Electronic Journal of Probability

Electron. J. Probab. **23** (2018), no. 35, 1–24.
ISSN: 1083-6489 https://doi.org/10.1214/18-EJP159

Decomposition of mean-field Gibbs distributions into product measures

Ronen Eldan* Renan Gross†

Abstract

We show that under a low complexity condition on the gradient of a Hamiltonian, Gibbs distributions on the Boolean hypercube are approximate mixtures of product measures whose probability vectors are critical points of an associated mean-field functional. This extends a previous work by the first author. As an application, we demonstrate how this framework helps characterize both Ising models satisfying a mean-field condition and the conditional distributions which arise in the emerging theory of nonlinear large deviations, both in the dense case and in the polynomially-sparse case.

Keywords: mean field; large deviation; Gibbs distribution; Ising model; Gaussian width; sparse random graphs.
AMS MSC 2010: 60K35; 60F10; 82B20; 05C80.
Submitted to EJP on September 15, 2017, final version accepted on March 20, 2018.
Supersedes arXiv:1708.05859.

1 Introduction

Let $n > 0$ and let $f : \{-1,1\}^n \to \mathbb{R}$ be a function. A probability measure ν on $\{-1,1\}^n$ is called a *Gibbs distribution with Hamiltonian* f if for $X \sim \nu$,

$$\Pr[X = x] = \exp(f(x))/Z,$$

where Z is a normalizing constant. We denote such a distribution by X_n^f. Gibbs distributions are central to statistical physics, and appear in applications in computer science, statistics, and economics. However, many important Hamiltonians are far from being analytically tractable.

One method to tackle the difficulties entrenched in such Hamiltonians is via mean-field approximations. This method goes back to Curie and Weiss and has long been widely used by physicists. More recently, such approximations were established in rigor, see for example [1].

For the case of Gibbs distributions on the Boolean hypercube, [4] showed that if the image of the gradient of the Hamiltonian f has small enough Gaussian-width and

*Weizmann Institute of Science, Israel. E-mail: ronen.eldan@weizmann.ac.il
†Weizmann Institute of Science, Israel. E-mail: renan.gross@weizmann.ac.il

Lipschitz constants, then the partition function can be approximated by applying the mean-field variant of the Gibbs variational principle. Further, under the same conditions, X_n^f can be approximated by a mixture of product measures. This improves an earlier result by Chatterjee and Dembo [2] who consider a slightly different notion of complexity.

In this paper, we extend the framework introduced in [4] by showing that if the discrete gradient ∇f also has a small enough Lipschitz constant, then the product measures described above are close to critical points of an associated variational functional which corresponds to the so-called *mean-field equations*. This gives a more precise characterization of the mixture.

An interesting feature of our framework is that it allows us to effectively bypass the need to obtain an accurate approximation of the normalizing constant in the route to understanding the Gibbs distribution. Even though the approximations to the normalizing constant obtained by the framework are far from sharp (they miss by a factor of $e^{o(n)}$ as seen in the examples in [4]), our results still manage to give information about the set where most of the mass resides.

The following is an overview of our results.

- In **Theorem 3.1**, we show that if the Hamiltonian f has low complexity and satisfies a Lipschitz condition, the corresponding Gibbs distribution behaves like a mixture of densities of vectors whose entries are i.i.d Bernoulli random variables, and whose expectations X satisfy
$$\left\|X - \tanh\left(\nabla f\left(X\right)\right)\right\|_1 = o\left(n\right),$$
where the $\tanh$ is applied entrywise.
- As an example of using this bound, we demonstrate in **Corollaries 3.4 and 3.5** that Ising models satisfying a mean-field assumption can be decomposed into product measures.
- **Theorem 3.6** concerns compositions: If a function $h : \mathbb{R} \to \mathbb{R}$ has small enough derivatives, then the function $h \circ f$ also satisfies Theorem 3.1.
- As an example of this composition, we demonstrate in **Theorem 3.8** that the conditional distribution $\Pr\left[Y = y \mid f\left(Y\right) \geq tn\right]$ arising in large deviation theory can be approximated by a smoothed-cutoff distribution that can be decomposed into product measures, each satisfying an equation which arises from the Lagrange multiplier problem associated with the rate function.

In the sequel work [5], we apply Theorem 3.1 to exponential random graphs, improving a previously known characterization.

2 Background and notation

We denote the Boolean hypercube by $\mathcal{C}_n = \{-1, 1\}^n$ and the continuous hypercube by $\overline{\mathcal{C}_n} = [0, 1]^n$. The uniform measure on $\mathcal{C}_n$ is denoted by μ. The space of all product measures on $\mathcal{C}_n$ is denoted $\mathcal{PM}_n$. For a vector $x \in \mathbb{R}^n$, we denote its one-norm by
$$\left\|x\right\|_1 = \sum_{i=1}^{n} \left|x_i\right|.$$

2.1 Two motivating examples of Hamiltonians

2.1.1 The Ising model

An Ising model on n sites can be described as follows: Let $x \in \mathcal{C}_n$ represent n interacting sites that can be in one of two states. Let $A \in \mathbb{R}^{n \times n}$ be a real symmetric matrix with 0

on the diagonal representing the intensity of interaction between the sites, so that the interaction between site i and site j is A_{ij}. Let $\mu \in \mathbb{R}^n$ be a vector representing magnetic field strengths, so that site i feels a magnetic field μ_i. The Hamiltonian for the system is then defined as

$$f(x) = \langle x, Ax \rangle + \langle \mu, x \rangle .$$

If $\mathrm{Tr} A^2 = o(n)$, we say that the model satisfies the *mean-field* assumption [1]. We also assume that both $\mu_{\max}$ and $\max_{i \in [n]} \sum_{j \in [n]} |A_{ij}|$ are $O(1)$, which amounts to the force acting on a single site being bounded.

2.1.2 Nonlinear large deviations

Let $f : \mathcal{C}_n \to \mathbb{R}$ be a Hamiltonian. For $0 \le p \le 1$, define μ_p to be the measure on $\mathcal{C}_n$ where every entry is an i.i.d Bernoulli random variable with success probability p. Let $t \in \mathbb{R}$ be a real number. The two central questions in the field of large deviation theory are:

1. For $Y \sim \mu_p$, what is the probability $\Pr[f(Y) \ge tn]$?
2. For $Y \sim \mu_p$, what is the conditional distribution $\Pr[Y = y \mid f(Y) \ge tn]$?

One line of approach to answering these questions is to approximate $\Pr[f(Y) \ge tn]$ and $\Pr[Y = y \mid f(Y) \ge tn]$ by using Gibbs distributions. For example, observe that the conditional distribution $\Pr[Y = y \mid f(y) \ge tn]$ may be obtained from a Gibbs distribution with a "cutoff Hamiltonian" $\tilde{f}$, defined by

$$\tilde{f}(y) = \begin{cases} \prod_{i=1}^{n} \log\left(\frac{1}{2}(1 - y_i + 2py_i)\right) & f(y) \ge tn \\ -\infty & f(y) < tn. \end{cases} \tag{2.1}$$

All y with $f(y) \ge tn$ are thus weighted according to μ_p, and all y with $f(y) < tn$ have probability 0. Unfortunately, $\tilde{f}$ is not smooth enough in order to be applicable for the existing large deviation frameworks. However, it is possible to get approximations of $X_n^{\tilde{f}}$ by using Hamiltonians which approximate $\tilde{f}$. Such a "smooth-cutoff" Hamiltonian should give a large mass to "good" vectors y such that $f(y) \ge tn$ and a small mass to "bad" vectors y such that $f(y) < tn$. Both [4] and [2] follow this approach in order to tackle item (1).

2.2 Boolean functions

Definition 2.1 (Discrete gradient, Lipschitz constant)**.** *Let $f : \mathcal{C}_n \to \mathbb{R}$ be a real function on the Boolean hypercube. The derivative of f at coordinate i is defined as*

$$\partial_i f(y) = \frac{1}{2}\left(f(y_1, \dots, y_{i-1}, 1, y_{+1}, \dots y_n) - f(y_1, \dots, y_{i-1}, -1, y_{+1}, \dots y_n)\right).$$

With this we define both the the discrete gradient:

$$\nabla f(y) = (\partial_1 f(y), \dots, \partial_n f(y)),$$

and the Lipschitz constant of f:

$$\mathit{Lip}(f) = \max_{i \in [n], y \in \{-1,1\}^n} |\partial_i f(y)| .$$

Every Boolean function $f : \mathcal{C}_n \to \mathbb{R}$ has a unique Fourier decomposition into monomials [7]:

$$f(x) = \sum_{S \subseteq [n]} \hat{f}(S) \prod_{i \in S} x_i .$$

This defines an extension of f from the discrete hypercube $\mathcal{C}_n$ into the continuous hypercube $\overline{\mathcal{C}_n} = [-1,1]^n$ by computing the value of the polynomial $\sum_{S\subseteq[n]} \hat{f}(S) \prod_{i\in S} x_i$ for $x \in \overline{\mathcal{C}_n}$. It can be shown that this is the same extension as the harmonic extension defined in [4, Section 3.1.1]. By Fact 14 in [4], the extension of $\partial_i f$ agrees with the i-th partial derivative (in the real-differentiable sense) of the extension of f. Throughout this text, we will always assume that f, and therefore ∇f as well, are extended to $\overline{\mathcal{C}_n}$.

Definition 2.2 (Gaussian width, gradient complexity)**.** *The* Gaussian-width *of a set* $K \subseteq \mathbb{R}^n$ *is defined as*

$$\mathbf{GW}(K) = \mathbb{E}\left[\sup_{x\in K} \langle x, \Gamma\rangle\right]$$

where $\Gamma \sim N(0, Id)$ *is a standard Gaussian vector in* $\mathbb{R}^n$. *For a function* $f : \mathcal{C}_n \to \mathbb{R}$*, the* gradient complexity *of* f *is defined as*

$$\mathcal{D}(f) = \mathbf{GW}(\{\nabla f(y) : y \in \mathcal{C}_n\} \cup \{0\}).$$

For a measure ν *on* $\mathcal{C}_n$*, by slight abuse of notation, we define its complexity as*

$$\mathcal{D}(\nu) = \mathcal{D}\left(\log \frac{d\nu}{d\mu}\right).$$

2.3 Mixture models

Definition 2.3 (ρ-mixtures)**.** *For* $z \in [-1,1]^n$*, denote by* $X(z)$ *the unique random vector in* $\mathcal{C}_n$ *whose coordinates are independent and whose expectation is* $\mathbb{E}X(z) = z$*. Let* ρ *be a measure on* $[-1,1]^n$*. We define the random vector* $X(\rho)$ *by*

$$Pr[X(\rho) = x] = \int Pr[X(z) = x]\, d\rho(z). \tag{2.2}$$

Definition 2.4 (Approximate mixture decomposition)**.** *Let* $\delta > 0$ *and let* ρ *be a measure on* $[-1,1]^n$*. A random variable* X *is called a* (ρ,δ)-mixture *if there exists a coupling between* $X(\rho)$ *and* X *such that*

$$\mathbb{E}\|X(\rho) - X\|_1 \le \delta n.$$

A result of [4] roughly states that low complexity Gibbs distributions are (ρ,δ)-mixtures for $\delta = o(1)$ and where ρ is such that most of the entropy comes from the individual $X(z)$ rather than from the mixture.

Definition 2.5 (Wasserstein distance)**.** *For two distributions* ν_1 *and* ν_2*, the* Wasserstein mass-transportation distance*, denoted* W_1*, is defined as*

$$W_1(\nu_1, \nu_2) = \inf_{\substack{(X,Y)\ s.t\\ X\sim\nu_1, Y\sim\nu_2}} \frac{1}{2}\mathbb{E}\|X - Y\|_1,$$

where the infimum is taken over all joint distributions whose marginals have the laws ν_1 *and* ν_2 *respectively.*

Definition 2.6 (Tilt of a distribution)**.** *For a vector* $\theta \in \mathbb{R}^n$*, the* tilt $\tau_\theta \nu$ *of the distribution* ν *is a distribution defined by*

$$\frac{d(\tau_\theta \nu)}{d\nu}(y) = \frac{e^{\langle\theta,y\rangle}}{\int_{\mathcal{C}_n} e^{\langle\theta,z\rangle} d\nu}.$$

With the notion of ρ-mixture and tilt, we define what it means for a random variable to break up into small tilts:

Definition 2.7 (Tilt decomposition). *Let $\delta, \varepsilon > 0$ and let ρ be a measure on $[-1,1]^n$. A random variable X with distribution ν is called a $(\rho, \delta, \varepsilon)$-*tilt-mixture* if there exists a probability measure m on $\mathbb{R}^n$ supported on $B(0, \varepsilon\sqrt{n}) \cap \left[-\frac{1}{4}, \frac{1}{4}\right]^n$ such that:*

1. For every $\varphi : \mathcal{C}_n \to \mathbb{R}$,
$$\int_{\mathcal{C}_n} \varphi d\nu = \int_{\mathbb{R}^n} \left(\int_{\mathcal{C}_n} \varphi d(\tau_\theta \nu) \right) dm(\theta).$$
2. For all but a δ-portion of the measure m, the tilt $\tau_\theta \nu$ is δn-close to a product measure in Wasserstein distance:
$$m\left(\{\theta \in \mathbb{R}^n : \exists \xi \in \mathcal{PM}_n \, s.t \, W_1(\tau_\theta \nu, \xi) \le \delta n\}\right) > 1 - \delta.$$
3. The measure ρ is the push-forward of the measure m under the map $\theta \mapsto \mathbb{E}_{X \sim \tau_\theta \nu}[X]$.

Fact 2.8. *Every $(\rho, \delta, \varepsilon)$-tilt-mixture is also a $(\rho, 4\delta)$-mixture.*

Proof. Define $\Theta = \{\theta \in \mathbb{R}^n : \exists \xi \in \mathcal{PM}_n \, s.t \, W_1(\tau_\theta \nu, \xi) \le \delta n\}$, and denote the distribution of X and of $X(\rho)$ by ν and σ respectively. Using item 1 in the definition of a tilt-mixture, we have

$$\begin{aligned} W_1(\nu, \sigma) &\le \int_{\mathbb{R}^n} W_1(\xi_\theta, \tau_\theta \nu) \, dm(\theta) \\ &\le \int_{\Theta} W_1(\xi_\theta, \tau_\theta \nu) \, dm(\theta) + m\left([-1/4, 1/4]^n \setminus \Theta\right) n. \end{aligned}$$

By item 2 in the definition of a tilt-mixture, there exists a coupling between X and $X(\rho)$ such that each term on the right hand side is bounded by δn. This gives a 4δ bound on the expectation $\mathbb{E}\left\|X - X(\rho)\right\|_1$. ☐

A tilt-mixture decomposition provides more information than generalρ-mixtures: It tells us something about the structure of the elements of the mixture, with the parameter ε in Definition 2.7 bounding the support of the tilts to a ball of radius $\varepsilon\sqrt{n}$. Some of our results will rely on the existence of tilt decompositions with small ε.

3 Results

Our main technical contribution is a characterization of the measure ρ described above: With high probability with respect to ρ, the vector z in equation (2.2) is nearly a critical point of a certain functional associated with f.

Theorem 3.1 (Main Structural Theorem). *Let $n > 0$, let $f : \mathcal{C}_n \to \mathbb{R}$ be a function and denote*

$$D = \mathcal{D}(f) \tag{3.1}$$
$$L_1 = \max\{1, \mathit{Lip}(f)\} \tag{3.2}$$
$$L_2 = \max\left\{1, \max_{x \ne y \in \mathcal{C}_n} \frac{\left\|\nabla f(x) - \nabla f(y)\right\|_1}{\left\|x - y\right\|_1}\right\}. \tag{3.3}$$

Denote by $\mathcal{X}_f$ the set

$$\mathcal{X}_f = \left\{X \in \overline{\mathcal{C}_n} : \left\|X - \tanh(\nabla f(X))\right\|_1 \le 5000 L_1 L_2^{3/4} D^{1/4} n^{3/4}\right\}, \tag{3.4}$$

where $\nabla f(X)$ is calculated by harmonically extending ∇f to $\overline{\mathcal{C}_n}$, and with the $\tanh$ *applied entrywise to the entries of $\nabla f(X)$. Then X_n^f is a $\left(\rho, \frac{3D^{1/4}}{n^{1/4}}, L_2^{3/4}\frac{D^{1/4}}{n^{1/4}}\right)$-tilt-mixture such that*

$$\rho(\mathcal{X}_f) \geq 1 - \frac{3D^{1/4}}{n^{1/4}}. \tag{3.5}$$

In particular, if $D = o(n)$, then X_n^f is a $(\rho, o(1))$-mixture with $\rho(\mathcal{X}_f) = 1 - o(1)$.

In other words, almost all the mass of the mixture resides on random vectors X which almost satisfy the fixed point equation

$$X = \tanh(\nabla f(X)). \tag{3.6}$$

Remark 3.2. One can check that the solutions of the fixed point equation are exactly the critical points of the functional $f(X) + H(X)$ where $H(X) = \sum_{i<,j} X_{ij}\log X_{ij} + (1 - X_{ij})\log(1 - X_{ij})$ is the entropy of X. This is a variant of the functional that arises in the variational problem in [3].

Remark 3.3. The following is an example application of Theorem 3.1 to Ising models, to be compared with the main result of [1].

Corollary 3.4 (Ising models)**.** *Let f be an Ising model Hamiltonian as described in Section 2.1.1, with interaction matrix $A \in \mathbb{R}^{n\times n}$ and a magnetic moment vector $\mu \in \mathbb{R}^n$. Denote*

$$\mathcal{X}_f = \left\{X \in \overline{\mathcal{C}_n} : \|X - \tanh(AX + \mu)\|_1 \leq 5000 L_1 L_2^{3/4} D^{1/4} n^{3/4}\right\},$$

where

$$D = \sqrt{n Tr A^2} + \sqrt{n}\mu_{\max}$$
$$L_1 = \max\left\{1, \mu_{\max} + \max_{i\in[n]}\sum_{j\in[n]}|A_{ij}|\right\}$$
$$L_2 = \max\left\{1, \max_{i\in[n]}\sum_{j\in[n]}|A_{ij}|\right\}.$$

Then X_n^f is a $\left(\rho, \frac{3D^{1/4}}{n^{1/4}}, L_2^{3/4}\frac{D^{1/4}}{n^{1/4}}\right)$-tilt-mixture such that

$$\rho(\mathcal{X}_f) \geq 1 - \frac{3D^{1/4}}{n^{1/4}}.$$

In particular, if $L_1 = O(1)$ and Tr$\left(A^2\right) = o(n)$ (the "mean-field assumption"), then $\mathcal{X}_f$ is $(\rho, o(1))$-mixture with $\rho(\mathcal{X}_f) = 1 - o(1)$.

The simplest example of an Ising model is the Curie-Weiss ferromagnet, for which we can use our framework as a toy example and rederive well-known properties about its distribution.

Corollary 3.5. *Let $\beta > 0$ and let $f : \mathcal{C}_n \to \mathbb{R}$ be the Curie-Weiss Hamiltonian, $f(x) = \frac{\beta}{n}\sum_{i\neq j} x_i x_j$. Denote*

$$\mathcal{X}_f = \left\{X \in \overline{\mathcal{C}_n} : \left\|X - \tanh\left(\frac{\beta \boldsymbol{J}}{n}X\right)\right\|_1 \leq 5001(1+\beta)^2 n^{7/8}\right\},$$

where $\boldsymbol{J}$ is the $n \times n$ all-1 matrix. Then X_n^f is a $\left(\rho, 3n^{-1/8}, 3n^{-1/8}\right)$-tilt-mixture, and $\rho(\mathcal{X}_f) \geq 1 - 3n^{-1/8}$. Further, if $\beta < 1$, then every $X \in \mathcal{X}_f$ satisfies

$$\|X\|_1 \leq 5001\frac{(1+\beta)^2}{1-\beta}n^{7/8}.$$

For a more detailed application of Theorem 3.1 for the case of exponential random graphs, see [5].

The following theorem finds sufficient conditions under which composing f with a real-valued function produces a Hamiltonian with aρ-mixture approximation:

Theorem 3.6 (Composition Theorem)**.** *Let $h : \mathbb{R} \to \mathbb{R}$ be a twice differentiable function satisfying*

$$\begin{aligned} \left|h'\left(x\right)\right| &< B_1 \quad \forall x \in \mathbb{R} \\ \left|h''\left(x\right)\right| &< B_2 \quad \forall x \in \mathbb{R}. \end{aligned}$$

Let $f : \mathcal{C}_n \to \mathbb{R}$ be a function with parameters D, L_1, and L_2 as described in Theorem 3.1. Denote by $\tilde{D}$, $\tilde{L}_1$, $\tilde{L}_2$ and $\tilde{L}_3$ the real numbers

$$\begin{aligned} \tilde{D} &= B_1 D + B_2 L_1^2 n \\ \tilde{L}_1 &= \max\left\{1, B_1 L_1\right\} \\ \tilde{L}_2 &= \max\left\{1, B_1 L_2 + 3B_2 L_1^2 n\right\} \\ \tilde{L}_3 &= 2B_2 L_1^2 n^{3/2} \end{aligned}$$

and denote by $\tilde{\mathcal{X}}_{h\circ f}$ the set

$$\tilde{\mathcal{X}}_{h\circ f} = \left\{X \in \overline{\mathcal{C}_n} : \left\|X - \tanh\left(h'\left(f\left(X\right)\right)\nabla f\left(X\right)\right)\right\|_1 \leq 5000\tilde{L}_1\tilde{L}_2^{3/4}\tilde{D}^{1/4}n^{3/4} + \tilde{L}_3\right\}, \quad (3.7)$$

where $\nabla f\left(X\right)$ is calculated by harmonically extending ∇f to $\overline{\mathcal{C}_n}$, and with the $\tanh$ *applied entrywise to the entries of $\nabla f\left(X\right)$. Then $X_n^{h\circ f}$ is a $\left(\rho, \frac{3\tilde{D}^{1/4}}{n^{1/4}}, \tilde{L}_2^{3/4}\frac{\tilde{D}^{1/4}}{n^{1/4}}\right)$-tilt-mixture such that*

$$\rho\left(\mathcal{X}_{h\circ f}\right) \geq 1 - \frac{3\tilde{D}^{1/4}}{n^{1/4}}.$$

Remark 3.7. Theorem 3.6 bounds the norm $\left\|X - \tanh\left(h'\left(f\left(X\right)\right)\nabla f\left(X\right)\right)\right\|_1$ rather than $\left\|X - \tanh\left(\nabla\left(h\circ f\right)\left(X\right)\right)\right\|_1$ (which is the analogue of the quantity arising in the main Theorem 3.1). This is a matter of practicality: For many known Hamiltonians f it is easy to compute ∇f and its extension to $\overline{\mathcal{C}_n}$, but it is not straightforward to compute $\nabla\left(h\circ f\right)\left(X\right)$ and its extension to $\overline{\mathcal{C}_n}$ for arbitrary h. In these cases, calculating $h'\left(f\left(X\right)\right)\nabla f\left(X\right)$ is a much simpler task. Further, as will be shown in Lemma 5.1, the two quantities $h'\left(f\left(X\right)\right)\nabla f\left(X\right)$ and $\nabla\left(h\circ f\right)\left(X\right)$ are close to each other.

As an example application of Theorem 3.6, we show that the conditional distribution $\Pr\left[Y = y \mid f\left(y\right) \geq tn\right]$ described in item (2) in Section 2.1.2 can be approximated by a "smoothed-out" distribution, which gives equal mass to vectors y satisfying $f\left(y\right) \geq nt$ and no mass to vectors y satisfying $f\left(y\right) < \left(t-\delta\right)n$. This "smoothed-out" distribution is obtained from a "smoothed-cutoff" approximation to the $\tilde{f}$ described in Section 2.1.2. Our framework can be applied to this "smoothed-cutoff" function, yielding an equation corresponding to the Lagrange multiplier problem associated with the rate function.

Theorem 3.8 (Large deviations)**.** *Let $t > 0$. Let $f : \mathcal{C}_n \to \mathbb{R}$ be a Hamiltonian with parameters D, L_1 and L_2 as described in Theorem 3.1, and assume that there exists $z \in \mathcal{C}_n$ such that $f\left(z\right) \geq tn$. Let $\delta > 0$. There exists a monotone function $h : \mathbb{R} \to \mathbb{R}$, such that for $\varphi = h\circ f$, we have that $\varphi\left(y\right) = 0$ if $f\left(y\right) < \left(t - 2\delta\right)n$, $\varphi\left(y\right) = 1$ if $f\left(y\right) \geq tn$ and such that the following holds. Denote by σ the measure defined by*

$$d\sigma = \frac{\varphi d\mu}{\int_{\mathcal{C}_n}\varphi d\mu},$$

and let X_φ be a random variable whose law is σ. Denote

$$\mathcal{X}_g = \Big\{ X \in \overline{\mathcal{C}_n} : \exists \lambda \in \mathbb{R}\, s.t.\ \|X - \tanh(\lambda \nabla f(X))\|_1 \le 5000 \tilde{L}_1 \tilde{L}_2^{3/4} \tilde{D}^{1/4} n^{3/4} + \tilde{L}_3 \\ \text{and } f(X) \in [(t - 6\delta) n, tn] \}$$

where

$$\begin{aligned} \tilde{D} &= \frac{2}{\delta} D + \frac{2}{\delta^2} L_1^2 \\ \tilde{L}_1 &= \max\left\{1, \frac{2}{\delta} L_1\right\} \\ \tilde{L}_2 &= \max\left\{1, \frac{2}{\delta} L_2^2 + 3\frac{2}{\delta^2} L_1^2\right\} \\ \tilde{L}_3 &= 2\frac{2}{\delta^2} L_1^2 n^{1/2}. \end{aligned}$$

Then X_φ is a $\left(\rho, 80\frac{\tilde{D}^{1/4}}{n^{1/4}} + 8 \cdot 2^{-n}\right)$-mixture such that

$$\rho(\mathcal{X}_g) \ge 1 - \frac{165 L_1 \tilde{D}^{1/4}}{n^{1/4} 2\delta} \left(1 - \frac{L_1}{2\delta\sqrt{n}} - 2^{-n}\right)^{-1}. \tag{3.8}$$

Note that the expression $X - \tanh(\lambda \nabla f(X))$ in the definition of the set $\mathcal{X}_g$ is closely related to the rate function: Consider the variational problem

$$\begin{aligned} &\text{minimize } H(Y) \\ &\text{subject to } \mathbb{E} f(Y) \ge tn \end{aligned}$$

where Y is a random vector in $\mathcal{C}_n$ whose entries are independent. By monotonicity, the minimum is attained on the boundary of the constraint. Denoting $\mathbb{E}Y = y$ and using the method of Lagrange multipliers, we obtain the equations

$$\begin{aligned} \nabla_y H(Y) &= \lambda \nabla f(y) \\ f(y) &= tn. \end{aligned} \tag{3.9}$$

Applying the fact that $\nabla_y H(Y) = \tanh^{-1}(y)$ on equation (3.9) gives exactly the equation $X - \tanh(\lambda \nabla f(X)) = 0$.

Example 3.9 (Large deviations for triangle counts)**.** Let $N > 0$ be an integer representing the number of vertices of a graph, and let $n = \binom{N}{2}$ be the number of possible edges in the graph. We treat each vector $v \in \mathcal{C}_n$ as a simple graph, with $v_e = 1$ if and only if the edge e appears in the graph. This in turns gives an adjacency matrix $(x_{ij})_{i,j=1}^N$ with $x_{ij} = 1$ if and only if $v_{\{ij\}} = 1$. In this setting, let f be a triangle-counting function,

$$f(x) = \frac{\beta}{N} \sum_{i \ne j \ne k} x_{ij} x_{jk} x_{ki}$$

for some real β. It is shown in [4] that $\mathcal{D}(f)$ is $O\left(n^{3/4}\right)$ and in [5] that L_1 and L_2 are bounded by $200|\beta|$. Thus we can apply Theorem 3.8 to f, concluding that for a fixed $t > 0$ there exists some $\delta = o(1)$ and a smoothed cutoff function h with $h(x) = 1$ for $x > tn$ and $h(x) = 0$ for $x < (t-\delta)n$ and such that the random graph G whose density is proportional to $h \circ f$ is a $(\rho, o(1))$-mixture such that $\rho(\mathcal{X}_g) = 1 - o(1)$, where

$$\mathcal{X}_g = \left\{ X \in \overline{\mathcal{C}_n} : \exists \lambda \in \mathbb{R}\, s.t.\ \left\|X - \tanh\left(\lambda X^2\right)\right\|_1 \le \varepsilon n \\ \text{and } f(X) \in [(t - 6\delta) n, tn] \right\}$$

for some $\varepsilon = o(1)$. Here $X \in \overline{\mathcal{C}_n}$ is treated as an $n \times n$ symmetric matrix with zeros on the diagonal, and we understand the expression X^2 as the usual matrix multiplication, with zeros on the diagonal as well. We conjecture that all of the points of the set $\mathcal{X}_g$ are close to the solutions obtained by Lubetzky and Zhao in [6].

Our results extend to triangle counts on sparse graphs as well. In this case, expected value of f is of order np^3, which decays to 0 as $p \to 0$. We should therefore take both t to be proportional to p^3 and δ to be $o\left(p^3\right)$. Since the bound on the vectors in $\mathcal{X}_g$ in Theorem 3.8 is polynomial in δ, we can consider large deviations for graphs whose edge probabilities are proportional to $p \sim n^{-c}$ for some constant c (for example, if we wish ε to be of order p, we can take $p \sim n^{-1/160}$).

The rest of this paper is organized as follows. Theorem 3.1 is proved in Section 4, while Theorem 3.6 is proved in Section 5. Corollaries 3.4 and 3.5 are proved in Section 6.1 and 3.8 is proved in Section 6.2.

4 Proof of main theorem

4.1 Notation and review

We use the notation from [4], and rely on the proofs therein. Here is a brief review of the required terms and bounds.

For a probability measure ν on $\mathcal{C}_n$, we define $f_\nu = \log\left(d\nu/d\mu\right)$, so that the Gibbs distribution with Hamiltonian f_ν is exactly ν. For every distribution ν on the hypercube (exponential or otherwise), we define

$$\mathcal{H}\left(\nu\right) = \int_{\mathcal{C}_n} \tanh\left(\nabla f_\nu\left(y\right)\right)^{\otimes 2} d\nu - \left(\int_{\mathcal{C}_n} \tanh\left(\nabla f_\nu\left(y\right)\right) d\nu\right)^{\otimes 2},$$

which should be thought of as the covariance matrix of the random variable $\nabla f_\nu\left(X\right)$ with $X \sim \nu$. We will use the following three results from [4].

Proposition 4.1 (Proposition 17 in [4]). *Let $\tilde{\nu}$ be a probability distribution on $\mathcal{C}_n$. Then there exists a product measure $\xi = \xi\left(\tilde{\nu}\right)$ such that*

$$W_1\left(\tilde{\nu}, \xi\right) \leq \sqrt{n \mathit{Tr}\left(\mathcal{H}\left(\tilde{\nu}\right)\right)}. \tag{4.1}$$

Moreover, one may take ξ to be the unique product measure whose center of mass lies at $\int_{\mathcal{C}_n} \tanh\left(\nabla f_{\tilde{\nu}}\left(y\right)\right) d\tilde{\nu}\left(y\right)$ where the $\tanh$ is applied entrywise.

Proposition 4.2 (Proposition 18 together with Lemma 16 in [4]). *Define $D = \mathcal{D}\left(f_\nu\right)$. Let $\varepsilon \in \left(0, 1/4\sqrt{\log\left(4n/D\right)}\right)$. Let ν be a probability measure on $\mathcal{C}_n$ and define $f = \log \frac{d\nu}{d\mu}$. Then there exists a measure m on $B\left(0, \varepsilon\sqrt{n}\right) \cap \left[-1/4, 1/4\right]^n$, such that ν admits the decomposition*

$$\int_{\mathcal{C}_n} \varphi d\nu = \int_{B\left(0, \varepsilon\sqrt{n}\right)} \left(\int_{\mathcal{C}_n} \varphi d\tau_\theta\left(\nu\right)\right) dm\left(\theta\right) \tag{4.2}$$

for every test function $\varphi : \mathcal{C}_n \to \mathbb{R}$, and which satisfies

$$m\left(\theta :\ \mathit{Tr}\left(\mathcal{H}\left(\tau_\theta \nu\right)\right) \leq 256 \frac{n^{1/3} D^{2/3}}{\varepsilon^{2/3}}\right) \geq 1 - \frac{3D^{1/3}}{n^{1/3}\varepsilon^{1/3}}. \tag{4.3}$$

Lemma 4.3 (Lemma 24 in [4]). *Let $\theta \in \mathbb{R}^n$ and let $\nu, \tilde{\nu}$ be probability measures on $\mathcal{C}_n$. Define*

$$A = \int_{\mathcal{C}_n} \tanh\left(\nabla f_\nu\left(y\right)\right)^{\otimes 2} d\tilde{\nu} - \left(\int_{\mathcal{C}_n} \tanh\left(\nabla f_\nu\left(y\right)\right) d\tilde{\nu}\right)^{\otimes 2}$$

and

$$B = \int_{\mathcal{C}_n} \tanh\left(\nabla f_{\tau_\theta \nu}\left(y\right)\right)^{\otimes 2} d\tilde{\nu} - \left(\int_{\mathcal{C}_n} \tanh\left(\nabla f_{\tau_\theta \nu}\left(y\right)\right) d\tilde{\nu}\right)^{\otimes 2}.$$

Then

$$e^{-4\|\theta\|_\infty} TrB \le TrA \le e^{4\|\theta\|_\infty} TrB.$$

We can now describe the general plan of our proof. Fix $\varepsilon > 0$, and let m be the measure obtained from Proposition 4.2. Denote by Θ the set

$$\Theta = \left\{ \theta \in \mathbb{R}^n : \operatorname{Tr}\left(\mathcal{H}\left(\tau_\theta \nu\right)\right) \le 256 \frac{n^{1/3} D^{2/3}}{\varepsilon^{2/3}} \right\}. \tag{4.4}$$

For every $\theta \in \mathbb{R}^n$, denote by ξ_θ the unique product measure with the same marginals as $\tau_\theta \nu$, and by $A(\theta)$ the vector

$$A(\theta) = \mathbb{E}_{X \sim \tau_\theta \nu}[X].$$

Denote by ρ the push-forward of the measure m under the map $\theta \mapsto A(\theta)$ and define

$$\mathcal{X} = \{A(\theta); \theta \in \Theta\}.$$

In order to prove Theorem 3.1, all we have to do is that show that for each $\theta \in \Theta$, the corresponding $A(\theta)$ is close in the one-norm to $\tanh(\nabla f(A(\theta)))$; this will show equation (3.5). In other words, we need the following proposition:

Proposition 4.4. *Let $\theta \in \Theta$ and let $Y \sim \xi_\theta$. Then for every $\varepsilon > 0$,*

$$\|\tanh(\nabla f(\mathbb{E}Y)) - \mathbb{E}Y\|_1 \le 41 L_1 \left(112 L_2 \frac{n^{2/3} D^{1/3}}{\varepsilon^{1/3}} + \varepsilon n\right).$$

Relying on the above, we can prove of Theorem 3.1.

Proof of Theorem 3.1. Define the measure ρ and the set $\mathcal{X}$ as above. Set $\varepsilon = \frac{D^{1/4} L_2^{3/4}}{n^{1/4}}$. Items (1)–(3) in Definition 2.7 follow immediately from Proposition 4.1 and 4.2 by choice of ε, δ and ρ. By Proposition 4.4 for all $\theta \in \Theta$, we have

$$\begin{aligned}\|\tanh(\nabla f(\mathbb{E}Y)) - \mathbb{E}Y\|_1 &\le 41 L_1 \left(113 L_2^{3/4} D^{1/4} n^{3/4}\right) \\ &\le 5000 L_1 L_2^{3/4} D^{1/4} n^{3/4}.\end{aligned}$$

This implies that $\mathcal{X} \subseteq \mathcal{X}_f$, and together with Proposition 4.2 and by choice of ε, this shows that $\rho(\mathcal{X}_f) \ge 1 - \frac{3D^{1/4}}{n^{1/4}}$, satisfying equation (3.5). $\square$

The rest of this section is devoted to proving Proposition 4.4.

4.2 Approximate fixed point

Let $\theta \in \Theta$ be a tilt and let ξ_θ be the product measure whose center of mass lies at $\int_{\mathcal{C}_n} \tanh(\nabla f_{\tau_\theta \nu}(y))\, d\tau_\theta \nu(y)$. Throughout the proof we will assume $X \sim \tau_\theta \nu$ and $Y \sim \xi_\theta$. A direct calculation shows that under this notation, $\mathbb{E}Y = \mathbb{E}\tanh(\nabla f(X) + \theta)$:

$$\begin{aligned}\mathbb{E}Y &= \int_{\mathcal{C}_n} \tanh(\nabla f_{\tau_\theta \nu}(y))\, d\tau_\theta \nu(y) \\ &= \int_{\mathcal{C}_n} \tanh\left(\nabla\left(\log\left(\frac{d\tau_\theta \nu}{d\nu}\right) + \log\left(\frac{d\nu}{d\mu}\right)\right)(y)\right) d\tau_\theta \nu(y) \\ &= \int_{\mathcal{C}_n} \tanh(\theta + \nabla f_\nu(y))\, d\tau_\theta \nu(y) \\ &= \mathbb{E}\tanh(\nabla f(X) + \theta).\end{aligned}$$

This gives

$$\begin{aligned}\left\|\mathbb{E}Y-\mathbb{E}\tanh\left(\nabla f\left(X\right)\right)\right\|_1 &\leq \left\|\mathbb{E}\tanh\left(\nabla f\left(X\right)\right)-\mathbb{E}\tanh\left(\nabla f\left(X\right)+\theta\right)\right\|_1\\ &\leq \left\|\theta\right\|_1\\ \left(\left\|\theta\right\|_2\leq\varepsilon\sqrt{n}\right) &\leq \varepsilon n.\end{aligned} \tag{4.5}$$

where in the second inequality we use the fact that

$$\left|\tanh\left(x\right)-\tanh\left(y\right)\right|\leq\left|x-y\right|. \tag{4.6}$$

Proposition 4.5. *Let $Y\sim\xi_\theta$. Then*

$$\mathbb{E}\left\|\tanh\left(\nabla f\left(Y\right)\right)-\mathbb{E}Y\right\|_1\leq 64L_2\frac{n^{2/3}D^{1/3}}{\varepsilon^{1/3}}+\varepsilon n.$$

Proof. For $X\sim\tau_\theta\nu$, consider the random variable $Z=\left\|\tanh\left(\nabla f(X)\right)-\mathbb{E}\tanh\left(\nabla f(X)\right)\right\|_2^2$. A short calculation shows that the expectation of Z is roughly $\mathrm{Tr}\mathcal{H}\left(\tau_\theta\nu\right)$:

$$\begin{aligned}\mathbb{E}Z &= \mathbb{E}\left\|\tanh\left(\nabla f\left(X\right)\right)-\mathbb{E}\tanh\left(\nabla f\left(X\right)\right)\right\|_2^2\\ &= \sum_{i=1}^n\mathbb{E}\left[\tanh\left(\nabla f\left(X\right)\right)_i^2\right]-\sum_{i=1}^n\left(\mathbb{E}\tanh\left(\nabla f\left(X\right)\right)_i\right)^2\\ &\leq 3\left(\sum_{i=1}^n\mathbb{E}\left[\tanh\left(\nabla f\left(X\right)+\theta\right)_i^2\right]-\sum_{i=1}^n\left(\mathbb{E}\tanh\left(\nabla f\left(X\right)+\theta\right)_i\right)^2\right)\\ &= 3\mathrm{Tr}\left(\mathcal{H}\left(\tau_\theta\nu\right)\right)\end{aligned}$$

where the inequality is by Lemma 4.3 with ν and $\tilde{\nu}=\tau_\theta\nu$ and the fact that $\left\|\theta\right\|_\infty\leq 1/4$. Thus by equation (4.3),

$$\mathbb{E}\left\|\tanh\left(\nabla f\left(X\right)\right)-\mathbb{E}\tanh\left(\nabla f\left(X\right)\right)\right\|_2^2\leq 3\cdot 256\frac{n^{1/3}D^{2/3}}{\varepsilon^{2/3}},$$

and together with the Cauchy-Schwarz inequality, we have that

$$\begin{aligned}\mathbb{E}\left\|\tanh\left(\nabla f\left(X\right)\right)-\mathbb{E}\tanh\left(\nabla f\left(X\right)\right)\right\|_1 &\leq \sqrt{n}\mathbb{E}\left\|\tanh\left(\nabla f\left(X\right)\right)-\mathbb{E}\tanh\left(\nabla f\left(X\right)\right)\right\|_2\\ &\leq 32\frac{n^{2/3}D^{1/3}}{\varepsilon^{1/3}}.\end{aligned} \tag{4.7}$$

By Proposition 4.1, there exists a coupling between $\tau_\theta\nu$ and ξ_θ such that

$$\begin{aligned}\mathbb{E}\left\|X-Y\right\|_1 &\leq 2\sqrt{n\mathrm{Tr}\mathcal{H}\left(\tau_\theta\nu\right)}\\ \text{(by equation (4.3))} &\leq 32\frac{n^{2/3}D^{1/3}}{\varepsilon^{1/3}}.\end{aligned}$$

Thus, since by equations (3.3) and (4.6),

$$\begin{aligned}\mathbb{E}\left\|\tanh\left(\nabla f\left(X\right)\right)-\tanh\left(\nabla f\left(Y\right)\right)\right\|_1 &\leq \mathbb{E}\left\|\nabla f\left(X\right)-\nabla f\left(Y\right)\right\|_1\\ &\leq L_2\mathbb{E}\left\|X-Y\right\|_1\\ &\leq 32L_2\frac{n^{2/3}D^{1/3}}{\varepsilon^{1/3}}.\end{aligned} \tag{4.8}$$

Combining equations (4.7), (4.5) and (4.8) together with the triangle inequality finally gives

$$\begin{aligned}\mathbb{E}\left\|\tanh\left(\nabla f\left(Y\right)\right)-\mathbb{E}Y\right\|_1 &\leq 32\left(1+L_2\right)\frac{n^{2/3}D^{1/3}}{\varepsilon^{1/3}}+\varepsilon n\\ &\leq 64L_2\frac{n^{2/3}D^{1/3}}{\varepsilon^{1/3}}+\varepsilon n\end{aligned}$$

as needed. ☐

Lemma 4.6. *Let Z be an almost-surely bounded random variable, $|Z| \le L$ with $L \ge 1$. Then*

$$\left|\tanh\left(\mathbb{E}Z\right) - \mathbb{E}\tanh\left(Z\right)\right| \le 20L \cdot \mathbb{E}\left|\tanh\left(Z\right) - \mathbb{E}\tanh\left(Z\right)\right|.$$

The proof is postponed to the appendix.

Claim 4.7. *Let ξ be a product measure on $\mathcal{C}_n$, let $Y \sim \xi$, and let $f : \mathcal{C}_n \to \mathbb{R}$ be a function on the hypercube. Then*

$$\mathbb{E}f\left(Y\right) = f\left(\mathbb{E}Y\right) \tag{4.9}$$

and

$$\mathbb{E}\nabla f\left(Y\right) = \nabla f\left(\mathbb{E}Y\right). \tag{4.10}$$

Proof. The extension of f to $\overline{\mathcal{C}_n}$ is defined by the Fourier decomposition

$$f\left(y\right) = \sum_{S\subseteq[n]} \hat{f}\left(S\right) \prod_{i\in S} y_i.$$

Thus, since ξ is a product measure,

$$\mathbb{E}f\left(Y\right) = \mathbb{E}\sum_{S\subseteq[n]} \hat{f}\left(S\right)\prod_{i\in S} Y_i = \sum_{S\subseteq[n]} \hat{f}\left(S\right)\prod_{i\in S}\mathbb{E}Y_i = f\left(\mathbb{E}Y\right).$$

Equation 4.10 is then obtained by applying equation 4.9 to every component of ∇f. □

Proof Proposition 4.4. By the triangle inequality,

$$\begin{aligned}
&\left\|\tanh\left(\nabla f\left(\mathbb{E}Y\right)\right) - \mathbb{E}Y\right\|_1 \\
&\quad\le \left\|\tanh\left(\nabla f\left(\mathbb{E}Y\right)\right) - \mathbb{E}\tanh\left(\nabla f\left(Y\right)\right)\right\|_1 + \left\|\mathbb{E}\tanh\left(\nabla f\left(Y\right)\right) - \mathbb{E}Y\right\|_1 \\
\text{(by Claim 4.7)} &\quad= \left\|\tanh\left(\mathbb{E}\nabla f\left(Y\right)\right) - \mathbb{E}\tanh\left(\nabla f\left(Y\right)\right)\right\|_1 + \left\|\mathbb{E}\tanh\left(\nabla f\left(Y\right)\right) - \mathbb{E}Y\right\|_1 \\
\text{(by convexity)} &\quad\le \left\|\tanh\left(\mathbb{E}\nabla f\left(Y\right)\right) - \mathbb{E}\tanh\left(\nabla f\left(Y\right)\right)\right\|_1 + \mathbb{E}\left\|\tanh\left(\nabla f\left(Y\right)\right) - \mathbb{E}Y\right\|_1.
\end{aligned} \tag{4.11}$$

Proposition 4.5 gives a bound on the second term in the right hand side.

For the first term, note that by equation (3.2), for every index $j \in [n]$,

$$\left|\nabla f\left(Y\right)_j\right| \le L_1.$$

We can therefore invoke Lemma 4.6 on every index, giving that

$$\begin{aligned}
\left\|\tanh\left(\mathbb{E}\nabla f\left(Y\right)\right) - \mathbb{E}\tanh\left(\nabla f\left(Y\right)\right)\right\|_1 &= \sum_{j=1}^{n}\left|\tanh\left(\mathbb{E}\nabla f\left(Y\right)\right)_j - \mathbb{E}\tanh\left(\nabla f\left(Y\right)\right)_j\right| \\
\text{(by Lemma 4.6)} &\le 20L_1\sum_{j=1}^{n}\mathbb{E}\left|\tanh\left(\nabla f\left(Y\right)\right)_j - \mathbb{E}\tanh\left(\nabla f\left(Y\right)_j\right)\right| \\
&= 20L_1\mathbb{E}\left\|\tanh\left(\nabla f\left(Y\right)\right) - \mathbb{E}\tanh\left(\nabla f\left(Y\right)\right)\right\|_1.
\end{aligned} \tag{4.12}$$

For this last term, we again use the triangle inequality and equation (4.5), giving

$$\begin{aligned}
\mathbb{E}\left\|\tanh\left(\nabla f\left(Y\right)\right) - \mathbb{E}\tanh\left(\nabla f\left(Y\right)\right)\right\|_1 &\le \mathbb{E}\left\|\tanh\left(\nabla f\left(Y\right)\right) - \mathbb{E}\tanh\left(\nabla f\left(X\right)\right)\right\|_1 + \\
&\quad + \mathbb{E}\left\|\mathbb{E}\tanh\left(\nabla f\left(X\right)\right) - \mathbb{E}\tanh\left(\nabla f\left(Y\right)\right)\right\|_1 \\
&\le \varepsilon n + \mathbb{E}\left\|\tanh\left(\nabla f\left(Y\right)\right) - \mathbb{E}Y\right\|_1 + \\
&\quad + \mathbb{E}\left\|\tanh\left(\nabla f\left(X\right)\right) - \tanh\left(\nabla f\left(Y\right)\right)\right\|_1
\end{aligned}$$

which, by Proposition 4.5 and equation (4.8), gives

$$\begin{aligned}\mathbb{E}\left\|\tanh\left(\nabla f\left(Y\right)\right)-\mathbb{E}\tanh\left(\nabla f\left(Y\right)\right)\right\|_1 &\leq \varepsilon n+64L_2\frac{n^{2/3}D^{1/3}}{\varepsilon^{1/3}}+\varepsilon n\\ &\quad+32L_2\frac{n^{2/3}D^{1/3}}{\varepsilon^{1/3}}\\ &\leq 96L_2\frac{n^{2/3}D^{1/3}}{\varepsilon^{1/3}}+2\varepsilon n.\end{aligned}$$

Putting this into equation (4.12), we have

$$\left\|\tanh\left(\mathbb{E}\nabla f\left(Y\right)\right)-\mathbb{E}\tanh\left(\nabla f\left(Y\right)\right)\right\|_1 \leq 40L_1\left(48L_2\frac{n^{2/3}D^{1/3}}{\varepsilon^{1/3}}+\varepsilon n\right).$$

And finally, plugging in the bounds into equation (4.11), we get

$$\begin{aligned}\left\|\tanh\left(\nabla f\left(\mathbb{E}Y\right)\right)-\mathbb{E}Y\right\|_1 &\leq 40L_1\left(48L_2\frac{n^{2/3}D^{1/3}}{\varepsilon^{1/3}}+\varepsilon n\right)\\ &\quad+64L_2\frac{n^{2/3}D^{1/3}}{\varepsilon^{1/3}}+\varepsilon n\\ &\leq 41L_1\left(112L_2\frac{n^{2/3}D^{1/3}}{\varepsilon^{1/3}}+\varepsilon n\right).\end{aligned}$$

□

5 Proof of composition theorem

We will use two lemmas concerning the relation between f and $h \circ f$. The first is a discrete chain rule which will be central to our calculations:

Lemma 5.1 (Chain rule for discrete gradient)**.** *Let $f : \mathcal{C}_n \to \mathbb{R}$ with Lip $(f) = L$ and let $h : \mathbb{R} \to \mathbb{R}$ with $|h''(x)| < B$. Then*

1. For ever $y \in \mathcal{C}_n$,
$$\left\|\nabla\left(h\circ f\right)\left(y\right)-h'\left(f\left(y\right)\right)\nabla f\left(y\right)\right\|_1 \leq BL^2 n \tag{5.1}$$
and
$$\left\|\nabla\left(h\circ f\right)\left(y\right)-h'\left(f\left(y\right)\right)\nabla f\left(y\right)\right\|_2 \leq BL^2\sqrt{n}. \tag{5.2}$$
2. For every $x \in \overline{\mathcal{C}_n}$,
$$\left\|\nabla\left(h\circ f\right)\left(x\right)-h'\left(f\left(x\right)\right)\nabla f\left(x\right)\right\|_1 \leq 2BL^2 n^{3/2}. \tag{5.3}$$

The second lemma concerns the parameters of the function $h \circ f$:

Lemma 5.2 (Composition parameters)**.** *Let $h : \mathbb{R} \to \mathbb{R}$ be a twice differentiable function satisfying*

$$\begin{aligned}\left|h'\left(x\right)\right| &\leq B_1\\ \left|h''\left(x\right)\right| &\leq B_2\end{aligned}$$

for all $x \in \mathbb{R}$. Let $f : \mathcal{C}_n \to \mathbb{R}$ be a function with parameters D, L_1, L_2 as described in Theorem 3.1. Then

$$\mathcal{D}\left(h\circ f\right) \leq B_1 D + B_2 L_1^2 n \tag{5.4}$$

$$\textit{Lip}\left(h\circ f\right) \leq B_1 L_1 \tag{5.5}$$

$$\max_{x\neq y\in\mathcal{C}_n}\frac{\left\|\nabla\left(h\circ f\right)\left(x\right)-\nabla\left(h\circ f\right)\left(y\right)\right\|_1}{\left\|x-y\right\|_1} \leq B_1 L_2 + 3B_2 L_1^2 n. \tag{5.6}$$

The proofs of both lemmas are postponed to the appendix.

Proof of Theorem 3.6. Denote by $\mathcal{X}_{h\circ f}$ the set

$$\mathcal{X}_{h\circ f} = \left\{ X \in [-1,1]^n : \left\| X - \tanh\left(\nabla\left(h\circ f\right)(X)\right)\right\|_1 \le 5000\tilde{L}_1\tilde{L}_2^{3/4}\tilde{D}^{1/4}n^{3/4} \right\}.$$

Note that for every $X \in \mathcal{X}_{h\circ f}$,

$$\begin{aligned} \left\| X - \tanh\left(h'(X)\nabla f(X)\right)\right\|_1 &\le \left\| X - \tanh\left(\nabla\left(h\circ f\right)(X)\right)\right\|_1 \\ &+ \quad \left\|\tanh\left(\nabla\left(h\circ f\right)(X)\right) - \tanh\left(h'(X)\nabla f(X)\right)\right\|_1 \\ \text{(by equation (5.3))} &\le 5000\tilde{L}_1\tilde{L}_2^{3/4}\tilde{D}^{1/4}n^{3/4} + 2B_2L_1^2n^{3/2} \end{aligned}$$

and so $\tilde{\mathcal{X}}_{h\circ f} \subseteq \mathcal{X}_{h\circ f}$. Applying Theorem 3.1 for the function $h\circ f$ with the bounds given by Lemma 5.2 gives the required results. □

Remark 5.3. The bound for compositions $h\circ f$ with domain $\overline{C_n}$, given in (5.3), is worse by a factor of $\sqrt{n}$ than that of compositions with domain $\mathcal{C}_n$, given in (5.1). This disparity is in fact tight. For example, consider the function

$$h(x) = \begin{cases} \frac{3}{4}x^3 - \frac{1}{4}x^5 & |x| < 1 \\ \frac{1}{2}x^2 & x \ge 1 \\ -\frac{1}{2}x^2 & x \le -1 \end{cases}$$

applied to the "counting" function

$$f(x) = \sum_{i=1}^n x_i.$$

The function h has a bounded second derivative and satisfies $h'(0) = 0$. For $x = 0$, we have $f(x) = 0$ and so $h'(f(x))\nabla f(x) = 0$ as well. However, a calculation shows that $\left\|\nabla\left(h\circ f\right)(x)\right\|_1 \sim n^{3/2}$, and so $\left\|\nabla\left(h\circ f\right)(x) - h'(f(x))\nabla f(x)\right\|_1 \sim n^{3/2}$ as well.

6 Example applications

6.1 The Ising model

Proof of Corollary 3.4. A short calculation shows that $\nabla f(x) = Ax + \mu$. The corollary will follow immediately from Theorem 3.1 once we have obtained the parameters D, L_1 and L_2 for f. The calculations for $\mathcal{D}(f)$ and $\mathrm{Lip}(f)$ are also found in [4, Section 1.3] but we repeat them here for completeness.

Denote $\mu_{\max} = \max_{i\in[n]}|\mu_i|$. We then have

1. The Gaussian-width is bounded by:

$$\begin{aligned} \mathcal{D}(f) &= \mathbb{E}\sup_{x\in\mathcal{C}_n}\langle Ax+\mu, \Gamma\rangle \\ &\le \mathbb{E}\sup_{x\in\mathcal{C}_n}\langle Ax, \Gamma\rangle + \mathbb{E}\left|\langle \mu, \Gamma\rangle\right| \\ &\le \sqrt{n}\mathbb{E}\sup_{x\in B(0,1)}\langle Ax, \Gamma\rangle + \|\mu\|_2 \\ &= \sqrt{n}\mathbb{E}\left\|A\Gamma\right\|_2 + \|\mu\|_2 \\ &\le \sqrt{n\mathbb{E}\left\|A\Gamma\right\|_2^2} + \|\mu\|_2 \\ &\le \sqrt{n\mathrm{Tr}A^2} + \sqrt{n}\mu_{\max}. \end{aligned}$$

2. The Lipschitz constant is bounded by

$$\begin{aligned}\text{Lip}(f) &\leq \mu_{\max} + \max_{i\in[n],x\in\mathcal{C}_n} \langle Ax, e_i\rangle \\ &\leq \mu_{\max} + \max_{i\in[n]} \sum_{j\in[n]} |A_{ij}|\,.\end{aligned}$$

3. Regarding the Lipschitz constant of the gradient, note that $\|\nabla f(x) - \nabla f(y)\|_1 = \|A(x-y)\|_1$. Suppose that x and y differ only in the i-th coordinate. Then $A(|x-y|)$ is just 2 times the i-th column of A. By the triangle inequality, we then have

$$\frac{\|\nabla f(x) - \nabla f(y)\|_1}{\|x-y\|_1} \leq \max_{i\in[n]} \sum_{j\in[n]} |A_{ij}|\,. \qquad \square$$

Proof of Corollary 3.5. The interactions described in Corollary 3.5 can be represented by an interaction matrix $A = \frac{\beta\mathbf{1}}{n}$, where $\mathbf{1}$ is the $n\times n$ matrix whose off-diagonal entries are 1 and whose diagonal is 0, and β is interpreted as the inverse temperature. Note that for every $x, y \in \mathcal{C}_n$,

$$\|\nabla f(x) - \nabla f(y)\|_1 = \|A(x-y)\|_1 \leq \beta\|x-y\|_1\,, \tag{6.1}$$

so that $L_2 \leq 1+\beta$. A simple calculation also shows that $D \leq \beta\sqrt{n}$ and $L_1 \leq 1+\beta$. Denoting

$$\mathcal{X} = \left\{X \in \overline{\mathcal{C}_n} : \left\|X - \tanh\left(\frac{\beta\mathbf{1}}{n}X\right)\right\|_1 \leq 5000(1+\beta)^2 n^{7/8}\right\},$$

by Corollary 3.4 we have that X_n^f is a $\left(\rho, 3n^{-1/8}, 3n^{-1/8}\right)$-tilt-mixture with $\rho(\mathcal{X}) \geq 1 - 3n^{-1/8}$. Denote by $\boldsymbol{J} = \mathbf{1} + \text{Id}$ the $n\times n$ matrix whose every entry is 1. Then every $X \in \mathcal{X}$ also satisfies

$$\begin{aligned}\left\|X - \tanh\left(\frac{\beta\boldsymbol{J}}{n}X\right)\right\|_1 &= \left\|X - \tanh\left(\frac{\beta\boldsymbol{J}}{n}X\right) - \tanh\left(\frac{\beta\mathbf{1}}{n}X\right) + \tanh\left(\frac{\beta\mathbf{1}}{n}X\right)\right\|_1 \\ &\leq 5000(1+\beta)^2 n^{7/8} + \left\|\tanh\left(\frac{\beta\mathbf{1}}{n}X\right) - \tanh\left(\frac{\beta(\mathbf{1}+\text{Id})}{n}X\right)\right\|_1 \\ &\leq 5000(1+\beta)^2(1+\beta)n^{7/8} + \left\|\frac{\beta\text{Id}}{n}X\right\|_1 \\ &\leq 5001(1+\beta)^2 n^{7/8}.\end{aligned}$$

Thus $\mathcal{X} \subseteq \mathcal{X}_f$ and the first part of Corollary 3.5 is proved. The fixed point equation $X = \tanh\left(\frac{\beta\boldsymbol{J}}{n}X\right)$ is easier to work with, since all of its exact solutions are constant: Indeed, every entry X_i of a solution satisfies $X_i = \tanh\left(\sum_{j=1}^n \frac{\beta}{n}X_j\right)$; every solution X is then of the form $X = (x, x, \dots, x)$, and the exact fixed point vector equation reduces to the scalar equation

$$x = \tanh(\beta x)\,.$$

The value $x_0 = 0$ is always a solution, corresponding to the case where the typical configuration is completely disordered.

For $\beta \leq 1$, this is also the only solution. In this case, for every $X \in \mathcal{X}_f$,

$$\begin{aligned}\|X\|_1 &= \left\|X - \tanh\left(\frac{\beta\boldsymbol{J}}{n}X\right) + \tanh\left(\frac{\beta\boldsymbol{J}}{n}X\right)\right\|_1 \\ &\leq 5001(1+\beta)^2 n^{7/8} + \left\|\tanh\left(\frac{\beta\boldsymbol{J}}{n}X\right)\right\|_1 \\ &\leq 5001(1+\beta)^2 n^{7/8} + \beta\|X\|_1\,.\end{aligned}$$

Rearranging, we get that every $X \in \mathcal{X}_f$ is close to 0:

$$\|X\|_1 \leq 5001 \frac{(1+\beta)^2}{1-\beta} n^{7/8}.$$

This represents the fact that for high temperatures, the system is always disordered.

For $\beta > 1$, there are two other solutions, $x_1 = -x_2$. These satisfy $|x_1|, |x_2| \to 1$ as $\beta \to \infty$, and correspond to the symmetry-broken phase where all spins tend to point in the same direction. Showing that every $X \in \mathcal{X}_f$ is close to either $(x_1, x_1, \ldots, x_1)$ or $(x_2, x_2, \ldots, x_2)$ can then be done by a standard counting argument, which we choose to omit. □

Adding a constant magnetic field $\mu = (\mu_0, \mu_0 \ldots, \mu_0)$ forces a non-zero constant solution for every $\beta > 0$, while shifting the values of x_1 and x_2.

6.2 Large deviations

In order to prove Theorem 3.8, we follow the approach mentioned in Section 2.1.2, and try to approximate function $\tilde{f}$ in equation (2.1) by a well-behaved Hamiltonian g.

Let $t \in \mathbb{R}$ and $\delta > 0$. Let $h : \mathbb{R} \to \mathbb{R}$ and $\psi : \mathbb{R} \to \mathbb{R}$ be defined as

$$h(x) = \begin{cases} 2x+1 & x \leq -1 \\ -x^2 & -1 \leq x \leq 0 \\ 0 & x \geq 0. \end{cases}$$

and

$$\psi(x) = n \cdot h\left(\left(\frac{x}{n} - t\right)/\delta\right).$$

Note that $|h'(x)| \leq 2$ and $|h''(x)| \leq 2$ for all $x \in \mathbb{R}$. Thus

$$\begin{aligned} |\psi'(x)| &= \left| n \cdot h'\left(\left(\frac{x}{n} - t\right)/\delta\right) \cdot \frac{1}{n\delta} \right| \\ &\leq \frac{2}{\delta} \end{aligned}$$

and

$$\begin{aligned} |\psi''(x)| &= \left| n \cdot h''\left(\left(\frac{x}{n} - t\right)/\delta\right) \cdot \frac{1}{n^2\delta^2} \right| \\ &\leq \frac{2}{n\delta^2}. \end{aligned}$$

Let $g : \mathcal{C}_n \to \mathbb{R}$ be defined as

$$g(y) = \psi(f(y)).$$

Denote by ν the measure defined by X_n^g. The function g is an approximation for $\tilde{f}$, in the sense that almost of all the mass of ν is supported on vectors on which f attains a large value.

Proposition 6.1. *Let $\delta' = \frac{\log 4 + 1}{2} \delta$ and define*

$$\mathcal{B} = \{y \in \mathcal{C}_n : \, f(y) \leq (t - \delta')n\}.$$

If there exists a $z \in \mathcal{C}_n$ such that $f(z) \geq tn$, then

$$\nu(\mathcal{B}) \leq 2^{-n}.$$

Proof. Let $y \in \mathcal{B}$. By definition of g,

$$\begin{aligned} g(y) &= n \cdot h\left(\left(\frac{f(y)}{n} - t\right)/\delta\right) \\ (h \text{ is increasing}) &\leq n \cdot h\left(\left(\frac{(t-\delta')\,n}{n} - t\right)/\delta\right) \\ &= n \cdot h\left(-\delta'/\delta\right) \\ (\text{since } \delta' > \delta) &= n\left(1 - 2\frac{\delta'}{\delta}\right) \\ &= -(\log 4)\,n. \end{aligned}$$

Let $z \in \mathcal{C}_n$ be such that $f(z) \geq tn$. Then under ν the probability for obtaining z is proportional to $e^{g(z)} = e^0 = 1$. On the other hand, for every $y \in \mathcal{B}$, the probability for obtaining y is proportional to a value smaller than $e^{-\log 4 \cdot n} = 4^{-n} = 2^{-2n}$. Since there are no more than 2^n possible vectors in $\mathcal{C}_n$, we thus obtain

$$\nu(\mathcal{B}) \leq \frac{\nu(\mathcal{B})}{\nu(z)} \leq 2^n 2^{-2n} = 2^{-n}. \qquad \square$$

Proposition 6.1 allows us to approximate ν with a distribution that does not give any mass at all to vectors $y \in \mathcal{C}_n$ with $f(y) < (t-\delta')\,n$. Define the function $\varphi : \mathcal{C}_n \to \mathbb{R}$ by

$$\varphi(y) = \begin{cases} 0 & f(y) < (t-\delta')\,n \\ e^{g(y)} & (t-\delta')\,n \leq f(y) < tn \\ 1 & f(y) \geq tn, \end{cases}$$

and observe that $\varphi(y)$ agrees with $e^{g(y)}$ for all y such that $f(y) \geq (t-\delta')\,n$. Denote by σ the measure defined by $d\sigma = \frac{\varphi d\mu}{\int_{\mathcal{C}_n} \varphi d\mu}$ and by X_φ a random variable whose law is σ.

Proposition 6.2. *Assume that there exists a $z \in \mathcal{C}_n$ such that $f(z) \geq tn$. Then there exists a coupling between X_n^g and X_φ such that*

$$\mathbb{E}\left\|X_n^g - X_\varphi\right\|_1 \leq 2n \cdot 2^{-n}.$$

We postpone the proof to the appendix.

Proof of Theorem 3.8. Applying Theorem 3.6 to g, there exists a ρ-mixture and a coupling between $X(\rho)$ and X_n^g such that

$$\rho(\mathcal{X}_g) \geq 1 - \frac{80\tilde{D}^{1/4}}{n^{1/4}} \tag{6.2}$$

and

$$\mathbb{E}\left\|X(\rho) - X_n^g\right\|_1 \leq 80n^{3/4}\tilde{D}^{1/4}.$$

Therefore by Proposition 6.2 there exists a coupling between $X(\rho)$ and X_φ such that

$$\mathbb{E}\left\|X(\rho) - X_\varphi\right\|_1 \leq 80n^{3/4}\tilde{D}^{1/4} + 2n \cdot 2^{-n}.$$

This shows that X_φ is a $\left(\rho, 80\frac{\tilde{D}^{1/4}}{n^{1/4}} + 8 \cdot 2^{-n}\right)$-mixture. To obtain equation (3.8), denote $\mathcal{Y}_g = \{X \in \mathcal{X}_g :\ f(X) < (t - 3\delta')\,n\}$, and let $X \in \mathcal{Y}_g$. Denote by ξ_X the product measure on $\mathcal{C}_n$ such that if $Y_X \sim \xi_X$ then $\mathbb{E}Y_X = X$. We then have

$$\begin{aligned}
\Pr\left[f\left(Y_X\right) \geq \left(t-2\delta'\right)n\right] &\leq \Pr\left[f\left(Y_X\right) \geq f\left(\mathbb{E}Y_X\right)+\delta' n\right] \\
&\leq \Pr\left[\left|f\left(Y_X\right)-f\left(\mathbb{E}Y_X\right)\right| \geq \delta' n\right] \\
\text{(by Markov's inequality)} &\leq \frac{\mathbb{E}\left|f\left(Y_X\right)-f\left(\mathbb{E}Y_X\right)\right|}{\delta' n} \\
\text{(by Proposition A.1)} &\leq \frac{L_1}{\delta'\sqrt{n}}. \qquad (6.3)
\end{aligned}$$

Denote by $\mathcal{A}_X$ the event

$$\mathcal{A}_X = \left\{f\left(Y_X\right) < \left(t-2\delta'\right)n\right\} \cap \left\{f\left(X_n^g\right) > \left(t-\delta'\right)n\right\}.$$

Equation (6.3) and Proposition 6.1 together imply that

$$\Pr\left[\mathcal{A}_X\right] \geq 1 - \frac{L_1}{\delta'\sqrt{n}} - 2^{-n}.$$

Under $\mathcal{A}_X$ we have that

$$\delta' n \leq f\left(X_n^g\right) - f\left(Y_X\right) \leq L_1 \left\|Y_X - X_n^g\right\|_1,$$

yielding

$$\left\|Y_X - X_n^g\right\|_1 \geq \frac{\delta' n}{L_1}. \qquad (6.4)$$

Since $\mathbb{E}\left\|X\left(\rho\right)-X_n^g\right\|_1$ is small, this inequality sets a constraint on the measure of $\mathcal{Y}_g$. Letting Z be a random variables with law ρ, coupled with $X\left(\rho\right)$ so that $X\left(\rho\right) \mid Z \sim Y_Z$, one has

$$\begin{aligned}
\mathbb{E}\left\|X\left(\rho\right)-X_n^g\right\|_1 &= \int_{\mathcal{C}_n} \mathbb{E}\left[\left\|Y_Z - X_n^g\right\|_1 \mid Z\right] d\rho\left(Z\right) \\
&\geq \int_{\mathcal{Y}_g} \mathbb{E}\left[\left\|Y_Z - X_n^g\right\|_1 \mid Z\right] d\rho\left(Z\right) \\
&\geq \int_{\mathcal{Y}_g} \mathbb{E}\left[\left\|Y_Z - X_n^g\right\|_1 \mid Z \wedge \mathcal{A}_Z\right] \Pr\left[\mathcal{A}_Z\right] d\rho\left(Z\right) \\
\text{(by equation (6.4))} &\geq \left(1-\frac{L_1}{\delta'\sqrt{n}}-2^{-n}\right)\int_{\mathcal{Y}_g} \frac{\delta' n}{L_1} d\rho\left(Z\right) \\
&= \left(1-\frac{L_1}{\delta'\sqrt{n}}-2^{-n}\right)\rho\left(\mathcal{Y}_g\right)\frac{\delta' n}{2L_1}.
\end{aligned}$$

We thus obtain

$$\rho\left(\mathcal{Y}_g\right) \leq \frac{80 L_1 \tilde{D}^{1/4}}{n^{1/4}\delta'}\left(1-\frac{L_1}{\delta'\sqrt{n}}-2^{-n}\right)^{-1}.$$

Together with equation (6.2), this gives

$$\rho\left(\mathcal{X}_g \backslash \mathcal{Y}_g\right) \geq 1 - \frac{161 L_1 \tilde{D}^{1/4}}{n^{1/4}\delta'}\left(1-\frac{L_1}{\delta'\sqrt{n}}-2^{-n}\right)^{-1}$$

as needed. □

Remark 6.3. A particular type of Hamiltonian that has been of considerable interest in the field of large deviations that of subgraph-counting functions. It was recently shown in [5] that for these types of Hamiltonians, $\nabla f\left(X\right)$ is close to a stochastic block matrix. Since $h'\left(\left(\frac{f(X)}{n}-t\right)/\delta\right)$ is a scalar, this implies that every $X \in \mathcal{X}_g$ is also close to a stochastic block matrix.

Remark 6.4. Theorem 3.8 corresponds to the unconditioned distribution μ_p with $p = 1/2$. To deal with the case $p \neq 1/2$, define $g(y)$ as

$$g(y) = \psi(f(y)) + \prod \log\left(\frac{1}{2}(1 - y_i + 2py_i)\right).$$

Analogues of Propositions 6.1 and 6.2 can then be proved following the same line.

A Appendix

Proof of Lemma 4.6. Denote $Y = \tanh Z$, and denote the bound of Y by $\alpha = \tanh L \geq \tanh(1)$. Under this notation, we wish to show that

$$\left|\tanh\left(\mathbb{E}\tanh^{-1} Y\right) - \mathbb{E}Y\right| \leq 20\tanh^{-1}(\alpha) \cdot \mathbb{E}|Y - \mathbb{E}Y|. \tag{A.1}$$

We will prove this inequality by considering it as a variational problem on the distribution μ of Y. Specifically, we will show that for every $a \in [-\alpha, \alpha]$, every distribution μ of Y satisfies

$$\left|\tanh\left(\mathbb{E}\tanh^{-1} Y\right) - a\right| \leq 20\tanh^{-1}(\alpha) \cdot \mathbb{E}|Y - a|. \tag{A.2}$$

Setting $a = \mathbb{E}Y$ then gives the desired result.

Suppose that $\mathbb{E}|Y - a|$ is fixed. Then the left hand side of (A.2) is maximized by the Y that gives $\tanh\left(\mathbb{E}\tanh^{-1} Y\right)$ an extremal value, conditioned on $b := \mathbb{E}|Y - a|$ being constant. Since $\tanh$ is monotone, this is equivalent to finding the extremal value of the integral

$$\int \tanh^{-1}(x)\, d\mu(x) \tag{A.3}$$

while maintaining the constraint

$$b = \mathbb{E}|Y - a|. \tag{A.4}$$

The constraint (A.4) is of the form $\int f(x)\, d\mu = b$, where $f(x) = |x - a|$. By Theorems 2.1 and 3.2 and Proposition 3.1 in [8], the extremal distributions which solve a system of n constraints of the form $\int f_i(x)\, d\mu = c_i$ are linear combinations of no more than $n + 1$ singletons, i.e delta distributions. We can therefore write the extremal μ as

$$\mu = p\delta(x) + (1 - p)\delta(y) \tag{A.5}$$

for some two real numbers $-\alpha \leq x, y \leq \alpha$ and $p \in [0, 1]$. Now, using the triangle inequality, we have that

$$\left|\tanh\left(\mathbb{E}\tanh^{-1} Y\right) - a\right| \leq \left|\tanh\left(\mathbb{E}\tanh^{-1} Y\right) - \mathbb{E}Y\right| + \mathbb{E}|Y - a|,$$

so it is in fact enough to show that

$$\left|\tanh\left(\mathbb{E}\tanh^{-1} Y\right) - \mathbb{E}Y\right| \leq 19\tanh^{-1}(\alpha) \cdot \mathbb{E}|Y - a|,$$

and since $\mathbb{E}|Y - \mathbb{E}Y| \leq 2\mathbb{E}|Y - a|$ for every a, it actually suffices to show that

$$\left|\tanh\left(\mathbb{E}\tanh^{-1} Y\right) - \mathbb{E}Y\right| \leq 9\tanh^{-1}(\alpha) \cdot \mathbb{E}|Y - \mathbb{E}Y|. \tag{A.6}$$

Plugging the decomposition (A.5) into (A.6), we need to prove that for every such x and y,

$$\frac{\left|\tanh\left(p\tanh^{-1}(x) + (1 - p)\tanh^{-1}(y)\right) - (px + (1 - p)y)\right|}{2p(1 - p)|x - y|\tanh^{-1}(\alpha)} \leq 9.$$

Assume without loss of generality that $x > 0$ and $x > |y|$. We will now show that inequality is correct for $0 < p \le \frac{1}{2}$. We omit the similar proof for $\frac{1}{2} \le p < 1$. For these values of p, it suffices to show that

$$\frac{\left|\tanh\left(p\tanh^{-1}(x) + (1-p)\tanh^{-1}(y)\right) - (px + (1-p)y)\right|}{p\tanh^{-1}(\alpha)(x-y)} \le 9. \tag{A.7}$$

For every fixed value of y, we treat the expression on the left hand side as a function of p for $p \in (0,1)$. This expression may attain its supremum either at $p \to 0^+$, $p = \frac{1}{2}$, or at values of p such that the derivative of the left hand side with respect to p is 0. We'll now consider each of these three cases.

Taking the derivative

Comparing the derivative to 0, one obtains the relation

$$\tanh\left(p\tanh^{-1}(x) + (1-p)\tanh^{-1}(y)\right) - (px + (1-p)y) = \\ \left(\frac{\tanh^{-1}(x) - \tanh^{-1}(y)}{\cosh^2\left(p\tanh^{-1}(x) + (1-p)\tanh^{-1}(y)\right)} - (x-y)\right)p.$$

Plugging this back into (A.7) and using the triangle inequality, it is enough to show that

$$\frac{\frac{\tanh^{-1}(x)-\tanh^{-1}(y)}{\cosh^2\left(p\tanh^{-1}(x)+(1-p)\tanh^{-1}(y)\right)} + (x-y)}{\tanh^{-1}(\alpha)(x-y)} \le 9. \tag{A.8}$$

Since $\tanh^{-1}(\alpha) \ge 1$, the expression $\frac{(x-y)}{\tanh^{-1}(\alpha)(x-y)}$ is bounded by 1, so it remains to show that

$$\frac{\tanh^{-1}(x) - \tanh^{-1}(y)}{\tanh^{-1}(\alpha)\cosh^2\left(p\tanh^{-1}(x) + (1-p)\tanh^{-1}(y)\right)(x-y)} \le 8. \tag{A.9}$$

If $y < 0$ and $x \ge \frac{1}{2}$, then $x - y > 1/2$ and we trivially have

$$\frac{\tanh^{-1}(x) - \tanh^{-1}(y)}{\tanh^{-1}(\alpha)} \frac{1}{\cosh^2\left(p\tanh^{-1}(x) + (1-p)\tanh^{-1}(y)\right)(x-y)} \le \frac{2}{\frac{1}{2}} = 4.$$

If $y < 0$ and $x < \frac{1}{2}$ then

$$\tanh^{-1}(x) - \tanh^{-1}(y) \le \frac{1}{1-x^2}(x-y) \le \frac{4}{3}(x-y)$$

and so

$$\frac{\tanh^{-1}(x) - \tanh^{-1}(y)}{\tanh^{-1}(\alpha)\cosh^2\left(p\tanh^{-1}(x) + (1-p)\tanh^{-1}(y)\right)(x-y)} \le \frac{2}{\tanh^{-1}(\alpha)} < 8.$$

For $y \ge 0$, the maximal w.r.t p value of the left hand side of (A.9) is attained when the argument of $\cosh^2$ is minimal, i.e at $p = 0$. Using the fact that $\cosh\left(\tanh^{-1}(y)\right) = 1/\sqrt{1-y^2} > 1/\sqrt{2(1-y)}$ and that $\tanh^{-1}(x) = \frac{1}{2}\log\frac{1+x}{1-x}$, it suffices to show that

$$\frac{\left(\log\frac{1+x}{1-x}\frac{1-y}{1+y}\right)(1-y)}{\tanh^{-1}(\alpha)(x-y)} \le 8. \tag{A.10}$$

We consider two cases. Suppose that $\frac{1-y}{1-x} \ge 2$. For any $z \ge 2$, it holds that $\log 2z \le 2\log z$, and since $x, y < 1$, it is enough to show that

$$2\frac{\left(\log\frac{1-y}{1-x}\right)(1-y)}{\tanh^{-1}(\alpha)(x-y)} \le 8. \tag{A.11}$$

Denote $1-x=e^{-a}$ and $1-y=e^{-b}$, with $a>b>0$; under this notation, the left hand side becomes $2\frac{(a-b)}{\tanh^{-1}(\alpha)\left(1-e^{-(a-b)}\right)}$. Note that $\tanh^{-1}(\alpha)=\frac{1}{2}\log\frac{1+\alpha}{1-\alpha}\geq\frac{1}{2}\log\frac{1}{1-x}=\frac{1}{2}a$. If $e^{-(a-b)}<\frac{1}{2}$, then $2\frac{(a-b)}{\tanh^{-1}(\alpha)\left(1-e^{-(a-b)}\right)}\leq 2\frac{(a-b)}{\frac{1}{2}a\left(\frac{1}{2}\right)}\leq 8$. Otherwise, if $e^{-(a-b)}\geq\frac{1}{2}$, then $a-b<\frac{3}{4}$. By Taylor's theorem, the $1-e^{-(a-b)}$ in the denominator can be bounded from below by $\frac{1}{2}(a-b)$, bounding the expression by $\frac{8}{\tanh^{-1}(\alpha)}\leq 8$.

Now suppose that $\frac{1-y}{1-x}<2$. Since $\log z\leq z-1$ for all z, we may then write the left hand side of (A.10) as

$$\begin{aligned}\frac{\left(\log\frac{1+x}{1+y}+\log\frac{1-y}{1-x}\right)(1-y)}{\tanh^{-1}(\alpha)(x-y)} &\leq \frac{1}{\tanh^{-1}(\alpha)}\left(\left(\frac{1+x}{1+y}-1\right)+\left(\frac{1-y}{1-x}-1\right)\right)\frac{1-y}{x-y}\\ &\leq \frac{1}{\tanh^{-1}(\alpha)}\left(\frac{1-y}{1+y}+\frac{1-y}{1-x}\right)\\ &\leq \frac{3}{\tanh^{-1}(\alpha)}<8.\end{aligned}$$

The case $p=0$

Using L'Hôpital's rule, the value of the left hand side of (A.7) attained as $p\to 0^+$ is

$$\frac{\left|\frac{\tanh^{-1}(x)-\tanh^{-1}(y)}{\cosh^2\left(\tanh^{-1}(y)\right)}-(x-y)\right|}{\tanh^{-1}(\alpha)(x-y)}.$$

For $y\geq 0$, this is the same expression obtained by setting $p=0$ in (A.8). The case $y<0$ is handled similarly as above.

The case $p=1/2$

In this case we must show that

$$\frac{\left|\tanh\left(\frac{1}{2}\tanh^{-1}(x)+\frac{1}{2}\tanh^{-1}(y)\right)-\left(\frac{1}{2}x+\frac{1}{2}y\right)\right|}{\tanh^{-1}(\alpha)(x-y)}\leq\frac{9}{2}.$$

This bound can be shown by differentiating with respect to y to the find the maximum of the left hand side. ☐

Proposition A.1. *Let $f:\mathcal{C}_n\to\mathbb{R}$, let ξ be a product measure over $\mathcal{C}_n$, and let $Y\sim\xi$. Then*

$$\mathbb{E}\left|f(Y)-f(\mathbb{E}Y)\right|\leq\sqrt{n}\mathit{Lip}(f).$$

Proof. Let $M_i=\mathbb{E}\left[f(Y)\mid Y_1,\ldots,Y_i\right]$. Then the variance of f can be bounded by

$$\mathrm{Var}\left[f(Y)\right]=\sum_{i=1}^{n}\mathbb{E}\left(M_i-M_{i-1}\right)^2\leq\mathrm{Lip}^2(f)\sum_{i=1}^{n}\mathrm{Var}\left[Y_i\right]\leq n\mathrm{Lip}^2(f).$$

By Jensen's inequality,

$$\mathbb{E}\left|f(Y)-f(\mathbb{E}Y)\right|=\mathbb{E}\sqrt{\left(f(Y)-f(\mathbb{E}Y)\right)^2}\leq\sqrt{\mathbb{E}\left(f(Y)-f(\mathbb{E}Y)\right)^2}=\sqrt{\mathrm{Var}\left[f(Y)\right]}.\quad ☐$$

Proof of the chain rule Lemma 5.1. For $y\in\mathcal{C}_n$ in the discrete hypercube, denote by $S_i(y)$ the vector which is equal to y everywhere, except for the i-th entry, so that

$$\left(S_i(y)\right)_j=\begin{cases}y_j & i\neq j\\ -y_j & i=j.\end{cases}$$

Using this notation, we have that

$$\left|\partial_i \left(h\circ f\right)(y) - h'\left(f\left(y\right)\right)\partial_i f\left(y\right)\right| = \left|-y_i \frac{h\left(f\left(S_i\left(y\right)\right)\right) - h\left(f\left(y\right)\right)}{2} - h'\left(f\left(y\right)\right)\partial_i f\left(y\right)\right|. \tag{A.12}$$

Using Taylor's theorem for h around $f\left(y\right)$ with the Lagrange remainder, there exists a $z \in \left[f\left(y\right), f\left(S_i\left(y\right)\right)\right]$ such that

$$h\left(f\left(S_i\left(y\right)\right)\right) - h\left(f\left(y\right)\right) = \left(f\left(S_i\left(y\right)\right) - f\left(y\right)\right)h'\left(f\left(y\right)\right) + \frac{1}{2}\left(f\left(S_i\left(y\right)\right) - f\left(y\right)\right)^2 h''\left(z\right).$$

Putting this into equation (A.12), we get

$$\begin{aligned}
&\left|\partial_i h\left(f\left(y\right)\right) - h'\left(f\left(y\right)\right)\partial_i f\left(y\right)\right| = \\
&= \left|-\frac{1}{2}y_i\left(\left(f\left(S_i\left(y\right)\right) - f\left(y\right)\right)h'\left(f\left(y\right)\right) + \frac{1}{2}\left(f\left(S_i\left(y\right)\right) - f\left(y\right)\right)^2 h''\left(z\right)\right) - \right. \\
&\qquad \left. - h'\left(f\left(y\right)\right)\partial_i f\left(y\right)\right| \\
&= \left|\partial_i f\left(y\right)h'\left(f\left(y\right)\right) - \frac{y_i}{4}\left(f\left(S_i\left(y\right)\right) - f\left(y\right)\right)^2 h''\left(z\right) - h'\left(f\left(y\right)\right)\partial_i f\left(y\right)\right| \\
&= \left|\frac{y_i}{4}\left(f\left(S_i\left(y\right)\right) - f\left(y\right)\right)^2 h''\left(z\right)\right| \\
&= \left|\partial_i f\left(y\right)\right|^2\left|h''\left(z\right)\right| \\
&\le BL^2.
\end{aligned}$$

Equations (5.1) and (5.2) then follow immediately.

For equation (5.3), let $x \in \overline{\mathcal{C}_n}$ and let ξ be the product measure on $\mathcal{C}_n$ such that for $Y \sim \xi$, $\mathbb{E}Y = x$. Applying equation (4.10) on ∇f and $\nabla\left(h\circ f\right)$, we have

$$\begin{aligned}
\left\|h'\left(f\left(\mathbb{E}Y\right)\right)\nabla f\left(\mathbb{E}Y\right) - \mathbb{E}\nabla\left(h\circ f\right)\left(Y\right)\right\|_1 &= \left\|h'\left(f\left(\mathbb{E}Y\right)\right)\mathbb{E}\nabla f\left(Y\right) - \mathbb{E}\nabla\left(h\circ f\right)\left(Y\right)\right\|_1 \\
&\le \left\|\mathbb{E}\left[h'\left(f\left(\mathbb{E}Y\right)\right)\nabla f\left(Y\right) - h'\left(f\left(Y\right)\right)\nabla f\left(Y\right)\right]\right\|_1 \\
&\quad + \left\|\mathbb{E}\left[h'\left(f\left(Y\right)\right)\nabla f\left(Y\right) - \nabla\left(h\circ f\right)\left(Y\right)\right]\right\|_1.
\end{aligned}$$

By equation (5.1), the second term on the right hand side is bounded by BL^2n. AS for the first term,

$$\begin{aligned}
\left\|\mathbb{E}\left[\left(h'\left(f\left(\mathbb{E}Y\right)\right) - h'\left(f\left(Y\right)\right)\right)\nabla f\left(Y\right)\right]\right\|_1 &\le B\mathbb{E}\left\|\left|f\left(\mathbb{E}Y\right) - f\left(Y\right)\right|\nabla f\left(Y\right)\right\|_1 \\
&\le B\mathbb{E}\left|f\left(\mathbb{E}Y\right) - f\left(Y\right)\right|nL \\
(\text{by Proposition A.1}) &\le BL^2n^{3/2}.
\end{aligned}$$

Thus $\left\|h'\left(f\left(\mathbb{E}Y\right)\right)\nabla f\left(\mathbb{E}Y\right) - \nabla\left(h\circ f\right)\left(\mathbb{E}Y\right)\right\|_1 \le 2BL^2n^{3/2}$. □

Proof of Lemma 5.2. For a vector $y \in \mathcal{C}_n$ and an index $i = 1, \ldots, n$, denote by y_i^+ the vector $y_i^+ = \left(y_1, y_2, \ldots, y_{i-1}, 1, y_{i+1}, \ldots, y_n\right)$, and by y_i^- the vector $y_i^- = \left(y_1, y_2, \ldots, y_{i-1}, -1, y_{i+1}, \ldots, y_n\right)$.

- **The gradient complexity**: Denote

  $$\mathcal{A}_f = \left\{\nabla f\left(y\right) : y \in \mathcal{C}_n\right\}, \quad \mathcal{A}_h = \left\{\nabla\left(h\circ f\right)\left(y\right) : y \in \mathcal{C}_n\right\}.$$

  By equation (5.2), we have that for every vector $v \in \mathbb{R}^n$,

  $$\sup_{u\in\mathcal{A}_h}\langle u, v\rangle \le \max\left(0, B_1\sup_{u\in\mathcal{A}_f}\langle u, v\rangle\right) + \sqrt{n}B_2L_1^2\left\|v\right\|_2.$$

  Since the expected norm of a Gaussian random vector Γ satisfies $\mathbb{E}\left\|\Gamma\right\|_2 \le \sqrt{n}$, we get that

  $$\mathcal{D}\left(h\circ f\right) = \mathbb{E}\sup_{u\in\mathcal{A}_h}\langle u, \Gamma\rangle \le B_1\mathcal{D}\left(f\right) + B_2L_1^2n.$$

- **The Lipschitz constant**: for every $y \in \mathcal{C}_n$ and every $i = 1, \dots, n$,

$$\begin{aligned} \left|\partial_i \left(h \circ f\right)(y)\right| &= \left|\frac{h\left(f\left(y_i^+\right)\right) - h\left(f\left(y_i^-\right)\right)}{2}\right| \\ &\leq B_1 \left|\frac{f\left(y_i^+\right) - f\left(y_i^-\right)}{2}\right| \leq B_1 L_1. \end{aligned}$$

  Thus $\mathrm{Lip}\left(h \circ f\right) \leq B_1 L_1$.
- **The Lipschitz constant of the gradient**: Let $x \neq y \in \mathcal{C}_n$. By Lemma 5.1:

$$\begin{aligned} &\left\|\nabla\left(h \circ f\right)(x) - \nabla\left(h \circ f\right)(y)\right\|_1 \\ &\quad = \left\|\nabla\left(h \circ f\right)(x) - h'\left(f(x)\right)\nabla f(x) + h'\left(f(x)\right)\nabla f(x)\right. \\ &\qquad \left. -\nabla\left(h \circ f\right)(y) - h'\left(f(y)\right)\nabla f(y) + h'\left(f(y)\right)\nabla f(y)\right\|_1 \\ &\quad \leq 2nB_2L_1^2 + \left\|h'\left(f(x)\right)\nabla f(x) - h'\left(f(y)\right)\nabla f(y)\right\|_1. \end{aligned}$$

  The last term on the right hand side can be bounded by

$$\begin{aligned} &\left\|h'\left(f(x)\right)\nabla f(x) - h'\left(f(y)\right)\nabla f(y)\right\| \\ &\quad \leq \left\|h'\left(f(x)\right)\nabla f(x) - h'\left(f(x)\right)\nabla f(y)\right\|_1 \\ &\qquad + \left\|h'\left(f(x)\right)\nabla f(y) - h'\left(f(y)\right)\nabla f(y)\right\|_1 \\ &\quad \leq B_1\left\|\nabla f(x) - \nabla f(y)\right\|_1 + B_2\left|f(x) - f(y)\right|\left\|\nabla f(y)\right\|_1 \\ &\quad \leq B_1L_2\left\|x-y\right\|_1 + B_2L_1\left\|x-y\right\|_1 L_1 n. \end{aligned}$$

  Putting the terms together, we get

$$\frac{\left\|\nabla\left(h \circ f\right)(x) - \nabla\left(h \circ f\right)(y)\right\|_1}{\left\|x-y\right\|_1} \leq B_1L_2 + 3B_2L_1^2 n. \qquad \square$$

Proof of Proposition 6.2. We will show that the total variation distance between X_n^g and X_φ satisfies

$$\mathrm{TV}\left(\nu, \sigma\right) \leq 2 \cdot 2^{-n};$$

the proof of the proposition then follows immediately. Denote by Z_g and Z_φ the normalizing constants of ν and σ, respectively. Then

$$Z_g = Z_\varphi + \sum_{y\ s.t\ f(y) \leq (t-\delta')n} e^{g(y)},$$

and by the proof of Proposition 6.1, this implies that

$$\varepsilon := \left|Z_g - Z_\varphi\right| \leq 2^{-n}.$$

The total variation distance is then given by

$$\begin{aligned} \mathrm{TV}\left(\nu,\sigma\right) &= \frac{1}{2}\sum_{y \in \mathcal{C}_n}\left|\frac{e^{g(y)}}{Z_g} - \frac{\varphi(y)}{Z_\varphi}\right| \\ &= \frac{1}{2}\sum_{f(y)<(t-\delta')n}\left|\frac{e^{g(y)}}{Z_g} - \frac{\varphi(y)}{Z_\varphi}\right| + \frac{1}{2}\sum_{f(y)\geq(t-\delta')n}\left|\frac{e^{g(y)}}{Z_g} - \frac{\varphi(y)}{Z_\varphi}\right|. \end{aligned}$$

By definition of φ and by Proposition 6.1, the first term on the right hand side is bounded by

$$\frac{1}{2}\sum_{f(y)<(t-\delta')n}\left|\frac{e^{g(y)}}{Z_g} - \frac{\varphi(y)}{Z_\varphi}\right| = \frac{1}{2}\sum_{f(y)<(t-\delta')n}\frac{e^{g(y)}}{Z_g} \leq \frac{1}{2}\cdot 2^{-n}$$

The second term is bounded by

$$\begin{aligned}
\frac{1}{2}\sum_{f(y)\geq(t-\delta')n}\left|\frac{e^{g(y)}}{Z_g}-\frac{\varphi\left(y\right)}{Z_\varphi}\right| &= \frac{1}{2}\sum\varphi\left(y\right)\left|\frac{1}{Z_\varphi+\varepsilon}-\frac{1}{Z_\varphi}\right| \\
&= \frac{1}{2}\sum\frac{\varphi\left(y\right)}{Z_\varphi}\left|\frac{1}{\left(1+\frac{\varepsilon}{Z_\varphi}\right)}-1\right| \\
&\leq \frac{1}{2}\sum\frac{\varphi\left(y\right)}{Z_\varphi}\left|\frac{\varepsilon}{Z_\varphi}+\frac{1}{2}\frac{\varepsilon^2}{Z_\varphi^2}\right| \\
&\leq \frac{1}{2}\sum\frac{\varphi\left(y\right)}{Z_\varphi}\left|\frac{2\varepsilon}{Z_\varphi}\right|\leq 2^{-n}. \qquad \square
\end{aligned}$$

References

[1] Anirban Basak and Sumit Mukherjee, *Universality of the mean-field for the potts model*, Probability Theory and Related Fields **168** (2017), no. 3, 557–600. MR-3663625

[2] Sourav Chatterjee and Amir Dembo, *Nonlinear large deviations*, arXiv:1401.3495 (2014). MR-3519474

[3] Sourav Chatterjee and Persi Diaconis, *Estimating and understanding exponential random graph models*, Ann. Statist. **41** (2013), no. 5, 2428–2461. MR-3127871

[4] Ronen Eldan, *Gaussian-width gradient complexity, reverse log-Sobolev inequalities and nonlinear large deviations*, arXiv:1612.04346 (2016).

[5] Ronen Eldan and Renan Gross, *Exponential random graphs behave like mixtures of stochastic block models*, arXiv:1707.01227 (2017).

[6] Eyal Lubetzky and Yufei Zhao, *On the variational problem for upper tails in sparse random graphs*, Random Structures Algorithms **50** (2017), no. 3, 420–436. MR-3632418

[7] Ryan O'Donnell, *Analysis of boolean functions*, Cambridge University Press, New York, NY, USA, 2014. MR-3443800

[8] Gerhard Winkler, *Extreme points of moment sets*, Mathematics of Operations Research **13** (1988), no. 4, 581–587. MR-0971911

Acknowledgments. The first author is grateful to Sourav Chatterjee for inspiring him to work on this topic and for an enlightening discussion. We thank Amir Dembo and Sumit Mukherjee for insightful discussions, and Yufei Zhao for his motivating comments on sparse bounds. Finally we thank the anonymous referee for comments improving the presentation of this work.