On Approximating the f-Divergence Between Two Ising Models

Weiming Feng
School of Computing and Data Science, The University of Hong Kong, China

Yucheng Fu
School of Computing and Data Science, The University of Hong Kong, China

Abstract

The f-divergence is a fundamental notion that measures the difference between two distributions. In this paper, we study the problem of approximating the f-divergence between two Ising models, which is a generalization of recent work on approximating the TV-distance. Given two Ising models ν and μ, which are specified by their interaction matrices and external fields, the problem is to approximate the f-divergence $D_f(\nu \,\|\, \mu)$ within an arbitrary relative error $\mathrm{e}^{\pm\varepsilon}$. For χ^α-divergence with a constant integer α, we establish both algorithmic and hardness results. The algorithm works in a parameter regime that matches the hardness result. Our algorithm can be extended to other f-divergences such as α-divergence, Kullback-Leibler divergence, Rényi divergence, Jensen-Shannon divergence, and squared Hellinger distance.

2012 ACM Subject Classification Theory of computation → Approximation algorithms analysis; Mathematics of computing → Markov networks

Keywords and phrases Ising model, f-divergence, approximation algorithms, randomized algorithms

Digital Object Identifier 10.4230/LIPIcs.ITCS.2026.59

Related Version *Full Version*: https://arxiv.org/abs/2509.05016

Funding *Weiming Feng*: Early Career Scheme of the Hong Kong Research Grants Council under grant number 27202725.
Yucheng Fu: 2025 Summer Research Internship Programme in the School of Computing and Data Science at The University of Hong Kong.

1 Introduction

Let ν and μ be two distributions with the same support Ω. Let $f : \mathbb{R}_{>0} \to \mathbb{R}_{\geq 0}$ be a convex function such that $f(1) = 0$ and f is strictly convex around 1. The f-divergence of ν from μ is defined by

$$D_f(\nu \,\|\, \mu) = \mathbf{E}_{X\sim\mu}\left[f\left(\frac{\nu(X)}{\mu(X)}\right)\right] = \sum_{\sigma\in\Omega}\mu(\sigma) f\left(\frac{\nu(\sigma)}{\mu(\sigma)}\right). \tag{1}$$

The f-divergence is a very general notion that measures the difference between two distributions. For instance, $f(x) = \frac{1}{2}|x-1|$ gives the *total variation distance (TV-distance)*, $f(x) = x^2 - x$ gives the *χ^2-divergence*, and $f(x) = x\ln x - x + 1$ gives the *Kullback-Leibler (KL) divergence*.

The *Ising model* is a fundamental graphical model in statistical physics, probability theory, and machine learning. Let $G = (V, E)$ be a graph. Let $J \in \mathbb{R}^{V\times V}$ be a *symmetric interaction matrix* such that $J_{uv} \neq 0$ only if $\{u, v\} \in E$. Let $h \in \mathbb{R}^V$ be the *external fields vector*. An Ising model specified by (G, J, h) defines a *Gibbs distribution* μ with support $\Omega = \{-1, +1\}^V$ such that

$$\forall \sigma \in \{-1,+1\}^V, \mu(\sigma) = \frac{w_\mu(\sigma)}{Z_\mu} = \frac{\exp\left(\frac{1}{2}\sigma^T J\sigma + h^T\sigma\right)}{Z_\mu}, \text{where } Z_\mu = \sum_{\sigma\in\{-1,+1\}^V} w_\mu(\sigma).$$

The function $w_\mu : \Omega \to \mathbb{R}_{\geq 0}$ is called the *weight function* of the Ising model and the normalization factor Z_μ is called the *partition function* of the Ising model.

Recently, the problem of computing the TV-distance between two high-dimensional distributions has received increasing attention. One interesting result [4] has proved that even for a pair of product distributions (Ising models on an empty graph), the *exact* computation of TV-distance is #P-hard. Later on, polynomial time *approximation algorithms* were proposed for product distributions [16, 18]. For general Ising models, both algorithmic and hardness results were established [6, 17] for approximating the TV-distance.

In this paper, we consider a more general problem of approximating the f-divergence between two Ising models. The problem is defined as follows.

▶ **Problem 1.** *Approximating the f-divergence for two Ising models.*

- ***Input**: Two Ising models (G, J^ν, h^ν) and (G, J^μ, h^μ) specifying two Gibbs distributions ν and μ respectively[1], a function f defining the f-divergence, and an error bound ε;*
- ***Output**: A number $\hat{D} \in \mathbb{R}$ such that $\mathrm{e}^{-\varepsilon} D_f(\nu \,\|\, \mu) \leq \hat{D} \leq \mathrm{e}^{\varepsilon} D_f(\nu \,\|\, \mu)$.*

A randomized algorithm is said to be an *FPRAS* (fully polynomial randomized approximation scheme) for Problem 1 if it runs in time polynomial in $n = |V|$ and $1/\varepsilon$ and with probability at least $\frac{2}{3}$, the output approximates the value of the f-divergence within relative error $\mathrm{e}^{\pm\varepsilon}$.

Our algorithmic result is a reduction from f-divergence approximation to sampling and approximate counting, which are two fundamental computational tasks for the Ising model. There is long-line of research on developing efficient algorithms [27] for sampling and approximate counting. We assume the following abstract oracles for sampling and approximate counting.

▶ **Definition 2** (sampling and approximate counting oracles)**.** *Let (G, J^μ, h^μ) be an Ising model with Gibbs distribution μ and partition function Z_μ. Let $T_G^{\mathrm{sp}}, T_G^{\mathrm{ct}} : \mathbb{R}_{>0} \to [|V| + |E|, \infty)$ be two non-increasing functions. Given any error bound $\varepsilon > 0$,*

- *The sampling oracle for (G, J^μ, h^μ) with cost function T_G^{sp} returns a random sample $X \in \{-1, +1\}^V$ in time $T_G^{\mathrm{sp}}(\varepsilon)$ with $d_{\mathrm{TV}}(X, \mu) \leq \varepsilon$, where $d_{\mathrm{TV}}(X, \mu)$ is the total variation distance between X and μ.*
- *The approximate counting oracle for (G, J^μ, h^μ) with cost function T_G^{ct} returns a random number $\hat{Z}_\mu$ in time $T_G^{\mathrm{ct}}(\varepsilon)$ with $\mathbf{Pr}\left[\mathrm{e}^{-\varepsilon} Z_\mu \leq \hat{Z}_\mu \leq \mathrm{e}^{\varepsilon} Z_\mu\right] \geq 0.99$.*

For functions $T_G^{\mathrm{sp}}(\varepsilon)$ and $T_G^{\mathrm{ct}}(\varepsilon)$, we add the index G to emphasize that the running time also depends on parameters of graph G such as the number of vertices/edges and the maximum degree. We assume the oracles need to read the whole graph G so that the cost is at least $|V| + |E|$.

We also require the following mild assumption on the Ising models.

▶ **Definition 3** (marginal lower bound)**.** *Let $b \geq 0$ be a constant. A Gibbs distribution μ is said to satisfy the b-marginal lower bound if for any $\Lambda \subseteq V$, any pinning $\sigma \in \{-1, +1\}^\Lambda$, any $v \in V \setminus \Lambda$, and any $c \in \{-1, +1\}$, it holds that $\mu_v^\sigma(c) \geq b$, where $\mu_v^\sigma(\cdot)$ denotes the marginal distribution on v projected from μ conditional on that the configuration on Λ is fixed as σ.*

The marginal lower bound condition is a natural and common assumption for the Ising model. The assumption is widely used in the literature of sampling and approximate counting [14], learning theory [10], and TV-distance approximation [17].

[1] Problem 1 assumes that ν and μ have the same underlying graph G. This assumption does not lose generality because the definition allows J_{uv}^ν and J_{uv}^μ to be 0 even if $\{u, v\} \in E$.

1.1 Algorithm and hardness results for χ^α-divergence approximation

Let $\alpha \geq 1$ be a constant integer. The $D_f(\nu \,\|\, \mu) = D_{\chi^\alpha}(\nu \,\|\, \mu)$ is called the χ^α-divergence if $f(x) = \frac{1}{2}|x-1|^\alpha$. We give the following approximation algorithm for the χ^α-divergence between two Ising models in Problem 1. Given two input Ising models (G, J^ν, h^ν) and (G, J^μ, h^μ), define the following family of Ising models on the graph G:

$$\mathcal{F}(\nu, \mu, \alpha) = \left\{ (G, J^{(k)}, h^{(k)}) \mid \begin{array}{l} J^{(k)} \triangleq kJ^\nu - (k-1)J^\mu \\ h^{(k)} \triangleq kh^\nu - (k-1)h^\mu \end{array} \text{ for integer } 0 \leq k \leq \alpha \right\}. \tag{2}$$

We remark that $(G, J^{(1)}, h^{(1)})$ and $(G, J^{(0)}, h^{(0)})$ are the same as the input Ising models (G, J^ν, h^ν) and (G, J^μ, h^μ) respectively. The following theorem says that if all Ising models in $\mathcal{F}$ admit sampling and approximate counting oracles, then the χ^α-divergence between two input Ising models can be approximated in polynomial time.

▶ **Theorem 4.** *Let $\alpha \geq 1$ be a constant integer and $b \in (0,1)$ be a constant. There exists an FPRAS that solves Problem 1 for χ^α-divergence in time*

$$O_{\alpha,b}\left(T_G^{\mathrm{ct}}\left(\frac{\eta_{\alpha,b}\cdot\varepsilon}{(n+m)^\alpha}\right) + \frac{(n+m)^{2\alpha}}{\varepsilon^2}\cdot T_G^{\mathrm{sp}}\left(\frac{\eta_{\alpha,b}\cdot\varepsilon^2}{(n+m)^{2\alpha}}\right)\right),$$

where $n = |V|$, $m = |E|$, and $\eta_{\alpha,b} > 0$ is a small constant depending only on α and b, if two input Ising Gibbs distributions μ and ν are both b-marginally bounded and all Ising models in $\mathcal{F}(\nu, \mu, \alpha)$ admit sampling and approximate counting oracles with cost functions $T_G^{\mathrm{sp}}(\cdot)$ and $T_G^{\mathrm{ct}}(\cdot)$ respectively.

Let us consider a simplified case of the Ising model. Suppose G has the max degree Δ. For an Ising model (G, J, h), if for any $e = \{u, v\} \in E$, if $J_{uv} = J_{vu} = \frac{\ln \beta}{2}$ take a unified value. In this case, an Ising model admits sampling and approximate counting oracles with cost function $T_G^{\mathrm{sp}}(\varepsilon) = \mathrm{poly}(n, \log \frac{1}{\varepsilon})$ and $T_G^{\mathrm{ct}}(\varepsilon) = \mathrm{poly}(n, \frac{1}{\varepsilon})$ if one of the following two conditions holds:

- Uniqueness condition: $\frac{\Delta-2}{\Delta} \leq \beta \leq \frac{\Delta}{\Delta-2}$ and an arbitrary external field $h \in \mathbb{R}^V$ [13].
- Uniqueness condition or ferromagnetic condition with zero external field: $\beta \geq \frac{\Delta-2}{\Delta}$ and the zero external field $h = \mathbf{0}$ [13, 21].

Consider Problem 1 where two input Ising models both have unified values $\frac{\ln \beta_\nu}{2}$ and $\frac{\ln \beta_\mu}{2}$ in the interaction matrices J^ν and J^μ. Assume β_ν, β_μ, $\|h^\nu\|_\infty$, $\|h^\mu\|_\infty$, and the max degree Δ are all constants. Then the marginal lower bound condition is satisfied. We have the following corollary.

▶ **Corollary 5.** *Let $\alpha \geq 1$ be a constant integer. For Ising models with unified values in interaction matrices, there exists an FPRAS that solves Problem 1 for χ^α-divergence in time $\mathrm{poly}(n, \frac{1}{\varepsilon})$ if β_ν, β_μ, $\|h^\nu\|_\infty$, $\|h^\mu\|_\infty$, and Δ are all constants and one of the following three conditions holds:*

- *Both β_μ and $(\frac{\beta_\nu}{\beta_\mu})^\alpha \beta_\mu$ are in $[\frac{\Delta-2}{\Delta}, \frac{\Delta}{\Delta-2}]$.*
- *$\beta_\nu \geq \beta_\mu \geq \frac{\Delta-2}{\Delta}$ and $h^\nu = h^\mu = \mathbf{0}$.*
- *$\beta_\nu < \beta_\mu$, $(\frac{\beta_\nu}{\beta_\mu})^\alpha \beta_\mu \geq \frac{\Delta-2}{\Delta}$, and $h^\nu = h^\mu = \mathbf{0}$.*

The corollary requires different conditions for the case $\beta_\nu \geq \beta_\mu$ and the case $\beta_\nu < \beta_\mu$. It is reasonable because for $\alpha > 1$, the χ^α-divergence is *not* symmetric $D_{\chi^\alpha}(\nu \,\|\, \mu) \neq D_{\chi^\alpha}(\mu \,\|\, \nu)$, and thus the roles of two input Ising models are different.

In Theorem 4, we show that χ^α-divergence can be approximated if the sampling and approximate counting oracles for all Ising models in $\mathcal{F}(\nu, \mu, \alpha)$ exist. In a special case when $\alpha = 1$, $\mathcal{F}(\nu, \mu, 1)$ only contains two input Ising models (G, J^ν, h^ν) and (G, J^μ, h^μ), where we recover the result in [17] for approximating the TV-distance. However, for general $\alpha > 1$, the family $\mathcal{F}(\nu, \mu, \alpha)$ contains other Ising models. It is natural to ask the following questions.

Questions

- Is it possible to approximate χ^α-divergence if we only assume the sampling and approximate counting oracles for the two input Ising models (G, J^ν, h^ν) and (G, J^μ, h^μ) exist?
- A more direct question: are the oracle assumptions in Theorem 4 necessary?

The above two questions are answered by the following hardness result. Let us restrict our attention to Ising models (G, J, h) with zero external field and interaction matrix with unified values. Formally, $h = \mathbf{0}$ and $J_{uv} = J_{vu} = \frac{\ln \beta}{2}$ for all $\{u, v\} \in E$. Denote the model by (G, β). Let (G, β_ν) and (G, β_μ) denote two input Ising models in Problem 1.

▶ **Theorem 6.** *Fix an integer $\alpha \geq 2$, an integer $\Delta \geq 3$, and two parameters $\beta_\mu > \beta_\nu \geq \frac{\Delta-2}{\Delta}$ such that $(\frac{\beta_\nu}{\beta_\mu})^\alpha \beta_\mu < \frac{\Delta-2}{\Delta}$. Unless $\mathsf{NP} = \mathsf{RP}$, there is no FPRAS for χ^α-divergence for two Ising models (G, β_ν) and (G, β_μ) on Δ-regular graph G.*

The above theorem considers two Ising models with zero external field and $\beta_\nu, \beta_\mu \geq \frac{\Delta-2}{\Delta}$. Two input Ising models both admit polynomial time sampling and approximate counting oracles because they are either in the uniqueness regime $\beta \in [\frac{\Delta-2}{\Delta}, \frac{\Delta}{\Delta-2}]$ [13] or the ferromagnetic regime $\beta \geq 1$ [21]. However, if $(\frac{\beta_\nu}{\beta_\mu})^\alpha \beta_\mu < \frac{\Delta-2}{\Delta}$, approximating χ^α-divergence is still hard. For a concrete example, the problem is hard when $\alpha \geq 2$, $\beta_\nu = \frac{\Delta-2}{\Delta}$, and $\beta_\mu = 2\beta_\nu = \frac{2(\Delta-2)}{\Delta}$.

Theorem 6 also shows that the condition in Theorem 4 is necessary. Note that since Δ, β_ν and β_μ are all constants and the external fields are zero, the hardness instances satisfy the marginal lower bound condition with $b = b(\Delta, \beta_\nu, \beta_\mu) = \Omega(1)$. Consider two Ising models (G, β_ν) and (G, β_μ) on Δ-regular graph G such that $\beta_\mu > \beta_\nu \geq \frac{\Delta-2}{\Delta}$. For this family of instances, our algorithmic result in Corollary 5 together with the hardness result in Theorem 6 discovers the following *computational phase transition* for approximating $D_{\chi^\alpha}(\nu \,\|\, \mu)$ when $\alpha \geq 2$:

- FPRAS exists when $(\frac{\beta_\nu}{\beta_\mu})^\alpha \beta_\mu \geq \frac{\Delta-2}{\Delta}$;
- Unless $\mathsf{NP} = \mathsf{RP}$, there is no FPRAS when $(\frac{\beta_\nu}{\beta_\mu})^\alpha \beta_\mu < \frac{\Delta-2}{\Delta}$.

Theorem 6 is proved via the hardness results of approximating partition functions of anti-ferromagnetic Ising models beyond the uniqueness regime [33, 19]. See Section 7 for details.

1.2 Algorithms for other f-divergences

We say an Ising model (G, J, h) admits polynomial time sampling and approximate counting oracles if it admits sampling and approximate counting oracles with cost function $T_G^{\mathrm{sp}}(\varepsilon) = \mathrm{poly}(\frac{n}{\varepsilon})$ and $T_G^{\mathrm{ct}}(\varepsilon) = \mathrm{poly}(\frac{n}{\varepsilon})$ respectively. Let (G, J^ν, h^ν) and (G, J^μ, h^μ) denote two input Ising models. We give algorithms for approximating the f-divergence $D_f(\nu \,\|\, \mu)$ for various divergence functions f. For many functions f, the divergence $D_f(\nu \,\|\, \mu)$ is *not* symmetric for ν and μ. We will refer to (G, J^ν, h^ν) as the first input Ising model and (G, J^μ, h^μ) as the second input Ising model.

By setting different functions f in $D_f(\nu \| \mu)$, we can obtain the following divergences:

$$f(x) = \begin{cases} x \ln x - x + 1, & \text{Kullback-Leibler divergence;} \\ -\ln x + x - 1, & \text{Rényi divergence;} \\ \frac{1}{2}\left(x \ln x - (x+1) \ln \frac{x+1}{2}\right) & \text{Jensen-Shannon divergence.} \end{cases}$$

▶ **Theorem 7.** *There exists an FPRAS for Problem 1 with KL-divergence, Rényi-divergence, and Jensen-Shannon divergence if two input Ising models are both marginally bounded and both admit polynomial time sampling and approximate counting oracles.*

Compared with the result in Theorem 4, the above theorem only requires the sampling and approximate counting oracles for the two input Ising models exist.

Next, consider the α-divergence. When $\alpha = 0$, the α-divergence refers to the KL-divergence. When $\alpha = 1$, the α-divergence refers to the Rényi-divergence. The α-divergences for other values of α are defined as follows. Let $\alpha \neq 0, 1$. The α-divergence is defined as $f(x) = \frac{x^\alpha - \alpha x - (1-\alpha)}{\alpha(\alpha-1)}$. Recall that in (2), $J^{(\alpha)} = \alpha J^\nu - (\alpha - 1)J^\mu$ and $h^{(\alpha)} = \alpha h^\nu - (\alpha - 1)h^\mu$. Here, α can be a *real* number.

▶ **Theorem 8.** *Let $\alpha \neq 0, 1$ be a constant real. There exists an FPRAS for Problem 1 with α-divergence if two input Ising models are both marginally bounded, the second input Ising model (G, J^μ, h^μ) admits a polynomial time sampling oracle, and both two input Ising models together with $(G, J^{(\alpha)}, h^{(\alpha)})$ admit polynomial time approximate counting oracles.*

The above theorem works for constant real α (assuming the computational model supports real arithmetic). Unlike χ^α-divergence, in addition to μ and ν we only require that the approximate counting oracle for $(G, J^{(\alpha)}, h^{(\alpha)})$ exists.

The squared Hellinger distance is defined by setting $f(x) = \frac{1}{2}(\sqrt{x} - 1)^2$. Given two input Ising models (G, J^ν, h^ν) and (G, J^μ, h^μ), define an averaged Ising model as $(G, J^{\mathsf{avg}}, h^{\mathsf{avg}})$, where $J^{\mathsf{avg}} = \frac{1}{2}(J^\nu + J^\mu)$ and $h^{\mathsf{avg}} = \frac{1}{2}(h^\nu + h^\mu)$. We have the following theorem.

▶ **Theorem 9.** *There exists an FPRAS for Problem 1 with squared Hellinger distance if two input Ising models are both marginally bounded, one of the input Ising models admits a polynomial time sampling oracle, and two input Ising models together with the averaged Ising model all admit polynomial time approximate counting oracles.*

The squared Hellinger distance is symmetric $D_f(\nu \| \mu) = D_f(\mu \| \nu)$. Hence, we only require one of the input Ising models admits a polynomial time sampling oracle. As a corollary, if two input Ising models are both in the uniqueness regime, then their averaged Ising model is also in the uniqueness regime and FPRAS exists for the squared Hellinger distance. For two Ising models (G, β_ν) and (G, β_μ) with zero external field, FPRAS exists when both $\beta_\nu, \beta_\mu \geq \frac{\Delta - 2}{\Delta}$.

1.3 Related work and open problems

Related work

The problem of computing the f-divergence between two (either discrete or continuous) distributions is well-motivated by the applications in machine learning and statistics. There are many works, both theoretical and practical, studying the algorithms in various settings. Many works study the error between the true divergence and the divergence computed by certain empirical distributions, e.g., [30, 22, 31, 34]. Embedding-base technique was also studied in [1], where algorithm embeds a large sample space into a smaller one while

preserving the f-divergence. We cannot directly use these techniques to Ising model as the size of the sample space is 2^n, which gives an exponential time algorithm. For graphical models, [26] shows that $\alpha\beta$-divergences can be computed exactly when the graphical model has a *bounded treewidth*.

Approximating the f-divergence is related to the testing problem, in which one or both two distributions are given by the access to different types of sampling oracles, and the algorithm is required to test certain divergence is large or small. There is a long line of work studying the testing problem. See [11, 12] for a comprehensive survey.

The TV-distance $d_{\mathrm{TV}}(\cdot,\cdot)$ is a special case of f-divergences. There are many theoretical works studying the approximation algorithms (either additive error or relative error) and hardness of approximation for the TV-distance between various types of high-dimensional distributions. For example, distributions specified by circuits [32], hidden Markov models [23], product distributions [4, 16, 18, 24], graphical models [8, 5, 17], and high-dimensional Gaussian distributions [8, 3]. Recently, there are some works focusing on approximating $1 - d_{\mathrm{TV}}(\cdot,\cdot)$ for two product distributions [25, 7].

Open problems

A natural direction is to consider more general distributions and more general divergences. Here are a few examples. For the Ising model, how to remove the marginal lower bound assumption? There are many other types of graphical models, such as Markov random fields with high-order interactions and Bayesian networks. It is interesting to consider other types of distributions, e.g., distributions generated by probabilistic circuits [2]. For the divergence, we focus on χ^α-divergence when α is an integer. One can also consider the case when α is a real number and give an algorithm that works in the tight parameter regime. Another question is to give a more general algorithm that works for a larger family of f-divergences.

The algorithm given in this paper is randomized. An open question is to find deterministic algorithms by exploring the deterministic approximate counting algorithms for the Ising model [28, 29]. Currently, deterministic algorithms are known for approximating the TV-distance between two product distributions [4, 18].

We can also study the same problem in different settings. In our setting, we assume that the input gives the description of two Ising models. It is interesting to consider another (perhaps more practical) setting that the algorithm can only access one or two models through random samples. This setting is closely related to the *Ising testing problem* in [15]. One can even consider abstract distributions with certain properties (e.g. approximate tensorization of f-divergence) and the distribution is given by the access to some oracles (e.g., oracle for querying conditional marginal distributions or querying probability masses). This abstract setting was considered by recent work in testing [9, 20].

2 Algorithm overview

We give an overview of our algorithm for approximating the χ^α-divergence between two Ising models. We start with the following definition of the *parameter distance* between two Ising models.

▶ **Definition 10** (parameter distance [17]). *For two Ising models (G, J^ν, h^ν) and (G, J^μ, h^μ), the parameter distance $d_{\mathrm{par}}(\nu,\mu)$ is defined by*

$$d_{\mathrm{par}}(\nu,\mu) \triangleq \max\left\{ \left\|J^\nu - J^\mu\right\|_{\max}, \max_v \frac{|h^\nu(v) - h^\mu(v)|}{\deg(v)+1} \right\},$$

where $\deg(v)$ is the degree of v in G.

In [17], the parameter distance is used to give a lower bound of TV-distance $d_{\mathrm{TV}}(\nu,\mu)$, with which the previous work gave an FPRAS for TV-distance. The following lemma says such a lower bound can be generalized to χ^α-divergence. The lemma is proved in Section 3.1, where the proof works for a general family of f-divergence.

▶ **Lemma 11** (χ^α-divergence lower bound). *For $f = \frac{1}{2}|x-1|^\alpha$, where $\alpha \geq 1$, it holds that*

$$D_{\chi^\alpha}(\nu \,\|\, \mu) \geq \frac{b^{2\alpha}}{2} \cdot d_{\mathrm{par}}(\nu,\mu)^\alpha.$$

Let $\theta = \frac{1}{\mathrm{poly}(n)}$ be a threshold. We estimate $D_{\chi^\alpha}(\nu \,\|\, \mu)$ in the following two cases.

- **Case** $d_{\mathrm{par}}(\nu,\mu) > \theta$. This case is trivial for TV-distance because there is a very simple additive-error approximation algorithm for the TV-distance [8]. Since TV-distance itself is lower bounded by $1/\mathrm{poly}(n)$, the additive error approximation can be easily transferred to relative error approximation. However, this simple algorithm only works for the TV-distance. Instead, we propose a new algorithm for approximating the general χ^α-divergence. The algorithm is outlined in Section 2.1.
- **Case** $d_{\mathrm{par}}(\nu,\mu) \leq \theta$. In this case, two Ising models are similar to each other. Previous work [17] has explored the similarity of two models and designed an algorithm for approximating their TV-distance. We substantially generalize their algorithm to approximate a family f-divergence (including χ^α-divergence). The algorithm is outlined in Section 2.2.

2.1 The algorithm for instances with large parameter distance

We first state the challenge of approximating the χ^α-divergence for general α compared to the TV-distance. It is well-known that the TV-distance can be written as follows:

$$d_{\mathrm{TV}}(\nu,\mu) = \frac{1}{2}\mathbf{E}_{X\sim\mu}\left[\left|\frac{\nu(X)}{\mu(X)} - 1\right|\right] \overset{(\star)}{=} \sum_{\sigma:\nu(\sigma)<\mu(\sigma)} \mu(\sigma)\left(1 - \frac{\nu(\sigma)}{\mu(\sigma)}\right). \tag{3}$$

It was observed in [8] that the TV-distance can be approximated by draw independent samples $X \sim \mu$ and then taking the average of $Y = \max\{0, 1 - \frac{\nu(X)}{\mu(X)}\}$[2]. The above equation shows $\mathbf{E}[Y] = d_{\mathrm{TV}}(\nu,\mu)$. It is easy to see $|Y| \leq 1$ and thus $\mathbf{Var}[Y] \leq 1$. The simple algorithm achieves the additive error approximation, which can be transferred to relative-error approximation because $d_{\mathrm{TV}}(\nu,\mu) \geq 1/\mathrm{poly}(n)$ is lower bounded in this case.

However, for the χ^α-divergence with general α, by the definition,

$$D_{\chi^\alpha}(\nu \,\|\, \mu) = \frac{1}{2}\sum_{\sigma\in\Omega}\mu(\sigma)\left|\frac{\nu(\sigma)}{\mu(\sigma)} - 1\right|^\alpha = \frac{1}{2}\sum_{\sigma\in\Omega}|\mu(\sigma)-\nu(\sigma)| \cdot \left|\frac{\nu(\sigma)}{\mu(\sigma)} - 1\right|^{\alpha-1}.$$

Due to the extra term $|\frac{\nu(\sigma)}{\mu(\sigma)} - 1|^{\alpha-1}$, we cannot apply a similar transformation as equation $(\star)$ in (3). Hence, it is not clear how to design an unbiased estimator Y such that variance of Y is small.

To overcome this challenge, we propose a new algorithm for approximating the χ^α-divergence. The following lemma gives a lower bound for the χ^α-divergence in this case.

[2] The value of Y can be computed approximately, which is sufficient for the purpose of approximation.

▶ **Lemma 12.** *Let $0 < \theta, b \le 1$. If $d_{\mathrm{par}}(\nu, \mu) \ge \theta$ and both ν and μ are b-marginally bounded, then*

$$D_{\chi^\alpha}(\nu \,\|\, \mu) \ge B_{\alpha,b}(\theta) \cdot \sum_{\sigma \in \{\pm\}^V} \mu(\sigma) \left(\frac{\nu(\sigma)}{\mu(\sigma)} + 1\right)^\alpha,$$

where $B_{\alpha,b}(\theta) = \frac{b^{2\alpha}\theta^\alpha}{2} \left(\left(\frac{b^{2\alpha}\theta^\alpha}{2}\right)^{1/(\alpha+1)} + 2\right)^{-\alpha} \left(2\left(\frac{b^{2\alpha}\theta^\alpha}{2}\right)^{1/(\alpha+1)} + 1\right)^{-1} = \Theta_{\alpha,b}(\theta^\alpha)$.

The proof of Lemma 12 separates all $\sigma \in \Omega$ into two parts, and separately lower bound their contributions to the χ^α-divergence. The detailed proof is deferred to Section 3.2.

Lemma 12 gives us a tool to control the error of approximation. Recall that our goal is to design an algorithm that approximates the χ^α-divergence with relative error $\mathrm{e}^{\pm\varepsilon}$. Equivalently, the algorithm should achieve the $O(\varepsilon) \cdot D_{\chi^\alpha}(\nu \,\|\, \mu)$ additive error. Using the above lemma, it is sufficient to show that the algorithm can achieve the $O(\varepsilon) \cdot B_{\alpha,b}(\theta) \cdot \sum_{\sigma\in\Omega} \mu(\sigma) \left(\frac{\nu(\sigma)}{\mu(\sigma)} + 1\right)^\alpha$ additive error approximation, which turns out to be much easier to prove. In addition to algorithms, Lemma 12 also plays an important role in the proof of the hardness result. See Section 7 for details.

2.1.1 The algorithm for even α

Our actual algorithm deals with all $\alpha \ge 1$ in a unified way. However, in the overview, we exhibit a simple algorithm for the case where α is even. The simple case will illustrate the intuition why we need to consider a family of Ising models $\mathcal{F}(\nu, \mu, \alpha)$ in (2) and why Lemma 12 can help us to control the error of approximation. When α is even, we can replace $\left|\frac{\nu(\sigma)}{\mu(\sigma)} - 1\right|^\alpha$ with $\left(\frac{\nu(\sigma)}{\mu(\sigma)} - 1\right)^\alpha$ and then use the binomial expansion to get the following equation:

$$D_{\chi^\alpha}(\nu \,\|\, \mu) = \frac{1}{2} \sum_{\sigma\in\Omega} \mu(\sigma) \left(\frac{\nu(\sigma)}{\mu(\sigma)} - 1\right)^\alpha = \frac{1}{2} \sum_{k=0}^{\alpha} \binom{\alpha}{k} (-1)^{\alpha-k} \sum_{\sigma\in\Omega} \frac{\nu^k(\sigma)}{\mu^{k-1}(\sigma)}. \tag{4}$$

To deal with the term $\frac{\nu^k(\sigma)}{\mu^{k-1}(\sigma)}$, recall that (2) defines a family of Ising models $(G, J^{(k)}, h^{(k)})$ such that $J^{(k)} \triangleq kJ^\nu - (k-1)J^\mu$ and $h^{(k)} \triangleq kh^\nu - (k-1)h^\mu$.

By our assumption in Theorem 4, for all $0 \le k \le \alpha$, the Ising model $(G, J^{(k)}, h^{(k)})$ admits sampling and approximate counting oracles. Let Z_k denote the *partition function* of the Ising model $(G, J^{(k)}, h^{(k)})$. We remark that $Z_\nu = Z_1$ and $Z_\mu = Z_0$ are the partition functions of input Ising model (G, J^ν, h^ν) and (G, J^μ, h^μ), respectively. Each term in the summation of (4) can be written as

$$\sum_{\sigma\in\Omega} \frac{\nu^k(\sigma)}{\mu^{k-1}(\sigma)} = \frac{Z_\mu^{k-1}}{Z_\nu^k} \sum_{\sigma\in\Omega} \frac{w_\nu^k(\sigma)}{w_\mu^{k-1}(\sigma)} = \frac{Z_\mu^{k-1}}{Z_\nu^k} \sum_{\sigma\in\Omega} \exp\left(\frac{\sigma^T J^{(k)} \sigma}{2} + \sigma^T h^{(k)}\right) = \frac{Z_\mu^{k-1} \cdot Z_k}{Z_\nu^k}. \tag{5}$$

The above ratio can be approximated by using approximate counting oracles for Z_ν, Z_μ, and Z_k. Note that $0 \le k \le \alpha$ is a constant. By choosing the relative approximation error $O(B_{\alpha,b}(\theta) \cdot \varepsilon) = \mathrm{poly}(\varepsilon/n)$ in the oracle small enough, where $B_{\alpha,b}(\theta) = \mathrm{poly}(1/n)$ is the parameter in Lemma 12, we can obtain a value $W_k \in \mathrm{e}^{\pm O(B_{\alpha,b}(\theta)\cdot\varepsilon)} \cdot \frac{Z_\mu^{k-1}\cdot Z_k}{Z_\nu^k}$ that achieves the following approximation guarantee

$$\left| W_k - \frac{Z_\mu^{k-1} \cdot Z_k}{Z_\nu^k} \right| \le B_{\alpha,b}(\theta) \cdot \varepsilon \cdot \frac{Z_\mu^{k-1} \cdot Z_k}{Z_\nu^k}, \tag{6}$$

where in the above inequality, we write the relative error in terms of the additive error. Finally, our algorithm outputs the number $\hat{D}$ such that

$$\hat{D} = \frac{1}{2}\sum_{k=0}^{\alpha}\binom{\alpha}{k}(-1)^{\alpha-k}W_k.$$

Note that $\hat{d}$ is an interlacing of plus and minus. Using (4) and (6), in the worst case (the additive error of every term takes the worst sign), the error of $\hat{d}$ can be upper bounded as follows

$$\begin{aligned}\left|\hat{D} - D_{\chi^\alpha}(\nu \,\|\, \mu)\right| &\le \varepsilon \cdot B_{\alpha,b}(\theta) \cdot \sum_{k=0}^{\alpha}\binom{\alpha}{k}\frac{Z_\mu^{k-1}\cdot Z_k}{Z_\nu^k} = \varepsilon \cdot B_{\alpha,b}(\theta)\cdot\sum_{k=0}^{\alpha}\binom{\alpha}{k}\sum_{\sigma\in\Omega}\frac{\nu^k(\sigma)}{\mu^{k-1}(\sigma)}\\ &= \varepsilon\cdot B_{\alpha,b}(\theta)\cdot\sum_{\sigma\in\Omega}\mu(\sigma)\left(\frac{\nu(\sigma)}{\mu(\sigma)}+1\right)^{\alpha} \overset{\text{Lemma 12}}{\le} \varepsilon\cdot D_{\chi^\alpha}(\nu\,\|\,\mu).\end{aligned}$$

2.1.2 The algorithm for general α

For general α, we need to consider the effect of the absolute value inside the divergence. The χ^α-divergence $D_{\chi^\alpha}(\nu\,\|\,\mu) = \frac{1}{2}\sum_{\sigma\in\Omega}\mu(\sigma)|\frac{\nu(\sigma)}{\mu(\sigma)}-1|^\alpha$ can be written as

$$D_{\chi^\alpha}(\nu\,\|\,\mu) = \frac{1}{2}\sum_{\sigma:\nu(\sigma)>\mu(\sigma)}\mu(\sigma)\left(\frac{\nu(\sigma)}{\mu(\sigma)}-1\right)^{\alpha} - \frac{1}{2}\sum_{\sigma:\nu(\sigma)<\mu(\sigma)}\mu(\sigma)\left(\frac{\nu(\sigma)}{\mu(\sigma)}-1\right)^{\alpha}.$$

Using the binomial expansion, the divergence can be written as

$$D_{\chi^\alpha}(\nu\,\|\,\mu) = \frac{1}{2}\sum_{k=0}^{\alpha}(-1)^{\alpha-k}\binom{\alpha}{k}\left(\sum_{\sigma:\nu(\sigma)>\mu(\sigma)}\frac{\nu^k(\sigma)}{\mu^{k-1}(\sigma)} + (-1)^\alpha\sum_{\sigma:\nu(\sigma)<\mu(\sigma)}\frac{\nu^k(\sigma)}{\mu^{k-1}(\sigma)}\right).$$

Fix an integer $0\le k\le\alpha$. Let X be a random sample from the Ising model $(G, J^{(k)}, h^{(k)})$. It holds that $\mathbf{Pr}[X=\sigma]\propto\frac{\nu^k(\sigma)}{\mu^{k-1}(\sigma)}$. Define W_k^+ and W_k^- to be random variables such that

$$W_k^+ = \mathbf{1}[\nu(X)>\mu(X)]\cdot\frac{Z_0^{k-1}\cdot Z_k}{Z_1^k} \quad\text{and}\quad W_k^- = \mathbf{1}[\nu(X)<\mu(X)]\cdot\frac{Z_0^{k-1}\cdot Z_k}{Z_1^k}. \tag{7}$$

With some calculation, we can verify that

$$\sum_{\sigma:\nu(\sigma)>\mu(\sigma)}\frac{\nu^k(\sigma)}{\mu^{k-1}(\sigma)} + (-1)^\alpha\sum_{\sigma:\nu(\sigma)<\mu(\sigma)}\frac{\nu^k(\sigma)}{\mu^{k-1}(\sigma)} = \mathbf{E}\left[W_k^+\right] + (-1)^\alpha\mathbf{E}\left[W_k^-\right].$$

Ideally, our algorithm wants to draw $T=\mathrm{poly}(\frac{\log n}{\varepsilon})$ independent random samples of W_k^+ and then uses their average value $\hat{W}_k^+$ to approximate $\mathbf{E}[W_k^+]$. Similarly, the algorithm computes $\hat{W}_k^-$ to approximate $\mathbf{E}[W_k^-]$. Finally, it outputs $\hat{D} = \frac{1}{2}\sum_{k=0}^{\alpha}(-1)^{\alpha-k}\binom{\alpha}{k}(\hat{W}_k^+ + (-1)^\alpha\hat{W}_k^-)$. However, to show the correctness of the algorithm, we need to deal with the following two types of errors.

- The algorithm cannot draw perfect samples of W_k^+ or W_k^-. One major issue[3] is that given a sample $X\sim\mathrm{Ising}(G, J^{(k)}, h^{(k)})$, the exact computation of the probability masses $\nu(X)$ and $\mu(X)$ is #P-hard [21]. Hence, the algorithm cannot perfectly distinguish the

[3] There are some other issues, e.g., the samples from the Ising model $(G, J^{(k)}, h^{(k)})$ are approximate and partition functions Z_0, Z_1, Z_k in (7) can only be approximated.

cases where $\nu(X) > \mu(X)$ or $\nu(X) < \mu(X)$. This issue can be solved by using the approximate counting oracle to approximate the probability masses $\nu(X)$ and $\mu(X)$. By choosing a small enough error parameter $\mathrm{poly}(n/\varepsilon)$ for approximate counting, the algorithm misclassifies $\nu(X) > \mu(X)$ or $\nu(X) < \mu(X)$ only if $|\frac{\nu(X)}{\mu(X)} - 1| \le \mathrm{poly}(\varepsilon/n)$. We need to show that such X does not contribute much to the error of approximation.

- We use the average of $\hat{W}_k^+$ and $\hat{W}_k^-$ to approximate $\mathbf{E}[W_k^+]$ and $\mathbf{E}[W_k^-]$ respectively. We need to control the concentration error. This error can be bounded via Lemma 12, using a similar technique as described in the even α case.

The detailed algorithm and analysis is in Section 5.1.

2.2 The algorithm for instances with small parameter distance

In this case, we give a general algorithm for abstract f-divergence. We briefly sketch the algorithm implemented for χ^α-divergence in the overview. The abstract one is in Section 4. By the definition,

$$D_{\chi^\alpha}(\nu \,\|\, \mu) = \frac{1}{2} \sum_{\sigma \in \{\pm\}^V} \mu(\sigma) \left| \frac{\nu(\sigma)}{\mu(\sigma)} - 1 \right|^\alpha = \frac{1}{2} \sum_{\sigma \in \{\pm\}^V} \mu(\sigma) \left| \frac{w_\nu(\sigma)}{w_\mu(\sigma)} \cdot \frac{Z_\mu}{Z_\nu} - 1 \right|^\alpha. \tag{8}$$

Define a random variable $W \triangleq \frac{w_\nu(X)}{w_\mu(X)}$, where $X \sim \mu$. It can be verified that $\mathbf{E}[W] = \frac{Z_\nu}{Z_\mu}$ and

$$D_{\chi^\alpha}(\nu \,\|\, \mu) = \mathbf{E}\left[\left| \frac{W}{\mathbf{E}[W]} - 1 \right|^\alpha \right].$$

The algorithm draws $T = \mathrm{poly}(n/\varepsilon)$ samples $W_1, \ldots, W_T$ from W independently, and then computes $\bar{W} = \frac{1}{T}\sum_{i=1}^T W_i$ and $\hat{D} = \frac{1}{T}\sum_{i=1}^T \frac{1}{2}\left|\frac{W_i}{\bar{W}} - 1\right|^\alpha$. Since the parameter distance is small, two weight functions $w_\nu(\cdot)$ and $w_\mu(\cdot)$ are very similar. Hence, the ratio $W = \frac{w_\nu(X)}{w_\mu(X)} \approx 1$. Formally,

$$\forall \sigma \in \Omega, \quad \left| \frac{w_\nu(\sigma)}{w_\mu(\sigma)} - \mathbf{E}[W] \right|^\alpha \le \mathrm{poly}(n) \cdot d_{\mathrm{par}}(\nu, \mu)^\alpha \overset{\text{Lemma 11}}{\le} \mathrm{poly}(n) \cdot D_{\chi^\alpha}(\nu \,\|\, \mu).$$

The above property guarantees that W is well-concentrated around its mean. By choosing the number of samples $T = \mathrm{poly}(n/\varepsilon)$ large enough, with some calculation, we can show that $\hat{D}$ approximates $D_{\chi^\alpha}(\nu \,\|\, \mu)$ with a relative error of $\mathrm{e}^{\pm O(\varepsilon)}$ with high probability. We remark that $\hat{D}$ is not necessarily an unbiased estimator for $D_{\chi^\alpha}(\nu \,\|\, \mu)$ but it is still a good approximation.

2.3 Organization of the paper

In Section 3, we prove some lower bounds for a broad family of f-divergences. In Section 4, we give a general algorithm for general f-divergences satisfying an abstract condition. In Section 5, we give the detailed algorithm for χ^α-divergence, where we focus on the large parameter distance case because the small parameter distance case is solved in Section 4. In Section 6, we show how to extend our algorithm to other f-divergences. Finally, in Section 7, we prove the hardness result for approximating χ^α-divergence in the parameter regime stated in Theorem 6.

3 Lower bounds of f-divergences

In this paper, we assume the function f in f-divergence satisfies the following assumption.

▶ **Assumption 1.** *The convex continuous function $f : (0, +\infty) \to [0, \infty)$ satisfies that $f(1) = 0$ and for any $x \neq 1$, f is twice differentiable at x such that $f''(x) \geq 0$. Furthermore, the derivative satisfies $f'(x) < 0$ if $0 < x < 1$ and $f'(x) > 0$ if $x > 1$.*

For any function f satisfying the above assumption, $x = 1$ is the unique point where $f(x) = 0$. It is straightforward to verify that f is strictly convex at around 1. The f-divergences we are interested in all satisfy the above assumption.

3.1 Lower bound in terms of parameter distance

Our proof uses the following lower bound of total variation distance.

▶ **Lemma 13** ([17, Lemma 15])**.** *For two Ising models, if both ν and μ are b-marginally bounded, then their total variation distance is lower bounded by*

$$d_{\mathrm{TV}}(\nu, \mu) \geq \frac{b^2}{2} \cdot d_{\mathrm{par}}(\nu, \mu).$$

▶ **Lemma 14.** *Suppose f satisfies Assumption 1. For two Gibbs distributions ν, μ of Ising models,*

$$\begin{aligned} D_f(\nu \,\|\, \mu) &\geq \max\{f(1 - d_{\mathrm{TV}}(\nu, \mu)), f(1 + d_{\mathrm{TV}}(\nu, \mu))\} \\ &\geq \max\left\{ f\left(1 - \frac{b^2}{2} d_{\mathrm{par}}(\nu, \mu)\right), f\left(1 + \frac{b^2}{2} d_{\mathrm{par}}(\nu, \mu)\right) \right\}. \end{aligned}$$

Proof of Lemma 14. Let $A \triangleq \{x \in \Omega : \nu(x) > \mu(x)\}$. Let $p = \mathbf{Pr}_{X \sim \nu}[X \in A]$ and $q = \mathbf{Pr}_{X \sim \mu}[X \in A]$. It is well-known that the total variation distance between ν and μ is $d_{\mathrm{TV}}(\nu, \mu) = p - q$. Since we consider the Ising model, $p, q > 0$. Consider a Markov kernel $K : \Omega \to \{0, 1\}$ such that given any $\sigma \in \Omega$, K deterministically transforms σ to 1 if and only if $\sigma \in A$. Note that νK and μK are Bernoulli distributions with parameters p and q respectively. We have

$$d_{\mathrm{TV}}(\nu, \mu) = d_{\mathrm{TV}}(\nu K, \mu K) = p - q.$$

By using the data processing inequality for f-divergence, we have

$$\begin{aligned} D_f(\nu \,\|\, \mu) &\geq D_f(\nu K \,\|\, \mu K) \\ &= q \cdot f\left(\frac{p}{q}\right) + (1 - q) \cdot f\left(\frac{1 - p}{1 - q}\right) \\ \left(f\left(\frac{1-p}{1-q}\right) \geq f(1) = 0\right) \quad &\geq q \cdot f\left(\frac{p}{q}\right) + (1 - q) \cdot f(1) \\ (\text{Jensen's inequality on } f) \quad &\geq f(1 + p - q) = f(1 + d_{\mathrm{TV}}(\nu, \mu)). \end{aligned}$$

Similarly, we use $f(p/q) \geq f(1)$ to get $D_f(\nu \,\|\, \mu) \geq f(1 - d_{\mathrm{TV}}(\nu, \mu))$.

Note that $f'(x) < 0$ if $0 < x < 1$ and $f'(x) > 0$ if $x > 1$. Combining with Lemma 13, we have

$$f(1 - d_{\mathrm{TV}}(\nu, \mu)) \geq f\left(1 - \frac{b^2}{2} d_{\mathrm{par}}(\nu, \mu)\right); f(1 + d_{\mathrm{TV}}(\nu, \mu)) \geq f\left(1 + \frac{b^2}{2} d_{\mathrm{par}}(\nu, \mu)\right). \quad \blacktriangleleft$$

Lemma 11 for χ^α-divergence can be proved as follows.

Proof of Lemma 11. For χ^α-divergence, we have $f(x) = \frac{1}{2}|x-1|^\alpha$. We can get a slightly better lower bound than the general result in Lemma 14. Note that $f(1-x) = f(1+x)$. Then $f(\frac{1-p}{1-q}) = f(\frac{1+p-2q}{1-q})$. Using the Jensen's inequality on $q \cdot f\left(\frac{p}{q}\right) + (1-q) \cdot f(\frac{1+p-2q}{1-q})$ implies the a lower bound $D_{\chi^\alpha}(\nu \,\|\, \mu) \geq f(1 + 2d_{\mathrm{TV}}(\nu,\mu)) = 2^{\alpha-1} d_{\mathrm{TV}}(\nu,\mu)^\alpha$. Then lemma follows from Lemma 13. ◀

3.2 A lower bound for χ^α-divergence

Proof of Lemma 12. Let $0 \leq t < 1$ be a parameter to be fixed later. We partition the whole space $\Omega = \{\pm\}^V$ into M and $\overline{M} = \Omega \setminus M$ such that

$$M = \left\{ \sigma \mid \left| \frac{\nu(\sigma)}{\mu(\sigma)} - 1 \right| \leq t \right\}.$$

Then, we bound the contribution of $\sigma \in M$ and $\sigma \in \overline{M}$ separately. By triangle inequality, we have $\frac{\nu(\sigma)}{\mu(\sigma)} + 1 \leq 2 + \left|\frac{\nu(\sigma)}{\mu(\sigma)} - 1\right|$. For the first case, we have

$$\sum_{\sigma \in M} \mu(\sigma) \left(\frac{\nu(\sigma)}{\mu(\sigma)} + 1 \right)^\alpha \leq \sum_{\sigma \in M} \mu(\sigma) (2+t)^\alpha \leq (2+t)^\alpha .$$

By our assumption, $d_{\mathrm{par}}(\nu,\mu) \geq \theta$. Using Lemma 11, we have $D_{\chi^\alpha}(\nu \,\|\, \mu) \geq \frac{b^{2\alpha}}{2}\theta^\alpha$. Hence,

$$\sum_{\sigma \in M} \mu(\sigma) \left(\frac{\nu(\sigma)}{\mu(\sigma)} + 1 \right)^\alpha \leq \frac{2 \cdot (2+t)^\alpha}{b^{2\alpha}\theta^\alpha} D_{\chi^\alpha}(\nu \,\|\, \mu) . \tag{9}$$

Consider the other set $\overline{M}$. For all $\sigma \in \overline{M}$, it holds that

$$\frac{\frac{\nu(\sigma)}{\mu(\sigma)} + 1}{\left|\frac{\nu(\sigma)}{\mu(\sigma)} - 1\right|} \leq 1 + \frac{2}{\left|\frac{\nu(\sigma)}{\mu(\sigma)} - 1\right|} \leq 1 + \frac{2}{t}.$$

Summing over all $\sigma \in \overline{M}$, we have

$$\begin{aligned} \sum_{\sigma \in \overline{M}} \mu(\sigma) \left(\frac{\nu(\sigma)}{\mu(\sigma)} + 1 \right)^\alpha &\leq \left(1 + \frac{2}{t}\right)^\alpha \sum_{\sigma \in \overline{M}} \mu(\sigma) \left| \frac{\nu(\sigma)}{\mu(\sigma)} - 1 \right|^\alpha \\ &\leq 2 \left(1 + \frac{2}{t}\right)^\alpha D_{\chi^\alpha}(\nu \,\|\, \mu) . \end{aligned} \tag{10}$$

Finally, adding the two inequalities (9) and (10), we have

$$\begin{aligned} \sum_{\sigma \in \Omega} \mu(\sigma) \left(\frac{\nu(\sigma)}{\mu(\sigma)} + 1 \right) &\leq \left(\frac{2 \cdot (2+t)^\alpha}{b^{2\alpha}\theta^\alpha} + 2 \left(1 + \frac{2}{t}\right)^\alpha \right) D_{\chi^\alpha}(\nu \,\|\, \mu) \\ &= g(t) \cdot D_{\chi^\alpha}(\nu \,\|\, \mu) . \end{aligned} \tag{11}$$

The above inequality holds for any $0 \leq t \leq 1$. We next find a value of t to minimize the value of $g(t)$.

Take the derivative of g, we get

$$g'(t) = 2\alpha \left(1 + \frac{2}{t}\right)^{\alpha-1} \left(-\frac{2}{t^2}\right) + \frac{2\alpha}{b^{2\alpha}\theta^\alpha}(t+2)^{\alpha-1} = 2\alpha(t+2)^{\alpha-1} \left(\frac{1}{b^{2\alpha}\theta^\alpha} - \frac{2}{t^{\alpha+1}} \right).$$

Thus when $t = \left(\frac{b^{2\alpha}\theta^{\alpha}}{2}\right)^{1/(\alpha+1)}$, where $0 < t < 1$,the function g obtains its minimum value:

$$\begin{aligned} g(t) &= \frac{2\cdot(2+t)^{\alpha}}{b^{2\alpha}\theta^{\alpha}} + 2\left(1+\frac{2}{t}\right)^{\alpha} = \frac{(2+t)^{\alpha}}{t^{\alpha+1}} + 2\left(1+\frac{2}{t}\right)^{\alpha} \\ &= \left(1+\frac{2}{t}\right)^{\alpha}\left(2+\frac{1}{t}\right) = \frac{(t+2)^{\alpha}(2t+1)}{t^{\alpha+1}} \\ &= \frac{2}{b^{2\alpha}\theta^{\alpha}}\left(\left(\frac{b^{2\alpha}\theta^{\alpha}}{2}\right)^{1/(\alpha+1)} + 2\right)^{\alpha}\left(2\left(\frac{b^{2\alpha}\theta^{\alpha}}{2}\right)^{1/(\alpha+1)} + 1\right) = \frac{1}{B_{\alpha,b}(\theta)}. \end{aligned}$$

Combining the above equation with (11) implies that $\sum_{\sigma\in\Omega}\mu(\sigma)\left(\frac{\nu(\sigma)}{\mu(\sigma)}+1\right) \le \frac{D_{\chi^{\alpha}}(\nu\,\|\,\mu)}{B_{\alpha,b}(\theta)}$. This proves the lower bound of the χ^{α}-divergence. ◀

4 Algorithm for f-divergence with small parameters distance

In this section, we generalize the algorithm in [17] to the case of small parameter distance. The new algorithm works for general f-divergence, where f satisfies the following abstract condition.

▶ **Condition 15.** *Let f be a function satisfying Assumption 1. There exists an function $F : \mathbb{R}^+ \to \mathbb{R}^+$ with $F(\zeta) \le \mathrm{poly}(\zeta)$ such that for any $\zeta \ge 1$, any $x \in (-\frac{1}{2\zeta}, 0)\cup(0, \frac{1}{2\zeta})$,*

$$\frac{xf'(1+\zeta x)}{f(1+x)} \le F(\zeta).$$

▶ **Theorem 16.** *Let f be a function satisfying Condition 15 with function F. There exists an algorithm such that given two Ising models (G, J^{ν}, h^{ν}) and (G, J^{μ}, h^{μ}), any $0 < \varepsilon < 1$, and f, b, F, if ν and μ are b-marginally bounded, $d_{\mathrm{par}}(\nu,\mu) < \theta = \frac{1}{10(n+3m)}$, and μ admits sampling oracle with cost functions $T_G^{\mathrm{sp}}(\cdot)$, then it returns a random number $\hat{D}$ in time $O(T\cdot T_G^{\mathrm{sp}}(\frac{1}{100T}))$, where $T = O(\frac{F(8(n+3m)/b^2)^2(n+m)^2}{\varepsilon^2})$, $n = |V|$, $m = |E|$, such that*

$$\mathbf{Pr}\left[e^{-\varepsilon}D_f(\nu\,\|\,\mu) \le \hat{D} \le e^{\varepsilon}D_f(\nu\,\|\,\mu)\right] \ge \frac{2}{3}.$$

▶ Remark 17. In Theorem 16, the algorithm for small parameter case only requires sampling oracles for the input Ising model (G, J^{μ}, h^{μ}) instead of sampling and approximate counting oracles for the whole family of Ising models $\mathcal{F}$.

By the definitions of f-divergence and Ising model, we have

$$D_f(\nu\,\|\,\mu) = \mathbf{E}_{\mu}\left[f\left(\frac{\nu(\sigma)}{\mu(\sigma)}\right)\right] = \mathbf{E}_{\mu}\left[f\left(\frac{w_{\nu}(\sigma)}{w_{\mu}(\sigma)}\cdot\frac{Z_{\mu}}{Z_{\nu}}\right)\right].$$

Define the parameter

$$T \triangleq \left\lceil \frac{2^{12}\cdot 10^3 F(8(n+3m)/b^2)^2(n+3m)^2}{b^4\varepsilon^2}\right\rceil.$$

Algorithm for small parameter distance case.

- Call the sampling oracle of μ with TV-distance error $\frac{1}{100T}$ to obtain independent random samples $\hat{\sigma}_1, \hat{\sigma}_2, \ldots, \hat{\sigma}_T$. For each $i \in [T]$, compute $\hat{W}_i = \frac{w_{\nu}(\hat{\sigma}_i)}{w_{\mu}(\hat{\sigma}_i)}$.
- Return $\hat{D} = \frac{1}{T}\sum_{i=1}^{T} f\left(\hat{W}_i/\bar{W}\right)$, where $\bar{W} = \frac{1}{T}\sum_{i=1}^{T}\hat{W}_i$.

The analysis of the algorithm is given in the full version of the paper.

5 Algorithm for χ^α-divergence

Our algorithm first computes the marginal lower bound b in time $O(n+m)$. Due to the conditional independence, the algorithm only need to enumerate all $v \in V$, any $c \in \{-1,+1\}$, and find a worst pinning $\tau = \tau(v,c)$ on all neighbor of v. Specifically, for any neighbor u of v, if $J_{uv} > 0$, then fix $\tau_u = -c$; otherwise, fix $\tau_u = c$. The algorithm compute $\mu_v^\tau(c)$. Let b be the minimum value over all $v \in V$ and $c \in \{-1,+1\}$. We now assume that the value b is known by the algorithm.

Let $\theta = \frac{1}{10(n+3m)}$ be the threshold parameter. We give two algorithms in Section 5.1 and Section 5.2 depending on whether $d_{\mathrm{par}}(\nu,\mu) > \theta$ or not. We prove Theorem 4 in Section 5.3.

5.1 Large parameters distance case

▶ **Lemma 18.** *Let $\alpha \geq 1$ be an integer and $b \in (0,1)$ be a constant. There exists an algorithm such that given two Ising models (G, J^ν, h^ν) and (G, J^μ, h^μ), any $0 < \varepsilon < 1$, and α, b, if ν and μ are b-marginally bounded, $d_{\mathrm{par}}(\nu,\mu) > \theta$, and every Ising model in $\mathcal{F}$ admits sampling and approximate counting oracles with cost function $T_G^{\mathrm{sp}}(\cdot)$ and $T_G^{\mathrm{ct}}(\cdot)$ respectively, then it returns a random $\hat{D}$ in time $O_{\alpha,b}\left(T_G^{\mathrm{ct}}(\delta) + T \cdot \left(T_G^{\mathrm{sp}}\left(\frac{1}{200T(\alpha+1)}\right)\right)\right)$, where $\delta = \Theta_{\alpha,b}(\theta^\alpha \varepsilon)$ and $T = \Theta_{\alpha,b}(\frac{1}{\varepsilon^2\theta^\alpha})$ such that*

$$\mathbf{Pr}\left[\mathrm{e}^{-\varepsilon} D_{\chi^\alpha}(\nu \,\|\, \mu) \leq \hat{D} \leq \mathrm{e}^{\varepsilon} D_{\chi^\alpha}(\nu \,\|\, \mu)\right] \geq \frac{2}{3}.$$

By the definition of χ^α-divergence, we can rewrite

$$\begin{aligned} D_{\chi^\alpha}(\nu \,\|\, \mu) &= \frac{1}{2} \sum_{\sigma:\nu(\sigma)>\mu(\sigma)} \mu(\sigma) \left(\frac{\nu(\sigma)}{\mu(\sigma)} - 1\right)^\alpha + \frac{1}{2} \sum_{\sigma:\nu(\sigma)<\mu(\sigma)} \mu(\sigma) \left(1 - \frac{\nu(\sigma)}{\mu(\sigma)}\right)^\alpha \\ &= \frac{1}{2} \sum_{k=0}^{\alpha} (-1)^{\alpha-k} \binom{\alpha}{k} \left(\sum_{\sigma:\nu(\sigma)>\mu(\sigma)} \frac{\nu^k(\sigma)}{\mu^{k-1}(\sigma)} + (-1)^\alpha \sum_{\sigma:\nu(\sigma)<\mu(\sigma)} \frac{\nu^k(\sigma)}{\mu^{k-1}(\sigma)} \right). \end{aligned}$$

Recall that Z_k is the partition function of the Ising model $(G, J^{(k)}, h^{(k)})$, where $J^{(k)} \triangleq kJ^\nu - (k-1)J^\mu$ and $h^{(k)} \triangleq kh^\nu - (k-1)h^\mu$. Note that $Z_\mu = Z_0$ and $Z_\nu = Z_1$. Define two random variables W_k^+ and W_k^- as follows. Let $\pi^{(k)}$ be the Gibbs distributions of the Ising model $(G, J^{(k)}, h^{(k)})$.

$$W_k^+ = \mathbf{1}[\nu(X) > \mu(X)] \frac{Z_0^{k-1} \cdot Z_k}{Z_1^k}, \quad \text{where} \quad X \sim \pi^{(k)}.$$
$$W_k^- = \mathbf{1}[\nu(Y) < \mu(Y)] \frac{Z_0^{k-1} \cdot Z_k}{Z_1^k}, \quad \text{where} \quad Y \sim \pi^{(k)}.$$

The expectation of W_k^+ can be calculated as follows

$$\begin{aligned} \mathbf{E}\left[W_k^+\right] &= \sum_{\sigma:\nu(\sigma)>\mu(\sigma)} \pi^{(k)}(\sigma) \frac{Z_0^{k-1} \cdot Z_k}{Z_1^k} = \sum_{\sigma:\nu(\sigma)>\mu(\sigma)} \frac{w_\nu^k(\sigma)/w_\mu^{k-1}(\sigma)}{Z_k} \frac{Z_0^{k-1} \cdot Z_k}{Z_1^k} \\ &= \sum_{\sigma:\nu(\sigma)>\mu(\sigma)} \frac{\nu^k(\sigma)}{\mu^{k-1}(\sigma)}. \end{aligned}$$

Similarly, we have $\mathbf{E}\left[W_k^-\right] = \sum_{\sigma:\nu(\sigma)<\mu(\sigma)} \frac{\nu^k(\sigma)}{\mu^{k-1}(\sigma)}$. In a high level, our algorithm draws approximate samples of W_k^+ and W_k^- and estimate $\mathbf{E}[W_k^+] + (-1)^\alpha \mathbf{E}[W_k^-]$. Define the parameter

$$T \triangleq \left\lceil \frac{8 \cdot 10^4(\alpha+1)}{\varepsilon^2 B_{\alpha,b}(\theta)^2} \right\rceil \quad \text{and} \quad \delta \triangleq \frac{B_{\alpha,b}(\theta)\varepsilon}{20(\alpha+1)},$$

where the function $B_{\alpha,b}(\theta)$ is defined in Lemma 12.

Algorithm for large parameter distance case.

- For $0 \le k \le \alpha$, call the approximate counting oracle on Ising model $(G, J^{(k)}, h^{(k)})$ for $O(\log \alpha)$ times independently with relative error δ, and take the median as $\hat{Z}_k$.
- For each k from 0 to α do:
  1. Draw $2T$ samples $\hat{\sigma}_1, \ldots, \hat{\sigma}_{2T} \sim \pi^{(k)}$ by the sampling oracle with error $\frac{1}{200T(\alpha+1)}$.
  2. For each $i \in [T]$, compute two values

$$\hat{W}_{k,i}^+ = \mathbf{1}[\hat{\nu}(\hat{\sigma}_i) > \hat{\mu}(\hat{\sigma}_i)] \frac{\hat{Z}_0^{k-1} \cdot \hat{Z}_k}{\hat{Z}_1^k}, \qquad \text{and}$$

$$\hat{W}_{k,i}^- = \mathbf{1}[\hat{\nu}(\hat{\sigma}_{T+i}) < \hat{\mu}(\hat{\sigma}_{T+i})] \frac{\hat{Z}_0^{k-1} \cdot \hat{Z}_k}{\hat{Z}_1^k},$$

  where for any $x \in \Omega$, $\hat{\nu}(x) = \frac{w_\nu(x)}{\hat{Z}_1}$ and $\hat{\mu}(x) = \frac{w_\mu(x)}{\hat{Z}_0}$.
- Return $\hat{D} = \frac{1}{2}\sum_{k=0}^{\alpha}(-1)^{\alpha-k}\binom{\alpha}{k}\left(\frac{1}{T}\sum_{i=1}^{T}\hat{W}_{k,i}^+ + (-1)^\alpha \frac{1}{T}\sum_{i=1}^{T}\hat{W}_{k,i}^-\right)$.

The analysis of the algorithm is given in the full version of the paper.

5.2 Small parameters distance case

With Theorem 16, we only need to verify Condition 15 for $f = \frac{1}{2}|x-1|^\alpha$. By definition,

$$\frac{xf'(1+\zeta x)}{f(1+x)} = \frac{\frac{\alpha}{2}|x|^\alpha|\zeta|^{\alpha-1}}{\frac{1}{2}|x|^\alpha} = \alpha\zeta^{\alpha-1}.$$

Set $F(\zeta) = \alpha\zeta^{\alpha-1}$, $F(\zeta)$ satisfies Condition 15. Thus

$$F(8(n+3m)/b^2) = \frac{2^{3\alpha-3}\alpha(n+3m)^{\alpha-1}}{b^{2\alpha-2}}.$$

Hence, using Theorem 16, there is an algorithm with running time $O(T \cdot T_G^{\mathrm{sp}}(\frac{1}{100T}))$, where $T = O(\frac{2^{6\alpha-6}\alpha^2(n+3m)^{2\alpha}}{b^{4\alpha-4}\varepsilon^2}) = O_{\alpha,b}(\frac{(n+3m)^{2\alpha}}{\varepsilon^2})$, for small parameter distance case $d_{\mathrm{par}}(\nu,\mu) < \theta = \frac{1}{10(n+3m)}$.

5.3 Putting everything together (Proof of Theorem 4)

Theorem 4 follows from Theorem 16 and Lemma 18. Define parameter $T = \Theta_{\alpha,b}(\frac{(n+m)^{2\alpha}}{\varepsilon^2})$ and $\delta = \Theta_{\alpha,b}(\frac{\varepsilon}{(n+m)^\alpha})$. By choosing constants, we can make T large enough and δ small enough. The preprocessing time for computing b and $d_{\mathrm{par}}(\nu,\mu)$ is $O(n+m)$, which is dominated by the time for sampling and approximate counting. The running time of the whole algorithm is at most

$$\max\left\{O_{\alpha,b}\left(T_G^{\mathrm{ct}}(\delta)+T\cdot T_G^{\mathrm{sp}}\left(\frac{1}{200T(\alpha+1)}\right)\right),O_{\alpha,b}\left(T\cdot T_G^{\mathrm{sp}}\left(\frac{1}{100T}\right)\right)\right\}$$
$$\leq O_{\alpha,b}\left(T_G^{\mathrm{ct}}\left(\frac{\eta_{\alpha,b}\cdot\varepsilon}{(n+m)^{\alpha}}\right)+\frac{(n+m)^{2\alpha}}{\varepsilon^2}\cdot T_G^{\mathrm{sp}}\left(\frac{\eta_{\alpha,b}\cdot\varepsilon^2}{(n+m)^{2\alpha}}\right)\right),$$

where the last inequality comes from the fact that both T_G^{ct} and T_G^{sp} are non-increasing functions and $\eta_{\alpha,b}>0$ is a small enough constant depending only on α and b.

6 Algorithms for other f-divergences

In this section, we briefly show the algorithms for other f-divergences. The algorithms prove the results in Theorem 7, Theorem 8, and Theorem 9.

Again, the algorithm first computes the parameter distance $d_{\mathrm{par}}(\nu,\mu)$. Then, it compares $d_{\mathrm{par}}(\nu,\mu)$ to the threshold $\theta=\frac{1}{10(n+3m)}$. If $d_{\mathrm{par}}(\nu,\mu)\leq\theta$, algorithms are given in Section 6.1. Otherwise $d_{\mathrm{par}}(\nu,\mu)>\theta$, algorithms are given in Section 6.2.

6.1 Algorithms for small parameter distance

In this case, we use the abstract algorithm in Theorem 16. We only need to verify Condition 15 for f-divergences. The following lemma gives a sufficient condition for Condition 15.

▶ **Lemma 19.** *Suppose $f:\mathbb{R}_{>0}\to\mathbb{R}_{\geq 0}$ is twice differentiable at every $x>0$ such that $f''(x)>0$ and $f'(1)=f(1)=0$. If there exists two positive constants $L,U>0$ such that $L\leq f''(x)\leq U$ for any $x\in[\frac{1}{2},\frac{3}{2}]$. Then, the function f satisfies Condition 15 with $F(\zeta)=\frac{2U}{L}\zeta$.*

Proof. Note that f satisfying the condition in the lemma satisfies Assumption 1. By condition that $f'(1)=f(1)=0$, we have the following two equalities ($\int_0^x=-\int_x^0$ if $x<0$):

$$\forall x\in\left(-\frac{1}{2},\frac{1}{2}\right),\ f'(1+x)=\int_0^x f''(1+u)\,du,\ f(1+x)=\int_0^x(x-u)f''(1+u)\,du. \quad (12)$$

Since $0<L\leq f''(\zeta x)\leq U$ for $\zeta>1$ and $x\in[-\frac{1}{2\zeta},0)\cup(0,\frac{1}{2\zeta}]$, we have

$$|f'(1+\zeta x)|=\left|\int_0^{\zeta x}f''(1+u)\,du\right|\leq\zeta|x|\cdot U,$$
$$f(1+x)=\int_0^x(x-u)f''(1+u)\,du\geq L\int_0^x(x-u)\,du=\frac{Lx^2}{2}.$$

Thus, Condition 15 can be verify as follows

$$\frac{xf'(1+\zeta x)}{f(1+x)}=\frac{|x|\cdot|f'(1+\zeta x)|}{f(1+x)}\leq\frac{|x|\cdot\zeta|x|U}{Lx^2/2}=\frac{2U}{L}\zeta.$$ ◀

Then, we using Lemma 19 verify Condition 15 for Kullback-Leibler, Rényi, Jensen-Shannon, α-divergence, and Squared Hellinger divergence. The result is summarized as the following table.

Divergence	$f(x)$	$f''(x)$	L	U	$F(\zeta)$
Kullback-Leibler	$x\ln x - x + 1$	$\frac{1}{x}$	$\frac{2}{3}$	2	6ζ
Rényi	$-\ln x + x - 1$	$\frac{1}{x^2}$	$\frac{4}{9}$	4	18ζ
Jensen-Shannon	$\frac{1}{2}(x\ln x - (x+1)\ln\frac{x+1}{2})$	$\frac{1}{2x(x+1)}$	$\frac{2}{15}$	$\frac{2}{3}$	10ζ
α-divergence ($\alpha \neq 0, 1$)	$\frac{x^\alpha - \alpha x - (1-\alpha)}{\alpha(\alpha-1)}$	$x^{\alpha-2}$	$2^{-\lvert\alpha-2\rvert}$	$2^{\lvert\alpha-2\rvert}$	$2\cdot 4^{\lvert\alpha-2\rvert}\zeta$
Squared Hellinger	$\frac{1}{2}(\sqrt{x}-1)^2$	$\frac{1}{2\sqrt{x}}$	$\frac{1}{\sqrt{6}}$	$\frac{1}{\sqrt{2}}$	$2\sqrt{3}\zeta$

We remark that Lemma 19 does not work for χ^α-divergence but it works for divergences above.

The algorithm then follows from Theorem 16 such that to approximate $D_f(\nu \,\|\, \mu)$, it only requires polynomial time sampling oracle for the input distribution μ. For squared Hellinger distance, as it is symmetric $D_f(\nu \,\|\, \mu) = D_f(\mu \,\|\, \nu)$, we can swap the roles of ν and μ in the algorithm to make sure μ admits a polynomial time sampling oracle.

6.2 Algorithms for large parameter distance

Now, assume that the parameter distance $d_{\mathrm{par}}(\nu,\mu) > \theta = \frac{1}{10(n+3m)}$. By Lemma 14, the f-divergence is at least

$$\max\left\{f\left(1-\frac{b^2}{2}d_{\mathrm{par}}(\nu,\mu)\right), f\left(1+\frac{b^2}{2}d_{\mathrm{par}}(\nu,\mu)\right)\right\} \geq \max\left\{f\left(1-\frac{b^2}{2}\theta\right), f\left(1+\frac{b^2}{2}\theta\right)\right\},$$

where the inequality holds because of the derivative assumptions in Assumption 1. Let $\delta = \frac{b^2}{2}\theta = \frac{1}{\mathrm{poly}(n)}$. We show that all divergences in the above table are at least $\frac{1}{\mathrm{poly}(n)}$. Kullback-Leibler and Rényi divergences can be verified by using Taylor series $\ln(1+\delta) = \delta - \frac{\delta^2}{2} + O(\delta^3)$. For Jensen-Shannon, one can rewrite $f(1+\delta) = (1+\delta)\ln(1+\frac{\delta}{2+\delta}) - \ln(1+\frac{\delta}{2})$ and then Taylor series shows $f(1+\delta) = \Omega(\delta^2)$. For α-divergence, expanding $(1+\delta)^\alpha$ for real α implies $f(1+\delta) = \frac{\delta^2}{2} \pm O(\delta^3) = \Omega(\delta^2)$. Finally, it is easy to verify that $f(1+\delta) = \Omega(\delta^2)$ for Squared Hellinger divergence. We have

$$D_f(\nu \,\|\, \mu) \geq \frac{1}{\mathrm{poly}(n)}. \tag{13}$$

Kullback-Leibler, Rényi, and Jensen-Shannon divergences

We give the algorithm for Jensen-Shannon divergence. Similar algorithms can be given for KL and Rényi divergences. The Jensen-Shannon divergence can be written as

$$\frac{1}{2}\sum_{\sigma\in\Omega}\nu(\sigma)\ln\frac{\nu(\sigma)}{\mu(\sigma)} - \frac{1}{2}\sum_{\sigma\in\Omega}\nu(\sigma)\ln\frac{\nu(x)+\mu(x)}{2\mu(x)} - \frac{1}{2}\sum_{\sigma\in\Omega}\mu(x)\ln\frac{\nu(x)+\mu(x)}{2\mu(x)}.$$

By (13), it suffices to give an algorithm with additive error $\frac{\varepsilon}{\mathrm{poly}(n)}$. We can estimate each term with an additive error. We give the algorithm for the first term. The other two terms can be estimated similarly. We use the sampling oracle to draw T samples σ from ν, then using counting oracles to approximate $h(\sigma) = \ln\frac{\nu(\sigma)}{\mu(\sigma)} = \ln(\frac{w_\nu(\sigma)}{w_\mu(\sigma)}\cdot\frac{Z_\mu}{Z_\nu})$, finally, return the average of $h(\sigma)$. The counting oracles estimate the partition functions with a relative error, and thus we can approximate $h(\sigma)$ with an additive error. Due to the marginal lower bound assumption, $|h(\sigma)| \leq n\log\frac{1}{b} = O(n)$. We can set $T = \mathrm{poly}(n, \frac{1}{\varepsilon})$ to achieve the additive error $\frac{\varepsilon}{\mathrm{poly}(n)}$ and our goal is achieved.

α-divergence

The α-divergence can be written as

$$\frac{1}{\alpha(\alpha-1)}\left(\sum_{\sigma}\frac{\nu(\sigma)^{\alpha}}{\mu(\sigma)^{\alpha-1}}\right)-\frac{1}{\alpha(\alpha-1)}=\frac{1}{\alpha(\alpha-1)}\left(\frac{Z_{\mu}^{\alpha-1}\cdot Z_{\alpha}}{Z_{\nu}^{\alpha}}-1\right),$$

where Z_α is the partition function of the Ising model $(G, J^{(\alpha)}, h^{(\alpha)})$. By calling approximate counting oracles with relative error $\mathrm{e}^{\pm\Theta_\alpha(\delta)}$, we estimate $\frac{Z_\mu^{\alpha-1}Z_\alpha}{Z_\nu^\alpha}$ with $\mathrm{e}^{\pm\delta}$ relative error, which is equivalent to $\pm O(\delta)\frac{Z_\mu^{\alpha-1}Z_\alpha}{Z_\nu^\alpha}$ additive error approximation. When the constant $\alpha > 1$ or $\alpha < 0$, by (13), $\frac{Z_\mu^{\alpha-1}Z_\alpha}{Z_\nu^\alpha}$ is lower bounded by $\frac{1}{\mathrm{poly}(n)}+1$. We set $\delta=\frac{\varepsilon}{\mathrm{poly}(n)}$, then $\frac{Z_\mu^{\alpha-1}Z_\alpha}{Z_\nu^\alpha}\geq\frac{\varepsilon}{\varepsilon-\delta}$, which implies

$$\delta\cdot\frac{Z_{\mu}^{\alpha-1}Z_{\alpha}}{Z_{\nu}^{\alpha}}\leq\varepsilon\cdot\left(\frac{Z_{\mu}^{\alpha-1}Z_{\alpha}}{Z_{\nu}^{\alpha}}-1\right).$$

When the constant $0 < \alpha < 1$, $\alpha(1-\alpha) < 0$. By (13), the α-divergence at least $\frac{1}{\mathrm{poly}(n)}$, $\frac{Z_\mu^{\alpha-1}Z_\alpha}{Z_\nu^\alpha}$ is upper bounded by $1-\frac{1}{\mathrm{poly}(n)}$. We still set $\delta=\frac{\varepsilon}{\mathrm{poly}(n)}$, and in this case,

$$\delta\cdot\frac{Z_{\mu}^{\alpha-1}Z_{\alpha}}{Z_{\nu}^{\alpha}}\leq\frac{\varepsilon}{\mathrm{poly}(n)}\leq\varepsilon\cdot\left(1-\frac{Z_{\mu}^{\alpha-1}Z_{\alpha}}{Z_{\nu}^{\alpha}}\right).$$

In both cases, our algorithm solves the problem with an additive error $O(\varepsilon)D_\alpha(\nu\,\|\,\mu)$ approximation, which implies a relative error $\mathrm{e}^{\pm\varepsilon}$ approximation.

Squared Hellinger distance

The Squared Hellinger distance can be written as

$$1-\sum_{\sigma\in\Omega}\mu(\sigma)\left(\frac{\nu(\sigma)}{\mu(\sigma)}\right)^{1/2}=1-\frac{\bar{Z}}{\sqrt{Z_\nu Z_\mu}},$$

where $\bar{Z}$ is the partition function of the averaged Ising model in Theorem 9. Again, by (13), it suffices to given an algorithm with additive error $\frac{\varepsilon}{\mathrm{poly}(n)}$. We can use approximate counting oracles to estimate $\frac{\bar{Z}}{Z_\nu Z_\mu}$ with a relative error. Note that for any $\delta > 0$, $\mathrm{e}^{\pm\delta}$ relative error approximation is the same as $\pm O(\delta)\frac{\bar{Z}}{Z_\nu Z_\mu}$ additive error approximation. Since $\frac{\bar{Z}}{Z_\nu Z_\mu}\leq 1$ (as the divergence is non-negative), we can achieve $\pm O(\delta)$ additive error. Then the whole term $1-\frac{\bar{Z}}{Z_\nu Z_\mu}$ can be approximated within an additive error $\pm O(\delta)$. The problem is solved by setting $\delta=\frac{\varepsilon}{\mathrm{poly}(n)}$.

7 Proof of the hardness result

In this section, we only consider Ising model (G, J, h) with zero external field and interaction matrix with unified values. Formally, $h = \mathbf{0}$ and $J_{uv} = J_{vu} = \frac{\ln\beta}{2}$ for all $\{u, v\} \in E$. It is easy to see the probability of a configuration σ is proportional to $\beta^{m(\sigma)}$, where $m(\sigma)$ is the number of monochromatic edges $m(\sigma) = |\{u, v\} \in E, \sigma_u = \sigma_v|$. We denote it by (G, β).

Let $\alpha \geq 2$, $\Delta \geq 3$, and $\beta_\mu > \beta_\nu \geq \frac{\Delta-2}{\Delta}$ such that $(\frac{\beta_\nu}{\beta_\mu})^\alpha \beta_\mu < \frac{\Delta-2}{\Delta}$ be constant parameters in Theorem 6. We define a constant parameter

$$\beta = \left(\frac{\beta_\nu}{\beta_\mu}\right)^\alpha \beta_\mu < \frac{\Delta-2}{\Delta}. \tag{14}$$

The following hardness result for approximating partition function $Z = \sum_{\sigma\in\{\pm\}^V} \beta^{m(\sigma)}$ of the anti-ferromagnetic Ising model beyond the uniqueness threshold is well-known.

▶ **Lemma 20** ([33, 19]). *Fix $\Delta \geq 3$ and $0 < \beta < \frac{\Delta-2}{\Delta}$. There exists $c = c(\Delta, \beta) > 0$ such that unless* NP = RP*, there is no polynomial time randomized algorithm to approximate the partition function within a factor of e^{cn} with probability at least $2/3$ for the zero external field Ising model (G, β) on Δ-regular n-vertex graphs G.*

For any Δ-regular graph G, let $Z(G, \beta)$ denote the partition function of the zero external field Ising model (G, β), where β is defined in (14). We prove the following lemma.

▶ **Lemma 21.** *Let $\Delta, \alpha, \beta_\nu, \beta_\mu$ be constants. There exists a constant $C = C(\Delta, \alpha, \beta_\nu, \beta_\mu) > 0$ such that for any graph G, let μ and ν be the Gibbs distributions of the Ising models (G, β_ν) and (G, β_μ),*

$$\frac{1}{C} \cdot Z(G,\beta) \leq D_{\chi^\alpha}(\nu \,\|\, \mu) \cdot \frac{Z(G,\beta_\nu)^\alpha}{Z(G,\beta_\mu)^{\alpha-1}} \leq C \cdot Z(G,\beta).$$

Assume Lemma 21 holds. We prove Theorem 6.

Proof of Theorem 6. Fix parameters $\Delta, \alpha, \beta_\nu, \beta_\mu$. Let β be defined in (14).

Suppose there exists an FPRAS for $D_{\chi^\alpha}(\nu \,\|\, \mu)$. By run algorithms independently and take median, we can approximate $D_{\chi^\alpha}(\nu \,\|\, \mu)$ within the factor of 2 in polynomial time with probability at least 0.99. By our assumption, $\beta_\nu, \beta_\mu \geq \frac{\Delta-2}{\Delta}$. We can compute $Z(G, \beta_\nu)$ within the factor of 2 by applying the algorithms in [13] (if $\beta_\nu \in [\frac{\Delta-2}{\Delta}, 1]$) or the algorithms in [21] (if $\beta_\nu > 1$) in polynomial time, where the algorithm succeeds with probability at least 0.99. Similarly, we can compute an approximation to $Z(G, \beta_\mu)$.

By Lemma 21, we get a constant approximation to $Z(G, \beta)$ in polynomial time with probability at least 0.99. The hardness result follows from Lemma 20. ◀

The rest of this section is devoted to prove Lemma 21. To prove the lemma, we consider a middle term $\sum_{\sigma\in\Omega} \mu(\sigma)(\frac{\nu(\sigma)}{\mu(\sigma)} + 1)^\alpha$. The following result is a corollary of Lemma 12.

▶ **Corollary 22.** *There exists $C' = C'(\Delta, \alpha, \beta_\nu, \beta_\mu)$ such that*

$$\frac{1}{C'} \cdot \sum_{\sigma\in\{\pm\}^V} \mu(\sigma) \left(\frac{\nu(\sigma)}{\mu(\sigma)} + 1\right)^\alpha \leq D_{\chi^\alpha}(\nu \,\|\, \mu) \leq C' \cdot \sum_{\sigma\in\{\pm\}^V} \mu(\sigma) \left(\frac{\nu(\sigma)}{\mu(\sigma)} + 1\right)^\alpha.$$

Proof. The second inequality is trivial. For the first inequality, by Lemma 12, we can set θ in the lemma to be $\frac{1}{2}(\ln \beta_\nu - \ln \beta_\mu)$, which is a constant depending on β_ν, β_μ. Furthermore, the underlying graph G has constant degree Δ, both β_μ and β_ν are constants. Hence, by the conditional independence property of the Ising model, both ν and μ have a constant marginal lower bound $b = b(\Delta, \beta_\nu, \beta_\mu)$. Hence $B_{\alpha,b}(\theta) = \Theta_{\alpha,\beta_\nu,\beta_\mu,\Delta}(1)$ is a constant. ◀

With Corollary 22, Lemma 21 is a straightforward corollary of the following lemma.

▶ **Lemma 23.** *It holds that*

$$Z(G,\beta) \le \frac{Z(G,\beta_\nu)^\alpha}{Z(G,\beta_\mu)^{\alpha-1}} \sum_{\sigma\in\{\pm\}^V} \mu(\sigma)\left(\frac{\nu(\sigma)}{\mu(\sigma)}+1\right)^\alpha \le 3^\alpha \cdot Z(G,\beta).$$

Proof. To simplify the notation, we fix the graph G in the proof. We denote the partition function $Z(G,\beta_\nu) = Z_\nu$ and $Z(G,\beta_\mu) = Z_\mu$. The family of Ising models in $\mathcal{F}(\nu,\mu,\alpha)$ in (2) can be written as (G,β_k), where $\beta_k = \frac{\beta_\nu^k}{\beta_\mu^{k-1}}$ for $0 \le k \le \alpha$. We use Z_k to denote the partition function of (G,β_k). We remark that $Z_0 = Z_\mu$, $Z_1 = Z_\nu$, β in (14) is the same as β_α, and $Z(G,\beta) = Z_\alpha$.

We use the binomial expansion that

$$\sum_{\sigma\in\{\pm\}^V} \mu(\sigma)\left(\frac{\nu(\sigma)}{\mu(\sigma)}+1\right)^\alpha = \sum_{k=0}^{\alpha}\binom{\alpha}{k}\sum_{\sigma\in\{\pm\}^V}\frac{\nu^k(\sigma)}{\mu^{k-1}(\sigma)} = \sum_{k=0}^{\alpha}\binom{\alpha}{k}\frac{Z_\mu^{k-1}}{Z_\nu^k}\cdot Z_k. \tag{15}$$

We claim the following inequality holds. For any $0 \le k \le \alpha-1$,

$$\frac{Z_\mu^k}{Z_\nu^{k+1}} Z_{k+1} \ge \frac{1}{2}\frac{Z_\mu^{k-1}}{Z_\nu^k} Z_k. \tag{16}$$

We prove (16) later. With this conclusion, combining with (15), we know that

$$\frac{Z_\mu^{\alpha-1}}{Z_\nu^\alpha} Z_\alpha \le \sum_{k=0}^{\alpha}\binom{\alpha}{k}\frac{Z_\mu^{k-1}}{Z_\nu^k}\cdot Z_k \le \frac{Z_\mu^{\alpha-1}}{Z_\nu^\alpha} Z_\alpha \cdot \sum_{k=0}^{\alpha}\binom{\alpha}{k}2^{\alpha-k} = 3^\alpha \frac{Z_\mu^{\alpha-1}}{Z_\nu^\alpha} Z_\alpha, \tag{17}$$

where the first inequality is trivial and the second inequality is due to (16). Rearranging the terms in (17), we get the desired result.

Finally, we prove (16). Recall that $m(\sigma)$ denotes the number of monochromatic edges under the configuration σ. Since $\beta_\mu > \beta_\nu$, the sequence β_k is monotonically decreasing as $\beta_k = \frac{\beta_\nu^k}{\beta_\mu^{k-1}}$. Fix an unordered pair of configurations $\{\sigma,\tau\} \in \{\pm\}^V \times \{\pm\}^V$ (it may be possible that $\sigma = \tau$). Without loss of generality, we assume $m(\sigma) \ge m(\tau)$, otherwise we can swap the two configurations.

We have the following inequality for all $1 \le k \le \alpha-1$,

$$\begin{aligned}
&\beta_\mu^{m(\sigma)}\beta_{k+1}^{m(\tau)} + \beta_\mu^{m(\tau)}\beta_{k+1}^{m(\sigma)} \overset{\text{(I)}}{\ge} \beta_\mu^{m(\sigma)}\beta_{k+1}^{m(\tau)} \overset{\text{(II)}}{=} \left(\frac{\beta_\mu}{\beta_\nu}\right)^{m(\sigma)}\beta_\nu^{m(\sigma)}\cdot\left(\frac{\beta_\nu}{\beta_\mu}\right)^{m(\tau)}\beta_k^{m(\tau)} \\
&= \left(\frac{\beta_\mu}{\beta_\nu}\right)^{m(\sigma)-m(\tau)}\beta_\nu^{m(\sigma)}\beta_k^{m(\tau)} \overset{\text{(III)}}{\ge} \beta_\nu^{m(\sigma)}\beta_k^{m(\tau)} \overset{\text{(IV)}}{\ge} \frac{1}{2}\left(\beta_\nu^{m(\sigma)}\beta_k^{m(\tau)} + \beta_\nu^{m(\tau)}\beta_k^{m(\sigma)}\right),
\end{aligned} \tag{18}$$

where the inequality (I) throws the second term, and the equality (II) is due to the definition of β_k, the inequality (III) is due to the assumption $m(\sigma) \ge m(\tau)$ and $\beta_\mu > \beta_\nu$, and the last inequality (IV) is due $m(\sigma) \ge m(\tau)$ and $\beta_\nu = \beta_1 \ge \beta_k$ for $k \ge 1$.

If $k = 0$, note that $\beta_0 = \beta_\mu$ and $\beta_1 = \beta_\nu$, we have

$$\begin{aligned}
\beta_\mu^{m(\sigma)}\beta_{k+1}^{m(\tau)} + \beta_\mu^{m(\tau)}\beta_{k+1}^{m(\sigma)} &= \beta_\nu^{m(\sigma)}\beta_k^{m(\tau)} + \beta_\nu^{m(\tau)}\beta_k^{m(\sigma)} \\
&\ge \frac{1}{2}\left(\beta_\nu^{m(\sigma)}\beta_k^{m(\tau)} + \beta_\nu^{m(\tau)}\beta_k^{m(\sigma)}\right).
\end{aligned} \tag{19}$$

To prove (16), we need to prove $Z_\mu Z_{k+1} \geq \frac{1}{2} Z_\nu Z_k$ for all $0 \leq k \leq \alpha - 1$. Recall that the partition function $Z_k = \sum_{\sigma \in \{\pm\}^V} \beta_k^{m(\sigma)}$. We have the following inequality

$$\begin{aligned} Z_\mu Z_{k+1} &= \sum_{\sigma,\tau \in \{\pm\}^V} \beta_\mu^{m(\sigma)} \beta_{k+1}^{m(\tau)} \\ &= \sum_{\substack{\text{unordered pair}\{\sigma,\tau\} \\ \sigma \neq \tau}} \left(\beta_\mu^{m(\sigma)} \beta_{k+1}^{m(\tau)} + \beta_\mu^{m(\tau)} \beta_{k+1}^{m(\sigma)} \right) + \sum_{\sigma \in \{\pm\}^V} \beta_\mu^{m(\sigma)} \beta_{k+1}^{m(\sigma)} \\ (*) \quad &\geq \frac{1}{2} \sum_{\substack{\text{unordered pair}\{\sigma,\tau\} \\ \sigma \neq \tau}} \left(\beta_\nu^{m(\sigma)} \beta_k^{m(\tau)} + \beta_\nu^{m(\tau)} \beta_k^{m(\sigma)} \right) + \frac{1}{2} \sum_{\sigma \in \{\pm\}^V} \beta_\nu^{m(\sigma)} \beta_k^{m(\sigma)} \\ &= \frac{1}{2} Z_\nu Z_k, \end{aligned}$$

where the inequality $(*)$ is due to (18) and (19) and two inequalities (18) and (19) hold even if $\sigma = \tau$. This proves the inequality in (16). ◀

References

1. Amirali Abdullah, Ravi Kumar, Andrew McGregor, Sergei Vassilvitskii, and Suresh Venkatasubramanian. Sketching, embedding and dimensionality reduction in information theoretic spaces. In *AISTATS*, volume 51, pages 948–956. JMLR.org, 2016. URL: `http://proceedings.mlr.press/v51/abdullah16.html`.
2. Antoine Amarilli, Marcelo Arenas, YooJung Choi, Mikaël Monet, Guy Van den Broeck, and Benjie Wang. A circus of circuits: Connections between decision diagrams, circuits, and automata. *arXiv preprint*, 2024. `doi:10.48550/arXiv.2404.09674`.
3. Arnab Bhattacharyya, Weiming Feng, and Piyush Srivastava. Approximating the total variation distance between Gaussians. In *AISTATS*, volume 258 of *Proceedings of Machine Learning Research*, pages 1846–1854. PMLR, 2025. URL: `https://proceedings.mlr.press/v258/bhattacharyya25a.html`.
4. Arnab Bhattacharyya, Sutanu Gayen, Kuldeep S. Meel, Dimitrios Myrisiotis, A. Pavan, and N. V. Vinodchandran. On approximating total variation distance. In *IJCAI*, pages 3479–3487. ijcai.org, 2023. `doi:10.24963/IJCAI.2023/387`.
5. Arnab Bhattacharyya, Sutanu Gayen, Kuldeep S. Meel, Dimitrios Myrisiotis, A. Pavan, and N. V. Vinodchandran. Total variation distance meets probabilistic inference. In *ICML*. OpenReview.net, 2024. URL: `https://openreview.net/forum?id=6OSLjErBhh`.
6. Arnab Bhattacharyya, Sutanu Gayen, Kuldeep S. Meel, Dimitrios Myrisiotis, A. Pavan, and N. V. Vinodchandran. Computational explorations of total variation distance. In *ICLR*. OpenReview.net, 2025. URL: `https://openreview.net/forum?id=xak8c9l1nu`.
7. Arnab Bhattacharyya, Sutanu Gayen, Kuldeep S Meel, Dimitrios Myrisiotis, A Pavan, and NV Vinodchandran. Algorithms and hardness for estimating statistical similarity. *arXiv preprint*, 2025. `arXiv:2502.10527`.
8. Arnab Bhattacharyya, Sutanu Gayen, Kuldeep S. Meel, and N. V. Vinodchandran. Efficient distance approximation for structured high-dimensional distributions via learning. In *NeurIPS*, 2020.
9. Antonio Blanca, Zongchen Chen, Daniel Stefankovic, and Eric Vigoda. Complexity of high-dimensional identity testing with coordinate conditional sampling. *ACM Trans. Algorithms*, 21(1):7:1–7:58, 2025. `doi:10.1145/3686799`.
10. Guy Bresler. Efficiently learning Ising models on arbitrary graphs. In *STOC*, pages 771–782. ACM, 2015. `doi:10.1145/2746539.2746631`.
11. Clément L. Canonne. *A Survey on Distribution Testing: Your Data is Big. But is it Blue?* Number 9 in Graduate Surveys. Theory of Computing Library, 2020. `doi:10.4086/toc.gs.2020.009`.

12 Clément L. Canonne. Topics and techniques in distribution testing: A biased but representative sample. *Found. Trends Commun. Inf. Theory*, 19(6):1032–1198, 2022. doi:10.1561/0100000114.

13 Xiaoyu Chen, Zongchen Chen, Yitong Yin, and Xinyuan Zhang. Rapid mixing at the uniqueness threshold. In *STOC*, pages 879–890. ACM, 2025. doi:10.1145/3717823.3718260.

14 Zongchen Chen, Kuikui Liu, and Eric Vigoda. Optimal mixing of Glauber dynamics: Entropy factorization via high-dimensional expansion. In *STOC*, pages 1537–1550. ACM, 2021. doi:10.1145/3406325.3451035.

15 Constantinos Daskalakis, Nishanth Dikkala, and Gautam Kamath. Testing Ising models. In *SODA*, pages 1989–2007. SIAM, 2018. doi:10.1137/1.9781611975031.130.

16 Weiming Feng, Heng Guo, Mark Jerrum, and Jiaheng Wang. A simple polynomial-time approximation algorithm for the total variation distance between two product distributions. *TheoretiCS*, 2, 2023. doi:10.46298/THEORETICS.23.7.

17 Weiming Feng, Hongyang Liu, and Minji Yang. Approximating the total variation distance between spin systems. In *COLT*, volume 291 of *Proceedings of Machine Learning Research*, pages 1974–2025. PMLR, 2025. URL: https://proceedings.mlr.press/v291/feng25a.html.

18 Weiming Feng, Liqiang Liu, and Tianren Liu. On deterministically approximating total variation distance. In *SODA*, pages 1766–1791. SIAM, 2024. doi:10.1137/1.9781611977912.70.

19 Andreas Galanis, Daniel Štefankovič, and Eric Vigoda. Inapproximability of the partition function for the antiferromagnetic Ising and hard-core models. *Combin. Probab. Comput.*, 25(4):500–559, 2016. doi:10.1017/S0963548315000401.

20 William Gay, William He, Nicholas Kocurek, and Ryan O'Donnell. Sampling and identity-testing without approximate tensorization of entropy. *arXiv preprint arXiv:2506.23456*, 2025. doi:10.48550/arXiv.2506.23456.

21 Mark Jerrum and Alistair Sinclair. Polynomial-time approximation algorithms for the Ising model. *SIAM J. Comput.*, 22(5):1087–1116, 1993. doi:10.1137/0222066.

22 Takafumi Kanamori, Taiji Suzuki, and Masashi Sugiyama. f-divergence estimation and two-sample homogeneity test under semiparametric density-ratio models. *IEEE Trans. Inf. Theory*, 58(2):708–720, 2012. doi:10.1109/TIT.2011.2163380.

23 Stefan Kiefer. On computing the total variation distance of hidden Markov models. In *ICALP*, volume 107 of *LIPIcs*, pages 130:1–130:13. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2018. doi:10.4230/LIPIcs.ICALP.2018.130.

24 Aryeh Kontorovich. On the tensorization of the variational distance. *Electron. Commun. Probab.*, 30:Paper No. 32, 10, 2025.

25 Aryeh Kontorovich and Ariel Avital. Sharp bounds on aggregate expert error. In *ALT*, volume 272 of *Proceedings of Machine Learning Research*, pages 653–663. PMLR, 2025. URL: https://proceedings.mlr.press/v272/kontorovich25a.html.

26 Loong Kuan Lee, Nico Piatkowski, François Petitjean, and Geoffrey I. Webb. Computing divergences between discrete decomposable models. In *AAAI*, pages 12243–12251. AAAI Press, 2023. doi:10.1609/AAAI.V37I10.26443.

27 David A Levin and Yuval Peres. *Markov chains and mixing times*, volume 107. American Mathematical Soc., 2017.

28 Liang Li, Pinyan Lu, and Yitong Yin. Correlation decay up to uniqueness in spin systems. In *SODA*, pages 67–84. SIAM, 2013. doi:10.1137/1.9781611973105.5.

29 Jingcheng Liu, Alistair Sinclair, and Piyush Srivastava. The Ising partition function: Zeros and deterministic approximation. *Journal of Statistical Physics*, 174(2):287–315, 2019.

30 Barnabás Póczos and Jeff G. Schneider. On the estimation of α-divergences. In *AISTATS*, volume 15 of *JMLR Proceedings*, pages 609–617. JMLR.org, 2011.

31 Paul K. Rubenstein, Olivier Bousquet, Josip Djolonga, Carlos Riquelme, and Ilya O. Tolstikhin. Practical and consistent estimation of f-divergences. In *NeurIPS*, pages 4072–4082, 2019. URL: https://proceedings.neurips.cc/paper/2019/hash/3147da8ab4a0437c15ef51a5cc7f2dc4-Abstract.html.

32 Amit Sahai and Salil Vadhan. A complete problem for statistical zero knowledge. *J. ACM*, 50(2):196–249, 2003. doi:10.1145/636865.636868.

33 Allan Sly and Nike Sun. Counting in two-spin models on d-regular graphs. *Ann. Probab.*, 42(6):2383–2416, 2014.

34 Sreejith Sreekumar and Ziv Goldfeld. Neural estimation of statistical divergences. *J. Mach. Learn. Res.*, 23:126:1–126:75, 2022. URL: https://jmlr.org/papers/v23/21-1212.html.