Probab. Theory Relat. Fields (2007) 138:305–321
DOI 10.1007/s00440-006-0029-y

# Stein's method for concentration inequalities

**Sourav Chatterjee**

Received: 17 May 2006 / Revised: 28 August 2006 /
Published online: 19 October 2006


**Abstract** We introduce a version of Stein's method for proving concentration and moment inequalities in problems with dependence. Simple illustrative examples from combinatorics, physics, and mathematical statistics are provided.

**Keywords** Concentration inequalities · Random permutations · Gibbs measures · Stein's method · Curie–Weiss model · Ising model

**Mathematics Subject Classification (2000)** 60E15 · 60C05 · 60K35 · 82C22

## 1 Introduction and results

Stein's method was introduced by Charles Stein [38] in the context of normal approximation for sums of dependent random variables. Stein's version of his method, best known as the "method of exchangeable pairs", attained maturity in his later work [39]. A reasonably large literature has developed around the subject, but it has almost exclusively developed as a method of proving distributional convergence with error bounds. Stein's attempts at getting large deviations in [39] did not, unfortunately, prove fruitful. Some progress for sums of dependent random variables was made by Raič [33]. A general version of Stein's method for concentration inequalities was introduced for the first time in the Ph.D. thesis [11] of the present author. The purpose of this paper is

S. Chatterjee (✉)
Department of Statistics, University of California, 367 Evans Hall #3860,
Berkeley, CA 94720-3860, USA
e-mail: sourav@stat.berkeley.edu
URL: http://www.stat.berkeley.edu/∼sourav

to explain the theory developed in [11] via examples. Another application is in [12].

This section is organized as follows: First, we give three examples, followed by the main abstract theorem; finally, towards the end of the section, we present very condensed overviews of Stein's method, concentration of measure, and the related literature. Proofs are in Sect. 2.

### 1.1 A generalized matching problem

Let $\{a_{ij}\}$ be an $n \times n$ array of real numbers. Let $\pi$ be chosen uniformly at random from the set of all permutations of $\{1, \ldots, n\}$, and let $X = \sum_{i=1}^{n} a_{i\pi(i)}$. This class of random variables was first studied by Hoeffding [24], who proved that they are approximately normally distributed under certain conditions. It is easy to see that various well-studied functions of random permutations, like the number of fixed points, the sum of a random sample picked without replacement from a finite population, and the function $\sum_i |i - \pi(i)|$ (known as Spearman's footrule [16]), are all instances of Hoeffding's statistic.

Hoeffding's statistic has a long history of association with Stein's method. In fact, in an unpublished work Stein introduced his method to treat the normal approximation problem for this object. Bolthausen [7] used Stein's method to give a Berry–Esseen bound. Bolthausen and Götze [8] gave multivariate central limit theorems under a further generalized setup. However, we have not seen large deviations or concentration bounds using any method.

Our version of Stein's method enables us to easily derive the following nice tail bound.

**Proposition 1.1** *Let $\{a_{ij}\}_{1 \le i,\, j \le n}$ be a collection of numbers from $[0, 1]$. Let $X = \sum_{i=1}^{n} a_{i\pi(i)}$, where $\pi$ is drawn from the uniform distribution over the set of all permutations of $\{1, \ldots, n\}$. Then*

$$\mathbb{P}\{|X - \mathbb{E}(X)| \ge t\} \le 2 \exp\left(-\frac{t^2}{4\mathbb{E}(X) + 2t}\right)$$

*for any $t \ge 0$.*

Note that the bound does not have an explicit dependence on $n$. Note also the automatic transition from Poissonian to gaussian tails as $\mathbb{E}(X)$ becomes large (when $\mathbb{E}(X)$ is small the bound is like $\exp(-Ct)$, whereas when $\mathbb{E}(X)$ is large, it is essentially a gaussian tail with standard deviation $\sqrt{\mathbb{E}(X)}$.). These two properties characterize it as a so-called "Bernstein type inequality", named after the classical Bernstein inequality (see [37], page 855) for sums of bounded independent random variables.

The classical result of Maurey [30] can only imply the weaker inequality $P(X > \mathbb{E}(X) + t) \le \mathrm{e}^{-t^2/4n}$. However, it is possible to derive a Bernstein bound similar to Proposition 1.1 (albeit with a significantly worse constant in the exponent) using Michel Talagrand's deep theorem about concentration of random

permutations (Theorem 5.1 in Sect. 5 of [40]; see also McDiarmid [31] and Luczak and McDiarmid [29]).

For a concrete application, let $X$ be the number of fixed points of a random permutation $\pi$. Then $X = \sum_{i=1}^{n} a_{i\pi(i)}$, where $a_{ij} = \mathbb{I}_{\{i=j\}}$. Since $\mathbb{E}(X) = 1$, Proposition 1.1 gives $\mathbb{P}\{|X - 1| \geq t\} \leq 2\exp(-t^2/(4 + 2t))$. Of course, we do not expect this to be the best possible bound in this very well-understood problem; this is just meant to be an illustration. In fact, the exact distribution of the the number of fixed points is known (see Feller [19], section IV.4), which gives a tail bound like $\exp(-Ct \log t)$.

Finally, we also have a "Burkholder-Davis-Gundy" type inequality for Hoeffding's statistic which does not require a bound on the $a_{ij}$'s.

**Proposition 1.2** *Let $\{a_{ij}\}_{1\leq i,\, j\leq n}$ be an arbitrary collection of real numbers. Let $\pi$ be a uniform random permutation, and let $X = \sum_{i=1}^{n} a_{i\pi(i)}$. Define*

$$\Delta = \frac{1}{4n} \sum_{i,j} (a_{i\pi(i)} + a_{j\pi(j)} - a_{i\pi(j)} - a_{j\pi(i)})^2.$$

*Then for every positive integer k, we have $\mathbb{E}(X - \mathbb{E}(X))^{2k} \leq (2k - 1)^k \mathbb{E}\Delta^k$.*

For a general exposition about the famous Burkholder–Davis–Gundy martingale inequalities we refer to the article by Burkholder [10].

## 1.2 Magnetization in the Curie–Weiss model

Fix any $\beta \geq 0$, $h \in \mathbb{R}$, and consider the probability mass function (the Gibbs measure) on $\{-1, 1\}^n$ given by

$$\mathbb{P}(\{\sigma\}) := Z^{-1} \exp\left(\frac{\beta}{n} \sum_{i<j} \sigma_i \sigma_j + \beta h \sum_i \sigma_i\right), \tag{1}$$

where $\sigma = (\sigma_1, \ldots, \sigma_n)$ is a typical element of $\{-1, 1\}^n$ and $Z$ is the normalizing constant (depends on $\beta$ and $h$). This is known as the 'Curie–Weiss model of ferromagnetic interaction' at inverse temperature $\beta$ and external field $h$. The $\sigma_i$'s stand for the spins of $n$ particles, each having a spin of $+1$ or $-1$. The ferromagnetic interaction between the particles is captured in a very simplistic manner by the first term in the hamiltonian.

The *magnetization* of the system, as a function of the configuration $\sigma$, is defined as $m(\sigma) := \frac{1}{n} \sum_{i=1}^{n} \sigma_i$. If $n$ is large and $\sigma$ is drawn from the Gibbs measure, then the magnetization satisfies

$$m(\sigma) \approx \tanh(\beta m(\sigma) + \beta h) \tag{2}$$

with high probability. The equation has a unique root for small values of $\beta$ and multiple solutions for $\beta$ above a critical value. In the physics parlance, this is

described by saying that the Curie–Weiss model exhibits "spontaneous magnetization" at low temperatures. For a formal discussion with rigorous proofs, we refer to Ellis [18], Sect. IV.4.

The following proposition formalizes (2) with finite sample tail bounds.

**Proposition 1.3** *Suppose $\sigma$ is drawn from the Gibbs measure* (1)*. Then, for any $\beta \geq 0$, $h \in \mathbb{R}$, $n \geq 1$, and $t \geq 0$, the magnetization $m := \frac{1}{n}\sum_i \sigma_i$ satisfies*

$$\mathbb{P}\left\{\left|m - \tanh(\beta m + \beta h)\right| \geq \frac{\beta}{n} + \frac{t}{\sqrt{n}}\right\} \leq 2\exp\left(-\frac{t^2}{4(1+\beta)}\right).$$

Although the Curie–Weiss model is a simple model of ferromagnetic interaction, we haven't encountered any result in the literature which gives an explicit bound like the above. In particular, the result shows concentration of $m(\sigma)$ around the *set* of roots of $x = \tanh(\beta x + \beta h)$, and not just its mean.

However, concentration inequalities for Gibbs measures without explicit constants under various mixing conditions have been obtained before. For a history of the literature and some significant recent progress, we refer to Chazottes et al. [14].

### 1.3 Least squares estimation in the Ising model

The Ising model is another model of ferromagnetic interaction. Given an undirected graph $G = (V, E)$ on the vertex set $V = \{1, \ldots, n\}$, the Ising model without external field assigns the following probability density on $\{-1, 1\}^n$:

$$\mathbb{P}(\{\sigma\}) = Z(\beta)^{-1}\exp\left(\beta \sum_{\{i,j\}\in E} \sigma_i\sigma_j\right). \tag{3}$$

Here, as before, $\beta$ is the inverse temperature and $Z(\beta)$ is the normalizing constant. A natural statistical problem in this model is the following: How to make inference about $\beta$ when your data is a single configuration generated from the Gibbs measure?

The classical maximum likelihood approach for this problem was first considered by Pickard [32]. Iterative methods for computing the maximum likelihood estimator (e.g. Geyer and Thompson [22], Jerrum and Sinclair [26]) are widely used nowadays. The Jerrum–Sinclair algorithm for computing the normalizing constant in the Ising model provably converges in polynomial time. However, it is not so clear whether the MLE is a good estimator at all, particularly at critical temperatures.

Here we investigate a method of estimating $\beta$ by minimizing an explicit sum-of-squares. First, let $\sigma$ be drawn from the Gibbs measure (3) on $\{-1, 1\}^n$, and for each $i$, let

$$m_i := \sum_{j:\{i,j\}\in E} \sigma_j.$$

For each $u \geq 0$, let

$$S(u) := \frac{1}{n} \sum_{i=1}^{n} (\sigma_i - \tanh(um_i))^2. \tag{4}$$

The 'least-squares estimate' of $\beta$ is defined to be

$$\hat{\beta}_{LS} := \operatorname{argmin}_{u \geq 0} S(u).$$

Note that it is practically very easy to compute $\hat{\beta}_{LS}$, because $S$ is a smooth function of a single variable.

The least-squares technique is well-known and commonly used in the analysis of gaussian Markov random field (GMRF) models (probably originating from Besag [6]), but rigorous results are scarce.

Proposition 1.4 (stated below) shows that the random function $S$ indeed attains an approximate global minimum near $\beta$. In fact, it gives

$$\mathbb{E} \left| S(\beta) - \min_{u \geq 0} S(u) \right| = O\left( \sqrt{\frac{r \log n}{n}} \right),$$

where $r$ is the maximum degree of the dependency graph $G$ (recall that the degree of a vertex is the number of neighbors of that vertex, and the maximum degree of a graph is the maximum vertex degree).

**Proposition 1.4** *Let $r$ be the maximum degree of the dependency graph $G$ in the Ising model* (3)*, and let $S(u)$ be defined as in* (4)*. Take any $t \geq 0$ and let*

$$\varepsilon = \sqrt{\frac{r(\log n + t)}{n}}.$$

*Then we have*

$$\mathbb{P} \left\{ S(\beta) \geq \min_{u \geq 0} S(u) + C\varepsilon \right\} \leq \exp(-Kt^2),$$

*where $C$ and $K$ are numerical constants.*

Although it is unclear whether Proposition 1.4 is useful from a statistical point of view, it seems to be interesting as a mathematical result. For instance, observe that the conclusion is valid at any temperature. This is quite remarkable, since the low temperature phase in the Ising model is notoriously intractable for most graphs.

Here we should also mention that the technique can be easily applied to the Ising model with an external field, but we prefer to restrict ourselves to the problem of estimating a single parameter (the temperature) for the sake of clarity.

### 1.4 The abstract result

The following theorem encapsulates the concentration and moment inequalities used to work out all the examples in this paper.

**Theorem 1.5** *Let $\mathcal{X}$ be a separable metric space and suppose $(X, X')$ is an exchangeable pair of $\mathcal{X}$-valued random variables. Suppose $f : \mathcal{X} \to \mathbb{R}$ and $F : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ are square-integrable functions such that $F$ is antisymmetric (i.e. $F(X, X') = -F(X', X)$ a.s.), and $\mathbb{E}(F(X, X') \mid X) = f(X)$ a.s. Let*

$$\Delta(X) := \frac{1}{2}\mathbb{E}\big(|(f(X) - f(X'))F(X, X')| \,\big|\, X\big).$$

*Then $\mathbb{E}(f(X)) = 0$, and the following concentration results hold for $f(X)$:*

(i) *If $\mathbb{E}(\Delta(X)) < \infty$, then $\mathrm{Var}(f(X)) = \frac{1}{2}\mathbb{E}((f(X) - f(X'))F(X, X'))$.*

(ii) *Assume that $\mathbb{E}(e^{\theta f(X)}|F(X, X')|) < \infty$ for all $\theta$. If there exists nonnegative constants $B$ and $C$ such that $\Delta(X) \le Bf(X) + C$ almost surely, then for any $t \ge 0$,*

$$\mathbb{P}\{f(X) \ge t\} \le \exp\left(-\frac{t^2}{2C + 2Bt}\right) \quad \text{and} \quad \mathbb{P}\{f(X) \le -t\} \le \exp\left(-\frac{t^2}{2C}\right).$$

(iii) *For any positive integer $k$, we have the following exchangeable pairs version of the Burkholder-Davis-Gundy inequality:*

$$\mathbb{E}(f(X)^{2k}) \le (2k - 1)^k \mathbb{E}(\Delta(X)^k).$$

To see how the exchangeable pairs are constructed and the theorem is applied in our examples, one has to look at the proofs in Sect. 2. However, for a quick illustration, we will now work out the inequalities for sums of independent random variables, taking care to spell out details.

### 1.5 Simplest example

Let $X = \sum_{i=1}^n Y_i$, where $Y_i$'s are independent square integrable random variables. Let $\mu_i = \mathbb{E}(Y_i)$ and $\sigma_i^2 = \mathrm{Var}(Y_i)$. An exchangeable pair is created by choosing a coordinate $I$ uniformly at random from $\{1, \ldots, n\}$, and defining

$$X' = \sum_{j \ne I} Y_j + Y_I',$$

where $Y_1', \ldots, Y_n'$ are independent copies of $Y_1, \ldots, Y_n$. Let

$$F(x, y) = n(x - y).$$

Then

$$\begin{aligned}\mathbb{E}(F(X,X') \mid Y_1,\dots,Y_n) &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(n(Y_i - Y_i') \mid Y_1,\dots,Y_n)\\ &= \sum_{i=1}^{n}(Y_i - \mu_i) = X - \mathbb{E}(X).\end{aligned}$$

Since the right hand side depends only on $X$, we have

$$f(X) = \mathbb{E}(F(X,X') \mid X) = X - \mathbb{E}(X).$$

Thus, from part (i) of Theorem 1.5 we get the elementary identity

$$\mathrm{Var}(X) = \frac{1}{2}\sum_{i=1}^{n}\mathbb{E}(Y_i - Y_i')^2 = \sum_{i=1}^{n}\sigma_i^2.$$

Now note that

$$\begin{aligned}\Delta(X) &= \frac{n}{2}\mathbb{E}((X - X')^2 \mid X)\\ &= \frac{1}{2}\sum_{i=1}^{n}\mathbb{E}((Y_i - Y_i')^2 \mid X).\end{aligned}$$

If $c_1,\dots,c_n$ are constants such that $|Y_i - \mu_i| \le c_i$ a.s. for each $i$, then

$$\begin{aligned}\mathbb{E}((Y_i - Y_i')^2 \mid X) &= \mathbb{E}((Y_i - \mu_i)^2 \mid X) + \mathbb{E}((Y_i' - \mu_i)^2)\\ &\le c_i^2 + \sigma_i^2.\end{aligned}$$

Part (ii) of Theorem 1.5 now implies that

$$\mathbb{P}\{|X - \mathbb{E}(X)| \ge t\} \le 2\exp\left(-\frac{t^2}{\sum_{i=1}^{n}(c_i^2 + \sigma_i^2)}\right).$$

This is similar to (but not exactly the same as) the classical Hoeffding inequality [25] for sums of bounded random variables.

Now suppose that $0 \le Y_i \le 1$ a.s. for each $i$. If the $\mu_i$'s are very small, then the Hoeffding bound is wasteful. A more careful analysis gives a better result, as follows. First, note that

$$\begin{aligned}\Delta(X) &= \frac{1}{2}\sum_{i=1}^{n}\mathbb{E}((Y_i - Y_i')^2 \mid X)\\ &= \frac{1}{2}\sum_{i=1}^{n}(\mathbb{E}Y_i^2 - 2\mu_i\mathbb{E}(Y_i \mid X) + \mathbb{E}(Y_i^2 \mid X)).\end{aligned}$$

Using the assumption that $0 \le Y_i \le 1$, we get

$$\Delta(X) \le \frac{1}{2}\sum_{i=1}^{n}(\mathbb{E}(Y_i) + \mathbb{E}(Y_i \mid X)) = \frac{1}{2}(\mathbb{E}(X) + X) = \frac{1}{2}f(X) + \mathbb{E}(X).$$

Thus, we can take $B = 1/2$ and $C = \mathbb{E}(X)$ in part (ii) of Theorem 1.5, which gives

$$\mathbb{P}\{|X - \mathbb{E}(X)| \ge t\} \le 2\exp\left(-\frac{t^2}{2\mathbb{E}(X) + t}\right).$$

Again, this is a version of the classical Bernstein inequality (see [37], page 855) for sums of independent random variables.

Finally observe that by part (iii) of Theorem 1.5 and an application of Jensen's inequality, we have for each positive integer $k$,

$$\begin{aligned}\mathbb{E}(X^{2k}) &\le (2k-1)^k \mathbb{E}\left(\frac{1}{2}\sum_{i=1}^{n}\mathbb{E}((Y_i - \mu_i)^2 + (Y_i' - \mu_i)^2 \mid X)\right)^k \\ &\le (2k-1)^k \mathbb{E}\left(\sum_{i=1}^{n}(Y_i - \mu_i)^2\right)^k.\end{aligned}$$

This is exactly what the Burkholder–Davis–Gundy inequality [10] would give us for sums of independent random variables (although in this case, it can be derived by easier methods).

In the remainder of this section, we give very short overviews of Stein's method and concentration of measure.

### 1.6 Stein's method

Suppose we want to show that a random variable $X$ taking value in some space $\mathcal{X}$ has approximately the same distribution as some other random variable $Z$. The classical version of Stein's method [38,39] involves four steps:

1. Identify a "characterizing operator" $T$ for $Z$, which has the defining property that for any function $g$ belonging to a fixed large class of functions, $\mathbb{E}Tg(Z) = 0$. For instance, if $\mathcal{X} = \mathbb{R}$ and $Z$ is a standard gaussian random variable, then $Tg(x) := g'(x) - xg(x)$ is a characterizing operator, acting on all locally absolutely continuous $g$ with subexponential growth at infinity.
2. Construct a random variable $X'$ such that $(X, X')$ is an exchangeable pair.
3. Find an operator $\alpha$ such that for any suitable $h : \mathcal{X} \to \mathbb{R}$, $\alpha h$ is an antisymmetric function (i.e. $\alpha h(x, y) \equiv -\alpha h(y, x)$) and

$$|\mathbb{E}(\alpha h(X, X')|X = x) - Th(x)| \leq \varepsilon_h,$$

where $\varepsilon_h$ is a small error depending only on $h$.

4. Take a function $g$ and find $h$ such that $Th(x) = g(x) - \mathbb{E}g(Z)$. By antisymmetry of $\alpha h$ and the exchangeability of $(X, X')$, it follows that $\mathbb{E}(\alpha h(X, X')) = 0$. Combining with the previous step, we have the error bound $|\mathbb{E}g(X) - \mathbb{E}g(Z)| \leq \varepsilon_h$.

There are other variants of Stein's method, most notably the generator method of Andrew Barbour [4], the dependency graph approach introduced by Chen [15] and Baldi and Rinott [3] and popularized by Arratia, Goldstein and Gordon [2], the size-biased coupling method of Barbour, Holst and Janson [5], and the zero-biased coupling method due to Goldstein and Reinert [23]. The recent applications to algebraic problems by Jason Fulman [20,21], and the quest for Berry–Esseen bounds by Rinott and Rotar [34] and Shao and Su [35] are also worthy of note.

However, it is not our purpose here to go deeply into the regular versions of Stein's method. For further references and exposition, we refer to the recent monograph [17]. For applications of the method of exchangeable pairs and other versions of Stein's method to Poisson approximation, one can look at the survey paper by Chatterjee, Diaconis & Meckes [13].

## 1.7 Concentration inequalities

The theory of concentration inequalities tries to answer the following question: Given a random variable $X$ taking value in some measure space $\mathcal{X}$ (which is usually some high dimensional Euclidean space), and a measurable map $f : \mathcal{X} \to \mathbb{R}$, what is a good explicit bound on $\mathbb{P}\{|f(X) - \mathbb{E}f(X)| \geq x\}$? Exact evaluation or accurate approximation is, of course, the central purpose of probability theory itself. In situations where this is not possible, concentration inequalities aim to do the next best job by providing rapidly decaying tail bounds.

The literature on concentration inequalities is huge – from the pioneering inequalities of Hoeffding [25] to the momentous work of Talagrand [40] – but most of it revolves around well-behaved functions of independent random variables. For a nearly complete account of the literature until the year 2001, we redirect the reader to the definitive resource in this subject – the monograph [28] by Michel Ledoux. The methods of Kim and Vu [27] and Boucheron, Lugosi, and Massart [9] are significant recent developments.

The techniques developed in [11] (and partially presented here) have some basic similarities with the concentration results of Schmuckenschläger [36], but go much beyond that in terms of applications. Other than that (and log-Sobolev inequalities, which are much harder to obtain anyway) there is very little – even in the vast concentration literature – about the concentration of functions of dependent random variables, particularly in the discrete setting. We hope that our version of Stein's method will partially fill this void.

## 2 Proofs

Before proving Theorem 1.5, let us see how it is applied to work out the three examples described in section 1.

*Proof of Proposition 1.1* Construct $X'$ as follows: Choose $I, J$ uniformly and independently at random from $\{1, \ldots, n\}$. Let $\pi' = \pi \circ (I, J)$, where $(I, J)$ denotes the transposition of $I$ and $J$. It can be easily verified that $(\pi, \pi')$ is an exchangeable pair. Hence if we let

$$X' := \sum_{i=1}^{n} a_{i\pi'(i)},$$

then $(X, X')$ is also an exchangeable pair. Now note that

$$\begin{aligned}\frac{1}{2}\mathbb{E}(n(X - X')|\pi) &= \frac{n}{2}\mathbb{E}(a_{I\pi(I)} + a_{J\pi(J)} - a_{I\pi(J)} - a_{J\pi(I)}|\pi) \\ &= \frac{1}{n}\sum_{i,j} a_{i\pi(i)} - \frac{1}{n}\sum_{i,j} a_{i\pi(j)} \\ &= X - \mathbb{E}(X).\end{aligned}$$

Thus, we can take $f(x) = x - \mathbb{E}(X)$ and $F(x, y) = \frac{1}{2}n(x - y)$. Now note that since $0 \le a_{ij} \le 1$ for all $i$ and $j$, we have

$$\begin{aligned}\frac{1}{2}\mathbb{E}\big(|(f(X) - f(X'))F(X, X')| \,\big|\, \pi\big) &= \frac{n}{4}\mathbb{E}((X - X')^2|\pi) \\ &= \frac{1}{4n}\sum_{i,j}(a_{i\pi(i)} + a_{j\pi(j)} - a_{i\pi(j)} - a_{j\pi(i)})^2 \\ &\le \frac{1}{2n}\sum_{i,j}(a_{i\pi(i)} + a_{j\pi(j)} + a_{i\pi(j)} + a_{j\pi(i)}) \\ &= X + \mathbb{E}(X) = f(X) + 2\mathbb{E}(X).\end{aligned}$$

Since the last quantity depends only on $X$ it follows that $\Delta(X) = f(X) + 2\mathbb{E}(X)$. Applying part (ii) of Theorem 1.5 with $B = 1$ and $C = 2\mathbb{E}(X)$ completes the proof. □

*Proof of Proposition 1.2* Follows directly from part (iii) of Theorem 1.5 and the computations done in the proof of Proposition 1.1. □

*Proof of Proposition 1.3* Suppose $\sigma$ is drawn from the Gibbs distribution. We construct $\sigma'$ by taking a step in the Gibbs sampler as follows: Choose a coordinate $I$ uniformly at random, and replace the $I$th coordinate of $\sigma$ by an element

drawn from the conditional distribution of the $I$th coordinate given the rest. It is well-known and easy to prove that $(\sigma, \sigma')$ is an exchangeable pair. Let

$$F(\sigma, \sigma') := \sum_{i=1}^{n} (\sigma_i - \sigma_i').$$

Now define

$$m_i(\sigma) := \frac{1}{n} \sum_{j \le n, j \ne i} \sigma_j, \quad i = 1, \ldots, n.$$

Since the Hamiltonian is a simple explicit function, the conditional distribution of the $i^{\text{th}}$ coordinate given the rest is easy to obtain. An easy computation gives $\mathbb{E}(\sigma_i | \{\sigma_j, j \ne i\}) = \tanh(\beta m_i + \beta h)$. Thus, we have

$$\begin{aligned} f(\sigma) = \mathbb{E}(F(\sigma, \sigma') | \sigma) &= \frac{1}{n} \sum_{i=1}^{n} (\sigma_i - \mathbb{E}(\sigma_i | \{\sigma_j, j \ne i\})) \\ &= m - \frac{1}{n} \sum_{i=1}^{n} \tanh(\beta m_i + \beta h). \end{aligned}$$

Now note that $|F(\sigma, \sigma')| \le 2$, because $\sigma$ and $\sigma'$ differ at only one coordinate. Also, since the map $x \mapsto \tanh x$ is 1-Lipschitz, we have

$$|f(\sigma) - f(\sigma')| \le |m(\sigma) - m(\sigma')| + \frac{\beta}{n} \sum_{i=1}^{n} |m_i(\sigma) - m_i(\sigma')| \le \frac{2(1+\beta)}{n}.$$

Thus, by part (ii) of Theorem 1.5 we have

$$\mathbb{P}\left\{ \left| m - \frac{1}{n} \sum_{i=1}^{n} \tanh(\beta m_i + \beta h) \right| \ge \frac{t}{\sqrt{n}} \right\} \le 2 \exp\left( -\frac{t^2}{4(1+\beta)} \right).$$

Finally note that for each $i$, by the Lipschitz nature of the tanh function, we get

$$\begin{aligned} &\left| \frac{1}{n} \sum_{i=1}^{n} \tanh(\beta m_i + \beta h) - \tanh(\beta m + \beta h) \right| \\ &\le \frac{1}{n} \sum_{i=1}^{n} |\tanh(\beta m_i + \beta h) - \tanh(\beta m + \beta h)| \\ &\le \frac{1}{n} \sum_{i=1}^{n} \beta |m_i - m| \le \frac{\beta}{n}. \end{aligned}$$

This completes the proof. □

*Proof of Proposition 1.4* As in the proof of Proposition 1.3, we produce $\sigma'$ by taking a step in the Gibbs sampler: A coordinate $I$ is chosen uniformly at random, and $\sigma_I$ is replace by $\sigma'_I$ drawn from the conditional distribution of the $I^{\text{th}}$ coordinate given $(\sigma_j)_{j\neq I}$. For each $i$, let

$$m_i = m_i(\sigma) := \sum_{j:\{i,j\}\in E} \sigma_j.$$

Now fix $u \geq 0$ and define

$$F(\sigma,\sigma') := (\sigma_I - \sigma'_I)(\tanh(\beta m_I) - \tanh(u m_I)).$$

Then $F(\sigma,\sigma') = -F(\sigma',\sigma)$ because $m_I(\sigma) = m_I(\sigma')$ . Now let

$$\begin{aligned} f(\sigma) &:= \mathbb{E}(F(\sigma,\sigma') \mid \sigma) \\ &= \frac{1}{n}\sum_{i=1}^{n}(\sigma_i - \tanh(\beta m_i))(\tanh(\beta m_i) - \tanh(u m_i)). \end{aligned}$$

Now, if $r$ is the maximum degree of $G$, then at most $r+1$ terms in the sums defining $f(\sigma)$ and $f(\sigma')$ are unequal, and they all lie in the interval $[-4,4]$. Thus, $|f(\sigma) - f(\sigma')| \leq 8(r+1)/n$. Also, evidently, $|F(\sigma,\sigma')| \leq 4$. Using all this information in part (ii) of Theorem 1.5, we get

$$\mathbb{P}\{f(\sigma) \leq -t\} \leq \exp\left(-\frac{nt^2}{32(r+1)}\right).$$

Now, a direct verification shows that

$$S(u) - S(\beta) = \frac{1}{n}\sum_{i=1}^{n}(\tanh \beta m_i) - \tanh(u m_i))^2 + 2f(\sigma).$$

Thus,

$$\mathbb{P}\{S(\beta) \geq S(u) + t\} \leq \mathbb{P}\{2f(\sigma) \leq -t\} \leq \exp\left(-\frac{nt^2}{128(r+1)}\right). \tag{5}$$

Now note that for any $u, v \geq 0$, we have

$$\begin{aligned} &|S(u) - S(v)| \\ &\leq \frac{1}{n}\sum_{i=1}^{n}|(2\sigma_i - \tanh(u m_i) - \tanh(v m_i))(\tanh(v m_i) - \tanh(u m_i))| \\ &\leq \frac{4}{n}\sum_{i=1}^{n}|\tanh(v m_i) - \tanh(u m_i)| \leq 4r|u - v|, \end{aligned}$$

since $|m_i(u-v)| \le r|u-v|$. Let $N = \lfloor \sqrt{nr\log n} \rfloor$, and let

$$u_k = k\sqrt{\frac{\log n}{nr}} \quad \text{for } k = 1, 2, \ldots, N.$$

Then, if $u_{k-1} \le u \le u_k$, the above inequality gives

$$|S(u) - S(u_k)| \le 4r|u - u_k| \le 4\sqrt{\frac{r\log n}{n}}.$$

Now take any $u \ge u_N$. Since $m_i \in \{0, \pm 1, \ldots, \pm r\}$, therefore $|\tanh(um_i) - \tanh(u_N m_i)| \le 1 - \tanh(u_N|m_i|) \le 1 - \tanh(u_N)$. Thus,

$$\begin{aligned} |S(u) - S(u_N)| &\le \frac{4}{n}\sum_{i=1}^{n} |\tanh(um_i) - \tanh(u_N m_i)| \\ &\le 4(1 - \tanh(u_N)) \le 4e^{-u_N} \le \frac{4e}{n}. \end{aligned}$$

If $n \ge 3$, then $\sqrt{\log n/n} \ge e/n$. Combining the steps, we see that for $n \ge 3$,

$$\min_{1\le k\le N} S(u_k) \le \min_{u\ge 0} S(u) + 4\sqrt{\frac{r\log n}{n}}.$$

Finally, combining this with (5), we get

$$\begin{aligned} &\mathbb{P}\left\{ S(\beta) \ge \min_{u\ge 0} S(u) + 4\sqrt{\frac{r\log n}{n}} + t \right\} \\ &\quad \le \mathbb{P}\big\{ S(\beta) \ge \min_{1\le k\le N} S(u_k) + t \big\} \\ &\quad \le \sum_{k=1}^{N} \mathbb{P}\big\{ S(\beta) \ge S(u_k) + t \big\} \le N\exp\left( -\frac{nt^2}{128(r+1)} \right). \end{aligned}$$

It is now easy to complete the proof by substituting the value of $N$ and choosing $t > \sqrt{Cr\log n/n}$ for sufficiently large $C$, so that the effect of $N$ washes out. □

Finally, let us prove our main result.

*Proof of Theorem 1.5* Let us begin with a useful general identity. Suppose $h : \mathcal{X} \to \mathbb{R}$ is any measurable map such that $\mathbb{E}|h(X)F(X, X')| < \infty$. Then clearly $\mathbb{E}(h(X)f(X)) = \mathbb{E}(h(X)F(X, X'))$. Using the exchangeability of $X$ and $X'$, and the antisymmetric nature of $F$, we have

$$\mathbb{E}(h(X)F(X, X')) = \mathbb{E}(h(X')F(X', X)) = -\mathbb{E}(h(X')F(X, X')).$$

Thus, we have

$$\mathbb{E}(h(X)f(X)) = \mathbb{E}(h(X)F(X,X')) = \frac{1}{2}\mathbb{E}((h(X) - h(X'))F(X,X')). \quad (6)$$

The above equation is the basis of all that follows. First, note that by putting $h \equiv 1$, we immediately get $\mathbb{E}(f(X)) = 0$, Similarly, part (i) of the Theorem follows by putting $h = f$. Next, let us start proving (ii). Let $m(\theta) := \mathbb{E}(\mathrm{e}^{\theta f(X)})$ be the moment generating function of $f(X)$. We can differentiate $m(\theta)$ and move the derivative inside the expectation because of the assumption that $\mathbb{E}(\mathrm{e}^{\theta f(X)}|F(X,X')|) < \infty$ for all $\theta$. Thus, by Eq. (6), we have

$$m'(\theta) = \mathbb{E}(\mathrm{e}^{\theta f(X)}f(X)) = \frac{1}{2}\mathbb{E}((\mathrm{e}^{\theta f(X)} - \mathrm{e}^{\theta f(X')})F(X,X')).$$

Now note that for any $x, y \in \mathbb{R}$,

$$\begin{aligned}\left|\frac{\mathrm{e}^x - \mathrm{e}^y}{x-y}\right| &= \int_0^1 \mathrm{e}^{tx+(1-t)y}\mathrm{d}t \\ &\le \int_0^1 (t\mathrm{e}^x + (1-t)\mathrm{e}^y)\mathrm{d}t = \frac{1}{2}(\mathrm{e}^x + \mathrm{e}^y). \end{aligned} \quad (7)$$

Using this inequality, and the exchangeability of $X$ and $X'$, we get

$$\begin{aligned}|m'(\theta)| &\le \frac{|\theta|}{4}\mathbb{E}((\mathrm{e}^{\theta f(X)} + \mathrm{e}^{\theta f(X')})|(f(X) - f(X'))F(X,X')|) \\ &= \frac{|\theta|}{2}\mathbb{E}(\mathrm{e}^{\theta f(X)}\Delta(X) + \mathrm{e}^{\theta f(X')}\Delta(X')) \\ &= |\theta|\mathbb{E}(\mathrm{e}^{\theta f(X)}\Delta(X)) \\ &\le |\theta|\mathbb{E}(\mathrm{e}^{\theta f(X)}(Bf(X) + C)) = B|\theta|m'(\theta) + C|\theta|m(\theta).\end{aligned}$$

Since $m$ is a convex function and $m'(0) = \mathbb{E}(f(X)) = 0$, therefore $m'(\theta)$ always has the same sign as $\theta$. Thus, for $0 \le \theta < 1/B$, the above inequality translates into

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\log m(\theta) \le \frac{C\theta}{1 - B\theta}.$$

Using this and recalling that $m(0) = 1$, we have

$$\log m(\theta) \le \int_0^\theta \frac{Cu}{1 - Bu}\mathrm{d}u \le \frac{C\theta^2}{2(1 - B\theta)}.$$

Putting $\theta = t/(C + Bt)$, we get

$$\mathbb{P}\{f(X) \geq t\} \leq \exp(-\theta t + \log m(\theta)) \leq \mathrm{e}^{-t^2/(2C+2Bt)}.$$

The lower tail can be done similarly; note that for $\theta \leq 0$, we have $m'(\theta) \leq 0$, and hence

$$|m'(\theta)| \leq B|\theta|m'(\theta) + C|\theta|m(\theta) \leq C|\theta|m(\theta),$$

and this is the reason why $B$ does not appear in the lower tail bound. This completes the proof of part (ii). For the moment inequalities in part (iii), first observe that by Eq. (6), we have

$$\mathbb{E}(f(X)^{2k}) = \frac{1}{2}\mathbb{E}((f(X)^{2k-1} - f(X')^{2k-1})F(X, X')).$$

By the inequality

$$|x^{2k-1} - y^{2k-1}| \leq \frac{2k-1}{2}(x^{2k-2} + y^{2k-2})|x - y|$$

which follows easily from a convexity argument very similar to (7), we have

$$\mathbb{E}(f(X)^{2k}) \leq (2k-1)\mathbb{E}(f(X)^{2k-2}\Delta(X)).$$

By Hölder's inequality, we get

$$\mathbb{E}(f(X)^{2k}) \leq (2k-1)(\mathbb{E}(f(X)^{2k}))^{(k-1)/k}(\mathbb{E}(\Delta(X)^k))^{1/k}.$$

The proof is completed by transferring $\mathbb{E}(f(X)^{2k})^{(k-1)/k}$ to the other side. □

**Acknowledgements** I am grateful to Persi Diaconis and Yuval Peres for many useful comments and suggestions. Thanks are also due to the two anonymous referees for pointing out several omissions and errors.

## References

1. Arratia, R., Goldstein, L., Gordon, L.: Two moments suffice for Poisson approximations: the Chen–Stein method. Ann. Probab. **17**(1), 9–25 (1989)
2. Arratia, R., Goldstein, L., Gordon, L.: Poisson approximation and the Chen-Stein method. Statist. Sci. **5**(4), 403–434 (1992)
3. Baldi, P., Rinott, Y.: On normal approximations of distributions in terms of dependency graphs. Ann. Probab. **17**(4), 1646–1650 (1989)
4. Barbour, A.D.: Stein's method for diffusion approximations. Probab. Theory Relat. Fields **84**(3), 297–322 (1990)
5. Barbour, A.D., Holst, L., Janson, S.: Poisson approximation. In: Oxford Studies in Probability, vol. 2 The Clarendon Press, Oxford University Press, New York (1992)
6. Besag, J.E.: Statistical analysis of non-lattice data. Statistician **24**, 179–195 (1975)

7. Bolthausen, E.: An estimate of the remainder in a combinatorial central limit theorem. Z. Wahrsch. Verw. Gebiete **66**(3), 379–386 (1984)
8. Bolthausen, E., Götze, F.: The rate of convergence for multivariate sampling statistics. Ann. Statist. **21**(4), 1692–1710 (1993)
9. Boucheron, S., Lugosi, G., Massart, P.: Concentration inequalities using the entropy method. Ann. Probab. **31**(3), 1583–1614 (2003)
10. Burkholder, D.L.: Distribution function inequalities for martingales. Ann. Probab. **1**, 19–42 (1973)
11. Chatterjee, S.: Concentration inequalities with exchangeable pairs. Ph.D. thesis, Department of Statistics, Stanford University. Available at http://arxiv.org/math.PR/0507526 (2005)
12. Chatterjee, S.: Concentration of Haar measures, with an application to random matrices. (Submitted, 2005) Available at http://arxiv.org/math.PR/0508518
13. Chatterjee, S., Diaconis, P., Meckes, E.: Exchangeable pairs and Poisson approximation. Probab. Surv. **2**, 64–106 (2005)
14. Chazottes, J.-R., Collet, P., Külske, C., Redig, F.: Concentration inequalities for random fields via coupling. (Submitted, 2006) Available at http://arxiv.org/math.PR/0503483
15. Chen, L.H.Y.: Poisson approximation for dependent trials. Ann. Probab. **3**(3), 534–545 (1975)
16. Diaconis, P., Graham, R.L.: Spearman's footrule as a measure of disarray. J. R. Statist. Soc. Ser. B **39**(2), 262–268 (1977)
17. Diaconis, P., Holmes, S. (eds.): Stein's method: expository lectures and applications. In: IMS Lecture Notes—Monograph Series, vol. 46 (2004)
18. Ellis, R.S.: Entropy, large deviations, and statistical mechanics. Grund. der Mathemat. Wissenschaften, vol. 271. Springer, New York
19. Feller, W.: An Introduction to Probability Theory and its Applications, vol. I, 3rd edn. Wiley New York (1968)
20. Fulman, J.: Stein's method and non-reversible Markov chains. In: Stein's Method: Expository Lectures and Applications, pp. 69–77. IMS Lecture Notes Monogr. Ser., vol. 46. IMS, Beachwood (2004)
21. Fulman, J.: Stein's method and Plancherel measure of the symmetric group. Trans. Am. Math. Soc. **357**(2), 555–570 (electronic) (2005)
22. Geyer, C.J., Thompson, E.A.: Constrained Monte Carlo maximum likelihood for dependent data. J. R. Statist. Soc. Ser. B **54**(3), 657–699 (1992)
23. Goldstein, L., Reinert, G.: Stein's method and the zero bias transformation with application to simple random sampling. Ann. Appl. Probab. **7**(4), 935–952 (1997)
24. Hoeffding, W.: A combinatorial central limit theorem. Ann. Math. Statist. **22**(4), 558–566 (1951)
25. Hoeffding, W.: Probability inequalities for sums of bounded random variables. J. Am. Stat. Assoc. **58**, 13–30 (1963)
26. Jerrum, M., Sinclair, A.: Polynomial-time approximation algorithms for the Ising model. SIAM J. Comput. **22**(5), 1087–1116 (1993)
27. Kim, J.H., Vu, V.H.: Divide and conquer martingales and the number of triangles in a random graph. Random Struct. Algor. **24**(2), 166–174 (2004)
28. Ledoux, M.: The Concentration of Measure Phenomenon. Am. Math. Soc., Providence, RI (2001)
29. Luczak, M.J., McDiarmid, C.: Concentration for locally acting permutations. Discrete Math. **265**(1–3), 159–171 (2003)
30. Maurey, B.: Construction de suites symétriques. C. R. Acad. Sci. Paris Sér. A-B **288**(14), A679–A681 (1979)
31. McDiarmid, C.: Concentration for independent permutations. Combin. Probab. Comput. **11**(2), 163–178 (2002)
32. Pickard, D.K.: Inference for discrete Markov fields: the simplest nontrivial case. J. Am. Statist. Assoc. **82**(397), 90–96 (1987)
33. Raič, M.: CLT-related large deviation bounds based on Stein's method (preprint, 2004)
34. Rinott, Y., Rotar, V.: On coupling constructions and rates in the CLT for dependent summands with applications to the antivoter model and weighted $U$-statistics. Ann. Appl. Probab. **7**(4), 1080–1105 (1997)
35. Shao, Q., Su, Z.: The Berry–Esseen bound for character ratios (preprint, 2004)
36. Schmuckenschläger, M.: Curvature of nonlocal Markov generators. In: Convex Geometric Analysis: MSRI Publications. vol. 34, pp. 189–197 (1998)

37. Shorack, G.R., Wellner, J.A.: Empirical processes with applications to statistics. Wiley New York (1986)
38. Stein, C.: A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In: Proceedings of the 6th Berkeley Symposium on Mathematical Statistics and Probability, vol. II: Probability theory, pp. 583–602 (1972)
39. Stein, C.: Approximate computation of expectations. IMS Lecture Notes—Monograph Series, vol. 7 (1986)
40. Talagrand, M.: Concentration of measure and isoperimetric inequalities in product spaces. Inst. Hautes Études Sci. Publ. Math. **81**, 73–205 (1995)