A transportation approach to the mean-field approximation

Fanny Augeri

Received: 11 May 2019 / Revised: 19 January 2021 / Accepted: 12 April 2021 / Published online: 3 May 2021
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2021

Abstract
We develop transportation-entropy inequalities which are saturated by measures such that their log-density with respect to the background measure is an affine function, in the setting of the uniform measure on the discrete hypercube and the exponential measure. In this sense, this extends the well-known result of Talagrand in the Gaussian case. By duality, these transportation-entropy inequalities imply a strong integrability inequality for Bernoulli and exponential processes. As a result, we obtain on the discrete hypercube a dimension-free mean-field approximation of the free energy of a Gibbs measure and a nonlinear large deviation bound with only a logarithmic dependence on the dimension. Applied to the Ising model, we deduce that the mean-field approximation is within $O(\sqrt{n}\|J\|_2)$ of the free energy, where $n$ is the number of spins and $\|J\|_2$ is the Hilbert–Schmidt norm of the interaction matrix. Finally, we obtain a reverse log-Sobolev inequality on the discrete hypercube similar to the one proved recently in the Gaussian case by Eldan and Ledoux.

Keywords Gibbs measures · Mean-field approximation · Transportation inequalities · Large deviations · Ising model

Mathematics Subject Classification 60F10 · 60E15 · 82B44

1 Introduction

A fundamental question in statistical Physics is to understand the behavior of Gibbs measures, in particular through the computation of their free energy. If $\mu^n$ is the uniform measure on the discrete hypercube $\{-1, 1\}^n$ and $f : \{-1, 1\}^n \to \mathbb{R}$ is a function, called the potential, one can consider the Gibbs measure associated to $f$,
defined as the probability measure
\[ \nu = \frac{1}{Z_f} e^f d\mu^n, \]
where \( Z_f = \int e^f d\mu^n \) is the partition function of \( \nu \). The logarithm of the partition function is called the free energy. The free energy encodes a rich information on the Gibbs measure. Unfortunately, the free energy is generally an intractable quantity, which in turn motivates the search for a meaningful large \( n \) approximation. The Gibbs variational principle (see [10, Lemma 6.2.13]) asserts that the free energy admits the following variational form:

\[
\log \int e^f d\mu^n = \sup_{\nu} \left\{ \int f d\nu - H(\nu|\mu^n) \right\},
\]

where the supremum runs over all probability measures \( \nu \) on \( \{-1, 1\}^n \), and \( H(\nu|\mu^n) \) denotes the relative entropy between \( \nu \) and \( \mu^n \). The mean-field approximation consists in restricting the above supremum over the special class of product probability measures (or more generally tilted measures, that is, measures whose log-density with respect to the background measure is an affine function). As product probability measures on the discrete hypercube are parametrized by their mean, the mean-field approximation reduces an optimization problem on probability measures on \( \{-1, 1\}^n \) into an optimization problem on \( [-1, 1]^n \), which can much more tractable. The question is then to understand under which condition on the potential \( f \), the mean-field approximation can be justified rigorously. The Gibbs variational principle implies that the mean-field approximation always gives a lower bound on the free energy, that is

\[
\log \int e^f d\mu^n \geq \sup_{y \in [-1,1]^n} \left\{ \int f d\mu_y - I(y) \right\},
\]

where \( \mu_y \) is the product measure on \( \{-1, 1\}^n \) with barycenter \( y \), and \( I(y) = H(\mu_y|\mu^n) \). Another way to reformulate the accuracy of the mean-field approximation is to say that the above inequality is approximately tight in the large \( n \) limit. Our main task in the present work will be to obtain quantitative upper bounds.

In a seminal paper [7], Chatterjee and Dembo showed that given a smooth extension of the potential \( f \) to the hypercube \( [-1, 1]^n \), the mean-field approximation is accurate if the set of gradients of \( f \) is of low complexity in a \( \ell^2 \)-metric entropy sense. However, the quantitative error bound from the mean-field approximation they obtained is rather intricate, and involves in particular the partial derivatives of \( f \) up to the second order.

In the case of the Ising model, where the potential \( f \) is a quadratic form
\[
\forall x \in \{-1, 1\}^n, \quad f(x) = \langle x, Jx \rangle,
\]
given in terms of an interaction matrix \( J \), the convergence of the free energy to the mean-field approximation has been shown in the context of dense graphs using the graphon framework in [5,6]. For general graphs, a first breakthrough was made by
Basak and Mukherjee [3] who showed the accuracy of the mean-field approximation under the condition that $||J||_2 = o(\sqrt{n})$, denoting by $||J||_2$ the Hilbert–Schmidt norm of $J$. As a result, they showed the universality of the limiting partition function for asymptotically regular graphs. However, this result does not give any information about the speed of convergence. A quantitative error bound from the mean-field approximation in $O((n||J||_2)^{\frac{2}{3}})$ up to a logarithmic factor was derived by Jain, Koehler and Mossel [16] using the Frieze–Kannan regularity lemma.

Another approach to this problem goes through the decomposition of the Gibbs measure itself into a mixture of measures where the coordinates are weakly correlated. This line of research was exploited by Jain et al. in [17] to remove the logarithmic correction in the mean-field approximation for the Ising model, and showed that,

$$\log\int e^{\langle x, Jx \rangle} d\mu^n(x) = \sup_{y \in [-1,1]^n} \{\langle y, Jy \rangle - I(y)\} + O((n||J||_2)^{\frac{2}{3}}).$$

(3)

In [11], Eldan proved structural theorems for general Gibbs measures on Gaussian space and for the discrete hypercube, which was refined later by Eldan and Gross [13]. He deduced an upper bound on the free energy where the complexity of the discrete gradient of the potential is assessed in terms of its Gaussian mean-width, namely,

$$g(V) = \mathbb{E} \sup_{\xi \in V} \langle \xi, \Gamma \rangle,$$

(4)

where $V = \nabla f((-1,1)^n)$ is the set of discrete gradients of $f$ and $\Gamma$ is a standard Gaussian random variable in $\mathbb{R}^n$. His approximation of the free energy [11, Corollary 2] takes the form,

$$\log\int e^f d\mu^n = \sup_{y \in [-1,1]^n} \left\{ \int f d\mu_y - I(y) \right\} + O\left(\text{Lip}(f)^{\frac{2}{3}} g(V)^{\frac{1}{3}} n^{\frac{2}{3}}\right),$$

where Lip($f$) is the Lipschitz constant of $f$ with respect to the Hamming metric. This approach was further developed by Austin [2] who proved a structure theorem for Gibbs measures on general product spaces and deduced a mean-field approximation of the free energy.

In [1], the author proved a mean-field approximation for Gibbs measures with respect to general compactly supported background measures which, using Sudakov minoration, implies (see [1, Remark 1.4]) in the case of the discrete hypercube that

$$\log\int e^f d\mu^n = \sup_{y \in [-1,1]^n} \left\{ \int f d\mu_y - I(y) \right\} + O\left(g(V)^{\frac{2}{3}} n^{\frac{1}{3}}\right).$$

In particular, this bound enables one to recover the bound (3) for the Ising model. In the present paper, we will remove the dimension dependence from the above estimate, and prove the dimension-free estimate,
\[
\log \int e^f \, d\mu^n = \sup_{y \in [-1, 1]^n} \left\{ \int f \, d\mu_y - I(y) \right\} + O(b(V)), \tag{5}
\]

where \(b(V) = \mathbb{E} \sup_{\xi \in V} \langle \xi, X \rangle\), and \(X\) is uniformly distributed on \([-1, 1]^n\).

In a recent work [12], Eldan proved a new decomposition theorem which allowed him to show in the case of the Ising model that for any \(p > 0\), the error on the free energy induced by the mean-field approximation is \(O\left(\frac{1+p}{p} (n||J||_p) \frac{1}{\sqrt{p}}\right)\), where \(|| \cdot ||_p\) denotes the \(p\)-Schatten norm. This bound recovers for \(p = 2\) the previous \(O((n||J||_2)^{\frac{3}{2}})\) error shown by Jain, Koehler and Risteski in [17], and can significantly improve upon this bound by an appropriate choice of \(p\).

The goal of this paper is to propose a transportation approach to justify the mean-field approximation of the free energy of Gibbs measures in the specific case of the discrete hypercube. The main interest of this approach is that it allows us to derive an approximation which is dimension-free, i.e (5).

We develop new transportation-entropy inequalities in the case of the Bernoulli and the exponential distribution. Originally, the transportation-entropy inequalities were put forward by Marton [20] and Talagrand [23]. They appear to have strong connections with concentration inequalities (see [18, Chapter 6], [9], [21, Chapter 8], or [15, section 4]). They also have many links with other functional inequalities. Quadratic transportation-entropy inequalities are known to imply a spectral gap inequality by [4, section 4.1] (see also [15, section 8.3]), and are weaker that logarithmic Sobolev inequalities by the result of Otto and Villani [22] (see also [4]).

The main feature of the transportation-entropy inequalities we will present is that they are saturated by tilts of the background measure, that is measures with an affine log-density. As mean-field variational problems are defined by restricting the Gibbs variational principle (1) over tilts, the fact that these measures saturate the transportation-entropy inequality will be particularly crucial.

Using duality, the main consequence we derive from these transportation-entropy inequalities is a strong integrability inequality for Bernoulli and exponential processes, similar to the Gaussian case. In turn, this will provide us the main ingredient to obtain a dimension-free mean-field approximation of the free energy of Gibbs measures and in a similar fashion, an almost dimension-free nonlinear large deviations bound on the discrete hypercube. Moreover, we outline a general approach to obtain such estimates for more general background measures through the investigation of transportation-entropy inequalities which are optimal for tilts.

In the setting of the Ising model on \([-1, 1]^n\), we deduce that the mean-field approximation is within \(O(\sqrt{n}||J||_2)\) of the free energy, improving the previous known bound (3) involving the Hilbert–Schmidt norm of \(J\). Finally, we prove a dimension-free reverse log-Sobolev inequality on the discrete hypercube similar as the one existing in the Gaussian case [14].
2 Main results

2.1 Mean-field approximation

We denote by \( \mu^n \) the uniform measure on \( \{-1, 1\}^n \) and we define the entropy function \( I \) by,

\[
I(y) = \sum_{i=1}^{n} \left( \frac{1 + y_i}{2} \log(1 + y_i) + \frac{1 - y_i}{2} \log(1 - y_i) \right),
\]

with the convention that \( 0 \log 0 = 0 \), and \( I(y) = +\infty \) if \( y \notin [-1, 1]^n \). One can check that \( I(y) \) is equal to the relative entropy of \( \mu_y \) with respect to \( \mu^n \), where \( \mu_y \) the unique product probability measure on \( \{-1, 1\}^n \) with mean \( y \).

Similarly as the Gaussian mean-width, we define for a given subset \( V \subset \mathbb{R}^n \), the Rademacher mean-width of \( V \), \( b(V) \), by

\[
b(V) = \int \sup_{\xi \in V} \langle \xi, x \rangle d\mu^n(x).
\]

With these notation, we have the following result.

**Theorem 1** Let \( f : \mathbb{R}^n \to \mathbb{R} \) be a differentiable function.

\[
\left| \log \int e^{f} d\mu^n - \sup_{y \in [-1, 1]^n} \{ f(y) - I(y) \} \right| \leq 4b(V),
\]

where \( V = \nabla f([-1, 1]^n) \).

**Remark 1** For a given function \( f : \{-1, 1\}^n \to \mathbb{R} \), a natural extension to \([-1, 1]^n\) is given by its harmonic extension defined by,

\[
\forall y \in [-1, 1]^n, \quad \hat{f}(y) = \int f \, d\mu_y,
\]

where \( \mu_y \) is the unique product probability measure on \([-1, 1]^n\) with mean \( y \). Writing explicitly the right-hand side of (8), one can actually obtain a smooth extension to \( \mathbb{R}^n \). By [11, Fact 14] we know that

\[
\forall y \in [-1, 1]^n, \quad \nabla \hat{f}(y) = \int \nabla f \, d\mu_y,
\]

where the gradient on the right-hand side is the discrete gradient of \( f \), defined for any \( x \in \{-1, 1\}^n \) by \( \nabla f(x) = (\partial_1 f(x), \ldots, \partial_n f(x)) \), with

\[
\forall i \in \{1, \ldots, n\}, \quad \partial_i f(x) = \frac{1}{2}(f(x_+) - f(x_-)),
\]

where \( x_+ = (x_1, \ldots, x_{i-1}, 1, x_{i+1}, \ldots, x_n) \) and \( x_- = (x_1, \ldots, x_{i-1}, -1, x_{i+1}, \ldots, x_n) \).
Thus, for the harmonic extension, the set of gradients $\nabla \hat{f}([-1, 1]^n)$ is included in the convex hull of the set of discrete gradients. As the Rademacher mean-width of a set is equal to the one of its convex hull, the error term of Theorem 1 is just $b(\nabla f([-1, 1]^n))$.

**Remark 2** In certain situations the harmonic extension described in the previous remark is not the most practical to work with. But as we will show in Lemma 4, if $f : \mathbb{R}^n \to \mathbb{R}$ is a differentiable function and $\hat{f} : [-1, 1]^n \to \mathbb{R}$ is the harmonic extension of $f|_{[-1, 1]^n}$, then

$$\sup_{y \in [-1, 1]^n} \left( f(y) - \hat{f}(y) \right) \vee \sup_{y \in [-1, 1]^n} \left( \hat{f}(y) - f(y) \right) \leq 4b(V),$$

where $V = \nabla f([-1, 1]^n)$. This implies that in the mean-field variational problem appearing in Theorem 1, the extension that one chooses does not matter as soon as its set of gradients is of low complexity.

Applying Theorem 1 to the Ising model, we obtain the following corollary.

**Corollary 1** Let $J$ be a Hermitian matrix of size $n$ such that $J_{i,i} = 0$ for any $i \in \{1, \ldots, n\}$, and $h \in \mathbb{R}^n$. Then,

$$\left| \log \int e^{\langle x, Jx \rangle + \langle h, x \rangle} d\mu_n(x) - \sup_{y \in [-1, 1]^n} \{\langle y, Jy \rangle + \langle h, y \rangle - I(y) \} \right| \leq 16||J||_2 \sqrt{n},$$

where $|| \cdot ||_2$ denotes the Hilbert–Schmidt norm, namely

$$||J||_2 = \left( \sum_{1 \leq i, j \leq n} |J_{i,j}|^2 \right)^{1/2}.$$

**Proof** Let $f(x) = \langle x, Jx \rangle + \langle h, x \rangle$ for any $x \in \mathbb{R}^n$, $V = \nabla f([-1, 1]^n)$ and $X$ be a random variable uniformly sampled on $\{-1, 1\}^n$. We have

$$b(V) = 2\mathbb{E} \sup_{x \in [-1, 1]^n} \langle Jx, X \rangle = 2\mathbb{E} \sup_{x \in [-1, 1]^n} \langle x, JX \rangle \leq 4\sqrt{n} \mathbb{E} ||JX||_{\ell_2}.$$

One can check that $\mathbb{E} ||JX||_{\ell_2}^2 = ||J||_2$. Thus, by Cauchy–Schwarz inequality, we obtain $b(V) \leq 4\sqrt{n}||J||_2$, which, using Theorem 1 ends the proof.  

In the interesting large deviations regime, the free energy is expected to be of order $n$. Thus, the above Corollary 1 gives a meaningful estimate whenever $||J||_2 = o(\sqrt{n})$. It recovers the qualitative result of Basak and Mukherjee [3] for the Ising model, and gives a quantitative error term which is strictly smaller than the one found in [17], i.e (3).
Example 1 (d-regular graphs) Let $G$ be a $d$-regular graph with $n$ vertices ($n$ and $d$ are implicitly taken such that $nd$ is even). Denote by $A$ the adjacency matrix of $G$, and consider the Ising model with the interaction matrix $J = \frac{1}{d}A$. This scaling is taken so that the free energy of this model is of order $n$. As $||J||_2 = \sqrt{n/d}$, Corollary 1 gives that for any inverse temperature $\beta > 0$,

$$\log \int e^{\beta \langle x, Jx \rangle} d\mu^n(x) = \sup_{y \in [-1,1]^p} \{\beta \langle y, Jy \rangle - I(y)\} + O\left(\frac{n}{\sqrt{d}}\right).$$

This bound improves the one of Eldan [12, Example 3] who showed that for a $d$-regular expander, that is such that the second largest eigenvalue $\lambda_2(A) = O(\sqrt{d})$, one has the error

$$\frac{n}{\sqrt{d}^{1-o(1)}},$$

over the mean-field approximation, in the regime where $\log d \ll \log n$.

Remark 3 In [12], Eldan proved that for any $p > 0$ the mean-field approximation is within $O\left(\frac{n}{p+1} (n||J||_p)^{\frac{p}{p+1}}\right)$ of the free energy of the Ising model with interaction matrix $J$, where || ||$_p$ is the $p$-Schatten norm. One can note that for $p \geq 2$ and in the regime where Eldan’s bound is meaningful, that is $||J||_p = O(n^{1/p})$, the inequality $||J||_2 \leq n^{\frac{1}{2} - \frac{1}{p}} ||J||_p$, yields that the error term of Corollary 1 is smaller than $O((n||J||_p)^{\frac{p}{p+1}})$. However, the real interest of Eldan’s bound is when $p \leq 2$ and in particular the regime when $p \to 0$ when $n$ grows to infinity. For $p < 2$, it seems that Eldan’s bound and the one given by Corollary 1 cannot be compared in general. For specific examples, like the Curie-Weiss model or the lattice with mesoscopic interactions (see [12, example 2]) where the eigenvalues of the interaction matrix decrease exponentially fast to 0, Eldan’s bound is better and yields only logarithmic errors, whereas Corollary 1 can only provide error terms depending polynomially on the dimension.

2.2 Nonlinear large deviations

The theory of nonlinear large deviations was introduced by Chatterjee and Dembo [7] in order to understand the large deviations of nonlinear functions of independent Bernoulli random variables, motivated by the study of sparse Erdős–Rényi graphs.

Given a function $f : \{-1,1\}^n \to \mathbb{R}$, and $X$ uniformly sampled in $\{-1,1\}^n$, one can wonder when the optimal changes of measure in the large deviations of $f(X)$ are given by product measures. The nonlinear large deviations theory aims at identifying which condition on $f$ can guarantee this mechanism of deviation to happen. Similarly as for the question of the mean-field approximation of the free energy, Chatterjee and Dembo showed in [7] that a sufficient condition is that the set of gradients of $f$ is of low complexity in a $\ell^2$-metric entropy sense.

Efforts have been put into improving the original non-asymptotic bound of [7], which has the inconvenient of involving error terms related to the smoothness of an
extension of $f$ to $[-1, 1]^n$. In [27], Yan generalizes to products of compactly supported measures on a general Banach space, the nonlinear large deviations bound of [7]. Eldan [11] removed most of the smoothness assumptions in the case of the discrete hypercube and proved a bound where the complexity of the gradient is assessed in term of its Gaussian mean-width. In [1], the author showed a general nonlinear large deviations estimate removing as well the smoothness assumptions of [7] but involving covering numbers of the convex hull of the set of gradients. In [8, Corollary 2.2], Cook and Dembo proposed a nonlinear large deviations bound which has the specificity of not relying on the complexity of the gradient but rather on an efficient covering of the space by convex sets.

We propose here a nonlinear large deviations bound in the specific case of the discrete hypercube which is dimension-free up to a logarithmic factor.

For any probability measure $\nu \in \mathcal{P}(\mathbb{R})$, we denote by $\nu^n$ its $n$-fold product measure. Let $p \in (0, 1)$ and $\mu_p = p\delta_1 + (1 - p)\delta_{-1}$. We define the function $I_p$ by

$$I_p(y) = \frac{1 + y}{2} \log \frac{1 + y}{2p} + \frac{1 - y}{2} \log \frac{1 - y}{2(1 - p)},$$

and $I_p(y) = +\infty$ if $y \notin [-1, 1]$. We extend $I_p$ to $\mathbb{R}^n$ by setting $\forall y \in \mathbb{R}^n, I_p(y) := \sum_{i=1}^n I_p(y_i)$. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a differentiable function. Define

$$\forall t \in \mathbb{R}, \phi_p(t) = \inf \{I_p(y) : f(y) \geq t, y \in \mathbb{R}^n\}.$$

With this notation, we have the following theorem.

**Theorem 2** Let $t \in \mathbb{R}$ and $\delta > 0$. Assume that for any $s > t - \delta$, $\phi_p(s) > \phi_p(t - \delta)$. Let $V = \nabla f([-1, 1]^n)$ and let $X$ be a random vector sampled according to $\mu^n$. If $b(V) \leq \delta/4$, where $b(V)$ is defined in (7), then

$$\log \mathbb{P}(f(X) \geq t) \leq -\phi_p(t - \delta) + 2 \log \left(\frac{n \log(p(1-p))}{\delta} \vee 1\right) + 3,$$

where $L = \sup_{y \in [-1, 1]^n} ||\nabla f(y)||_{\ell^2}$.

**Remark 4** The assumption that for any $s > t - \delta$, $\phi_p(s) > \phi_p(t - \delta)$ is a technical assumption which should hold in most cases, as the function $\phi_p$ is expected to be zero on some interval $(-\infty, t_0]$ and then increasing on $(t_0, +\infty)$ for some $t_0 \in \mathbb{R}$ corresponding to the typical value of $f(X)$.

### 2.3 Reverse log-Sobolev inequality on the discrete hypercube

The logarithmic Sobolev inequality on the discrete hypercube (see [21, Theorem 5.1]) says that for any $\nu = e \mu^n$ probability measure on $\{-1, 1\}^n$, 

$$H(\nu|\mu^n) \leq \frac{1}{2} \int ||\nabla f(x)||^2_{\ell^2} d\nu(x),$$

(11)
where $\nabla f$ denotes the discrete gradient defined in (9).

The inequality (11) can be improved by replacing the quadratic function $\| || \|_{L^2}^2/2$ by $I(\nabla \Lambda_\mu^n)$ where $I$ is defined in (6) and $\Lambda_\mu^n$ is the logarithmic Laplace transform of $\mu^n$, defined by

$$\forall \xi \in \mathbb{R}^n, \quad \Lambda_\mu^n(\xi) = \log \int e^{\langle \xi, x \rangle} d\mu^n(x),$$

which gives

$$H(\nu|\mu^n) \leq \int I(\nabla \Lambda_\mu^n(\nabla f(x))) d\nu(x). \quad (12)$$

From the inequality in dimension 1 that for any $\lambda \in \mathbb{R}$, $I(\Lambda_\mu'(\lambda)) \leq \lambda^2/2$, we see that (11) is indeed implied by (12).

The proof of inequality (12) goes over induction on the dimension. For $n = 1$, one can check that there is actually equality. For $n > 1$, one uses the sub-additivity of the relative entropy [18, Proposition 5.6]:

$$H(\nu|\mu^n) \leq \sum_{i=1}^n H(\nu_{x(i)}|\mu)d\nu(x),$$

where $\nu_{x(i)}$ is the conditional distribution of $x_i$ given $x^{(i)} = (x_j)_{j \neq i}$ under $\nu$, which is equal in our case to the probability measure proportional to $e^{\partial_i f(x)} d\mu(x_i)$.

The interest of stating the log-Sobolev inequality as (12) is that it is saturated for product measures. Thus, one can expect that whenever the set of gradients of $f$ is of low complexity, the inequality (12) above is almost an equality. We prove that it is indeed the case, and show a reverse log-Sobolev inequality, similar to the one proved by Eldan and Ledoux [14] in the Gaussian case.

**Proposition 1** Let $\nu = e^f d\mu^n$ be a probability measure on $\{-1, 1\}^n$. Let

$$\mathcal{I}(\nu) = \int I(\nabla \Lambda_\mu^n(\nabla f(x))) d\nu(x),$$

where $\nabla f$ is the discrete gradient of $f$ defined in (9) and $I$ as in (6). Then,

$$\mathcal{I}(\nu) \leq H(\nu|\mu^n) + 4b(V),$$

where $V = \nabla f(\{-1, 1\}^n)$ and $b(V)$ is defined in (7).

### 2.4 Transportation-entropy inequalities

The starting point of our approach to the results stated above is to prove a transportation-entropy inequality for the uniform measure on the discrete hypercube which is tight for product measures. More generally, we address the question of finding, given a reference probability measure on $\mathbb{R}^n$, a transportation-entropy inequality which is saturated by tilts, that is measures with an affine log-density.
Let $\mathcal{P}(\mathbb{R}^n)$ denote the set of probability measures on $\mathbb{R}^n$. For any $\mu, \nu \in \mathcal{P}(\mathbb{R}^n)$, and a lower semi-continuous cost function $c : \mathbb{R}^n \times \mathbb{R}^n \to [0, +\infty]$, one defines the transportation cost $W_c(\nu, \mu)$ by
\[
W_c(\nu, \mu) = \inf_{\pi} \int c(x, y) d\pi(x, y),
\]
where the infimum runs over all couplings between $\nu$ and $\mu$. The assumption that the cost function is lower semi-continuous ensures, by Prokhorov’s theorem (see [25, Theorem 4.1]), that the infimum defining the transportation cost (13) above is achieved for some coupling $\pi$ called an optimal coupling.

We say that a given measure $\mu \in \mathcal{P}(\mathbb{R}^n)$ satisfies a transportation-entropy inequality with cost function $c : \mathbb{R}^n \times \mathbb{R}^n \to [0, +\infty]$ if
\[
\forall \nu \in \mathcal{P}(\mathbb{R}^n), \ W_c(\nu, \mu) \leq H(\nu|\mu),
\]
where $H(\nu|\mu)$ denotes the relative entropy of $\nu$ with respect to $\mu$.

Let $\mu$ be a reference probability measure on $\mathbb{R}^n$. We define its logarithmic Laplace transform $\Lambda_\mu$ by
\[
\forall \xi \in \mathbb{R}^n, \ \Lambda_\mu(\xi) = \log \int e^{\langle \xi, x \rangle} d\mu(x) \in \mathbb{R} \cup \{+\infty\},
\]
and we assume that the domain of $\Lambda_\mu$, $\mathcal{D}_\mu = \{\xi : \Lambda_\mu(\xi) < +\infty\}$, has nonempty interior. We call tilted measures of $\mu$ the family of probability measures $(\mu_\xi)_{\xi \in \mathbb{R}^n}$ defined by
\[
\forall \xi \in \mathcal{D}_\mu, \ \mu_\xi = e^{\langle \xi, x \rangle - \Lambda_\mu(\xi)} d\mu(x).
\]
As one can observe, the tilts of the uniform measure on $\{-1, 1\}^n$ are all the product measures on $\{-1, 1\}^n$, whereas the tilts of the standard Gaussian measure on $\mathbb{R}^n$, denoted by $\gamma$, are pushforward of $\gamma$ by translations.

By Talagrand’s result [23], we know that the standard Gaussian measure on $\mathbb{R}^n$ satisfies a transportation-entropy inequality with cost function $(x, y) \mapsto \frac{1}{2}||x - y||^2_{\ell^2}$, where $|| \cdot ||_{\ell^2}$ denotes the $\ell^2$-norm, that is,
\[
\forall \nu \in \mathcal{P}(\mathbb{R}^n), \ W_{\frac{1}{2}|| \cdot ||_{\ell^2}}(\nu, \gamma) \leq H(\nu|\gamma).
\]
Moreover, one can see from [23, Section 2, (2.5)] that equality occurs in (17) if and only if $\nu$ is a push-forward of $\gamma$ by a translation, that is a tilt of the standard Gaussian measure.

We will present two new transportation-entropy inequalities which are saturated by tilts, when the reference measure is the exponential measure and the uniform measure on the discrete hypercube. In the following we denote for any function $\Lambda : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ by $\Lambda^*$ its Legendre transform defined by
\[
\forall y \in \mathbb{R}^n, \ \Lambda^*(y) = \sup_{\xi \in \mathbb{R}^n} \{\langle \xi, y \rangle - \Lambda(\xi)\}.
\]
In the case of the exponential measure \( \eta = e^{-x} \mathds{1}_{x \geq 0} dx \), we consider the following cost function,

\[
\forall x, y \in (0, +\infty)^n, \quad c(x, y) = \sum_{i=1}^{n} y_i A^*_\eta \left( \frac{x_i}{y_i} \right),
\]  

(19)

and \( c(x, y) = +\infty \) if \( x \) or \( y \) \( \notin (0, +\infty)^n \). As one can check for any \( t > 0 \), \( A^*_\eta(t) = t - 1 - \log t \).

With these definitions, we have the following transportation-entropy inequality.

**Proposition 2**  
Let \( \eta \) be the probability measure \( e^{-x} \mathds{1}_{x \geq 0} dx \).

\[
\forall \nu \in \mathcal{P}(\mathbb{R}^n), \quad \mathcal{W}_c(\nu, \eta^n) \leq H(\nu|\eta^n).
\]

Moreover, the equality holds if \( \nu \) is a tilt of \( \eta^n \).

**Remark 5**  
In [23], Talagrand proved that the symmetric exponential measure \( m = \frac{1}{2} e^{-|x|} dx \) satisfies a certain family of transportation-cost inequalities with costs \( c_t \) indexed by \( t \in (0, 1) \), defined by,

\[
\forall x \in \mathbb{R}, \quad c_t(x, y) = \left( \frac{1}{t} - 1 \right) \left( e^{-t|x-y|} + t|x-y| - 1 \right).
\]

This family of cost functions has the striking property that for *any* \( t \) there exists a probability measure which achieves the equality in the transportation-entropy inequality with cost function \( c_t \). In this sense, this is a family of optimal cost functions. However, the probability measures which saturate the inequality are not tilts of the exponential measure, but are more intricate measures whose monotonous rearrangements from the exponential measure satisfy a certain family of differential equations.

To deal with the singularity of the Bernoulli measure, we propose a variant of the transportation-entropy inequality (14) where we make it possible to enrich the transportation cost by considering another reference measure than the one of the relative entropy. More precisely, we will say that \( \mu \) satisfies a transportation-entropy inequality with cost function \( c : \mathbb{R}^n \times \mathbb{R}^n \to [0, +\infty] \) if

\[
\forall v \in \mathcal{P}(\mathbb{R}^n), \quad \mathcal{W}_c(v, \tilde{\mu}) \leq H(v|\mu),
\]

where \( \tilde{\mu} \) is a given probability measure on \( \mathbb{R}^n \).

Let \( p \in (0, 1) \) and define the cost function \( w_p : \{-1, 1\}^n \times [-1, 1]^n \to [0, +\infty] \) by

\[
\forall x \in \{-1, 1\}^n, y \in [-1, 1]^n, \quad w_p(x, y) = \sum_{i=1}^{n} 2 |I_p(y_i)| \mathds{1}_{x_i(y_i - h_0) > 0},
\]

(20)

where \( h_0 = 2p - 1 \) and \( I_p \) is defined in (10). With these definitions, we have the following transportation-entropy inequality.
Proposition 3 Let \( \mu_p = (1 - p)\delta_{-1} + p\delta_1 \) and \( \mathcal{U} \) be the uniform probability measure on \([-1, 1]\). For any probability measure \( v \) on \([-1, 1]^n\),

\[
\mathcal{W}_{w_p}(v, \mathcal{U}^n) \leq H(v|\mu_p^n),
\]

and equality holds if \( v \) is a product measure.

2.5 Strong integrability of empirical processes

The main consequence we will derive from the transportation-entropy of the previous section consists in the strong integrability of Bernoulli and exponential empirical processes. By empirical process, we mean any process of the form,

\[
((\langle \xi, X \rangle))_{\xi \in V},
\]

where \( V \) is some subset of \( \mathbb{R}^n \), \( X \) is a random vector in \( \mathbb{R}^n \) with independent and identically distributed coordinates, and \( \langle \cdot, \cdot \rangle \) denotes the standard inner product in \( \mathbb{R}^n \).

In the Gaussian case, it is known that for any set \( V \subset \mathbb{R}^n \),

\[
\log \int e^{\sup_{\xi \in V} \{\langle \xi, x \rangle - \frac{1}{2}||\xi||^2_2\}} d\gamma(x) \leq \int \sup_{\xi \in V} \langle \xi, x \rangle d\gamma(x). \tag{21}
\]

This inequality can be traced back to [26], [24]. It can also be seen as a consequence of Talagrand’s transportation-entropy for the Gaussian measure (17).

We show that a similar estimate holds for the uniform measure on \([-1, 1]^n\) and the exponential measure, where the quadratic cost \( \frac{1}{2}||\xi||^2_2 \) is replaced by the logarithmic Laplace transform of the measure considered.

The following estimate of the exponential moments of supremum of Bernoulli empirical process will be one of the key elements in our approach to the mean-field approximation.

Proposition 4 Let \( \mu^n \) be the uniform probability measure on \([-1, 1]^n\). For any subset \( V \subset \mathbb{R}^n \),

\[
\log \int e^{\sup_{\xi \in V} \{\langle \xi, x \rangle - \Lambda_{\mu^n}(\xi)\}} d\mu^n(x) \leq 4b(V),
\]

where \( b(V) \) is defined in (7).

Remark 6 The above proposition remains true if \( \mu^n \) is replaced by any product measure on \([-1, 1]^n\) (with \( \Lambda_{\mu^n} \) replaced by \( \Lambda_v \)). Indeed, if \( v \) is a product measure whose support is \([-1, 1]^n\), one can find \( \theta \in \mathbb{R}^n \) such that \( \log dv/d\mu^n = \langle \theta, x \rangle - \Lambda_{\mu^n}(\theta) \), and we have for any \( \xi \in \mathbb{R}^n \), \( \Lambda_v(\xi) = \Lambda_{\mu^n}(\xi + \theta) - \Lambda_{\mu^n}(\theta) \). Moreover, for any \( V \subset \mathbb{R}^n \), \( b(V + \theta) = b(V) \). The case of general product measures on \([-1, 1]^n\) is then obtained by density.

Similarly, we get in the case of the exponential measure the following result.
Proposition 5 Let \( \eta \) be the probability measure \( 1_{x \geq 0} e^{-x} \, dx \). For any subset \( V \subset \mathbb{R}^n \),

\[
\log \int e^{\sup_{\xi \in V} \{ \langle \xi, x \rangle - \Lambda_\eta(\xi) \}} \, d\eta^n(x) \leq \int \sup_{\xi \in V} \langle \Lambda_\eta(\xi), x - u \rangle \, d\eta^n(x),
\]

where \( u \) denotes the vector \((1, 1, \ldots, 1)\), and \( \Lambda_\eta(\xi) = (\Lambda_\eta(\xi_1), \ldots, \Lambda_\eta(\xi_n)) \).

### 3 Transportation-entropy inequalities

Before proving the transportation-entropy inequalities of Propositions 2 and 3, we recall in this section some standard features of transportation-entropy inequalities which we will use in the sequel.

#### 3.1 Tensorization properties

A first important property of transportation-entropy inequalities is that they tensorize in a certain way which we recall in the following lemmas.

**Lemma 1** ([15, Proposition 1.3]) For \( i \in \{1, 2\} \), let \( \mu_i \in \mathcal{P}(\mathbb{R}^{d_i}) \) where \( d_i \in \mathbb{N} \). If for \( i \in \{1, 2\} \), \( \mu_i \) satisfies the inequality,

\[
\forall v \in \mathcal{P}(\mathbb{R}^{d_i}), \quad W_{c_i}(v, \tilde{\mu}_i) \leq H(v|\mu_i),
\]

where \( c_i : \mathbb{R}^{d_i} \times \mathbb{R}^{d_i} \to [0, +\infty] \) is lower semi-continuous and \( \tilde{\mu}_i \in \mathcal{P}(\mathbb{R}^{d_i}) \), then

\[
\forall v \in \mathcal{P}(\mathbb{R}^{d_1} \times \mathbb{R}^{d_2}), \quad W_{c_1 \oplus c_2}(v, \tilde{\mu}_1 \otimes \tilde{\mu}_2) \leq H(v|\mu_1 \otimes \mu_2),
\]

with \( c_1 \oplus c_2 \) defined for any \( x = (x_1, x_2) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_2} \) and \( y = (y_1, y_2) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_2} \) by,

\[
c_1 \oplus c_2(x, y) = c_1(x_1, y_1) + c_2(x_2, y_2). \tag{22}
\]

Moreover, one can check that when dealing with product measures, the relative entropy and the transportation cost tensorize exactly, as described in the following lemma, whose proof is left to the reader.

**Lemma 2** For \( i \in \{1, 2\} \), let \( \mu_i \) and \( v_i \) be probability measures on \( \mathbb{R}^{d_i} \), and \( c_i : \mathbb{R}^{d_i} \times \mathbb{R}^{d_i} \to [0, +\infty] \) be lower semi-continuous functions. Then,

\[
H(v_1 \otimes v_2|\mu_1 \otimes \mu_2) = H(v_1|\mu_1) + H(v_2|\mu_2),
\]

and

\[
W_{c_1 \oplus c_2}(v_1 \otimes v_2, \mu_1 \otimes \mu_2) = W_{c_1}(v_1, \mu_1) + W_{c_2}(v_2, \mu_2),
\]

where \( c_1 \oplus c_2 \) is defined in (22).
Note that when the background measure is a product measure, its tilts are also product measures, as one can see from the definition (16). Thus, the question of finding a transportation-entropy inequality for the $n$-fold product measure $\mu^n$ which is saturated for tilts reduces itself to a 1-dimensional problem by the two tensorization properties stated in Lemmas 1 and 2.

### 3.2 Infimum-convolution inequalities

The main aspect of transportation-entropy inequalities we will use is their dual functional form, which consists of infimum-convolution inequalities. This duality relies on the Kantorovich duality (see [25, Theorem 5.10]) which states that if $c : \mathbb{R}^n \times \mathbb{R}^n \to [0, +\infty]$ is lower semi-continuous, then for any $\nu, \mu \in \mathcal{P}(\mathbb{R}^n)$,

$$\mathcal{W}_c(\nu, \mu) = \sup_{(\phi, \psi) \in A} \left\{ \int \phi d\nu - \int \psi d\mu \right\}, \tag{23}$$

where the supremum holds over the set $A$ of admissible pairs $(\phi, \psi) \in L^1(\nu) \times L^1(\mu)$ such that for $\nu$-almost all $x$ and $\mu$-almost all $y$

$$\phi(x) - \psi(y) \leq c(x, y). \tag{24}$$

Moreover, by [25, Theorem 5.10, (ii)], if $\mathcal{W}_c(\nu, \mu) < +\infty$, then a coupling $\pi$ between $\nu$ and $\mu$ is optimal if and only if there exists $\phi : \mathbb{R}^n \to \mathbb{R} \cup \{-\infty\}$ such that $\pi$-almost everywhere,

$$\phi(x) - \phi^c(y) = c(x, y),$$

where $\phi^c$ the $c$-conjugate of $\phi$ is defined by,

$$\forall y \in \mathbb{R}^n, \phi^c(y) = \sup_{x \in \mathbb{R}^n} \{\phi(x) - c(x, y)\}.$$ 

One calls the pair $(\phi, \phi^c)$ a solution to the dual Kantorovitch problem.

In the next lemma, we recall the equivalence between transportation-entropy inequalities and infimum-convolution inequalities and we show that an equality case in the transportation-entropy inequality can be translated into an equality case for the corresponding infimum-convolution inequality. In the following, we denote for any probability measure $\nu$ on $\mathbb{R}^n$ by $\text{supp}(\nu)$ its support.

**Proposition 6** Let $\mu, \tilde{\mu} \in \mathcal{P}(\mathbb{R}^n)$ and $c : \mathbb{R}^n \times \mathbb{R}^n \to [0, +\infty]$ be a lower semi-continuous function. The following statements are equivalent.

(i). $\mu$ satisfies the transportation-entropy inequality,

$$\forall \nu \in \mathcal{P}(\mathbb{R}^n), \mathcal{W}_c(\nu, \tilde{\mu}) \leq H(\nu | \mu). \tag{25}$$
(ii). \( \mu \) satisfies the infimum-convolution inequality,

\[
\log \int e^f \, d\mu \leq \int f^c \, d\tilde{\mu},
\]

for any \( f : \mathbb{R}^n \to \mathbb{R} \) measurable such that \( f^c \in L^1(\tilde{\mu}) \).

Assume (i) or (ii) is satisfied. Let \( \nu \in \mathcal{P}(\mathbb{R}^n) \) be such that \( H(\nu|\mu) < +\infty \) and let \( f = \log \frac{d\nu}{d\mu} \). Assume further that for \( \tilde{\mu} \)-almost all \( y \), the function \( f - c(., y) \) is continuous on \( \text{supp}(\nu) \) and equal to \( -\infty \) on \( \text{supp}(\nu)^c \). Then,

\[
\mathcal{W}_c(\nu, \tilde{\mu}) = H(\nu|\mu) \iff f^c \in L^1(\tilde{\mu}), \quad \int f^c \, d\tilde{\mu} = 0.
\]

**Remark 7** In general, if (i) or (ii) is satisfied, then the equality \( \mathcal{W}_c(\nu, \tilde{\mu}) = H(\nu|\mu) \) holds if and only if there exists \( g : \mathbb{R}^n \to \mathbb{R} \cup \{-\infty\} \) such that \( f = g \nu \)-almost surely and

\[
g^c \in L^1(\tilde{\mu}), \quad \int g^c \, d\tilde{\mu} = 0.
\]

**Proof** A proof of the equivalence between (i) and (ii) can be found in [15, Corollary 3.1] or [25, Theorem 5.26]. We are now left to prove the equivalence between the equality cases. From now on we assume that (i) or equivalently (ii) holds. Let \( \nu \in \mathcal{P}(\mathbb{R}^n) \) be such that \( H(\nu|\mu) < +\infty \) and let \( f = \log \frac{d\nu}{d\mu} \). We assume that for \( \tilde{\mu} \)-almost all \( y \), \( f - c(., y) \) is continuous on \( \text{supp}(\nu) \) and equal to \( -\infty \) on \( \text{supp}(\nu)^c \).

**Necessary condition:** Assume that we have the equality

\[
\mathcal{W}_c(\nu, \tilde{\mu}) = H(\nu|\mu).
\]

By [25, Theorem 4.1], we know that there exists \( \pi \) an optimal coupling between \( \nu \) and \( \tilde{\mu} \). Namely

\[
\int c(x, y) \, d\pi(x, y) = \mathcal{W}_c(\nu, \tilde{\mu}).
\]

As \( H(\nu|\mu) < +\infty \), we have that \( \mathcal{W}_c(\nu, \tilde{\mu}) < +\infty \). Therefore, by [25, Theorem 5.10, (ii)] there exists a function \( \phi : \mathbb{R}^n \to \mathbb{R} \cup \{-\infty\} \) such that \( \pi \)-almost everywhere

\[
\phi(x) - \phi^c(y) = c(x, y).
\]

Our goal is to show that \( \nu \)-almost surely \( \phi \) is equal to \( f \) up to some additive constant. Indeed, assume for now that \( \nu \)-almost surely \( \phi = f + \alpha \) for some constant \( \alpha \in \mathbb{R} \), then using the fact that \( (\phi - \alpha)^c = \phi^c - \alpha \) and letting \( g = \phi - \alpha \), we get that \( \nu \)-almost surely \( g = f \) and \( \pi \)-almost surely,

\[
g(x) - g^c(y) = c(x, y).
\]
Since $H(v|\mu) < +\infty$ we have $f \in L^1(v)$ and therefore $g \in L^1(v)$. As $W_c(v, \tilde{\mu}) < +\infty$ we have that $c \in L^1(\pi)$. Using (30) we get that $g^c \in L^1(\tilde{\mu})$. Integrating equality (30) with respect to $\pi$ yields

$$\int g^c(y)d\tilde{\mu}(y) = \int f(x)d\nu(x) - \int c(x, y)d\pi(x, y). \quad (31)$$

Using (27) and (28), we get

$$\int f d\nu = H(v|\mu) = \int c(x, y)d\pi(x, y).$$

Thus, by (31), $\int g^c d\tilde{\mu} = 0$. To conclude, we will show that $\tilde{\mu}$-almost surely, $g^c = f^c$. Note that (30) and the fact that $g = f$ $\nu$-almost surely entail that $\tilde{\mu}$-almost surely $g^c \leq f^c$. To prove the converse inequality, let $A$ be the subset of $\text{supp}(v)$ where $f = g$. As $\nu(A) = 1$ and $A \subset \text{supp}(\nu)$, we deduce that $A$ is dense in $\text{supp}(\nu)$. As we assumed that for $\tilde{\mu}$-almost all $y$, $f - c(., y)$ is continuous on $\text{supp}(\nu)$, and equal to $-\infty$ on $\text{supp}(\nu)^c$, we can write for $\tilde{\mu}$-almost all $y$,

$$f^c(y) = \sup_{x \in \text{supp}(\nu)} \{f(x) - c(x, y)\} = \sup_{x \in A} \{f(x) - c(x, y)\} = \sup_{x \in A} \{g(x) - c(x, y)\}.$$ 

Therefore, for $\tilde{\mu}$-almost all $y$, $f^c(y) \leq g^c(y)$, and thus $f^c = g^c$ $\tilde{\mu}$-almost surely, which would end the proof the first part of the equivalence between the equality cases.

We are thus reduced to prove that $\nu$-almost surely $\phi$ and $f$ are equal up to an additive constant. Let $\psi : \mathbb{R}^n \rightarrow \mathbb{R}$ be a measurable and bounded function such that

$$\int \psi d\nu = 0, \quad \text{and} \quad \int |\psi\phi|d\nu < +\infty.$$ 

As $\psi$ is bounded, we can define for $\delta > 0$ small enough the probability measure

$$\nu_\delta = (1 + \delta \psi)d\nu.$$ 

We will prove that $\delta > 0$ small enough

$$W_c(\nu_\delta, \tilde{\mu}) - W_c(\nu, \tilde{\mu}) \geq \int \phi d(\nu_\delta - \nu). \quad (32)$$

The above inequality is the consequence of Kantorovitch duality (23) and the fact that the pair $(\phi, \phi^c)$ is a solution to the dual Kantorovitch problem of $W_c(v, \tilde{\mu})$. The only difficulty lies in the fact that $\phi$ and $\phi^c$ are not separately integrable a priori. Following [25, proof of Theorem 5.10], we use a truncation argument to overcome this issue. For
Let \( m \in \mathbb{N} \) denote by \( T_m \) the function

\[
\forall x \in \mathbb{R} \cup \{ \pm \infty \}, \quad T_m(x) = \begin{cases} 
 x & \text{if } |x| \leq m, \\
 m & \text{if } x > m, \\
 -m & \text{if } x < -m.
\end{cases}
\]

For any \( x, y \in \mathbb{R}^n \), define

\[
\xi(x, y) = \phi(x) - \phi_c(y), \quad \xi_m(x, y) = T_m(\phi(x)) - T_m(\phi_c(y)).
\]

One can check that if \( \xi(x, y) \geq 0 \) then \( \xi_m(x, y) \) is non-negative and non-decreasing to \( \xi(x, y) \), and if \( \xi(x, y) \leq 0 \), then \( \xi_m(x, y) \) is non-positive and non-increasing to \( \xi(x, y) \). Therefore,

\[
\xi_m \leq \xi \leq c,
\]

so that \( (T_m \circ \phi, T_m \circ \phi_c) \) is non-decreasing and non-negative on \( \{ \xi \geq 0 \} \), we get by monotonous convergence,

\[
\int \xi_m(x, y) d\pi(x, y) \rightarrow_{m \rightarrow +\infty} \int \xi(x, y) d\pi(x, y).
\]

As \( \xi(x, y) = c(x, y) \geq 0 \), \( \pi \)-almost surely,

\[
\int \xi_m(x, y) d\pi(x, y) = \int_{\{\xi \geq 0\}} \xi_m(x, y) d\pi(x, y).
\]

Since \( (\xi_m)_m \) is non-decreasing and non-negative on \( \{ \xi \geq 0 \} \), we get by monotonous convergence,

\[
\int \xi_m(x, y) d\pi(x, y) \rightarrow_{m \rightarrow +\infty} \int c(x, y) d\pi(x, y) = \mathcal{W}_c(v, \tilde{\mu}),
\]

where we used (28). Therefore, by taking \( m \rightarrow +\infty \) in (33), we obtain by dominated convergence the claim (32). Using that (i) holds and that \( \mathcal{W}_c(v, \tilde{\mu}) = H(v|\mu) \), we get from (32) that for any \( \delta > 0 \) small enough

\[
H(v_\delta|\mu) - H(v|\mu) \geq \int \phi d(v_\delta - v).
\]
Note that $H(v_\delta|\mu) = \int \log((1 + \delta \psi)e^f) dv_\delta$ and $H(v|\mu) = \int f dv$. Therefore, by Jensen’s inequality and the fact that $\int \psi dv = 0$, we obtain

$$H(v_\delta|\mu) - H(v|\mu) \leq \int \log ((1 + \delta \psi)e^f) d(v_\delta - v).$$

As $d(v_\delta - v) = \delta \psi dv$, we get by dividing (34) by $\delta$:

$$\int \log(1 + \delta \psi) \psi dv + \int f \psi dv \geq \int \phi \psi dv. \tag{35}$$

Taking $\delta \to 0$ we conclude by dominated convergence, that $\int f \psi dv \geq \int \phi \psi dv$. By symmetry, we can conclude that for any $\psi$ bounded, measurable such that $\int \psi dv = 0$ and $\int |\phi\psi| dv < +\infty$

$$\int f \psi dv = \int \phi \psi dv. \tag{35}$$

Define for $r > 0$ the set $A_r = \{x : |\phi(x)| \leq r\}$. As $c < +\infty \pi$-almost surely, we deduce from (29) that $|\phi| < +\infty v$-almost surely. Thus, for $r$ large enough we can define $v_r$ the conditional measure of $v$ given $A_r$. From (35), we deduce that for any bounded measurable $\psi$,

$$\int (f - \phi - \alpha_r) \psi dv_r = 0,$$

where $\alpha_r = \int (f - \phi) dv_r$. This implies that $v_r$-almost surely, $\psi = f + \alpha_r$. As $(A_r)_{r>0}$ is a non-decreasing sequence of sets for the inclusion such that $\cup_{r>0} A_r = \mathbb{R}^n$, we can conclude that $v$-almost surely, $\phi = f + \alpha$, for some $\alpha \in \mathbb{R}$.

**Sufficient condition:** We assume that $f^c \in L^1$ and $\int f^c d\tilde{\mu} = 0$. By definition,

$$H(v|\mu) = \int f dv.$$

Let $\pi$ be a coupling between $v$ and $\tilde{\mu}$. We can write,

$$H(v|\mu) = \int f(x) dv(x) - \int \sup_{x \in \mathbb{R}^n} \{|f(x) - c(x, y)| d\tilde{\mu}(y) \leq \int c(x, y) d\pi(x, y).$$

As $\pi$ is arbitrary, we get that $H(v|\mu) \leq \mathcal{W}_c(v, \tilde{\mu})$. Since from (i) we know that the converse inequality holds, this proves the second part of the equivalence. \qed

### 4 The discrete hypercube

We examine in this section the case of the uniform measure $\mu^n$ on the discrete hypercube $\{-1, 1\}^n$. We start by proving Proposition 3, namely we show that $\mu^n$, as well as...
the products of biased Bernoulli distributions, satisfy transportation-entropy inequalities which are saturated by all the product measures on \([-1, 1]^n\). Using duality and comparison arguments between supremum of empirical processes, we then give a proof of the strong integrability inequality of empirical processes of Proposition 4.

Next, we derive the dimension-free mean-field approximation of Theorem 1 and the nonlinear large deviations bound of Theorem 2. The key step consists of Lemmas 5 and 6 respectively, which show that these questions boil down to controlling the deviations of an empirical process of the form,

\[
\left( \langle \xi, X \rangle - \Lambda_{\mu^n}(\xi) \right)_{\xi \in V},
\]

where \( V \subset \mathbb{R}^n \) and \( X \) is uniformly sampled on \([-1, 1]^n\). In turn, the deviations of processes of the form (36) can be precisely controlled through the strong integrability inequality of Proposition 4.

### 4.1 Proof of Proposition 3

Let \( p \in (0, 1) \) and \( \mu_p = (1 - p)\delta_{-1} + p\delta_1 \). By the tensorization property of transportation-entropy inequalities (see Lemma 1), and the fact that equality cases tensorizes exactly (see Lemma 2), it suffices to prove Proposition 3 for \( n = 1 \). This is the content of the following lemma. Recall that the cost function \( w_p \) is defined for any \( x \in \{-1, 1\} \) and \( y \in [-1, 1] \) by \( w_p(x, y) = 2|I'_p(y)|\mathbb{1}_{x(y - h_0) > 0} \), where \( h_0 = 2p - 1 \)

**Lemma 3** For any \( \nu \) probability measure on \([-1, 1]\),

\[
\mathcal{W}_{w_p}(\nu, \mathcal{U}) = H(\nu|\mu_p),
\]

where \( \mathcal{U} \) denotes the uniform measure on \([-1, 1]\).

**Proof** Let \( h \) denote the mean of \( \nu \). Let \( \pi \) be the law of

\[
(\text{sg}(h - U), U),
\]

where \( U \) is uniformly distributed on \([-1, 1]\), and \( \text{sg}(x) = 1 \) if \( x \geq 0 \) and \(-1 \) otherwise. By definition of \( \pi \),

\[
\int w_p(x, y) d\pi(x, y) = 2\mathbb{E}\left(|I'_p(U)|\mathbb{1}_{U \in (h, h_0)}\right).
\]

Since \( I_p(h_0) = 0 \), \( I_p \) is decreasing on \((-1, h_0)\) and increasing on \((h_0, 1)\), we get

\[
\int w_p(x, y) d\pi(x, y) = I_p(h).
\]

Besides, \( H(\nu|\mu_p) = I_p(h) \), which proves the inequality

\[
\mathcal{W}_{w_p}(\nu, \mathcal{U}) \leq H(\nu|\mu_p).
\]
Let now $\pi$ be an arbitrary coupling between $\nu$ and $U$. We will show that $\int w_p(x, y) d\pi(x, y) \geq I_p(h)$, thus ending the proof, together with (38), of the equality (37). Since $\nu$ has mean $h$, there exists a measurable function $q : [-1, 1] \to [0, 1]$ such that $\int q(y) dy = h + 1$ and

$$\pi = (q(y)\delta_1 + (1 - q(y))\delta_{-1}) U(dy).$$

We have,

$$\int w_p(x, y) d\pi(x, y) = \int_{-1}^{1} |I_p'(y)| 1_{\{y > h_0\}} q(y) dy + \int_{-1}^{1} |I_p'(y)| 1_{\{y < h_0\}} (1 - q(y)) dy = \int_{-1}^{1} q(y) I_p'(y) dy + I_p(-1).$$

Let $C$ be the set of measurable functions $q : [-1, 1] \to [0, 1]$ such that $\int q(y) dy = h + 1$. This set is a convex, weak-* compact subset of $L^\infty([-1, 1])$. Thus, the infimum

$$\inf_{q \in C} \int_{-1}^{1} q(y) I_p'(y) dy,$$

is achieved at one of the extreme points of $C$. As one can check, the extreme points of $C$ consists in the functions $1_A$, where $A$ is a Borel subset of $[-1, 1]$ with Lebesgue measure $h + 1$. Thus, it suffices to prove that

$$\inf_{A \in \mathcal{B}([-1, 1])} \int_{A} I_p'(y) dy \geq I_p(h) - I_p(-1),$$

where $\mathcal{B}([-1, 1])$ is the set of Borel subsets of $[-1, 1]$ and $\lambda$ denotes the Lebesgue measure. Since $I_p'$ is increasing, we deduce by the rearrangement inequality (see [19, Theorem 3.4]) that

$$\inf_{A \in \mathcal{B}([-1, 1])} \int_{A} I_p'(y) dy \geq \int_{-1}^{h} I_p'(y) dy = I_p(h) - I_p(-1),$$

which ends the proof. $\square$

4.2 Proof of Proposition 4

In the next proposition, we show in a general setting, how a transportation-entropy inequality which is saturated by tilts implies a strong integrability inequality for empirical processes.
Proposition 7  Let $\mu$ be a probability measure on $\mathbb{R}$ with support of diameter less than 1. Let $\tilde{\mu}$ be a probability measure on $\mathbb{R}$ and let $c : \mathbb{R} \times \mathbb{R} \to [0, +\infty]$ be a lower semi-continuous function such that for $\tilde{\mu}$-almost all $y$, $c(., y)$ is continuous on $\text{supp}(\mu)$. Assume that

$$\forall \nu \in \mathcal{P}(\mathbb{R}), \; W_c(v, \tilde{\mu}) \leq H(v|\mu), \quad (39)$$

and equality holds for the tilts of $\mu$. Then, for any subset $V \subset \mathbb{R}^n$,

$$\log \int e^{\sup_{\xi \in V} \{\langle x, \xi \rangle - \Lambda_{\mu_n}(\xi)\}} d\mu_n(x) \leq 2b(V),$$

where $b(V)$ is defined in (7).

Proof  Note that (39) is meaningful only when $H(v|\mu) < +\infty$, in particular, only when the support of $\nu$ is included in the one of $\mu$. Therefore, we can assume without loss of generality that $c(x, y) = +\infty$ whenever $x \notin \text{supp}(\mu)$. Let $K \subset \mathbb{R}$ denote the convex hull of the support of $\mu$. By the tensorization property of transportation-entropy inequalities (see Lemma 1), $\mu_n$ satisfies the transportation-entropy inequality,

$$\forall \nu \in \mathcal{P}(\mathbb{R}^n), \; W_c(v, \tilde{\mu}^n) \leq H(v|\mu^n),$$

where $c$ is extended to $\mathbb{R}^n \times \mathbb{R}^n$, by setting for any $x, y \in \mathbb{R}^n$, $c(x, y) := \sum_{i=1}^n c(x_i, y_i)$. Let $V$ be a subset of $\mathbb{R}^n$. By Proposition 6, we have

$$\log \int e^{\sup_{\xi \in V} \{\langle x, \xi \rangle - \Lambda_{\mu_n}(\xi)\}} d\mu_n(x) \leq \int \phi d\tilde{\mu}^n,$$

where

$$\forall y \in \mathbb{R}^n, \; \phi(y) = \sup_{\xi \in V, x \in K^n} \sup_{x \in K^n} \{\langle x, \xi \rangle - c(x, y) - \Lambda_{\mu_n}(\xi)\}.$$

Here, we used the fact that $c(x, y) = +\infty$ if $x \notin K^n$. Thus, it remains to compute the expectation of a supremum of a certain empirical process. To this end, we will use a symmetrization and a contraction argument. We write in probabilistic notation,

$$\int \phi d\tilde{\mu}^n = \mathbb{E} \sup_{\xi \in V} Z_\xi,$$

where for any $\xi \in \mathbb{R}^n$,

$$Z_\xi = \sum_{i=1}^n T_{\xi i}, \; \text{with} \; T_{\xi i} = \sup_{x \in K} \{\xi_i x - c(x, Y_i)\} - \Lambda_{\mu}(\xi_i),$$

and $Y = (Y_1, \ldots, Y_n)$ is sampled according to $\tilde{\mu}_n$. Let $\xi, \zeta \in \mathbb{R}^n$. We claim that for any $i \in \{1, \ldots, n\}$,

$$T_{\xi_i} - T_{\zeta_i} \leq |\xi_i - \zeta_i|. \quad (40)$$
For any $i \in \{1, \ldots, n\}$, we have by convexity of $\Lambda_\mu$,

$$T_{\xi_i} - T_{\zeta_i} \leq \sup_{x \in K} \{ (x - \Lambda'_\mu(\zeta_i))(\xi_i - \zeta_i) \}.$$  

Observe that differentiating (15) yields that $\Lambda'_\mu(\zeta_i)$ is the barycenter of the probability measure $\mu_{\xi_i} = e^{\xi_i x - \Lambda_\mu(\zeta_i)} d\mu(x)$. As $K$ is a convex set containing the support of $\mu$, we have $\Lambda'_\mu(\zeta_i) \in K$. Since $K$ has diameter less than 1, we get the claim (40).

The fundamental fact about the process $(T_{\xi_i})_{1 \leq i \leq n}, \xi \in \mathbb{R}^n$ is that it is centered. This is due to the fact that equality in the transportation-entropy inequality with cost function $c$ holds for the tilts of $\mu$. Indeed, as for $\tilde{\mu}$-almost all $y$, $c(., y)$ is continuous on $\text{supp}(\mu)$ and is equal to $+\infty$ on $\text{supp}(\mu)^c$, we deduce by Proposition 6 that equality holds in the corresponding infimum-convolution inequality for linear forms, which exactly says that, for any $\xi \in \mathbb{R}^n$ and $i \in \{1, \ldots, n\}$, $E T_{\xi_i} = 0$. Since the process $(T_{\xi_i})_{1 \leq i \leq n}, \xi \in \mathbb{R}^n$ is centered, by the symmetrization principle (see [21, Lemma 11.4]), we have

$$E \sup_{\xi \in V} \sum_{i=1}^n T_{\xi_i} \leq 2E \sup_{\xi \in V} \sum_{i=1}^n X_i T_{\xi_i},$$

where $X_1, \ldots, X_n$ are independent Rademacher random variables independent of $Y$. Observe that since $E T_0 = 0$ and $T_0 \leq 0$ almost surely, we have $T_0 = 0$ almost surely. Using the fact that the maps $\xi_i \mapsto T_{\xi_i}$ are contractions for every $i$ and $T_0 = 0$ almost surely, we deduce by [21, Theorem 11.6] that

$$E \left( \sup_{\xi \in V} \sum_{i=1}^n X_i T_{\xi_i} | Y \right) \leq E \sup_{\xi \in V} \sum_{i=1}^n X_i \xi_i,$$

which ends the proof. \qed

We can now end the proof of Proposition 4. By Proposition 3, the uniform measure $\mu^n$ on the discrete hypercube satisfies a transportation-entropy inequality which is saturated by tilted measures, namely all the product measures of the discrete hypercube. Using Proposition 7 with the measure $\hat{\mu} = \frac{1}{2} \delta_{-1/2} + \frac{1}{2} \delta_{1/2}$ and the set $2V$, where $V$ is a subset of $\mathbb{R}^n$, we get

$$\log \int e^{\sup_{\xi \in V} \{ (x, \xi) - \Lambda_\mu^n(\xi) \}} d\mu^n(x) \leq 2b(2V) = 4b(V).$$

Using the same comparison arguments between supremum of empirical processes as in the proof of Proposition 7, we obtain the following lemma which enables us to estimate the error between a given extension to $[-1, 1]^n$ of a function defined on $\{-1, 1\}^n$ and the harmonic extension.
Lemma 4 Let $f : \mathbb{R}^n \to \mathbb{R}$ be a differentiable function and $\hat{f}$ the harmonic extension of $f|_{\{-1, 1\}^n}$ to $[-1, 1]^n$ defined as in remark 1. Then,

$$\sup_{y \in [-1, 1]^n} (f(y) - \hat{f}(y)) \lor \sup_{y \in [-1, 1]^n} (\hat{f}(y) - f(y)) \leq 4b(V),$$

where $V = \nabla f([-1, 1]^n)$ and $b(V)$ is defined as in (7).

**Proof** Let $y \in [-1, 1]^n$. By definition,

$$\hat{f}(y) - f(y) = \mathbb{E} f(X_y) - f(y),$$

where $X_y$ is a random vector in $\{-1, 1\}^n$ with independent coordinates and mean $y$. By the mean-value theorem, we have

$$\hat{f}(y) - f(y) \leq \mathbb{E} \sup_{\xi \in V} \langle \xi, X_y - y \rangle.$$

Let $X$ be uniformly sampled on $\{-1, 1\}^n$ and independent of $X_y$. Using again the symmetrization and contraction principles (see [21, Lemma 11.4, Theorem 11.5]), we obtain

$$\hat{f}(y) - f(y) \leq 2 \mathbb{E} \sup_{\xi \in V} \sum_{i=1}^n \xi_i (X_y - y)_i X_i \leq 4 \mathbb{E} \sup_{\xi \in V} \sum_{i=1}^n \xi_i X_i = 4b(V),$$

where we used the fact that almost surely, for any $i$, $|(X_y - y)_i| \leq 2$. With the same argument we obtain $\sup_{y \in [-1, 1]^n} (f(y) - \hat{f}(y)) \leq 4b(V).$ \qed

### 4.3 Proof of Theorem 1

Building on the strong integrability of Bernoulli empirical processes, we give here a proof of the mean-field approximation of the free energy of Gibbs measures on the discrete hypercube. Note that the following lower bound is always true by the Gibbs variational principle (2):

$$\log \int e^f d\mu^n \geq \sup_{y \in [-1, 1]^n} \{ \hat{f}(y) - I(y) \},$$

where $\hat{f}$ denotes the harmonic extension of $f|_{\{-1, 1\}^n}$ defined in (8). By Lemma 4, we deduce that

$$\log \int e^f d\mu^n \geq \sup_{y \in [-1, 1]^n} \{ f(y) - I(y) \} - 4b(V),$$

where $V = \nabla f([-1, 1]^n).$
Thus, it only remains to prove the upper bound of Theorem 1. First, we identify the error term induced by the mean-field approximation with the help of the following lemma. A proof of this result can be found in [1, Proposition 1.1].

Lemma 5 Let $\nu$ be a compactly supported probability measure on $\mathbb{R}^n$. Denote by $K$ the convex hull of its support. For any differentiable function $f : \mathbb{R}^n \to \mathbb{R}$,

$$
\log \int e^{f} \, d\nu \leq \sup \{ f - \Lambda_{\nu}^* \} + \log \int e^{\sup_{\xi \in V} \{ \langle \xi, x \rangle - \Lambda_{\nu}(\xi) \}} \, d\nu(x),
$$

where $V$ is the convex hull of $\nabla f(K)$, and $\Lambda_{\nu}^*$ denotes the Legendre transform of $\Lambda_{\nu}$, defined in (18).

Combining the above result with the strong integrability inequality of Proposition 4, we obtain

$$
\log \int e^{f} \, d\mu^n \leq \sup \{ f - \Lambda_{\mu^n}^* \} + 4b(V).
$$

Observing that $I$ is the Legendre transform of $\Lambda_{\mu^n}$, this ends the proof of Theorem 1. In fact, putting together Lemma 5 and Proposition 7, we have the following more general result which states that a dimension-free mean-field approximation holds as soon as a transportation-entropy inequality saturated by tilts exists.

Proposition 8 Let $\mu$ be a probability measure on $\mathbb{R}$ whose support is included in a convex set $K$ of diameter less than 1. Assume there exist $\tilde{\mu}$ a probability measure on $\mathbb{R}$ and $c : \mathbb{R} \times \mathbb{R} \to [0, +\infty]$ a lower semi-continuous function such that for $\tilde{\mu}$-almost all $y$, $c(\cdot, y)$ is continuous on $\text{supp}(\mu)$. Assume that

$$
\forall \nu \in \mathcal{P}(\mathbb{R}), \quad W_{c}(\nu, \tilde{\mu}) \leq H(\nu|\mu),
$$

and equality holds for the tilts of $\mu$. Then, for any differentiable function $f : \mathbb{R}^n \to \mathbb{R}$,

$$
\log \int e^{f} \, d\mu^n \leq \sup \{ f - \Lambda_{\mu^n}^* \} + 2b(V),
$$

where $V = \nabla f(K^n)$ and $b(V)$ is defined in (7).

4.4 Proof of Theorem 2

Contrary to the nonlinear large deviations bounds shown in the previous works [7,11, 27], the proof of Theorem 2 will not rely on the computation of exponential moments of functions with a low complexity set of gradients. Instead, similarly as in the proof of Theorem 1 we use a result of [1] which states that in the general setting where $X$ is a random vector in $\mathbb{R}^n$ distributed according to a compactly supported measure $\nu$, and $f : \mathbb{R}^n \to \mathbb{R}$ is a smooth function, one can reformulate the deviations of $f(X)$ in terms of the deviations of the process $(\langle \theta \xi, X \rangle - \Lambda_{\nu}(\theta \xi))_{\xi \in V, \theta > 0}$. Then, in the
A transportation approach to the mean-field approximation

specific case of the uniform measure on $[-1, 1]^n$, we will use the strong integrability inequality of Bernoulli processes of Proposition 4 to control the deviations of the latter process.

In the next lemma, we reformulate in geometrical terms the large deviation event $\{f(X) \geq t\}$. This is the direct consequence of [1, Lemmas 2.1, 2.2].

**Lemma 6** Let $\nu$ be a compactly supported probability measure on $\mathbb{R}^n$, whose support is not included in a hyperplane. Denote by $K$ the convex hull of its support. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a differentiable function and let $W$ be the convex hull of $\nabla f(K)$. Define the function,

$$\forall t \in \mathbb{R}, \phi(t) = \inf \{ \Lambda^*_\nu(y) : f(y) \geq t, y \in \mathbb{R}^n \}.$$ 

Let $t \in \mathbb{R}$ and $\delta > 0$. Assume that for any $s > t - \delta$, $\phi(s) > \phi(t - \delta)$. If $x \in K$ and $f(x) \geq t$, then

$$\sup_{\xi \in W} \{ \langle \theta \xi, x \rangle - \Lambda^*_\nu(\theta \xi) - \theta \delta \} \geq \phi(t - \delta),$$

where $\theta_0 = \Lambda^*_\nu(x)/\delta$.

**Proof** Define the “rate function” $\psi$ as

$$\forall t \in \mathbb{R}, \psi(t) = \inf \{ \Lambda^*_\nu(y) : f(y) = t, y \in \mathbb{R}^n \}.$$

For any $\delta > 0$ and $A \subset \mathbb{R}$ we denote by $V_\delta(A)$ the $\delta$-neighborhood of $A$, defined by $V_\delta(A) = \{ x \in \mathbb{R} : \inf_{y \in A} |x - y| < \delta \}$. By [1, Lemmas 2.1 and 2.2], we know that for any $\delta, r > 0$, if $x \in K$ and $f(x) \notin V_\delta(\{ \psi \leq r \})$, then

$$\sup_{\xi \in W} \{ \langle \theta \xi, x \rangle - \Lambda^*_\nu(\theta \xi) - \theta \delta \} \geq r.$$  \hspace{1cm} (41)

Let now $t \in \mathbb{R}$ and $\delta > 0$. We claim that the hypothesis that for any $s > t - \delta$, $\phi(s) > \phi(t - \delta)$ implies that

$$\{ \psi \leq \phi(t - \delta) \} \subset (\infty, t - \delta].$$  \hspace{1cm} (42)

Indeed, if $\psi(s) \leq \phi(t - \delta)$ then in particular $\phi(s) \leq \phi(t - \delta)$. By our assumption on $\phi$ we deduce that $s \leq t - \delta$, which shows the claim (42). Thus, if $f(x) \geq t$ then $f(x) \notin V_\delta(\{ \psi \leq \phi(t - \delta) \})$, which using (41) ends the proof. \hfill $\square$

We can now give a proof of Theorem 2. For simplicity, we denote by $\Lambda^p_\mu$ the log-Laplace transform of $\mu^n_p$. We observe that its Legendre transform $\Lambda^*_\mu$ is equal to $I_p$ and

$$\forall x \in [-1, 1]^n, I_p(x) \leq n \log \left( \frac{1}{p(1 - p)} \right).$$
Using Lemma 6, we obtain,

\[ P(f(X) \geq t) \leq P \left( \sup_{\xi \in W, 0 \leq \theta \leq \theta_0} \{(\theta \xi, X) - \Lambda_p(\theta \xi) - \theta \delta\} \geq \phi_p(t - \delta) \right), \]

where \( \theta_0 = -n \log(p(1 - p)) / \delta \), and \( W \) is the convex hull of \( \nabla f([-1, 1]^n) \). We now perform a net argument on \( \theta \). Let \( D \) be a subset of \([0, \theta_0]\) such that for any \( \theta \in [0, \theta_0] \), there exists \( \theta' \in D \) satisfying \( 0 \leq \theta - \theta' \leq 1 / (2\sqrt{n}L) \) where

\[ L = \sup_{y \in K} ||\nabla f(y)||_\ell^2 = \sup_{\xi \in W} ||\xi||_\ell^2. \]

One can find a net \( D \) such that,

\[ |D| = [2\sqrt{n}L\theta_0] + 1 \leq 3 \left( \frac{n\sqrt{n} \log(p(1 - p))L}{\delta} \lor 1 \right). \]

For \( X \in \{-1, 1\}^n \) fixed, define the function

\[ G : \theta \in \mathbb{R}_+ \mapsto \sup_{\xi \in W} \{(\theta \xi, X) - \Lambda_p(\theta \xi) - \theta \delta\}. \]

We claim that for any \( \theta' \leq \theta \),

\[ G(\theta) - G(\theta') \leq 2(\theta - \theta')L\sqrt{n}. \] (43)

Indeed,

\[ G(\theta) - G(\theta') \leq \sup_{\xi \in W} \{(\theta - \theta') \langle \xi, X \rangle - \Lambda_p(\theta \xi) + \Lambda_p(\theta' \xi) - (\theta - \theta')\delta\}. \]

Using that \( \theta' \leq \theta \) and the convexity of \( \Lambda_p \), we get

\[ G(\theta) - G(\theta') \leq \sup_{\xi \in W} \{(\theta - \theta') \langle \xi, X - \nabla \Lambda_p(\theta' \xi) \rangle \}. \]

Since \( \sup_{\xi \in W} ||\xi||_\ell^2 = L \) and \( \nabla \Lambda_p(\theta' \xi) \in [-1, 1]^n \) as one can easily check, we get the claim (43).

Thus, using a union bound we get,

\[ \mathbb{P} \left( \sup_{\xi \in W, 0 \leq \theta \leq \theta_0} \{(\theta \xi, X) - \Lambda_p(\theta \xi) - \theta \delta\} \geq \phi_p(t - \delta) \right) \]

\[ \leq \sum_{\theta \in D} \mathbb{P} \left( \sup_{\xi \in W} \{(\theta \xi, X) - \Lambda_p(\theta \xi) - \theta \delta\} \geq \phi_p(t - \delta) - 1 \right). \]
Now, fix $\theta \in \mathcal{D}$. By Chernoff’s inequality, we have

$$
\log \mathbb{P}\left( \sup_{\xi \in W} \{ \langle \theta \xi, X \rangle - \Lambda_p(\theta \xi) - \theta \delta \} \geq \phi_p(t - \delta) - 1 \right) \leq -\phi_p(t - \delta) + 1
$$

$$
+ \log \mathbb{E} \mathbb{e}^{\sup_{\xi \in W} \{ \langle \theta \xi, X \rangle - \Lambda_p(\theta \xi) \} - \theta \delta}.
$$

As observed in remark 6, Proposition 4 remains true if we replace $\mu^n$ by $\mu^n_p$. Therefore,

$$
\log \mathbb{E} e^{\sup_{\xi \in W} \{ \langle \theta \xi, X \rangle - \Lambda_p(\theta \xi) \} \leq 4b(\theta W),
$$

where $b(\theta W)$ is defined in (7). Observe that as the Rademacher mean-width involves of a supremum of a linear form, the Rademacher mean-width of a set is the same as the one of its convex hull. Thus, $b(\theta W) = \theta b(W) = \theta b(V)$. We finally get

$$
\log \mathbb{P}\left( \sup_{\xi \in W} \{ \langle \theta \xi, X \rangle - \Lambda_p(\theta \xi) \} \geq \phi_p(t - \delta) - 1 \right) \leq -\phi_p(t - \delta) + 1
$$

$$
+ \theta (4b(V) - \delta).
$$

Therefore, if $b(W) \leq \delta/4$, we obtain

$$
\mathbb{P}(f(X) \geq t) \leq |\mathcal{D}| e^{-\phi_p(t-\delta)+1}.
$$

One can easily check that

$$
\log |\mathcal{D}| \leq 2 \log \left( n |\log(p(1-p))|L \right) + 2,
$$

which ends the proof.

### 4.5 Proof of Proposition 1

In this section, we give a proof of the reverse Sobolev inequality of Proposition 1. We will essentially follow the argument in the Gaussian case (see [14, Theorem 1]) which was based on the Gaussian integration by parts formula and Talagrand’s transportation-entropy inequality [23].

Let $v = e^f d\mu^n$ be a probability measure on $[-1, 1]^n$. As $I$ is the Legendre transform of $\Lambda_{\mu^n}$, we have for any $\xi \in \mathbb{R}^n$, $I(\nabla \Lambda_{\mu^n}(\xi)) = \langle \nabla \Lambda_{\mu^n}(\xi), \xi \rangle - \Lambda_{\mu^n}(\xi)$. Therefore,

$$
\mathcal{I}(v) = \int \left( \langle \nabla \Lambda_{\mu^n}(\nabla f(x)), \nabla f(x) \rangle - \Lambda_{\mu^n}(\nabla f(x)) \right) e^f(x) d\mu^n(x).
$$
Let \( i \in \{1, \ldots, n\} \) and \((x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) \in (-1, 1)^{n-1}\). Since \( \partial_i f(x) \) does not depend on \( x_i \), we have
\[
\int A'_\mu(\partial_i f(x))\partial_i f(x)e^{f(x)}d\mu(x_i) = A'_\mu(\partial_i f(x))\partial_i f(x)\left(\frac{1}{2}e^{f(x_+)} + \frac{1}{2}e^{f(x_-)}\right),
\]
where \( x_+ = (x_1, \ldots, x_{i-1}, 1, x_{i+1}, \ldots, x_n) \) and \( x_- = (x_1, \ldots, x_{i-1}, -1, x_{i+1}, \ldots, x_n) \). As \( A'_\mu = \tanh \), we have
\[
A'_\mu(\partial_i f(x)) = \frac{e^{f(x_+)} - e^{f(x_-)}}{e^{f(x_+)} + e^{f(x_-)}}.
\]
Therefore,
\[
\int A'_\mu(\partial_i f(x))\partial_i f(x)e^{f(x)}d\mu(x_i) = \frac{1}{2}(e^{f(x_+)} - e^{f(x_-)})\partial_i f(x)
= \int x_i\partial_i f(x)e^{f(x)}d\mu(x_i).
\]
Integrating the above equality with respect to \((x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)\) under \( \mu^{n-1} \), and summing over \( i \in \{1, \ldots, n\} \), we obtain
\[
\int \langle \nabla A_\mu^n(\nabla f(x)), \nabla f(x) \rangle e^{f(x)}d\mu^n(x) = \int \langle x, \nabla f(x) \rangle e^{f(x)}d\mu^n(x).
\]
This formula can be seen as the discrete analogue of the Gaussian integration by parts formula. We have thus proved
\[
\mathcal{I}(\nu) = \int \left( \langle x, \nabla f(x) \rangle - A_\mu^n(\nabla f(x)) \right) d\nu(x).
\]
In particular,
\[
\mathcal{I}(\nu) \leq \int \sup_{\xi \in V} \left\{ \langle x, \xi \rangle - A_\mu^n(\xi) \right\} d\nu(x).
\]
But, the Gibbs variational principle (1) implies that
\[
\int \sup_{\xi \in V} \left\{ \langle x, \xi \rangle - A_\mu^n(\xi) \right\} d\nu(x) \leq H(\nu|\mu^n) + \log \int e^{\sup_{\xi \in V} \left\{ \langle \xi, x \rangle - A_\mu^n(\xi) \right\}} d\mu^n(x).
\]
Using Proposition 4, we can conclude the proof.
5 The exponential measure

In this last section we investigate the case of the exponential measure and give a proof of Propositions 2 and 5. We consider the cost function defined by,

$$\forall x, y \in (0, +\infty)^n, \ c(x, y) = \sum_{i=1}^{n} y_i \Lambda_{\eta}^*(\frac{X_i}{y_i}),$$  \hspace{1cm} (44)

and by $c(x, y) = +\infty$ if $x$ or $y \notin (0, +\infty)^n$. We have $\Lambda_{\eta}^*(t) = t - 1 - \log t$ for any $t > 0$ and $\Lambda_{\eta}^* = +\infty$ otherwise. The form of this cost function can be explained by the natural coupling of all the tilts $(\eta_\lambda)_{\lambda > 0}$ of the exponential measure, where

$$\forall \lambda > 0, \ \eta_\lambda = 1_{x \geq 0} \lambda e^{-x/\lambda} dx.$$

There is a simple way to transport $\nu$ onto $\nu_\lambda$ by the map $x \mapsto \log x$. This fact explains the shape of the cost function (44) as essentially a function of the ratio $y/x$.

5.1 Proof of Proposition 2

By the tensorization properties of the transportation-entropy inequalities (see Lemmas 1 and 2), it is sufficient to prove the statement for $n = 1$. Let $\nu$ be a probability measure such that $H(\nu | \eta) < +\infty$. Let $\tilde{\nu}$ and $\tilde{\eta}$ be the push-forward of respectively $\nu$ and $\eta$ by the map $x \mapsto \log x$. One can check that

$$\tilde{\eta} = e^{-\xi(x)} dx,$$

with $\xi(x) = e^x - x$, which is a convex function. From [23, section 2] we know that $\tilde{\eta}$ satisfies a transportation-entropy inequality with cost function $\tilde{c}$ defined by,

$$\forall x, y \in \mathbb{R}, \ \tilde{c}(x, y) = \xi(x) - \xi(y) - \xi'(y)(x - y),$$

that is, $\tilde{c}(x, y) = e^y \Lambda_{\eta}^*(e^{y-x})$. Therefore $W_{\tilde{c}}(\tilde{\nu}, \tilde{\eta}) \leq H(\tilde{\nu} | \tilde{\eta})$. But as $x \mapsto \log x$ is invertible, we have on the one hand, $H(\tilde{\nu} | \tilde{\eta}) = H(\nu | \eta)$, and on the other hand, $W_{\tilde{c}}(\tilde{\nu}, \tilde{\eta}) = W_{c}(\nu, \eta)$, which proves that $W_{c}(\nu, \eta) \leq H(\nu | \eta)$.

It only remains to show that if $\nu$ is a tilt of $\eta^n$ then it achieves the equality in the transportation-entropy inequality of Proposition 2. Let $\lambda = (\lambda_1, \ldots, \lambda_n)$ with $\lambda_i > 0$. Denote by $\eta_{\lambda} = \eta_{\lambda_1} \otimes \ldots \otimes \eta_{\lambda_n}$. Let $\pi$ be a coupling between $\eta_{\lambda}$ and $\eta^n$. Since for any $i \in \{1, \ldots, n\}$, we have

$$\int y_i d\pi(x, y) = 1,$$
we deduce by Jensen’s inequality that,

\[
\int \sum_{i=1}^{n} v_i A^*_\eta\left(\frac{x_i}{y_i}\right) d\pi(x, y) \geq \sum_{i=1}^{n} A^*_\eta\left(\int x_i d\pi(x, y)\right) = \sum_{i=1}^{n} A^*_\eta\left(\int x_i d\eta^*(x_i)\right) = A^*_\eta(\lambda_i).
\]

As one can check (by a direct computation or using the Gibbs variational principle (1)), we have \(H(\eta_\lambda|\eta^n) = A^*_\eta(n).\) Thus,

\[
\mathcal{W}_c(\eta_\lambda|\eta^n) \geq H(\eta_\lambda|\eta^n),
\]

which ends the proof of the equality case.

### 5.2 Proof of Proposition 5

We now show how the transportation-entropy inequality of Proposition 2 entails by duality Proposition 5. Let \(V \subset \mathbb{R}^n\) be a subset. Define the function \(g\) by

\[
\forall x \in \mathbb{R}^n, \quad g(x) = \sup_{\xi \in V} \{\langle \xi, x \rangle - A^*_\eta(n)\}.
\]

By Propositions 2 and 6 we have,

\[
\log \int e^x d\eta^n \leq \int \sup_{x \in \mathbb{R}^n_+} \{g(x) - c(x, y)d\eta^n(y),
\]

where \(c\) is defined in (44), and we used the fact that \(c(x, y) = +\infty\) if \(x \leq 0\). But, for any \(y \in (0, +\infty)^n\),

\[
\sup_{x \in \mathbb{R}^n_+} \{g(x) - c(x, y)\} = \sup_{x \in \mathbb{R}^n_+} \{g(x) - \sum_{i=1}^{n} y_i A^*_\eta(n)(x_i/y_i)\} = \sup_{t \in \mathbb{R}^n_+} \{g(ty) - \langle y, A^*_\eta(t)\rangle\},
\]

where \(tx = (t_1 x_1, \ldots, t_n x_n)\) and \(A^*_\eta(t) = (A^*_\eta(t_1), \ldots, A^*_\eta(t_n))\). Therefore,

\[
\sup_{x \in \mathbb{R}^n_+} \{g(x) - c(x, y)\} = \sup_{\xi \in V} \sup_{t \in \mathbb{R}^n_+} \{\langle \xi, ty \rangle - \langle y, A^*_\eta(t)\rangle - A^*_\eta(n)\}.
\]

Fix \(\xi \in V\). We have

\[
\sup_{t \in \mathbb{R}^n_+} \{\langle \xi, ty \rangle - \langle y, A^*_\eta(t)\rangle\} = \sum_{i=1}^{n} \sup_{t > 0} (t\xi_i - A^*_\eta(t)) y_i = \sum_{i=1}^{n} A^*_\eta(\xi_i) y_i.
\]
where we used the fact that $\Lambda_\eta$ is the Legendre transform of $\Lambda_\eta^*$. From (45) we obtain

$$\log \int e^{\sup_{\xi \in V} \{\langle \xi, x \rangle - \Lambda_\eta^*(\xi)\}} d\eta^n(x) \leq \int \sup_{\xi \in V} \sum_{i=1}^n \Lambda_\eta(\xi_i)(y_i - 1) d\eta^n(y),$$

which gives the claim.

Acknowledgements I thank Ofer Zeitouni for many fruitful discussions which helped me build the present paper, as well as his valuable comments on an earlier version of the manuscript. I am grateful to Ronen Eldan for several influential and helpful discussions. Finally, I thank Michel Ledoux for his precious remarks which helped me improving this article. Data sharing not applicable to this article as no datasets were generated or analysed during the current study.

References

1. Augeri, F.: Nonlinear large deviation bounds with applications to Wigner matrices and sparse Erdős-Rényi graphs. Ann. Probab. 48(5), 2404–2448 (2020). https://doi.org/10.1214/20-AOP1427
2. Austin, T.: The structure of low-complexity Gibbs measures on product spaces. Ann. Probab. 47(6), 4002–4023 (2019). https://doi.org/10.1214/19-aop1352
3. Basak, A., Mukherjee, S.: Universality of the mean-field for the Potts model. Probab. Theory Related Fields 168(3–4), 557–600 (2017)
4. Bobkov, S., Gentil, I., Ledoux, M.: Hypercontractivity of Hamilton-Jacobi equations. J. Math. Pures Appl. (9) 80(7), 669–696 (2001)
5. Borgs, C., Chayes, J.T., Cohn, H., Zhao, Y.: An $L^p$ theory of sparse graph convergence II: LD convergence, quotients and right convergence. Ann. Probab. 46(1), 337–396 (2018)
6. Borgs, C., Chayes, J.T., Lovász, L., Sós, V.T., Vesztergombi, K.: Convergent sequences of dense graphs II Multiway cuts and statistical physics. Ann. Math. (2) 176(1), 151–219 (2012)
7. Chatterjee, S., Dembo, A.: Large deviations of subgraph counts for sparse Erdős-Rényi graphs. Adv. Math. (2020). https://doi.org/10.1016/j.aim.2020.107289
8. Dembo, A.: Information inequalities and concentration of measure. Ann. Probab. 25(2), 927–939 (1997)
9. Dembo, A., Zeitouni, O.: Large deviations techniques and applications, *Stochastic Modelling and Applied Probability*, vol. 38. Springer, Berlin (2010). https://doi.org/10.1007/978-3-642-03311-7. Corrected reprint of the second (1998) edition
10. Eldan, R.: Gaussian-width gradient complexity, reverse log-Sobolev inequalities and nonlinear large deviations. Geom. Funct. Anal. 28(6), 1548–1596 (2018). https://doi.org/10.1007/s00039-018-0461-z
11. Eldan, R.: Taming correlations through entropy-efficient measure decompositions with applications to mean-field approximation (2018, Preprint). https://arxiv.org/pdf/1811.11530.pdf
12. Eldan, R., Gross, R.: Decomposition of mean-field gibbs distributions into product measures. Electron. J. Probab. 23, 24 pp. (2018)
13. Eldan, R., Ledoux, M.: A dimension-free reverse logarithmic Sobolev inequality for low-complexity functions in Gaussian space (2018, Preprint). arXiv:1903.07093
14. Gozlan, N., Léonard, C.: Transport inequalities—a survey. Markov Process. Related Fields 16, 635–736 (2010)
15. Jain, V., Koehler, F., Mossel, E.: The mean-field approximation: Information inequalities, algorithms, and complexity. In: Conference On Learning Theory, COLT 2018, Stockholm, Sweden, 6-9 July 2018., pp. 1326–1347 (2018). http://proceedings.mlr.press/v75/jain18b.html
16. Jain, V., Koehler, F., Risteski, A.: Mean-field approximation, convex hierarchies, and the optimality of correlation rounding: a unified perspective (2018, Preprint). arXiv:1808.07226
17. Ledoux, M.: The Concentration of Measure Phenomenon, Mathematical Surveys and Monographs, vol. 89. American Mathematical Society, Providence (2001)
19. Lieb, E.H., Loss, M.: Analysis, Graduate Studies in Mathematics, vol. 14. American Mathematical Society, Providence (1997). https://doi.org/10.2307/3621022
20. Marton, K.: Bounding $d$-distance by informational divergence: a method to prove measure concentration. Ann. Probab. 24(2), 857–866 (1996)
21. Massart, P., Lugosi, G., Boucheron, S.: Concentration Inequalities?: A Nonasymptotic Theory of Independence. Oxford University Press, Oxford (2013)
22. Otto, F., Villani, C.: Generalization of an inequality by Talagrand and links with the logarithmic Sobolev inequality. J. Funct. Anal. 173(2), 361–400 (2000)
23. Talagrand, M.: Transportation cost for Gaussian and other product measures. Geom. Funct. Anal. 6(3), 587–600 (1996)
24. Tsirel’son, B.: A geometric approach to maximum likelihood estimation for an infinite-dimensional gaussian location. i. Teor. Veroyatnost. i Primenen. 27, 388–395 (1982)
25. Villani, C.: Optimal transport, Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], vol. 338. Springer-Verlag, Berlin (2009). Old and new
26. Vitale, R.: The Wills functional and Gaussian processes. Ann. Probab. 24(4), 2172–2178 (1996)
27. Yan, J.: Nonlinear large deviations: beyond the hypercube. Ann. Appl. Probab. 30(2), 812–846 (2020). https://doi.org/10.1214/19-AAP1516

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.