THE STRUCTURE OF LOW-COMPLEXITY GIBBS MEASURES ON PRODUCT SPACES

BY TIM AUSTIN

University of California, Los Angeles

Let $K_1, \ldots, K_n$ be bounded, complete, separable metric spaces. Let $f : \prod_i K_i \to \mathbb{R}$ be a bounded and continuous potential function, and let $\mu \propto e^f$ be the associated Gibbs distribution.

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

This paper shows that a ‘low-complexity’ assumption on $\nabla f$ implies that $\mu$ can be approximated by a mixture of other measures, relatively few in number, and most of them close in a natural transportation distance to product measures. 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 $\nabla_x f$ 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 $\{\nabla_x f : x \in \prod_i K_i\}$, 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 $(K_1, d_{K_1}), \ldots, (K_n, d_{K_n})$ be nonempty, complete, separable metric spaces, all with diameter at most one. Let $C_b(K_i)$ and $\text{Prob}(K_i)$ denote the spaces of bounded continuous functions and Borel probability measures on $K_i$, respectively, and similarly for other topological spaces. Let $\lambda_i \in \text{Prob}(K_i)$ be a fixed reference measure for each $i$.

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

$$\mu(dx) := \frac{1}{Z} e^{f(x)} \prod_{i=1}^n \lambda_i(dx_i)$$

be the resulting Gibbs measure, where $Z$ is the normalizing constant that makes $\mu$ a probability measure (the ‘partition function’ of statistical mechanics).
Inside $C_b(\prod_i K_i)$ lies the subspace of functions that have the form

$$x \mapsto f_1(x_1) + \cdots + f_n(x_n)$$

for some $f_1 \in C_b(K_1), \ldots, f_n \in C_b(K_n)$. We denote this subspace by $\Sigma_i C_b(K_i)$. We write $\| \cdot \|$ for the uniform norm on $C_b(\prod_i K_i)$ and for its restriction to this subspace. 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

$$\frac{1}{Z} \prod_{i=1}^n (e^{f_i(x_i)} \lambda_i(dx_i)),$$

and so this Gibbs measure is a product measure. In this special case we denote this Gibbs measure by $\xi_f$. Similarly, if $g \in C_b(K_i)$ for some $i$, then we denote by $\xi_g$ the Gibbs measure on $K_i$ constructed from $g$ and $\lambda_i$.

Consider again a general $f \in C_b(\prod_i K_i)$. In the main result of this paper, we assume that $f$ has ‘low complexity’ relative to the subspace $\Sigma_i C_b(K_i)$, and deduce a structure theorem for $\mu$ 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 [CD16] introducing the theory of ‘nonlinear large deviations’. Chatterjee and Dembo’s main result gave an approximation of the partition function $Z$ under such an assumption on $f$. More recently, works by Eldan [Eld] and Eldan and Gross [EG18] have uncovered the approximate structure of the measure $\mu$ 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 $x \in \prod_i K_i$, $i \in [n]$, and $y \in K_i$, we define

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

Thus, we replace the $i$th coordinate of $x$ 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 $i$th 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.

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

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

For a fixed point $x \in \prod_i K_i$, the function $\nabla f(x, \cdot)$ is a member of $\Sigma_i C_b(K_i)$. We refer to it as the discrete gradient of $f$ at the point $x$, and usually write $\nabla_x f$ instead of $\nabla f(x, \cdot)$. 
LOW-COMPLEXITY GIBBS MEASURES

If $f \in C_b(\prod_i K_i)$ has the property that $\nabla_x f$ is the same function for every $x \in \prod_i K_i$, then an easy exercise shows that $f$ lies in $\Sigma_i C_b(K_i)$. Beyond this case, we can make the weaker assumption that $f$ has relatively few different gradients $\nabla_x f$ as $x$ 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 $d_{K_i}$:

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

If each $K_i$ is just a finite alphabet $A$ 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

$$(3) \quad d_n(\mu, \nu) := \inf_\lambda \int d_n(x, y) \lambda(dx, dy) \quad (\mu, \nu \in \text{Prob}(\prod_i K_i)), $$

where $\lambda$ ranges over all couplings of $\mu$ and $\nu$. This is a standard and well-studied construction of a metric on probability measures: see, for instance, [Dud02, Section 11.8] for a textbook account.

The second concept we need is a parameter which measures the multi-variate correlation for a joint distribution on an $n$-fold product space. There are many such parameters, but the one we need is called ‘dual total correlation’ or ‘DTC’. It goes back to work of Han in information theory [Han75]. The recent paper [Ausb] studies DTC in some depth, and shows that a small value of DTC implies that the joint distribution is close in $d_n$ to a mixture of relatively few product measures. The relevance of DTC to the present paper is therefore not surprising, but it is interesting that the proof of our main theorem automatically yields an upper bound on the DTC of the Gibbs measure (1) as a by-product.

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

**Theorem (A).** Let $P$ be a partition of $\prod_i K_i$ with the property that

$$\|\nabla_x f - \nabla_y f\| < \delta_n$$

whenever $x, y$ lie in the same part of $P$. Then:
TIM AUSTIN

\[ a) \ \text{DTC}(\mu) < H_\mu(P) + \delta n, \text{ where } H_\mu(P) \text{ is the Shannon entropy of the partition } P \text{ according to } \mu, \]

\[ b) \text{ we have} \]

\[ \sum_{P \in P} \mu(P) \cdot \int D(\mu_{|P} \| \xi_{\nabla y f}) \mu_{|P}(dy) < H_\mu(P) + \delta n, \]

where \( D \) denotes Kullback–Leibler divergence, and

\[ c) \text{ we have} \]

\[ \sum_{P \in P} \mu(P) \cdot \int d_n(\mu_{|P}, \xi_{\nabla y f}) \mu_{|P}(dy) < \sqrt{\frac{1}{2}(H_\mu(P) + n + \delta)}. \]

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 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 concerns only parts (a) and (b).

The definition of \( \nabla_x f \) 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 \( \nabla'_x f \) for the alternative discrete gradients that result, then these different discrete gradients are related by the cocycle equation

\[ \nabla'_x f(y) = \nabla_x f(y) - \nabla_x f(*'_{1}, \ldots, *'_{n}). \]

As a result, if \( P \) satisfies (4) for the discrete gradients \( \nabla_x f \), then it satisfies the analogous property for the discrete gradients \( \nabla'_x f \) with \( \delta n \) 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 \( H_\mu(P) \) is small compared to \( n \). This is implied if we have enough control on \( |P| \) itself: for instance, if we know that

\[ \text{cov}_{\delta n} \left( \left\{ \nabla_x f : x \in \prod_i K_i \right\} , \| \cdot \| \right) \leq e^{\varepsilon n} \]

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

**Corollary (A').** If \( f \) satisfies (5), then there are (i) a partition \( P \) of \( \prod_i K_i \) into at most \( e^{\varepsilon n} \) parts and (ii) a selection of functions \( g_P \in \Sigma_i C_b(K_i) \) for \( P \in P \) such that

\[ \sum_{P \in P} \mu(P) \cdot D(\mu_{|P} \| \xi_{g_P}) < (\varepsilon + \delta)n \]
\[
\sum_{P \in \mathcal{P}} \mu(P) \cdot d_n(\mu|_P, \xi_{g_P}) < \sqrt{(\varepsilon + \delta)/2}.
\]

(Informally: ‘most of the mass of \(\mu\) 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 adds 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 [EG18] 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 subject of [CD16], and also re-appears in Eldan’s method [Eld].

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 [CD16] and Eldan and Gross [Eld, EG18] are confined to that case.

In [CD16], 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 \(\nabla f\) 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 [Yan]. 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 [Eld], 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, his ‘low-complexity’ assumption applies directly to the discrete gradient \(\nabla f\). But his approach uses the embedding \(\{-1, 1\}^n \subset \mathbb{R}^n\) in another very essential. 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 fairly directly to other product spaces. However, at some places Eldan’s estimates seem to be sharper than ours; we compare them more carefully in Subsection 3.3. In a subsequent work \([EG18]\) Eldan and Gross also give a characterization of which product measures appear in the structure theorem in terms of the function \( f \); Proposition 4.1 below is similar to their result.

Chatterjee and Dembo’s results in \([CD16]\) 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 \([Cha17]\), 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 \([LZ17, BGLZ17]\). Some related applications which require more than two-point alphabets are described by Yan in \([Yan]\). Applications with two-point alphabets are treated again by Eldan and Gross in \([Eld, EG18, EG]\), where they obtain refinements of some of the Chatterjee–Dembo estimates using Eldan’s new structural results. They also discuss a few other applications in probabilistic combinatorics. Another application discussed in \([Eld]\) is to large deviations of the number of arithmetic progressions in a random arithmetic set, and the more recent paper \([BGSZ]\) gives a more complete analysis of this problem by building on Eldan’s approach.

I do not know whether the results of the present paper 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 \([DC]\) achieves substantial improvements over the other works cited above by carefully finding and exploiting certain convex sets related to that problem. We do not discuss these issues further here.

Acknowledgements. I am grateful to Sourav Chatterjee, Amir Dembo and Ofer Zeitouni for some insightful conversations. Along with Ronen Eldan, they also made valuable suggestions about an early draft of this paper.

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
KL divergence can be defined for any pair of probability measures $\mu, \gamma$ on a general measurable space: $D(\mu \parallel \gamma)$ is infinite unless $\mu \ll \gamma$, and in that case it is defined by

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

(which may still be infinite). This generalization and its properties can be found in [Pin64, Chapters 2 and 3] or [DZ10, 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 measures $\mu, \gamma \in \text{Prob}(K)$, we have

$$D(\mu \parallel \gamma) = -H_\mu(\mathcal{P}) + \sum_{P \in \mathcal{P}} \mu(P) D(\mu|_P \parallel \gamma|_P).$$

**Proof.** The usual chain rule for KL divergence gives

$$D(\mu \parallel \gamma) = D([\mu]|_\mathcal{P} \parallel [\gamma]|_\mathcal{P}) + \sum_P \mu(P) D(\mu|_P \parallel \gamma|_P),$$

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

$$D(\mu|_P \parallel \gamma|_P) = \int_P \log \frac{d\mu|_P}{d\gamma|_P} d\mu|_P = \int_P \log \left( \gamma(P) \frac{d\mu|_P}{d\gamma} \right) d\mu|_P = D(\mu|_P \parallel \gamma) + \log \gamma(P).$$

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

$$D([\mu]|_\mathcal{P} \parallel [\gamma]|_\mathcal{P}) + \sum_P \mu(P) \log \gamma(P) = \sum_P \mu(P) \log \frac{\mu(P)}{\gamma(P)} + \sum_P \mu(P) \log \gamma(P) = -H_\mu(\mathcal{P}).$$

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

$$\mu(dx) := \frac{e^{f(x)}}{M_\lambda(f)} \lambda(dx), \quad \text{where } M_\lambda(f) := \int e^f d\lambda.$$

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 $\mu$ is the Gibbs measure associated to $f$, then

$$D(\mu \parallel \lambda) - \int f \, d\mu = -\log M_\lambda(f).$$

For any other $\nu \in \text{Prob}(K)$, we have

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

\[\square\]

If $K$ is finite and $\lambda$ is uniform, then (8) appears within the calculation in [CT06, equations (12.5)–(12.12)]. The same calculation gives the general case upon replacing $H(\cdot)$ with $-D(\cdot \parallel \lambda)$ throughout. In that generality it can be found in the proof of [DZ10, Lemma 6.2.13], and in the application of [Csi75, Equation (2.6)] in Section 3, case (A) of that paper.

2.3. Marton’s transportation-entropy inequality. A classical inequality of Marton [Mar86, Mar96] 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 inequality for product measures). If $\mu, \nu \in \text{Prob}(\prod_i K_i)$ and $\nu$ is a product measure then

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

\[\square\]

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, one simply changes the constants in the inequality.) See [Mar96, Mar98, Dem97, Sam00] for various extensions of Proposition 2.3, and [Tal96] for the 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

$$\text{DTC}(\xi_1, \ldots, \xi_n) := H(\xi_1, \ldots, \xi_n) - \sum_{i=1}^n H(\xi_i \mid \xi_1, \ldots, \xi_{i-1}, \xi_{i+1}, \ldots, \xi_n).$$
For more general random variables valued in measurable spaces \( K_1, \ldots, K_n \), their DTC is

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

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 \( \xi_i \) 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 [Han78, Theorem 4.1] includes the fact that DTC is always non-negative. DTC is studied carefully in [Ausb], 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 \( \lambda_i \in \text{Prob}(K_i) \) as in the Introduction. The next lemma is part (b) of [Ausb, Proposition 6.1] (along with the first remark following its proof in that paper).

**Lemma 2.4.** If \( D(\mu \| \lambda_1 \times \cdots \times \lambda_n) < \infty \) then

\[
DTC(\mu) = \sum_i \int D(\mu_{i,z} \| \lambda_i) \mu_{[n] \setminus i}(dz) - D(\mu \| \lambda_1 \times \cdots \times \lambda_n),
\]

where

- \( \mu_{[n] \setminus i} \) is the projection of \( \mu \) 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 \( \mu \) of the \( i \)th coordinate given the other coordinates.

In case \( \mu \) is the Gibbs measure of \( f \), this lemma turns into a simple expression for \( DTC(\mu) \) in terms of \( \nabla f \). It plays a crucial role later in the paper.

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

(9) \[
DTC(\mu) = \int D(\xi_{\nabla f} \| \lambda_1 \times \cdots \times \lambda_n) \mu(dx) - D(\mu \| \lambda_1 \times \cdots \times \lambda_n).
\]

**Proof.** The additivity of KL divergence for product measures gives

\[
\int D(\xi_{\nabla f} \| \lambda_1 \times \cdots \times \lambda_n) \mu(dx) = \sum_{i=1}^n \int D(\xi_{\partial_i f(x, \cdot)} \| \lambda_i) \mu(dx).
\]

Using this, the right-hand side of (9) matches the formula for \( DTC(\mu) \) in Lemma 2.4, because \( \xi_{\partial_i f(x, \cdot)} \) is precisely the conditional distribution of \( x_i \) under \( \mu \) given the other coordinates \( x_{[n] \setminus i} \).
The fact that we really need in the proof of Theorem A is that the difference on the right-hand side of (9) is non-negative. However, Theorem A also provides a bound on \( DTC(\mu) \). This actually offers an alternative route to a decomposition of \( \mu \) into near-product measures: the main result of [Ausb] 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 \( \mu \) and discrete gradient \( \nabla f \) 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

\[
\int \nabla_x f(x) \mu(dx) = \int \int \nabla_x f(y) \xi \nabla_x f(dy) \mu(dx).
\]

**Proof.** It suffices to prove that

\[
\int \partial_i f(x,x_i) \mu(dx) = \int \int \partial_i f(x,y) \xi \partial_i f(x,\cdot) (dy) \mu(dx),
\]

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

\[
\int \partial_i f(x,y) \xi \partial_i f(x,\cdot) (dy) = E_{\mu} [\partial_i f(x,x_i) | x_{[n]\setminus i}],
\]

and (10) is the tower property of iterated conditional expectations.

To prove Theorem A, we must bound both \( DTC(\mu) \) and also the expression

\[
\sum_P \mu(P) \int \eta_n(\mu|P,\xi \nabla y f) \mu|P(dy).
\]

The proof focuses on this expression, but naturally yields a bound on \( DTC(\mu) \) as well.

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

\[
D(\nu \parallel \prod_{i \in S} \lambda_i) \text{ to } D(\nu)
\]

whenever \( S \subseteq [n] \) and \( \nu \in \text{Prob}(\prod_{i \in S} K_i) \).
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 \( \nabla_x f \) may be identified with a vector in \( \mathbb{R}^n \): its \( i \)th entry is

\[
(12) \quad f(x_1, x_2, \ldots, x_{i-1}, 1, x_{i+1}, \ldots, x_n) - f(x_1, x_2, \ldots, x_{i-1}, 0, x_{i+1}, \ldots, x_n).
\]

Conversely, any vector \( u \in \mathbb{R}^n \) defines a linear function on \( \{0, 1\}^n \) using the Euclidean inner product: \( u(y) := \sum_i u_i y_i \). If \( u \) is the vector given by \((12)\), then \( u(y) \) agrees with our earlier definition of \( \nabla_x f(y) \) if we choose the reference points \( *_i = 0 \) for each \( i \).

To simplify this sketch further, let us also imagine that the discrete gradient \( \nabla_x f \) takes only two distinct values in \( \mathbb{R}^n \), say \( u \) and \( v \). (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

\[
\mathcal{P} := \{ x : \nabla_x f = u \} \quad \text{and} \quad \mathcal{P}^c := \{ x : \nabla_x f = v \}.
\]

Assume further that neither of the values \( \mu(\mathcal{P}) \), \( \mu(\mathcal{P}^c) \) 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 \( D(\mu | \mu) \parallel \xi_u) \) and \( D(\mu | \mu) \parallel \xi_v) \) are both small relative to \( n \).

Here are the steps in the proof:

- Gibbs’ variational principle lets us interpret this requirement another way: \( \mu | \mathcal{P} \) must come close to saturating the variational inequality that characterizes the product measure \( \xi_u \), meaning that

\[
(13) \quad D(\mu | \mathcal{P}) - \int u \, d\mu | \mathcal{P} \quad \text{is not much larger than} \quad D(\xi_u) - \int u \, d\xi_u
\]

relative to \( n \), and similarly when comparing \( \mu | \mathcal{P}^c \) with \( \xi_u \).

- For each of \( \mu | \mathcal{P} \) and \( \mu | \mathcal{P}^c \) separately, we have no clear way to control the gap in \( (13) \). But we can control the average of those gaps over \( \mathcal{P} \) and \( \mathcal{P}^c \), keeping in mind that \( \nabla_x f \) equals \( u \) on \( \mathcal{P} \) and \( v \) on \( \mathcal{P}^c \) respectively. By Lemma 2.1, the average of the KL divergences is

\[
\mu(\mathcal{P}) \cdot D(\mu | \mathcal{P}) + \mu(\mathcal{P}^c) \cdot D(\mu | \mathcal{P}^c) = D(\mu) + H_\mu(\mathcal{P}),
\]

and by Corollary 2.5 this is equal to

\[
\left[ \int D(\xi_{\nabla_x f}) \mu(dx) - DTC(\mu) \right] + H_\mu(\mathcal{P})
= \mu(\mathcal{P}) \cdot D(\xi_u) + \mu(\mathcal{P}^c) \cdot D(\xi_v) - DTC(\mu) + H_\mu(\mathcal{P}).
\]
The average of the required integrals is

\[ \mu(P) \int u \, d\mu_P + \mu(P^c) \int v \, d\mu_{P^c} \]

\[ = \int_P u \, d\mu + \int_{P^c} v \, d\mu \]

\[ = \int \nabla_x f(x) \, \mu(dx) \]

\[ = \iint \nabla_x f \, d\xi \nabla_x f \, \mu(dx) \]

\[ = \mu(P) \int u \, d\xi_u + \mu(P^c) \int v \, d\xi_v. \]

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

\[ \mu(P) \left[ D(\mu_P | P) - \int u \, d\mu_P \right] - \mu(P) \left[ D(\xi_u) - \int u \, d\xi_u \right] \]

\[ + \mu(P^c) \left[ D(\mu_{P^c} | P^c) - \int v \, d\mu_{P^c} \right] - \mu(P^c) \left[ D(\xi_v) - \int v \, d\xi_v \right] \]

\[ = -\text{DTC}(\mu) + H_\mu(P). \]

- Since DTC is non-negative, the last line above is at most \( H_\mu(P) \leq \log |P| = \log 2 \). This is indeed small compared to \( n \), and hence so are both of those KL divergences. 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 \( \text{DTC}(\mu) \leq \log 2 \).

Now we give the careful proof in full.

**Proof of Theorem A.** *Step 1: formula for the average KL divergence.* Consider a single \( P \in \mathcal{P} \). For any \( g \in \Sigma_1 C_b(K) \), Proposition 2.2 gives

\[ D(\mu_P \| \xi_g) = \left[ D(\mu_P) - \int g \, d\mu_P \right] - \left[ D(\xi_g) - \int g \, d\xi_g \right]. \]

Let us average this identity over \( g = \nabla_x f \) where \( x \sim \mu_P \). The result is

\[ \int D(\mu_P \| \nabla_x f) \, \mu_P(dx) \]

\[ = D(\mu_P) - \iint \nabla_x f(y) \, \mu_{P}(dy) \, \mu_P(dx) \]

\[ - \int D(\nabla_x f) \, \mu_P(dx) + \iint \nabla_x f(y) \, \xi_{\nabla_x f}(dy) \, \mu_P(dx). \]
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:

$$\sum_P \mu(P) \int D(\xi \nabla_x f) \mu|_P(dx) = \int D(\xi \nabla_x f) \mu(dx).$$

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

$$\sum_P \mu(P) \int D(\mu \parallel \xi \nabla_y f) \mu|_P(dy)$$

$$= \sum_P \mu(P) \cdot D(\mu|_P) - \sum_P \mu(P) \int \int \nabla_x f(y) \mu|_P(dy) \mu|_P(dx)$$

(14)$$- \int D(\xi \nabla_x f) \mu(dx) + \int \int \nabla_x f(y) \xi \nabla_x(dy) \mu(dx).$$

Next we bound the the right-hand side of (14) by considering separately (i) the KL-divergence terms and (ii) the double-integral terms.

**Step 2: difference of the KL-divergence terms.** By Lemma 2.1 and Corollary 2.5, we have

$$\sum_P \mu(P) \cdot D(\mu|_P) - \int D(\xi \nabla_x f) \mu(dx) = [D(\mu) + H_{\mu}(\mathcal{P})] - [D(\mu) + \text{DTC}(\mu)]$$

$$= H_{\mu}(\mathcal{P}) - \text{DTC}(\mu).$$

**Step 3: difference of the double-integral terms.** Using the crucial Lemma 3.1 to substitute for the second double integral in (14), and then using the law of total probability, the difference of those double-integral terms is equal to

$$- \sum_P \mu(P) \int \int \nabla_x f(y) \mu|_P(dy) \mu|_P(dx) + \int \nabla_y f(y) \mu(dy)$$

$$= \sum_P \mu(P) \int \int (\nabla_y f(y) - \nabla_x f(y)) \mu|_P(dy) \mu|_P(dx).$$

(15)$$- \sum_P \mu(P) \int \int \nabla_x f(y) \mu|_P(dy) \mu|_P(dx) + \int \nabla_y f(y) \mu(dy)$$

$$= \sum_P \mu(P) \int \int (\nabla_y f(y) - \nabla_x f(y)) \mu|_P(dy) \mu|_P(dx).$$

By assumption, whenever $x$ and $y$ lie in the same cell of $\mathcal{P}$, we have

$$\|\nabla_x f - \nabla_y f\| < \delta n$$

and hence also

$$|\nabla_y f(y) - \nabla_x f(y)| < \delta n.$$

Substituting this fact into (15), we obtain that the difference of those double integrals is less than $\delta n$. 

(16)$$||\nabla_x f - \nabla_y f|| < \delta n$$

and hence also

$$|\nabla_y f(y) - \nabla_x f(y)| < \delta n.$$
Step 4: completion. Inserting the calculations from Steps 2 and 3 back into (14) and re-arranging, we are left with

\[(17) \quad \text{DTC}(\mu) + \sum_{P} \mu(P) \int \text{D}(\mu|P \parallel \xi_{\nabla_y f}) \mu_P(d\mathbf{y}) < H_\mu(P) + \delta n.\]

Both terms on the left are non-negative, 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:

\[
\sum_{P} \mu(P) \int \text{d}_n(\mu|P, \xi_{\nabla_y f}) \mu_P(d\mathbf{y}) \leq \sum_{P} \mu(P) \int \sqrt{\frac{1}{2n} \text{D}(\mu|P \parallel \xi_{\nabla_y f})} \mu_P(d\mathbf{y}) \leq \sqrt{\sum_{P} \mu(P) \int \frac{1}{2n} \text{D}(\mu|P \parallel \xi_{\nabla_y f}) \mu_P(d\mathbf{y})}.
\]

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 $\lambda_i$ are highly biased, for instance when $\lambda_i\{1\} = p \ll 1$ for each $i$?

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

**Proof of Corollary A’ from Theorem A.** Let $Q$ be a partition of

\[\left\{\nabla_x f : x \in \prod_i K_i\right\}\]

into sets of $\| \cdot \|$-diameter less than $\delta n$, and let $P$ be the pullback of $Q$ under the map $x \mapsto \nabla_x f$. Now pick an element $y_P$ in each $P \in P$ which minimizes $\text{d}_n(\mu|P, \xi_{\nabla_y f})$, and set $g_P := \nabla_{y_P} f$. 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. \qed
3.1. Possible variations on Theorem A. If one follows the proof of Theorem A as far as inequality (17), except without substituting the bound (16) into (15), then one arrives at an identity instead:

\[
\text{DTC}(\mu) + \sum_P \mu(P) \int D(\mu_{|P} \parallel \xi \nabla_y f) \mu_{|P}(dy) \\
= H_\mu(P) + \sum_P \mu(P) \int \int (\nabla_y f(y) - \nabla_x f(y)) \mu_{|P}(dy) \mu_{|P}(dx).
\]

In the proof of Theorem A, our assumptions on \( \mathcal{P} \) imply that both right-hand terms are small simultaneously, and this yields the desired bounds on both left-hand terms because they are non-negative.

As Amir Dembo has emphasized to me, the specific identity (18) may have other interesting 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 \( \nabla_y f(y) - \nabla_x f(y) \). Are there other ways to bound this term, perhaps by exploiting further the special structure of the Gibbs measure \( \mu \)? The reward could be a similar result to Theorem A but with weaker hypotheses.

In a different direction, one could try replacing the use of Marton’s inequality (Proposition 2.3) in Theorem A with a different transportation-entropy inequality. In [Dem97], Dembo proves several analogs of Proposition 2.3 that replace \( d_n \) with various alternative transportation-like quantities. Those analogs then yield some powerful variations on product-space measure concentration that were previously discovered and used by Talagrand [Tal95]. I do not know whether a version of Theorem A part (c) for one of those alternative transportation-like quantities could be useful.

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 [Ausb, Theorem A], conclusion (a) by itself implies something like conclusion (c). To be precise, if \( \mu \) is a probability measure on a product space \( \prod_i K_i \) for which \( \text{DTC}(\mu) \leq r^3 n \), then [Ausb, Theorem A] asserts that \( \mu \) can be a represented as a mixture

\[
\mu = \int_L \mu_y \nu(dy)
\]

whose mutual information is at most \( \text{DTC}(\mu) \) and such that there is a measurable family \( (\xi_y : y \in L) \) of product measures on \( \prod_i K_i \) satisfying

\[
\int_L d_n(\mu_y, \xi_y) \nu(dy) < 2r.
\]
In our setting, if we know that
\[ H_\mu(p) + \delta n \leq r^3 n, \]
then this gives such a mixture representation of \( \mu \) with mutual information at most \( H_\mu(p) + \delta n \). 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 [Ausa, 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, \( C \) is used to stand for several universal constants which we do not compute explicitly. Its value may change from one appearance to the next.

In [Eld], Eldan considers a Gibbs measure \( \mu \) as in (1) on \( \{-1, 1\}^n \). In this case, each discrete gradient \( \nabla_x f \) may be identified with a vector in \( \mathbb{R}^n \); this is similar to (12) except using 1 and \(-1\) instead of 1 and 0. (Actually 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:
\[ D(f) := GW\left(\{\nabla_x f : x \in \{-1, 1\}^n\} \cup \{0\}\right). \]

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

Let us write \( \gamma := D(f)/n \) for convenience. Specifically, after fixing \( \varepsilon > 0 \) and also picking an auxiliary parameter \( \alpha > 1 \), [Eld, Theorem 3] provides a mixture
\[ \mu = \int \mu_\theta m(d\theta) \]
over some space of parameters \( \theta \) such that
\begin{enumerate}
  \item \( H(\mu) - \int H(\mu_\theta) m(d\theta) < C \varepsilon n \) and
  \item there is a subset \( \Theta \) of values of \( \theta \) such that \( m(\Theta) > 1 - \frac{1}{n} - \frac{1}{\alpha} \) and such that every \( \theta \in \Theta \) satisfies
    \[ d_n(\mu_\theta, \xi_{v_\theta}) \leq C \sqrt{\alpha \gamma / \varepsilon} \]
\end{enumerate}
for some product measure \( \xi_{v_\theta} \).
Eldan’s theorem gives other conclusions as well. It derives bounds on the vectors $v_\theta$; and the ingredients in Eldan’s mixture are ‘tilts’ of $\mu$, 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 (see, for instance, the proof of [Eld, Theorem 1]).

By Sudakov’s minoration [LT11, Theorem 3.18] and the inequality $\| \cdot \|_1 \leq \sqrt{n} \| \cdot \|_2$, the covering numbers of a nonempty subset $K \subseteq \mathbb{R}^n$ are related to Gaussian width according to

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

Also, our identification of each discrete gradient with a vector in $\mathbb{R}^n$ satisfies $\| \nabla_x f \| \leq \| \nabla_x f \|_1$, where $\| \cdot \|$ denotes the uniform norm for functions on $\{-1, 1\}^n$ as before. Therefore, with $\gamma$ as above, we have

$$\log \text{cov}_{\delta n}(\{ \nabla_x f : x \in \{-1, 1\}^n \}, \| \cdot \|) \leq C \frac{D(f)^2}{\delta^2 n} = C \frac{\gamma^2}{\delta^2 n} \quad \forall \delta > 0. \quad (19)$$

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

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

$$H(\mu) - \sum_P \mu(P) H(\mu|_P) = H_\mu(P) \leq C \varepsilon n \quad (20)$$

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

$$\sum_P \mu(P) \cdot d_{\mu|_P}(\mu|_P, \xi_{g_P}) \leq C \sqrt{\varepsilon + \frac{\gamma}{\sqrt{\varepsilon}}} \leq C \max \left\{ \sqrt{\varepsilon}, \sqrt{\gamma/\sqrt{\varepsilon}} \right\}. \quad (21)$$

If $\varepsilon \leq \gamma^{2/3}$ 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 $\varepsilon$ is very small.

However, if $\varepsilon > \gamma^{2/3}$, then the right-hand side of (21) is $C \sqrt{\varepsilon}$. In this case we may reduce $\varepsilon$ (equivalently, increase $\delta$) 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.
4. A characterization of the terms in the mixture. In this section, we suppose that the discrete gradient function

$$\prod_i K_i \rightarrow \Sigma_i C_b(K_i) : x \mapsto \nabla_x f$$

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

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

be the Lipschitz constant of the discrete gradient (not of $f$ itself).

Let $\mathcal{P}$ be a partition of $\prod_i K_i$ as in Corollary A', and let $y_P \in P$ for $P \in \mathcal{P}$ be a selection of points which minimize the distances $d_n(\mu_P, \xi_{\nabla y_P f})$. Recall that the functions $g_P$ of Corollary A' are then given by $g_P := \nabla y_P f$.

**Proposition 4.1.** These data satisfy

$$\sum_P \mu(P) \| g_P - \int \nabla_x f \xi_{\nabla g_P}(dx) \| \leq \delta n + L \sqrt{\frac{1}{2} \left( \frac{H_{\mu}(P)}{n} + \delta \right)}.$$  

Thus, most of the probability vector $(\mu(P) : P \in \mathcal{P})$, when transferred to the list of product measures $(\xi_{g_P} : P \in \mathcal{P})$, is on terms that satisfy the 'approximate fixed point' equation

$$g_P \approx \int \nabla f \, d\xi_{g_P}.$$  

**Proof.** Since $y_P \in P$, our assumption (4) about $\mathcal{P}$ gives

$$\| g_P - \int \nabla_x f \mu_{\mu_P}(dx) \| < \delta n \quad \forall P \in \mathcal{P}.$$  

On the other hand, the definition of $L$ gives

$$\left\| \int \nabla_x f \mu_{\mu_P}(dx) - \int \nabla_x f \xi_{\nabla g_P}(dx) \right\| \leq L \cdot d_n(\mu_{\mu_P}, \xi_{g_P}) \quad \forall P \in \mathcal{P}.$$  

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

In the special case of measures on $\{-1, 1\}^n$, we can derive from Proposition 4.1 a bound in terms of the Gaussian-width gradient complexity $D(f)$ studied by Eldan. Let $\gamma := D(f)/n$, as previously; consider an auxiliary parameter $\delta > 0$; and
choose a partition $P$ as promised by inequality (19) for this $\delta$. Then (20) lets us turn Proposition 4.1 into

$$\sum_P \mu(P) \left\| g_P - \int \nabla_x f \xi_{g_P}(dx) \right\| \leq \delta n + CL \sqrt{\frac{\gamma^2}{\delta^2}} + \delta \quad \forall \delta > 0$$

(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 $\gamma$ is small, we are left with

$$\sum_P \mu(P) \left\| g_P - \int \nabla_x f \xi_{g_P}(dx) \right\| \leq \gamma^{2/3} n + CL \gamma^{1/3} \leq CL \gamma^{1/3}.$$  

A result of this flavour is obtained by Eldan and Gross in [EG18, 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 $\nabla f$), 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 $\int e^f d\lambda$. By Proposition 2.2, it satisfies

$$\log \int e^f d\lambda \geq \int f d\nu - D(\nu \parallel \lambda)$$

for any probability measure $\nu$ 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 $\nu$ 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 $d$, and satisfies (5), then

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

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

**Remark.** In case each $K_i$ is finite and $\lambda_i$ is uniform, standard manipulations turn (23) into

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

This is the relevant form for many applications of nonlinear large deviations.
Proof. Let $P$ be the partition implied by (5). Gibbs’ identity and Corollary 2.5 give
\[
\log \int e^f \, d\lambda = \int f \, d\mu - D(\mu \| \lambda)
\]
\[
= \int f \, d\mu + \text{DTC}(\mu) - \int D(\xi_{V_y} \| \lambda) \, \mu(dy).
\]
(24)

Using part (b) of Theorem A, we have
\[
\int f \, d\mu = \sum_P \mu(P) \int f \, d\mu_P
\leq \int \int f \, d\xi_{V_y} \mu(dy) + L \sum_P \mu(P) \int d_n(\mu_P, \xi_{V_y}) \mu(dy)
\]
\[
\leq \int \int f \, d\xi_{V_y} \mu(dy) + \sqrt{\frac{\varepsilon + \delta}{2}} L.
\]
(25)

Inserting this bound and also part (a) of Theorem A into (24), we obtain
\[
\log \int e^f \, d\lambda \leq \int \left[ \int f \, d\xi_{V_y} - D(\xi_{V_y} \| \lambda) \right] \mu(dy) + (\varepsilon + \delta)n + \sqrt{\frac{\varepsilon + \delta}{2}} L.
\]

Now we bound the integral over $y$ by the supremum over all product measures $\xi$, as in (23).

Partition-function estimation was the heart of the original paper [CD16] 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 [Yan], 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 [Eld, Corollary 2]. One can compare Proposition 5.1 with Eldan’s results using the ideas from Subsection 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 Subsection 3.3 above. In the notation of that discussion, the appropriate choice of $\varepsilon$ for the proof of [Eld, Corollary 2] is
\[
C \left( \frac{L}{n} \right)^{2/3} \gamma^{1/3},
\]
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 Subsection 3.3.
REFERENCES

[Ausa] Tim Austin. Measure concentration and the weak Pinsker property. To appear. *Publ. Math. Inst. Hautes Études Sci*. Preprint available online at arXiv.org: 1705.00302.

[Ausb] Tim Austin. Measures of correlation and mixtures of product measures. Preprint, available online at arXiv.org: 1809.10272.

[BGLZ17] Bhaswar B. Bhattacharya, Shirshendu Ganguly, Eyal Lubetzky, and Yufei Zhao. Upper tails and independence polynomials in random graphs. *Adv. Math.*, 319:313–347, 2017.

[BGSZ] Bhaswar B. Bhattacharya, Shirshendu Ganguly, Xuancheng Shao, and Yufei Zhao. Upper tail large deviations for arithmetic progressions in a random set. Preprint, available online at arXiv.org: 1605.02994.

[CD16] Sourav Chatterjee and Amir Dembo. Nonlinear large deviations. *Adv. Math.*, 299:396–450, 2016.

[Cha17] Sourav Chatterjee. *Large deviations for random graphs*, volume 2197 of Lecture Notes in Mathematics. Springer, Cham, 2017. Lecture notes from the 45th Probability Summer School held in Saint-Flour, June 2015, École d’Été de Probabilités de Saint-Flour. [Saint-Flour Probability Summer School].

[Csi75] I. Csiszár. I-divergence geometry of probability distributions and minimization problems. *Ann. Probability*, 3:146–158, 1975.

[CT06] Thomas M. Cover and Joy A. Thomas. *Elements of information theory*. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, second edition, 2006.

[DC] Amir Dembo and Nicholas Cook. Large deviations of subgraph counts in Erdős–Rényi random graphs. Preprint, available online at arXiv.org: 1809.11148.

[Dem97] Amir Dembo. Information inequalities and concentration of measure. *Ann. Probab.*, 25(2):927–939, 1997.

[Dud02] R. M. Dudley. *Real analysis and probability*, volume 74 of Cambridge Studies in Advanced Mathematics. Cambridge University Press, Cambridge, 2002. Revised reprint of the 1989 original.

[DZ10] Amir Dembo and Ofer Zeitouni. *Large deviations techniques and applications*, volume 38 of Stochastic Modelling and Applied Probability. Springer-Verlag, Berlin, 2010. Corrected reprint of the second (1998) edition.

[EG] Ronen Eldan and Renan Gross. Exponential random graphs behave like mixtures of stochastic block models. Preprint, available online at arXiv.org: 1707.01227.

[EG18] Ronen Eldan and Renan Gross. Decomposition of mean-field Gibbs distributions into product measures. *Electron. J. Probab.*, 23:Paper No. 35, 24, 2018.

[Eld] Ronen Eldan. Gaussian-width gradient complexity, reverse log-sobolev inequalities and nonlinear large deviations. Preprint, available online at arXiv.org: 1612.04346.

[Han75] Té Sun Han. Linear dependence structure of the entropy space. *Information and Control*, 29(4):337–368, 1975.

[Han78] Té Sun Han. Nonnegative entropy measures of multivariate symmetric correlations. *Information and Control*, 36(2):133–156, 1978.

[LT11] Michel Ledoux and Michel Talagrand. *Probability in Banach spaces*. Classics in Mathematics. Springer-Verlag, Berlin, 2011. Isoperimetry and processes, Reprint of the 1991 edition.

[LZ17] Eyal Lubetzky and Yufei Zhao. On the variational problem for upper tails in sparse random graphs. *Random Structures Algorithms*, 50(3):420–436, 2017.

[Mar86] K. Marton. A simple proof of the blowing-up lemma. *IEEE Trans. Inform. Theory*, 32(3):445–446, 1986.
[Mar96] K. Marton. Bounding $d$-distance by informational divergence: a method to prove measure concentration. *Ann. Probab.*, 24(2):857–866, 1996.

[Mar98] Katalin Marton. Measure concentration for a class of random processes. *Probab. Theory Related Fields*, 110(3):427–439, 1998.

[Pin64] M. S. Pinsker. *Information and information stability of random variables and processes*. Holden-Day, Inc., San Francisco-California, 1964.

[Sam00] Paul-Marie Samson. Concentration of measure inequalities for Markov chains and $\Phi$-mixing processes. *Ann. Probab.*, 28(1):416–461, 2000.

[Tal95] Michel Talagrand. Concentration of measure and isoperimetric inequalities in product spaces. *Inst. Hautes Études Sci. Publ. Math.*, (81):73–205, 1995.

[Tal96] M. Talagrand. Transportation cost for Gaussian and other product measures. *Geom. Funct. Anal.*, 6(3):587–600, 1996.

[Yan] Jun Yan. Nonlinear large deviations: beyond the hypercube. Preprint, available online at arXiv.org:1703.08887.