Smooth compression, Gallager bound and Nonlinear sparse-graph codes

Andrea Montanari
EE and Statistics Departments
Stanford University
montanari@stanford.edu

Elchanan Mossel
Statistics and CS Departments
UC Berkeley
mossel@stat.berkeley.edu

***Abstract*— A data compression scheme is defined to be *smooth* if its image (the codeword) depends gracefully on the source (the data). Smoothness is a desirable property in many practical contexts, and widely used source coding schemes lack of it.**

We introduce a family of smooth source codes based on sparse graph constructions, and prove them to achieve the (information theoretic) optimal compression rate for a dense set of iid sources. As a byproduct, we show how Gallager bound on sparsity can be overcome using non-linear function nodes.

I. Introduction

Data compression schemes that achieve optimal compression rate often present an inconvenient. Their image (the stored codeword) depends *chaotically* on the source (the data): Changing a single source symbol may lead to an unbounded change in stored data. This paper presents a construction that overcomes this problem for many (but not all) iid sources.

The rest of this Section provides a more formal definition of the problem, its motivation, and discusses related themes. Sections II and III present our main results and their proofs.

A. Smoothness

Let $x = (x_1, \dots, x_n)$ be a string of symbols taking values in a finite alphabet $\mathcal{X} \ni x_i$. If such symbols are iid random variables, we can encode them in a shorter string of symbols, say $z = (z_1, \dots, z_m)$, and incur a vanishing (as $n \to \infty$) error probability as long as $m \geq nH(X_i) + o(n)$. Several source-coding schemes achieve this goal [1].

Imagine now that the data x change because of some external cause. This might be because x depends on a data stream, or it is an image tracked by a camera, or a vector of temperatures recorded by a sensor network. In particular, our work was motivated by the the approach to internet traffic measurement proposed in [2], [3]. In such applications, x changes very rapidly over time, but each time it does change, the new vector is close to the old one in an appropriate metric. We want to keep in memory only a compressed version of the *current* vector x, and minimize the required memory. In most data compression schemes, this requires to recompute the whole coded vector z eaach time x changes. This requires at least $\Theta(n)$ operations, which is way too much.

The problem can be ascribed to the 'chaotic' behavior of most information-theoretic optimal source coding schemes. Formally, we describe such a scheme as a couple couple $(\mathsf{F}, \widehat{\mathsf{F}})$ of encoding/decoding maps $\mathsf{F} : \mathcal{X}^n \to \mathcal{Y}^m$, and $\widehat{\mathsf{F}} : \mathcal{Y}^m \to \mathcal{X}^n$ (its rate being $R \equiv m/n$.) We further denote by $||z - z'||_0$ the Hamming distance between two vectors z, z'. Then, one has $||\mathsf{F}(x) - \mathsf{F}(x')||_0 \gg 1$ even if x and x' differ only in one position. In order to eliminate this problem, we will require the source code to be smooth.

Definition I.1. *Given a constant $L \geq 0$ source code $(\mathsf{F}, \widehat{\mathsf{F}})$ is said to be L-*smooth *if F has Lipschitz norm smaller than L, i.e. if, for any x, $x' \in \mathcal{X}^n$*

$$||\mathsf{F}(x) - \mathsf{F}(x')||_0 \leq L||x - x'||_0 \,. \tag{1}$$

In other words, the source code is L smooth if, whenever one entry of the data changes, their image changes in at most L positions. Smoothness is a desirable property, but clearly constrains the code choice. In this paper we want to understand how much we have to pay (in terms of compression rate) in order to get it. The above definition naturally suggests a dual one:

Definition I.2. *Given a constant $L \geq 0$, the source code $(\mathsf{F}, \widehat{\mathsf{F}})$ is said to be L-*robust *if $\widehat{\mathsf{F}}$ has Lipschitz norm smaller than L, i.e. if, for any z, $z' \in \mathcal{Y}^m$*

$$||\widehat{\mathsf{F}}(z) - \widehat{\mathsf{F}}(z')||_0 \leq L||z - z'||_0 \,. \tag{2}$$

The origin of the name, as well as the practical interest of this notion are easily explained. If a source is compressed using a robust code, transmitted through a channel, and uncompressed, a small channel noise translates into a mild distortion of the data. While in this paper we focus on *smoothness*, similar techniques apply to *robustness* and will be the object of a forthcoming publication. Let us however emphasize that optimal (with respect to compression rate) smooth codes are typically not robust and vice-versa.

B. Sparse graphs and Gallager bound

There is a simple way to enforce smoothness: have each coordinate of the input affect a bounded number of coordinates of the output. Formally, let F_a, for $a \in [m]$, be the a-th component of the encoding function, and assume it depends on x only through the vector $x_{\partial a} = \{x_i : i \in \partial a\}$ for some $\partial a \subseteq [n]$. Viceversa, for $i \in [n]$, we let ∂i be the set of output coordinates that depend on i: $\partial i \equiv \{a \in [m] : i \in \partial a\}$.

Definition I.3. *Given a constant $d \geq 0$, a source code $(\mathsf{F}, \widehat{\mathsf{F}})$ is said to be d-*sparse *if $|\partial i| \leq d$ for any $i \in [n]$. In particular, such a code is d-smooth.*

A sparse code is conveniently represented by *factor graph* G over vertex sets $[n]$ ('variable nodes') and $[m]$ ('factor nodes') and and including an edge (i, a) (for $i \in [n]$, $a \in [m]$) whenever $i \in \partial a$. The code is d-sparse if and only if the variable nodes degree is bounded by d.

Sparse graphs were first used for source coding in [6], where the encoding function F was chosen to be linear (thus computing the syndrome of a low-density parity check code.) For this construction, a classic result by Gallager [4] implies that optimal compression rate cannot be achieved for bounded degree d. Therefore, in the linear setting, there is an unavoidable trade-off between smoothness and rate.

Below we will generalize Gallager bound to non-linear functions F, and show how it can be bypassed. This allows to construct d-sparse source codes that achieve optimal compression rate for iid sources.

Sparse-graph constructions were applied to lossy source coding in [9]. In particular Ref. [10] advocates the use of non-linear nodes, albeit in a different scheme and with objectives different from ours.

C. Geometry, embeddings and finitary codes

The above definitions have a suggestive geometrical interpretation. Consider the set $\mathsf{Typ}_n(p)$ of iid sequences $x \in \mathcal{X}^n$ that are *typical* with respect to some single letter distribution $\{p(x) : x \in \mathcal{X}\}$. Formally, we will define $\mathsf{Typ}_n(p)$ as the set of sequences x whose type θ_x satisfies[1] $||\theta_x - p||_{\text{TV}} \le n^{-\alpha}$ for some $\alpha \in (0, 1/2)$, say $\alpha = 0.1$. The volume of this set $|\mathsf{Typ}_n(n)| \doteq 2^{nH(p)}$ is about the same as the one of an Hamming hypercube $\{0,1\}^m$, of dimension $m = nH(p) + o(n)$. Suppose now that an L-smooth source code $\mathsf{F} : \mathcal{X}^n \to \{0,1\}^m$ exists with rate $R \approx H(p)$. This means that $\mathsf{Typ}_n(p)$ can be embedded in $\{0,1\}^m$ without 'folding' it onto itself, and without 'too much stretching.' If the code is both L-smooth and L-robust, then $\mathsf{Typ}_n(p)$ is roughly isometric to $\{0,1\}^m$.

Low-distortion embeddings of finite metric spaces in Banach spaces have been intensively studied in theoretical computer science [11]. They are a rich source of approximation algorithms. Here we propose instead to study embeddings of finite metric spaces in finite metric spaces.

Finitary coding [12] (in ergodic theory) is another topic related to the theme of this paper. Given two 'Bernoulli shifts' $X = \{X_i\}_{i\in\mathbb{Z}}$ and $Z = \{Z_i\}_{i\in\mathbb{Z}}$ (i.e. two sequences of iid random variable endowed with translations), with the same entropy rate, there exists a deterministic invertible mapping F such that: (a) $\mathsf{F}(X)$ is distributed as Z, (b) Each coordinate of $\mathsf{F}(X)$ depends on *finitely many* of the X_i's.

The problem treated in this paper is in some regards a finite-blocklength version of finitary coding.

II. Main results

This section provides formal definitions, some simple non-achievability results (including a generalization of Gallager bound) and the statement of our main result. Although our approach can be generalized to arbitrary (finite) input and output alphabets $\mathcal{X}$, $\mathcal{Y}$, to simplify the presentation we will assume throughout $\mathcal{Y} = \mathcal{X}$ and $|\mathcal{X}| = q$ to be a prime number. We will often identify $\mathcal{X}$ with the field of integers modulo q. Entropies will be measured in base q.

As a running example, we shall use a binary source: $\mathcal{X} = \{0,1\}$, with $p(0) = 1 - p(1) = \epsilon < 1/2$.

A. Definitions

Definition II.1. *Let p be a probability distribution over the finite alphabet $\mathcal{X}$. The* rate-Lipschitz constant pair (R, L) *is said to be* achievable for p *if there exists a sequence of L-smooth source codes $(\mathsf{F}_n, \widehat{\mathsf{F}}_n)$, $\mathsf{F}_n : \mathcal{X}^n \to \mathcal{X}^{nR}$, $\widehat{\mathsf{F}}_n : \mathcal{X}^{nR} \to \mathcal{X}^n$ such that*

$$\lim_{n\to\infty} \mathbb{P}\{\widehat{\mathsf{F}}_n(\mathsf{F}_n(X)) \neq X\} = 0\,. \tag{3}$$

Analogously, the rate-degree pair (R, d) *is* achievable for p *if the above holds for a sequence of d-sparse source codes $(\mathsf{F}_n, \widehat{\mathsf{F}}_n)$.*

Obviously, if (R, L) is achievable for p and $R' \ge R$, $L' \ge L$, then (R', L') is achievable as well. The set of achievable pairs is thus characterized by the *Lipschitz entropy* function

$$H_{\text{Lip}}(p; L) = \inf\{R : \ (R, L) \text{ is achievable for } p\}\,. \tag{4}$$

Similarly we define the *bounded degree* entropy function $H_{\deg}(p; d)$ by replacing the pair (R, L) with the rate-degree pair (R, d).

By the above remarks, $H_{\text{Lip}}(p; L)$ and $H_{\deg}(p; d)$ are non-increasing function of L, with $H_{\deg}(p; L) \ge H_{\text{Lip}}(p; L) \ge H(p)$ for any L.

B. Lower bounds (non-achievability)

For small L is not hard to derive a lower bound on $H(p; L)$ from geometrical considerations.

Proposition II.1. *Let $r(p) \equiv 1 - \sum_{x\in\mathcal{X}} p(x)^2$, $\mathcal{H}_q(x) \equiv -x\log_q(x/(q-1)) - (1-x)\log_q(1-x)$ denote the q-ary entropy function and $\overline{\mathcal{H}}_q(x) \equiv \mathcal{H}_q(\min(x, 1-1/q))$. Then*

$$H_{\text{Lip}}(p; L)\overline{\mathcal{H}}_q(Lr(p)/H_{\text{Lip}}(p; L)) \ge H(p)\,. \tag{5}$$

Proof. Fix a typical source sequence $x^0 = (x_1^0, \dots, x_n^0) \in \mathsf{Typ}_n(p)$ and let $z^0 = \mathsf{F}(x^0)$. Notice that most of the sequences in $\mathsf{Typ}_n(p)$ have distance at most $nr(p) + o(n)$ from x^0. Therefore $\mathsf{F}(\mathsf{Typ}_n(p))$ must lie within a ball of radius $nr(p)L + o(n)$ from z^0. The volume of this ball is $q^{m\overline{\mathcal{H}}(nr(p)L/m)+o(n)}$ Since $|\mathsf{Typ}_n(p)| \doteq q^{nH(p)}$, we must have $m\overline{\mathcal{H}}(nr(p)L/m) \ge nH(p) + o(n)$ which concludes the proof. □

Using $H_{\text{Lip}}(p; L) \ge H(p)$, this result implies that the Lipschitz entropy is bounded away from the Shannon entropy whenever $L \le L_*(p) \equiv (q-1)H(p)/qr(p)$. For our running example, we get $L_*(\epsilon) = \mathcal{H}_2(\epsilon)/4\epsilon(1-\epsilon) = 4\log(1/\epsilon) + O(\epsilon)$.

A better result is obtained for bounded degree entropy by generalizing Gallager bound.

[1] For $s \in \mathcal{X}$, $\theta_x(s)$ is the fraction of indexes $i \in [n]$ such that $x_i = s$. For two distributions q_1, q_2, we let $||q_1 - q_2||_{\text{TV}} \equiv \frac{1}{2}\sum_x |q_1(x) - q_2(x)|$ be their total variation distance.

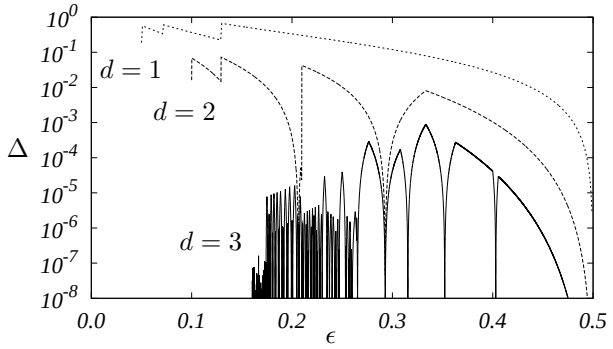

Fig. 1. Generalized Gallager bound for compression rate using non-linear d-sparse codes (from Eq. (6) with $\rho = 0.9$.) Here we plot the lower bound on the relative overhead $\Delta(\epsilon; d)$ due to sparsity for iid Bernoulli(ϵ).

Proposition II.2 (Generalized Gallager bound)**.** *Let $d \geq 1$ be an integer and $A(p, d)$ be the supremum value of $h \in [0, 1]$ such that there exists a function $\mathsf{f} : \mathcal{X}^d \to \{0, 1\}$ satisfying $H(\mathsf{f}(X_1, \dots, X_d)) = h$, for $X_1, \dots X_d$ iid with distribution p. Then, for any $\rho \in (0, 1)$*

$$H_{\deg}(p; d) \geq \frac{H(p)}{1 - \rho + \rho\, A(p, \lfloor d/\rho H(p) \rfloor)}\,. \tag{6}$$

It is easily observed that for all d, almost all distributions p satisfy $A(p, d) < 1$.

Proof. Consider the factor graph associated to F. The average degree of the factor node side is $nd/m \leq d/H(p)$. Therefore, fixing $\rho \in (0, 1)$, at least a fraction ρ of these node have degree $d/\rho H(p)$ or smaller. The entropy of the corresponding entries of the compressed vector $H(Z_i) \leq A(p, \lfloor d/\rho H(p) \rfloor)$. Since we always have $H(Z_i) \leq 1$, we get:

$$H(Z) \leq \sum_{i=1}^{m} H(Z_i) \leq m(1 - \rho) + m\rho\, A(p, \lfloor d/\rho H(p) \rfloor)\,.$$

By Fano inequality, the error probability cannot vanish unless $H(Z) \geq H(X) - o(n)$ which implies the thesis. □

We can define the relative overhead $\Delta(p; d)$ due to sparsity by letting $H_{\deg}(p; d) \equiv H(p)[1 + \Delta(p; d)]$. In Fig. 1 we plot the lower bound on the relative overhead that is implied by the last Proposition. As suggested by this plot, the lower bound vanishes on a (countable) set of 'special values' of ϵ, that increases in d.

This is in fact a general (and intrinsic) phenomenon. For any alphabet $\mathcal{X}$ and any integer $k \geq 1$, let $\mathcal{D}_k(\mathcal{X})$ denote the set of probability distributions p over $\mathcal{X}$ such that there exists a function $\mathsf{f} : \mathcal{X}^k \to \mathcal{X}$ for which the following holds. If X_1,..., X_k are iid with distribution p, then $\mathsf{f}(X_1, \dots, X_k)$ is uniform in $\mathcal{X}$. For instance $(1 - \epsilon, \epsilon) \in \mathcal{D}_2(\{0, 1\})$ for $\epsilon = 1/\sqrt{2}$ since $\mathsf{f}(X_1, X_2) = X_1 \,\mathsf{AND}\, X_2$ is unbiased when X_1 and X_2 are iid Bernoulli(ϵ). Clearly $\mathcal{D}_k(\mathcal{X})$ is finite and increasing in k, and $\mathcal{D}(\mathcal{X}) \equiv \cup_k \mathcal{D}_k(\mathcal{X})$ is dense in the $(|\mathcal{X}| - 1)$-dimensional simplex of probability distributions over $\mathcal{X}$.

If $p \in \mathcal{D}_k(\mathcal{X})$ then, for $d \geq k\, H(p)$, the lower bound (6) reduces to the trivial one $H_{\deg}(p; d) \geq H(p)$. This raises a natural question: Is $H_{\deg}(p; d) = H(p)$ for $p \in \mathcal{D}(\mathcal{X})$ and d sufficiently large (but bounded)? Concretely: is there a bounded d and a sequence of d-sparse codes that achieve information theoretically optimal compression rate?

C. Upper bounds and code construction

Let us review the compression rate achieved by linear codes.

Proposition II.3. *For any distribution p over $\mathcal{X}$, there exist $\lambda(p) \in (0, 1)$ and $C > 0$ such that*

$$H_{\deg}(p; d) \leq H(p) + C\, \lambda(p)^d\,. \tag{7}$$

Proof. The elements of of $\mathcal{X}$ are identified with the group of integers modulo $|\mathcal{X}| \equiv q$. One then takes $\mathsf{F}(x) = \mathbb{H}x \mod q$ where $\mathbb{H}$ is the parity check matrix of a random LDPC code. The upper bound follows from the analysis of Ref. [8]. □

Theorem II.1. *If $p \in \mathcal{D}_k(\mathcal{X})$ then there exists $d_*(p)$ bounded such that, for any $d \geq d_*(p)$, $H_{\deg}(p; d) = H(p)$.*

The proof of this statement is deferred no the next Section. Here we limit ourselves to describing the construction that achieves compression rate equal to the source entropy $H(p)$.

The encoding map is obtained by summing two functions $\mathsf{L} : \mathcal{X}^n \to \mathcal{X}^m$ and $\mathsf{N} : \mathcal{X}^n \to \mathcal{X}^m$:

$$\mathsf{F}(x) = \mathsf{L}(x) \oplus \mathsf{N}(x)\,, \tag{8}$$

where $\oplus$ denotes component-wise sum modulo q. The function L is linear:

$$\mathsf{L}(x) = \mathbb{H}x \qquad \mod q\,, \tag{9}$$

with $\mathbb{H}$ the parity check matrix of a (q-ary) LDPC code with blocklength n. In order to define the non-linear component, we fix $\mathsf{f} : \mathcal{X}^k \to \{0, 1\}$ such that, if $X_1, \dots, X_k$ are iid with distribution p, $\mathsf{f}(X_1, \dots, X_k)$ is uniformly random in $\mathcal{X}$. Then, for each $a \in [m]$ we fix r_a ordered subsets $a(1), \dots, a(r_a) \subseteq [n]$ of the variable indexes, with common size $|a(s)| = k$. Denoting by $x_{a(s)} = \{x_i : i \in a(s)\}$, we let

$$\mathsf{N}_a(x) = \mathsf{f}(x_{a(1)}) \oplus \cdots \oplus \mathsf{f}(x_{a(r)})\,. \tag{10}$$

In order to complete the definition, we need to specify the sparse matrix $\mathbb{H}$ as well as the indexes sets $a(s)$. We will consider a random construction and prove that the condition (3) holds in expectation over the ensemble. More precisely, we use for $\mathbb{H}$ a standard irregular LDPC ensemble, which corresponds to drawing the graph uniformly given the node degrees. A simple choice consists in having regular variable nodes of degree v, and irregular check nodes with two consecutive degrees c and $c + 1$ in such a way to match the number of edges on the two sides.

We proceed analogously for the subsets $a(s)$. We draw a random factor graph on the same vertices sets, with regular variable nodes of degree l, and irregular check nodes with two degrees kr and $k(r + 1)$ such to match the number of edges. Then, for any $a \in [m]$ we let $a(1)$ be the set of the first[2] k variable nodes it is connected to, $a(2)$ the following k nodes, and so on (thus $r_a = r$ or $r + 1$).

[2]According to the standard definition of LDPC ensembles, edges are labeled.

In the next Section we will show that, for any rate above $H(p)$, we can choose v, c, l, r bounded uniformly in the rate, in such a way to achieve vanishing error probability.

III. Proof of the main theorem

In order to prove our main theorem, we need to upper bound the probability of a 'collision', i.e. the probability that, for a random source vector x, there exists a different vector $y \in \mathsf{Typ}_n(p)$ with $\mathsf{F}(x) = \mathsf{F}(y)$. The basic idea is that the linear component of the code will resist collisions among atypically close sources x, y, while the non-linear part will take care of x, y at typical distance.

Lemma III.1. *Let $\mathsf{L} : \mathcal{X}^n \to \mathcal{X}^m$, $\mathsf{N} : \mathcal{X}^n \to \mathcal{X}^m$ be two independent random functions (non necessary linear, or non-linear), and $\mathsf{F} \equiv \mathsf{L} \oplus \mathsf{N}$ (the sum being modulo q.) Then, for any $x \in \mathsf{Typ}_n(p)$ and any partition $\{A, B\}$ of $\mathsf{Typ}_n(p) \setminus \{x\} \equiv T$:*

$$\sum_{y \in T} \mathbb{P}\{\mathsf{F}(y) = \mathsf{F}(x)\} \le \sum_{y \in A} \sup_{u \in \mathcal{X}^m} \mathbb{P}\{\mathsf{L}(y) = \mathsf{L}(x) \oplus u\} + \sum_{y \in B} \sup_{u \in \mathcal{X}^m} \mathbb{P}\{\mathsf{N}(x) = \mathsf{N}(y) \oplus u\} .$$

Proof. Clearly the sum to be bounded is equal to

$$\sum_{y \in A} \mathbb{P}\{\mathsf{F}(y) = \mathsf{F}(x)\} + \sum_{y \in B} \mathbb{P}\{\mathsf{F}(y) = \mathsf{F}(x)\} .$$

For $y \in A$, we write $\mathbb{P}\{\mathsf{F}(y) = \mathsf{F}(x)\} = \mathbb{P}\{\mathsf{L}(y) = \mathsf{L}(x) \oplus u\}$ where $u = \mathsf{N}(x) \ominus \mathsf{N}(y)$, whence the sum over A can be upper bounded as in the statement. The sum over $y \in B$ is bounded analogously. □

For any two source vectors $x, y \in \mathcal{X}^n$, we let θ_x, θ_y denote their *type*, and θ_{xy} their *joint type* (thus $\theta_{xy}(s,t)$ is the number of entries such that $x_i = s$ and $y_i = t$.) Finally, given two distributions q_1, q_2 on $\mathcal{X}$, we denote by $q_1 \times q_2$ the product distribution on $\mathcal{X} \times \mathcal{X}$ whose marginals are q_1 and q_2.

Proposition III.1. *For all $x \in \mathsf{Typ}_n(p)$, $\epsilon > 0$, define*

$$A_{x,\epsilon} = \{y \in \mathsf{Typ}_n(p) : ||\theta_{xy} - p \times p||_{\text{TV}} < \epsilon\}. \qquad (11)$$

Then there exist ϵ, and for any $R \ge H(p)$, l, r uniformly bounded, such that

$$\lim_{n \to \infty} \sum_{y \in A_{x,y}} \sup_{u \in \mathcal{X}^m} \mathbb{P}\{\mathsf{N}(x) = \mathsf{N}(y) \oplus u\} = 0 . \qquad (12)$$

Proof. By the definition of N, $\mathbb{P}\{\mathsf{N}(x) = \mathsf{N}(y) \oplus u\}$ only depends on x, y through their joint type $\theta_{x,y}$ and on u through its type. To simplify our presentation, we shall first estimate this probability for $u = 0$, and then explain the differences or a general u. Further, we will assume that the graph that defines $\mathsf{N}(\,\cdot\,)$ is regular. Although this is typically not true, the proof in the irregular case only requires a slightly lenghtier calculation.

Let $\mathrm{P}(\theta_{x,y}) \equiv \mathbb{P}\{\mathsf{N}(x) = \mathsf{N}(y)\}$, and define $D_{x,\epsilon}$ to be the set of probability distributions θ over $\mathcal{X} \times \mathcal{X}$ such that $\sum_t \theta(s,t) = \theta_x(s)$ $||\sum_s \theta(s,\,\cdot\,) - p(\,\cdot\,)||_{\text{TV}} \le n^{-\alpha}$, and $||\theta - p \times p||_{\text{TV}} \le n\epsilon$. Then, by reordering the sum over y according to the joint type of x, y, ad using standard bounds for binomial coefficients, we get

$$\sum_{y \in A} \mathbb{P}\{\mathsf{N}(x) = \mathsf{N}(y)\} \le n^q \sup_{\theta \in D_{x,\epsilon}} q^{n[H(\theta) - H(\theta_x)]} \, \mathrm{P}(\theta) .$$

We claim that $\mathrm{P}(\theta) \le n^C \, q^{-nR\, I(\theta)}$, where, for θ close to $p \times p$, $I(\theta) = 1 + O(||\theta - p \times p||^3)$. Before proving the claim, let us show that it implies the thesis. First we write the supremum as a supremum over distributions ξ on $\mathcal{X}$, such that $||\xi - p||_{\text{TV}} \le n^{-\alpha}$ and then over $\theta \in D_{x,\epsilon}$ such that $\sum_s \theta(s,t) = \xi(t)$. The above is therefore bounded by $n^{q+C} q^{-nJ}$ where, for some constants C, D

$$J = \sup_{\xi,\theta} \left\{ H(\theta) - H(\theta_x) - R + D||\theta - p \times p||^3 \right\}$$
$$\le -R + \sup_{\xi,\theta} \left\{ H(\xi) - C||\theta - \theta_x \times \xi||^2 + D||\theta - p \times p||^3 \right\} ,$$

where the sup is taken under the constraints mentioned above. By triangular inequality the right hand side is upper bounded by $-R + H(p) + O(n^{-\alpha}) + \sup_{\xi,\theta}\{-C||\theta - p \times p||^2 + D||\theta - p \times p||^3\}$. Now, for ϵ small enough (but uniformly in R!), the sup is realized at $\theta = p \times p$. The left hand side of Eq. (12) is therefore bounded by $q^{-n[R - H(p) + O(n^{-\alpha})]} \to 0$.

We now pass to proving the claim $\mathrm{P}(\theta) \le n^C \, q^{-nR\, I(\theta)}$ with $I(\theta) = 1 + O(||\theta - p \times p||^3)$. Let $K \equiv kr$. To each factor node in the graph defining N, we can associate the an element $\vec{s} = (x_1, \dots, x_K, y_1, \dots, y_K) \in \mathcal{X}^{2K}$ listing the values of the adjacent variables in the two vectors x, y. Denote by $\omega(\vec{s})$ the type of the vector $(\vec{s}_1, \dots, \vec{s}_m)$. The probability (over the graph distribution) of seeing a type ω when the joint type of x and y is θ_{xy}, is given by a standard combinatorial calculation

$$\frac{1}{(nl)!} \; \frac{m!}{\prod_{\vec{s} \in \mathcal{X}^{2K}} (m\omega(\vec{s}))!} \prod_{(s,t) \in \mathcal{X}^2} (nl\theta(s,t))! \le \qquad (13)$$
$$\le n^{C'} \, q^{m[H(\omega) - kH(\theta)]} .$$

For $\vec{x} = (x_1, \dots, x_K) \in \mathcal{X}^K$, let $\varphi(\vec{x})$ denote the function that is computed at factor nodes by the encoder. Explicitly

$$\varphi(\vec{x}) = \mathsf{f}(x_1, \dots, x_k) \oplus \cdots \oplus \mathsf{f}(x_{k(r-1)+1}, \dots, x_{kr}) . \qquad (14)$$

Then $\mathrm{P}(\theta) \le n^C q^{-nR\, I(\theta)}$, where

$$I(\theta) = \inf \{ kH(\theta) - H(\omega) : \omega \in \Omega \} . \qquad (15)$$

Here the minimization domain Ω is the set of all types such that $\omega(\vec{x}, \vec{y}) = 0$ unless $\varphi(\vec{x}) = \varphi(\vec{y})$ and $\sum_{\vec{x}^{(i)}, \vec{y}^{(i)}} \omega(\vec{x}, \vec{y}) = \theta(x_i, y_i)$ (here the sum excludes x_i, y_i.) It is a minimization problem for a concave function on a convex domain, and can be solved introducing Lagrange multipliers. The optimal ω has the form

$$\omega(\vec{x}, \vec{y}) = \frac{1}{Z(\rho)} \prod_{i=1}^{K} \rho(x_i, y_i) \, \mathbb{I}\{\varphi(\vec{x}) = \varphi(\vec{y})\} ,$$

where $\rho(x, y)$ can be chosen as a probability distribution over $\mathcal{X} \times \mathcal{X}$. Denote by $\mathbb{P}_\rho$ probability over iid couples (X_1, Y_1), (X_2, Y_2), …, (X_K, Y_K). The stationarity condition for ρ is

$$\frac{\rho(x_i, y_i)}{\theta(x_i, y_i)} = \frac{\mathbb{P}_\rho\{\varphi(X_1 \dots x_i \dots X_k) = \varphi(Y_1 \dots y_i \dots Y_k)\}}{\mathbb{P}_\rho\{\varphi(X_1 \dots X_k) = \varphi(Y_1 \dots Y_k)\}}$$

The exponential rate $I(\theta)$ can be written in terms of ρ satisfying this equation:

$$I(\theta) = -kD(\theta||\rho) - \log_q \mathbb{P}_\rho\{\varphi(X_1 \dots X_K) = \varphi(Y_1 \dots Y_K)\} \,. \tag{16}$$

If we take $\theta = p \times p$, then the above equation is solved by $\rho = p \times p$. Indeed this implies that $\varphi(X_1 \dots X_K)$ is uniform in $\mathcal{X}$ and independent from $\varphi(Y_1 \dots Y_K)$, and hence $\mathbb{P}_\rho\{\varphi(X_1 \dots X_K) = \varphi(Y_1 \dots Y_K)\} = 1/q$. However, if $r \geq 2$, $\mathbb{P}_\rho\{\varphi(X_1 \dots x_i \dots X_K) = \varphi(Y_1 \dots y_i \dots Y_K)\} = 1/q$ for any x_i, y_i as well (this follows from the definition of φ, cf. Eq. (14)), and therefore $\rho = p \times p$ solves the stationarity condition. On this point we get $I(\theta) = 1$.

Consider now $\theta \neq p \times p$. For ρ close to $p \times p$ one can show that $\mathbb{P}_\rho\{\varphi(X_1 \dots X_K) = \varphi(Y_1 \dots Y_K)\} = 1/q + O(||\rho - p \times p||^r)$ and $\mathbb{P}_\rho\{\varphi(X_1 \dots x_i \dots X_K) = \varphi(Y_1 \dots y_i \dots Y_K)\} = 1/q + O(||\rho - p \times p||^{r-1})$. Applying perturbation theory to the stationarity condition, one obtains $\rho(x,y)/\theta(x,y) = 1 + O(||\theta - p \times p||^{r-1})$. Substituting in Eq. (16) one obtains $I(\theta) = 1 + O(||\theta - p \times p||^{r-1})$ which proves our claim for $r \geq 4$.

Let us finally comment on the changes in the last part of the proof for irregular ensembles and $u \neq 0$. In these cases $\mathbb{P}\{\mathsf{N}(x) = \mathsf{N}(y) \oplus u\}$ also depends on the type of u and on the fraction of factor nodes of each degree (here we think that variable nodes are still regular.) The calculation presented above can be repeated provide a different type ω is introduced for each value class of factor nodes, depending on their degree and value of the vector u. Despite these changes, the main steps, as long as the thesis remain true. □

Our last auxiliary result concerns the linear code.

Proposition III.2. *Given $x \in \mathsf{Typ}_n(p)$ and $\epsilon > 0$, define $A_{x,\epsilon}$ as in Eq. (11) and $B_{x,\epsilon} \equiv \mathsf{Typ}_n(p) \setminus A_{x,\epsilon} \cup \{x\}$. Then there exists degrees c, v uniformly bounded for $R \geq H(p)$, such that*

$$\lim_{n\to\infty} \sum_{y \in A \setminus A_{x,y}} \sup_{u \in \mathcal{X}^m} \mathbb{P}\{\mathsf{L}(x) = \mathsf{L}(y) \oplus u\} = 0 \,. \tag{17}$$

Proof. The proof is completely analogous to the one in Proposition II.3, and to the analysis of q-ary LDPC codes in [8]. We omit details for lack of space and only stress a point for the reader who is familiar with this type of derivation. Excluding the region $A_{x,\epsilon}$ from the sum in Eq. (17) is equivalent (in the channel coding interpretation) to exclude the dominant error type for large degrees. This allows to achieve vanishing error probability at any rate above $H(p)$ (in the channel coding language, any rate below capacity), with bounded degree. □

We are finally in position of proving our main result.

Proof of Theorem II.1. First of all we need to specify the decoding map. For $z \in \mathcal{X}^m$ we let

$$\widehat{\mathsf{F}}(z) = \begin{cases} x & \text{if } \mathsf{Typ}_n(p) \cap \mathsf{F}^{-1}(z) = \{x\} \,, \\ \mathsf{error} & \text{if } |\mathsf{Typ}_n(p) \cap \mathsf{F}^{-1}(z)| \neq 1 \,. \end{cases} \tag{18}$$

The average error probability is then upper bounded by

$$\mathbb{P}\{|\mathsf{Typ}_n(p) \cap \mathsf{F}^{-1}(\mathsf{F}(X))| \neq 1, \ X \in \mathsf{Typ}_n(p)\} + \mathbb{P}\{X \notin \mathsf{Typ}_n(p)\} \,.$$

Since the second term vanishes as $n \to \infty$, it is sufficient to show that, for any $x \in \mathsf{Typ}_n(p)$, $\mathbb{P}\{|\mathsf{Typ}_n(p) \cap \mathsf{F}^{-1}(\mathsf{F}(x))| \neq 1\}$ vanishes as $n \to \infty$. This is in turn upper bounded as

$$\mathbb{P}\{\exists y \in \mathsf{Typ}_n(p) \setminus \{x\} : \mathsf{F}(y) = \mathsf{F}(x)\} \leq \leq \sum_{y \in \mathsf{Typ}_n(p) \setminus \{x\}} \mathbb{P}\{\mathsf{F}(y) = \mathsf{F}(x)\} \,.$$

The proof is completed by partitioning $\mathsf{Typ}_n(p) \setminus \{x\}$ into $A_{x,\epsilon}$ and $B_{x,\epsilon}$ as described above, and ubber bounding it as in Lemma III.1. Then we fix $r \geq 4$ and ϵ in such a way for Proposition III.1 to hold, which implies that the sum over $A_{x,\epsilon}$ vanishes. Finally c, v are chosen using Proposition III.2, in such a way that the sum over $B_{x,\epsilon}$ vanishes as well. □

ACKNOWLEDGMENT

We thank Yuval Peres and Martin Wainwright for stimulating discussions on the topic of this paper. A.M. is partially supported by a Terman award and a Filo and Yang fellowship. E.M. is partially supported by a Sloan fellowship in mathematics, DOD ONR grant N0014-07-1-05-06, NSF Career Award DMS 0548249, BSF grant 2004105 and NSF grant DMS 0528488. Part of this work was carried out during a visit of the authors to Microsoft Research, Redmond.

REFERENCES

[1] T. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley Interscience, New York, 1991

[2] S. Dharmapurikar, A. Kabbani, Y. Lu, A. Montanari and B. Prabhakar. "Counter Braids: A Novel Counter Architecture for Per-Flow Measurement." Sigmetrics, June 2008

[3] Y. Lu, A. Montanari, and B. Prabhakar. "Detailed Network Measurements Using Sparse Graph Counters: The Theory" Proc. 45$^{\text{th}}$ Allerton Conference, Monticello IL, October 2007

[4] R. G. Gallager, *Low-Density Parity-Check Codes*. MIT Press, Cambridge, Massachussetts, 1963.

[5] T. Richardson and R. Urbanke, *Modern Coding Theory*, draft available at `http://lthcwww.epfl.ch/mct/index.php`

[6] G. Caire, S. Shamai, and S. Verdú. "Noiseless data compression with low density parity check codes." In P. Gupta and G. Kramer, editors, *Dimacs Series in Mathematics and Theoretical Computer Science*, pages 224–235. AMS, 2004.

[7] G. Miller and D. Burshtein, "Asymptotic enumeration method for analyzing LDPC codes," *IEEE Trans. Inform. Theory*, vol. 50, no. 6, pp. 1115–1131, June 2004.

[8] A. Bennatan and D. Burshtein. "On the application of LDPC Codes to Arbitrary Discrete-Memoryless Channels." *IEEE Trans. Inform. Theory*, 50:417–438, 2004.

[9] M. J. Wainwright and E. Maneva, "Lossy source encoding via message-passing and decimation over generalized codewords of LDGM codes," *Proc. IEEE Int. Symp. on Inf. Theory*, Adelaide, September 2005

[10] S. Ciliberti, M. Mézard and R. Zecchina, "Lossy data compression with random gates," Phys. Rev. Lett. 95 (2005) 038701

[11] N. Linial, "Finite Metric Spaces - Combinatorics, Geometry and Algorithms," Proc. Int. Cong. Math. III, 573–586 Beijing, 2002

[12] M. Keane and M. Smorodinsky, "Bernoulli schemes of the same entropy are finitarily isomormphic," Ann. of Math. 109 (1970) 397-406