

# Sampling and Identity-Testing Without Approximate Tensorization of Entropy

William Gay<sup>\*</sup>   William He<sup>†</sup>   Nicholas Kocurek<sup>‡</sup>   Ryan O'Donnell<sup>§</sup>

July 5, 2025

## Abstract

Certain tasks in high-dimensional statistics become easier when the underlying distribution satisfies a local-to-global property called *approximate tensorization of entropy* (ATE). For example, the Glauber dynamics Markov chain of an ATE distribution mixes fast and can produce approximate samples in a small amount of time, since such a distribution satisfies a *modified log-Sobolev inequality*. Moreover, identity-testing for an ATE distribution requires few samples if the tester is given *coordinate conditional* access to the unknown distribution, as shown by Blanca, Chen, Štefankovič, and Vigoda [BCSV23].

A natural class of distributions that do *not* satisfy ATE consists of *mixtures* of (few) distributions that do satisfy ATE. We study the complexity of identity-testing and sampling for these distributions. Our main results are the following:

1. We show fast mixing of Glauber dynamics from a *data-based initialization*, with optimal sample complexity, for mixtures of distributions satisfying modified log-Sobolev inequalities. This extends work of Huang, Koehler, Lee, Mohanty, Rajaraman, Vuong, and Wu [KLV24, HMRW24] for mixtures of distributions satisfying Poincaré inequalities.
2. Answering an open question posed by Blanca et al., we give efficient identity-testers for mixtures of ATE distributions in the coordinate-conditional sampling access model. We also give some simplifications and improvements to the original algorithm of Blanca et al.

---

<sup>\*</sup>Carnegie Mellon University. [wgay@andrew.cmu.edu](mailto:wgay@andrew.cmu.edu)

<sup>†</sup>Carnegie Mellon University. [wrhe@cs.cmu.edu](mailto:wrhe@cs.cmu.edu)

<sup>‡</sup>Carnegie Mellon University. [nkocurek@andrew.cmu.edu](mailto:nkocurek@andrew.cmu.edu)

<sup>§</sup>Carnegie Mellon University. [odonnell@cs.cmu.edu](mailto:odonnell@cs.cmu.edu). Supported in part by a grant from Google Quantum AI.

# 1 Introduction

## 1.1 Approximate Tensorization of Entropy

Let  $\mu$  be a distribution on the discrete product set  $\Sigma^n$ . If  $f$  is a (non-negative) real-valued function on  $\Sigma^n$ , then the following functional captures the amount of local variation that  $f$  has, where “local” means with respect to varying a single component of  $\Sigma^n$ :

**Definition 1.** We write  $\mathcal{L}_\mu$  for the functional on functions  $f : \Sigma^n \rightarrow \mathbb{R}_{\geq 0}$  given by

$$\mathcal{L}_\mu[f] = \sum_{i \in [n]} \mathbf{E}_{\mathbf{x} \sim \mu} \left[ \mathbf{Ent}_{\mathbf{y} \sim \mu|_{\mathbf{x}_{\setminus i}}} [f(\mathbf{y})] \right].$$

Here  $\mathbf{Ent}_\mu[\cdot]$  is the standard entropy functional with respect to  $\mu$ . See [Section 2](#) for a full definition.

Every product distribution  $\mu$  on  $\Sigma^n$  satisfies *tensorization of entropy*, meaning  $\mathcal{L}_\mu[f] \geq \mathbf{Ent}_\mu[f]$  for all  $f$ . Much work in the context of Markov chain mixing is focused on establishing an approximate version of this inequality for distributions of interest:

**Definition 2.** A distribution  $\mu$  on  $\Sigma^n$  satisfies *approximate tensorization of entropy (ATE)* with constant  $c^*$  if for all  $f : \Sigma^n \rightarrow \mathbb{R}_{\geq 0}$ ,

$$\mathbf{Ent}_\mu[f] \leq c^* \cdot \mathcal{L}_\mu[f].$$

Note that  $c^* \geq 1$  unless  $\mu$  is a point mass (in which case both sides are always 0).

A motivation for [Definition 2](#) is the fact that many statistical tasks related to  $\mu$  become easier when  $\mu$  satisfies  $c^*$ -ATE with a small  $c^*$  (close to 1).

**Sampling.** One such statistical task is that of approximately sampling from  $\mu$ , given oracle access to  $\mu$  up to proportionality. When  $\mu$  satisfies  $c^*$ -ATE, then it is well-known that the Glauber dynamics Markov chain for  $\mu$  mixes in time  $\tilde{O}(c^*n)$ , and Markov chain Monte Carlo techniques are very effective for drawing approximate samples from  $\mu$ . Thus, establishing ATE is instrumental for obtaining optimal mixing times for Glauber dynamic chains for natural distributions on high-dimensional spaces, such as Gibbs distributions of certain spin systems at high temperatures. See, for example, [\[BT06, CMT15, AJK<sup>+</sup>21, CLV21, CE22, BCC<sup>+</sup>22, HS23, CMM23, CGG<sup>+</sup>24\]](#).

**Identity-Testing.** Another statistical task where ATE helps significantly is identity-testing, also known as hypothesis-testing or goodness-of-fit testing. In this setting, a testing algorithm is given some distance measure  $D$  between distributions, a threshold parameter  $\varepsilon$ , access to some kind of description of a known (“visible”) distribution  $\mu$ , and some kind of sample access to an unknown distribution  $\pi$ . The tester must then satisfy the following performance guarantees:

1. If  $\pi = \mu$  then the tester accepts  $\pi$  with high probability.
2. If  $D(\pi \parallel \mu) \geq \varepsilon$  then the tester rejects  $\pi$  with high probability.

The study of the complexity of this problem has a long history in statistics [\[Pea00, Fis66\]](#). This is a fundamental problem in science, where one wants to confirm that the behavior of some system conforms to that of a purported model for that system.

Motivated by previous works on identity-testing with alternative access models (see [Section 1.4](#) for a more detailed account of the literature), Blanca, Chen, Stefankovic, and Vigoda [BCSV23] studied the model of coordinate conditional access, which is a common relaxation of subcube conditioning access and pairwise conditional access when  $\Sigma$  is binary. In this access model, the tester gets access to samples from the unknown distribution  $\pi$ , and also gets access to  $\pi|_{x_{\setminus i}}$  for any  $x \in \Sigma^n$ . Here  $x_{\setminus i} = \{y \in \Sigma^n : |x - y|_0 \leq 1\}$ . Blanca et al. proved that distributions  $\mu$  satisfying ATE admit very efficient identity testers.

More precisely, let the “General Oracle” be an oracle that, when queried, outputs a sample drawn from the unknown distribution  $\pi$ , and let the “Coordinate Oracle” be the oracle that, when queried with a pair  $(x, i)$ , outputs a sample from  $\pi|_{x_{\setminus i}}$ .

**Theorem 3** ([BCSV23], Theorem 4.1). Let  $\mu$  be a distribution on  $\Sigma^n$  that is  $\eta$ -balanced (see [Definition 27](#)), fully supported, and ATE with constant  $c^*$ . Assuming

$$c^* \leq \text{poly}(n), \quad \eta \geq \exp(-\text{poly}(n)),$$

there is a testing algorithm for  $\mu$  with access to the Coordinate Oracle and General Oracle having

$$\text{sample complexity} \leq O\left(\frac{c^*n}{\varepsilon}\right) \cdot \log^3\left(\frac{n}{\varepsilon}\right) \cdot f(\eta), \quad \text{where } f(\eta) = \begin{cases} \log(1/\eta) & \text{if } |\Sigma| = 2, \\ \frac{1}{\sqrt{\eta}} & \text{if } |\Sigma| \geq 3. \end{cases}$$

Also, in the  $|\Sigma| \geq 3$  case one can reduce the dependence on  $1/\eta$  back to logarithmic at the expense of being quadratic in the other parameters; precisely, one can also achieve

$$\text{sample complexity} \leq O\left(\frac{c^*n}{\varepsilon}\right)^2 \cdot \log^2\left(\frac{n}{\varepsilon}\right) \cdot \sqrt{|\Sigma|} \cdot \log(1/\eta).$$

## 1.2 Mixtures of ATE Distributions and Our Results

As described, certain statistical tasks become much easier when the related distribution  $\mu$  satisfies approximate tensorization of entropy. This leads us to consider cases in which this property fails. A very natural class of distributions that do not satisfy ATE is the class of mixtures of distributions that *do* satisfy ATE. We consider distributions  $\mu$  of the form  $\sum_{a \in [k]} \rho(a) \mu_a$ , where  $\rho$  is a distribution on  $[k]$  and each  $\mu_a$  satisfies  $c^*$ -ATE. We call the  $\mu_a$ ’s the *mixture components*.

Our first result concerns the sampling task. It is easy to see that certain mixtures of ATE distributions have exponentially large mixing times for Glauber dynamics (for example the equal mixture of the 0.1- and 0.9-biased distributions on  $\{0, 1\}^n$ ). However, this is a lower bound on the mixing time from a *worst-case initialization*. We instead show that when the chain is initialized with a *data-based initialization*, it still experiences fast mixing, given that there is enough data for the initialization:

**Theorem 4.** Let  $\mu = \sum_{a=1}^k \rho(a) \mu_a$  be a mixture on  $\Sigma^n$  with each component satisfying  $c^*$ -ATE (or, more weakly, satisfying  $\frac{1}{c^*n}$ -MLSI; see [Definition 15](#) and [Remark 17](#)). Let  $\pi$  be an empirical distribution induced by

$$m = O(k/\varepsilon + \log(1/\delta)/\varepsilon) \tag{1}$$

independent samples from  $\mu$ . Then with probability at least  $1 - \delta$  over these samples, the Glauber dynamics for  $\mu$  warm-started at  $\pi$  mixes to  $D_{\text{KL}}(P^t \pi \parallel \mu) \leq \varepsilon$  in continuous-time  $t = c^*n \cdot O(\log \log(1/\min_x \mu(x)) + \log(1/\varepsilon))$ .

Our next result concerns the identity-testing task. We show that with the same access model studied by [BCSV23], we can efficiently identity-test any  $\mu$  that is a mixture of few ATE distributions, answering the open question posed in that paper:

**Theorem 5.** Let  $\mu = \sum_{a=1}^k \rho(a) \mu_a$  be a mixture on  $\Sigma^n$  with each component satisfying  $c^*$ -ATE. Assume also that  $\mu$  is  $\eta$ -balanced (see Definition 27). Then there is an identity-testing algorithm for  $\mu$  that uses

$$O\left(\frac{c^*n}{\varepsilon}\right) \cdot \log^2\left(\frac{c^*n}{\varepsilon}\right) \cdot \log(1/\eta) \cdot \log \log(1/\eta) \cdot \sqrt{|\Sigma|} + O\left(\frac{\sqrt{k} \cdot \log(1/\rho^*)}{\varepsilon}\right) \quad (2)$$

calls to the General Oracle and Coordinate Oracle.

**Remark 6.** Our algorithm, and the algorithm of [BCSV23], do not use the full power of the Coordinate Oracle. The algorithms even work in the setting where the set of calls that the tester can make to the Coordinate Oracle is predetermined as a random set of pairs  $\mathbf{x} \sim \pi$  and  $\mathbf{i} \in [n]$ . Moreover, given an efficiently accessible description of the distribution  $\mu$ , the test has similarly efficient computational complexity. As mentioned in [BCSV23], this efficient description is equivalent to the ability to efficiently implement single steps of Glauber dynamics for the distribution  $\mu$ .

As mentioned before, in the case where  $k = 1$ , Theorems 4 and 5 are essentially known from previous work. In analyzing MCMC algorithms for sampling from  $\mu$ , one wants to show that the KL-divergence between the state of the Markov chain and  $\mu$  is contracting. When analyzing identity-testing algorithms for  $\mu$  one wants show that if an unknown distribution  $\pi$  has large KL-divergence from  $\mu$ , then this is captured by the divergences between neighbors in  $\Sigma^n$ . The ATE inequality then relates the contraction of KL-divergence through the Markov chain and the divergences between neighbors to the original KL-divergence.

However, when  $\mu$  is merely a mixture of distributions that individually satisfy ATE, such a local-to-global property fails. That is, the variation between  $\mu$  and some alternative distribution  $\pi$  is not conserved when zooming in from the entire distribution  $\mu$  to  $\mu|_{\mathbf{x} \setminus \mathbf{i}}$ . In fact, when the variation between  $\mu$  and  $\pi$  is well-captured by the local variation, we are essentially done.

In order to deal with distributions  $\pi$  whose variations are not captured locally, one needs to identify where exactly the variation is captured, and then deal with such cases accordingly. The chain rule for entropy (Lemma 11) allows us to identify where the variation is captured, and we develop new techniques to deal with such cases.

### 1.3 Related Work in Sampling

Our result Theorem 4 can be seen as an entropic analogue of the works of Koehler, Lee, and Vuong [KLV24] and Huang, Mohanty, Rajaraman, and Wu [HMRW24], both of which show a similar result when each component satisfies Poincaré inequality rather than a modified log-Sobolev inequality.

More concretely, suppose each mixture component in  $\mu$  satisfies a Poincaré inequality. Using higher-order spectral gaps for the Glauber dynamics chain, [KLV24] were able to establish that reasonably fast mixing in TV distance occurs from a warm-start having a near-minimal number of samples. However, due to their use of Poincaré inequalities, the mixing time they conclude can be suboptimal. They also prove a version of their main theorem with a faster mixing time using standard log-Sobolev inequalities, but this relies on the hypothesis that each mixture component satisfies a log-Sobolev inequality, which is a stronger assumption than an MLSI.

[HMRW24] obtains a result similar to [KLV24], again under the assumption that each mixture component satisfies a Poincaré inequality. However, their result is quantitatively much weaker in sample complexity: (1) it has a larger dependence on  $k$  and  $\varepsilon$ ; and, (2) it has a dependence on the minimum mixing weight  $\rho_a$ , which does not appear in [KLV24]. [HMRW24] also suffers from suboptimal mixing time due to use of Poincaré inequalities, but it does achieve mixing in  $\chi^2$ -divergence, which is stronger than mixing in TV distance.

Our [Theorem 4](#) simultaneously achieves: (1) optimal mixing times in TV distance in the case that  $\mu$  is a product distribution, by using MLSIs rather than Poincaré inequalities; and, (2) optimal sample complexity in the case where  $\mu$  is a mixture of isolated point masses. A key point in our analysis is that we avoid using standard Chernoff bounds, which seem to necessarily introduce a dependence on the minimum mixing weight, and instead employ a bound on the moment generating function (m.g.f.) for the KL-divergence of an empirical distribution from the true distribution.

## 1.4 Related Work in Testing

Our [Theorem 5](#) fits into a broader line of work on testing distributions with alternative sampling models, which is motivated by the fact that testing algorithms with access to i.i.d. samples from a high-dimensional distribution often require an exponential number of samples.

In the direction of providing more powerful models of access to distributions in identity-testing, concurrent works of Canonne, Ron, and Servedio [CRS14] and Chakraborty, Fischer, Goldhirsch, and Matsliah [CFG13] introduced the problem of identity-testing with access to conditional samples from  $\pi$ . That is, instead of getting access to i.i.d. samples from the unknown distribution  $\pi$ , the tester can specify a subset  $S$  of the underlying set of outcomes and receive samples from  $\pi|_S$ . These papers showed that *any* distribution admits an identity-tester in this model whose sample complexity is  $\text{poly}(1/\varepsilon)$ , where  $\varepsilon$  denotes the minimum distance of distributions that should be rejected with high probability. Subsequent work [FJO<sup>+</sup>15] continued the study of testers with general conditional samples.

Other access models to the unknown distribution  $\pi$  have also been considered. For example, [CR14] considered the dual query and cumulative dual query models in which one can explicitly query the probability density function on points and subsets of the universe, respectively. See also later works [CKOS15, NT23].

A model in which one can sample from  $\pi$  conditioned on an arbitrary subset of the universe is rather strong; in many settings it might be unclear how to simulate/obtain such samples. There are, however, natural weakenings of this access model in the high-dimensional setting. In particular, suppose  $\Sigma^n$  is the set of configurations of a system of  $n$  particles (or individuals, organisms, etc.). Then it might be significantly more reasonable to obtain samples from “subcubes” of  $\Sigma^n$ , meaning conditional distributions in which a subset of the particles have fixed states. Indeed, this subcube conditioning model was introduced by Bhattacharyya and Chakraborty [BC18], who showed that  $\tilde{O}(n^2)$  subcube-conditioned samples suffice for identity-testing of arbitrary distributions on  $\Sigma^n$ . Subsequent works [CCK<sup>+</sup>21, CJLW21] provided improvements in cases where assumptions on the visible distribution  $\mu$  are made.

Another realistic relaxation of arbitrary conditioning studied was that of pairwise conditioning, or conditioning under subsets of size 2, which was also studied in [CRS14]. Narayanan [Nar21] provided testing algorithms for arbitrary distributions with complexity  $\tilde{O}\left(\frac{\sqrt{n}}{\varepsilon^2}\right)$ .

We provide some further motivation for the coordinate conditional sampling access model not mentioned in [BCSV23]. Our motivation is based on practical matters, and we suggest a situation in which one might be able to easily simulate coordinate conditional access in cases where subcube conditional access and pairwise conditional access might not be easily simulable. In our situation,

we again regard  $\Sigma^n$  as a configuration of  $n$  particles, and we think of  $\pi$  as some distribution on the set of configurations of those  $n$  particles, potentially the Gibbs distribution for some set of interactions between the particles.

A natural model for the evolution of the configuration over time is that of Glauber dynamics. In the Glauber dynamics process, given a current configuration  $x \in \Sigma^n$ , the next configuration is chosen from  $\pi|_{x_{\setminus i}}$ . Assume the ability to i) arbitrary fix configurations of particles; and ii) simulate Glauber dynamics for a distribution  $\pi$ . Then to simulate access to  $\pi|_{x_{\setminus i}}$ , one can repeatedly initialize the system to  $x$  and simulate one step of Glauber dynamics for  $\pi$  starting from  $x$ . Any step that updates the  $j$ th particle for  $j \neq i$  is ignored, and the result given by an update to the  $i$ th particle is a sample from  $\pi|_{x_{\setminus i}}$ . Note here that we assume that we can tell when a site undergoes resampling, even if the resampling does not result in a new state.

## 2 The Chain Rule for Entropy

### 2.1 Entropies and Divergences

The following notion of  $\Phi$ -entropy was introduced in [Cha04]:

**Definition 7.** Let  $\Phi$  be a smooth and convex function mapping some interval of real numbers to the nonnegative real numbers. Let  $\mu$  be a probability distribution on a finite set  $\Omega$ . The  $\Phi$ -entropy of a function  $f : \Omega \rightarrow \mathbb{R}$  with respect to  $\mu$  is defined to be

$$\mathbf{Ent}_{\mu}^{\Phi}[f] := \mathbf{E}_{x \sim \mu} [\Phi(f(x))] - \Phi\left(\mathbf{E}_{x \sim \mu} [f(x)]\right).$$

**Fact 8.** If  $\Phi$  is convex then  $\mathbf{Ent}_{\mu}^{\Phi}[f] \geq 0$ .

**Remark 9.** Assume  $\Phi(1) = 0$  and let  $\pi$  and  $\mu$  be probability distributions on  $\Omega$ . Then the  $\Phi$ -entropy functional

$$\mathbf{Ent}_{\mu}^{\Phi}\left[\frac{\pi}{\mu}\right] = D_{\Phi}(\pi \parallel \mu)$$

is also known as the  $\Phi$ -divergence<sup>1</sup> between  $\pi$  and  $\mu$ . For example,

$$\mathbf{Ent}_{\mu}^{u \rightarrow u \log u}\left[\frac{\pi}{\mu}\right] = D_{\text{KL}}(\pi \parallel \mu).$$

As usual, we use the convention  $0 \log 0 = 0$ , and that this function is defined on  $\mathbb{R}_{\geq 0}$ . The quantity  $D_{\text{KL}}(\pi \parallel \mu)$  may be  $\infty$  (when  $\pi \not\leq \mu$ ).

We are most often interested in the case  $\Phi(u) = u \log u$  throughout the paper, so when we drop the  $\Phi$  in the superscript, the case  $\mathbf{Ent}_{\mu} = \mathbf{Ent}_{\mu}^{u \log u}$  is assumed. An important property of this specific choice of  $\Phi$ , used in our identity-testing result, is that the resulting entropy functional is 1-homogeneous:

**Fact 10.**  $\mathbf{Ent}$  is 1-homogeneous: That is, if  $\alpha$  is a nonnegative scalar, then  $\mathbf{Ent}_{\mu}[\alpha f] = \alpha \mathbf{Ent}_{\mu}[f]$ .

<sup>1</sup>The notation of “ $f$ -divergence” is more common, but to avoid notational overload we use the symbol “ $\Phi$ .”

## 2.2 The Chain Rule

When  $\Phi(u) = u^2$ , the following [Lemma 11](#) is known as the law of total variance. When  $\Phi(u) = u \log u$ , it is known as the chain rule for entropy. Both of these tools are of great use in establishing Poincaré and log-Sobolev inequalities. See, for example, [\[LY98, Sal21\]](#).

The fact itself is standard (see, e.g., [\[BG18\]](#)), but we give a proof in [Appendix A](#) for the convenience of the reader:

**Lemma 11.** If  $\mu = \sum_{a=1}^k \rho(a)\mu_a$ , then  $\mathbf{Ent}_\mu^\Phi[f] = \mathbf{Ent}_{a \sim \rho}^\Phi[\mathbf{E}_{\mu_a}[f]] + \mathbf{E}_{a \sim \rho}[\mathbf{Ent}_{\mu_a}^\Phi[f]]$ .

[Lemma 11](#) is especially useful when a distribution  $\mu$  is a mixture of many distributions. In the case where each of the mixture components satisfies approximate tensorization of entropy, we can use [Lemma 11](#) to show that the local entropy of some function  $f$  under the distribution  $\mu$  is lower-bounded by the portion of the entropy of  $f$  that arises as “intra-component” entropy.

**Lemma 12.** If  $\mu = \sum_{a=1}^k \rho(a)\mu_a$ , where each  $\mu_a$  satisfies  $c^*$ -ATE, then for any distribution  $\pi$  on  $\Sigma^n$  we have

$$c^* \cdot \mathcal{L}_\mu[f] \geq \mathbf{E}_{a \sim \rho} \left[ \mathbf{Ent}_{x \sim \mu_a} [f(x)] \right].$$

[Lemma 12](#) is proved in [Appendix A](#). [Lemma 12](#) essentially shows that the intra-component contribution to the entropy of  $f$  is captured by the local entropy in the case where the mixture components satisfy ATE. In the context of our sampling result, this implies that any initialization that is evenly balanced across the components will experience fast mixing. In the context of our identity-testing result, this implies that any unknown distribution that is evenly balanced across the components yet still far from the target distribution will be rejected by a local tester.

The main contribution of this paper is showing how to also handle the inter-component contribution in our applications. That is, we need to show that data-based initializations are evenly balanced with high probability and that our identity-testing algorithm will reject unbalanced distributions.

For this purpose it will be helpful to characterize the inter-component entropy  $\mathbf{Ent}_{a \sim \rho}[\mathbf{E}_{\mu_a}[f]]$  in the case that  $f = \pi/\mu$  is a density function. We can notice that  $g$  defined by  $a \mapsto \mathbf{E}_{\mu_a}[f]$  is itself a density on  $[k]$  vis-a-vis  $\rho$ . More explicitly, define  $\rho_\pi$  to be the probability distribution on  $[k]$  induced by sampling  $x \sim \pi$ , and then drawing  $a \sim \rho_x$  where  $\rho_x$  is the posterior of  $x$  from  $\mu$  with respect to  $\rho$ . In other words

$$\rho_\pi(a) = \sum_{x \in \Omega} \pi(x) \cdot \frac{\rho(a)\mu_a(x)}{\mu(x)}.$$

Now we can observe that  $g$  is indeed the density  $\rho_\pi/\rho$ ; hence:

**Fact 13.** For any mixture  $\mu = \sum_{a=1}^k \rho(a)\mu_a$  we have  $\mathbf{Ent}_{a \sim \rho}^\Phi[\mathbf{E}_{\mu_a}[\pi/\mu]] = D_\Phi(\rho_\pi \parallel \rho)$ .

## 3 Sampling from Data-Based Initializations

In this section we prove [Theorem 4](#). An important observation made in [\[HMRW24\]](#) is that the Glauber dynamics chain for a mixture of distributions satisfying either Poincaré or modified log-Sobolev inequalities satisfies a *weak Poincaré* or *weak modified log-Sobolev inequality* respectively. They then use these weak functional inequalities to infer fast mixing. We first generalize this result to the setting of  $\Phi$ -Sobolev inequalities.

**Definition 14.** Let  $P$  be a reversible discrete-time Markov operator on  $\Omega$  with stationary distribution  $\mu$ . Then the Dirichlet form with respect to  $P$  is defined on functions  $f, g : \Omega \rightarrow \mathbb{R}$  as

$$\mathcal{E}_P(f, g) = \mathbf{E}_{\mathbf{x} \sim \mu} \mathbf{E}_{\mathbf{y} \sim P\mathbf{x}} [(f(\mathbf{x}) - f(\mathbf{y}))(g(\mathbf{x}) - g(\mathbf{y}))].$$

We will study inequalities relating the Dirichlet form, which is a measure of local variation, to measures of global variation in the case where the associated Markov chain is the Glauber dynamics chain with respect to a distribution  $\mu$  on  $\Sigma^n$ . The Glauber dynamics chain is the chain in which given a current state  $\mathbf{x}^{(t)} \in \Sigma^n$  the next state is sampled by sampling a uniform random  $i \in [n]$  and then sampling the next state  $\mathbf{x}^{(t)} \sim \mu|_{\mathbf{x}_{\setminus i}^{(t)}}$ . That is, the chain resamples single coordinates at a time in a way that ensures that  $\mu$  is stationary.

**Definition 15.** A distribution  $\mu$  on  $\Sigma^n$  satisfies a  $\Phi$ -Sobolev inequality with constant  $c^*$  if for all  $f : \Sigma^n \rightarrow \mathbb{R}_{\geq 0}$ ,

$$\mathbf{Ent}_{\mu}^{\Phi}[f] \leq c^* \cdot \mathcal{E}_P(f, \Phi'(f)). \quad (3)$$

Here  $P$  is the Glauber dynamics chain associated to  $\mu$ .

The importance of this notion comes from the fact that the right-hand side of Equation (3) governs the decay of the  $\Phi$ -entropy between a Markov chain's current distribution and the stationary distribution through time.

**Remark 16.** Let  $\Phi(x) = x \log x$  and notice that  $\Phi' = x \mapsto 1 + \log x$ . Since  $\mathcal{E}_P$  is translation-invariant, we have  $\mathcal{E}_P(f, \Phi'(f)) = \mathcal{E}_P(f, \log f)$ , and therefore Equation (3) is the *modified log-Sobolev inequality*. Similarly, when  $\Phi(x) = x^2$  we have  $\mathcal{E}_P(f, \Phi'(f)) = 4\mathcal{E}(f, f)$ , and Equation (3) is (up to factor 4) the *Poincaré inequality*.

**Remark 17.** It is known that  $c^*$ -ATE implies  $\frac{2}{c^*n}$ -MLSI. See Proposition 1.1 in [CMT15].

The following Lemma 18 is essentially a  $\Phi$ -entropic generalization of Theorem 4.5 from [HMRW24]. See Appendix A for the proof.

**Lemma 18.** Let  $\mu = \sum_{a=1}^k \rho(a)\mu_a$  be a mixture of distributions on  $\Sigma^n$  with each  $\mu_a$  satisfying a  $\Phi$ -Sobolev inequality with constant  $c^*$ . Let  $P$  be the transition matrix for the Glauber dynamics for  $\mu$  and  $P_t$  be the associated continuous-time Markov operator. Then, for any initial distribution  $\pi$  we have

$$D_{\Phi}(P_t\pi \parallel \mu) \leq (1 - 1/c^*n)^t \cdot D_{\Phi}(\pi \parallel \mu) + \mathbf{E}_{\mathbf{s}} \left[ \mathbf{Ent}_{\mathbf{a} \sim \rho}^{\Phi} \left[ \mathbf{E}_{P_{\mathbf{s}}\pi} \left[ \frac{\mu_{\mathbf{a}}}{\mu} \right] \right] \right], \quad (4)$$

where  $\mathbf{s}$  is some random variable supported on  $[0, t]$ .

Lemma 18 shows that as long as one initializes Glauber dynamics for  $\mu$  at a distribution  $\pi$  so that the second term on the right-hand side in Equation (4) is small, the chain will experience fast mixing at the rate of the individual mixtures. The question is then how to choose  $\pi$  such that this quantity is indeed small. We show that when  $\pi$  is a “data-based initialization”, meaning the empirical distribution formed by some number of i.i.d. samples from  $\mu$ , the inter-component entropy is indeed small by concentration.

Recall [Fact 13](#), which motivates us to study the inter-component entropy as the  $\Phi$ -divergence between the empirical posterior distribution  $\rho_\pi = \frac{1}{m} \sum_{j=1}^m \rho_{\mathbf{x}_j}$  and the mixing weights. If the mixture components are separated, the task becomes to learn a distribution on  $[k]$  from samples up to  $\varepsilon$  error in  $\Phi$ -divergence. Furthermore, by exactly characterizing the m.g.f. of the empirical estimator's KL-divergence, we can use a convexity argument to handle the case when components are not separated. While we focus here on the well-studied case of KL-divergence, to establish a version of [Theorem 4](#) for any  $\Phi$ -divergence one generally only needs to establish the sample complexity of the learning task for a reasonable estimator (see [\[Can20\]](#) for more on such learning tasks).

### 3.1 The Case of KL-Divergence

Our main result, [Theorem 4](#), is to apply this paradigm in the case of KL-divergence. When all  $\rho_{\mathbf{x}_j}$  are indicator vectors (corresponding to the case of disjointly supported mixture components) the result then directly follows by taking  $m = \Theta(k/\varepsilon + \log(1/\delta)/\varepsilon)$  in the recent result of Agrawal given here:

**Theorem 19** ([\[Agr20\]](#), Theorem I.2). For  $\varepsilon > \frac{k-1}{m}$  we have that

$$\Pr_{\mathbf{a}_1, \dots, \mathbf{a}_m \sim \rho} \left[ D_{\text{KL}}(\text{avg}_{j \in [m]} \delta_{\mathbf{a}_j} \parallel \rho) > \varepsilon \right] \leq e^{-\varepsilon m} \cdot \left( \frac{e\varepsilon m}{k-1} \right)^{k-1}.$$

This result is established immediately by the following m.g.f. bound when  $\lambda = m - \frac{k-1}{\varepsilon}$ .

**Theorem 20** ([\[Agr20\]](#), Theorem I.3). For  $0 \leq \lambda < m$  we have that

$$\mathbf{E}_{\mathbf{a}_1, \dots, \mathbf{a}_m \sim \rho} \left[ \exp \left( \lambda \cdot D_{\text{KL}}(\text{avg}_{j \in [m]} \delta_{\mathbf{a}_j} \parallel \rho) \right) \right] \leq \left( \frac{1}{1 - \lambda/m} \right)^{k-1}.$$

We apply a simple convexity argument to arrive at the same m.g.f. bound for the general case of overlapping mixture components:

**Lemma 21.** Let  $\pi = \frac{1}{m} \sum_{j=1}^m \delta_{\mathbf{x}_j}$ , where each  $\mathbf{x}_j$  is sampled i.i.d. from  $\mu = \sum_{a=1}^k \rho(a) \mu_a$ . Then we have the following moment generating function bound:

$$\mathbf{E}_{\mathbf{x}_1, \dots, \mathbf{x}_m} \left[ \exp(\lambda \cdot D_{\text{KL}}(\rho_\pi \parallel \rho)) \right] \leq \left( \frac{1}{1 - \lambda/m} \right)^{k-1}.$$

*Proof.* For all  $\lambda > 0$  we have

$$\begin{aligned} \mathbf{E}_{\mathbf{x}_1, \dots, \mathbf{x}_m} \left[ \exp(\lambda \cdot D_{\text{KL}}(\rho_\pi \parallel \rho)) \right] &= \mathbf{E}_{\mathbf{x}_1, \dots, \mathbf{x}_m} \left[ \exp \left( \lambda \cdot D_{\text{KL}}(\text{avg}_{j \in [m]} \rho_{\mathbf{x}_j} \parallel \rho) \right) \right] \\ &= \mathbf{E}_{\mathbf{x}_1, \dots, \mathbf{x}_m} \left[ \exp \left( \lambda \cdot D_{\text{KL}}(\mathbf{E}_{\mathbf{a}_j \sim \rho_{\mathbf{x}_j}} \text{avg}_{j \in [m]} \delta_{\mathbf{a}_j} \parallel \rho) \right) \right] \\ &\leq \mathbf{E}_{\mathbf{x}_1, \dots, \mathbf{x}_m} \mathbf{E}_{\mathbf{a}_j \sim \rho_{\mathbf{x}_j}} \left[ \exp \left( \lambda \cdot D_{\text{KL}}(\text{avg}_{j \in [m]} \delta_{\mathbf{a}_j} \parallel \rho) \right) \right] \end{aligned}$$

$$= \mathbf{E}_{\mathbf{a}_1, \dots, \mathbf{a}_m \sim \rho} \left[ \exp \left( \lambda \cdot D_{\text{KL}} \left( \text{avg}_{j \in [m]} \delta_{\mathbf{a}_j} \parallel \rho \right) \right) \right].$$

The inequality follows from the convexity of the KL-divergence in the left component and the exponential function when  $\lambda > 0$ . Applying [Theorem 20](#) completes the proof.  $\square$

We can transfer this m.g.f. bound (and accordingly the tail bound) to that of a convex combination of these quantities over the  $s$ :

**Lemma 22.** Let  $\pi = \frac{1}{m} \sum_{j=1}^m \delta_{\mathbf{x}_j}$ , where the  $\mathbf{x}_j$  are sampled i.i.d. from  $\mu = \sum_{a=1}^k \rho(a) \mu_a$ . Then we have the following tail bound for any  $\mathbf{s}$  whenever  $\varepsilon > \frac{k-1}{m}$ :

$$\Pr_{\mathbf{x}_1, \dots, \mathbf{x}_m} \left[ \mathbf{E}_{\mathbf{s}} [D_{\text{KL}}(\rho_{P_s \pi} \parallel \rho) > \varepsilon] \right] \leq e^{-\varepsilon m} \cdot \left( \frac{e \varepsilon m}{k-1} \right)^{k-1}.$$

*Proof.* In general, let  $\mathbf{X} = \mathbf{E}_{\mathbf{s}} [\mathbf{X}_s]$  where each  $\mathbf{X}_i$  is distributed identically. Fixing  $\lambda \geq 0$ , we have:

$$\mathbf{E} [e^{\lambda \mathbf{X}}] = \mathbf{E} [e^{\lambda \mathbf{E}_{\mathbf{s}} [\mathbf{X}_s]}] \leq \mathbf{E} \left[ \mathbf{E}_{\mathbf{s}} [e^{\lambda \mathbf{X}_s}] \right] = \mathbf{E}_{\mathbf{s}} \left[ \mathbf{E} [e^{\lambda \mathbf{X}_s}] \right] = \mathbf{E} [e^{\lambda \mathbf{X}_0}].$$

The inequality follows via convexity, and  $\mathbf{X}_0$  is a copy of  $\mathbf{X}_s$ .

Set  $\mathbf{X}_s = D_{\text{KL}}(\rho_{P_s \pi} \parallel \rho) = D_{\text{KL}}(\text{avg}_{j \in [m]} \rho_{P_s \delta_{\mathbf{x}_j}} \parallel \rho)$  for all  $s$ . Note that each  $\mathbf{X}_s$  is marginally distributed according to the distribution given by  $D_{\text{KL}}(\text{avg}_{j \in [m]} \rho_{\delta_{\mathbf{x}_j}} \parallel \rho)$  since each  $\mathbf{x}_j \sim \mu$ , and  $\mu$  is stationary with respect to  $P$ . Whenever  $0 \leq \lambda < m$  we apply the above and get the bound

$$\mathbf{E} [e^{\lambda \mathbf{E}_{\mathbf{s}} [D_{\text{KL}}(\rho_{P_s \pi} \parallel \rho)]] \leq \mathbf{E} [e^{\lambda D_{\text{KL}}(\rho_{\pi} \parallel \rho)}] \stackrel{\text{Lemma 21}}{\leq} \left( \frac{1}{1 - \lambda/m} \right)^{k-1}.$$

Taking  $\lambda = m - \frac{k-1}{\varepsilon}$  as in [Theorem 19](#) finishes the proof.  $\square$

Combining [Lemma 18](#) and [Lemma 22](#) immediately implies this formal version of [Theorem 4](#):

**Theorem 23.** Let  $\mu = \sum_{a=1}^k \rho(a) \mu_a$  be a mixture distribution with each component satisfying a modified log-Sobolev inequality with constant  $c^*$ . Let  $P_t$  be the continuous-time Markov operator induced by the Glauber dynamics for  $\mu$  and  $\pi = \frac{1}{m} \sum_{j=1}^m \delta_{\mathbf{x}_j}$  for  $\mathbf{x}_1, \dots, \mathbf{x}_m \sim \mu$ . Then

$$\Pr_{\mathbf{x}_1, \dots, \mathbf{x}_m} [D_{\text{KL}}(P_t \pi \parallel \mu) > \varepsilon] \leq \delta,$$

for  $m = \Theta(k/\varepsilon + \log(1/\delta)/\varepsilon)$  and  $t = c^* n \cdot \Theta(\log \log(1/\min_x \mu(x)) + \log(1/\varepsilon))$ .

*Proof.* By [Lemma 18](#), the fact that  $D_{\text{KL}}(\pi \parallel \mu) \leq \log(1/\min_x \mu(x))$ , and the setting of  $t$  it suffices to show that for the claimed sample complexity  $m$  we have

$$\Pr_{\mathbf{x}_1, \dots, \mathbf{x}_m} \left[ \mathbf{E}_{\mathbf{s}} [D_{\text{KL}}(\rho_{P_s \pi} \parallel \rho)] > 0.5\varepsilon \right] \leq \delta.$$

The use of the correct constant in the setting of  $m$  and [Lemma 21](#) imply the desired bound.  $\square$

## 4 Identity-Testing with Coordinate Conditional Sampling

Our second application deals with the problem of identity-testing distributions on  $\Sigma^n$ , but with access to coordinate conditional samples from the target distribution  $\mu$ , which is the mixture of some number of distributions sampling ATE. We prove [Theorem 5](#).

**Proof Overview.** As in [Section 3](#), if  $\mu$  itself satisfies ATE then our result follows, and all that needs to be done is to handle the case where inter-component entropy exists. Note that [Algorithm 1](#) is more or less the same as the algorithm of [\[BCSV23\]](#), except for [Step 2](#) (and the use of an improved KL identity-tester). To motivate the design of our algorithm, consider a distribution  $\pi$  that is far from  $\mu$  in KL-divergence. Then, by the chain rule ([Lemma 11](#)), we have that either  $\mathbf{E}_{a \sim \rho} \mathbf{Ent}_{\mu_a} \left[ \frac{\pi}{\mu} \right]$  is large, or  $\mathbf{Ent}_{a \sim \rho} \mathbf{E}_{\mu_a} \left[ \frac{\pi}{\mu} \right]$  is large. Intuitively, either the intra-component entropy is large, or the inter-component entropy is large.

In the first case, the algorithm of [\[BCSV23\]](#) rejects  $\pi$  with high probability. This does not directly follow from the guarantee of [\[BCSV23\]](#), but another application of the chain rule allows us to deduce this. See [Section 4.1.1](#) for the analysis of this part of the algorithm.

In the second case, the inter-component entropy is large. However, it may be the case that the intra-component entropy is small. For example,  $\pi$  could be a mixture of the  $\mu_a$ , but with incorrect mixture weights. In this case, it is not clear how to use the coordinate oracle to detect this discrepancy in a generic way, especially if there is some intra-component entropy. However, in this case, we note that we can use posterior sampling of  $a \sim \rho|x$ , where  $x$  are samples from  $\pi$ , to infer what the effective weights on each component are. Here  $\rho|x$  is the distribution on  $[k]$  where  $(\rho|x)(a) = \frac{\rho(a)\mu_a(x)}{\mu(x)}$ .

More formally, the inter-component entropy is equal to the KL-divergence between this posterior distribution when  $x \sim \pi$  and when  $x \sim \mu$ . Given knowledge of the true distribution  $\rho$ , we can again use our base KL-divergence tester from [Theorem 24](#) to reject this  $\pi$ . This is the content of [Step 1](#) in [Algorithm 1](#), analyzed in [Section 4.1.2](#).

As a subroutine, we will need an identity-testing algorithm with respect to KL-divergence error  $\varepsilon$ , for  $\eta$ -balanced distributions on domains of size  $d$  (we will use this algorithm for the cases  $d = k$  and  $d = |\Sigma|$ ). In [\[BCSV23\]](#), two such testers were given and analyzed; a main one with sample complexity  $O\left(\min\left\{\frac{\sqrt{d}\ln(1/\eta)}{\varepsilon^2}, \frac{1}{\sqrt{\eta}\varepsilon}\right\}\right)$ ,<sup>2</sup> and one for the special case of  $k = 2$  with sample complexity  $O\left(\frac{\ln(1/\eta)}{\varepsilon}\right)$ . The main effort there was to get  $\log(1/\eta)$  dependence, rather than  $1/\eta$  dependence; however, the general case suffers from a quadratic dependence on  $1/\varepsilon$ . Here we note a tester with both linear dependence on  $1/\varepsilon$  and logarithmic dependence on  $1/\eta$  can be recovered by combining two known results from the literature; we prove this in [Appendix A](#):

**Theorem 24.** Let  $q$  be a distribution over a universe  $D$  of size  $d$ , with each outcome in  $D$  having probability at least  $\eta$  under  $q$ . There is an algorithm  $\text{KL-TEST}(p, q, \varepsilon, \delta)$  that given input  $\varepsilon, \delta > 0$  and access to samples from an unknown distribution  $p$  on  $D$ , draws

$$O\left(\frac{\sqrt{d} \cdot \log(1/\eta) \cdot \log(1/\delta)}{\varepsilon}\right)$$

samples from  $p$  and has the following performance guarantee:

1. If  $p = q$  then  $\text{KL-TEST}(p, q, \varepsilon, \delta)$  accepts with probability at least  $1 - \delta$ .
2. If  $D_{\text{KL}}(p \parallel q) \geq \varepsilon$  then  $\text{KL-TEST}(p, q, \varepsilon)$  accepts with probability at most  $\delta$ .

---

<sup>2</sup>We remark that the first expression in the min essentially always dominates. Even when its  $\varepsilon$  dependence is made linear, as in our [Theorem 24](#), the second expression is smaller only in the regime  $\frac{1}{d \log^2 d} \lesssim \eta \leq \frac{1}{d}$ , and then only by a  $\log d$  factor.

**Remark 25.** The [BCSV23] work shows that in [Theorem 24](#), we can actually take the minimum of the sample complexity with  $O(\frac{1}{\sqrt{\eta\varepsilon}})$ . For simplicity of exposition, we will not carry around this ‘min’ in our subsequent complexity bounds (for the reasons described in [Footnote 2](#)).

With this tester in hand, we may state our algorithm.

**Input:** Coordinate Oracle and General Oracle access (See [Remark 6](#)) to distribution  $\pi$  on  $\Sigma^n$ , known mixing weights  $\rho(1), \dots, \rho(k)$ , descriptions of  $c^*$ -ATE distributions  $\mu_1, \dots, \mu_k$  on  $\Sigma^n$ , and the assumption that distribution  $\mu = \sum_{a=1}^k \rho(a)\mu_a$  is  $\eta$ -balanced.

**Output:** “Accept” or “Reject”.

PRODUCT-SET-KL-TEST( $\pi, \mu, \varepsilon$ ):

1. Independently draw  $T_1 = O\left(\frac{c^*n}{\varepsilon}\right)$  pairs  $(\mathbf{x}, \mathbf{i})$ , where  $\mathbf{x} \sim \pi$  and  $\mathbf{i} \in [n]$  is uniformly random. For each pair  $(\mathbf{x}, \mathbf{i})$ , reject if  $\text{KL-TEST}\left(\pi|_{\mathbf{x}_{\setminus \mathbf{i}}}, \mu|_{\mathbf{x}_{\setminus \mathbf{i}}}, \boldsymbol{\theta}, 0.05 \cdot \frac{1}{T_1}\right)$  rejects, where  $\boldsymbol{\theta} \sim \text{Unif}\left([0.05\varepsilon/c^*, \log(1/\eta)]\right)$ . If the total number of Coordinate Oracle calls needed to perform these calls to KL-TEST exceeds

$$T = 100 \cdot T_1 \cdot \underbrace{c_{\text{KL-TEST}} \cdot \sqrt{|\Sigma|} \cdot \log(1/\eta) \cdot \log(20T_1) \cdot 10 \log\left(\frac{\log(1/\eta)}{\varepsilon/c^*}\right)}_{\text{bounds expected number of Coordinate Oracle uses per call to KL-TEST}},$$

then reject. Here  $c_{\text{KL-TEST}}$  is the constant hidden in the  $O$ -notation of [Theorem 24](#).

2. Consider the distribution  $\rho_\pi$  on  $[k]$  defined by drawing  $\mathbf{x} \sim \pi$  and outputting  $a \in [k]$  with probability  $\frac{\rho(a)\mu_a(\mathbf{x})}{\mu(\mathbf{x})}$ . (Note the algorithm can simulate draws from  $\rho_\pi$  using the General Oracle for  $\pi$ .) Reject if  $\text{KL-TEST}(\rho_\pi, \rho, 0.5\varepsilon, 0.1)$  rejects.
3. Accept.

**Algorithm 1:** Identity-testing of  $\mu$  with Coordinate Oracle and General Oracle access.

We now prove that [Algorithm 1](#) is a good identity-tester for the distribution  $\mu$ , deferring proofs of the auxiliary lemmas [Lemma 30](#) and [Lemma 31](#) to [Section 4.1](#).

**Theorem 26** ([Theorem 5](#), restated). Let  $\mu = \sum_{a=1}^k \rho(a)\mu_a$  be  $\eta$ -balanced, where  $\rho^* = \min_a \rho(a)$  and each  $\mu_a$  satisfies  $c^*$ -ATE. Then [Algorithm 1](#) uses

$$\underbrace{O\left(\frac{c^*n}{\varepsilon} \cdot \log^2\left(\frac{c^*n}{\varepsilon}\right) \cdot \sqrt{|\Sigma|} \cdot \log(1/\eta) \cdot \log \log(1/\eta)\right)}_{\text{from Step 1}} + \underbrace{O\left(\frac{\sqrt{k} \cdot \log(1/\rho^*)}{\varepsilon}\right)}_{\text{from Step 2}}$$

calls to the General Oracle and Coordinate Oracle and satisfies the following performance guarantees:

1. If  $\pi = \mu$  then COORDINATE-ORACLE-TEST( $\mu, \pi, \varepsilon$ ) rejects with probability at most 0.4.
2. If  $D_{\text{KL}}(\pi \parallel \mu) \geq \varepsilon$  then COORDINATE-ORACLE-TEST( $\mu, \pi, \varepsilon$ ) rejects with probability at least 0.6.

*Proof.* Recalling that  $T_1 = O\left(\frac{c^*n}{\varepsilon}\right)$ , **Step 1** contributes at most the following amount of calls to the oracles:

$$\begin{aligned} T &= 1000 \cdot c_{\text{KL-TEST}} \cdot T_1 \cdot \sqrt{|\Sigma|} \cdot \log(1/\eta) \cdot \log(20T_1) \cdot \log\left(\frac{c^* \log(1/\eta)}{\varepsilon}\right). \\ &= O\left(\frac{c^*n}{\varepsilon}\right) \cdot \log\left(\frac{c^*n}{\varepsilon}\right) \cdot \sqrt{|\Sigma|} \cdot \log(1/\eta) \cdot \log\left(\frac{c^* \log(1/\eta)}{\varepsilon}\right) \\ &= O\left(\frac{c^*n}{\varepsilon} \cdot \log^2\left(\frac{c^*n}{\varepsilon}\right) \cdot \sqrt{|\Sigma|} \cdot \log(1/\eta) \cdot \log \log(1/\eta)\right). \end{aligned}$$

**Step 2** in **Algorithm 1** uses the following amount of oracle calls:

$$O\left(\frac{\sqrt{k} \cdot \log(1/\rho^*)}{\varepsilon}\right).$$

Summing these gives the final number of oracle calls.

Now we bound the failure probabilities. Suppose that  $\pi = \mu$  so that  $D_{\text{KL}}(\pi \parallel \mu) = 0$ . Then the probability that **Step 1** in **Algorithm 1** rejects is at most 0.1 by **Lemma 30**. The probability that **Step 2** in **Algorithm 1** rejects is at most 0.1 by **Lemma 31**. Therefore,  $\text{KL-TEST}(\mu, \pi, \varepsilon)$  accepts with probability at least 0.8.

Now suppose that  $D_{\text{KL}}(\pi \parallel \mu) = \mathbf{Ent}_{\mu} \left[ \frac{\pi}{\mu} \right] \geq \varepsilon$ . Using the chain rule (**Lemma 11**), we have

$$\mathbf{Ent}_{\mu} \left[ \frac{\pi}{\mu} \right] = \mathbf{E}_{a \sim \rho} \left[ \mathbf{Ent}_{\mu_a} \left[ \frac{\pi}{\mu} \right] \right] + \mathbf{Ent}_{a \sim \rho} \left[ \mathbf{E}_{\mu_a} \left[ \frac{\pi}{\mu} \right] \right],$$

and we find that one of the two summands on the right-hand side is at least  $0.5\varepsilon$ . In the first case, we have  $\mathbf{E}_{a \sim \rho} \left[ \mathbf{Ent}_{\mu_a} \left[ \frac{\pi}{\mu} \right] \right] \geq 0.5\varepsilon$  and then **Lemma 30** shows that **Step 1** rejects with probability at least 0.9. Otherwise if  $\mathbf{Ent}_{a \sim \rho} \left[ \mathbf{E}_{\mu_a} \left[ \frac{\pi}{\mu} \right] \right] \geq 0.5\varepsilon$  then **Lemma 31** shows that **Step 2** rejects with probability at least 0.9.  $\square$

#### 4.1 Proofs of **Lemmas 30** and **31**

In this section we prove **Lemmas 30** and **31** which were the main tools needed in our application to identity-testing.

As in [BCSV23], our sample complexity has a dependence on the balancedness of our visible distributions:

**Definition 27.** We say that a distribution  $\mu$  on  $\Sigma^n$  is  $\eta$ -balanced if for all  $x \in \Sigma^n$  we have that the distribution  $\mu|_{x_{\setminus i}}$  has minimum probability  $\eta$ .

We always regard the distribution  $\mu$  on  $\Sigma^n$  as a mixture of distributions  $\mu = \sum_{a=1}^k \rho(a) \mu_a$ , where each  $\mu_a$  satisfies approximate tensorization of entropy with constant  $c_a$ . Moreover, we assume that  $\mu$  is  $\eta$ -balanced. The following **Fact 28** also shows that  $\mu$  is fully supported whenever  $\eta > 0$ :

**Fact 28.** Let  $\mu$  be an  $\eta$ -balanced distribution on  $\Sigma^n$ . Then for all  $x \in \Sigma^n$ , we have  $\mu(x) \geq \eta^n$ .

*Proof.* We induct on  $n$ . In the base case  $n = 1$  and the conclusion is immediate.

For the inductive step let  $\mu$  be an  $\eta$ -balanced distribution on  $\Sigma^n$ . For each  $b \in \Sigma$ , let  $S_b = \{x \in \Sigma^n : x_n = b\}$ . It must be the case that for all  $b \in \Sigma$ , we have  $\mu(S_b) \geq \eta$ , since otherwise by averaging there would be  $y \in \Sigma^{n-1}$  such that  $\mu|_{y \setminus \Sigma}$  is not  $\eta$ -balanced. By the inductive hypothesis, since the distribution  $\mu|_{S_b}$  is  $\eta$ -balanced, for any  $x \in S_b$  we have  $\mu|_{S_b}(x) \geq \eta^{n-1}$ . Then  $\mu(x) = \mu(S_b)\mu|_{S_b}(x) \geq \eta \cdot \eta^{n-1} = \eta^n$ .  $\square$

Throughout this section let  $0 < \varepsilon < n \log\left(\frac{1}{\eta}\right)$ , which is without loss of generality since if  $\mu$  is  $\eta$ -balanced then by [Fact 28](#) we have  $\min_{x \in \Sigma^n} \mu(x) \geq \eta^n$ . For any other distribution  $\pi$ , we have

$$D_{\text{KL}}(\pi \parallel \mu) = \mathbf{E}_{\pi} \left[ \log \left( \frac{\pi}{\mu} \right) \right] \leq \max_x \left( \log(\pi(x)) \mu(x) \right) \leq n \log \left( \frac{1}{\eta} \right).$$

Let  $\pi$  be an arbitrary distribution on  $\Sigma^n$ , and recall that  $\mu = \sum_{a=1}^k \rho(a) \mu_a$  is a mixture of  $k$  distributions  $\mu_1, \dots, \mu_k$ . Let  $\rho^* = \min_a \rho(a)$ .

#### 4.1.1 Rejection by Local Testing

**Lemma 29.** The probability that the number of calls to the Coordinate Oracle and General Oracle needed exceeds the limit stated in [Step 1](#) is at most 0.01.

*Proof.* With certainty over the choice of  $\mathbf{x}$  and  $\mathbf{i}$ , the distribution  $\mu|_{\mathbf{x}_{\setminus \mathbf{i}}}$  has minimum probability at least  $\eta$  by definition of  $\eta$ -balancedness. Then a call to  $\text{KL-TEST}(\pi|_{\mathbf{x}_{\setminus \mathbf{i}}}, \mu|_{\mathbf{x}_{\setminus \mathbf{i}}}, \boldsymbol{\theta}, 0.05 \cdot \frac{1}{T_1})$ , by [Theorem 24](#) requires

$$c_{\text{KL-TEST}} \cdot \frac{\sqrt{|\Sigma|} \cdot \log(1/\eta) \cdot \log(20T_1)}{\boldsymbol{\theta}}$$

calls to the Coordinate Oracle. To compute the total expected number of samples, we first compute

$$\begin{aligned} \mathbf{E}_{\boldsymbol{\theta}} \left[ \frac{1}{\boldsymbol{\theta}} \right] &= \frac{1}{\log(1/\eta) - 0.05\varepsilon/c^*} \int_{\frac{0.05\varepsilon}{c^*n}}^{\log(1/\eta)} \frac{1}{\theta} d\theta \\ &= \frac{1}{\log(1/\eta) - 0.05\varepsilon/c^*} \ln \left( \frac{c^* \log(1/\eta)}{0.05\varepsilon} \right) \\ &\leq \frac{10}{\log(1/\eta)} \log \left( \frac{c^* \log(1/\eta)}{\varepsilon} \right). \end{aligned}$$

Here we use that  $c^* \geq 1$ , and that  $\varepsilon \leq n \log\left(\frac{1}{\eta}\right)$  so that

$$\frac{0.05c^*\varepsilon}{n} \leq \frac{0.05n \log\left(\frac{1}{\eta}\right)}{n} = 0.05 \log\left(\frac{1}{\eta}\right).$$

Then, using that  $\eta \leq \frac{1}{2}$ , we bound this by

$$\mathbf{E}_{\boldsymbol{\theta}} \left[ \frac{1}{\boldsymbol{\theta}} \right] \leq 10 \log \left( \frac{\log(1/\eta)}{c^*\varepsilon} \right).$$

Therefore, the expected number of samples, over all  $T_1$  calls to  $\text{KL-TEST}$  is at most

$$T_1 \cdot 10 \cdot c_{\text{KL-TEST}} \cdot \sqrt{|\Sigma|} \cdot \log(1/\eta) \cdot \log(20T_1) \cdot \log \left( \frac{\log(1/\eta)}{c^*\varepsilon} \right).$$

Markov's inequality shows the result.  $\square$

**Lemma 30.** Assume that  $\mu$  is  $\eta$ -balanced. Then Algorithm 1 satisfies the following performance guarantees:

1. If  $\pi = \mu$  then Step 1 in  $\text{KL-TEST}(\mu, \pi, \varepsilon)$  rejects with probability at most 0.1.
2. If  $\mathbf{E}_{\mathbf{a} \sim \rho} \left[ \mathbf{Ent}_{\mathbf{x} \sim \mu_{\mathbf{a}}} \left[ \frac{\pi(\mathbf{x})}{\mu(\mathbf{x})} \right] \right] \geq 0.5\varepsilon$  then Step 1 in  $\text{KL-TEST}(\mu, \pi, \varepsilon)$  rejects with probability at least 0.9.

*Proof.* Define the random variable  $\mathbf{Y}$  to be  $\mathbf{Ent}_{\mathbf{z} \sim \mu|_{\mathbf{x}_{\setminus i}}} \left[ \frac{\pi|_{\mathbf{x}_{\setminus i}}(\mathbf{z})}{\mu|_{\mathbf{x}_{\setminus i}}(\mathbf{z})} \right]$  for  $\mathbf{x} \sim \pi$  and  $i \in [n]$  uniform.

Consider the version of Step 1 without the sample limit. If  $\pi = \mu$  then by the guarantee of Theorem 24 shows that the rejection probability in each of the  $T_1$  iterations using this unlimited tester in Step 1 is at most  $0.05 \cdot \frac{1}{T_1}$ . Since by Lemma 29 the limit on the number of Coordinate Oracle calls changes the behavior with probability at most 0.01, the overall rejection probability is by a union bound at most  $T_1 \cdot \frac{0.05}{T_1} + 0.01 \leq 0.1$ .

Now suppose that  $\mathbf{E}_{\mathbf{a} \sim \rho} \left[ \mathbf{Ent}_{\mathbf{x} \sim \mu_{\mathbf{a}}} \left[ \frac{\pi(\mathbf{x})}{\mu(\mathbf{x})} \right] \right] \geq 0.5\varepsilon$ . Then by Lemma 32 and Lemma 12 we have

$$\mathbf{E}[\mathbf{Y}] = \frac{1}{n} \sum_{i \in [n]} \mathbf{E}_{\mathbf{x} \sim \pi} \left[ \mathbf{Ent}_{\mathbf{z} \sim \mu|_{\mathbf{x}_{\setminus i}}} \left[ \frac{\pi|_{\mathbf{x}_{\setminus i}}(\mathbf{z})}{\mu|_{\mathbf{x}_{\setminus i}}(\mathbf{z})} \right] \right] = \frac{1}{n} \mathcal{L}_{\mu} \left[ \frac{\pi}{\mu} \right] \geq \frac{0.5\varepsilon}{c^*n},$$

where  $c^* = \max_a c_a$  is a constant such that all  $\mu_a$  satisfy  $c^*$ -ATE.

Using that  $\mathbf{Y} \leq \log(1/\eta)$  with certainty,

$$\begin{aligned} \Pr_{\theta, \mathbf{Y}}[\mathbf{Y} \geq \theta] &= \frac{1}{\log(1/\eta) - \frac{0.05\varepsilon}{c^*n}} \int_0^{\log(1/\eta)} \Pr_{\mathbf{Y}}[\mathbf{Y} \geq \theta] d\theta \\ &\geq \frac{1}{\log(1/\eta)} \int_0^{\log(1/\eta)} \Pr_{\mathbf{Y}}[\mathbf{Y} \geq \theta] d\theta - \frac{1}{\log(1/\eta)} \int_0^{\frac{0.05\varepsilon/c^*}{n}} \Pr_{\mathbf{Y}}[\mathbf{Y} \geq \theta] d\theta \\ &= \mathbf{E}[\mathbf{Y}] - \frac{1}{\log(1/\eta)} \int_0^{0.05\varepsilon/c^*} \Pr_{\mathbf{Y}}[\mathbf{Y} \geq \theta] d\theta \\ &\geq \mathbf{E}[\mathbf{Y}] - \frac{0.05\varepsilon}{c^*n} \\ &\geq \frac{0.9}{n} \cdot \mathbf{E}[\mathbf{Y}] \\ &\geq \frac{0.45\varepsilon}{c^*n}. \end{aligned}$$

Therefore, for  $T_1$  draws of  $\mathbf{x} \sim \pi$ ,  $i \sim [n]$  uniform, and  $\theta$ , the probability that none satisfy  $\theta \leq \mathbf{Ent}_{\mathbf{z} \sim \mu|_{\mathbf{x}_{\setminus i}}} \left[ \frac{\pi|_{\mathbf{x}_{\setminus i}}(\mathbf{z})}{\mu|_{\mathbf{x}_{\setminus i}}(\mathbf{z})} \right]$  is at most

$$\left( 1 - \frac{0.45\varepsilon}{c^*n} \right)^{T_1} \leq 0.04$$

for a correct choice of constant in the definition of  $T_1$ .

Let  $\mathcal{A}$  be the event that this occurs so that  $\Pr[\mathcal{A}] \leq 0.01$ . Then let  $\mathcal{B}$  be the event that for some call  $\text{KL-TEST}(\pi|_{\mathbf{x}_{\setminus i}}, \mu|_{\mathbf{x}_{\setminus i}}, \theta, 0.05 \cdot \frac{1}{T_1})$  for which  $\theta \leq \mathbf{Ent}_{\mathbf{z} \sim \mu|_{\mathbf{x}_{\setminus i}}} \left[ \frac{\pi|_{\mathbf{x}_{\setminus i}}(\mathbf{z})}{\mu|_{\mathbf{x}_{\setminus i}}(\mathbf{z})} \right]$ , the test accepted.

The probability that a single call of this form failed is at most  $0.05 \cdot \frac{1}{T_1}$ , so a union bound gives  $\Pr[\mathcal{B}] \leq 0.05$ .

By a union bound, the probability of accepting  $\pi$  is then

$$\Pr[\mathcal{A}] + \Pr[\mathcal{B}] \leq 0.1.$$

Thus, the performance guarantee is satisfied.  $\square$

#### 4.1.2 Rejection by Posterior Weight Estimation

**Lemma 31.** *Algorithm 1* satisfies the following performance guarantees:

1. If  $\pi = \mu$  then *Step 2* in  $\text{KL-TEST}(\mu, \pi, \varepsilon)$  rejects with probability at most 0.1.
2. If  $\mathbf{Ent}_{\mathbf{a} \sim \rho} \left[ \mathbf{E}_{x \sim \mu_{\mathbf{a}}} \left[ \frac{\pi(x)}{\mu(x)} \right] \right] \geq 0.5\varepsilon$  then *Step 2* in  $\text{KL-TEST}(\mu, \pi, \varepsilon)$  rejects with probability at least 0.9.

*Proof.* By *Fact 13* we have that

$$\mathbf{Ent}_{\mathbf{a} \sim \rho} \left[ \mathbf{E}_{x \sim \mu_{\mathbf{a}}} \left[ \frac{\pi(x)}{\mu(x)} \right] \right] = \mathbf{Ent}_{\mathbf{a} \sim \rho} \left[ \frac{\rho_{\pi}(\mathbf{a})}{\rho(\mathbf{a})} \right] = D_{\text{KL}}(\rho_{\pi} \parallel \rho).$$

Then the guarantee of *Theorem 24* shows that if  $\pi = \mu$  the step *Step 2* rejects with probability at most 0.1. Otherwise, if  $D_{\text{KL}}(\rho_{\pi}, \rho) \geq 0.5\varepsilon$  then *Step 2* in *Algorithm 1* rejects with probability at least 0.9.  $\square$

**Lemma 32.** The following equality holds for all distribution  $\pi$  and  $\mu$  on  $\Sigma^n$ :

$$\mathcal{L}_{\pi} \left[ \frac{\pi}{\mu} \right] = \sum_{i \in [n]} \mathbf{E}_{x \sim \pi} \left[ \mathbf{Ent}_{z \sim \mu|_{x_{\setminus i}}} \left[ \frac{\pi|_{x_{\setminus i}}(z)}{\mu|_{x_{\setminus i}}(z)} \right] \right].$$

*Proof.* We directly compute

$$\begin{aligned} \sum_{i \in [n]} \mathbf{E}_{x \sim \pi} \left[ \mathbf{Ent}_{z \sim \mu|_{x_{\setminus i}}} \left[ \frac{\pi|_{x_{\setminus i}}(z)}{\mu|_{x_{\setminus i}}(z)} \right] \right] &= \frac{1}{|\Sigma|} \sum_{i \in [n]} \sum_x \pi(x_{\setminus i}) \mathbf{Ent}_{z \sim \mu|_{x_{\setminus i}}} \left[ \frac{\pi|_{x_{\setminus i}}(z)}{\mu|_{x_{\setminus i}}(z)} \right] \\ &= \frac{1}{|\Sigma|} \sum_{i \in [n]} \sum_x \pi(x_{\setminus i}) \cdot \frac{\mu(x_{\setminus i})}{\pi(x_{\setminus i})} \cdot \mathbf{Ent}_{y \sim \mu|_{x_{\setminus i}}} \left[ \frac{\pi(y)}{\mu(y)} \right] \\ &= \frac{1}{|\Sigma|} \sum_{i \in [n]} \sum_x \mu(x_{\setminus i}) \cdot \mathbf{Ent}_{y \sim \mu|_{x_{\setminus i}}} \left[ \frac{\pi(y)}{\mu(y)} \right] \\ &= \mathcal{L}_{\pi} \left[ \frac{\pi}{\mu} \right]. \end{aligned}$$

The second equality follows by 1-homogeneity of  $\mathbf{Ent}[\cdot]$  (*Fact 10*).  $\square$

**Remark 33.** *Lemma 32* is the only place we require the  $\Phi$ -entropy we use to be  $\Phi(u) = u \log u$ , since this is the only place we need 1-homogeneity. We leave it open for future work whether one can obtain testers for different  $\Phi$ -entropies (i.e., other divergences besides KL-divergence) by bypassing the need for *Lemma 32*.

## References

- [ADK15] Jayadev Acharya, Constantinos Daskalakis, and Gautam Kamath. Optimal testing for properties of distributions. *Advances in Neural Information Processing Systems*, 28, 2015.
- [Agr20] Rohit Agrawal. Finite-sample concentration of the multinomial in relative entropy. *IEEE Transactions on Information Theory*, 66(10):6297–6302, 2020.
- [AJK<sup>+</sup>21] Nima Anari, Vishesh Jain, Frederic Koehler, Huy Tuan Pham, and Thuy-Duong Vuong. Entropic independence I: Modified log-Sobolev inequalities for fractionally log-concave distributions and high-temperature Ising models. *arXiv preprint arXiv:2106.04105*, 2021.
- [BC18] Rishiraj Bhattacharyya and Sourav Chakraborty. Property testing of joint distributions using conditional samples. *ACM Transactions on Computation Theory (TOCT)*, 10(4):1–20, 2018.
- [BCC<sup>+</sup>22] Antonio Blanca, Pietro Caputo, Zongchen Chen, Daniel Parisi, Daniel vStefankovic, and Eric Vigoda. On mixing of Markov chains: Coupling, spectral independence, and entropy factorization. In *Proceedings of the 2022 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 3670–3692. SIAM, 2022.
- [BCSV23] Antonio Blanca, Zongchen Chen, Daniel Stefankovic, and Eric Vigoda. Complexity of high-dimensional identity testing with coordinate conditional sampling. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 1774–1790. PMLR, 2023.
- [BG18] Salman Beigi and Amin Gohari.  $\Phi$ -Entropic Measures of Correlation. *IEEE Transactions on Information Theory*, 64(4):2193–2211, 2018.
- [BM98] Lucien Birgé and Pascal Massart. Minimum contrast estimators on sieves: exponential bounds and rates of convergence. *Bernoulli*, 4(3):329–375, 1998.
- [BOW19] Costin Bădescu, Ryan O’Donnell, and John Wright. Quantum state certification. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pages 503–514, 2019.
- [BT06] Sergey G Bobkov and Prasad Tetali. Modified logarithmic Sobolev inequalities in discrete settings. *Journal of Theoretical Probability*, 19:289–336, 2006.
- [Can20] Clément L. Canonne. A short note on learning discrete distributions, 2020.
- [CCK<sup>+</sup>21] Clément L. Canonne, Xi Chen, Gautam Kamath, Amit Levi, and Erik Waingarten. Random restrictions of high dimensional distributions and uniformity testing with sub-cube conditioning. In *Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 321–336. SIAM, 2021.
- [CE22] Yuansi Chen and Ronen Eldan. Localization schemes: A framework for proving mixing bounds for Markov chains. In *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 110–122. IEEE, 2022.

- [CFG13] Sourav Chakraborty, Eldar Fischer, Yonatan Goldhirsh, and Arie Matsliah. On the power of conditional samples in distribution testing. In *Proceedings of the 4th conference on Innovations in Theoretical Computer Science*, pages 561–580, 2013.
- [CGG<sup>+</sup>24] Zongchen Chen, Andreas Galanis, Leslie Ann Goldberg, Heng Guo, Andrés Herrera-Poyatos, Nitya Mani, and Ankur Moitra. Fast sampling of satisfying assignments from random-SAT with applications to connectivity. *SIAM Journal on Discrete Mathematics*, 38(4):2750–2811, 2024.
- [Cha04] Djalil Chafaï. Entropies, Convexity, and Functional Inequalities, On  $\Phi$ -Entropies and  $\Phi$ -Sobolev Inequalities. *Journal of Mathematics of Kyoto University*, 44(2):325–363, 2004.
- [CJLW21] Xi Chen, Rajesh Jayaram, Amit Levi, and Erik Waingarten. Learning and testing junta distributions with sub cube conditioning. In *Conference on Learning Theory*, pages 1060–1113. PMLR, 2021.
- [CKOS15] Cafer Caferov, Barış Kaya, Ryan O’Donnell, and A.C. Cem Say. Optimal bounds for estimating entropy with pmf queries. In *International Symposium on Mathematical Foundations of Computer Science*, pages 187–198. Springer, 2015.
- [CLV21] Zongchen Chen, Kuikui Liu, and Eric Vigoda. Optimal mixing of Glauber dynamics: Entropy factorization via high-dimensional expansion. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 1537–1550, 2021.
- [CMM23] Zongchen Chen, Nitya Mani, and Ankur Moitra. From algorithms to connectivity and back: finding a giant component in random k-SAT. In *Proceedings of the 2023 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 3437–3470. SIAM, 2023.
- [CMT15] Pietro Caputo, Georg Menz, and Prasad Tetali. Approximate tensorization of entropy at high temperature. In *Annales de la Faculté des sciences de Toulouse: Mathématiques*, volume 24, pages 691–716, 2015.
- [CR14] Clément Canonne and Ronitt Rubinfeld. Testing probability distributions underlying aggregated data. In *International Colloquium on Automata, Languages, and Programming*, pages 283–295. Springer, 2014.
- [CRS14] Clément Canonne, Dana Ron, and Rocco A. Servedio. Testing equivalence between distributions using conditional samples. In *Proceedings of the twenty-fifth annual ACM-SIAM symposium on Discrete algorithms*, pages 1174–1192. SIAM, 2014.
- [DKW18] Constantinos Daskalakis, Gautam Kamath, and John Wright. Which distribution distances are sublinearly testable? In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2747–2764. SIAM, 2018.
- [Fis66] Ronald Aylmer Fisher. *The design of experiments*, volume 21. Springer, 1966.
- [FJO<sup>+</sup>15] Moein Falahatgar, Ashkan Jafarpour, Alon Orlitsky, Venkatadheeraj Pichapati, and Ananda Theertha Suresh. Faster algorithms for testing under conditional sampling. In *Conference on Learning Theory*, pages 607–636. PMLR, 2015.

- [FO24] Steven T. Flammia and Ryan O’Donnell. Quantum chi-squared tomography and mutual information testing. *Quantum*, 8:1381, 2024.
- [HMRW24] Brice Huang, Sidhanth Mohanty, Amit Rajaraman, and David X. Wu. Weak Poincaré inequalities, simulated annealing, and sampling from spherical spin glasses. *CoRR*, abs/2411.09075, 2024.
- [HS23] Jonathan Hermon and Justin Salez. Modified log-Sobolev inequalities for strong-Rayleigh measures. *The Annals of Applied Probability*, 33(2):1501–1514, 2023.
- [KLV24] Frederic Koehler, Holden Lee, and Thuy-Duong Vuong. Efficiently learning and sampling multimodal distributions with data-based initialization. 2024.
- [LY98] Tzong-Yow Lee and Horng-Tzer Yau. Logarithmic Sobolev inequality for some models of random walks. *The Annals of Probability*, 26(4):1855–1873, 1998.
- [Nar21] Shyam Narayanan. On tolerant distribution testing in the conditional sampling model. In *Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 357–373. SIAM, 2021.
- [NT23] Shyam Narayanan and Jakub Tetek. Estimating the effective support size in constant query complexity. In *Symposium on Simplicity in Algorithms (SOSA)*, pages 242–252. SIAM, 2023.
- [Pea00] Karl Pearson. X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302):157–175, 1900.
- [Sal21] Justin Salez. A sharp log-Sobolev inequality for the multislice. *Annales Henri Lebesgue*, 4:1143–1161, 2021.

## A Deferred Proofs

### A.1 Proof of Lemma 11

*Proof of Lemma 11.* We directly compute:

$$\begin{aligned}
& \mathbf{Ent}_{a \sim \rho}^{\Phi} \left[ \mathbf{E}_{x \sim \mu_a} [f] \right] + \mathbf{E}_{a \sim \rho} \left[ \mathbf{Ent}_{x \sim \mu_a}^{\Phi} [f] \right] \\
&= \mathbf{E}_{a \sim \rho} \left[ \Phi \left( \mathbf{E}_{x \sim \mu_a} [f] \right) \right] - \Phi \left( \mathbf{E}_{a \sim \rho} \mathbf{E}_{x \sim \mu_a} [f] \right) + \mathbf{E}_{a \sim \rho} \left[ \mathbf{E}_{x \sim \mu_a} [\Phi(f(x))] \right] - \mathbf{E}_{a \sim \rho} \left[ \Phi \left( \mathbf{E}_{x \sim \mu_a} [f] \right) \right] \\
&= \mathbf{E}_{a \sim \rho} \left[ \mathbf{E}_{x \sim \mu_a} [\Phi(f(x))] \right] - \Phi \left( \mathbf{E}_{a \sim \rho} \mathbf{E}_{x \sim \mu_a} [f] \right) \\
&= \mathbf{E}_{x \sim \mu} [\Phi(f(x))] - \Phi \left( \mathbf{E}_{x \sim \mu} [f] \right) \\
&= \mathbf{Ent}_{\mu}^{\Phi} [f].
\end{aligned}$$

□

## A.2 Proof of Lemma 12

*Proof of Lemma 12.* We compute

$$\mathcal{L}_\mu[f] = \sum_{i \in [n]} \mathbf{E}_{\mathbf{x} \sim \mu} \left[ \mathbf{Ent}_{\mathbf{y} \sim \mu|_{\mathbf{x} \setminus i}} [f(\mathbf{y})] \right] = \sum_{i \in [n]} \mathbf{E}_{\mathbf{a} \sim \rho, \mathbf{x} \sim \mu_{\mathbf{a}}} \left[ \mathbf{Ent}_{\mathbf{y} \sim \mu|_{\mathbf{x} \setminus i}} [f(\mathbf{y})] \right].$$

Since  $\mu|_{\mathbf{x} \setminus i} = \mathbf{E}_{\mathbf{a}' \sim \rho|_{\mathbf{x} \setminus i}} [\mu_{\mathbf{a}'}|_{\mathbf{x} \setminus i}]$ , we have by the chain rule (Lemma 11) that the above is bounded below by

$$\begin{aligned} \sum_{i \in [n]} \mathbf{E}_{\mathbf{a} \sim \rho, \mathbf{x} \sim \mu_{\mathbf{a}}} \left[ \mathbf{E}_{\mathbf{a}' \sim \rho|_{\mathbf{x} \setminus i}} \left[ \mathbf{Ent}_{\mathbf{y} \sim \mu_{\mathbf{a}'}|_{\mathbf{x} \setminus i}} [f(\mathbf{y})] \right] \right] &= \sum_{i \in [n]} \mathbf{E}_{\mathbf{a} \sim \rho, \mathbf{x} \sim \mu_{\mathbf{a}}, \mathbf{a}' \sim \rho|_{\mathbf{x} \setminus i}} \left[ \mathbf{Ent}_{\mathbf{y} \sim \mu_{\mathbf{a}'}|_{\mathbf{x} \setminus i}} [f(\mathbf{y})] \right] \\ &= \sum_{i \in [n]} \mathbf{E}_{\mathbf{a} \sim \rho, \mathbf{x} \sim \mu_{\mathbf{a}}} \left[ \mathbf{Ent}_{\mathbf{y} \sim \mu_{\mathbf{a}}|_{\mathbf{x} \setminus i}} [f(\mathbf{y})] \right]. \end{aligned}$$

By applying ATE for each  $\mu_{\mathbf{a}}$ , we can lower-bound this by

$$\mathbf{E}_{\mathbf{a} \sim \rho} \left[ \sum_{i \in [n]} \mathbf{E}_{\mathbf{x} \sim \mu_{\mathbf{a}}} \left[ \mathbf{Ent}_{\mathbf{y} \sim \mu_{\mathbf{a}}|_{\mathbf{x} \setminus i}} [f(\mathbf{y})] \right] \right] \geq \mathbf{E}_{\mathbf{a} \sim \rho} \left[ \frac{1}{c^*} \cdot \mathbf{Ent}_{\mathbf{x} \sim \mu_{\mathbf{a}}} [f(\mathbf{x})] \right]. \quad \square$$

## A.3 Proof of Lemma 18

*Proof of Lemma 18.* Let  $\pi$  be any initial distribution and let  $f = \pi/\mu$  be the density function of  $\pi$  with respect to  $\mu$ . The impetus behind the switch to continuous-time  $P_t$  rather than discrete-time  $P$  is that we immediately get the following continuous characterization of  $\Phi$ -divergence contraction:

**Lemma 34** ([Cha04], Proposition 1). Let  $f_t = P_t f$ . Then,

$$\frac{d}{dt} D_\Phi(P_t \pi \parallel \mu) = -\mathcal{E}_P(f_t, \Phi'(f_t)).$$

This turns out to be the only piece of the proof missing towards a generalization of Theorem 4.5 from [HMRW24]. All that is left is to establish a *weak  $\Phi$ -Sobolev inequality* for mixtures.

**Definition 35.** A distribution  $\mu$  on  $\Sigma^n$  satisfies a *weak  $\Phi$ -Sobolev inequality* with constant  $c^*$  and error  $g : \mathbb{R}_{\geq 0}^{\Sigma^n} \rightarrow \mathbb{R}_{\geq 0}$  if for all  $f : \Sigma^n \rightarrow \mathbb{R}_{\geq 0}$ ,

$$\mathbf{Ent}_\mu^\Phi[f] \leq c^* \cdot \mathcal{E}_P(f, \Phi'(f)) + g(f).$$

Let  $f : \Sigma^n \rightarrow \mathbb{R}_{\geq 0}$  and observe

$$\begin{aligned} \mathbf{Ent}_\mu^\Phi[f] &= \mathbf{E}_{\mathbf{a} \sim \rho} \left[ \mathbf{Ent}_{\mathbf{x} \sim \mu_{\mathbf{a}}}^\Phi[f] \right] + \mathbf{Ent}_{\mathbf{a} \sim \rho}^\Phi \left[ \mathbf{E}_{\mathbf{x} \sim \mu_{\mathbf{a}}} [f] \right] \\ &\leq c^* \cdot \mathbf{E}_{\mathbf{a} \sim \rho} [\mathcal{E}_{P_{\mathbf{a}}}(f, \Phi'(f))] + \mathbf{Ent}_{\mathbf{a} \sim \rho}^\Phi \left[ \mathbf{E}_{\mathbf{x} \sim \mu_{\mathbf{a}}} [f] \right] \\ &\leq c^* \cdot \mathcal{E}_P(f, \Phi'(f)) + \mathbf{Ent}_{\mathbf{a} \sim \rho}^\Phi \left[ \mathbf{E}_{\mathbf{x} \sim \mu_{\mathbf{a}}} [f] \right]. \end{aligned}$$

The first line is the chain rule ([Lemma 11](#)). The second line applies the component-wise  $\Phi$ -Sobolev inequality. The third line invokes the concavity of the Dirichlet form for Glauber dynamics. That is, we use that

$$\begin{aligned}\mathcal{E}_P(f, \Phi'(f)) &= \mathbf{E}_{\mathbf{x}} \mathbf{E}_{\mathbf{y} \sim P_{\mathbf{x}}} [(f(\mathbf{x}) - f(\mathbf{y}))(\Phi'(f(\mathbf{x})) - \Phi'(f(\mathbf{y})))] \\ &= \frac{1}{n} \sum_{x \sim y} \frac{\mu(x)\mu(y)}{\mu(x) + \mu(y)} (f(x) - f(y))(\Phi'(f(x)) - \Phi'(f(y))) \\ &\geq \frac{1}{n} \sum_{x \sim y} \mathbf{E}_{\mathbf{a}} \left[ \frac{\mu_{\mathbf{a}}(x)\mu_{\mathbf{a}}(y)}{\mu_{\mathbf{a}}(x) + \mu_{\mathbf{a}}(y)} \right] (f(x) - f(y))(\Phi'(f(x)) - \Phi'(f(y))).\end{aligned}$$

Here the inequality follows from concavity of the map  $(a, b) \mapsto \frac{ab}{a+b}$  and the fact that  $\Phi'$  is increasing in  $f(\cdot)$ , so the summands are all positive.

With this weak  $\Phi$ -Sobolev inequality and [Lemma 34](#) the result follows by the exact proof of Theorem 4.5 from [\[HMRW24\]](#), replacing KL-divergence with  $\Phi$ -divergence and  $\mathcal{E}_P(f_t, \log f_t)$  with  $\mathcal{E}_P(f_t, \Phi'(f_t))$ .  $\square$

#### A.4 Proof of [Theorem 24](#)

*Proof of [Theorem 24](#).* Irrespective of  $q$  having minimum probability  $\eta$ , Theorem 1 of [\[DKW18\]](#)<sup>3</sup> gives an algorithm we'll call  $H^2$ -TEST( $p, q, \varepsilon, \delta$ ) that — given  $q, \varepsilon, \delta$  and samples from an unknown  $p$  on  $D$  — has sample complexity  $O\left(\frac{\sqrt{d} \cdot \log(1/\delta)}{\varepsilon}\right)$  and the following guarantee:

1. If  $\chi^2(p||q) \leq 0.5\varepsilon$  (e.g., if  $p = q$ ),  $H^2$ -TEST( $p, q, \varepsilon, \delta$ ) accepts with probability at least  $1 - \delta$ .
2. If  $H^2(p||q) \geq \varepsilon$ ,  $H^2$ -TEST( $p, q, \varepsilon, \delta$ ) accepts with probability at most  $\delta$ .

We also have the following inequality relating Hellinger distance and KL-divergence:

$$D_{H^2}(p \parallel q) \geq \frac{1}{\log(e^2/\eta)} \cdot D_{\text{KL}}(p \parallel q).$$

(With a slightly different constant factor, this inequality appears in, e.g., [\[BM98\]](#). See [\[FO24, Prop. 2.12\]](#) for the version above.) It follows that we can simply run  $H^2$ -TEST( $\varepsilon/\log(e^2/\eta), p, q, \delta$ ) to obtain the desired KL-TEST.  $\square$

---

<sup>3</sup>See also [\[ADK15, BOW19\]](#).