Computational Implications of Reducing Data to Sufficient Statistics

Andrea Montanari[*]

September 12, 2014

Abstract

Given a large dataset and an estimation task, it is common to pre-process the data by reducing them to a set of sufficient statistics. This step is often regarded as straightforward and advantageous (in that it simplifies statistical analysis). I show that –on the contrary– reducing data to sufficient statistics can change a computationally tractable estimation problem into an intractable one. I discuss connections with recent work in theoretical computer science, and implications for some techniques to estimate graphical models.

1 Introduction

The idea of sufficient statistics is a cornerstone of statistical theory and statistical practice. Given a dataset, evaluating a set of sufficient statistics yields a concise representation that can be subsequently used to design (for instance) optimal statistical estimation procedures. To quote a widely adopted textbook [LC98]:

> 'It often turns out that some part of the data carries no information about the unknown distribution and that $\boldsymbol{X}$ can therefore be replaced by some statistic $\boldsymbol{T} = \boldsymbol{T}(\boldsymbol{X})$ without loss of information.'

The main point of the present paper is the following. While optimal statistical estimation can be performed solely on the basis of sufficient statistics, *reduction to sufficient statistics can lead to an explosion in computational complexity.* This phenomenon is so dramatic that –after reduction to sufficient statistics– a tractable estimation task can become entirely intractable.

To be concrete, we shall consider the problem of estimating the parameters of an exponential family over $\boldsymbol{x} \in \{0,1\}^p$:

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}) = \frac{1}{Z(\boldsymbol{\theta})}\, h(\boldsymbol{x})\, e^{\langle \boldsymbol{\theta}, \boldsymbol{x}\rangle}\,. \tag{1}$$

Here $h : \{0,1\}^p \to \mathbb{R}_{>0}$ will be assumed strictly positive, $\boldsymbol{\theta} \in \mathbb{R}^p$ and $\langle \boldsymbol{a}, \boldsymbol{b}\rangle = \sum_{i=1}^d a_i b_i$ is the standard scalar product in $\mathbb{R}^d$. We assume to be given n i.i.d. samples $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n \sim p_{\boldsymbol{\theta}}$ and want to estimate $\boldsymbol{\theta}$. A vector of sufficient statistics is clearly given by the empirical average

$$\boldsymbol{T}(\boldsymbol{X}^{(n)}) = \frac{1}{n}\sum_{\ell=1}^{n} \boldsymbol{X}_\ell\,, \tag{2}$$

[*] Department of Electrical Engineering and Department of Statistics, Stanford University

where we introduced the notation $\boldsymbol{X}^{(n)} = (\boldsymbol{X}_1, \dots, \boldsymbol{X}_n)$ for the dataset of n samples.

Let us reconsider the standard argument used to prove that a reduction in complexity entails no loss of information [LC98], since this is already suggestive of a possible explosion in computational complexity. Given any estimator $\hat{\boldsymbol{\theta}}(\boldsymbol{X}^{(n)})$ of the parameters $\boldsymbol{\theta}$, the argument constructs a randomized estimator $\hat{\boldsymbol{\theta}}^{\text{new}}(\boldsymbol{T})$ that only depends on the sufficient statistics $\boldsymbol{T}$, and has the same distribution as $\hat{\boldsymbol{\theta}}(\boldsymbol{X}^{(n)})$, in particular the same risk. The construction is fairly simple. Given $\boldsymbol{T}$, we can sample $\widetilde{\boldsymbol{X}}^{(n)}$ from the law $p_{\boldsymbol{\theta}}(\,\cdot\,|\boldsymbol{T})$, conditional on the observed sufficient statistics $\boldsymbol{T}(\widetilde{\boldsymbol{X}}^{(n)}) = \boldsymbol{T}$. By definition of sufficient statistics, the conditional law $p_{\boldsymbol{\theta}}(\,\cdot\,|\boldsymbol{T})$ does not depend on $\boldsymbol{\theta}$, and hence we can sample $\widetilde{\boldsymbol{X}}^{(n)}$ without knowing $\boldsymbol{\theta}$. We then define $\hat{\boldsymbol{\theta}}^{\text{new}}(\boldsymbol{T}) \equiv \hat{\boldsymbol{\theta}}(\widetilde{\boldsymbol{X}}^{(n)})$. This has clearly the same risk as $\hat{\boldsymbol{\theta}}(\boldsymbol{X}^{(n)})$. In words, we were able to generate new data as informative as the original one using the sufficient statistics.

The problem with this argument is that sampling from the conditional distribution of $p_{\boldsymbol{\theta}}$ given $\boldsymbol{T}(\widetilde{\boldsymbol{X}}^{(n)}) = \boldsymbol{T}$ can be computationally hard. Indeed simple examples can be given for the weight $h(\,\cdot\,)$ (see Section 3) that make sampling from $p_{\boldsymbol{\theta}}(\,\cdot\,|\boldsymbol{T})$ –even approximately– impossible unless P = NP (see below). In other words, this argument is based on a reduction that is not computationally efficient.

The rest of the paper is organized as follows. In Section 2 we state formally our main results. Under technical assumptios this shows that, if there is a computationally efficient estimator $\hat{\boldsymbol{\theta}}(\boldsymbol{T})$ that use only sufficient statistics, then the partition function $Z(\boldsymbol{0})$ can be approximated, also efficiently. In Section 3 we construct a family of functions $h(\,\cdot\,)$ for which approximating $Z(\boldsymbol{0})$ in polynomial time is impossible unless P = NP. As a consequence of our main theorem, no efficient parameter estimator from sufficient statistics exists for these models (unless P = NP). Remarkably, we show that a simple consistent estimator can be developed using the data $\boldsymbol{X}^{(n)}$, for the same models.

The example in Section 3 is an undirected graphical model, and indeed intractability of computing approximations to $Z(\boldsymbol{0})$ is quite generic in this context. In Section 4 we discuss relations to the literature in algorithms and graphical models estimation. While several closely related

1.1 Notations

We will use $[n] = \{1, 2, \dots, n\}$ to denote the set of first n integers. Whenever possible, we will follow the convention of denoting deterministic values by lowercase letters (e.g. $x, y, z, \dots$) and random variables by uppercase letters (e.g. $X, Y, Z, \dots$). We reserve boldface for vectors and matrices, e.g. $\boldsymbol{x}$ is a deterministic vector or matrix, and $\boldsymbol{X}$ a random vector or matrix. For a vector $\boldsymbol{x} \in \mathbb{R}^m$, and a set $A \subseteq [m]$, $\boldsymbol{x}_A$ denotes the subvector indexed by A. The components of $\boldsymbol{x}$ are denoted by $(x_1, x_2, \dots, x_m)$.

Given $f : \mathbb{R}^m \to \mathbb{R}$, we denote by $\nabla f(\boldsymbol{x}) \in \mathbb{R}^m$ its gradient at $\boldsymbol{x}$, and by $\nabla^2 f(\boldsymbol{x}) \in \mathbb{R}^{m\times m}$ it Hessian. Whenever useful, we add as a subscript the variable with respect to which we are differentiating as, for instance, in $\nabla_{\boldsymbol{x}} f(\boldsymbol{x})$.

2 Computational hardness of consistent parameter estimation

As mentioned above, we consider the problem of consistent estimation of the parameters $\boldsymbol{\theta}$ of the model (1), from the sufficient statistics (2). Throughout we will assume $h(\boldsymbol{x}) > 0$ strictly for all

$\boldsymbol{x} \in \{0,1\}^n$.

As $n \to \infty$, $\boldsymbol{T}(\boldsymbol{X}^{(n)})$ converges –by the law of large numbers– to the population mean

$$\boldsymbol{\tau}_*(\boldsymbol{\theta}) \equiv \mathbb{E}_{\boldsymbol{\theta}}\{\boldsymbol{X}\} = \frac{1}{Z(\boldsymbol{\theta})} \sum_{\boldsymbol{x}\in\{0,1\}^n} \boldsymbol{x}\, h(\boldsymbol{x})\, e^{\langle \boldsymbol{\theta}, \boldsymbol{x}\rangle}\,. \tag{3}$$

Here and below $\mathbb{E}_{\boldsymbol{\theta}}$ denotes expectation with respect to the distribution $p_{\boldsymbol{\theta}}$. We use the $*$ subscript to emphasize that this $\boldsymbol{\tau}_*$ is a function, and distinguish it from a specific value $\boldsymbol{\tau} \in (0,1)^p$. A consistent estimator based on the sufficient statistics $\boldsymbol{T}$ must necessarily invert this function in the limit $n \to \infty$. This motivates the following definition.

Definition 2.1. *For $p \in \mathbb{N}$, let $\mathcal{H}_p = \{h(\,\cdot\,)\}$ be a set of functions $h : \{0,1\}^p \to \mathbb{R}_{>0}$ A* polynomial consistent estimator from sufficient statistics *for the model (1) with $h \in \mathcal{H}_p$ is an algorithm that given $\boldsymbol{\tau} \in (0,1)^p$, and a precision parameter $\xi > 0$, returns $\hat{\boldsymbol{\theta}}(\boldsymbol{\tau}, \xi)$ such that*

1. *$\|\hat{\boldsymbol{\theta}}(\boldsymbol{\tau}, \xi) - \boldsymbol{\theta}\|_2 \le \xi$ for some $\boldsymbol{\theta}$ such that $\boldsymbol{\tau}_*(\boldsymbol{\theta}) = \boldsymbol{\tau}$.*
2. *The algorithm terminates in time polynomial in $1/\xi$, p, the maximum description length of any of the τ_i, and the description length of $h \in \mathcal{H}_p$.*

Two remaks are in order. The terminology adopted here (in particular, the use of ‘consistent’ and ‘polynomial’) is motivated by the following remarks. First, in the terminology of complexity theory, this is a ‘fully polynomial time approximation scheme.’

Second, as discussed below, there is indeed a unique, continuous, function $\boldsymbol{\theta}_* : (0,1)^n \to \mathbb{R}^p$ such that $\boldsymbol{\tau}_*(\boldsymbol{\theta}_*(\boldsymbol{\tau})) = \boldsymbol{\tau}$. This implies that $\boldsymbol{\theta}$ is indeed a consistent estimator in the sense that for a sequence $\xi_n \to 0$, we have, almost surely and hence in probability,

$$\lim_{n\to\infty} \hat{\boldsymbol{\theta}}(\boldsymbol{T}(\boldsymbol{X}^{(n)}), \xi_n) = \boldsymbol{\theta}\,. \tag{4}$$

Finally, in the following we shall occasionally drop the qualification ‘from sufficient statistics’ when it is clear that we are considering estimators that only use sufficient statistics.

We next state our main theorem.

Theorem 1. *Assume $\mathcal{H}_p$, $p \ge 1$, to be a family of weight functions such that the the following three conditions hold:*

C1. *There exists $\delta \in (0, 1/2)$ such that, for any $p \in \mathbb{N}$, any $h \in \mathcal{H}_p$, and any $i \in [p]$*

$$\tau_i(\boldsymbol{\theta} = \mathbf{0}) \in (\delta, 1-\delta)\,. \tag{5}$$

C2. *There exists a polynomial $L(p)$ such that, for any $p \in \mathbb{N}$, any $h \in \mathcal{H}_p$, we have, for all $\boldsymbol{\theta}$ such that $\boldsymbol{\tau}_*(\boldsymbol{\theta}) \in [\delta, 1-\delta]^p$,*

$$\mathrm{Cov}_{\boldsymbol{\theta}}(\boldsymbol{X}) \equiv \mathbb{E}_{\boldsymbol{\theta}}\{\boldsymbol{X}\boldsymbol{X}^{\mathsf{T}}\} - \mathbb{E}_{\boldsymbol{\theta}}\{\boldsymbol{X}\}\mathbb{E}_{\boldsymbol{\theta}}\{\boldsymbol{X}\}^{\mathsf{T}} \succeq \frac{1}{L(p)}\, \mathrm{I}_{p\times p}\,. \tag{6}$$

C3. *There exists a polynomial $K(p)$ such that, for any $p \in \mathbb{N}$, any $h \in \mathcal{H}_p$*

$$\max_{\boldsymbol{x},\boldsymbol{y}\in\{0,1\}^p} \big|\log h(\boldsymbol{x}) - \log h(\boldsymbol{y})\big| \le K(p)\,. \tag{7}$$

If there exists a polynomial consistent estimator for the model $\mathcal{H}$, then, for any $\varepsilon > 4p\delta(K(p) + \log(4/\delta))$, there exists an algorithm polynomial in the description length of h, and in $(1/\varepsilon)$, returning $\widehat{Z}$ such that

$$e^{-\varepsilon}\, Z(\mathbf{0}) \le \widehat{Z} \le e^{\varepsilon}\, Z(\mathbf{0})\,. \tag{8}$$

The rest of this section is devoted to the proof of this theorem.

The main implication of this theorem is negative. If the problem of approximating $Z(\mathbf{0})$ is intractable, it follows from the above that there is no polynomial consistent estimator. We will show in Section 3 that we can use recent results in complexity theory to construct a fairly simple class $\mathcal{H} = \{h\}$ such that:

- An approximation scheme does not exist for $Z(\mathbf{0})$ under the standard complexity theory assumption P $\neq$ NP.
- Remarkably, a consistent and computationally efficient estimator $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(\boldsymbol{X}^{(n)})$ of the model parameters exists! However, this is not a function only of the sufficient statistics.

As a direct consequence, we have the following.

Corollary 2.2. *There are classes of models $\mathcal{H}$ for which no polynomial consistent estimator of the parameters from sufficient statistics exists, unless* P = NP*.*

Before outlining the proof of Theorem 1, it is useful to recall a few well-known properties of exponential families, as specialized to the present setting. While these are consequences of fairly standard general statements (see for instance [Efr78, LC98]), we present self-contained proofs in Appendix A for the readers' convenience.

We will use standard statistics notations for the *cumulant generating function* (also known as *log-partition function*), $A : \mathbb{R}^p \to \mathbb{R}$,

$$A(\boldsymbol{\theta}) \equiv \log Z(\boldsymbol{\theta}) = \log \Big\{ \sum_{\boldsymbol{x} \in \{0,1\}^n} h(\boldsymbol{x})\, e^{\langle \boldsymbol{\theta}, \boldsymbol{x} \rangle} \Big\}\,, \tag{9}$$

and its Legendre-Fenchel transform $F : (0,1)^p \to \mathbb{R}$, which we shall call the *free energy*,

$$F(\boldsymbol{\tau}) = \inf_{\boldsymbol{\theta} \in \mathbb{R}^p} \big(A(\boldsymbol{\theta}) - \langle \boldsymbol{\tau}, \boldsymbol{\theta} \rangle \big)\,. \tag{10}$$

Proposition 2.3. *Given a strictly positive $h : \{0,1\}^n \to \mathbb{R}_{>0}$, the following hold:*

Fact1. *The function $\boldsymbol{\tau}_* : \mathbb{R}^p \to (0,1)^p$ is $C_\infty(\mathbb{R}^p)$.*

Fact2. *Recalling that $\mathrm{Cov}_{\boldsymbol{\theta}}(\boldsymbol{X}) \in \mathbb{R}^{p \times p}$ denotes the covariance matrix of the law $p_{\boldsymbol{\theta}}$, we have*

$$\nabla_{\boldsymbol{\theta}} A(\boldsymbol{\theta}) = \boldsymbol{\tau}_*(\boldsymbol{\theta})\,, \tag{11}$$

$$\nabla^2_{\boldsymbol{\theta}} A(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} \boldsymbol{\tau}_*(\boldsymbol{\theta}) = \mathrm{Cov}_{\boldsymbol{\theta}}(\boldsymbol{X})\,. \tag{12}$$

Fact3. $\mathrm{Cov}_{\boldsymbol{\theta}}(\boldsymbol{X}) \succeq c(\boldsymbol{\theta})\, \mathrm{I}_p$ *for some continuous strictly positive $c(\boldsymbol{\theta}) > 0$.*

Fact4. $\boldsymbol{\tau}_* : \mathbb{R}^p \to (0,1)^p$ *is a bijection. We will denote by $\boldsymbol{\theta}_* : (0,1)^p \to \mathbb{R}^p$ the inverse mapping.*

Fact5. *The function $F : (0,1)^p \to \mathbb{R}^p$ is concave and $C_\infty((0,1)^p)$ with*

$$A(\boldsymbol{\theta}) = \max_{\boldsymbol{\tau}\in(0,1)^p} \big\{F(\boldsymbol{\tau}) + \langle \boldsymbol{\tau}, \boldsymbol{\theta}\rangle\big\} = F(\boldsymbol{\tau}_*(\boldsymbol{\theta})) + \langle \boldsymbol{\tau}_*(\boldsymbol{\theta}), \boldsymbol{\theta}\rangle\,, \tag{13}$$

$$\nabla_{\boldsymbol{\tau}} F(\boldsymbol{\tau}) = -\boldsymbol{\theta}_*(\boldsymbol{\tau}) \tag{14}$$

$$\nabla^2_{\boldsymbol{\tau}} F(\boldsymbol{\tau}) = -\nabla_{\boldsymbol{\tau}}\boldsymbol{\theta}_*(\boldsymbol{\tau}) = -\big[\nabla^2_{\boldsymbol{\theta}} A(\boldsymbol{\theta}_*(\boldsymbol{\tau}))\big]^{-1}\,. \tag{15}$$

Fact6. *Let $\mathcal{P}_p = \mathcal{P}(\{0,1\}^p)$ be the set of probability distributions over $\{0,1\}^p$, and denote by $H(q) = -\sum_{\boldsymbol{x}\in\{0,1\}^p} q(\boldsymbol{x}) \log q(\boldsymbol{x})$ the Shannon entropy of $q \in \mathcal{P}_p$. Then*

$$F(\boldsymbol{\tau}) = \max_{q\in\mathcal{P}_p} \Big\{ H(q) + \mathbb{E}_q \log h(\boldsymbol{X}) \;\; \textit{such that } \mathbb{E}_q\{\boldsymbol{X}\} = \boldsymbol{\tau}\Big\}\,. \tag{16}$$

We are now in position to prove Theorem 1: we present here a version of the proof that omits some technical step, and complete the details in Appendix B.

Proof of Theorem 1. By Eq. (13) we have, letting $D_p(\delta) \equiv [\delta, 1-\delta]^p$,

$$\log Z(\mathbf{0}) = \max_{\boldsymbol{\tau}\in(0,1)^p} F(\boldsymbol{\tau}) \tag{17}$$

$$= \max_{\boldsymbol{\tau}\in D_p(\delta)} F(\boldsymbol{\tau}) = F(\boldsymbol{\tau}_*(\mathbf{0}))\,, \tag{18}$$

where the second equality follows from assumption C1.

The problem of computing $Z(\mathbf{0})$ is then reduced to the one of maximizing the concave function $F(\boldsymbol{\tau})$ over the convex set $D_p(\delta)$. Note that, by assumption C2 and Eq. (15), the gradient of $F(\boldsymbol{\tau})$ has Lipshitz modulus bounded by $L(p)$ on $D_p(\delta)$. Namely, for all $\boldsymbol{\tau}_1, \boldsymbol{\tau}_2 \in D_p(\delta)$,

$$\big\|\nabla_{\boldsymbol{\tau}} F(\boldsymbol{\tau}_1) - \nabla_{\boldsymbol{\tau}} F(\boldsymbol{\tau}_2)\big\|_2 \le L(p)\, \|\boldsymbol{\tau}_1 - \boldsymbol{\tau}_2\|_2\,. \tag{19}$$

We will maximize $F(\boldsymbol{\tau})$ by a standard projected gradient algorithm. We will work here under the assumption that we have access to an oracle that given a point $\boldsymbol{\tau} \in D_p(\delta)$, returns $\boldsymbol{\theta}_*(\boldsymbol{\tau}) = -\nabla_{\boldsymbol{\tau}} F(\boldsymbol{\tau})$. While in reality we do not have access to such an oracle, Definition 2.1 (and the Theorems assumptions) imply that there exists an efficient approximation scheme for this oracle. We will show in Appendix B that indeed we can replace the oracle by such an approximation scheme.

Given the oracle, the projected gradient algorithm is defined by the iteration (with the superscript t indicating the iteration number, and letting $L = L(p)$)

$$\boldsymbol{\tau}^0 = \Big(\frac{1}{2}, \frac{1}{2}, \dots, \frac{1}{2}\Big)^{\mathsf{T}}\,, \tag{20}$$

$$\boldsymbol{\tau}^{t+1} = \mathsf{P}_\delta\Big(\boldsymbol{\tau}^t - \frac{1}{L}\boldsymbol{\theta}_*(\boldsymbol{\tau}^t)\Big)\,. \tag{21}$$

Here $\mathsf{P}_\delta(\boldsymbol{x})$ is the orthogonal projector on $D_p(\delta)$, which can be computed efficiently since, for each $i \in \{1, 2, \dots, p\}$, and $\boldsymbol{u} \in \mathbb{R}^p$,

$$\mathsf{P}_\delta(\boldsymbol{u})_i = \begin{cases} \delta & \text{if } u_i < \delta\,,\\ u_i & \text{if } u_i \in [\delta, 1-\delta]\,,\\ 1-\delta & \text{if } u_i > 1-\delta\,. \end{cases} \tag{22}$$

We run $t_0 = t_0(p, \varepsilon) \equiv \lceil 2p\, L(p)/\varepsilon \rceil$ iterations of projected gradient. By [BT09, Theorem 3.1] we have

$$0 \le F(\boldsymbol{\tau}_*(\mathbf{0})) - F(\boldsymbol{\tau}^{t_0}) \le \frac{L\|\boldsymbol{\tau}^0 - \boldsymbol{\tau}_*(\mathbf{0})\|_2^2}{2t_0} \le \frac{\varepsilon}{4}\,. \tag{23}$$

Hence, $F(\boldsymbol{\tau}^{t_0})$ provides a good approximation of $F(\boldsymbol{\tau}_*(\mathbf{0})) = \log Z(\mathbf{0})$.

We are left with the task of evaluating $F(\boldsymbol{\tau}^{t_0})$. The idea is to 'integrate' the derivative of $F(\boldsymbol{\tau})$ along a path that starts at a point $\boldsymbol{\tau}^{(0)}$ where F can be easily approximated. We will use again $\nabla_{\boldsymbol{\tau}} F(\boldsymbol{\tau}) = -\boldsymbol{\theta}_*(\boldsymbol{\tau})$ and assume that $\boldsymbol{\theta}_*$ is exactly computed by an oracle. Again we will see in Appendix B that this oracle can be replaced by the estimator in Definition 2.1

Let $\boldsymbol{\tau}^{(0)} \equiv (\delta, \dots, \delta)^{\mathsf{T}}$. Next consider Eq. (16). For any $q \in \mathcal{P}_p$ such that $\mathbb{E}_q(\boldsymbol{X}) = \boldsymbol{\tau}^{(0)}$, we have, letting $s(x) \equiv -x \log x - (1-x)\log(1-x)$

$$0 \le H(q) \le p\, s(\delta)\,, \tag{24}$$

$$\Big|\mathbb{E}_p \log \frac{h(\boldsymbol{X})}{h(\mathbf{0})}\Big| \le K(p)\, \mathbb{P}_q(\boldsymbol{X} \neq \mathbf{0}) \le K(p) p \delta\,. \tag{25}$$

Where the second inequality in Eq. (24) follows since the joint entropy is not larger than the sum of the entropy of the marginals. The first inequality Eq. (25) follows from assumption C3, and the last inequality in Eq. (25) is Markov's. Using these bounds in Eq. (16), we get

$$\Big|F(\boldsymbol{\tau}^{(0)}) - \log h(\mathbf{0})\Big| \le p\, s(\delta) + pK(p)\,\delta \le \frac{\varepsilon}{4}\,, \tag{26}$$

where the last inequality follows from the assumption $\varepsilon > 4p\delta(K(p) + \log(4/\delta))$.

For an integer m, we let $\boldsymbol{\tau}^{(m)} = \boldsymbol{\tau}^{t_0}$ be the output of projected gradient, and, for $\ell \in \{1, \dots, m-1\}$, $\boldsymbol{\tau}^{(\ell)} \equiv \boldsymbol{\tau}^{(0)} + \ell\,(\boldsymbol{\tau}^{(m)} - \boldsymbol{\tau}^{(0)})/m$ be given by linearly interpolating between $\boldsymbol{\tau}^{(0)}$ and $\boldsymbol{\tau}^{(m)}$. Note that, since $\boldsymbol{\tau}^{(0}, \boldsymbol{\tau}^{(m)} \in D_p(\delta)$, we have $\boldsymbol{\tau}^{(\ell)} \in D_p(\delta)$ as well, by convexity.

We finally construct our approximation $\widehat{Z}$ by letting

$$\log \widehat{Z}^{\rm or} \equiv \log h(\mathbf{0}) - \sum_{\ell=1}^{m} \langle \boldsymbol{\theta}_*(\boldsymbol{\tau}^{(\ell)}), \boldsymbol{\tau}^{(\ell)} - \boldsymbol{\tau}^{(\ell-1)} \rangle\,. \tag{27}$$

(We introduced the superscript or to emphasize that this approximation makes use of the oracle $\boldsymbol{\theta}_*$. In Appendix B we control the additional error induced by the use of $\hat{\boldsymbol{\theta}}$.) Let us bound the approximation error:

$$\begin{aligned} \Big|\log \frac{\widehat{Z}^{\rm or}}{Z(\mathbf{0})}\Big| \le & \big|F(\boldsymbol{\tau}^{(0)}) - \log h(\mathbf{0})\big| \\ & + \sum_{\ell=1}^{m} \big|F(\boldsymbol{\tau}^{(\ell)}) - F(\boldsymbol{\tau}^{(\ell-1)}) - \langle \nabla_{\boldsymbol{\tau}} F(\boldsymbol{\tau}^{(\ell)}), \boldsymbol{\tau}^{(\ell)} - \boldsymbol{\tau}^{(\ell-1)} \rangle\big| \\ & + \big|F(\boldsymbol{\tau}_*(\mathbf{0})) - F(\boldsymbol{\tau}^{(m)})\big|\,. \end{aligned} \tag{28}$$

The first term is bounded by Eq. (26) and the last by Eq. (23). As for the middle term, by the intermediate value theorem there exists, for each ℓ, a point $\tilde{\boldsymbol{\tau}}^{(\ell)} \in [\boldsymbol{\tau}^{(\ell-1)}, \boldsymbol{\tau}^{(\ell)}]$ such that

$$\begin{aligned} \big|F(\boldsymbol{\tau}^{(\ell)}) - F(\boldsymbol{\tau}^{(\ell-1)}) - \langle \nabla_{\boldsymbol{\tau}} F(\boldsymbol{\tau}^{(\ell)}), \boldsymbol{\tau}^{(\ell)} - \boldsymbol{\tau}^{(\ell-1)} \rangle\big| &= \big|\langle \nabla_{\boldsymbol{\tau}} F(\tilde{\boldsymbol{\tau}}^{(\ell)}) - \nabla_{\boldsymbol{\tau}} F(\boldsymbol{\tau}^{(\ell)}), \boldsymbol{\tau}^{(\ell)} - \boldsymbol{\tau}^{(\ell-1)} \rangle\big| \\ &\le L(p)\, \|\boldsymbol{\tau}^{(\ell)} - \boldsymbol{\tau}^{(\ell-1)}\|_2^2 \le \frac{L(p)}{m^2}\, \|\boldsymbol{\tau}^{(m)} - \boldsymbol{\tau}^{(0)}\|_2^2 \le \frac{L(p)p}{m^2}\,. \end{aligned} \tag{29}$$

Hence, choosing $m = m_0(p,\varepsilon) \equiv \lceil 4L(p)p/\varepsilon \rceil$, we can bound the sum in Eq. (28) by $\varepsilon/4$. Hence the approximation error is bounded by

$$\Big|\log \frac{\widehat{Z}^{\rm or}}{Z(\mathbf{0})}\Big| \le \frac{\varepsilon}{4} + \frac{\varepsilon}{4} + \frac{\varepsilon}{4} \le \frac{3\varepsilon}{4}\,, \tag{30}$$

which concludes our proof outline. □

3 An example

In this section we describe a simple class of functions $\mathcal{H}_p$ for which it is impossible to estimate $Z(\mathbf{0})$ in polynomial time unless P = NP. Using our Theorem 1, we will show that consistent parameter estimation using sufficient statistics is intractable for these models. We will then show that –for the same models– it is quite easy to estimate the parameters from i.i.d. samples $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n \sim p_{\boldsymbol{\theta}}$.

3.1 Intractability of estimation from sufficient statistics

For $\beta > 0$ a fixed number, and $G = (V = [p], E)$ a simple graph, we let $h_{G,\beta} : \{0,1\}^p \to \mathbb{R}_{>0}$ be defined by

$$h_{G,\beta}(\boldsymbol{x}) \equiv \exp\Big\{2\beta \sum_{(i,j)\in E} \mathbb{I}(x_i \neq x_j)\Big\}\,. \tag{31}$$

This is also known as the *anti-ferromagnetic Ising model* on graph G. Fixing $k \in \mathbb{N}$, and $\beta \in \mathbb{R}_{>0}$, we introduce the class of functions

$$\mathcal{H}_p(k,\beta) \equiv \big\{h_{G,\beta} : G \text{ is a simple regular graph of degree } k \text{ over } p \text{ vertices } \big\}\,. \tag{32}$$

(Recall that a regular graph is a graph with the same degree at all vertices. The set of regular graphs is non-empty as soon as pk is even, and $p \ge p_0(k)$.) Intractability of approximating $Z(\mathbf{0})$ was characterized in [SS12, GSV12]. We restate the main result of [SS12], adapting it to the present setting[1].

Theorem 2 (Sly, Sun, 2012)**.** *For any $k \ge 3$ and $\beta > \operatorname{atanh}(1/(k-1))$ there exists $\varepsilon_0 = \varepsilon_0(k,\beta) > 0$ such that the following holds. Unless* P = NP*, there is no polynomial algorithm taking as input $h \in \mathcal{H}_p(k,\beta)$ and returning $\widehat{Z}$ such that*

$$e^{-\varepsilon_0 p} Z(\mathbf{0}) \le \widehat{Z} \le e^{\varepsilon_0 p} Z(\mathbf{0}) \tag{33}$$

where we recall that $Z(\mathbf{0}) \equiv \sum_{\boldsymbol{x}\in\{0,1\}^p} h(\boldsymbol{x})$.

We can now state (and prove) and more concrete form of Corollary 2.2. In Section 3.2 we will show that the model parameters can be consistently estimated in polynomial time from i.i.d. samples $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n \sim p_{\boldsymbol{\theta}}$.

[1] The present statement differs from the original one in [SS12] in two technical aspects. First, we state that $Z(\mathbf{0})$ cannot be approximated within a ratio $e^{\varepsilon_0 p}$, while [SS12] state their result in terms of FPRAS. However, the authors of [SS12] also mention that their same proof yields impossibility to achieve the approximation ratio stated here (which is in fact easy to check). Second, the complexity theoretic assumption in [SS12] is RP≠NP instead of P≠NP. Since we are focusing here on impossibility of *deterministic* algorithms, assuming P≠NP is sufficient.

Corollary 3.1. *Fix $k \geq 3$ and $\beta > \text{atanh}(1/(k-1))$. Unless* $\mathrm{P} = \mathrm{NP}$*, then there exists no polynomial consistent estimator from sufficient statistics for the model* $\{\mathcal{H}_p(k,\beta)\}_{p \geq p_0(k)}$.

Proof. The proof consists in checking that the assumptions C1, C2, C3 of Theorem 1 apply. It then follows from Theorem 2 that either P = NP or there is no polynomial consistent estimator from sufficient statistics. Throughout, we will write C or c for generic strictly positive constants that can depend on β, k.

Let us start with condition C1. We fix a graph $G = (V = [p], E)$ on p vertices. For a vertex $i \in [p]$, we let $\partial i = \{j \in [p] : (i,j) \in E\}$ be the set of neighbors of i. Writing $\mathbb{P}_{\mathbf{0}}$ for $\mathbb{P}_{\boldsymbol{\theta}=\mathbf{0}}$, we have $\tau_{*,i}(\mathbf{0}) = \mathbb{P}_{\mathbf{0}}(X_i = 1)$.

Note that $p_{\boldsymbol{\theta}}$ is a Markov Random Field with graph G. We therefore have

$$\min_{\boldsymbol{x}_{\partial i} \in \{0,1\}^{\partial i}} \mathbb{P}_{\mathbf{0}}(X_i = 1 | \boldsymbol{X}_{\partial i} = \boldsymbol{x}_{\partial i}) \leq \mathbb{P}_{\mathbf{0}}(X_i = 1) \leq \max_{\boldsymbol{x}_{\partial i} \in \{0,1\}^{\partial i}} \mathbb{P}_{\mathbf{0}}(X_i = 1 | \boldsymbol{X}_{\partial i} = \boldsymbol{x}_{\partial i}) \, . \tag{34}$$

A simple direct calculation with Eq. (31) yields

$$\mathbb{P}_{\mathbf{0}}(X_i = 1 | \boldsymbol{X}_{\partial i} = \boldsymbol{x}_{\partial i}) = \frac{e^{2\beta \, n_0(\boldsymbol{x}_{\partial i})}}{e^{2\beta \, n_0(\boldsymbol{x}_{\partial i})} + e^{2\beta \, n_1(\boldsymbol{x}_{\partial i})}} \, , \tag{35}$$

where $n_0(\boldsymbol{x}_{\partial i})$ and $n_0(\boldsymbol{x}_{\partial i})$ are the number of zeros and ones in the vector $\boldsymbol{x}_{\partial i}$.

The right hand side of Eq. (35) is maximized when $n_0(\boldsymbol{x}_{\partial i}) = k - n_1(\boldsymbol{x}_{\partial i}) = k$, and minimized when $n_0(\boldsymbol{x}_{\partial i}) = k - n_1(\boldsymbol{x}_{\partial i}) = 0$. We then have

$$\frac{1}{1 + e^{2\beta k}} \leq \mathbb{P}_{\mathbf{0}}(X_i = 1) \leq \frac{e^{2\beta k}}{1 + e^{2\beta k}} \, . \tag{36}$$

We conclude that there exists $c = c(k, \beta) > 0$ such that condition C1 is satisfied for all $\delta \in (0, c)$. We will select the value of δ after checking condition C3.

Consider condition C3. From Eq. (31), it follows that $h_{G,\beta}(\boldsymbol{x}) \geq 1$ (because the argument in the exponent is non-negative) and $h_{G,\beta}(\boldsymbol{x}) \leq \exp(2\beta |E|)$ with $|E|$ the number of edges (the upper bound is saturated if the graph is bipartite). Since $|E| = kp/2$, Eq. (7) follows with $K(p) = \beta \, kp$.

At this point we choose $\delta = 1/(10pK(p))$. Assuming, without loss of generality, $K(p), p \geq 10$, this implies that we can take $\varepsilon = 2$ in Eq. (8), which yields the desired contradiction with Theorem 2, provided we can verify condition C2 with the stated value of δ.

To conclude our proof, consider condition C2. First we claim that $\boldsymbol{\tau}_*(\boldsymbol{\theta}) \in [\delta, 1-\delta]^p$ implies $\|\boldsymbol{\theta}\|_\infty \leq C \, \log p$ for some constant $C = C(k, \beta)$. Indeed, let us fix $i \in [p]$ and prove that $\theta_i \leq C \, \log p$ (the lower bound follows from an analogous argument). Using again the fact that $p_{\boldsymbol{\theta}}$ is a Markov Random Field, we have (since, by definition $\tau_{*,i}(\boldsymbol{\theta}) = \mathbb{P}_{\boldsymbol{\theta}}(X_i = 1)$)

$$1 - \delta \geq \mathbb{P}_{\boldsymbol{\theta}}(X_i = 1) \geq \min_{\boldsymbol{x}_{\partial i} \in \{0,1\}^{\partial i}} \mathbb{P}_{\boldsymbol{\theta}}(X_i = 1 | \boldsymbol{X}_{\partial i} = \boldsymbol{x}_{\partial i}) \, . \tag{37}$$

Proceeding as in the case of condition C1, we see that the last conditional probability is minimized when all the neighbors of i are in state 0 (i.e. $\boldsymbol{x}_{\partial i} = \mathbf{0}$). This yields

$$1 - \delta \geq \frac{e^{\theta_i}}{e^{\theta_i} + e^{2\beta k}}, \tag{38}$$

which is equivalent to $\theta_i \leq 2\beta k + \log((1-\delta)/\delta)$. Substituting $\delta = 1/(10pK(p))$ with $K(p)$ a polynomial yields the desired bound (recall that β and k are constants).

Next, we claim that, there exists a polynomial $L_0(p)$ such that, for each $i \in [p]$, all $\boldsymbol{x}_{\partial i}$ and all $\boldsymbol{\theta}$ with $|\theta_i| \leq C \log p$, we have

$$\mathsf{Var}_{\boldsymbol{\theta}}(X_i|\boldsymbol{X}_{\partial i} = \boldsymbol{x}_{\partial i}) \geq \frac{1}{L_0(p)}\,. \tag{39}$$

The calculation is essentially the same as the one already carried out above for $\mathbb{P}_{\boldsymbol{\theta}}(X_i = 1|\boldsymbol{X}_{\partial i} = \boldsymbol{x}_{\partial i})$ and therefore we omit it.

Finally, we want to prove Eq. (6) for some polynomial $L(p)$. Equivalently, we need to prove that $\mathsf{Var}_{\boldsymbol{\theta}}(\langle \boldsymbol{v}, \boldsymbol{X}\rangle) \geq 1/L(p)$ for any vector $\boldsymbol{v} \in \mathbb{R}^p$, $\|\boldsymbol{v}\|_2 = 1$. Fix one such vector $\boldsymbol{v}$. Let $i(1)$ be the index of the component of $\boldsymbol{v}$ with largest magnitude (i.e. $|v_{i(1)}| \geq |v_j|$ for all $j \in [p] \setminus \{i(1)\}$). Let $i(2)$ be the of the component of $\boldsymbol{v}$ with largest magnitude *excluded* $i(1)$ *and* $\partial i(1)$ (i.e. $|v_{i(2)}| \geq |v_j|$ for all $j \in [p] \setminus \{i(1), \partial i(1), i(2)\}$), and so on. Namely, for each ℓ, we let $i(\ell)$ be the index of the component of $\boldsymbol{v}$ with the largest magnitude *excluded* $i(1), \dots, i(\ell-1)$ *and their neighbors in* G. Denote by $m \geq n/(k+1)$ the total number of vertices selected in this manner, and let $S = \{i(1), \dots, i(m)\}$. It is immediate to see that

$$\sum_{i \in S} v_i^2 \geq \frac{\|\boldsymbol{v}\|_2^2}{k+1}\,. \tag{40}$$

Further, letting $S^c = [p] \setminus S$,

$$\begin{aligned} \mathsf{Var}_{\boldsymbol{\theta}}(\langle \boldsymbol{v}, \boldsymbol{X}\rangle) &\geq \mathsf{Var}_{\boldsymbol{\theta}}(\langle \boldsymbol{v}, \boldsymbol{X}\rangle | \boldsymbol{X}_{S^c}) & (41)\\ &= \sum_{i \in S} \mathsf{Var}_{\boldsymbol{\theta}}(v_i X_i | \boldsymbol{X}_{S^c}) & (42)\\ &= \sum_{i \in S} v_i^2 \,\frac{1}{L_0(p)} \geq \frac{\|\boldsymbol{v}\|_2^2}{(k+1)L_0(p)}\,. & (43) \end{aligned}$$

Here the identity in Eq. (42) follows because the $(X_i)_{i \in S}$ are conditionally independent given $\boldsymbol{X}_{S^c}$ (note that S is an independent set in G, i.e. there is no edge connecting two vertices in S), and because $(X_i)_{i \in S^c}$ constant given $\boldsymbol{X}_{S^c}$ The expressions in Eq. (42) follow from Eqs. (39) and (40).

We therefore established also condition $\mathsf{C2}$, with $L(p) = (k+1)L_0(p)$. This finishes the proof. □

3.2 Tractable estimation from samples

In this section we assume to be given n i.i.d. samples $\boldsymbol{X}_1, \boldsymbol{X}_2, \dots, \boldsymbol{X}_n \sim p_{\boldsymbol{\theta}}$ and denote by $\boldsymbol{X}^{(n)} = (\boldsymbol{X}_1, \boldsymbol{X}_2, \dots, \boldsymbol{X}_n)$ the entire dataset. We will seek an estimator $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(\boldsymbol{X}^{(n)}; \xi)$ that can be computed efficiently, and is consistent in the sense that, for a sequence $\xi_n \to 0$

$$\hat{\boldsymbol{\theta}}(\boldsymbol{X}^{(n)}; \xi_n) \xrightarrow{p} \boldsymbol{\theta}\,, \tag{44}$$

in $p_{\boldsymbol{\theta}}$-probability.

It is indeed fairly easy to construct such an estimator: we will use an approach developed in [AKN06]. Fix $i \in [p]$ and say we want to estimate θ_i. Let $N_i(x_i; \boldsymbol{x}_{\partial i})$ be number of samples ℓ such that $X_i^{(\ell)} = x_i$, $\boldsymbol{X}_{\partial i}^{(\ell)} = \boldsymbol{x}_{\partial i}$. In formulae

$$N_i(x_i; \boldsymbol{x}_{\partial i}) = \#\Big\{\ell \in [n]: \;\; X_i^{(\ell)} = x_i, \boldsymbol{X}_{\partial i}^{(\ell)} = \boldsymbol{x}_{\partial i}\Big\}\,. \tag{45}$$

Then by the law of large numbers we have, almost surely and in probability

$$\lim_{n\to\infty} \frac{N_i(1;\mathbf{0})}{N_i(0;\mathbf{0})} = \frac{\mathbb{P}_{\boldsymbol{\theta}}(X_i = 1, \boldsymbol{X}_{\partial i} = 0)}{\mathbb{P}_{\boldsymbol{\theta}}(X_i = 0, \boldsymbol{X}_{\partial i} = 0)} = \frac{\mathbb{P}_{\boldsymbol{\theta}}(X_i = 1 | \boldsymbol{X}_{\partial i} = 0)}{\mathbb{P}_{\boldsymbol{\theta}}(X_i = 0 | \boldsymbol{X}_{\partial i} = 0)} \, . \tag{46}$$

Now, an immediate generalization of Eq. (35) yields

$$\frac{\mathbb{P}_{\boldsymbol{\theta}}(X_i = 1 | \boldsymbol{X}_{\partial i} = 0)}{\mathbb{P}_{\boldsymbol{\theta}}(X_i = 0 | \boldsymbol{X}_{\partial i} = 0)} = e^{2\beta k + \theta_i} \, . \tag{47}$$

This immediately suggests the estimator

$$\hat{\theta}_i \equiv -2\beta k + \log \frac{N_i(1;\mathbf{0})}{N_i(0;\mathbf{0})} \, . \tag{48}$$

This can be obviously evaluated for all vertices $i \in [p]$, in time linear in p and in n. We next provide a non-asymptotic estimate on the number of samples necessary to achieve a desired level of accuracy.

Proposition 3.2. *Let $\xi > 0$ be a precision parameter and Δ an error probability, and let $\boldsymbol{\theta}$ be such that $\|\boldsymbol{\theta}\|_\infty \le \theta_{\max}$, $\theta_{\max} \ge 1$. Then, letting $\hat{\boldsymbol{\theta}}$ denote the estimator (48), we have*

$$\mathbb{P}_{\boldsymbol{\theta}}\Big(\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_\infty \le \xi \Big) \ge 1 - \Delta \, , \tag{49}$$

provided $n \ge e^{C_* \theta_{\max}} \xi^{-2} \log(p/\Delta)$, where $C_* = C_*(k, \beta)$ is a constant.

Proof. Fix $i \in [p]$, and let, for $x \in \{0, 1\}$,

$$q_x \equiv \mathbb{P}_{\boldsymbol{\theta}}(X_i = x; \boldsymbol{X}_{\partial i} = \mathbf{0}) \, . \tag{50}$$

Letting $D \subseteq [p]$, $|D| \le k(k-1)$, be the set of neighbors of $\{i\} \cup \partial i$ in G, we have, by the Markov property

$$\min_{\boldsymbol{x}_D \in \{0,1\}^D} \mathbb{P}_{\boldsymbol{\theta}}\big(X_i = x; \boldsymbol{X}_{\partial i} = \mathbf{0} \big| \boldsymbol{X}_D = \boldsymbol{x}_D\big) \le q_x \le \max_{\boldsymbol{x}_D \in \{0,1\}^D} \mathbb{P}_{\boldsymbol{\theta}}\big(X_i = x; \boldsymbol{X}_{\partial i} = \mathbf{0} \big| \boldsymbol{X}_D = \boldsymbol{x}_D\big) \, . \tag{51}$$

An explicit calculation shows that there exists a constant $C = C(k, \beta)$ such that

$$e^{-C\theta_{\max}} \le q_x \le 1 - e^{-C\theta_{\max}} \, . \tag{52}$$

Note that, for $x \in \{0, 1\}$, $N_i(x; \mathbf{0})$ is a binomial random variable $\text{Binom}(n, q_x)$. Standard tail bounds on binomial random variables yield (for $\varepsilon \le 1/2$)

$$\mathbb{P}_{\boldsymbol{\theta}}\Big(\big|N_i(x;\mathbf{0}) - \mathbb{E}N_i(x;\mathbf{0})\big| \ge \varepsilon \mathbb{E}N_i(x;\mathbf{0}) \Big) \le 2\, \exp\Big\{ - \frac{n\varepsilon^2}{8\min(q_x, 1 - q_x)} \Big\} \, . \tag{53}$$

Using Eq. (47) and the definition (48), together with Eq. (52), we get

$$\mathbb{P}_{\boldsymbol{\theta}}\Big(\big|e^{\hat{\theta}_i} - e^{\theta_i}\big| \ge \varepsilon e^{\theta_i} \Big) \le 2\, \exp\Big\{ - n e^{C\theta_{\max}} \varepsilon^2 \Big\} \, . \tag{54}$$

Passing to logarithms and using $|\log(1 + x)| \le C|x|$ for all $x \le 1/2$, this yields, for all $\varepsilon < 1/4$, and eventually a different constant C,

$$\mathbb{P}_{\boldsymbol{\theta}}\Big(\big|\hat{\theta}_i - \theta_i\big| \ge \xi \Big) \le 2\, \exp\Big\{ - n e^{C\theta_{\max}} \xi^2 \Big\} \, . \tag{55}$$

The proof is completed by choosing n so that the right hand side is upper bounded by Δ/p and using the union bound over the vertices p. □

4 Discussion and related literature

From a mathematical point of view, the phenomenon highlighted by Theorem 1 is not new. It can be traced back to two well-understood facts, and one simple remark:

1. First well-understood fact. The log partition function is the value of a convex optimization problem:

$$\log Z(\mathbf{0}) = \max \Big\{ F(\boldsymbol{\tau}) \; : \; \boldsymbol{\tau} \in [0,1]^p \Big\} . \tag{56}$$

2. Second well-understood fact. Recall that a weak evaluation oracle for F is an oracle that –on input $\boldsymbol{\tau}, \varepsilon$, with $\boldsymbol{\tau} \in [0,1]^p$ and $\varepsilon > 0$– returns $\widehat{F}$ such that $|F(\boldsymbol{\tau}) - \widehat{F}| \leq \varepsilon$. Given such an oracle, then the optimization problem (56) can be solved in polynomial time with accuracy $C(p)\varepsilon$, with $C(p)$ a polynomial.

   (See for instance [Lov87], Lemma 2.2.4 and Theorem 2.2.14 or [GLS81].)

3. Simple remark. We know that F is differentiable with Lipshitz continuous gradient for $\boldsymbol{\tau} \in [\delta, 1-\delta]^p$, by assumption $\mathsf{C2}$. Further, we showed quite easily that, letting $\boldsymbol{\tau}^{(0)} = (\delta, \dots, \delta)^{\mathsf{T}}$, we have $F(\boldsymbol{\tau}^{(0)}) \approx F(\mathbf{0}) = \log h(\mathbf{0})$. It follows that $F(\boldsymbol{\tau})$ can be –approximately– evaluated by 'integrating' the gradient ∇F between $\boldsymbol{\tau}^{(0)}$ and $\boldsymbol{\tau}$.

   Since we assume to have access to a gradient oracle (i.e. an oracle for $\nabla F(\boldsymbol{\tau}) = -\boldsymbol{\theta}_*(\boldsymbol{\tau})$), we can construct an evaluation oracle. Hence, by the previous facts we can approximate $\log Z(\mathbf{0})$.

In other words, we could have proven Theorem 1 by using the construction at point 3, and then referring to standard results in convex optimization [GLS81, Lov87]. We preferred a self-contained proof, that uses additional structure of the problem (differentiability of F and existence of an oracle for ∇F).

The specific example discussed in Section 3 can be regarded as a graphical model. While we used a specific class of models, namely antiferromagnetic Ising models defined by Eq. (31), approximating the partition function of a graphical model is –in many cases– intractable [GJP03, Sly10, SS12, GSV12, CCGL12]. Hence, the conclusion is that –generally speaking– *it is intractable to estimate the parameters of a graphical model from sufficient statistics* (unless the model has special structure).

The literature on estimating parameters and structure of a graphical model is quite vast, and we can only provide a few pointers here. Traditional methods [HS83, AHS85] attempt at maximizing the likelihood function by gradient ascent. These approaches are necessarily based on sufficient statistics, and hence covered by our Theorem 1: in general, they cannot be implemented in polynomial time. The bottleneck is quite apparent in this case: evaluating the likelihood function or its gradient requires to compute expectations with respect to $p_{\boldsymbol{\theta}}$ which –in general– is NP-hard. Notice that this difficulty is not circumvented by regularizing the likelihood function, as in the graphical LASSO [BEGd08].

A possible approach to overcome this problem was proposed in [RWL^{+}10]: the basic idea is to reconstruct the neighborhood of a node i by performing a logistic regression against the other nodes. This approach generalizes to discrete graphical models a method developed in [MB06] for Gaussian graphical models. As shown in [BM09], the logistic regression method fails unless the graphical model satisfies a 'weak dependence' condition. Under the same type of condition, the log partition function can be computed efficiently.

Several papers [AKN06, CT06, BMS08, RSS12, ATHW12] develop algorithm to learn parameters and graph structures, with consistency guarantees under weak assumptions on the graphical model. As stressed in Section 3.2, these algorithms do not make use exclusively of the sufficient statistics and instead effectively estimate the joint distributions of subsets of k variables, with k depending on the maximum degree.

On the impossibility side, Bogdanov, Mossel and Vadhan [BMV08] showed that estimating graphical models with hidden nodes is intractable. Singh and Vishnoi [SV13] establish a result that is very similar to Theorem 1 with a slightly different oracle assumption. Their proof uses the ellipsoid algorithm instead of projected gradient, and their motivation is related to maximum entropy rounding techniques in optimization algorithms. Roughgarden and Kearns [RK13] establish equivalence of various computational tasks in graphical models. Again, they use convex duality and the ellipsoid algorithm.

Finally, I recently became aware of unpublished work by Bresler, Gamarnik and Shah [BGS14], that also proves intractability of learning graphical models from sufficient statistics. Their proof strategy is broadly similar to the one presented here but differs in the solution of several technical obstacles.

Acknowledgements

This work was partially supported by the NSF grant CCF-1319979 and the grants AFOSR/DARPA FA9550-12-1-0411 and FA9550-13-1-0036. I am grateful to Guy Bresler for discussing [BGS14] prior to publication.

A Simple properties of exponential families

In this section we provide a self-contained proof of Proposition 2.3. Fact1 and Fact2 are immediate and we omit their proof.

Fact3. For any vector $\boldsymbol{v} \in \mathbb{R}^p$, $\|\boldsymbol{v}\|_2 = 1$, we have

$$\langle \boldsymbol{v}, \mathrm{Cov}_{\boldsymbol{\theta}}(\boldsymbol{X})\boldsymbol{v}\rangle = \mathsf{Var}_{\boldsymbol{\theta}}\Big(\langle \boldsymbol{v}, \boldsymbol{X}\rangle\Big) = \frac{1}{2}\,\mathbb{E}_{\boldsymbol{\theta}}\Big\{\langle \boldsymbol{v}, \boldsymbol{X} - \boldsymbol{X}'\rangle^2\Big\}\,, \tag{57}$$

with $\boldsymbol{X}$, $\boldsymbol{X}' \sim p_{\boldsymbol{\theta}}$ independent. Letting $\boldsymbol{s} = \mathrm{sign}(\boldsymbol{v})$, and continuing the above chain of inequalities, we get

$$\langle \boldsymbol{v}, \mathrm{Cov}_{\boldsymbol{\theta}}(\boldsymbol{X})\boldsymbol{v}\rangle \geq 2\cdot\frac{1}{2} p_{\boldsymbol{\theta}}(\boldsymbol{s})\, p_{\boldsymbol{\theta}}(-\boldsymbol{s})\langle \boldsymbol{v}, \boldsymbol{s} - (-\boldsymbol{s})\rangle^2 \tag{58}$$

$$\geq p_{\boldsymbol{\theta}}(\boldsymbol{s})\, p_{\boldsymbol{\theta}}(-\boldsymbol{s})\|\boldsymbol{v}\|_1^2 \geq \min_{\boldsymbol{x}\in\{0,1\}^n} p_{\boldsymbol{\theta}}(\boldsymbol{x})^2\,, \tag{59}$$

which yields the desired claim.

Fact4. It is convenient to extend by continuity $\boldsymbol{\tau}_* : \overline{\mathbb{R}} \to [0,1]^p$, where $\overline{\mathbb{R}} \equiv \mathbb{R}\cup\{+\infty,-\infty\}$ is the extended real line. It is a simple analysis exercise to check that this extension exists and is well defined. Indeed, if $\boldsymbol{\theta} \in \overline{\mathbb{R}}^n$ is such that $\theta_i = +\infty$ for all $i \in S_+ \subseteq [n]$, $\theta_i = -\infty$ for all $i \in S_- \subseteq [n]$, and $\theta_i \in (-\infty, +\infty)$ for all $i \in S_0 \equiv [n] \setminus (S_+ \cup S_-)$, then we have

$$\boldsymbol{\tau}_*(\boldsymbol{\theta}) = \frac{1}{\widetilde{Z}(\boldsymbol{\theta})} \sum_{\substack{\boldsymbol{x}\in\{0,1\}^n \text{ s.t.}\\ \boldsymbol{x}_{S_+}=1, \boldsymbol{x}_{S_-}=0}} \boldsymbol{x}\; h(\boldsymbol{x})\; \exp\Big\{\langle \boldsymbol{\theta}_{S_0}, \boldsymbol{x}_{S_0}\rangle\Big\}\,, \tag{60}$$

$$\widetilde{Z}(\boldsymbol{\theta}) \equiv \sum_{\substack{\boldsymbol{x}\in\{0,1\}^n \text{ s.t.}\\ \boldsymbol{x}_{S_+}=1, \boldsymbol{x}_{S_-}=0}} h(\boldsymbol{x})\; \exp\Big\{\langle \boldsymbol{\theta}_{S_0}, \boldsymbol{x}_{S_0}\rangle\Big\}\,, \tag{61}$$

We next prove that $\boldsymbol{\tau}_*$ is injective. Indeed, assume by contradiction that $\boldsymbol{\tau}_*(\boldsymbol{\theta}_1) = \boldsymbol{\tau}_*(\boldsymbol{\theta}_2)$ for $\boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_2$. Then, by the intermediate value theorem, there exists $\lambda \in [0,1]$ such that, letting $\boldsymbol{\theta}_\lambda = \lambda\, \boldsymbol{\theta}_1 + (1-\lambda)\, \boldsymbol{\theta}_2$,

$$\boldsymbol{\tau}_*(\boldsymbol{\theta}_1) - \boldsymbol{\tau}_*(\boldsymbol{\theta}_2) = \nabla_{\boldsymbol{\theta}} \boldsymbol{\tau}_*(\boldsymbol{\theta}_\lambda)(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)\,, \tag{62}$$

and hence, letting $\boldsymbol{v} = \boldsymbol{\theta}_1 - \boldsymbol{\theta}_2$,

$$0 = \langle \boldsymbol{v}, \boldsymbol{\tau}_*(\boldsymbol{\theta}_1) - \boldsymbol{\tau}_*(\boldsymbol{\theta}_2)\rangle = \langle \boldsymbol{v}, \mathrm{Cov}_{\boldsymbol{\theta}}(\boldsymbol{\theta}_\lambda)\boldsymbol{v}\rangle\,, \tag{63}$$

which is impossible by Fact 3 above.

In order to prove that $\boldsymbol{\tau}_* : \overline{\mathbb{R}}^p \to [0,1]^p$ is surjective, we will proceed by induction over the problem's dimension p. The claim is obvious for $p = 1$. We assume next that it holds for all dimensions up to $(p-1)$ and prove it for dimension p. This claim follows by continuity, after proving that the image of $\boldsymbol{\tau}_*$ contains the boundary $B_p \equiv [0,1]^p \setminus (0,1)^p$. Namely, for each $\boldsymbol{\tau} \in B_p$, there exists $\boldsymbol{\theta} \in \overline{\mathbb{R}}^p$ such that $\boldsymbol{\tau}_*(\boldsymbol{\theta}) = \boldsymbol{\tau}$.

To see that this is the case, fix one such $\boldsymbol{\tau}$. Let $S_+ \equiv \{i \in [n] : \tau_i = 1\}$, $S_- \equiv \{i \in [n] : \tau_i = 0\}$, and $S_0 \equiv [p] \setminus (S_+ \cup S_-)$ and note that, by assumption $|S_0| \leq n-1$. Take $\boldsymbol{\theta}$ such that $\theta_i = +\infty$ for all $i \in S_+$ and $\theta_i = -\infty$ for all $i \in S_-$. By Eq. (60), we have $\boldsymbol{\tau}_*(\boldsymbol{\theta})_{S_+ \cup S_-} = \boldsymbol{\tau}_{S_+ \cup S_-}$ and it is therefore sufficient to check the values of $\boldsymbol{\tau}_*(\boldsymbol{\theta})$ on S_0. Let $h_{S_+,S_-}(\boldsymbol{x}_{S_0}) = h(\boldsymbol{x})$ where $\boldsymbol{x}_{S_+} = 1_{S_+}$ and $\boldsymbol{x}_{S_-} = 0_{S_-}$. Then, again by Eq. (60), we have

$$\boldsymbol{\tau}_*(\boldsymbol{\theta})_{S_0} = \frac{1}{\widetilde{Z}(\boldsymbol{\theta})} \sum_{\boldsymbol{x}_{S_0}\in\{0,1\}^{S_0}} \boldsymbol{x}_{S_0}\; h_{S_+,S_-}(\boldsymbol{x}_{S_0})\; \exp\Big\{\langle \boldsymbol{\theta}_{S_0}, \boldsymbol{x}_{S_0}\rangle\Big\}\,, \tag{64}$$

$$\widetilde{Z}(\boldsymbol{\theta}) = \sum_{\boldsymbol{x}_{S_0}\in\{0,1\}^{S_0}} h_{S_+,S_-}(\boldsymbol{x}_{S_0})\; \exp\Big\{\langle \boldsymbol{\theta}_{S_0}, \boldsymbol{x}_{S_0}\rangle\Big\}\,. \tag{65}$$

Comparison with Eq. (3) shows that –by the induction hypothesis– there exists $\boldsymbol{\theta}_{S_0}$ such that $\boldsymbol{\tau}_*(\boldsymbol{\theta})_{S_0} = \boldsymbol{\tau}_{S_0}$.

Fact5. This is a standard exercise with Legendre-Fenchel transforms and we omit its proof.

Fact6. Recall the Gibbs variational principle [MM09]

$$A(\boldsymbol{\theta}) = \max_{q\in\mathcal{P}_p} \Big\{ H(q) + \mathbb{E}_q\big(\log h(\boldsymbol{X}) + \langle \boldsymbol{\theta}, \boldsymbol{X}\rangle\big)\Big\}\,. \tag{66}$$

The claim follows by comparing this expression with Eq. (13).

B Finishing the proof of Theorem 1

In this appendix we complete the proof of Theorem 1. In the simplified version presented in Section 2, we assumed to have access to an oracle returning $\boldsymbol{\theta}_*(\boldsymbol{\tau})$ when queried with value $\boldsymbol{\tau}$. We will show that our claim remains correct if the oracle is replaced by a polynomial consistent estimator $\hat{\boldsymbol{\theta}}(\,\cdot\,,\,\cdot\,)$. By Definition 2.1, for any $\xi > 0$ we have

$$\|\hat{\boldsymbol{\theta}}(\boldsymbol{\tau}, \xi) - \boldsymbol{\theta}_*(\boldsymbol{\tau})\|_2 \le \xi\,. \tag{67}$$

The oracle $\boldsymbol{\theta}_*(\,\cdot\,)$ is used in two points in the proof of Theorem 1:

1. In the implementation of the projected gradient algorithm, cf. Eq. (21). It is queried a total of $t_0(p, \varepsilon) \equiv 2p\, L(p)/\varepsilon$ times for this purpose.

2. In calculating the approximation of $Z(\mathbf{0})$, cf. Eq. (27). It is queried a total of $m_0(p, \varepsilon) \equiv \lceil 4L(p)p/\varepsilon \rceil$ times for this purpose.

We will replace these queries with queries to $\hat{\boldsymbol{\theta}}(\,\cdot\,, \xi)$ with $1/\xi$ polynomial in p and $1/\varepsilon$. Since the total number of calls is also polynomial in p and $1/\xi$, this yields an algorithm that is polynomial in p and $1/\varepsilon$.

Queries for projected gradient. We denote by $\boldsymbol{\sigma}^0$, $\boldsymbol{\sigma}^1$,$\dots \boldsymbol{\sigma}^{t_0}$ the sequence generated by the projected gradient algorithm, with $\nabla_{\boldsymbol{\tau}} F(\boldsymbol{\tau})$ approximated by $-\hat{\boldsymbol{\theta}}(\boldsymbol{\tau}, \xi)$. Namely, we have $\boldsymbol{\sigma}^0 = \boldsymbol{\tau}^0$ and, for all $t \ge 0$,

$$\boldsymbol{\sigma}^{t+1} = \mathsf{P}_\delta\Big(\boldsymbol{\sigma}^t - \frac{1}{L}\hat{\boldsymbol{\theta}}(\boldsymbol{\sigma}^t, \xi)\Big)\,. \tag{68}$$

Comparing with Eq. (21), we have (dropping the argument ξ of $\hat{\boldsymbol{\theta}}$ in order to simplify the notation):

$$\|\boldsymbol{\tau}^{t+1} - \boldsymbol{\sigma}^{t+1}\|_2 \le \Big\|\Big(\boldsymbol{\tau}^t - \frac{1}{L}\hat{\boldsymbol{\theta}}(\boldsymbol{\tau}^t)\Big) - \Big(\boldsymbol{\sigma}^t - \frac{1}{L}\hat{\boldsymbol{\theta}}(\boldsymbol{\sigma}^t)\Big)\Big\|_2 \tag{69}$$

$$\le \Big\|\boldsymbol{\tau}^t + \frac{1}{L}\nabla_{\boldsymbol{\tau}} F(\boldsymbol{\tau}^t) - \boldsymbol{\sigma}_t - \frac{1}{L}\nabla_{\boldsymbol{\tau}} F(\boldsymbol{\sigma}^t)\Big\|_2 \tag{70}$$

$$+ \frac{1}{L}\|\hat{\boldsymbol{\theta}}(\boldsymbol{\sigma}_t) - \boldsymbol{\theta}_*(\boldsymbol{\sigma}_t)\|_2 + \frac{1}{L}\|\hat{\boldsymbol{\theta}}(\boldsymbol{\tau}_t) - \boldsymbol{\theta}_*(\boldsymbol{\tau}_t)\|_2 \tag{71}$$

$$\le \Big\|\Big(\mathrm{I} + \frac{1}{L}\nabla^2 F(\boldsymbol{\rho}^t)\Big)(\boldsymbol{\tau}^t - \boldsymbol{\sigma}_t)\|_2 + \frac{2\xi}{L}\,, \tag{72}$$

where the last inequality follows from the definition of $\hat{\boldsymbol{\theta}}$, and by applying the intermediate value theorem, with $\boldsymbol{\rho}^t$ a point on the interval $[\boldsymbol{\tau}^t, \boldsymbol{\sigma}^t]$. Now, by Eq. (12), (15) and assumption $\mathsf{C2}$, we have $-L \preceq \nabla^2 F(\boldsymbol{\rho}^t) \preceq \mathbf{0}$, and therefore $\|\mathrm{I} + L^{-1}\nabla^2 F(\boldsymbol{\rho}^t)\|_2 \le 1$ (in operator norm). We get therefore

$$\|\boldsymbol{\tau}^{t+1} - \boldsymbol{\sigma}^{t+1}\|_2 \le \|\boldsymbol{\tau}^t - \boldsymbol{\sigma}^t\|_2 + \frac{2\xi}{L}\,, \tag{73}$$

which, of course, implies

$$\|\boldsymbol{\tau}^{t_0} - \boldsymbol{\sigma}^{t_0}\|_2 \le \frac{2t_0\,\xi}{L}\,. \tag{74}$$

Notice that, for any $\boldsymbol{\tau} \in D_p(\delta)$,

$$\|\nabla_{\boldsymbol{\tau}} F(\boldsymbol{\tau})\|_2 = \|\nabla_{\boldsymbol{\tau}} F(\boldsymbol{\tau}) - \nabla_{\boldsymbol{\tau}} F(\boldsymbol{\tau}_*)\|_2 \le L\,\|\boldsymbol{\tau} - \boldsymbol{\tau}_*\|_2 \le L\sqrt{p}\,. \tag{75}$$

Hence

$$\big|F(\boldsymbol{\tau}^{t_0}) - F(\boldsymbol{\sigma}^{t_0})\big| \le L\sqrt{p}\,\|\boldsymbol{\tau}^{t_0} - \boldsymbol{\sigma}^{t_0}\|_2 \le 2t_0\,\sqrt{p}\xi\,. \tag{76}$$

In particular, by choosing $\xi \le \varepsilon/(16t_0(p,\varepsilon)\sqrt{p})$, it follows that $|F(\boldsymbol{\tau}^{t_0}) - F(\boldsymbol{\sigma}^{t_0})| \le \varepsilon/8$.

Queries for evaluating Eq. (27). We let $\boldsymbol{\sigma}^{(0)} = \boldsymbol{\tau}^{(0)}$, $\boldsymbol{\sigma}^{(m)} = \sigma^{t_0}$ and $\boldsymbol{\sigma}^{(\ell)}$, $\ell \in \{1, \dots, m-1\}$ be given by interpolating linearly. The approximation $\widehat{Z}$ in Eq. (27) is replaced by

$$\log \widehat{Z} = \log h(\mathbf{0}) - \sum_{\ell=1}^{m} \langle \boldsymbol{\theta}_*(\boldsymbol{\tau}^{(\ell)}), \boldsymbol{\tau}^{(\ell)} - \boldsymbol{\tau}^{(\ell-1)}\rangle\,. \tag{77}$$

The approximation error is bounded analogously to Eq. (28). We get two additional error terms

$$\begin{aligned}
\Big|\log\frac{\widehat{Z}}{Z}\Big| &\le \frac{3\varepsilon}{4} + |F(\boldsymbol{\tau}^{t_0}) - F(\boldsymbol{\sigma}^{t_0})| + \Big|\sum_{\ell=1}^{m} \langle \hat{\boldsymbol{\theta}}(\boldsymbol{\sigma}^{(\ell)}) - \boldsymbol{\theta}_*(\boldsymbol{\sigma}^{\ell}), \boldsymbol{\sigma}^{(\ell)} - \boldsymbol{\sigma}^{(\ell-1)}\rangle\Big| && (78)\\
&\le \frac{7\varepsilon}{8} + \|\boldsymbol{\sigma}^{(m)} - \boldsymbol{\sigma}^{(0)}\|_2 \max_{\ell\in[m]} \|\hat{\boldsymbol{\theta}}(\boldsymbol{\sigma}^{(\ell)}) - \boldsymbol{\theta}_*(\boldsymbol{\sigma}^{\ell})\|_2 && (79)\\
&\le \frac{7\varepsilon}{8} + \xi\sqrt{p}\,. && (80)
\end{aligned}$$

The latter is bounded by ε as soon as $\xi \le 1/(8\varepsilon\sqrt{p})$, which is guaranteed since we already required $\xi \le \varepsilon/(16t_0(p,\varepsilon)\sqrt{p})$.

We conclude that the desired approximation error is guaranteed by using the oracle $\hat{\boldsymbol{\theta}}$, with accuracy parameter $\xi \le \varepsilon/(16t_0(p,\varepsilon)\sqrt{p})$. This concludes the proof, since it corresponds to $(1/\xi)$ polynomial in p and $(1/\varepsilon)$.

References

[AHS85] David H Ackley, Geoffrey E Hinton, and Terrence J Sejnowski, *A learning algorithm for boltzmann machines*, Cognitive science **9** (1985), no. 1, 147–169.

[AKN06] Pieter Abbeel, Daphne Koller, and Andrew Y Ng, *Learning factor graphs in polynomial time and sample complexity*, The Journal of Machine Learning Research **7** (2006), 1743–1788.

[ATHW12] Animashree Anandkumar, Vincent YF Tan, Furong Huang, and Alan S Willsky, *High-dimensional structure estimation in Ising models: Local separation criterion*, The Annals of Statistics **40** (2012), no. 3, 1346–1375.

[BEGd08] Onureena Banerjee, Laurent El Ghaoui, and Alexandre d'Aspremont, *Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data*, The Journal of Machine Learning Research **9** (2008), 485–516.

[BGS14] Guy Bresler, David Gamarnik, and Devavrat Shah, *Hardness of parameter estimation in graphical models*, unpublished, 2014.

[BM09] José Bento and Andrea Montanari, *Which graphical models are difficult to learn?*, Neural Information Processing Systems (Vancouver), December 2009.

[BMS08] Guy Bresler, Elchanan Mossel, and Allan Sly, *Reconstruction of markov random fields from samples: Some observations and algorithms*, Approximation, Randomization and Combinatorial Optimization. Algorithms and Techniques, Springer, 2008, pp. 343–356.

[BMV08] Andrej Bogdanov, Elchanan Mossel, and Salil Vadhan, *The complexity of distinguishing markov random fields*, Approximation, Randomization and Combinatorial Optimization. Algorithms and Techniques, Springer, 2008, pp. 331–342.

[BT09] Amir Beck and Marc Teboulle, *A fast iterative shrinkage-thresholding algorithm for linear inverse problems*, SIAM Journal on Imaging Sciences **2** (2009), no. 1, 183–202.

[CCGL12] Jin-Yi Cai, Xi Chen, Heng Guo, and Pinyan Lu, *Inapproximability after uniqueness phase transition in two-spin systems*, Combinatorial Optimization and Applications, Springer, 2012, pp. 336–347.

[CT06] Imre Csiszár and Zsolt Talata, *Consistent estimation of the basic neighborhood of Markov random fields*, The Annals of Statistics (2006), 123–145.

[Efr78] Bradley Efron, *The geometry of exponential families*, The Annals of Statistics **6** (1978), no. 2, 362–376.

[GJP03] Leslie Ann Goldberg, Mark Jerrum, and Mike Paterson, *The computational complexity of two-state spin systems*, Random Structures & Algorithms **23** (2003), no. 2, 133–154.

[GLS81] Martin Grötschel, László Lovász, and Alexander Schrijver, *The ellipsoid method and its consequences in combinatorial optimization*, Combinatorica **1** (1981), no. 2, 169–197.

[GSV12] Andreas Galanis, Daniel Stefankovic, and Eric Vigoda, *Inapproximability of the partition function for the antiferromagnetic ising and hard-core models*, arXiv:1203.2226 (2012).

[HS83] GE Hinton and TJ Sejnowski, *Analysing cooperative computation*, Proceedings of the Fifth Annual Conference of the Cognitive Science Society, 1983.

[LC98] EL Lehmann and George Casella, *Theory of point estimation*, 2 ed., Springer, 1998.

[Lov87] László Lovász, *An algorithmic theory of numbers, graphs and convexity*, vol. 50, SIAM, 1987.

[MB06] Nicolai Meinshausen and Peter Bühlmann, *High-dimensional graphs and variable selection with the lasso*, Ann. Statist. **34** (2006), 1436–1462.

[MM09] Marc Mézard and Andrea Montanari, *Information, Physics and Computation*, Oxford, 2009.

[RK13] Tim Roughgarden and Michael Kearns, *Marginals-to-models reducibility*, Advances in Neural Information Processing Systems, 2013, pp. 1043–1051.

[RSS12] Avik Ray, Sujay Sanghavi, and Sanjay Shakkottai, *Greedy learning of graphical models with small girth*, Communication, Control, and Computing (Allerton), 2012 50th Annual Allerton Conference on, IEEE, 2012, pp. 2024–2031.

[RWL$^+$10] Pradeep Ravikumar, Martin J Wainwright, John D Lafferty, et al., *High-dimensional Ising model selection using ℓ_1-regularized logistic regression*, The Annals of Statistics **38** (2010), no. 3, 1287–1319.

[Sly10] Allan Sly, *Computational transition at the uniqueness threshold*, Foundations of Computer Science (FOCS), 2010 51st Annual IEEE Symposium on, IEEE, 2010, pp. 287–296.

[SS12] Allan Sly and Nike Sun, *The computational hardness of counting in two-spin models on d-regular graphs*, Foundations of Computer Science (FOCS), 2012 IEEE 53rd Annual Symposium on, IEEE, 2012, pp. 361–369.

[SV13] Mohit Singh and Nisheeth K Vishnoi, *Entropy, optimization and counting*, arXiv preprint arXiv:1304.8108 (2013).