'\right) \le \sum_{\ell=1}^{L} d_{\mathrm{TV}}\left(H^{(\ell-1)}, H^{(\ell)}\right) \le L \cdot \frac{16\varepsilon}{n} \le \frac{1}{4},$$

as claimed. □

It remains to prove Lemma 6.6. Inspecting the definitions of $H^{(\ell-1)}$ and $H^{(\ell)}$, we see that they only differ locally at one place, which we describe as follows. For $0 \le i \le L$ let $H_i = ((\tau_1, x_1), \ldots, (\tau_i, x_i))$ denote the first i entries of a random hybrid query history (notice that $H_0 = \emptyset$). We write $H^{(\ell-1)}$ and $H^{(\ell)}$ in the following form:

Generation of $H^{(\ell-1)}$:
(1) $H_0 \xrightarrow{\mathsf{Alg},\mathsf{Ora}[u]} H_{\ell-1}$;
(2) $\pi \sim \mathrm{unif}(\mathcal{B})$;
(3) $H_{\ell-1} \xrightarrow{\mathsf{Alg}} \tau_\ell \xrightarrow{\mathsf{Ora}[\pi]} x_\ell$;
(4) $H_\ell \leftarrow H_{\ell-1}$ append (τ_ℓ, x_ℓ);
(5) $H_\ell \xrightarrow{\mathsf{Alg},\mathsf{Ora}[\pi]} H_L = H^{(\ell-1)}$.

Generation of $H^{(\ell)}$:
(1) $H_0 \xrightarrow{\mathsf{Alg},\mathsf{Ora}[u]} H_{\ell-1}$;
(2) $H_{\ell-1} \xrightarrow{\mathsf{Alg}} \tau_\ell \xrightarrow{\mathsf{Ora}[u]} x_\ell$;
(3) $H_\ell \leftarrow H_{\ell-1}$ append (τ_ℓ, x_ℓ);
(4) $\pi \sim \mathrm{unif}(\mathcal{B})$;
(5) $H_\ell \xrightarrow{\mathsf{Alg},\mathsf{Ora}[\pi]} H_L = H^{(\ell)}$.

In fact the ordering of the steps (2)–(4) can be changed appropriately without having any influence on the final distribution of both $H^{(\ell-1)}$ and $H^{(\ell)}$, which will be helpful for a coupling argument. We rewrite the generating processes of $H^{(\ell-1)}$ and $H^{(\ell)}$ equivalently as follows:

Generation of $H^{(\ell-1)}$:
(1) $H_0 \xrightarrow{\mathsf{Alg},\mathsf{Ora}[u]} H_{\ell-1}$;
(2) $H_{\ell-1} \xrightarrow{\mathsf{Alg}} \tau_\ell$;
(3) $\pi \sim \mathrm{unif}(\mathcal{B})$, $\tau_\ell \xrightarrow{\mathsf{Ora}[\pi]} x_\ell$;
(4) $H_\ell \leftarrow H_{\ell-1}$ append (τ_ℓ, x_ℓ);
(5) $H_\ell \xrightarrow{\mathsf{Alg},\mathsf{Ora}[\pi]} H_L = H^{(\ell-1)}$.

Generation of $H^{(\ell)}$:
(1) $H_0 \xrightarrow{\mathsf{Alg},\mathsf{Ora}[u]} H_{\ell-1}$;
(2) $H_{\ell-1} \xrightarrow{\mathsf{Alg}} \tau_\ell$;
(3) $\pi \sim \mathrm{unif}(\mathcal{B})$, $\tau_\ell \xrightarrow{\mathsf{Ora}[u]} x_\ell$;
(4) $H_\ell \leftarrow H_{\ell-1}$ append (τ_ℓ, x_ℓ);
(5) $H_\ell \xrightarrow{\mathsf{Alg},\mathsf{Ora}[\pi]} H_L = H^{(\ell)}$.

Note that before and after the third step, the two processes have exactly the same steps. In the third step for $H^{(\ell-1)}$, we pick a bad distribution $\pi \in \mathcal{B}$ uniformly at random, and $\mathsf{Ora}[\pi]$ receives the pinning τ_ℓ as input and generates $x_\ell \in \{0,1\}^n$ according to π conditioned on τ_ℓ. Meanwhile, in the third step for $H^{(\ell)}$, we still pick a bad distribution $\pi \in \mathcal{B}$ but do not use it (for now), and $\mathsf{Ora}[u]$ receives τ_ℓ as input and generates $x_\ell \in \{0,1\}^n$ according to u instead of π. It is enough to show that, in this step, conditional on that $H_{\ell-1}$ and τ_ℓ are the same, the x_ℓ generated in the two processes are the same with high probability. Since before and after this step the two processes are doing the same thing, we can then couple these two processes to produce the same hybrid query history with high probability, i.e., $H^{(\ell-1)} = H^{(\ell)}$.

The following technical lemma bounds the probability that x_ℓ's are the same in both processes in the third step, which is crucial to us as explained earlier. The proof of it can be found in Section 6.2.2.

Lemma 6.7. *Let $\tau \in \mathcal{T}$ be an arbitrary pinning on some subset $\Lambda \subseteq V$ of size m. Then for a random distribution π chosen uniformly at random from $\mathcal{B}$, we have*

$$\mathbb{E}_{\pi \sim \mathrm{unif}(\mathcal{B})}\left[d_{\mathrm{TV}}\left(u\left(\cdot \mid \tau\right), \pi\left(\cdot \mid \tau\right)\right)\right] \le \frac{16\varepsilon}{n}.$$

We give below the proof of Lemma 6.6.

Proof of Lemma 6.6. We construct a coupling of $H^{(\ell-1)}$ and $H^{(\ell)}$ via coupling step-by-step the two processes generating $H^{(\ell-1)}$ and $H^{(\ell)}$. Initially $H_0 = \emptyset$ for both processes. Then we can couple $H_{\ell-1}$ and τ_ℓ since they are generated in the same way in both processes. For the third step, the bad distribution π can be chosen to be the same and we deduce from Lemma 6.7 that x_ℓ's can be

coupled with probability at least $1 - \varepsilon/n$. After that, suppose we couple x_ℓ, H_ℓ and then the final outputs are coupled. Hence, for this coupling $\mathbb{P}$ we have

$$d_{\mathrm{TV}}\left(H^{(\ell-1)}, H^{(\ell)}\right) \le \mathbb{P}\left(H^{(\ell-1)} \neq H^{(\ell)}\right) \le \frac{16\varepsilon}{n},$$

as wanted. □

6.2.2 Proof of Lemma 6.7. Here we give the proof of the technical lemma, Lemma 6.7.

Proof of Lemma 6.7. The distribution $\pi \in \mathcal{B}$ depends on A and σ. We will show that for any choice of $A \in \binom{[n]}{t}$ one has

$$\mathbb{E}_\sigma\left[d_{\mathrm{TV}}\left(u\left(\cdot \mid \tau\right), \pi_{A,\sigma}\left(\cdot \mid \tau\right)\right)\right] \le \frac{16\varepsilon}{n},$$

where σ is a uniformly random configuration in $\{0,1\}^n$.

Suppose $|\Lambda| = \ell$. Suppose $|A \cap \Lambda| = j$ and hence $|A \setminus \Lambda| = t - j$. Notice that $j \le \min\{t, \ell\}$. We partition $\mathcal{X} = \{0,1\}^n$ into three disjoint subsets.

Case 1. $\mathcal{X}_1 = \{\sigma \in \{0,1\}^n : \sigma_{A\cap\Lambda} \neq \tau_{A\cap\Lambda}\}$. We have

$$\Pr\left(\sigma \in \mathcal{X}_1\right) = \frac{|\mathcal{X}_1|}{2^n} = 1 - \frac{1}{2^j},$$

and also

$$d_{\mathrm{TV}}\left(u\left(\cdot \mid \tau\right), \pi_{A,\sigma}\left(\cdot \mid \tau\right)\right) = 0, \quad \forall \sigma \in \mathcal{X}_1.$$

Case 2. $\mathcal{X}_2 = \{\sigma \in \{0,1\}^n : \sigma_{A\cap\Lambda} = \tau_{A\cap\Lambda},\ \sigma_{\Lambda\setminus A} \neq \tau_{\Lambda\setminus A}\}$. We have

$$\Pr_\sigma\left(\sigma \in \mathcal{X}_2\right) = \frac{|\mathcal{X}_2|}{2^n} = \frac{1}{2^j} - \frac{1}{2^\ell}.$$

By definition we have

$$\pi_{A,\sigma}\left(x \mid \tau\right) = \begin{cases} 0, & \text{if} x_{A\setminus\Lambda} = \sigma_{A\setminus\Lambda}; \\ \dfrac{1}{2^{n-\ell} - 2^{n-\ell-t+j}}, & \text{if} x_{A\setminus\Lambda} \neq \sigma_{A\setminus\Lambda}. \end{cases}$$

It follows that

$$d_{\mathrm{TV}}\left(u\left(\cdot \mid \tau\right), \pi_{A,\sigma}\left(\cdot \mid \tau\right)\right) = \frac{1}{2^{t-j}}, \quad \forall \sigma \in \mathcal{X}_2.$$

Case 3. $\mathcal{X}_3 = \{\sigma \in \{0,1\}^n : \sigma_\Lambda = \tau_\Lambda\}$. We have

$$\Pr_\sigma\left(\sigma \in \mathcal{X}_3\right) = \frac{|\mathcal{X}_3|}{2^n} = \frac{1}{2^\ell}.$$

By definition we have

$$\pi_{A,\sigma}\left(x \mid \tau\right) = \begin{cases} \dfrac{\frac{1}{2^n}}{\frac{1}{2^t} + \frac{1}{2^\ell} - \frac{1}{2^{t+\ell-j}}}, & \text{if} x_{A\setminus\Lambda} \neq \sigma_{A\setminus\Lambda}; \\ 0, & \text{if} x_{A\setminus\Lambda} = \sigma_{A\setminus\Lambda} \text{and} x_{[n]\setminus\Lambda\setminus A} \neq \sigma_{[n]\setminus\Lambda\setminus A}; \\ \dfrac{\frac{1}{2^t}}{\frac{1}{2^t} + \frac{1}{2^\ell} - \frac{1}{2^{t+\ell-j}}}, & \text{if} x_{[n]\setminus\Lambda} = \sigma_{[n]\setminus\Lambda}. \end{cases}$$

It follows that

$$d_{\mathrm{TV}}\left(u\left(\cdot \mid \tau\right), \pi_{A,\sigma}\left(\cdot \mid \tau\right)\right) = \frac{\frac{1}{2^t}}{\frac{1}{2^t} + \frac{1}{2^\ell} - \frac{1}{2^{t+\ell-j}}} - \frac{1}{2^{n-\ell}} = \frac{2^\ell}{2^t + 2^\ell - 2^j} - \frac{1}{2^{n-\ell}}, \quad \forall \sigma \in \mathcal{X}_3.$$

Therefore, combining all three cases we get from the law of total expectation that

$$\mathbb{E}_\sigma\left[d_{\mathrm{TV}}\left(u\left(\cdot \mid \tau\right), \pi_{A,\sigma}\left(\cdot \mid \tau\right)\right)\right] = \left(1 - \frac{1}{2^j}\right)0 + \left(\frac{1}{2^j} - \frac{1}{2^\ell}\right)\frac{1}{2^{t-j}} + \frac{1}{2^\ell}\left(\frac{2^\ell}{2^t + 2^\ell - 2^j} - \frac{1}{2^{n-\ell}}\right)$$
$$\leq \frac{1}{2^t} + \frac{1}{2^t + 2^\ell - 2^j}.$$

Note that the second term is monotone increasing in j and by definition $j \leq \min\{t, \ell\}$. Hence, we deduce that

$$\frac{1}{2^t + 2^\ell - 2^j} \leq \frac{1}{2^t + 2^\ell - 2^{\min\{t,\ell\}}} = \frac{1}{2^{\max\{t,\ell\}}} \leq \frac{1}{2^t}.$$

We conclude that for any A,

$$\mathbb{E}_\sigma\left[d_{\mathrm{TV}}\left(u\left(\cdot \mid \tau\right), \pi_{A,\sigma}\left(\cdot \mid \tau\right)\right)\right] \leq \frac{1}{2^{t-1}} \leq \frac{16\varepsilon}{n},$$

where in the last inequality we recall that $t = \left\lceil \log_2(n/\varepsilon)\right\rceil - 3 \geq \log_2(n/\varepsilon) - 3$. □

6.3 Uniformity Testing with Coordinate Oracle and General Oracle for TV Distance

In this section we consider uniformity testing over the binary hypercube for TV distance when we have access to Coordinate Oracle and General Oracle. We assume the binary hypercube is denoted by $\mathcal{X}_n = \{+1, -1\}^n$ instead of $\{0, 1\}^n$, since our bad distributions will be Ising models where $+1, -1$ are more often used.

Let Alg denote an arbitrary uniformity testing algorithm (possibly randomized and adaptive) with Coordinate Oracle and General Oracle access. We assume that Alg receives L independent full samples from the General Oracle and is allowed to make L queries to the Coordinate Oracle. For ease of notation we denote by $\mathsf{Ora}[\pi]$ the Subcube Oracle with respect to a distribution π over $\{+1, -1\}^n$.

Definition 6.8 (Query History for Coordinate Oracle and General Oracle). Let $\mathcal{T}$ denote the set of all pinnings on $n - 1$ coordinates (which is exactly all possible inputs to the Coordinate Oracle). For integer $L \in \mathbb{N}^+$, we define the *query history* with respect to Alg and $\mathsf{Ora}[\pi]$ of length $2L$ to be the random vector in $\mathcal{X}_n^L \times (\mathcal{T} \times \{+1, -1\})^L$ generated as follows:

- Let $x_1, \ldots, x_L$ be L independent samples from π;
- For $i = 1, \ldots, L$:
  - Alg receives $(x_1, \ldots, x_L)$ and $((\tau_1, a_1), \ldots, (\tau_{i-1}, a_{i-1}))$ as input and generates $\tau_i \in \mathcal{T}$ (randomly) as output;
  - $\mathsf{Ora}[\pi]$ receives τ_i as input and generates $a_i \in \{+1, -1\}$ as output.
- The (coordinate and general) query history is $H = (x_1, \ldots, x_L; (\tau_1, a_1), \ldots, (\tau_L, a_L))$.

Definition 6.8 is analogous to (in fact, a special case of) Definition 6.2; throughout this subsection, we consider query history only with respect to Coordinate Oracle and General Oracle.

The output of Alg with sample complexity $2L$ is a (randomized) function of the query history H of length $2L$. Our main theorem is then stated as follows.

Theorem 6.9. *There exists a universal constant $c > 0$ such that the following holds. Let $n \in \mathbb{N}^+$ be a sufficiently large integer and $\varepsilon > 0$ be a real. Let $u = u_n$ denote the uniform distribution over $\{+1, -1\}^n$. Then there is no algorithm which can achieve the following properties using L samples from* General Oracle *and L queries from* Coordinate Oracle *where $L \leq cn/\varepsilon^2$:*

- $\Pr_H(\mathsf{output} = \mathsf{Yes}) \geq 2/3$ *for a random query history H of length $2L$ with respect to* Alg *and* $\mathsf{Ora}[u]$;

—$\Pr_{H'}(\mathsf{output} = \mathsf{No}) \geq 2/3$ *for a random query history* H' *of length* $2L$ *with respect to* Alg *and* $\mathsf{Ora}[\pi]$ *where* π *is any distribution such that* $d_{\mathrm{TV}}(\pi, u) \geq \varepsilon$.

We observe that Theorem 1.5 follows immediately from Theorem 6.9.

In [31, Theorem 14] it was shown that $\Omega(n/\varepsilon^2)$ samples are necessary for uniformity testing with only General Oracle access but assuming the hidden distribution π is an Ising model. See also [18, Theorem 14] for very similar lower bounds in the setting of Bayesian networks. Note that though Theorem 14 from [31] is stated for *symmetric* KL divergence, it actually works for TV distance as well; see [31, Remark 4]. We use the same constructions from [18, 31] for the family of bad distributions for our purpose. Assume that n is even; the case of odd n can be easily reduced to even n by adding an extra uniform, independent coordinate. Suppose M is a perfect matching of n coordinates, i.e., M is a collection of $n/2$ pairs of coordinates such that each coordinate appears in exactly one pair. Let $\mathcal{M}$ be the set of all perfect matchings on $[n]$. Each bad distribution π_M where $M \in \mathcal{M}$ corresponds to an Ising model on the graph $G = ([n], M)$ of $n/2$ edges, with the edge coupling set to be $\beta = \rho\varepsilon/\sqrt{n}$ where ρ is a universal constant sufficiently large. The following are established in [18, 31].

Claim 6.10. ([18, 31]).

(1) For $\rho > 0$ *sufficiently large, for all* $M \in \mathcal{M}$*, it holds*

$$d_{\mathrm{TV}}(\pi_M, u) \geq \varepsilon.$$

(2) For any $\rho > 0$ *there exists* $c_1 = c_1(\rho) > 0$ *such that the following holds. Suppose* $L \leq c_1 n/\varepsilon$*. Let* $X = (x_1, \ldots, x_L)$ *be* L *independent samples from* u*. Independently, let* $M \in \mathcal{M}$ *be chosen uniformly at random, and let* $X' = (x'_1, \ldots, x'_L)$ *be* L *independent samples from* π_M*. Then* $d_{\mathrm{TV}}(X, X') \leq 0.98$.

Proof. (1) follows from Lemma 8 in [18]. (2) is proved in Section 8.3.2 in [31]. See also in Section 8.1 from [18] the same result for a slightly different construction of π_M, where every edge is set to be ferromagnetic with probability $1/2$ and anti-ferromagnetic otherwise. □

Define H to be the random query history of length $2L$ with respect to Alg and $\mathsf{Ora}[u]$, and let output denote the random output with respect to H and Alg. Define H' to be the random query history of length $2L$ generated by

—Pick $M \in \mathcal{M}$ uniformly at random and let $\pi = \pi_M$;
—Let H' be the random query history of length $2L$ with respect to Alg and $\mathsf{Ora}[\pi]$.

Further, let output′ denote the random output with respect to H' and Alg. Then we can show the following key lemma.

Lemma 6.11. *For query histories* H, H' *of length* $2L$ *where* $L \leq cn/\varepsilon$ *and* output, output′ *defined as above, we have*

$$d_{\mathrm{TV}}(\mathsf{output}, \mathsf{output}') \leq d_{\mathrm{TV}}(H, H') \leq 0.99.$$

Proof. The first inequality follows from the data processing inequality. We focus on the second one. For $M \in \mathcal{M}$ and $t \in \{0, 1\}$, let $\pi_{M,t}$ denote the Ising model on $G = ([n], M)$ with edge coupling $t\beta = t\rho\varepsilon/\sqrt{n}$. Observe that $\pi_{M,0} = u$ and $\pi_{M,1} = \pi_M$. We rewrite the process for generating the query histories H and H' of length $2L$ in the following equivalent form:

—Let $M \in \mathcal{M}$ be chosen uniformly at random from $\mathcal{M}$;
—Let $X_t = (x_1, \ldots, x_L) \in \mathcal{X}_n^L$ be L independent samples from $\pi_{M,t}$;

—Let $R_t = (r_1, \ldots, r_L) \in \{0,1\}^L$ be L independent Bernoulli random variables with mean $(1 + \tanh(t\rho\varepsilon/\sqrt{n}))/2$;
—For $i = 1, \ldots, L$:
– Alg receives $(x_1, \ldots, x_L)$ and $((\tau_1, a_1), \ldots, (\tau_{i-1}, a_{i-1}))$ as input and generates $\tau_i \in \mathcal{T}$ (randomly) as output;
– $\mathsf{Ora}[\pi]$ receives τ_i as input, which fixes all coordinates but one say j, and suppose j' is matched to j in M, then $\mathsf{Ora}[\pi]$ outputs $a_i = (\tau_i)_{j'}$ (the j'th coordinate of τ_i) as the sampled value at the jth coordinate if $r_i = 1$, and outputs $a_i = -(\tau_i)_{j'}$ otherwise;
—The query history is $H_t = (x_1, \ldots, x_L; (\tau_1, a_1), \ldots, (\tau_L, a_L))$.

Observe that if $t = 0$, then the final query history H_0 is distributed as H; meanwhile, if $t = 1$, then it is distributed as H'. Moreover, the process mentioned earlier can be viewed as a random mapping from the vector (M, X_t, R_t) to the query history H_t where, for fixed (M, X_t, R_t), the randomness purely comes from the decision-making of Alg. Therefore, we can apply the data processing inequality and obtain

$$d_{\mathrm{TV}}(H, H') \le d_{\mathrm{TV}}((M, X_0, R_0), (M, X_1, R_1)) \le d_{\mathrm{TV}}(X_0, X_1) + d_{\mathrm{TV}}(R_0, R_1).$$

Note that $d_{\mathrm{TV}}(X_0, X_1) \le 0.98$ by Claim 6.10. For the second term, we have

$$d_{\mathrm{TV}}(R_0, R_1) = d_{\mathrm{TV}}\left(\mathrm{Bin}\left(L, \frac{1}{2}\right), \mathrm{Bin}\left(L, \frac{1}{2}\left(1 + \tanh\frac{\rho\varepsilon}{\sqrt{n}}\right)\right)\right) \le c' \cdot \sqrt{L} \cdot \frac{\rho\varepsilon}{\sqrt{n}} \le 0.01,$$

where $c' > 0$ is a universal large constant, and $L \le cn/\varepsilon^2$ for c sufficiently small. Therefore, we deduce that $d_{\mathrm{TV}}(H, H') \le 0.98 + 0.01 = 0.99$ as claimed. □

We end this section with the proof of Theorem 6.9.

Proof of Theorem 6.9. Suppose for sake of contradiction that Alg satisfies both properties as in Theorem 6.9. Then by a standard amplification technique for failure probability, one can decrease the failure probability from 1/3 to 0.001 with the number of samples needed increases only by a constant factor; see [16, Lemma 1.1.1]. In particular, for query histories H, H' of length $2L$ where $L \le cn/\varepsilon$ and $\mathsf{output}, \mathsf{output}'$ defined as earlier, we have

$$\Pr(\mathsf{output} = \mathsf{Yes}) \ge 0.999 \quad \text{and} \quad \Pr(\mathsf{output}' = \mathsf{Yes}) \le 0.001.$$

This implies $d_{\mathrm{TV}}(\mathsf{output}, \mathsf{output}') \ge 0.998$ which contradicts Lemma 6.11. □

7 Identity Testing with Subcube Oracle

In this section we give our algorithmic results for identity testing with access to the Subcube Oracle. In particular, we establish slightly more general versions of Theorems 1.6 and 1.8 which relax the assumption that μ is fully supported and only require that the support of π is a subset of the support of μ. For the case when μ is not fully supported we instead require the slightly stronger assumption that μ is b-marginally bounded (this notion is equivalent to balancedness when μ is fully supported).

7.1 Identity Testing with Exact Conditional Marginal Distributions

Recall that $[i] = \{1, \ldots, i\}$ for an integer $i \in \mathbb{N}^+$. The following factorization of (relative) entropy is well-known; see, e.g., [21, 22, 56].

Lemma 7.1. *For any distribution π over $\mathcal{K}^n$ such that $\pi \ll \mu$ we have*

$$D_{\mathrm{KL}}(\pi \,\|\, \mu) = \sum_{i=1}^{n} \mathbb{E}_{x \sim \pi_{[i-1]}}\left[D_{\mathrm{KL}}(\pi_i(\cdot \mid x) \,\|\, \mu_i(\cdot \mid x))\right]. \tag{20}$$

We now give our testing algorithm with Subcube Oracle.

Theorem 7.2. *Let $k = k(n)$ be an integer and let $b = b(n) \in (0, 1/2]$ be a real. Suppose that $\log\log(1/b) = O(\log n)$. There is an identity testing algorithm for all b-marginally bounded distributions with query access to* Subcube Oracle *and for KL divergence with distance parameter $\varepsilon > 0$. The query complexity of the identity testing algorithm is*

$$O\left(\min\left\{\frac{1}{\sqrt{b}} \cdot \frac{n}{\varepsilon}\log^3\left(\frac{n}{\varepsilon}\right),\ \sqrt{k}\log\left(\frac{1}{b}\right) \cdot \frac{n^2}{\varepsilon^2}\log^2\left(\frac{n}{\varepsilon}\right)\right\}\right).$$

The running time of the algorithm is polynomial in all parameters and also proportional to the time of computing the conditional marginal distributions $\mu_i(\cdot \mid x)$ for any $i \in [n]$ and any feasible $x \in \mathcal{K}^{[i-1]}$. Furthermore, if $k = 2$, i.e., we have a binary domain $\mathcal{K} = \{0, 1\}$, the query complexity can be improved to

$$O\left(\log\left(\frac{1}{b}\right) \cdot \frac{n}{\varepsilon}\log^3\left(\frac{n}{\varepsilon}\right)\right).$$

Proof. We observe that (20) can be equivalently written as

$$D_{\mathrm{KL}}\left(\pi \,\|\, \mu\right) = n\,\mathbb{E}_{(i,x)}\left[D_{\mathrm{KL}}\left(\pi_i(\cdot \mid x) \,\|\, \mu_i(\cdot \mid x)\right)\right],$$

where $i \in [n]$ is a uniformly random coordinate and x is generated from $\pi_{[i-1]}$. Therefore, Algorithm 1 still works once we generate the pair (i, x) in Line 1 from the correct distribution as just described, and define $p_i^x = \pi_i(\cdot \mid x)$, $q_i^x = \mu_i(\cdot \mid x)$ correspondingly. The analysis is exactly the same with the constant C for approximate tensorization replaced by 1. We omit the proofs here and only highlight the differences: The coordinate balancedness η is now replaced by the marginal boundedness b, and the running time depends on the time to compute the conditional marginal distributions $\mu_i(\cdot \mid x)$ for any $i \in [n]$ and any $x \in \mathcal{K}^{[i-1]}$ such that $\mu_{[i-1]}(x) > 0$. □

Remark 7.3. We remark that the assumption of marginal boundedness can be relaxed to the following slightly weaker version: for a fixed ordering of the coordinates, for every $i \in [n]$, every $x \in \mathcal{K}^{[i-1]}$ with $\mu_{[i-1]}(x) > 0$, and every $a \in \mathcal{K}$, one has

$$\text{either } \mu_i(a \mid x) = 0, \quad \text{or } \mu_i(a \mid x) \geq b.$$

In some circumstances, this weaker notion of marginal boundedness can give a better bound on the sample complexity.

Theorem 7.2 that identity testing can be done efficiently for a wide variety of families of distributions with the power of Subcube Oracle, assuming that one can efficiently compute the exact marginal probabilities under any conditioning. Below we give a few examples where Theorem 7.2 applies:

- Consider any undirected graphical model (e.g., Ising model, Potts model) defined on trees of constant degrees. Then the distributions are $\Omega(1)$-marginally bounded, and one can efficiently compute the marginal probabilities under any pinning via, e.g., Belief Propagation. Hence, there is a polynomial-time identity testing algorithm for undirected graphical models on bounded-degree trees with Subcube Oracle access. The sample complexity is $O((n/\varepsilon)\log^3(n/\varepsilon))$ where n is the number of vertices. If the degree is unbounded, then the marginal bound b can be as small as $e^{-\Theta(n)}$. Still, by the second bound in Theorem 7.2 the number of samples needed is at most $O((n^3/\varepsilon^2)\log^2(n/\varepsilon))$.
- Consider the Bayesian network on a **Directed Acyclic Graph (DAG)**, and assume without loss of generality that $[n] = \{1, \ldots, n\}$ is the topological ordering of the DAG. In particular, all conditional marginal probabilities at any coordinate $i \in [n]$ and conditioned on any feasible pinning $x \in \mathcal{K}^{[i-1]}$ are given by the Bayesian network. If these conditional marginal

probabilities are lower bounded by $b = \Omega(1)$, then there is a polynomial-time identity testing algorithm for such Bayesian networks with Subcube Oracle access, and the sample complexity is $O((n/\varepsilon)\log^3(n/\varepsilon))$. If b is exponentially small, then similarly as before the sample complexity is $O((n^3/\varepsilon^2)\log^2(n/\varepsilon))$. See also Remark 7.3 described earlier on relaxing the marginal boundedness condition to specifically the topological ordering.

— Consider mixtures of polynomially many product distributions, each of which has $\eta(\mu) = \Omega(1)$ as defined in Section 4.4.1. One can efficiently compute the conditional marginal probabilities by the simple nature of mixtures of product distributions. Then by Theorem 7.2, we have an efficient identity testing algorithm with Subcube Oracle access and the sample complexity is $O((n/\varepsilon)\log^3(n/\varepsilon))$. Similarly as before, the sample complexity becomes $O((n^3/\varepsilon^2)\log^2(n/\varepsilon))$ when the minimum $\eta(\mu)$ is exponentially small.

7.2 Identity Testing with Approximate Conditional Marginal Distributions

In Theorem 7.2 we assume that one can compute exactly any conditional marginal distribution in polynomial time. In some applications the exact computation is not possible and one can get, at the best, an estimator of the conditional marginal probabilities. As we will show in this subsection, identity testing can still be done efficiently in this setting.

We first need more robust versions of Lemmas 4.3 and 4.10. We say there is an FPRAS for a distribution q over $\mathcal{K}$ if for any $\varepsilon > 0$ and $\delta \in (0,1)$, one can compute a distribution $\hat{q}$ over $\mathcal{K}$ as an approximation of q such that, with probability $1-\delta$, we have that for every $a \in \mathcal{K}$,

$$e^{-\varepsilon} \le \frac{\hat{q}(a)}{q(a)} \le e^{\varepsilon},$$

and $\hat{q}$ can be computed with running time polynomial in k, $1/\varepsilon$, $\log(1/\delta)$, and the input size of q (e.g., the number of parameters representing q). We remark that if $q(a) = 0$ then $\hat{q}(a) = 0$.

Lemma 7.4. *Let $k \in \mathbb{N}^+$ be an integer, and let $\varepsilon > 0$, $b \in (0, 1/2]$ be reals. Given an FPRAS for a target distribution q over domain $\mathcal{K}$ of size k such that either $q(a) = 0$ or $q(a) \ge b$ for any $a \in \mathcal{K}$, and given sample access to an unknown distribution $p \ll q$ over $\mathcal{K}$, there exists a polynomial-time identity testing algorithm that distinguishes with probability at least $2/3$ between the two cases*

$$p = q \quad \text{and} \quad D_{\mathrm{KL}}(p \,\|\, q) \ge \varepsilon. \tag{21}$$

For $k \ge 3$, the sample complexity of the identity testing algorithm is

$$O\left(\min\left\{\frac{1}{\varepsilon\sqrt{b}}, \frac{\sqrt{k}\ln(1/b)}{\varepsilon^2}\right\}\right).$$

For $k = 2$, the sample complexity of the identity testing algorithm is

$$O\left(\frac{\ln(1/b)}{\varepsilon}\right).$$

Proof. Let m be an upper bound for the number of samples required in Lemmas 4.3 and 4.10, with the assumption being either $q(a) = 0$ or $q(a) \ge b/2$ for any $a \in \mathcal{K}$, distance parameter $\varepsilon/2$, and failure probability $1/10$. Let $\xi = O(\min\{\varepsilon, 1/m\})$ be a small constant, and let $\hat{q}$ be an approximation of q such that with probability $9/10$ we have $e^{-\xi} \le \hat{q}(a)/q(a) \le e^{\xi}$ for every $a \in \mathcal{K}$. Notice that if this holds then

$$|D_{\mathrm{KL}}(p \,\|\, \hat{q}) - D_{\mathrm{KL}}(p \,\|\, q)| \le \sum_{a\in\mathcal{K}} p(a)\left|\ln\left(\frac{q}{\hat{q}}\right)\right| \le \xi.$$

We then apply the identity testing algorithm $\mathcal{A}_{\text{KL-ID}}$ from Lemmas 4.3 and 4.10 to the distributions $p, \hat{q}$ with distance parameter $\varepsilon/2$ and failure probability $1/10$, and returns the output of $\mathcal{A}_{\text{KL-ID}}$ as our output. Note that $\hat{q}(a) = 0$ if $q(a) = 0$ and $\hat{q}(a) \geq e^{-\xi} q(a) \geq b/2$ if $q(a) \geq b$, assuming $\hat{q}$ is a ξ-approximation of q. Thus, the number of samples required by $\mathcal{A}_{\text{KL-ID}}$ is at most m. If $D_{\text{KL}}(p \,\|\, q) \geq \varepsilon$, then

$$D_{\text{KL}}(p \,\|\, \hat{q}) \geq D_{\text{KL}}(p \,\|\, q) - |D_{\text{KL}}(p \,\|\, \hat{q}) - D_{\text{KL}}(p \,\|\, q)| \geq \varepsilon - \xi \geq \frac{\varepsilon}{2}.$$

Hence, the testing algorithm wrongly outputs Yes only if at least one of the following happens:

(1) $\hat{q}$ is not a ξ-approximation of q, which happens with probability at most $1/10$;
(2) $\mathcal{A}_{\text{KL-ID}}$ makes a mistake, which happens with probability at most $1/10$.

This shows that the failure probability is at most $1/5$. If $p = q$, then notice that

$$d_{\text{TV}}(p, \hat{q}) = d_{\text{TV}}(q, \hat{q}) = O(\xi) \leq \frac{1}{10m}.$$

We consider an optimal coupling between m independent samples from p and m independent samples from $\hat{q}$, so the probability that these two sets of m samples are not exactly the same is at most $1/10$. One can think of the testing process as follows: We try to send m samples from $\hat{q}$ to $\mathcal{A}_{\text{KL-ID}}$, and it succeeds only when the samples are coupled with those from p. Therefore, the failure probability, in addition to (1) and (2) above, also includes this uncoupled probability, and hence is at most $3/10$. Finally, the number of samples needed, m, is bounded in Lemmas 4.3 and 4.10. □

Lemma 7.4, combined with the proof of Theorem 7.2, immediately implies the following theorem. See also Remark 7.3 for the discussion on relaxing marginal boundedness.

Theorem 7.5. *Let $k = k(n)$ be an integer and let $b = b(n) \in (0, 1/2]$ be a real. Suppose that $\log\log(1/b) = O(\log n)$. There is an identity testing algorithm for all b-marginally bounded distributions with query access to* Subcube Oracle *and for KL divergence with distance parameter $\varepsilon > 0$. The query complexity of the identity testing algorithm is*

$$O\left(\min\left\{\frac{1}{\sqrt{b}} \cdot \frac{n}{\varepsilon}\log^3\left(\frac{n}{\varepsilon}\right),\ \sqrt{k}\log\left(\frac{1}{b}\right) \cdot \frac{n^2}{\varepsilon^2}\log^2\left(\frac{n}{\varepsilon}\right)\right\}\right).$$

The running time of the algorithm is polynomial in all parameters assuming that there is an FPRAS for the conditional marginal distributions $\mu_i(\cdot \mid x)$ for any $i \in [n]$ and any feasible $x \in \mathcal{K}^{[i-1]}$. Furthermore, if $k = 2$, i.e., we have a binary domain $\mathcal{K} = \{0, 1\}$, the query complexity can be improved to

$$O\left(\log\left(\frac{1}{b}\right) \cdot \frac{n}{\varepsilon}\log^3\left(\frac{n}{\varepsilon}\right)\right).$$

Again, we give a few examples as applications of Theorem 7.5, omitting all the technical details:

—Consider the Ising model with the interaction matrix J (with entries being β_{uv}'s and assumed to be positive semi-definite). We know from recent works [2, 36, 54] that one can efficiently estimate all conditional marginal probabilities when $\|J\|_2 < 1$ under any external fields. There are two special features for this application. The first is that the marginal bounds could potentially be as small as $e^{-\Theta(\sqrt{n})}$. The second is that we can only approximate the conditional marginal probabilities rather than get the exact values, and hence we should apply Theorem 7.5 instead of Theorem 7.2. With access to the Subcube Oracle, one can obtain a polynomial-time identity testing algorithm for this family of Ising models with sample complexity $O((n^{3/2}/\varepsilon)\log^3(n/\varepsilon))$ (note that $k = 2$).

—Consider the monomer–dimer model (weighted matchings) on arbitrary (unbounded-degree) graphs. We know from the classical work [51] that one can approximate the conditional marginal distributions for all pinnings. Similar to the previous example, the marginal probabilities can be exponentially small (in the number of vertices) and one can at the best approximate them efficiently rather than computing them exactly. Still, we can apply Theorem 7.5 to obtain an efficient identity testing algorithm with access to the Subcube Oracle with sample complexity $O((mn/\varepsilon)\log^3(n/\varepsilon))$ where m is the number of edges of the graph (which is the dimension) and n is the number of vertices (note that $k = 2$).

7.3 Estimating KL Divergence with Additive Error

With access to the Subcube Oracle, we can also estimate the KL divergence from an unknown distribution π to a given distribution μ within an arbitrary additive error in polynomial time. This corresponds to the *tolerant identity testing* problem for KL divergence, that is, given $s, \varepsilon > 0$, we want to distinguish between $D_{\mathrm{KL}}(\pi \,\|\, \mu) \le s$ and $D_{\mathrm{KL}}(\pi \,\|\, \mu) \ge s + \varepsilon$.

We first consider estimating KL divergence for distributions on a finite domain of size k.

Lemma 7.6. *Let $k \in \mathbb{N}^+$ be an integer, and let $\varepsilon > 0$, $b \in (0, 1/2]$ be reals. Given an FPRAS for a target distribution q over domain $\mathcal{K}$ of size k such that either $q(a) = 0$ or $q(a) \ge b$ for any $a \in \mathcal{K}$, and given sample access to an unknown distribution $p \ll q$ over $\mathcal{K}$, there exists a polynomial-time algorithm that computes $\widehat{R}$ such that with probability at least $2/3$ it holds*

$$\left|\widehat{R} - D_{\mathrm{KL}}(p \,\|\, q)\right| \le \varepsilon, \tag{22}$$

with sample complexity

$$O\left(\frac{k}{\varepsilon \log(k/\varepsilon)} + \frac{\log^2(1/b)}{\varepsilon^2}\right).$$

For a distribution p over a finite domain $\mathcal{K}$, the (Shannon) entropy of p is defined as

$$H(p) = \sum_{a \in \mathcal{K}} p(a) \ln\left(\frac{1}{p(a)}\right).$$

Observe that if $p \ll q$ are two distributions over $\mathcal{K}$, then

$$D_{\mathrm{KL}}(p \,\|\, q) = \sum_{a \in \mathcal{K}} p(a) \ln\left(\frac{1}{q(a)}\right) - \sum_{a \in \mathcal{K}} p(a) \ln\left(\frac{1}{p(a)}\right) = \mathbb{E}_{a \sim p}\left[\ln\left(\frac{1}{q(a)}\right)\right] - H(p). \tag{23}$$

It suffices to estimate the two terms on the right-hand side of (23), respectively, with good enough accuracy.

We need the following well-known result from [68] for estimating the entropy of an unknown distribution from samples; see also [52, 66, 69].

Lemma 7.7 ([68]). *Let $k \in \mathbb{N}^+$ be an integer, and let $\varepsilon > 0$ be a real. Given sample access to an unknown distribution p over domain $\mathcal{K}$ of size k, there exists a polynomial-time algorithm that computes $\widehat{H}$ such that with probability at least $9/10$ it holds*

$$\left|\widehat{H} - H(p)\right| \le \varepsilon, \tag{24}$$

with sample complexity

$$O\left(\frac{k}{\varepsilon \log(k/\varepsilon)} + \frac{\log^2 k}{\varepsilon^2}\right).$$

For the first term in (23), we show the following estimator.

Lemma 7.8. *Let $k \in \mathbb{N}^+$ be an integer, and let $\varepsilon > 0$, $b \in (0, 1/2]$ be reals. Given an FPRAS for a target distribution q over domain $\mathcal{K}$ of size k such that either $q(a) = 0$ or $q(a) \geq b$ for any $a \in \mathcal{K}$, and given sample access to an unknown distribution $p \ll q$ over $\mathcal{K}$, there exists a polynomial-time algorithm that computes $\widehat{G}$ such that with probability at least $4/5$ it holds*

$$\left|\widehat{G} - \mathbb{E}_{a\sim p}\left[\ln\left(\frac{1}{q(a)}\right)\right]\right| \leq \varepsilon, \tag{25}$$

with sample complexity

$$O\left(\frac{\ln^2(1/b)}{\varepsilon^2}\right).$$

Proof. Compute an approximation $\hat{q}$ of q such that, with probability $9/10$, we have that $e^{-\varepsilon/2} \leq \hat{q}(a)/q(a) \leq e^{\varepsilon/2}$ for every $a \in \mathcal{K}$ with $q(a) > 0$, and $\hat{q}(a) = 0$ for $q(a) = 0$. Generate m independent samples from p, denoted by $a_1, \ldots, a_m$. Then our estimator is defined as

$$\widehat{G} = \frac{1}{m}\sum_{j=1}^{m} \ln\left(\frac{1}{\hat{q}(a_j)}\right).$$

We will show that $\widehat{G}$ satisfies (25) with probability at least $4/5$ for

$$m = \left\lceil \frac{8\ln^2(1/b)}{\varepsilon^2} \right\rceil.$$

Observe that

$$\begin{aligned}
&\left|\widehat{G} - \mathbb{E}_{a\sim p}\left[\ln\left(\frac{1}{q(a)}\right)\right]\right| \\
&\leq \left|\frac{1}{m}\sum_{j=1}^{m} \ln\left(\frac{1}{\hat{q}(a_j)}\right) - \frac{1}{m}\sum_{j=1}^{m} \ln\left(\frac{1}{q(a_j)}\right)\right| + \left|\frac{1}{m}\sum_{j=1}^{m} \ln\left(\frac{1}{q(a_j)}\right) - \mathbb{E}_{a\sim p}\left[\ln\left(\frac{1}{q(a)}\right)\right]\right|.
\end{aligned} \tag{26}$$

Assuming $\hat{q}$ is an $(\varepsilon/2)$-approximation of q, we can upper bound the first term in (26) by

$$\frac{1}{m}\sum_{j=1}^{m}\left|\ln\left(\frac{\hat{q}(a_j)}{q(a_j)}\right)\right| \leq \frac{1}{m}\sum_{j=1}^{m}\frac{\varepsilon}{2} = \frac{\varepsilon}{2}.$$

Meanwhile, for the second term in (26), since $0 \leq \ln(1/q(a)) \leq \ln(1/b)$ for all $a \in \mathcal{K}$ with $q(a) > 0$ and since $p \ll q$, we deduce from Hoeffding's inequality that

$$\Pr\left(\left|\frac{1}{m}\sum_{j=1}^{m} \ln\left(\frac{1}{q(a_j)}\right) - \mathbb{E}_{a\sim p}\left[\ln\left(\frac{1}{q(a)}\right)\right]\right| \geq \frac{\varepsilon}{2}\right) \leq 2\exp\left(-\frac{\varepsilon^2 m}{2\ln^2(1/b)}\right) \leq \frac{1}{10},$$

provided $m \geq (8/\varepsilon^2)\ln^2(1/b)$. Therefore, we deduce from (26) that

$$\left|\widehat{G} - \mathbb{E}_{a\sim p}\left[\ln\left(\frac{1}{q(a)}\right)\right]\right| \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon$$

with failure probability at most $1/10 + 1/10 = 1/5$ by the union bound, as wanted. □

Lemma 7.6 then follows easily from Lemmas 7.7 and 7.8.

Proof Lemma 7.6. Since the sample complexity upper bound we want to show is monotone increasing in k, we can safely assume without loss of generality that q is fully supported on $\mathcal{K}$, i.e., $q(a) \geq b$ for each $a \in \mathcal{K}$. In particular, this implies that $b \leq 1/k$. Take $\widehat{G}$ from Lemma 7.8 and $\widehat{H}$ from Lemma 7.7, and let $\widehat{R} = \widehat{G} - \widehat{H}$. We then deduce from (23) that

$$\left|\widehat{R} - D_{\mathrm{KL}}\left(p \,\|\, q\right)\right| \leq \left|\widehat{G} - \mathbb{E}_{a\sim p}\left[\ln\left(\frac{1}{q(a)}\right)\right]\right| + \left|\widehat{H} - H(p)\right| \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon,$$

which fails with probability at most $1/5 + 1/10 = 3/10$ by the union bound. The running time is polynomial in all parameters and depends on the given FPRAS for q. The sample complexity is given by

$$O\left(\frac{k}{\varepsilon \log(k/\varepsilon)} + \frac{\log^2 k}{\varepsilon^2}\right) + O\left(\frac{\log^2(1/b)}{\varepsilon^2}\right) = O\left(\frac{k}{\varepsilon \log(k/\varepsilon)} + \frac{\log^2(1/b)}{\varepsilon^2}\right),$$

since we have $k \leq 1/b$. □

We now give our main theorem for estimating KL divergence with Subcube Oracle access.

Theorem 7.9. *Let $k = k(n)$ be an integer and let $b = b(n) \in (0, 1/2]$ be a real. Suppose that $k = O(n)$ and $\log\log(1/b) = O(\log n)$. Given a visible distribution μ over $\mathcal{K}^n$ that is b-marginally bounded, and given access to* Subcube Oracle *for a hidden distribution $\pi \ll \mu$ over $\mathcal{K}^n$, there is an algorithm that for any $\varepsilon > 0$ computes $\widehat{S}$ such that with probability at least $2/3$ it holds*

$$\left|\widehat{S} - D_{\mathrm{KL}}\left(\pi \,\|\, \mu\right)\right| \leq \varepsilon. \tag{27}$$

The query complexity of the algorithm is

$$O\left(\log^4\left(\frac{1}{b}\right) \cdot \frac{n^4}{\varepsilon^4} \log\left(\frac{n}{\varepsilon}\right)\right).$$

The running time of the algorithm is polynomial in all parameters assuming that there is an FPRAS for the conditional marginal distributions $\mu_i(\cdot \mid x)$ for any $i \in [n]$ and any feasible $x \in \mathcal{K}^{[i-1]}$.

Proof. From (20) we observe that

$$D_{\mathrm{KL}}\left(\pi \,\|\, \mu\right) = n\, \mathbb{E}_{(i,x)}\left[D_{\mathrm{KL}}\left(\pi_i(\cdot \mid x) \,\|\, \mu_i(\cdot \mid x)\right)\right],$$

where (i, x) is a random pair generated by taking a uniformly random coordinate $i \in [n]$ and sampling $x \in \mathcal{K}^{[i-1]}$ from the marginal of π on the first $i-1$ coordinates. Hence, it suffices to estimate $\mathbb{E}_{(i,x)}\left[D_{\mathrm{KL}}\left(\pi_i(\cdot \mid x) \,\|\, \mu_i(\cdot \mid x)\right)\right]$ with additive error ε/n.

Let $(i_1, x_1), \ldots, (i_L, x_L)$ be L independent random pairs generated via the General Oracle (which is contained in the power of Subcube Oracle), where we define

$$L = \left\lceil \frac{8n^2 \ln^2(1/b)}{\varepsilon^2} \right\rceil.$$

For $1 \leq \ell \leq L$, we let

$$R_\ell = D_{\mathrm{KL}}\left(\pi_{i_\ell}(\cdot \mid x_\ell) \,\|\, \mu_{i_\ell}(\cdot \mid x_\ell)\right).$$

Furthermore, for each ℓ let $\widehat{R}_\ell$ be an estimate of R_ℓ which is obtained from Lemma 7.6 via the Subcube Oracle, such that

$$\Pr\left(\left|\widehat{R}_\ell - R_\ell\right| \geq \frac{\varepsilon}{2n}\right) \leq \frac{1}{10L}.$$

Then, by the union bound we have

$$\Pr\left(\left|\frac{1}{L}\sum_{\ell=1}^{L}\widehat{R}_\ell - \frac{1}{L}\sum_{\ell=1}^{L}R_\ell\right| \geq \frac{\varepsilon}{2n}\right) \leq \sum_{\ell=1}^{L}\Pr\left(\left|\widehat{R}_\ell - R_\ell\right| \geq \frac{\varepsilon}{2n}\right) \leq L \cdot \frac{1}{10L} = \frac{1}{10}.$$

Note that using the standard amplification technique for the failure probability, the number of samples we need for each ℓ is

$$O\left(\frac{kn}{\varepsilon \log(kn/\varepsilon)} + \frac{n^2 \log^2(1/b)}{\varepsilon^2}\right) \cdot O\left(\log L\right) = O\left(\log^2\left(\frac{1}{b}\right) \cdot \frac{n^2}{\varepsilon^2}\log\left(\frac{n}{\varepsilon}\right)\right),$$

where we use the assumptions $k = O(n)$ and $\log\log(1/b) = O(\log n)$.

Meanwhile, we observe $0 \leq D_{\mathrm{KL}}\left(\pi_i(\cdot \mid x) \,\|\, \mu_i(\cdot \mid x)\right) \leq \ln(1/b)$ for any feasible pair (i, x) since μ is b-marginally bounded and $\pi \ll \mu$. Hence, Hoeffding's inequality implies that

$$\Pr\left(\left|\frac{1}{L}\sum_{\ell=1}^{L}R_\ell - \mathbb{E}_{(i,x)}\left[D_{\mathrm{KL}}\left(\pi_i(\cdot \mid x) \,\|\, \mu_i(\cdot \mid x)\right)\right]\right| \geq \frac{\varepsilon}{2n}\right) \leq 2\exp\left(-\frac{\varepsilon^2 L}{2n^2 \ln^2(1/b)}\right) \leq \frac{1}{10},$$

provided $L \geq (8n^2/\varepsilon^2)\ln^2(1/b)$.

Therefore, by letting our estimator to be

$$\widehat{S} = \frac{n}{L}\sum_{\ell=1}^{L}\widehat{R}_\ell,$$

we deduce that

$$\begin{aligned}
&\left|\widehat{S} - D_{\mathrm{KL}}\left(\pi \,\|\, \mu\right)\right| \\
&\quad\leq n\left|\frac{1}{L}\sum_{\ell=1}^{L}\widehat{R}_\ell - \frac{1}{L}\sum_{\ell=1}^{L}R_\ell\right| + n\left|\frac{1}{L}\sum_{\ell=1}^{L}R_\ell - \mathbb{E}_{(i,x)}\left[D_{\mathrm{KL}}\left(\pi_i(\cdot \mid x) \,\|\, \mu_i(\cdot \mid x)\right)\right]\right| \\
&\quad\leq n \cdot \frac{\varepsilon}{2n} + n \cdot \frac{\varepsilon}{2n} = \varepsilon,
\end{aligned}$$

with failure probability at most $1/10 + 1/10 = 1/5$ by the union bound. Finally, the query complexity is given by

$$O\left(\log^2\left(\frac{1}{b}\right) \cdot \frac{n^2}{\varepsilon^2}\log\left(\frac{n}{\varepsilon}\right)\right) \cdot L = O\left(\log^4\left(\frac{1}{b}\right) \cdot \frac{n^4}{\varepsilon^4}\log\left(\frac{n}{\varepsilon}\right)\right),$$

as claimed. □

We remark that Theorem 7.9 is applicable to all the examples mentioned in Sections 7.1 and 7.2. See also Remark 7.3 on relaxing the marginal boundedness condition.

8 Conclusion and Open Problems

In this article we give efficient algorithms for identity testing for the Coordinate Oracle model, and also establish matching computational hardness and information-theoretical lower bounds. Our algorithmic result builds on the fact that the visible distribution satisfies approximate tensorization of entropy. While we show that for the anti-ferromagnetic Ising model, there is no polynomial-time identity testing algorithm when approximate tensorization fails, it is in general unclear if one can get a testing algorithm running in polynomial time without approximate tensorization, using either Coordinate Oracle or Subcube Oracle in addition to General Oracle. One important example is the ferromagnetic Ising model at all temperatures. We know that approximate tensorization fails at low temperature (large β) since the Glauber dynamics has exponential mixing time. We do not know

whether an efficient identity testing algorithm exists or not even with access to the more powerful Subcube Oracle. Note that our Theorem 7.2 does not apply to ferromagnetic Ising models since we cannot estimate conditional marginal probabilities under an arbitrary pinning (corresponding to ferromagnetic Ising models with inconsistent local fields). Another important example is mixtures of product distributions. It is easy to show that approximate tensorization could fail even for a mixture of two product distributions with equal weights. We know from Theorem 7.2 that there is an efficient identity testing algorithm for the family of mixtures of polynomially many balanced product distributions given access to the Subcube Oracle. It is unclear to us, however, that if there is a polynomial-time testing algorithm using only the weaker Coordinate Oracle.

References

[1] Jayadev Acharya, Constantinos Daskalakis, and Gautam Kamath. 2015. Optimal testing for properties of distributions. *Advances in Neural Information Processing Systems (NeurIPS)* 28 (2015), 3591–3599.

[2] Nima Anari, Vishesh Jain, Frederic Koehler, Huy Tuan Pham, and Thuy-Duong Vuong. 2022. Entropic independence: Optimal mixing of down-up random walks. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing (STOC)*, 1418–1430.

[3] Nima Anari, Kuikui Liu, and Shayan Oveis Gharan. 2020. Spectral independence in high-dimensional expanders and applications to the hardcore model. In *Proceedings of the 61st Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, 1319–1330.

[4] Tugkan Batu, Eldar Fischer, Lance Fortnow, Ravi Kumar, Ronitt Rubinfeld, and Patrick White. 2001. Testing random variables for independence and identity. In *Proceedings of the 42nd Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, 442–451.

[5] Ivona Bezáková, Antonio Blanca, Zongchen Chen, Daniel Štefankovič, and Eric Vigoda. 2020. Lower bounds for testing graphical models: Colorings and antiferromagnetic Ising models. *Journal of Machine Learning Research* 21 (2020), 25:1–25:62.

[6] Arnab Bhattacharyya, Clément L. Canonne, and Joy Qiping Yang. 2022. Independence testing for bounded degree Bayesian network. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 35, 15027–15038.

[7] Arnab Bhattacharyya, Sutanu Gayen, Saravanan Kandasamy, and N. V. Vinodchandran. 2021. Testing product distributions: A closer look. In *Proceedings of the Algorithmic Learning Theory (ALT)*, 367–396.

[8] Rishiraj Bhattacharyya and Sourav Chakraborty. 2018. Property testing of joint distributions using conditional samples. *ACM Trans. Comput. Theory* 10, 4 (Dec 2018), 1–20. DOI: https://doi.org/10.1145/3241377

[9] Eric Blais, Clément L. Canonne, and Tom Gur. 2019. Distribution testing lower bounds via reductions from communication complexity. *ACM Transactions on Computation Theory* 11, 2, Article 6 (2019), 1–37.

[10] Antonio Blanca, Pietro Caputo, Zongchen Chen, Daniel Parisi, Daniel Štefankovič, and Eric Vigoda. 2022. On mixing of Markov chains: Coupling, spectral independence, and entropy factorization. *Electronic Journal of Probability* 27 (2022), 1–42.

[11] Antonio Blanca, Zongchen Chen, Daniel Štefankovič, and Eric Vigoda. 2021. Hardness of identity testing for restricted Boltzmann machines and Potts models. *Journal of Machine Learning Research* 22 (2021), 152:1–152:56.

[12] Sergej G. Bobkov and Friedrich Götze. 1999. Exponential integrability and transportation cost related to logarithmic Sobolev inequalities. *Journal of Functional Analysis* 163, 1 (1999), 1–28.

[13] Gerandy Brito, Ioana Dumitriu, and Kameron Decker Harris. 2022. Spectral gap in random bipartite biregular graphs and applications. *Combinatorics, Probability and Computing* 31, 2 (2022), 229–267.

[14] Jin-Yi Cai, Andreas Galanis, Leslie Ann Goldberg, Heng Guo, Mark Jerrum, Daniel Štefankovič, and Eric Vigoda. 2016. #BIS-hardness for 2-spin systems on bipartite bounded degree graphs in the tree non-uniqueness region. *Journal of Computer and System Sciences* 82, 5 (2016), 690–711.

[15] Clément L. Canonne. 2020. A survey on distribution testing: Your data is big. But is it blue? *Theory of Computing Library Graduate Surveys* 9 (2020), 1–100.

[16] Clément L. Canonne. 2022. Topics and techniques in distribution testing: A biased but representative sample. *Foundations and Trends in Communications and Information Theory* 19, 6 (2022), 1032–1198.

[17] Clement L. Canonne, Xi Chen, Gautam Kamath, Amit Levi, and Erik Waingarten. 2021. Random restrictions of high dimensional distributions and uniformity testing with subcube conditioning. In *Proceedings of the 32nd Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 321–336.

[18] Clément L. Canonne, Ilias Diakonikolas, Daniel M. Kane, and Alistair Stewart. 2020. Testing Bayesian networks. *IEEE Transactions on Information Theory* 66, 5 (2020), 3132–3170.

[19] Clément L. Canonne, Dana Ron, and Rocco A. Servedio. 2015. Testing probability distributions using conditional samples. *SIAM Journal on Computing* 44, 3 (2015), 540–616.
[20] Pietro Caputo, Georg Menz, and Prasad Tetali. 2015. Approximate tensorization of entropy at high temperature. *Annales de la Faculté des Sciences de Toulouse: Mathématiques* 6, 24, 4 (2015), 691–716.
[21] Pietro Caputo and Daniel Parisi. 2021. Block factorization of the relative entropy via spatial mixing. *Communications in Mathematical Physics* 388, 2 (2021), 793–818.
[22] Filippo Cesi. 2001. Quasi-factorization of the entropy and logarithmic Sobolev inequalities for Gibbs random fields. *Probability Theory and Related Fields* 120, 4 (2001), 569–584.
[23] Sourav Chakraborty, Eldar Fischer, Yonatan Goldhirsh, and Arie Matsliah. 2016. On the power of conditional samples in distribution testing. *SIAM Journal on Computing* 45, 4 (2016), 1261–1296.
[24] Siu-On Chan, Ilias Diakonikolas, Paul Valiant, and Gregory Valiant. 2014. Optimal algorithms for testing closeness of discrete distributions. In *Proceedings of the 25th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 1193–1203.
[25] Xi Chen, Rajesh Jayaram, Amit Levi, and Erik Waingarten. 2021b. Learning and testing junta distributions with subcube conditioning. In *Proceedings of the 34th Conference on Learning Theory (COLT)*, Vol. 134, 1060–1113.
[26] Zongchen Chen, Andreas Galanis, Daniel Štefankovič, and Eric Vigoda. 2021a. Rapid mixing for colorings via spectral independence. In *Proceedings of the 32nd Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 1548–1557.
[27] Zongchen Chen, Kuikui Liu, and Eric Vigoda. 2020. Rapid mixing of Glauber dynamics up to uniqueness via contraction. In *Proceedings of the 61st Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, 1307–1318.
[28] Zongchen Chen, Kuikui Liu, and Eric Vigoda. 2021c. Optimal mixing of Glauber dynamics: Entropy factorization via high-dimensional expansion. In *Proceedings of the 53rd Annual ACM Symposium on Theory of Computing (STOC)*, 1537–1550.
[29] Zongchen Chen, Kuikui Liu, and Eric Vigoda. 2021d. Spectral independence via stability and applications to Holant-type problems. In *Proceedings of the 62nd IEEE Annual Symposium on Foundations of Computer Science (FOCS)*, 149–160.
[30] Zongchen Chen, Nitya Mani, and Ankur Moitra. 2023. From algorithms to connectivity and back: Finding a giant component in random k-SAT. In *Proceedings of the 34th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 3437–3470.
[31] Constantinos Daskalakis, Nishanth Dikkala, and Gautam Kamath. 2019. Testing Ising models. *IEEE Transactions on Information Theory* 65, 11 (2019), 6829–6852.
[32] Constantinos Daskalakis, Gautam Kamath, and John Wright. 2018. Which distribution distances are sublinearly testable? In *Proceedings of the 29th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2747–2764.
[33] Constantinos Daskalakis and Qinxuan Pan. 2017. Square Hellinger subadditivity for Bayesian networks and its applications to identity testing. In *Proceedings of the Conference on Learning Theory (COLT)*, 697–703.
[34] Ilias Diakonikolas, Themis Gouleakis, Daniel M. Kane, John Peebles, and Eric Price. 2021. Optimal testing of discrete distributions with high probability. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing (STOC)*, 542–555.
[35] Ilias Diakonikolas and Daniel M. Kane. 2016. A new approach for testing properties of discrete distributions. In *Proceedings of the 57th IEEE Annual Symposium on Foundations of Computer Science (FOCS)*, 685–694.
[36] Ronen Eldan, Frederic Koehler, and Ofer Zeitouni. 2022. A spectral condition for spectral gap: Fast mixing in high-temperature Ising models. *Probability Theory and Related Fields* 182 (2022), 1035–1051.
[37] Moein Falahatgar, Ashkan Jafarpour, Alon Orlitsky, Venkatadheeraj Pichapati, and Ananda Theertha Suresh. 2015. Faster algorithms for testing under conditional sampling. In *Proceedings of the 28th Conference on Learning Theory (COLT)*, Vol. 40, 607–636.
[38] Weiming Feng, Heng Guo, Yitong Yin, and Chihao Zhang. 2022. Rapid mixing from spectral independence beyond the Boolean domain. *ACM Transactions on Algorithms* 18, 3 (2022), Article 28, 1–32 pages.
[39] Joel N. Franklin. 2012. *Matrix Theory*. Dover Publications, Inc.
[40] Tobias Friedrich, Andreas Göbel, Martin S. Krejca, and Marcus Pappik. 2022. A spectral independence view on hard spheres via block dynamics. *SIAM Journal on Discrete Mathematics* 36, 3 (2022), 2282–2322.
[41] Andreas Galanis, Leslie Ann Goldberg, Heng Guo, and Andrés Herrera-Poyatos. 2022. Fast sampling of satisfying assignments from random k-SAT. arXiv:2206.15308.
[42] Andreas Galanis, Daniel Štefankovič, and Eric Vigoda. 2016. Inapproximability of the partition function for the antiferromagnetic Ising and hard-core models. *Combinatorics, Probability and Computing* 25, 4 (2016), 500–559.
[43] David Galvin and Jeff Kahn. 2004. On phase transition in the hard-core model on Z^d. *Combinatorics, Probability and Computing* 13, 2 (2004), 137–164.
[44] C. D. Godsil. 1984. Spectra of trees. In *North-Holland Mathematics Studies*, Vol. 87. North-Holland, 151–159.
[45] Leslie Ann Goldberg and Mark Jerrum. 2007. The complexity of ferromagnetic Ising with local fields. *Combinatorics, Probability and Computing* 16, 1 (2007), 43–61.

[46] Oded Goldreich. 2004. *Foundations of Cryptography. II: Basic Applications*. Cambridge University Press.
[47] Oded Goldreich. 2017. *Introduction to Property Testing*. Cambridge University Press.
[48] Oded Goldreich. 2020. The uniform distribution is complete with respect to testing identity to a fixed distribution. In *Computational Complexity and Property Testing. Lecture Notes in Computer Science*, Vol. 12050. Springer, Cham., 152–172.
[49] Friedrich Götze, Holger Sambale, and Arthur Sinulis. 2019. Higher order concentration for functions of weakly dependent random variables. *Electronic Journal of Probability* 24 (2019), 1–19.
[50] Matthew Jenssen, Peter Keevash, and Will Perkins. 2020. Algorithms for #BIS-hard problems on expander graphs. *SIAM Journal on Computing* 49, 4 (2020), 681–710.
[51] Mark Jerrum and Alistair Sinclair. 1989. Approximating the permanent. *SIAM Journal on Computing* 18, 6 (1989), 1149–1178.
[52] Jiantao Jiao, Kartik Venkat, Yanjun Han, and Tsachy Weissman. 2015. Minimax estimation of functionals of discrete distributions. *IEEE Transactions on Information Theory* 61, 5 (2015), 2835–2885.
[53] Adam Klivans and Raghu Meka. 2017. Learning graphical models using multiplicative weights. In *Proceedings of the 58th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, 343–354.
[54] Frederic Koehler, Holden Lee, and Andrej Risteski. 2022. Sampling approximately low-rank Ising models: MCMC meets variational methods. In *Proceedings of the 35th Conference on Learning Theory (COLT)*, 4945–4988.
[55] Kuikui Liu. 2021. From coupling to spectral independence and blackbox comparison with the down-up walk. In *Proceedings of the Randomization and Approximation Techniques in Computer Science (RANDOM)*, 32:1–32:21.
[56] Fabio Martinelli, Alistair Sinclair, and Dror Weitz. 2003. The Ising model on trees: Boundary conditions and mixing time. In *Proceedings of the 44th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*. IEEE, 628–639.
[57] Katalin Marton. 2019. Logarithmic Sobolev inequalities in discrete product spaces. *Combinatorics, Probability and Computing* 28, 6 (2019), 919–935.
[58] Michael S. O. Molloy, Hanna Robalewska, Robert W. Robinson, and Nicholas C. Wormald. 1997. 1-Factorizations of random regular graphs. *Random Structures & Algorithms* 10, 3 (1997), 305–321.
[59] Ravi Montenegro and Prasad Tetali. 2006. Mathematical aspects of mixing times in Markov chains. *Foundations and Trends® in Theoretical Computer Science* 1, 3 (2006), 237–354.
[60] Shyam Narayanan. 2021. on tolerant distribution testing in the conditional sampling model. In *Proceedings of the 32nd Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 357–373.
[61] Liam Paninski. 2008. A coincidence-based test for uniformity given very sparsely sampled discrete data. *IEEE Transactions on Information Theory* 54, 10 (2008), 4750–4755.
[62] James Gary Propp and David Bruce Wilson. 1996. Exact sampling with coupled Markov chains and applications to statistical mechanics. *Random Structures & Algorithms* 9, 1–2 (1996), 223–252.
[63] Allan Sly. 2008. Uniqueness thresholds on trees versus graphs. *The Annals of Applied Probability* 18, 5 (2008), 1897–1909.
[64] Allan Sly. 2010. Computational transition at the uniqueness threshold. In *Proceedings of the 51st Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, 287–296.
[65] Allan Sly and Nike Sun. 2014. Counting in two-spin models on d-regular graphs. *Annals of Probability* 42, 6 (2014), 2383–2416.
[66] Gregory Valiant and Paul Valiant. 2011. The power of linear estimators. In *Proceedings of the 52nd IEEE Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE, 403–412.
[67] Gregory Valiant and Paul Valiant. 2017a. An automatic inequality prover and instance optimal identity testing. *SIAM Journal on Computing* 46, 1 (2017), 429–455.
[68] Gregory Valiant and Paul Valiant. 2017b. Estimating the unseen: Improved estimators for entropy and other properties. *Journal of the ACM (JACM)* 64, 6 (2017), 1–41.
[69] Yihong Wu and Pengkun Yang. 2016. Minimax rates of entropy estimation on large alphabets via best polynomial approximation. *IEEE Transactions on Information Theory* 62, 6 (2016), 3702–3720.

Received 11 July 2023; revised 6 May 2024; accepted 21 July 2024