The Annals of Probability
2019, Vol. 47, No. 6, 4002–4023
https://doi.org/10.1214/19-AOP1352

THE STRUCTURE OF LOW-COMPLEXITY GIBBS MEASURES ON PRODUCT SPACES

BY TIM AUSTIN

University of California, Los Angeles

Let $K_1, \dots, K_n$ be bounded, complete, separable metric spaces. Let λ_i be a Borel probability measure on K_i for each i. Let $f : \prod_i K_i \longrightarrow \mathbb{R}$ be a bounded and continuous potential function, and let

$$\mu(\mathrm{d}\boldsymbol{x}) \propto \mathrm{e}^{f(\boldsymbol{x})}\lambda_1(\mathrm{d}x_1)\cdots\lambda_n(\mathrm{d}x_n)$$

be the associated Gibbs distribution.

At each point $\boldsymbol{x} \in \prod_i K_i$, one can define a 'discrete gradient' $\nabla f(\boldsymbol{x}, \cdot)$ by comparing the values of f at all points which differ from $\boldsymbol{x}$ in at most one coordinate. In case $\prod_i K_i = \{-1, 1\}^n \subset \mathbb{R}^n$, the discrete gradient $\nabla f(\boldsymbol{x}, \cdot)$ is naturally identified with a vector in $\mathbb{R}^n$.

This paper shows that a 'low-complexity' assumption on ∇f implies that μ can be approximated by a mixture of other measures, relatively few in number, and most of them close to product measures in the sense of optimal transport. This implies also an approximation to the partition function of f in terms of product measures, along the lines of Chatterjee and Dembo's theory of 'nonlinear large deviations'.

An important precedent for this work is a result of Eldan in the case $\prod_i K_i = \{-1, 1\}^n$. Eldan's assumption is that the discrete gradients $\nabla f(\boldsymbol{x}, \cdot)$ all lie in a subset of $\mathbb{R}^n$ that has small Gaussian width. His proof is based on the careful construction of a diffusion in $\mathbb{R}^n$ which starts at the origin and ends with the desired distribution on the subset $\{-1, 1\}^n$. Here our assumption is a more naive covering-number bound on the set of gradients $\{\nabla f(\boldsymbol{x}, \cdot) : \boldsymbol{x} \in \prod_i K_i\}$, and our proof relies only on basic inequalities of information theory. As a result, it is shorter, and applies to Gibbs measures on arbitrary product spaces.

1. Introduction. Let $(K_1, d_{K_1}), \dots, (K_n, d_{K_n})$ be nonempty, complete, separable metric spaces, all with diameter at most one. Let $C_{\mathrm{b}}(K_i)$ and $\mathrm{Prob}(K_i)$ denote the spaces of bounded continuous functions and Borel probability measures on K_i, respectively, and similarly for other topological spaces. Let $\|\cdot\|$ denote the uniform norm on any space of real-valued functions. Let $\lambda_i \in \mathrm{Prob}(K_i)$ be a fixed reference measure for each i.

Received December 2018; revised January 2019.
MSC2010 subject classifications. Primary 60B99; secondary 60G99, 82B20, 94A17.
Key words and phrases. Nonlinear large deviations, Gibbs measures, gradient complexity, dual total correlation, mixtures of product measures.

Let $f : \prod_{i=1}^{n} K_i \longrightarrow \mathbb{R}$ be a bounded continuous function, and call it the 'potential'. Let

$$\mu(\mathrm{d}\boldsymbol{x}) := \frac{1}{Z}\mathrm{e}^{f(\boldsymbol{x})} \prod_{i=1}^{n} \lambda_i(\mathrm{d}x_i) \tag{1}$$

be the resulting Gibbs measure, where Z is the normalizing constant that makes μ a probability measure. In the language of statistical mechanics, Z is the 'partition function' of f.

Inside $C_{\mathrm{b}}(\prod_i K_i)$ lies the vector subspace of functions that have the form

$$\boldsymbol{x} \mapsto f_1(x_1) + \cdots + f_n(x_n) \tag{2}$$

for some $f_1 \in C_{\mathrm{b}}(K_1), \ldots, f_n \in C_{\mathrm{b}}(K_n)$. We call such functions *additively separable*. The choice of $f_1, \ldots, f_n$ representing the function in (2) is unique up additive constants. If f is the function in (2), then its Gibbs measure factorizes as

$$\frac{1}{Z}\prod_{i=1}^{n}\bigl(\mathrm{e}^{f_i(x_i)}\lambda_i(\mathrm{d}x_i)\bigr),$$

so it is a product measure. In this special case we denote this Gibbs measure by ξ_f. Similarly, if $g \in C_{\mathrm{b}}(K_i)$ for some i, then we denote by ξ_g the Gibbs measure on K_i constructed from g and λ_i.

Consider again a general $f \in C_{\mathrm{b}}(\prod_i K_i)$. In the main result of this paper, we assume that f has 'low complexity' relative to the subspace of additively separable functions, and deduce a structure theorem for μ in terms of mixtures of product measures. In case $|K_i| = 2$ for each i, theorems of this kind originate in Chatterjee and Dembo's work [7] introducing the theory of 'nonlinear large deviations'. Chatterjee and Dembo's main result gives an approximation to the partition function Z under such an assumption on f. More recently, works by Eldan [14] and Eldan and Gross [16] have uncovered the approximate structure of the measure μ itself, again in case $|K_i| = 2$. Our main theorem fits a similar template to Eldan's, but it applies to general product spaces and its proof is quite different.

To explain the relevant notion of 'low complexity', fix a reference element $*_i \in K_i$ for each i. Given $\boldsymbol{x} \in \prod_i K_i$, $i \in [n]$, and $y \in K_i$, we define

$$\partial_i f(\boldsymbol{x}, y) := f(x_1, \ldots, x_{i-1}, y, x_{i+1}, \ldots, x_n) - f(x_1, \ldots, x_{i-1}, *_i, x_{i+1}, \ldots, x_n).$$

Thus, we replace the ith coordinate of $\boldsymbol{x}$ twice, first with y and then with $*_i$, and take the difference of the resulting values of f. This should be regarded as a discrete analog of the partial derivative of f in the ith coordinate. Beware that the definition of $\partial_i f$ depends on the choice of the reference point $*_i$. We suppress this dependence in our notation, but return to this point after the statement of the main theorem below.

We assemble these new functions $\partial_i f$ into the single function

$$\nabla f(\boldsymbol{x}, \boldsymbol{y}) := \sum_{i=1}^{n} \partial_i f(\boldsymbol{x}, y_i) \quad \Big(\boldsymbol{x}, \boldsymbol{y} \in \prod_i K_i\Big).$$

For a fixed point $\boldsymbol{x} \in \prod_i K_i$, the function $\nabla f(\boldsymbol{x}, \cdot)$ is additively separable. We refer to it as the *discrete gradient* of f at the point $\boldsymbol{x}$.

If $f \in C_{\mathrm{b}}(\prod_i K_i)$ has the property that $\nabla f(\boldsymbol{x}, \cdot)$ is the same function for every $\boldsymbol{x} \in \prod_i K_i$, then an easy exercise shows that f itself is additively separable. Beyond this case, we can make the weaker assumption that f has relatively few different gradients $\nabla f(\boldsymbol{x}, \cdot)$ as $\boldsymbol{x}$ varies in $\prod_i K_i$, at least up to small errors. This is the notion of 'low complexity' that we need.

The statement of our main theorem involves two other concepts. The first is a mode of approximation between measures on a product space, provided by an appropriate transportation metric. Endow the product space $\prod_i K_i$ with the normalized Hamming average of the metrics d_{K_i}:

$$d_n(\boldsymbol{x}, \boldsymbol{y}) := \frac{1}{n} \sum_{i=1}^{n} d_{K_i}(x_i, y_i) \quad \Big(\boldsymbol{x}, \boldsymbol{y} \in \prod_i K_i\Big).$$

If each K_i is just a finite alphabet with the discrete metric, then d_n is the usual Hamming metric, normalized to have diameter 1. Now define the transportation metric over d_n by

$$\overline{d_n}(\mu, \nu) := \inf_{\lambda} \int d_n(\boldsymbol{x}, \boldsymbol{y}) \lambda(\mathrm{d}\boldsymbol{x}, \mathrm{d}\boldsymbol{y}) \quad \Big(\mu, \nu \in \mathrm{Prob}\Big(\prod_i K_i\Big)\Big), \tag{3}$$

where λ ranges over all couplings of μ and ν. This is a standard and well-studied construction of a metric on probability measures: see, for instance, [13], Section 11.8.

The second concept we need is a quantity which measures the multi-variate correlation in a joint distribution on an n-fold product space. There are many such quantities, but the one we need is called 'dual total correlation' ('DTC'), which goes back to work of Han in information theory [17]. The definition of DTC is recalled in Section 2.4 below. The recent paper [2] studies DTC in some depth. According to the main result of that paper, a small value of DTC implies that the joint distribution is close in $\overline{d_n}$ to a mixture of relatively few product measures. The relevance of DTC to the present paper is therefore not surprising, but in fact we do not call on that result from [2] in our work below. Rather, the proof of the main theorem in the present paper is relatively self-contained, and happens to yield an upper bound on the DTC of the Gibbs measure (1) as a by-product.

Our main theorem decomposes μ as a mixture in a very concrete way: by partitioning $\prod_i K_i$ and conditioning on the parts. Given any probability space (K, μ) and measurable subset $P \subseteq K$ of positive measure, we write $\mu_{|P}$ for the conditioned measure $\mu(\cdot \mid P)$.

THEOREM A. *Let $\mathcal{P}$ be a partition of $\prod_i K_i$ with the property that*

$$\|\nabla f(\boldsymbol{x},\cdot) - \nabla f(\boldsymbol{y},\cdot)\| = \sup_z |\nabla f(\boldsymbol{x},z) - \nabla f(\boldsymbol{y},z)| < \delta n \tag{4}$$

whenever $\boldsymbol{x}$, $\boldsymbol{y}$ lie in the same part of $\mathcal{P}$. Then:

(a) $\mathrm{DTC}(\mu) < \mathrm{H}_\mu(\mathcal{P}) + \delta n$, *where $\mathrm{H}_\mu(\mathcal{P})$ is the Shannon entropy of the partition $\mathcal{P}$ according to μ*,

(b) *we have*

$$\sum_{P\in\mathcal{P}} \mu(P) \cdot \int \mathrm{D}(\mu_{|P} \parallel \xi_{\nabla f(\boldsymbol{y},\cdot)})\mu_{|P}(\mathrm{d}\boldsymbol{y}) < \mathrm{H}_\mu(\mathcal{P}) + \delta n,$$

where D *denotes Kullback–Leibler divergence, and*

(c) *we have*

$$\sum_{P\in\mathcal{P}} \mu(P) \cdot \int \overline{d_n}(\mu_{|P}, \xi_{\nabla f(\boldsymbol{y},\cdot)})\mu_{|P}(\mathrm{d}\boldsymbol{y}) < \sqrt{\frac{1}{2}\left(\frac{\mathrm{H}_\mu(\mathcal{P})}{n} + \delta\right)}.$$

In this theorem, part (c) follows directly from part (b) using Marton's transportation-entropy inequality (Proposition 2.3 below) and Hölder's inequality. We include both of these parts in the statement because both kinds of approximation by product measures have an intrinsic interest and potential applications. But most of the proof of Theorem A goes towards parts (a) and (b).

The definition of ∇f depends on the choice of the reference points $*_i$. However, if we write $*_1', \ldots, *_n'$ for an alternative choice of reference points and write $\nabla' f$ for the alternative discrete gradients that result, then these discrete gradients are related by the cocycle equation

$$\nabla' f(\boldsymbol{x},\boldsymbol{y}) = \nabla f(\boldsymbol{x},\boldsymbol{y}) - \nabla f\big(\boldsymbol{x},(*_1',\ldots,*_n')\big).$$

As a result, if $\mathcal{P}$ satisfies (4) for the discrete gradients $\nabla f(\boldsymbol{x},\cdot)$, then it satisfies the analogous bounds for the discrete gradients $\nabla' f(\boldsymbol{x},\cdot)$ with δn loosened to $2\delta n$. For this reason, the particular choice of the reference points $*_i$ has little effect on Theorem A.

Theorem A is valuable in case $\mathrm{H}_\mu(\mathcal{P})$ is small compared to n. This is implied if we have enough control on $|\mathcal{P}|$ itself: for instance, if we know that

$$\mathrm{cov}_{\delta n}\left(\left\{\nabla f(\boldsymbol{x},\cdot) : \boldsymbol{x} \in \prod_i K_i\right\}, \|\cdot\|\right) \le \mathrm{e}^{\varepsilon n} \tag{5}$$

for some small ε, where $\mathrm{cov}_{\delta n}(\cdot,\|\cdot\|)$ denotes the smallest cardinality of a covering by sets of $\|\cdot\|$-diameter less than δn.

COROLLARY A′. *If f satisfies* (5), *then there are* (i) *a partition $\mathcal{P}$ of $\prod_i K_i$ into at most $\mathrm{e}^{\varepsilon n}$ parts and* (ii) *a selection of additively separable functions g_P for $P \in \mathcal{P}$ such that*

$$\sum_{P\in\mathcal{P}} \mu(P) \cdot \mathrm{D}(\mu_{|P} \parallel \xi_{g_P}) < (\varepsilon + \delta)n$$

and

$$\sum_{P\in\mathcal{P}} \mu(P)\cdot \overline{d_n}(\mu_{|P},\xi_{g_P}) < \sqrt{(\varepsilon+\delta)/2}.$$

(*Informally*: *'most of the mass of* μ *is on conditioned measures* $\mu_{|P}$, $P\in\mathcal{P}$, *that are close to products'*.)

The argument from Theorem A to Corollary A′ is very short, but we include it explicitly after the proof of Theorem A in Section 3.

The proof of Theorem A brings together four basic results of information theory. All are well known up to some routine manipulations, but Section 2 lays them out carefully. Then Section 3 completes the proof of Theorem A.

Subsequent sections explore some consequences and related results.

First, Section 4 gives an approximate description of *which* product measures appear in the mixture promised by Corollary A′: see Proposition 4.1. The description takes the form of an approximate fixed point equation for those measures in terms of f. It follows almost immediately from Theorem A itself. It is an analog of the main result in [16] for the case of the Hamming cube, although that paper provides various finer details and applications that we do not pursue here.

Next, Section 5 turns Theorem A into an approximation for the normalizing constant Z in (1). This topic is the original concern of [7], and reappears among the consequences of Eldan's main result in [14].

Some further comparison with previous works. The simplest setting for Theorem A is a product of two-point alphabets. The works of Chatterjee and Dembo [7] and Eldan and Gross [14, 16] are confined to that case.

In [7], the authors assume that f is defined on the whole of $\mathbb{R}^n$, and then restrict it to the product subset $\{0,1\}^n$. Then they quantify the 'complexity' of f in terms of its classical gradient in the sense of calculus, rather than the discrete quantity ∇f that we discuss above. Their main result is an estimate on the normalizing constant Z, roughly in terms of a variational formula over product measures. This work has now been generalized to other alphabets by Yan [28]. For Yan's generalization, he retains the feature that the alphabets K_i are subsets of some given Banach spaces V_i, and that the complexity of f is measured by its Fréchet derivative as a function on $\prod_i V_i$. It should be straightforward to move between this notion of 'low complexity' and ours, but we do not explore this further here.

In [14], Eldan also regards his state space as a subset of $\mathbb{R}^n$. He uses the subset $\{-1,1\}^n$, which is more convenient for his proofs. Like the assumption in Theorem A, Eldan's 'low-complexity' assumption applies directly to the discrete gradient ∇f, not its relative from calculus. But his approach uses the embedding $\{-1,1\}^n\subset\mathbb{R}^n$ in another very essential way. It relies on a diffusion process in $\mathbb{R}^n$ which starts at the origin and ends with the desired distribution on $\{-1,1\}^n$. This approach seems quite different from ours. One can regard the main contribution of

the present paper as an alternative proof of something like Theorem A which (i) is simpler in the special case $K_i = \{-1, 1\}$ and (ii) generalizes easily to other product spaces. However, at some points Eldan's estimates seem to be sharper than ours; we compare them more carefully in Section 3.3. In a subsequent work [16], Eldan and Gross have characterized which product measures appear in Eldan's structure theorem in terms of the function f; Proposition 4.1 below is similar to their result.

Chatterjee and Dembo's results in [7] were initially motivated by statistical physics or large deviations questions in certain highly symmetric models of random graphs. Chatterjee gives a careful introduction to this area in his monograph [6], where Chapter 8 is given to the 'nonlinear large deviations approach'. After applying the basic theory of nonlinear large deviations, these questions about random graphs lead to a class of highly nontrivial variational problems, which are successfully analyzed in [4, 20]. Some related applications which require more than two-point alphabets are described by Yan in [28]. Applications with two-point alphabets are treated again by Eldan and Gross in [14–16], where they obtain refinements of some of the Chatterjee–Dembo estimates using Eldan's new structural results. Another application discussed in [7] is to large deviations of the number of arithmetic progressions in a random arithmetic set, and the recent paper [5] gives a more complete analysis of this problem using Eldan's approach.

In this paper, we do not investigate whether our new results lead to any further improvements in those applications. In the case of large deviations of subgraph counts in Erdős–Rényi random graphs, the recent preprint [8] achieves substantial improvements over the other works cited above by carefully finding and exploiting certain convex sets related to that problem.

Another variation on Chatterjee and Dembo's original partition-function estimates appears in Augeri's recent preprint [1]. She considers an arbitrary compactly supported probability measure λ on $\mathbb{R}^n$ and a continuously differentiable potential function $f : \mathbb{R}^n \longrightarrow \mathbb{R}$. By an elegant application of elementary convex analysis, she obtains a new upper bound on the partition function $\log \int \mathrm{e}^f \, \mathrm{d}\lambda$ in terms of a covering-number estimate on the range of the derivative Df. From this upper bound, she then derives improved estimates in several of the applications of nonlinear large deviations, including again large deviations for cycle counts in Erdős–Rényi random graphs.

2. The four principles in the proof of Theorem A. Our proof of Theorem A combines four basic principles from probability and information theory. This section recalls these in turn, and the next section assembles them into the proof.

2.1. *A modified chain rule for Kullback–Leibler divergence.* We assume familiarity with the basic properties of Shannon entropy and Kullback–Leibler divergence. A standard reference in the setting of discrete probability distributions is [9], Chapter 2. The KL divergence $\mathrm{D}(\mu \parallel \gamma)$ can be defined for any pair of probability measures μ, γ on a general measurable space: it is $+\infty$ unless $\mu \ll \gamma$, and

in that case it is defined by

$$\mathrm{D}(\mu \parallel \gamma) := \int \log \frac{\mathrm{d}\mu}{\mathrm{d}\gamma}\,\mathrm{d}\mu$$

(which may still equal $+\infty$). This generalization and its properties can be found in [24], Chapters 2 and 3, or [12], Appendix D.3.

For KL divergence, we need the following modification of the usual chain rule.

LEMMA 2.1 (Modified chain rule). *For any measurable space K, finite measurable partition $\mathcal{P}$ of K, and probability measures μ and γ on K, we have*

$$\mathrm{D}(\mu \parallel \gamma) = -\mathrm{H}_\mu(\mathcal{P}) + \sum_{P\in\mathcal{P}} \mu(P)\mathrm{D}(\mu_{|P} \parallel \gamma). \tag{6}$$

PROOF. The usual chain rule for KL divergence gives

$$\mathrm{D}(\mu \parallel \gamma) = \mathrm{D}\big([\mu]_{\mathcal{P}} \parallel [\gamma]_{\mathcal{P}}\big) + \sum_{P} \mu(P)\mathrm{D}(\mu_{|P} \parallel \gamma_{|P}), \tag{7}$$

where $[\mu]_{\mathcal{P}}$ denotes the stochastic vector $(\mu(P))_{P\in\mathcal{P}}$. Now observe that

$$\begin{aligned}\mathrm{D}(\mu_{|P} \parallel \gamma_{|P}) &= \int_P \log \frac{\mathrm{d}\mu_{|P}}{\mathrm{d}\gamma_{|P}}\,\mathrm{d}\mu_{|P} \\ &= \int_P \log\left(\gamma(P)\frac{\mathrm{d}\mu_{|P}}{\mathrm{d}\gamma}\right)\mathrm{d}\mu_{|P} = \mathrm{D}(\mu_{|P} \parallel \gamma) + \log\gamma(P).\end{aligned}$$

Inserting this into (7), we are left with the second right-hand term of (6), together with the quantity

$$\begin{aligned}\mathrm{D}\big([\mu]_{\mathcal{P}} \parallel [\gamma]_{\mathcal{P}}\big) + \sum_P \mu(P)\log\gamma(P) &= \sum_P \mu(P)\log\frac{\mu(P)}{\gamma(P)} + \sum_P \mu(P)\log\gamma(P) \\ &= -\mathrm{H}_\mu(\mathcal{P}).\end{aligned}$$

□

2.2. *Gibbs' variational principle.* Fix a probability space (K,λ) and let $f : K \longrightarrow \mathbb{R}$ be a bounded measurable function. As in the Introduction, the *Gibbs measure* on K associated to f and λ is defined by

$$\mu(\mathrm{d}x) := \frac{\mathrm{e}^{f(x)}}{\int \mathrm{e}^f\,\mathrm{d}\lambda}\lambda(\mathrm{d}x). \tag{8}$$

The importance of Gibbs measures is intimately related to Gibbs' famous variational principle. Here, we use it in the following form.

PROPOSITION 2.2. *If μ is the Gibbs measure associated to f, then*

$$\mathrm{D}(\mu \parallel \lambda) - \int f\,\mathrm{d}\mu = -\log\int \mathrm{e}^f\,\mathrm{d}\lambda.$$

For any other probability measure ν on K, we have

$$\left[\mathrm{D}(\nu \parallel \lambda) - \int f \,\mathrm{d}\nu\right] = \left[\mathrm{D}(\mu \parallel \lambda) - \int f \,\mathrm{d}\mu\right] + \mathrm{D}(\nu \parallel \mu). \tag{9}$$

If K is finite and λ is uniform, then (9) appears within the calculation in [9], equations (12.5)–(12.12). The same calculation gives the general case upon replacing $\mathrm{H}(\cdot)$ with $-\mathrm{D}(\cdot \parallel \lambda)$ throughout. In that generality it can be found in the proof of [12], Lemma 6.2.13, and in the application of [10], equation (2.6), in Section 3, case (A) of that paper.

2.3. *Marton's transportation-entropy inequality.* A classical inequality of Marton [21, 22] provides the basic link between KL divergence relative to a product measure and the transportation metric (3). Most of the proof of Theorem A concerns information theoretic estimates, before Marton's inequality turns these into a transportation-distance bound at the last step.

PROPOSITION 2.3 (Marton's transportation-entropy inequality). *Let (K_i, d_{K_i}), $i = 1, 2, \ldots, n$, be complete and separable metric spaces of diameter at most 1. If $\mu, \nu \in \mathrm{Prob}(\prod_i K_i)$ and ν is a product measure then*

$$\overline{d_n}(\mu, \nu) \le \sqrt{\frac{1}{2n}\mathrm{D}(\mu \parallel \nu)}.$$

Marton's original presentation is only for measures on finite sets, but the proof works without change for products of general complete separable metric spaces. The only requirement is that they all have diameter at most one: this is why we assume that from the beginning of this paper. (If the diameters are larger but finite, this simply changes the factor of $1/2$ in the inequality.) See [11, 22, 23, 25] for various extensions of Proposition 2.3, and [27] for a Gaussian analog.

2.4. *DTC and a version of Han's inequality.* Given finite-valued random variables $\xi_1, \ldots, \xi_n$, their *dual total correlation* or *DTC* is

$$\mathrm{DTC}(\xi_1, \ldots, \xi_n) := \mathrm{H}(\xi_1, \ldots, \xi_n) - \sum_{i=1}^{n} \mathrm{H}(\xi_i \mid \xi_1, \ldots, \xi_{i-1}, \xi_{i+1}, \ldots, \xi_n).$$

For more general random variables valued in measurable spaces $K_1, \ldots, K_n$, their DTC is

$$\mathrm{DTC}(\xi_1, \ldots, \xi_n) := \sup_{\mathcal{P}_1, \ldots, \mathcal{P}_n} \mathrm{DTC}\big([\xi_1]_{\mathcal{P}_1}, \ldots, [\xi_n]_{\mathcal{P}_n}\big),$$

where the supremum runs over all tuples of finite measurable partitions $\mathcal{P}_i$ of the spaces K_i, and $[\xi_i]_{\mathcal{P}_i}$ denotes the quantization of ξ_i by $\mathcal{P}_i$.

DTC is one of several possible ways to quantify the correlation among n random variables. A classical family of inequalities due to Han [18], Theorem 4.1, includes the fact that DTC is always nonnegative. DTC is studied carefully in [2], together with its simpler relative TC. In this paper we use DTC via an alternative formula in the setting of Gibbs measures.

Consider the spaces K_i and reference measures $\lambda_i \in \mathrm{Prob}(K_i)$ as in the Introduction. The next lemma is [2], Proposition 6.1, part (b), along with the first remark following its proof in that paper.

LEMMA 2.4. *If* $\mathrm{D}(\mu \parallel \lambda_1 \times \cdots \times \lambda_n) < \infty$ *then*

$$\mathrm{DTC}(\mu) = \sum_{i=1}^{n} \int \mathrm{D}(\mu_{i,z} \parallel \lambda_i)\mu_{[n]\setminus i}(\mathrm{d}z) - \mathrm{D}(\mu \parallel \lambda_1 \times \cdots \times \lambda_n),$$

where:

- $\mu_{[n]\setminus i}$ *is the projection of* μ *to* $\prod_{j\in[n]\setminus i} K_j$, *and*
- $(\mu_{i,z} : z \in \prod_{j\in[n]\setminus i} K_j)$ *is a conditional distribution under* μ *of the* i*th coordinate given the other coordinates.*

In the form given by this lemma, the nonnegativity of DTC is a sharpening of the usual logarithmic Sobolev inequality for product spaces. Indeed, this sharpening already appears implicitly inside standard proofs of that logarithmic Sobolev inequality: see the discussion at the end of [2], Section 6.

In case μ is the Gibbs measure of f, Lemma 2.4 turns into a simple expression for $\mathrm{DTC}(\mu)$ in terms of ∇f. It plays a crucial role later in this paper.

COROLLARY 2.5. *The Gibbs measure in* (1) *satisfies*

$$\mathrm{DTC}(\mu) = \int \mathrm{D}(\xi_{\nabla f(\boldsymbol{x},\cdot)} \parallel \lambda_1 \times \cdots \times \lambda_n)\mu(\mathrm{d}\boldsymbol{x}) - \mathrm{D}(\mu \parallel \lambda_1 \times \cdots \times \lambda_n). \tag{10}$$

PROOF. The additivity of KL divergence for product measures gives

$$\int \mathrm{D}(\xi_{\nabla f(\boldsymbol{x},\cdot)} \parallel \lambda_1 \times \cdots \times \lambda_n)\mu(\mathrm{d}\boldsymbol{x}) = \sum_{i=1}^{n} \int \mathrm{D}(\xi_{\partial_i f(\boldsymbol{x},\cdot)} \parallel \lambda_i)\mu(\mathrm{d}\boldsymbol{x}).$$

Using this, the right-hand side of (10) matches the formula for $\mathrm{DTC}(\mu)$ in Lemma 2.4, because $\xi_{\partial_i f(\boldsymbol{x},\cdot)}$ is precisely the conditional distribution of x_i under μ given the other coordinates $\boldsymbol{x}_{[n]\setminus i}$. □

In the proof of Theorem A, the fact we really need is that the difference on the right-hand side of (10) is nonnegative. However, Theorem A also provides a bound on $\mathrm{DTC}(\mu)$. This actually offers an alternative route to a decomposition of μ into near-product measures: the main result of [2] obtains such a decomposition precisely from a bound on DTC. We discuss this further at the end of Section 3.

3. Proof of Theorem A. We now return to the setting of Theorem A.

The measure μ and discrete gradient ∇f are both constructed from the potential function f. The proof of Theorem A depends on the following link between them.

LEMMA 3.1. *We have*

$$\int \nabla f(\boldsymbol{x}, \boldsymbol{x})\mu(\mathrm{d}\boldsymbol{x}) = \iint \nabla f(\boldsymbol{x}, \boldsymbol{y})\, \xi_{\nabla f(\boldsymbol{x},\cdot)}(\mathrm{d}\boldsymbol{y})\, \mu(\mathrm{d}\boldsymbol{x}).$$

PROOF. It suffices to prove that

$$\int \partial_i f(\boldsymbol{x}, x_i)\mu(\mathrm{d}\boldsymbol{x}) = \iint \partial_i f(\boldsymbol{x}, y)\, \xi_{\partial_i f(\boldsymbol{x},\cdot)}(\mathrm{d}y)\, \mu(\mathrm{d}\boldsymbol{x}) \tag{11}$$

for each $i \in [n]$, for then we can just sum over i. The function $\partial_i f(\boldsymbol{x}, \cdot)$ depends only on $\boldsymbol{x}_{[n]\setminus i}$, and $\xi_{\partial_i f(\boldsymbol{x},\cdot)}$ is the conditional distribution of x_i under μ given the other coordinates $\boldsymbol{x}_{[n]\setminus i}$. Therefore

$$\int \partial_i f(\boldsymbol{x}, y)\, \xi_{\partial_i f(\boldsymbol{x},\cdot)}(\mathrm{d}y) = E_\mu\big[\partial_i f(\boldsymbol{x}, x_i) \mid \boldsymbol{x}_{[n]\setminus i}\big],$$

and (11) is the tower property of iterated conditional expectations. □

To prove Theorem A, we must bound both $\mathrm{DTC}(\mu)$ and the expression

$$\sum_P \mu(P) \int \mathrm{D}(\mu_{|P} \parallel \xi_{\nabla f(\boldsymbol{y},\cdot)})\mu_{|P}(\mathrm{d}\boldsymbol{y}). \tag{12}$$

The proof focuses on this expression, but also yields a bound on $\mathrm{DTC}(\mu)$ as a by-product. This feature of our work bears a curious resemblance to another part of Eldan's paper [14]. For a Gibbs measure μ defined relative to a standard Gaussian distribution on $\mathbb{R}^n$, rather than on a hypercube, [14], Theorem 4, gives an upper bound on the deficit in the Gaussian logarithmic Sobolev inequality satisfied by μ in terms of the gradient complexity (measured using Gaussian width) of its potential function. In view of the relationship between DTC and deficits in logarithmic Sobolev inequalities, already remarked following Lemma 2.4 above, our bound on $\mathrm{DTC}(\mu)$ (Theorem A part (a)) could be an analog of that deficit bound in [14]. We do not explore this connection further here.

To lighten notation during the rest of this section, let us abbreviate

$$\mathrm{D}\Big(\nu \Big\| \prod_{i\in S} \lambda_i\Big) \quad \text{to } \mathrm{D}(\nu)$$

whenever $S \subseteq [n]$ and $\nu \in \mathrm{Prob}(\prod_{i\in S} K_i)$.

Before the formal proof, let us give an informal sketch highlighting the relevance of Lemma 3.1. In this sketch, we assume that $K_i = \{0, 1\}$ for each i. In this case each discrete gradient $\nabla f(\boldsymbol{x}, \cdot)$ may be identified with a vector in $\mathbb{R}^n$: its ith entry is

$$f(x_1, x_2, \ldots, x_{i-1}, 1, x_{i+1}, \ldots, x_n) - f(x_1, x_2, \ldots, x_{i-1}, 0, x_{i+1}, \ldots, x_n). \tag{13}$$

Conversely, any vector $\boldsymbol{u} \in \mathbb{R}^n$ defines a linear function on $\{0,1\}^n$ using the Euclidean inner product: $\boldsymbol{u}(\boldsymbol{y}) := \sum_i u_i y_i$. If $\boldsymbol{u}$ is the vector given by (13), then $\boldsymbol{u}(\boldsymbol{y})$ agrees with our earlier definition of $\nabla f(\boldsymbol{x}, \boldsymbol{y})$ if we choose the reference points $*_i = 0$ for each i.

To simplify this sketch further, let us also imagine that the discrete gradient $\nabla f(\boldsymbol{x}, \cdot)$ takes only two distinct values in $\mathbb{R}^n$, say $\boldsymbol{u}$ and $\boldsymbol{v}$. (This is really a fantasy: a short exercise shows that, if f is not linear itself, then its discrete gradient function takes at least four distinct values. But the story is simplest if we pretend there are just two.)

Under these assumptions, we take the partition $\mathcal{P}$ to consist of

$$P := \{\boldsymbol{x} : \nabla f(\boldsymbol{x}, \cdot) = \boldsymbol{u}\} \quad \text{and} \quad P^{\mathrm{c}} := \{\boldsymbol{x} : \nabla f(\boldsymbol{x}, \cdot) = \boldsymbol{v}\}.$$

Assume further that neither of the values $\mu(P)$, $\mu(P^{\mathrm{c}})$ is close to zero. (If one of them is close to zero, then a slightly degenerate version of the ensuing discussion applies.) So we need to show that $\mathrm{D}(\mu_{|P} \parallel \xi_{\boldsymbol{u}})$ and $\mathrm{D}(\mu_{|P^{\mathrm{c}}} \parallel \xi_{\boldsymbol{v}})$ are both small relative to n.

Here are the steps in the proof:

- Gibbs' variational principle lets us interpret this requirement another way: $\mu_{|P}$ must come close to saturating the variational inequality that characterizes the product measure $\xi_{\boldsymbol{u}}$, meaning that

$$\mathrm{D}(\mu_{|P}) - \int \boldsymbol{u}\, \mathrm{d}\mu_{|P} \quad \text{is not much larger than} \quad \mathrm{D}(\xi_{\boldsymbol{u}}) - \int \boldsymbol{u}\, \mathrm{d}\xi_{\boldsymbol{u}} \tag{14}$$

  relative to n, and similarly when comparing $\mu_{|P^{\mathrm{c}}}$ with $\xi_{\boldsymbol{v}}$.
- For each of $\mu_{|P}$ and $\mu_{|P^{\mathrm{c}}}$ separately, we have no clear way to control the gap in (14). But we can control the average of those two gaps over P and P^{c}, keeping in mind that $\nabla f(\boldsymbol{x}, \cdot)$ equals $\boldsymbol{u}$ on P and $\boldsymbol{v}$ on P^{c} respectively. By Lemma 2.1, the average of the KL divergences is

$$\mu(P) \cdot \mathrm{D}(\mu_{|P}) + \mu\big(P^{\mathrm{c}}\big) \cdot \mathrm{D}(\mu_{|P^{\mathrm{c}}}) = \mathrm{D}(\mu) + \mathrm{H}_\mu(\mathcal{P}),$$

  and by Corollary 2.5 this is equal to

$$\begin{aligned} &\left[\int \mathrm{D}(\xi_{\nabla f(\boldsymbol{x},\cdot)})\mu(\mathrm{d}\boldsymbol{x}) - \mathrm{DTC}(\mu)\right] + \mathrm{H}_\mu(\mathcal{P}) \\ &\quad = \mu(P) \cdot \mathrm{D}(\xi_{\boldsymbol{u}}) + \mu\big(P^{\mathrm{c}}\big) \cdot \mathrm{D}(\xi_{\boldsymbol{v}}) - \mathrm{DTC}(\mu) + \mathrm{H}_\mu(\mathcal{P}). \end{aligned}$$

  The average of the required integrals is

$$\begin{aligned} &\mu(P) \int \boldsymbol{u}\, \mathrm{d}\mu_{|P} + \mu\big(P^{\mathrm{c}}\big) \int \boldsymbol{v}\, \mathrm{d}\mu_{|P^{\mathrm{c}}} \\ &\quad = \int_P \boldsymbol{u}\, \mathrm{d}\mu + \int_{P^{\mathrm{c}}} \boldsymbol{v}\, \mathrm{d}\mu \\ &\quad = \int \nabla f(\boldsymbol{x}, \boldsymbol{x}) \mu(\mathrm{d}\boldsymbol{x}) \end{aligned}$$

$$= \iint \nabla f(\boldsymbol{x}, \boldsymbol{y}) \xi_{\nabla f(\boldsymbol{x},\cdot)}(\mathrm{d}\boldsymbol{y}) \mu(\mathrm{d}\boldsymbol{x})$$

$$= \mu(P) \int \boldsymbol{u}(\boldsymbol{y}) \xi_{\boldsymbol{u}}(\mathrm{d}\boldsymbol{y}) + \mu\big(P^{\mathrm{c}}\big) \int \boldsymbol{v}(\boldsymbol{y}) \xi_{\boldsymbol{v}}(\mathrm{d}\boldsymbol{y}).$$

The equality of the third and fourth lines here is the crucial appearance of Lemma 3.1. Combining these calculations of averages, we arrive at

$$\begin{aligned} &\mu(P) \cdot \mathrm{D}(\mu_{|P} \parallel \xi_{\boldsymbol{u}}) + \mu\big(P^{\mathrm{c}}\big) \cdot \mathrm{D}(\mu_{|P^{\mathrm{c}}} \parallel \xi_{\boldsymbol{v}}) \\ &\quad = \mu(P)\bigg[\mathrm{D}(\mu_{|P}) - \int \boldsymbol{u} \, \mathrm{d}\mu_{|P}\bigg] - \mu(P)\bigg[\mathrm{D}(\xi_{\boldsymbol{u}}) - \int \boldsymbol{u} \, \mathrm{d}\xi_{\boldsymbol{u}}\bigg] \\ &\qquad + \mu\big(P^{\mathrm{c}}\big)\bigg[\mathrm{D}(\mu_{|P^{\mathrm{c}}}) - \int \boldsymbol{v} \, \mathrm{d}\mu_{|P^{\mathrm{c}}}\bigg] - \mu\big(P^{\mathrm{c}}\big)\bigg[\mathrm{D}(\xi_{\boldsymbol{v}}) - \int \boldsymbol{v} \, \mathrm{d}\xi_{\boldsymbol{v}}\bigg] \\ &\quad = -\mathrm{DTC}(\mu) + \mathrm{H}_{\mu}(\mathcal{P}). \end{aligned} \tag{15}$$

- Since DTC is nonnegative, the last line above is at most $\mathrm{H}_{\mu}(\mathcal{P}) \le \log|\mathcal{P}| = \log 2$. This is indeed small compared to n, and hence so are both $\mathrm{D}(\mu_{|P} \parallel \xi_{\boldsymbol{u}})$ and $\mathrm{D}(\mu_{|P^{\mathrm{c}}} \parallel \xi_{\boldsymbol{v}})$. Moreover, the last line above cannot be negative, since it is a positive combination of differences in the variational principle. So we also find that $\mathrm{DTC}(\mu) \le \log 2$.

Now we give the careful proof in full.

PROOF OF THEOREM A. Start by considering a single $P \in \mathcal{P}$. For any additively separable function g, Proposition 2.2 gives

$$\mathrm{D}(\mu_{|P} \parallel \xi_g) = \bigg[\mathrm{D}(\mu_{|P}) - \int g \, \mathrm{d}\mu_{|P}\bigg] - \bigg[\mathrm{D}(\xi_g) - \int g \, \mathrm{d}\xi_g\bigg].$$

Let us average this identity over $g = \nabla f(\boldsymbol{x}, \cdot)$ where $\boldsymbol{x} \sim \mu_{|P}$. The result is

$$\begin{aligned} &\int \mathrm{D}(\mu_{|P} \parallel \xi_{\nabla f(\boldsymbol{x},\cdot)}) \mu_{|P}(\mathrm{d}\boldsymbol{x}) \\ &\quad = \mathrm{D}(\mu_{|P}) - \iint \nabla f(\boldsymbol{x}, \boldsymbol{y}) \mu_{|P}(\mathrm{d}\boldsymbol{y}) \mu_{|P}(\mathrm{d}\boldsymbol{x}) \\ &\qquad - \int \mathrm{D}(\xi_{\nabla f(\boldsymbol{x},\cdot)}) \mu_{|P}(\mathrm{d}\boldsymbol{x}) + \iint \nabla f(\boldsymbol{x}, \boldsymbol{y}) \xi_{\nabla f(\boldsymbol{x},\cdot)}(\mathrm{d}\boldsymbol{y}) \mu_{|P}(\mathrm{d}\boldsymbol{x}). \end{aligned}$$

Now we average this equality over $P \in \mathcal{P}$ with the weights $\mu(P)$. The third right-hand term simplifies according to the law of total probability:

$$\sum_P \mu(P) \int \mathrm{D}(\xi_{\nabla f(\boldsymbol{x},\cdot)}) \mu_{|P}(\mathrm{d}\boldsymbol{x}) = \int \mathrm{D}(\xi_{\nabla f(\boldsymbol{x},\cdot)}) \mu(\mathrm{d}\boldsymbol{x}).$$

The fourth right-hand term simplifies similarly. After these simplifications, we are left with

$$\sum_P \mu(P) \int \mathrm{D}(\mu_{|P} \parallel \xi_{\nabla f(\boldsymbol{y},\cdot)}) \mu_{|P}(\mathrm{d}\boldsymbol{y})$$

$$= \sum_P \mu(P) \cdot \mathrm{D}(\mu_{|P}) - \sum_P \mu(P) \iint \nabla f(\boldsymbol{x}, \boldsymbol{y}) \mu_{|P}(\mathrm{d}\boldsymbol{y}) \mu_{|P}(\mathrm{d}\boldsymbol{x}) \tag{16}$$
$$- \int \mathrm{D}(\xi_{\nabla f(\boldsymbol{x},\cdot)}) \mu(\mathrm{d}\boldsymbol{x}) + \iint \nabla f(\boldsymbol{x}, \boldsymbol{y}) \xi_{\nabla f(\boldsymbol{x},\cdot)}(\mathrm{d}\boldsymbol{y}) \mu(\mathrm{d}\boldsymbol{x}).$$

Next we rewrite the right-hand side of (16) by considering separately (i) the KL-divergence terms and (ii) the double-integral terms:

(i) By Lemma 2.1 and Corollary 2.5, we have

$$\sum_P \mu(P) \cdot \mathrm{D}(\mu_{|P}) - \int \mathrm{D}(\xi_{\nabla f(\boldsymbol{x},\cdot)}) \mu(\mathrm{d}\boldsymbol{x})$$
$$= \big[\mathrm{D}(\mu) + \mathrm{H}_\mu(\mathcal{P})\big] - \big[\mathrm{D}(\mu) + \mathrm{DTC}(\mu)\big] = \mathrm{H}_\mu(\mathcal{P}) - \mathrm{DTC}(\mu).$$

(ii) Using Lemma 3.1 to substitute for the second double integral in (16), and then using the law of total probability, the difference of those double-integral terms is equal to

$$\sum_P \mu(P) \Big[- \iint \nabla f(\boldsymbol{x}, \boldsymbol{y}) \mu_{|P}(\mathrm{d}\boldsymbol{y}) \mu_{|P}(\mathrm{d}\boldsymbol{x}) + \int \nabla f(\boldsymbol{y}, \boldsymbol{y}) \mu_{|P}(\mathrm{d}\boldsymbol{y}) \Big]$$
$$= \sum_P \mu(P) \iint \big(\nabla f(\boldsymbol{y}, \boldsymbol{y}) - \nabla f(\boldsymbol{x}, \boldsymbol{y})\big) \mu_{|P}(\mathrm{d}\boldsymbol{y}) \mu_{|P}(\mathrm{d}\boldsymbol{x}).$$

Inserting these calculations into (16) and rearranging slightly, we arrive at the identity

$$\mathrm{DTC}(\mu) + \sum_P \mu(P) \int \mathrm{D}(\mu_{|P} \parallel \xi_{\nabla f(\boldsymbol{y},\cdot)}) \mu_{|P}(\mathrm{d}\boldsymbol{y})$$
$$= \mathrm{H}_\mu(\mathcal{P}) + \sum_P \mu(P) \iint \big(\nabla f(\boldsymbol{y}, \boldsymbol{y}) - \nabla f(\boldsymbol{x}, \boldsymbol{y})\big) \mu_{|P}(\mathrm{d}\boldsymbol{y}) \mu_{|P}(\mathrm{d}\boldsymbol{x}). \tag{17}$$

This identity generalizes equation (15) in the proof-sketch above.

By our assumption (4), we have

$$\big|\nabla f(\boldsymbol{y}, \boldsymbol{y}) - \nabla f(\boldsymbol{x}, \boldsymbol{y})\big| < \delta n$$

whenever $\boldsymbol{x}$ and $\boldsymbol{y}$ lie in the same cell of $\mathcal{P}$. This implies that the average of double integrals on the right-hand side of (17) is less than δn, and so

$$\mathrm{DTC}(\mu) + \sum_P \mu(P) \int \mathrm{D}(\mu_{|P} \parallel \xi_{\nabla f(\boldsymbol{y},\cdot)}) \mu_{|P}(\mathrm{d}\boldsymbol{y}) < \mathrm{H}_\mu(\mathcal{P}) + \delta n. \tag{18}$$

Both terms on the left-hand side of (18) are nonnegative, so conclusions (a) and (b) of Theorem A follow immediately. Conclusion (b) implies conclusion (c) using

Marton's inequality (Proposition 2.3) and Hölder's inequality:

$$\begin{aligned}
&\sum_P \mu(P) \int \overline{d_n}(\mu_{|P}, \xi_{\nabla f(\boldsymbol{y},\cdot)})\mu_{|P}(\mathrm{d}\boldsymbol{y}) \\
&\quad \le \sum_P \mu(P) \int \sqrt{\frac{1}{2n}\mathrm{D}(\mu_{|P} \parallel \xi_{\nabla f(\boldsymbol{y},\cdot)})}\mu_{|P}(\mathrm{d}\boldsymbol{y}) \\
&\quad \le \sqrt{\sum_P \mu(P) \int \frac{1}{2n}\mathrm{D}(\mu_{|P} \parallel \xi_{\nabla f(\boldsymbol{y},\cdot)})\mu_{|P}(\mathrm{d}\boldsymbol{y})}.
\end{aligned}$$

□

The proof above is quite insensitive to the structure of the spaces K_i. When $K_i = \{-1, 1\}$, it yields similar results to those obtained from Eldan's diffusion approach. But one can imagine 'even smaller' marginal spaces:

QUESTION 3.2. *Can the structural results be improved when* $K_i = \{-1, 1\}$ *and the reference measures* λ_i *are highly biased, for instance when*

$$\lambda_i\{1\} = p \ll 1 \quad \textit{for each } i?$$

Next let us fill in the proof of Corollary A′ from Theorem A.

PROOF OF COROLLARY A′ FROM THEOREM A. Let $\mathcal{Q}$ be a partition of

$$\left\{\nabla f(\boldsymbol{x}, \cdot) : \boldsymbol{x} \in \prod_i K_i\right\}$$

into sets of $\|\cdot\|$-diameter less than δn, and let $\mathcal{P}$ be the pullback of $\mathcal{Q}$ under the map $\boldsymbol{x} \mapsto \nabla f(\boldsymbol{x}, \cdot)$. Now pick an element $\boldsymbol{y}_P$ in each $P \in \mathcal{P}$ which minimizes $\overline{d_n}(\mu_{|P}, \xi_{\nabla f(\boldsymbol{y}_P,\cdot)})$, and set $g_P := \nabla f(\boldsymbol{y}_P, \cdot)$. The first desired inequality follows from part (b) of Theorem A, and then the second follows by the same use of Proposition 2.3 as above. □

3.1. *Possible variations on Theorem A.* The proof of Theorem A pivots around the identity (17). As Amir Dembo has emphasized to me, this identity may have other valuable consequences. The double integral on the right seems to deserve particular attention. In the proof above we simply use a uniform bound on the integrand $\nabla f(\boldsymbol{y}, \boldsymbol{y}) - \nabla f(\boldsymbol{x}, \boldsymbol{y})$. Are there other ways to bound this double integral, perhaps by exploiting further the special structure of the Gibbs measure μ? The reward could be a result similar to Theorem A but with weaker hypotheses.

In a separate direction, one could try replacing the use of Marton's inequality (Proposition 2.3) in Theorem A with a different transportation-entropy inequality. In [11], Dembo proves several analogs of Proposition 2.3 that replace $\overline{d_n}$ with various alternative transportation-like quantities. Those analogs then recover some

powerful variations on product-space measure concentration that were previously discovered and used by Talagrand [26]. It might be worth exploring a version of Theorem A part (c) for one of those alternative transportation-like quantities.

Further possibilities arise if we know more about the coordinate spaces K_i. We have already mentioned the case of two-point spaces in connection with the foundational papers [7] and [14], but other special examples are also worth considering. For example, if $K_i = [0, 1]$ for each i and if $f : [0, 1]^n \longrightarrow \mathbb{R}$ is differentiable, then we can imagine using its gradient Df in the usual sense of calculus as a substitute for our discrete gradient ∇f. Regarding Df as a function from $[0, 1]^n$ to $\mathbb{R}^n$, we can still measure its 'complexity' in terms of covering numbers, and seek a description of the corresponding Gibbs measures μ if this 'complexity' is small enough. Some of the steps above should adapt with little change, but a few sticking points stand out. Perhaps most important is Lemma 3.1, which depends on recognizing $\xi_{\partial_i(\boldsymbol{x}\cdot)}$ as the conditional distribution of x_i under μ given the other coordinates $\boldsymbol{x}_{[n]\setminus i}$. This identification does not carry over simply to a more 'localized' notion of gradient than ∇f. As a result, Lemma 3.1 would have to be replaced with a different calculation, probably involving some extra terms whose control requires new methods or assumptions (perhaps on the second derivatives of f, by analogy with some of the estimates in [7]).

3.2. *An alternative approach using the DTC bound.* We have seen that conclusion (a) of Theorem A emerges naturally during the course of proving conclusion (b), and that conclusion (b) implies conclusion (c). However, according to the main result of [2] (labelled Theorem A in that paper), conclusion (a) by itself implies something like conclusion (c). To be precise, if μ is a probability measure on a product space $\prod_i K_i$ for which $\mathrm{DTC}(\mu) \leq r^3 n$, then that result from [2] asserts that μ can be a represented as a mixture

$$\mu = \int_L \mu_y \nu(\mathrm{d}y)$$

such that (i) the mutual information in the mixture is at most $\mathrm{DTC}(\mu)$ and (ii) there is a measurable family $(\xi_y : y \in L)$ of product measures on $\prod_i K_i$ satisfying

$$\int_L \overline{d_n}(\mu_y, \xi_y)\nu(\mathrm{d}y) < 2r.$$

In our setting, if we know that

$$\mathrm{H}_\mu(\mathcal{P}) + \delta n \leq r^3 n,$$

then this gives such a mixture representation of μ with mutual information at most $\mathrm{H}_\mu(\mathcal{P}) + \delta n$. This is a softer route with a similar conclusion to Theorem A part (c). It gives only a mixture with controlled mutual information, rather than a controlled number of terms, but one could then obtain the latter using [3], Theorem 9.5. However, this approach gives weaker estimates than the proof of Theorem A above, which exploits more directly the special nature of our Gibbs measures.

3.3. *Comparison with Eldan's main structure theorem.* In this subsection C stands for several universal constants which are not computed explicitly. The value of C may change from one appearance to the next.

In [14], Eldan considers a Gibbs measure μ as in (1) on the space $\{-1, 1\}^n$. In this case, each discrete gradient $\nabla f(\boldsymbol{x}, \cdot)$ may be identified with a vector in $\mathbb{R}^n$: this is similar to (13) except using 1 and -1 instead of 1 and 0. (To be precise, this differs from Eldan's convention by a factor of $1/2$, which affects a few of the constants below.) Eldan measures the 'complexity' of f relative to linear functions by the Gaussian width of the set of its discrete gradients:

$$\mathcal{D}(f) := \mathrm{GW}\big(\{\nabla f(\boldsymbol{x}, \cdot) : \boldsymbol{x} \in \{-1, 1\}^n\} \cup \{\mathbf{0}\}\big).$$

This quantity is called the 'Gaussian-width gradient complexity' of f. Eldan's main structure theorem [14], Theorem 3, represents a Gibbs measure μ as a mixture of other measures which (i) are mostly close to product measures in $\overline{d_n}$ and (ii) on average carry almost as much entropy as μ itself, where the quality of these estimates depends on $\mathcal{D}(f)$.

Let us write $\gamma := \mathcal{D}(f)/n$ for convenience. Specifically, after fixing $\varepsilon > 0$ and also picking an auxiliary parameter $\alpha > 1$, [14], Theorem 3, provides a mixture

$$\mu = \int \mu_\theta \, m(\mathrm{d}\theta)$$

over some space of parameters θ such that:

(i) $\mathrm{H}(\mu) - \int \mathrm{H}(\mu_\theta) m(\mathrm{d}\theta) < C\varepsilon n$ and

(ii) there is a subset Θ of values of θ such that $m(\Theta) > 1 - \frac{1}{n} - \frac{1}{\alpha}$ and such that every $\theta \in \Theta$ satisfies

$$\overline{d_n}(\mu_\theta, \xi_{\boldsymbol{v}_\theta}) \le C\sqrt{\alpha\gamma/\varepsilon}$$

for some product measure $\xi_{\boldsymbol{v}_\theta}$.

Eldan's theorem gives other conclusions as well, such as bounds on the vectors $\boldsymbol{v}_\theta$. The ingredients in Eldan's mixture are 'tilts' of μ, whereas in our work they are the conditioned measures $\mu_{|P}$. We do not explore these features further here, although the special structure of tilts does seem to give an advantage in some of Eldan's applications of his method (particularly the proof of [14], Theorem 1).

By Sudakov's minoration [19], Theorem 3.18, and the inequality $\|\cdot\|_1 \le \sqrt{n}\|\cdot\|_2$, the covering numbers of a nonempty subset $K \subseteq \mathbb{R}^n$ are related to Gaussian width according to

$$\delta\sqrt{n}\sqrt{\log \mathrm{cov}_{\delta n}\big(K, \|\cdot\|_1\big)} \le C \cdot \mathrm{GW}(K) \quad \forall \delta > 0.$$

Also, our identification of each discrete gradient with a vector in $\mathbb{R}^n$ satisfies $\|\nabla f(\boldsymbol{x}, \cdot)\| \le \|\nabla f(\boldsymbol{x}, \cdot)\|_1$, where $\|\cdot\|$ denotes the uniform norm for functions on $\{-1, 1\}^n$ as previously. Therefore, with γ as above, we have

$$\log \mathrm{cov}_{\delta n}\big(\{\nabla f(\boldsymbol{x}, \cdot) : \boldsymbol{x} \in \{-1, 1\}^n\}, \|\cdot\|\big) \le C\frac{\mathcal{D}(f)^2}{\delta^2 n} = C\frac{\gamma^2}{\delta^2} n \tag{19}$$

for all $\delta > 0$. For a given δ, this gives (5) with $\varepsilon = \gamma^2/\delta^2$ up to a multiplicative constant. Thus, assuming a bound on γ does not force a particular choice of ε and δ in (5), but sets up a trade-off between them.

After making a choice of δ in (19) and letting $\varepsilon := \gamma^2/\delta^2$, the partition in Corollary A′ satisfies

$$\mathrm{H}(\mu) - \sum_P \mu(P)\mathrm{H}(\mu_{|P}) = \mathrm{H}_\mu(\mathcal{P}) \le C\varepsilon n \tag{20}$$

for the approximation of entropies, and the conclusion of Corollary A′ gives

$$\sum_P \mu(P) \cdot \overline{d_n}(\mu_{|P}, \xi_{g_P}) \le C\sqrt{\varepsilon + \frac{\gamma}{\sqrt{\varepsilon}}} \le C \max\{\sqrt{\varepsilon}, \sqrt{\gamma/\sqrt{\varepsilon}}\}. \tag{21}$$

If $\varepsilon \le \gamma^{2/3}$, then (20) matches Eldan's conclusion (i) above up to a constant, and the right-hand side of (21) is $C\sqrt{\gamma/\sqrt{\varepsilon}}$, which improves on Eldan's conclusion (ii) when ε is very small.

However, if $\varepsilon > \gamma^{2/3}$ (equivalently, $\delta < \gamma^{2/3}$), then the right-hand side of (21) is $C\sqrt{\varepsilon}$. In this case we may reduce ε (equivalently, increase δ) and improve both of the bounds (20) and (21). So there is no value in using our estimates with $\varepsilon > \gamma^{2/3}$. By contrast, Eldan's bounds are still potentially useful in this range, which enables them to perform better than ours in some applications. We discuss an example at the end of Section 5.

This comparison brings up an interesting puzzle:

QUESTION 3.3. *Suppose we begin with a bound on γ and then use (19) to establish the trade-off $\varepsilon = \gamma^2/\delta^2$. Can we vary the proof of Theorem A so that there is some benefit to choosing $\delta < \gamma^{2/3}$? This would mean that at least one of the left-hand sides of (20) and (21) is bounded more effectively when $\delta < \gamma^{2/3}$ than when $\delta \ge \gamma^{2/3}$. (I would guess that any improvement must be to (21).)*

4. A characterization of the terms in the mixture. In this section, we suppose that the discrete gradient function

$$\prod_i K_i \longrightarrow C_{\mathrm{b}}\Big(\prod_i K_i\Big) : \boldsymbol{x} \mapsto \nabla f(\boldsymbol{x}, \cdot)$$

is Lipschitz from d_n to the uniform norm $\|\cdot\|$. Let

$$L := \sup\left\{ \frac{\|\nabla f(\boldsymbol{x}, \cdot) - \nabla f(\boldsymbol{y}, \cdot)\|}{d_n(\boldsymbol{x}, \boldsymbol{y})} : \boldsymbol{x}, \boldsymbol{y} \in \prod_i K_i \text{ distinct} \right\}$$

be its Lipschitz constant (beware: L is not the Lipschitz constant of f itself).

Let $\mathcal{P}$ be a partition of $\prod_i K_i$ as in Corollary A′, and let $\boldsymbol{y}_P \in P$ for $P \in \mathcal{P}$ be a selection of points that minimize the distances $\overline{d_n}(\mu_{|P}, \xi_{\nabla f(\boldsymbol{y}_P, \cdot)})$. Recall that the functions g_P of Corollary A′ are then given by $g_P := \nabla f(\boldsymbol{y}_P, \cdot)$.

PROPOSITION 4.1. *These data satisfy*

$$\sum_P \mu(P)\left\| g_P - \int \nabla f(\boldsymbol{x},\cdot)\xi_{g_P}(\mathrm{d}\boldsymbol{x})\right\| \le \delta n + L\sqrt{\frac{1}{2}\left(\frac{\mathrm{H}_\mu(\mathcal{P})}{n}+\delta\right)}.$$

Assuming that the right-hand side here is small, this means that most of the probability vector $(\mu(P) : P \in \mathcal{P})$, when transferred to the list of functions $(g_P : P \in \mathcal{P})$, is on terms that satisfy the 'approximate fixed point' equation

$$g_P \approx \int \nabla f(\boldsymbol{x},\cdot)\xi_{g_P}(\mathrm{d}\boldsymbol{x}). \tag{22}$$

PROOF. Since $\boldsymbol{y}_P \in P$, our assumption (4) about $\mathcal{P}$ gives

$$\left\| g_P - \int \nabla f(\boldsymbol{x},\cdot)\mu_{|P}(\mathrm{d}\boldsymbol{x})\right\| < \delta n \quad \forall P \in \mathcal{P}.$$

On the other hand, the definition of L gives

$$\left\| \int \nabla f(\boldsymbol{x},\cdot)\mu_{|P}(\mathrm{d}\boldsymbol{x}) - \int \nabla f(\boldsymbol{x},\cdot)\xi_{g_P}(\mathrm{d}\boldsymbol{x})\right\| \le L\cdot\overline{d_n}(\mu_{|P},\xi_{g_P}) \quad \forall P \in \mathcal{P}.$$

Averaging over $\mathcal{P}$ with weights $\mu(P)$, the result follows by conclusion (c) of Theorem A and the triangle inequality. □

In the special setting of $\{-1,1\}^n$, the approximate equation (22) admits an alternative form which resembles its counterpart in [16] more closely. In that setting, we can identify each discrete gradient $\nabla f(\boldsymbol{x},\cdot)$ with a vector in $\mathbb{R}^n$, as previously. Also, any product measure on $\{-1,1\}^n$ is uniquely specified by its barycentre in $[-1,1]^n$. Let $\boldsymbol{m}_P$ be the barycentre of ξ_{g_P} for each P. After these identifications, equation (22) is equivalent to

$$\boldsymbol{m}_P \approx \tanh\left(\int \nabla f(\boldsymbol{x},\cdot)\xi_{g_P}(\mathrm{d}\boldsymbol{x})\right) = \tanh\bigl(D\widetilde{f}(\boldsymbol{m}_P)\bigr), \tag{23}$$

where $\widetilde{f}$ is the harmonic extension of f to $[-1,1]^n$ (see [14], Section 3.1.1), $D\widetilde{f}$ is its derivative in the usual sense of calculus, and tanh is applied coordinate-wise. Equation (23) is essentially the same as [16], equation (8).

Let us return to the formulation in terms of g_P rather than $\boldsymbol{m}_P$, but derive from Proposition 4.1 a bound in terms of the Gaussian-width gradient complexity $\mathcal{D}(f)$ studied by Eldan. Let $\gamma := \mathcal{D}(f)/n$, as previously; consider an auxiliary parameter $\delta > 0$; and choose a partition $\mathcal{P}$ as promised by inequality (19) for this δ. Then (20) lets us turn Proposition 4.1 into

$$\sum_P \mu(P)\left\| g_P - \int \nabla f(\boldsymbol{x},\cdot)\xi_{g_P}(\mathrm{d}\boldsymbol{x})\right\| \le \delta n + CL\sqrt{\frac{\gamma^2}{\delta^2}+\delta} \quad \forall \delta > 0$$

(where once again C is a universal constant that we do not estimate explicitly, and may change value when it appears again below).

In the regime $L \geq Cn$, which seems to be more relevant to applications, the second right-hand term above is the more significant. To minimize it, we make the choice $\delta := \gamma^{2/3}$. Assuming also that γ is small, we are left with

$$\sum_P \mu(P) \left\| g_P - \int \nabla f(\boldsymbol{x}, \cdot) \xi_{g_P}(\mathrm{d}\boldsymbol{x}) \right\| \leq \gamma^{2/3} n + CL\gamma^{1/3} \leq CL\gamma^{1/3}.$$

A result of this flavour is obtained by Eldan and Gross in [16], Theorem 9. But their estimates have quite a different shape from ours (for instance, they also involve the Lipschitz constant of f itself, not just that of ∇f), and a direct comparison seems difficult.

5. Approximation of partition functions. One of the main potential applications of Theorem A is to the estimation of the partition function $\int \mathrm{e}^f \, \mathrm{d}\lambda$. By Proposition 2.2, it satisfies

$$\log \int \mathrm{e}^f \, \mathrm{d}\lambda \geq \int f \, \mathrm{d}\nu - \mathrm{D}(\nu \parallel \lambda) \tag{24}$$

for any probability measure ν on $\prod_i K_i$, with equality if and only if $\nu = \mu$, the Gibbs measure associated to f. A structural result like Theorem A allows one to achieve approximate equality using a product measure ν on the right-hand side. This can help us estimate the expression on the left.

PROPOSITION 5.1. *If f is L-Lipschitz for the metric d_n and satisfies (5), then*

$$\log \int \mathrm{e}^f \, \mathrm{d}\lambda \leq \sup_\xi \left[\int f \, \mathrm{d}\xi - \mathrm{D}(\xi \parallel \lambda) \right] + (\varepsilon + \delta)n + \sqrt{\frac{\varepsilon + \delta}{2}} L, \tag{25}$$

where the supremum runs over product measures on $\prod_i K_i$.

REMARK. In case each K_i is finite and λ_i is uniform, standard manipulations turn (25) into

$$\log \sum_{\boldsymbol{x}} \mathrm{e}^{f(\boldsymbol{x})} \leq \sup_\xi \left[\mathrm{H}(\xi) + \int f \, \mathrm{d}\xi \right] + (\varepsilon + \delta)n + \sqrt{\frac{\varepsilon + \delta}{2}} L.$$

This is the relevant form for many applications of nonlinear large deviations.

PROOF OF PROPOSITION 5.1. Let $\mathcal{P}$ be the partition implied by (5). Gibbs' identity and Corollary 2.5 give

$$\begin{aligned} \log \int \mathrm{e}^f \, \mathrm{d}\lambda &= \int f \, \mathrm{d}\mu - \mathrm{D}(\mu \parallel \lambda) \\ &= \int f \, \mathrm{d}\mu + \mathrm{DTC}(\mu) - \int \mathrm{D}(\xi_{\nabla f(\boldsymbol{y}, \cdot)} \parallel \lambda) \mu(\mathrm{d}\boldsymbol{y}). \end{aligned} \tag{26}$$

Using part (c) of Theorem A, we have

$$\begin{aligned}\int f\,\mathrm{d}\mu &= \sum_P \mu(P)\int f\,\mathrm{d}\mu_{|P}\\ &\le \iint f\,\mathrm{d}\xi_{\nabla f(\boldsymbol{y},\cdot)}\mu(\mathrm{d}\boldsymbol{y}) + L\sum_P \mu(P)\int \overline{d_n}(\mu_{|P},\xi_{\nabla f(\boldsymbol{y},\cdot)})\mu_{|P}(\mathrm{d}\boldsymbol{y}) \qquad (27)\\ &\le \iint f\,\mathrm{d}\xi_{\nabla f(\boldsymbol{y},\cdot)}\mu(\mathrm{d}\boldsymbol{y}) + \sqrt{\frac{\varepsilon+\delta}{2}}L.\end{aligned}$$

Inserting this bound and also part (a) of Theorem A into (26), we obtain

$$\log\int \mathrm{e}^f\,\mathrm{d}\lambda \le \int\biggl[\int f\,\mathrm{d}\xi_{\nabla f(\boldsymbol{y},\cdot)} - \mathrm{D}(\xi_{\nabla f(\boldsymbol{y},\cdot)}\parallel\lambda)\biggr]\mu(\mathrm{d}\boldsymbol{y}) + (\varepsilon+\delta)n + \sqrt{\frac{\varepsilon+\delta}{2}}L.$$

Now we bound the integral over $\boldsymbol{y}$ by the supremum over all product measures ξ, as in (25). □

Partition-function estimation was the heart of the original paper [7] which instigated research into nonlinear large deviations. A general discussion of the usefulness of such estimates is given in the Introduction to that paper. In [28], Yan extends Chatterjee and Dembo's estimates from the cube $\{0,1\}^n$ to a more general class of products of subsets of Banach spaces. In the setting of $\{-1,1\}^n$, Eldan shows how partition-function estimates can be derived from his main structure theorem in [14], Corollary 2. One can compare Proposition 5.1 with Eldan's results using the ideas from Section 3.3, but I have not been able to recover the full strength of Eldan's estimate as a consequence of Theorem A. This seems to be related to the discussion at the end of Section 3.3 above. In the notation of that discussion, the appropriate choice of ε for the proof of [14], Corollary 2, is

$$C\biggl(\frac{L}{n}\biggr)^{2/3}\gamma^{1/3},$$

which is generally much larger than the threshold $\varepsilon=\gamma^{2/3}$ that marks the end of the usefulness of our estimates, as discussed in Section 3.3.

Acknowledgments. I am grateful to Sourav Chatterjee, Amir Dembo and Ofer Zeitouni for some insightful conversations. Along with Ronen Eldan and an anonymous referee, they also made valuable suggestions about earlier versions of this paper.

REFERENCES

[1] AUGERI, F. Nonlinear large deviation bounds with applications to traces of Wigner matrices and cycles counts in Erdős–Rényi graphs. Preprint. Available at arXiv:1810.01558.

[2] AUSTIN, T. Multi-variate correlation and mixtures of product measures. Preprint. Available at arXiv:1809.10272.

[3] AUSTIN, T. (2018). Measure concentration and the weak Pinsker property. *Publ. Math. Inst. Hautes Études Sci.* **128** 1–119. MR3905465

[4] BHATTACHARYA, B. B., GANGULY, S., LUBETZKY, E. and ZHAO, Y. (2017). Upper tails and independence polynomials in random graphs. *Adv. Math.* **319** 313–347. MR3695877

[5] BHATTACHARYA, B. B., GANGULY, S., SHAO, X. and ZHAO, Y. Upper tail large deviations for arithmetic progressions in a random set. Preprint. Available at arXiv:1605.02994.

[6] CHATTERJEE, S. (2017). *Large Deviations for Random Graphs*. *Lecture Notes in Math.* **2197**. Springer, Cham. MR3700183

[7] CHATTERJEE, S. and DEMBO, A. (2016). Nonlinear large deviations. *Adv. Math.* **299** 396–450. MR3519474

[8] COOK, N. and DEMBO, A. Large deviations of subgraph counts in Erdős–Rényi random graphs. Preprint. Available at arXiv:1809.11148.

[9] COVER, T. M. and THOMAS, J. A. (2006). *Elements of Information Theory*, 2nd ed. Wiley-Interscience, Hoboken, NJ. MR2239987

[10] CSISZÁR, I. (1975). I-divergence geometry of probability distributions and minimization problems. *Ann. Probab.* **3** 146–158. MR0365798

[11] DEMBO, A. (1997). Information inequalities and concentration of measure. *Ann. Probab.* **25** 927–939. MR1434131

[12] DEMBO, A. and ZEITOUNI, O. (2010). *Large Deviations Techniques and Applications*. *Stochastic Modelling and Applied Probability* **38**. Springer, Berlin. Corrected reprint of the second (1998) edition. MR2571413

[13] DUDLEY, R. M. (2002). *Real Analysis and Probability*. *Cambridge Studies in Advanced Mathematics* **74**. Cambridge Univ. Press, Cambridge. Revised reprint of the 1989 original. MR1932358

[14] ELDAN, R. (2018). Gaussian-width gradient complexity, reverse log-Sobolev inequalities and nonlinear large deviations. *Geom. Funct. Anal.* **28** 1548–1596. Preprint. Available at arXiv:1612.04346. MR3881829

[15] ELDAN, R. and GROSS, R. (2018). Exponential random graphs behave like mixtures of stochastic block models. *Ann. Appl. Probab.* **28** 3698–3735. Preprint. Available at arXiv:1707.01227. MR3861824

[16] ELDAN, R. and GROSS, R. (2018). Decomposition of mean-field Gibbs distributions into product measures. *Electron. J. Probab.* **23** Paper No. 35, 24. MR3798245

[17] HAN, T. S. (1975). Linear dependence structure of the entropy space. *Inf. Control* **29** 337–368. MR0453264

[18] HAN, T. S. (1978). Nonnegative entropy measures of multivariate symmetric correlations. *Inf. Control* **36** 133–156. MR0464499

[19] LEDOUX, M. and TALAGRAND, M. (2011). *Probability in Banach Spaces*: *Isoperimetry and Processes*. *Classics in Mathematics*. Springer, Berlin. Reprint of the 1991 edition. MR2814399

[20] LUBETZKY, E. and ZHAO, Y. (2017). On the variational problem for upper tails in sparse random graphs. *Random Structures Algorithms* **50** 420–436. MR3632418

[21] MARTON, K. (1986). A simple proof of the blowing-up lemma. *IEEE Trans. Inform. Theory* **32** 445–446. MR0838213

[22] MARTON, K. (1996). Bounding $\overline{d}$-distance by informational divergence: A method to prove measure concentration. *Ann. Probab.* **24** 857–866. MR1404531

[23] MARTON, K. (1998). Measure concentration for a class of random processes. *Probab. Theory Related Fields* **110** 427–439. MR1616492

[24] PINSKER, M. S. (1964). *Information and Information Stability of Random Variables and Processes*. *Translated and Edited by Amiel Feinstein*. Holden-Day, San Francisco, CA. MR0213190

[25] SAMSON, P.-M. (2000). Concentration of measure inequalities for Markov chains and Φ-mixing processes. *Ann. Probab.* **28** 416–461. MR1756011

[26] TALAGRAND, M. (1995). Concentration of measure and isoperimetric inequalities in product spaces. *Publ. Math. Inst. Hautes Études Sci.* **81** 73–205. MR1361756

[27] TALAGRAND, M. (1996). Transportation cost for Gaussian and other product measures. *Geom. Funct. Anal.* **6** 587–600. MR1392331

[28] YAN, J. Nonlinear large deviations: Beyond the hypercube. Preprint. Available at arXiv:1703.08887.

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF CALIFORNIA, LOS ANGELES
BOX 951555
LOS ANGELES, CALIFORNIA 90095-1555
USA
E-MAIL: tim@math.ucla.edu