Recent advances in approximating f-divergences between two Ising models

Weiming Feng

(Hong Kong U)

This talk is based on two joint works with

Hongyang Liu (Nanjing U), **Minji Yang** (Hong Kong U), **Yucheng Fu** (Hong Kong U)

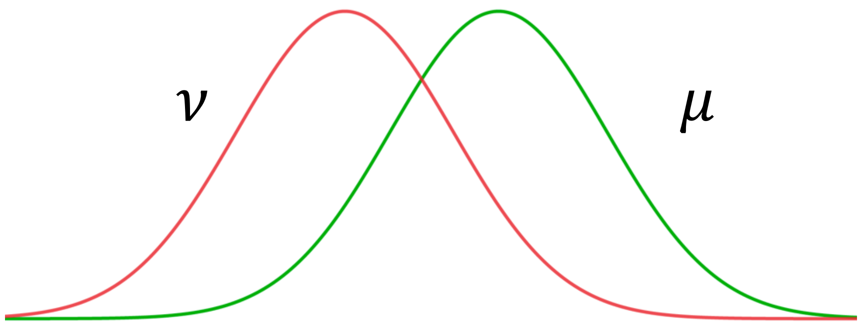
Seminar Talk

Nanyang Technological University, 25 Sep 2025

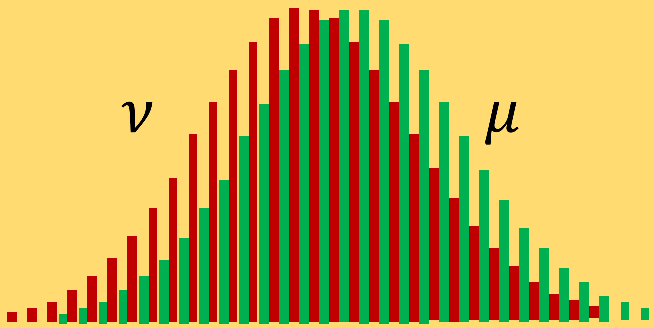
Difference between two distributions

Input: two distributions ν and μ over state space Ω

Question: how to measure the difference between ν and μ ?



continuous distributions

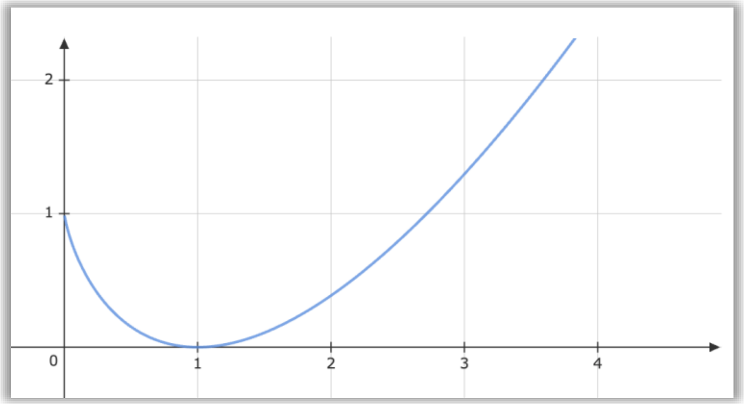


discrete distributions

The f-divergence between two distributions

Let $f: \mathbb{R}_+ \to \mathbb{R}_{\geq 0}$ be a ***convex*** function s.t. $f(1) = 0$

f-divergence: $D_f(\nu \parallel \mu) = \mathbb{E}_{X\sim\mu}\left[f\left(\frac{\nu(X)}{\mu(X)}\right)\right]$



- χ^α divergence $f(x) = \frac{1}{2}|x-1|^\alpha$ for $\alpha \geq 1$

  $\alpha = 1$ gives *total variation (TV) distance* $D_{TV}(\nu \parallel \mu) = \frac{1}{2}\sum_{x\in\Omega}|\nu(x) - \mu(x)|$

- α divergence $f(x) = \frac{x^\alpha - \alpha x - (1-\alpha)}{\alpha(\alpha-1)}$ for $\alpha \in \mathbb{R}$

  $\alpha = 1$ gives *Kullback–Leibler (KL) divergence* $D_{KL}(\nu \parallel \mu) = \frac{1}{2}\sum_{x\in\Omega}\nu(x)\ln\frac{\nu(x)}{\mu(x)}$

  $\alpha = 0$ gives *Rényi divergence* $D_R(\nu \parallel \mu) = D_{KL}(\mu \parallel \nu)$

- Squared Hellinger distance $f(x) = \frac{1}{2}(\sqrt{x} - 1)^2$

Compute value the f-divergence

- **Input**: descriptions of two distributions ν, μ over Ω and a function f
- **Output**: the f-divergence $D_f(\nu \parallel \mu)$ between ν and μ

for instance, TV distance: $D_{TV}(\nu \parallel \mu) = \frac{1}{2}\sum_{x\in\Omega}|\nu(x) - \mu(x)|$

Trivial algorithm: enumerate all $x \in \Omega$ and add $\frac{1}{2}|\nu(x) - \mu(x)|$ together

Challenge:ν and μ have ***succinct descriptions (structured distribution)***

- $|\Omega|$ can be ***exponentially large*** w.r.t. the size of input
- It can be ***challenging*** to evaluate the value of $\nu(x)$ and $\mu(x)$

Examples: probabilistic graphical models, probabilistic circuits

Warm-up: Product distributions

Product distribution μ over $\{-,+\}^n$

$$\mu = \mu_1 \times \mu_2 \times \cdots \times \mu_n$$

μ_i is a distribution over $\{-,+\}$.

- μ can be described by n marginals
- Size the ***input*** $2n$
- Size of ***sample space*** $|\Omega| = 2^n$

Random sample $X = (X_1, X_2, \dots, X_n) \sim \mu$

$X \in \{-,+\}^n$: n-dimensional random vector

$X_i \in \{-,+\}$: independent sample from μ_i

Compute TV distance between product distributions

[Bhattacharyya, Gayen, Meel, Myrisiotis,Pavan,Vinodchandran, 2022]

- **Input**: distributions $\{\nu_i, \mu_i | 1 \le i \le n\}$ specifying ν and μ over $\{\pm\}^n$
- **Output**: the total variation distance between ν and μ

Results for computing TV distance between product distributions

Theorem [BGMMPV22]: the exact computing is **#P-complete.**

FPTAS (Full Poly-time Approximation Scheme)

A *deterministic* algorithm outputs a $\hat{d}$ in time $\mathrm{poly}(n, 1/\epsilon)$

$$(1-\epsilon)D_{TV}(\nu \parallel \mu) \leq \hat{d} \leq (1+\epsilon)D_{TV}(\nu \parallel \mu)$$

FPRAS (Full Poly-time Randomised Approximation Scheme)

A *randomized* algorithm outputs a random $\hat{d}$ in time $\mathrm{poly}(n, 1/\epsilon)$

$$\Pr\left[(1-\epsilon)D_{TV}(\nu \parallel \mu) \leq \hat{d} \leq (1+\epsilon)D_{TV}(\nu \parallel \mu)\right] \geq 2/3$$

Results for computing TV distance between product distributions

Theorem [BGMMPV22]: the exact computing is **#P-complete.**

Theorem [BGMMPV22] ***FPTAS/FPRAS*** exists ***one of*** the following condition holds

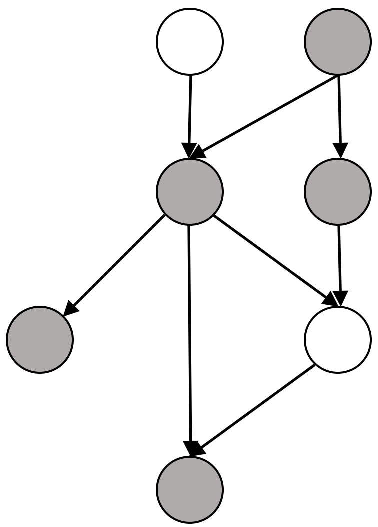
- μ has *constant number* of distinct marginals (e.g. uniform distribution over $\{-,+\}^n$)
- $\forall i \in [n], \underbrace{\nu_i(1) \geq \mu_i(1)}_{\textit{break symmetry}}$ and $\underbrace{\nu_i(1) \geq 1/2}_{\textit{lower bound}}$

Theorem [**F**GJW23 and **F**LL23]

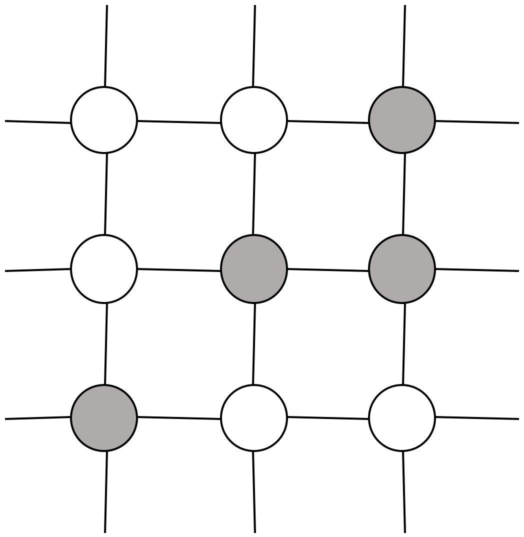
General product distributions ν, μ and error bound $0 < \epsilon < 1$

- FPTAS running time: $\tilde{O}\left(\frac{n^2}{\epsilon} \log \frac{1}{D_{TV}(\nu \parallel \mu)}\right)$
- FPRAS running time : $\tilde{O}\left(\frac{n^2}{\epsilon^2}\right)$

Beyond product distribution: graphical models



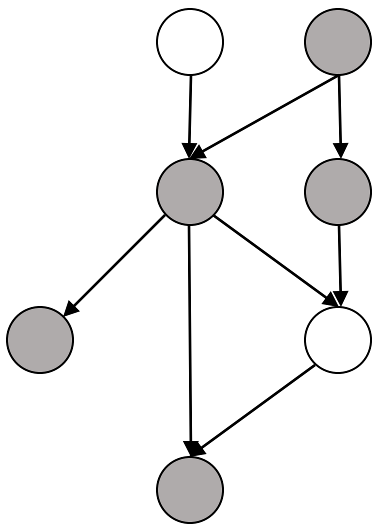
Bayesian network



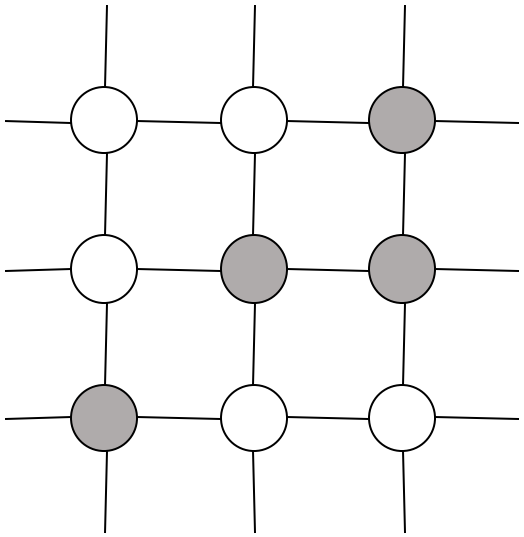
Ising models

- ***Vertices*** are ***random variables*** and ***edges*** model ***local interactions***
- ***Graphical models*** define joint distributions with ***complex correlations***

Beyond product distribution: graphical models



Bayesian network



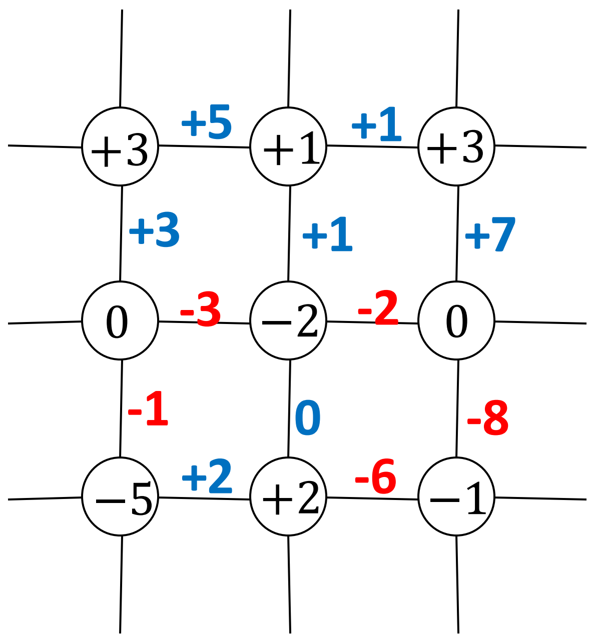
Ising models

FPRAS for two ***Bayesian networks*** with bounded treewidth
[Bhattacharyya, Gayen, Meel, Myrisiotis, Pavan, Vinodchandran 2025]

Focus of this talk

Ising model (G, J, h)

Graph $G = (V, E)$, weighted adjacent matrix $J \in \mathbb{R}^{V \times V}$, and external fields $h \in \mathbb{R}^V$



Ising model (G, J, h)

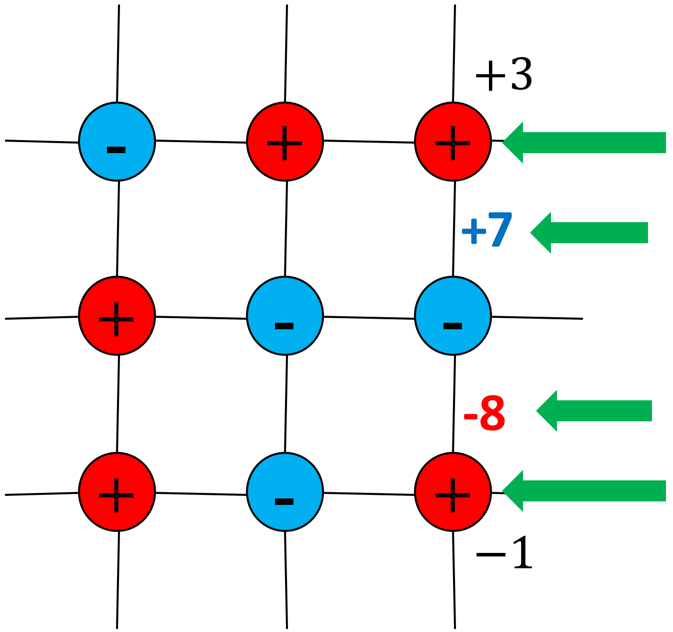
Graph $G = (V, E)$, weighted adjacent matrix $J \in \mathbb{R}^{V \times V}$, and external fields $h \in \mathbb{R}^{V}$

$\forall$ configurtation
$\sigma \in \{-,+\}^V$

Weight $$w(\sigma) = \exp\left(\frac{\sigma^T J \sigma}{2}\right) = \exp\left(\sum_{\{u,v\}\in E} \sigma_u \sigma_v J_{uv} + \sum_{v \in V} \sigma_v h_v\right)$$

Probability $$\mu(\sigma) = \frac{w(\sigma)}{Z}$$

Partition Function $$Z = \sum_{x \in \{-1,+1\}^V} w(x)$$



$h(v) > 0$: σ_v prefer to take the ***+*** value

$J(u, v) > 0$: σ_u and σ_v prefer to take the ***same*** value

$J(u, v) < 0$: σ_u and σ_v prefer to take ***different*** values

$h(v) < 0$: σ_v prefer to take the **-** value

Ising model (G, J, h)

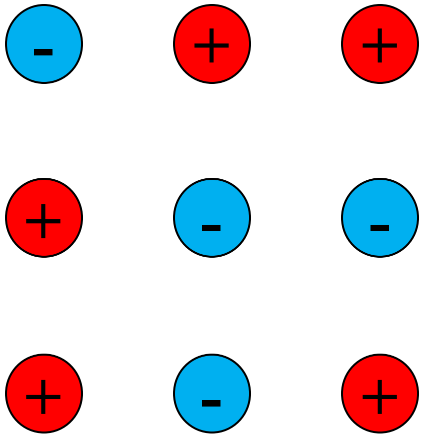
Graph $G = (V, E)$, weighted adjacent matrix $J \in \mathbb{R}^{V \times V}$, and external fields $h \in \mathbb{R}^V$

∀ configurtation
$\sigma \in \{-,+\}^V$

Weight $$w(\sigma) = \exp\left(\frac{\sigma^T J \sigma}{2}\right) = \exp\left(\sum_{\{u,v\} \in E} \sigma_u \sigma_v J_{uv} + \sum_{v \in V} \sigma_v h_v\right)$$

Probability $$\mu(\sigma) = \frac{w(\sigma)}{Z}$$

Partition Function $$Z = \sum_{x \in \{-1,+1\}^V} w(x)$$



Empty Graph
$J = 0$

→

Product Distribution

Ising model (G, J, h)

Graph $G = (V, E)$, weighted adjacent matrix $J \in \mathbb{R}^{V\times V}$, and external fields $h \in \mathbb{R}^V$

∀ configurtation $\sigma \in \{-,+\}^V$

Weight $$w(\sigma) = \exp\left(\frac{\sigma^T J\sigma}{2}\right) = \exp\left(\sum_{\{u,v\}\in E} \sigma_u\sigma_v J_{uv} + \sum_{v\in V} \sigma_v h_v\right)$$

Probability $$\mu(\sigma) = \frac{w(\sigma)}{Z}$$

Partition Function $$Z = \sum_{x\in\{-1,+1\}^V} w(x)$$

Simplified Ising model (G, β)

- All edges have a ***unified value*** in interaction matrix $J(u, v) = \beta$ for all $\{u, v\} \in E$
- All vertices have ***zero external field*** $h(v) = 0$ for all $v \in V$

$$\mu(\sigma) \propto \prod_{\{u,v\}\in E} \exp(\sigma_u\sigma_v\beta) \propto \prod_{\{u,v\}\in E:\sigma_u=\sigma_v} \exp(2\beta) = \exp(2\beta \cdot \#\text{monochromatic edges})$$

Approximating the χ^α-divergence between two Ising models

- **Input**: ***two Ising models*** (G, J^ν, h^ν) and (G, J^μ, h^μ) defining ν and μ

  integer parameter α and ***error bound*** $\epsilon > 0$

- **Output**: $D \in (1+\varepsilon) \cdot D_{\chi^\alpha}(\nu \parallel \mu)$ for χ^α-divergence $D_{\chi^\alpha}(\nu \parallel \mu)$

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

Reduction

Sampling: draw ***random sample*** $X \sim \mu$ from the law of $\mu = \mathrm{Ising}(G, J, h)$

Approximate Counting: compute an estimate $\hat{Z}$ the ***partition function*** of $\mathrm{Ising}(G, J, h)$

$$(1-\varepsilon)Z \le \hat{Z} \le (1+\varepsilon)Z$$

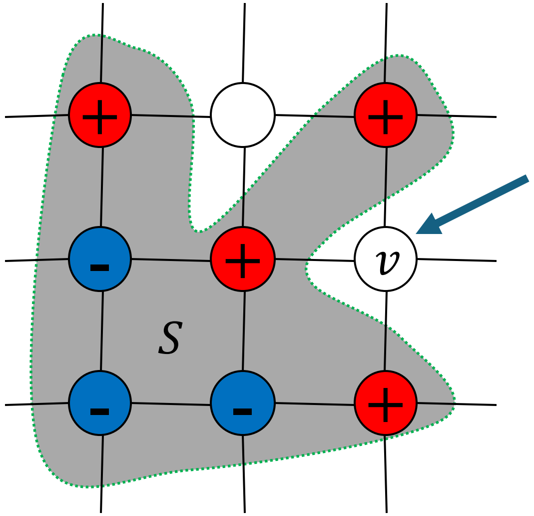
Our result: total variation distance ($\alpha = 1$)

Definition **Marginal lower bound for Ising model**

For any subset $S \subseteq V$, any vertex $v \in V \backslash S$, any pinning $\tau \in \{-1,+1\}^S$,

$$\forall c \in \{-1,+1\}, \qquad \mu_v(\, c \mid \tau\,) = \Omega(1)$$

*Under **any conditional**, the **marginal distribution** on one vertex **cannot be too biased***



marginal distribution at v

$$\mu_v(+|\tau), \mu_v(-|\,\tau\,) = \Omega(1)$$

The assumption also appeared in ***learning*** [Bresler15], ***sampling and counting*** [CLV21]

Our result: total variation distance ($\alpha = 1$)

Definition **Marginal lower bound for Ising model**

For any subset $S \subseteq V$, any vertex $v \in V \setminus S$, any pinning $\tau \in \{-1,+1\}^S$,

$$\forall c \in \{-1,+1\}, \qquad \mu_v(\, c \mid \tau \,) = \Omega(1)$$

*Under **any conditional**, the **marginal distribution** on one vertex **cannot be too biased***

Theorem: total variation distance ($\alpha = 1$) [F, Liu, Yang, 2025]

Two Ising models $\nu = \mathrm{Ising}(G, J^\nu, h^\nu)$ and $\mu = \mathrm{Ising}(G, J^\mu, h^\mu)$ with marginal lower bound

Two models both admit $\mathrm{poly}(n/\epsilon)$-time algos for
- sampling
- approximate counting

$\Rightarrow$

$\mathrm{poly}(n/\epsilon)$-time algorithms for approximate $D_{TV}(\nu \parallel \mu)$

- Our result also work for ***other graphical models***
- The marginal lower bound can be ***removed*** in some graphical models

Simplified Ising model (G, β)

- All edges have a ***unified value*** in interaction matrix $J(u,v) = \beta$ for all $\{u,v\} \in E$
- All vertices have ***zero external field*** $h(v) = 0$ for all $v \in V$

$$\mu(\sigma) \propto \prod_{\{u,v\}\in E} \exp(\sigma_u \sigma_v \beta) \propto \prod_{\{u,v\}\in E:\sigma_u=\sigma_v} \exp(2\beta)$$

Computational Phase Transition for Sampling and Approx. Counting

Max degree of graph Δ. ***Uniqueness threshold*** $\beta_c = \beta_c(\Delta) < 0$ s.t. $\exp(2\beta_c) = \frac{\Delta-2}{\Delta}$

- ***Polynomial time*** sampling and approx. counting if $\beta \geq \beta_c$ [JS93, CCYZ25]
- Sampling and approx. counting is ***hard*** (unless NP=RP) if $\beta < \beta_c$ [SS14, GŠV16]

$$\beta_c = \frac{1}{2}\ln\left(\frac{\Delta-2}{\Delta}\right)$$

Simplified Ising model (G, β) with constant β, Δ $\Longrightarrow$ marginal lower bound

Corollary [F, Liu, Yang, 2025]

Two Ising models $\nu = \mathrm{Ising}(G, \beta_\nu)$ and $\mu = \mathrm{Ising}(G, \beta_\mu)$

Two models ***both above the threshold***
$$\min\{\beta_\nu, \beta_\mu\} \geq \beta_c(\Delta)$$

$\Longrightarrow$

FPRAS for the
TV distance $D_{TV}(\nu \,\|\, \mu)$

Hardness result [Bhattacharyya, Gayen, Meel, Myrisiotis, Pavan, Vinodchandran, 2025]

Two Ising models $\nu = \mathrm{Ising}(G, \beta_\nu)$ and $\mu = \mathrm{Ising}(G, \beta_\mu)$

Two models ***both below the threshold***
$$\max\{\beta_\nu, \beta_\mu\} < \beta_c(\Delta)$$

$\Longrightarrow$

No FPRAS for TV-distance
unless NP=RP

Our result: χ^{α}-divergence

- **Input**: an integer $\alpha \geq 1$, two Ising models (G, J^{ν}, h^{ν}) and (G, J^{μ}, h^{μ}), an error bound $\epsilon > 0$
- **Output**: $D \in (1 \pm \epsilon) D_{\chi^{\alpha}}(\nu \parallel \mu)$ for χ^{α}-divergence

$$D_{\chi^{\alpha}}(\nu \parallel \mu) = \sum_{x \in \{\pm\}^V} \mu(x) \left| 1 - \frac{\nu(x)}{\mu(x)} \right|^{\alpha}$$

Our result: χ^α-divergence

Theorem: approximation algorithm [F and Fu, 2025]

Two Ising models $\nu = \mathrm{Ising}(G, J^\nu, h^\nu)$ and $\mu = \mathrm{Ising}(G, J^\mu, h^\mu)$ with marginal lower bound

A ***family*** of Ising models $\mathcal{F} = \{(G, J^{(k)}, h^{(k)}) \mid \text{integer } 0 \le k \le \alpha\}$, where

$$J^{(k)} = kJ^\nu - (k-1)J^\mu$$

$$h^{(k)} = kh^\nu - (k-1)h^\mu$$

All Ising models in $\mathcal{F}$ admit $\mathrm{poly}(n/\epsilon)$-time algos for
- sampling
- approximate counting

$\Rightarrow$

$\mathrm{poly}(n/\epsilon)$-time algorithms for approximate $D_{\chi^\alpha}(\nu \parallel \mu)$

χ^α-divergence with $\alpha = 1$ $\Rightarrow$ $D_{TV} = D_{\chi^\alpha}$

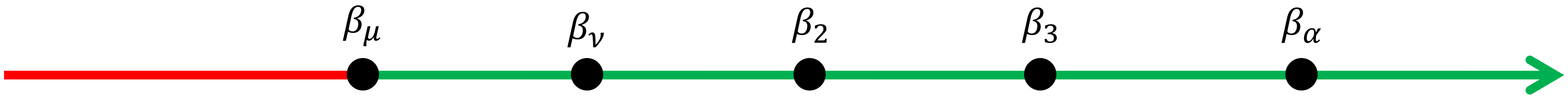
$\mathcal{F}$ ***only*** contains ***two input*** Ising ν and μ $\Rightarrow$ Recover TV-distance result

Corollary: simplified Ising model [F, Fu, 2025]

Two ***zero field*** Ising models (G, β_ν) and (G, β_μ) with ***unified non-zero values*** in interaction matrices

$$\mathcal{F} = \{(G, \beta_k) \mid \text{integer } 0 \le k \le \alpha\}, \text{ with } \beta_k = \beta_\mu + k(\beta_\nu - \beta_\mu) \text{ (note } \beta_0 = \beta_\mu \text{ and } \beta_1 = \beta_\nu\text{)}$$

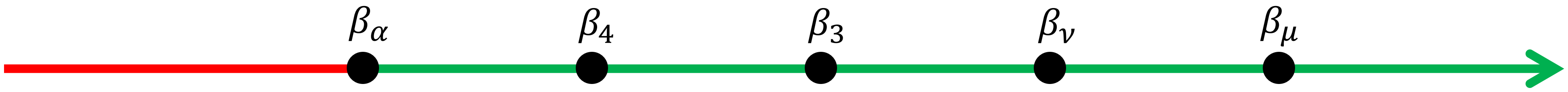
Case $\beta_\nu \ge \beta_\mu$: poly-time algorithm for χ^α-divergence exist if $\beta_\mu \ge \beta_c$



$$\beta_c = \frac{1}{2}\ln\left(\frac{\Delta - 2}{\Delta}\right)$$

Is it necessary to assume oracles for all Ising models in $\mathcal{F}$?

Case $\beta_\mu > \beta_\nu$: poly-time algorithm for χ^α-divergence exist if $\beta_\alpha = \beta_\mu + k(\beta_\nu - \beta_\mu) \ge \beta_c$



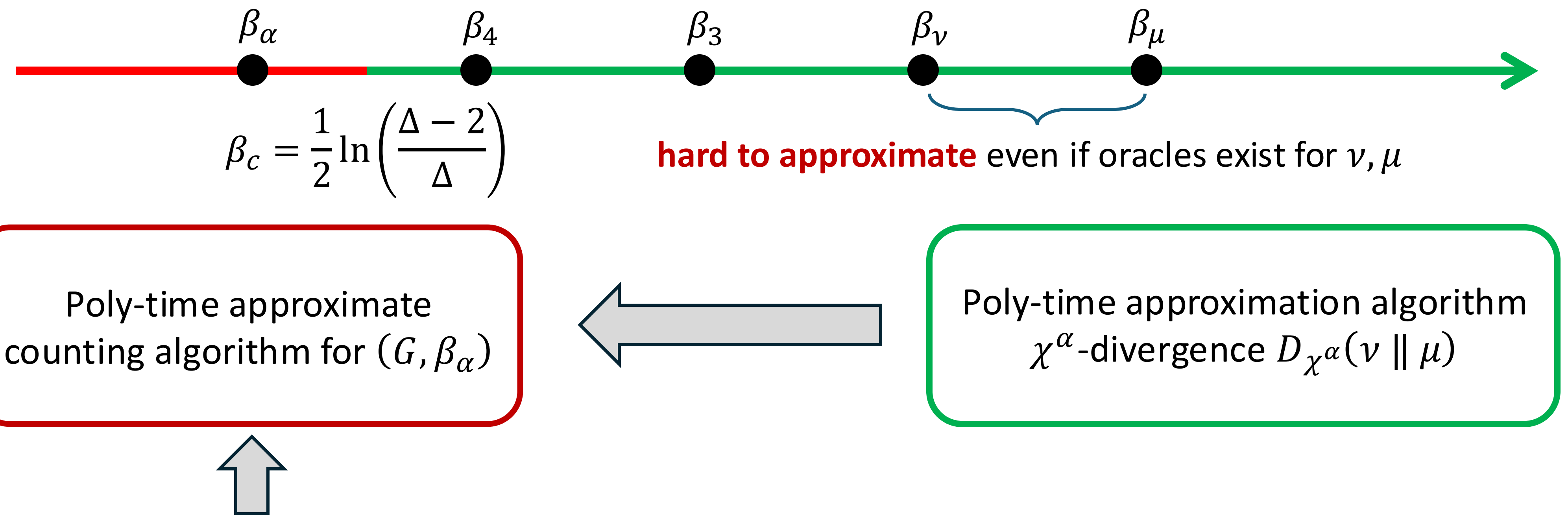
$$\beta_c = \frac{1}{2}\ln\left(\frac{\Delta - 2}{\Delta}\right)$$

Theorem: hardness of approximation [F. and Fu, 2025]

Fix integers $\alpha \geq 2$ and $\Delta \geq 3$. Fix $\beta_\mu > \beta_\nu \geq \beta_c(\Delta)$ such that

$$\beta_\alpha = \beta_\mu + k(\beta_\nu - \beta_\mu) < \beta_c(\Delta).$$

Unless ***NP=RP***, ***no FPRAS*** for χ^α-divergence between (G, β_ν) and (G, β_μ) on Δ-regular graphs G



Approximate counting is ***hard*** for $\beta_\alpha < \beta_c$ [Sly and Sun 14, Galanis, Štefankovič and Vigoda 16]

Theorem: hardness of approximation [F. and Fu, 2025]

Fix integers $\alpha \geq 2$ and $\Delta \geq 3$. Fix $\beta_\mu > \beta_\nu \geq \beta_c(\Delta)$ such that

$$\beta_\alpha = \beta_\mu + \alpha(\beta_\nu - \beta_\mu) < \beta_c(\Delta).$$

Unless ***NP=RP***, ***no FPRAS*** for χ^α-divergence between (G, β_ν) and (G, β_μ) on Δ-regular graphs G

Computational Phase Transition

Corollary: approximation algorithms [F. and Fu, 2025]

Fix integers $\alpha \geq 2$ and $\Delta \geq 3$. Fix $\beta_\mu > \beta_\nu \geq \beta_c(\Delta)$ such that

$$\beta_\alpha = \beta_\mu + \alpha(\beta_\nu - \beta_\mu) \geq \beta_c(\Delta).$$

FPRAS exists for χ^α-divergence between (G, β_ν) and (G, β_μ) on graph G with max degree Δ

Summary of algorithmic results

Divergence	Function f	Existence of oracles for sampling / counting
χ^α for $\alpha \in \mathbb{N}$	$f(x) = \frac{1}{2}\lvert x-1\rvert^\alpha$	$\{(G, J^{(k)}, h^{(k)}) \mid 0 \le k \le \alpha\}$
α-divergence for $\alpha \neq 0,1$	$f(x) = \frac{t^\alpha - \alpha t - (1-\alpha)}{\alpha(\alpha-1)}$	$\{(G, J^{(k)}, h^{(k)}) \mid k = 0,1,\alpha\}$
Kullback–Leibler Rényi Jensen-Shannon	$f(x) = x\ln x - x + 1$ $f(x) = -\ln x + x - 1$ $f(x) = \frac{1}{2}\left(x\ln x - (x+1)\ln\frac{x+1}{2}\right)$	$(G, J^{(\nu)}, h^{(\nu)})$ and $(G, J^{(\mu)}, h^{(\mu)})$
Squared Hellinger	$f(x) = \frac{1}{2}(\sqrt{x}-1)^2$	$(G, J^{(\nu)}, h^{(\nu)}), (G, J^{(\mu)}, h^{(\mu)})$ and $\left(G, \frac{J^{(\nu)}+J^{(\mu)}}{2}, \frac{h^{(\nu)}+h^{(\mu)}}{2}\right)$

Algorithm for χ^α-divergence between two Ising models

Theorem: approximation algorithm [F and Fu, 2025]

Two Ising models $\nu = \mathrm{Ising}(G, J^\nu, h^\nu)$ and $\mu = \mathrm{Ising}(G, J^\mu, h^\mu)$ with marginal lower bound

A ***family*** of Ising models $\mathcal{F} = \{(G, J^{(k)}, h^{(k)}) \mid \text{integer } 0 \le k \le \alpha\}$, where

$$J^{(k)} = kJ^\nu - (k-1)J^\mu$$

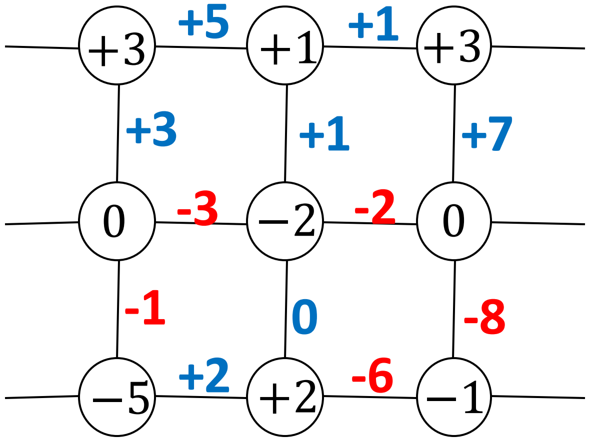
$$h^{(k)} = kh^\nu - (k-1)h^\mu$$

All Ising models in $\mathcal{F}$ admit $\mathrm{poly}(n/\epsilon)$-time algos for
- sampling
- approximate counting

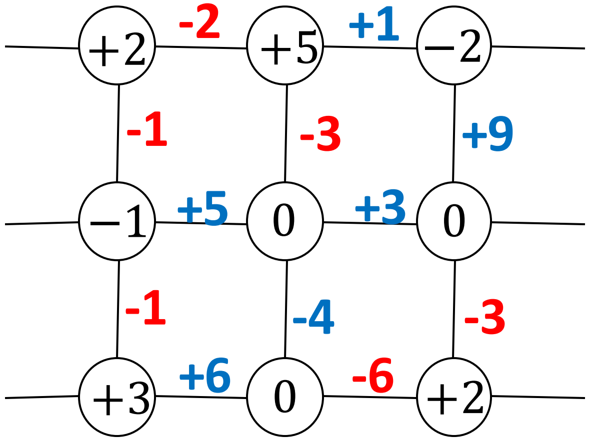
$\Rightarrow$

$\mathrm{poly}(n/\epsilon)$-time algorithms for approximate $D_{\chi^\alpha}(\nu \,\|\, \mu)$

Parameter distance



(G, J^{ν}, h^{ν})



(G, J^{μ}, h^{μ})

$$d_{\text{par}}(\nu, \mu) = \max\{\text{edge diff}, \text{vertex diff}\}$$

- Edge diff: $\max_{e \in E} |J^{\nu}(e) - J^{\mu}(e)|$
- Vertex diff: $\max_{v \in V} \frac{|h^{\nu}(v) - h^{\mu}(v)|}{\deg_G(v) + 1}$

What is the relation between ***parameter distance*** and ***TV-distance/χ^{α}-divergence***?

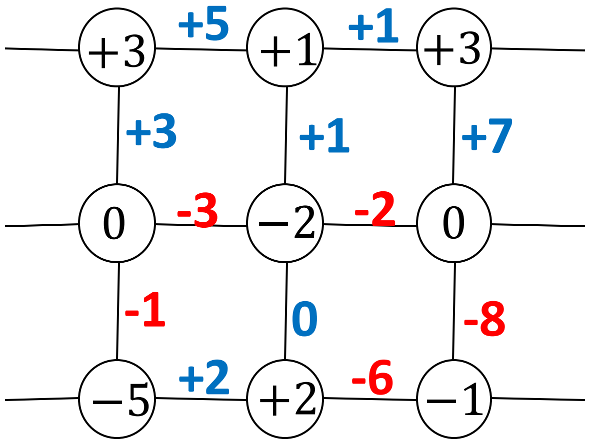
$d_{\text{par}}(\nu, \mu)$ is small enough

$D_{TV}(\nu \,||\, \mu)$ is small

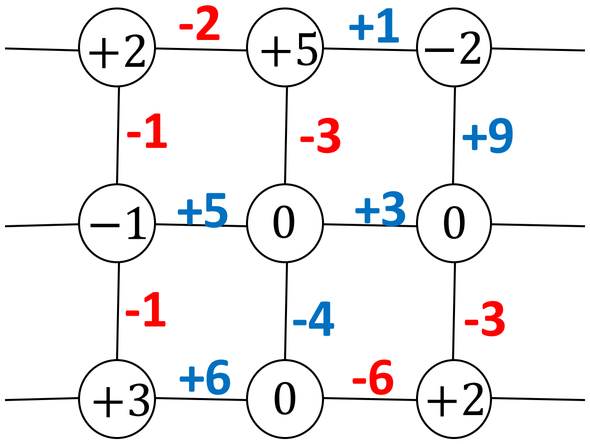
If two Ising model are different, then we can blame it on a node or an edge

a similar observation (but with different definitions of distances) was made in
"*Test Ising Models*" [Daskalakis, Dikkala and Kamath 19]

Parameter distance



(G, J^{ν}, h^{ν})



(G, J^{μ}, h^{μ})

$d_{\text{par}}(\nu, \mu) = \max\{$edge diff, vertex diff$\}$

- Edge diff: $\max_{e \in E} |J^{\nu}(e) - J^{\mu}(e)|$
- Vertex diff: $\max_{v \in V} \frac{|h^{\nu}(v) - h^{\mu}(v)|}{\deg_G(v) + 1}$

What is the relation between ***parameter distance*** and ***TV-distance/χ^{α}-divergence***?

$d_{\text{par}}(\nu, \mu)$ is ***large***

marginal lower bound $\Longrightarrow$

[Feng, Liu and Yang 2025]

$D_{TV}(\nu \,||\, \mu)$ is ***large***

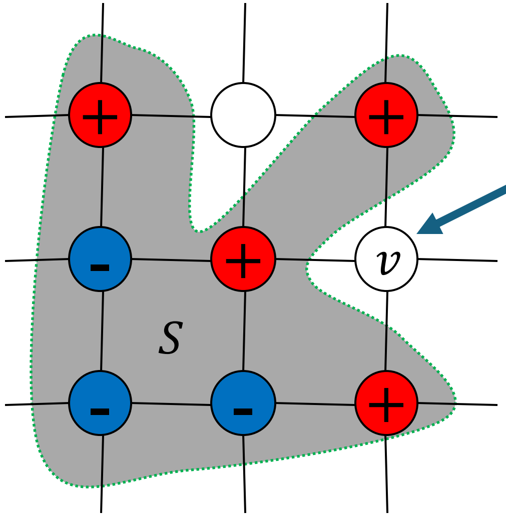
Our result: total variation distance ($\alpha = 1$)

Definition **Marginal lower bound for Ising model**

For any subset $S \subseteq V$, any vertex $v \in V \setminus S$, any pinning $\tau \in \{-1,+1\}^S$,

$$\forall c \in \{-1,+1\}, \qquad \mu_v(\, c \mid \tau\,) = \Omega(1)$$

*Under **any conditional**, the **marginal distribution** on one vertex **cannot be too biased***

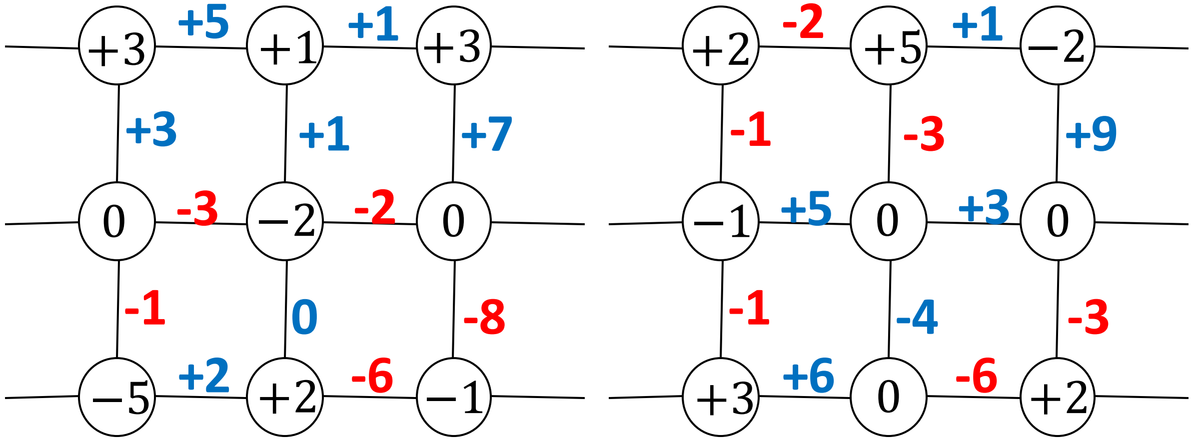


marginal distribution at v

$$\mu_v(+|\tau), \mu_v(\,-|\,\tau\,) = \Omega(1)$$

The assumption also appeared in ***learning*** [Bresler15], ***sampling and counting*** [CLV21]

Parameter distance



(G, J^{ν}, h^{ν}) (G, J^{μ}, h^{μ})

$$d_{\text{par}}(\nu, \mu) = \max\{\text{edge diff}, \text{vertex diff}\}$$

- Edge diff: $\max_{e \in E} |J^{\nu}(e) - J^{\mu}(e)|$
- Vertex diff: $\max_{v \in V} \frac{|h^{\nu}(v) - h^{\mu}(v)|}{\deg_G(v) + 1}$

What is the relation between ***parameter distance*** and ***TV-distance/χ^{α}-divergence***?

$d_{\text{par}}(\nu, \mu)$ is ***large*** $\xrightarrow{\text{marginal lower bound}}$ $D_{TV}(\nu \,\|\, \mu)$ is ***large***

[Feng, Liu and Yang 2025]

- Different edges/nodes may differ in ***different directions***. Overall, all differences ***may be cancelled out***
- The cancellation ***cannot happen*** for Ising models with marginal lower bounds

For two Ising models ν and μ both with **marginal lower bound** $b = \Omega(1)$

- The TV-distance: $D_{TV}(\nu \parallel \mu) \geq \frac{b^2}{2} d_{\mathrm{par}}(\nu, \mu)$ [Feng, Liu and Yang 2025]
- The χ^α-divergence: $D_{\chi^\alpha}(\nu \parallel \mu) \geq \frac{b^{2\alpha}}{2} d_{\mathrm{par}}^{\alpha}(\nu, \mu)$ [Feng and Fu 2025]
- A family of f-divergence can also be lower bounded in terms of b [Feng and Fu 2025]

Compute and check ***whether*** $d_{\mathrm{par}}(\nu, \mu) \leq \frac{1}{\mathrm{poly}(n)}$?

- **Yes**. All parameters in (J^ν, J^μ) and (h^ν, h^μ) are ***similar*** to each other

  use the similarity of parameters to design ***well-concentrated*** estimator

- **No**. Then the $D_{\chi^\alpha}(\nu \| \mu)$ is ***large***, at least $\frac{1}{\mathrm{poly}(n)}$

  relative-error approximation $\Rightarrow$ we can tolerate certain large error

The algorithm for small parameter distance $d_{\mathrm{par}}(\nu,\mu) \le \frac{1}{\mathrm{poly}(n)}$

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

$$W = \frac{w_\nu(X)}{w_\mu(X)} = \exp\Big(X^T\underbrace{(J^\nu - J^\mu)}_{\approx 0}X + X^T\underbrace{(h^\nu - h^\mu)}_{\approx 0}\Big) \text{ for } X\sim\mu \quad \Longrightarrow \quad \mathbb{E}[W] = \frac{Z_\nu}{Z_\mu}$$

- Draw random samples of W to estimate $\frac{1}{\mathbb{E}[W]} = \frac{Z_\mu}{Z_\nu}$
- Draw random samples of W to estimate the expectation of $\frac{1}{2}\left|1 - W\cdot\frac{Z_\mu}{Z_\nu}\right|^\alpha$

- How ***many*** samples do we need?
- How well are W and $\left|1 - W\cdot\frac{Z_\mu}{Z_\nu}\right|^\alpha$ ***concentrated*** around their mean?

The algorithm for small parameter distance $d_{\text{par}}(\nu, \mu) \leq \frac{1}{\text{poly}(n)}$

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

$$d_{\text{par}}(\nu, \mu) \leq \frac{1}{\text{poly}(n)}$$

$$\frac{w_\nu(x)}{w_\mu(x)} \in 1 \pm n \cdot d_{\text{par}}$$

$$\frac{Z_\mu}{Z_\nu} \in 1 \pm n \cdot d_{\text{par}}$$

$$0 \leq \left| 1 - \frac{w_\nu(x)}{w_\mu(x)} \cdot \frac{Z_\mu}{Z_\nu} \right|^\alpha \lesssim n^\alpha \cdot d_{\text{par}}^\alpha$$

$$\leq \text{poly}(n) \cdot D_{\chi^\alpha}(\nu \parallel \mu)$$

Divergence is at least $\Omega(d_{\text{par}}^\alpha)$

$$\mathbf{Var}\left[\left| 1 - W \cdot \frac{Z_\mu}{Z_\nu} \right|^\alpha \right] \leq \text{poly}(n) \cdot D_{\chi^\alpha}^2(\nu \parallel \mu)$$

$\text{poly}\left(\frac{n}{\varepsilon}\right)$ samples are enough to achieve $(1 \pm \varepsilon)$ ***relative error*** approximation

Algorithm ***sketch*** for large parameter distance $d_{\mathrm{par}}(\nu, \mu) > \frac{1}{\mathrm{poly}(n)}$

$d_{\mathrm{par}}(\nu, \mu)$ is ***large*** $\xrightarrow[\text{[Feng, Liu and Yang 2025]}]{\text{marginal lower bound}}$ $D_{\chi^\alpha}(\nu, \mu)$ is ***large***

Algorithm *sketch* for *large divergence* $D_{\chi^\alpha}(\nu \,\|\, \mu) > \frac{1}{\mathrm{poly}(n)}$

If $D_{\chi^\alpha}(\nu \,\|\, \mu) > \frac{1}{\mathrm{poly}(n)}$, then the following lower bound holds

$$D_{\chi^\alpha}(\nu \,\|\, \mu) = \frac{1}{2} \sum_{x \in \{\pm\}^V} \mu(x) \left| 1 - \frac{\nu(x)}{\mu(x)} \right|^\alpha \geq \frac{1}{\mathrm{poly}(n)} \cdot \sum_{x \in \{\pm\}^V} \mu(x) \left(1 + \frac{\nu(x)}{\mu(x)} \right)^\alpha$$

α is even → $|\cdot|^\alpha \to (\cdot)^\alpha$ → $D_{\chi^\alpha}(\nu \,\|\, \mu) = \frac{1}{2} \sum_{0 \leq k \leq \alpha} \binom{\alpha}{k} (-1)^{\alpha - k} \sum_{x \in \{\pm\}^V} \frac{\nu^k(x)}{\mu^{k-1}(x)}$

$$\sum_{\sigma \in \{\pm\}^V} \frac{\nu^k(\sigma)}{\mu^{k-1}(\sigma)} = \frac{Z_\mu^{k-1}}{Z_\nu^k} \cdot \sum_{x \in \{\pm\}^V} \exp\left(x^T (kJ^\nu - (k-1)J^\mu) x + x^T (kh^\nu - (k-1)h^\mu) \right)$$

Algorithm for χ^α-divergence between two Ising models

Theorem: approximation algorithm [F and Fu, 2025]

Two Ising models $\nu = \mathrm{Ising}(G, J^\nu, h^\nu)$ and $\mu = \mathrm{Ising}(G, J^\mu, h^\mu)$ with marginal lower bound

A ***family*** of Ising models $\mathcal{F} = \left\{\left(G, J^{(k)}, h^{(k)}\right) \mid \text{integer } 0 \le k \le \alpha\right\}$, where

$$J^{(k)} = kJ^\nu - (k-1)J^\mu$$

$$h^{(k)} = kh^\nu - (k-1)h^\mu$$

All Ising models in $\mathcal{F}$ admit $\mathrm{poly}(n/\epsilon)$-time algos for

- sampling
- approximate counting

$\Rightarrow$

$\mathrm{poly}(n/\epsilon)$-time algorithms for approximate $D_{\chi^\alpha}(\nu \,\|\, \mu)$

Algorithm *sketch* for *large divergence* $D_{\chi^\alpha}(\nu \parallel \mu) > \frac{1}{\mathrm{poly}(n)}$

If $D_{\chi^\alpha}(\nu \parallel \mu) > \frac{1}{\mathrm{poly}(n)}$, then the following lower bound holds

$$D_{\chi^\alpha}(\nu \parallel \mu) = \frac{1}{2} \sum_{x \in \{\pm\}^V} \mu(x) \left|1 - \frac{\nu(x)}{\mu(x)}\right|^\alpha \geq \frac{1}{\mathrm{poly}(n)} \cdot \sum_{x \in \{\pm\}^V} \mu(x) \left(1 + \frac{\nu(x)}{\mu(x)}\right)^\alpha$$

α is even → $|\cdot|^\alpha \to (\cdot)^\alpha$ → $D_{\chi^\alpha}(\nu \parallel \mu) = \frac{1}{2} \sum_{0 \leq k \leq \alpha} \binom{\alpha}{k} (-1)^{\alpha-k} \sum_{x \in \{\pm\}^V} \frac{\nu^k(x)}{\mu^{k-1}(x)}$

$$\sum_{\sigma \in \{\pm\}^V} \frac{\nu^k(\sigma)}{\mu^{k-1}(\sigma)} = \frac{Z_\mu^{k-1}}{Z_\nu^k} \cdot \sum_{x \in \{\pm\}^V} \exp\left(x^T J^{(k)} x + x^T h^{(k)}\right) = \frac{Z_\mu^{k-1}}{Z_\nu^k} \cdot Z^{(k)}$$

Algorithm *sketch* for *large divergence* $D_{\chi^\alpha}(\nu \,\|\, \mu) > \frac{1}{\mathrm{poly}(n)}$

If $D_{\chi^\alpha}(\nu \,\|\, \mu) > \frac{1}{\mathrm{poly}(n)}$, then the following lower bound holds

$$D_{\chi^\alpha}(\nu \,\|\, \mu) = \frac{1}{2} \sum_{x \in \{\pm\}^V} \mu(x) \left|1 - \frac{\nu(x)}{\mu(x)}\right|^\alpha \geq \frac{1}{\mathrm{poly}(n)} \cdot \sum_{x \in \{\pm\}^V} \mu(x) \left(1 + \frac{\nu(x)}{\mu(x)}\right)^\alpha$$

α is even $\rightarrow$ $|\cdot|^\alpha \to (\cdot)^\alpha$ $\rightarrow$ $D_{\chi^\alpha}(\nu \,\|\, \mu) = \frac{1}{2} \sum_{0 \leq k \leq \alpha} \binom{\alpha}{k} (-1)^{\alpha-k} \sum_{x \in \{\pm\}^V} \frac{\nu^k(x)}{\mu^{k-1}(x)}$

- Approx. each $\sum_{x \in \{\pm\}^V} \frac{\nu^k(x)}{\mu^{k-1}(x)}$ with ***relative err*** $\frac{\varepsilon}{\mathrm{poly}(n)}$ = ***additive err*** $\frac{\varepsilon}{\mathrm{poly}(n)} \cdot \sum_{x \in \{\pm\}^V} \frac{\nu^k(x)}{\mu^{k-1}(x)}$
- Approx. whole sum $D_{\chi^\alpha}(\nu \,\|\, \mu)$ with ***additive error***

$$\frac{\varepsilon}{\mathrm{poly}(n)} \cdot \sum_{0 \leq k \leq \alpha} \binom{\alpha}{k} \sum_{x \in \{\pm\}^V} \frac{\nu^k(x)}{\mu^{k-1}(x)} = \frac{\varepsilon}{\mathrm{poly}(n)} \cdot \sum_{x \in \{\pm\}^V} \mu(x) \left(1 + \frac{\nu(x)}{\mu(x)}\right)^\alpha \leq \varepsilon \cdot D_{\chi^\alpha}(\nu \,\|\, \mu)$$

relative err ε

Algorithm *sketch* for *large divergence* $D_{\chi^\alpha}(\nu \parallel \mu) > \frac{1}{\mathrm{poly}(n)}$

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

$$= \frac{1}{2}\sum_{0\le k\le\alpha} (-1)^{\alpha-k}\binom{\alpha}{k}\left(\sum_{x:\nu(x)>\mu(x)} \frac{\nu^k(x)}{\mu^{k-1}(x)} + (-1)^\alpha \sum_{x:\nu(x)<\mu(x)} \frac{\nu^k(x)}{\mu^{k-1}(x)}\right)$$

- ***Sample*** $X \sim \mathrm{Ising}\left(G, J^{(k)}, h^{(k)}\right)$
- ***Estimator***: $W_k = \mathbf{1}[\nu(X) > \mu(X)] \cdot \frac{Z_\mu^{k-1}}{Z_\nu^k} Z^{(k)}$

$\mathbb{E}[W_k] =$

We can only approximate $\nu(X)$ and $\mu(X)$, but ***cannot*** exactly compute $\mathbf{1}[\nu(X) > \mu(X)]$

- We only make mistake when $\nu(X)$ is very close to $\mu(X)$

We use random samples to ***estimate*** the expectation and we need to put all terms together

- The concentration error can be bounded in a similar way as that for even α case

Open problems

- remove the marginal lower bound assumption.
- more general graphical models or general distributions
- χ^{α}-divergence for real number α or other divergences
- deterministic approximation algorithms
- faster algorithms (current algorithm require $\alpha = O(1)$ with running time $n^{O(\alpha)}$)
- connections or applications in learning and testing

Thank You