Concentration of Measure Inequalities and Their Communication and Information-Theoretic Applications

Maxim Raginsky Igal Sason

Abstract

During the last two decades, concentration of measure has been a subject of various exciting developments in convex geometry, functional analysis, statistical physics, high-dimensional statistics, probability theory, information theory, communications and coding theory, computer science, and learning theory. One common theme which emerges in these fields is probabilistic stability: complicated, nonlinear functions of a large number of independent or weakly dependent random variables often tend to concentrate sharply around their expected values. Information theory plays a key role in the derivation of concentration inequalities. Indeed, both the entropy method and the approach based on transportation-cost inequalities are two major information-theoretic paths toward proving concentration.

This brief survey is based on a recent monograph of the authors in the *Foundations and Trends in Communications and Information Theory*, and a tutorial given by the authors at ISIT 2015. It introduces information theorists to three main techniques for deriving concentration inequalities: the martingale method, the entropy method, and the transportation-cost inequalities. Some applications in information theory, communications, and coding theory are used to illustrate the main ideas.

I. Introduction

Concentration inequalities bound from above the probability that a random variable Z deviates from its mean, median or some other typical value by a given amount. These inequalities have been studied for several decades, with some fundamental and substantial contributions during the last two decades. Very roughly speaking, the concentration-of-measure phenomenon can be stated in the following simple way: "A random variable that depends in a smooth way on many independent random variables (but not too much on any of them) is essentially constant" [1]. Informally, this amounts to saying that such a random variable Z concentrates around its expected value, $\mathbb{E}[Z]$, in such a way that the probability of the event $\{|Z - \mathbb{E}[Z]| \geq t\}$, for a given $t > 0$, decays exponentially in some power of t. Detailed treatments of the concentration-of-measure phenomenon, including historical accounts, can be found, e.g., in [2]–[9].

In recent years, concentration inequalities have been intensively studied and used as a powerful tool in various areas. These include convex geometry, functional analysis, statistical physics, probability theory, statistics, information theory, communications and coding theory, learning theory, and computer science. Several techniques have been developed so far to prove concentration inequalities. This survey paper focuses on three such techniques which are studied in our tutorial [9] and references therein:

- The martingale method (see, e.g., [6], [10], [11], [8, Chapter 7], [12], [13]), and its information-theoretic applications (see, e.g., [14] and references therein, [15]).
- The entropy method and logarithmic Sobolev inequalities (see, e.g., [3, Chapter 5], [4] and references therein).

Maxim Raginsky is with Department of Electrical and Computer Engineering, Coordinated Science Laboratory, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA (e-mail: maxim@illinois.edu).

I. Sason is with the Department of Electrical Engineering, Technion–Israel Institute of Technology, Haifa 32000, Israel (e-mail: sason@ee.technion.ac.il).

- Transportation-cost inequalities which originated from information theory (see, e.g., [3, Chapter 6], [16], [17] and references therein).

Our goal here is to give the reader a quick preview of the vast field of concentration inequalities and their applications in information theory, communications and coding. Therefore, we state most of the theorems and lemmas without proofs; occasionally, we provide sketches or brief outlines. More details can be found in our monograph [9] and the slides of our ISIT'15 tutorial.[1]

II. The basic toolbox

Our objective is to derive tight upper bounds on the tail probabilities

$$\mathbb{P}[Z \geq \mathbb{E}[Z] + t] \text{ and } \mathbb{P}[Z \leq \mathbb{E}[Z] - t], \qquad \forall t > 0$$

where $Z = f(X_1, \ldots, X_n)$ is an arbitrary function of n independent random variables $X_1, \ldots, X_n$. To get an idea of what we can expect, let us first recall Chebyshev's inequality:

$$\mathbb{P}[|Z - \mathbb{E}[Z]| \geq t] \leq \frac{\mathrm{Var}[Z]}{t^2}, \qquad \forall t > 0.$$

This inequality shows that the tail probability decays with t, and that the rate of decay is proportional to the variance of Z. Thus, the variance of Z gives an idea about how tightly Z concentrates around its mean. In fact, if Z takes values in a bounded interval, then we can upper-bound the variance of Z only in terms of the length of this interval:

Lemma 1. *Let Z be a random variable taking values in an interval $[a, b]$. Then*

$$\mathrm{Var}[Z] \leq \tfrac{1}{4}(b-a)^2. \tag{1}$$

This bound is sharp: if Z only takes the two values a and b with equal probability, then $\mathrm{Var}[Z] = \frac{1}{4}(b-a)^2$.

Proof: Recall that $\mathrm{Var}[Z] \leq \mathbb{E}[(Z-c)^2]$ for all $c \in \mathbb{R}$. Letting $c = \frac{a+b}{2}$, we obtain (1). The case of equality is an easy calculation. ∎

Thus, for a bounded Z in an interval $[a, b]$, Chebyshev's inequality gives

$$\mathbb{P}[|Z - \mathbb{E}[Z]| \geq t] \leq \frac{(b-a)^2}{4t^2}.$$

Much stronger concentration inequalities can be derived, however, for bounded random variables. Using Markov's inequality, for every $\lambda > 0$ we have

$$\begin{aligned}\mathbb{P}\left[Z - \mathbb{E}[Z] \geq t\right] &= \mathbb{P}\left[e^{\lambda(Z-\mathbb{E}[Z])} \geq e^{\lambda t}\right] \\ &\leq e^{-(\lambda t - \psi(\lambda))},\end{aligned}$$

where $\psi(\lambda) \triangleq \log \mathbb{E}[e^{\lambda(Z-\mathbb{E}[Z])}]$ is the *logarithmic moment-generating function* of Z. Optimizing over λ, we get the *Chernoff bound*

$$\mathbb{P}\left[Z \geq \mathbb{E}[Z] + t\right] \leq e^{-\psi^\star(t)},$$

where $\psi^\star(t) \triangleq \sup_{\lambda \geq 0}\left[\lambda t - \psi(\lambda)\right]$ is the Legendre dual of ψ. For example, if $Z \sim N(0, \sigma^2)$ (Gaussian with mean 0 and variance σ^2), we have $\psi(\lambda) = \lambda^2\sigma^2/2$, and $\psi^\star(t) = t^2/2\sigma^2$. With

[1] Part 1 (The martingale method):
http://webee.technion.ac.il/people/sason/raginsky_sason_ISIT_2015_tutorial_part_1.pdf.
Part 2 (The entropy method and transportation-cost inequalities):
http://webee.technion.ac.il/people/sason/raginsky_sason_ISIT_2015_tutorial_part_2.pdf.

this in mind, we say that a random variable Z is σ^2-*subgaussian* if $\psi(\lambda) \leq \lambda^2\sigma^2/2$. For a subgaussian random variable, we obtain $\psi^\star(t) \geq t^2/2\sigma^2$, which gives the tail bound

$$\mathbb{P}[Z \geq \mathbb{E}[Z] + t] \leq e^{-t^2/2\sigma^2}, \qquad \forall t > 0.$$

Thus, the whole affair hinges on our ability to prove that the random variable Z of interest is subgaussian.

To start with, a bounded random variable is subgaussian:

Lemma 2 (Hoeffding [11])**.** *For a random variable Z taking values in an interval $[a, b]$, we have*

$$\log \mathbb{E}[e^{\lambda(Z-\mathbb{E}[Z])}] \leq \tfrac{1}{8}\,\lambda^2(b-a)^2. \tag{2}$$

Proof: We give a simple probabilistic proof, which has the additional benefit of highlighting the role of the tilted distribution. Let $P = \mathcal{L}(Z)$,[2] and introduce its *exponential tilting* $P^{(t)}$: for an arbitrary sufficiently regular function $f\colon \mathbb{R} \to \mathbb{R}$,

$$\mathbb{E}_{P^{(t)}}[f(Z)] \triangleq \frac{\mathbb{E}_P[f(Z)e^{tZ}]}{\mathbb{E}_P[e^{tZ}]}.$$

Since Z is supported on $[a, b]$ under P, the same holds under $P^{(t)}$ as well. Therefore, by Lemma 1,

$$\mathrm{Var}_{P^{(t)}}[Z] \leq \tfrac{1}{4}\,(b-a)^2.$$

On the other hand,

$$\begin{aligned}\mathrm{Var}_{P^{(t)}}[Z] &= \frac{\mathbb{E}_P[Z^2e^{tZ}]}{\mathbb{E}_P[e^{tZ}]} - \left(\frac{\mathbb{E}_P[Ze^{tZ}]}{\mathbb{E}_P[e^{tZ}]}\right)^2 \\ &= \psi''(t).\end{aligned}$$

Therefore,

$$\psi''(t) \leq \tfrac{1}{4}\,(b-a)^2$$

for all t. Integrating and using the fact that

$$\psi(0) = \psi'(0) = 0,$$

we get (2). ∎

Both the martingale method and the entropy method are just elaborations of these basic tools, which are applicable to an arbitrary bounded real-valued random variable. However, one should keep in mind that concentration of measure is a *high-dimensional* phenomenon: we are interested in situations when Z is a function of many independent random variables $X_1, \ldots, X_n$, and we can often quantify the "sensitivity" of f to changes in each of its arguments while the others are kept fixed. This suggests that we may get a handle on the high-dimensional concentration properties of Z by breaking up the problem into n one-dimensional subproblems involving only one of the X_i's at a time. Whenever such a divide-and-conquer approach is possible, we speak of *tensorization*, by which we mean that some quantity involving the distribution of

$$Z = f(X_1, \ldots, X_n)$$

(e.g., variance or relative entropy) can be related to the sum of similar quantities involving the *conditional* distribution of each X_i given

$$\overline{X}^i \triangleq (X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n).$$

[2]The notation $\mathcal{L}(Z)$ stands for the law, or probability distribution, of the random variable Z.

III. THE MARTINGALE METHOD

The basic idea behind the martingale method is to start with the *Doob martingale decomposition*

$$Z - \mathbb{E}[Z] = \sum_{k=1}^{n} \xi_k, \tag{3}$$

where

$$\xi_k \triangleq \mathbb{E}[Z|X^k] - \mathbb{E}[Z|X^{k-1}] \tag{4}$$

with

$$X^k \triangleq (X_1, \ldots, X_k)$$

and then to exploit any information about the sensitivity of f to local changes in its arguments in order to control the sizes of the increments ξ_k. As a warm-up, consider the following inequality, first obtained in a restricted setting by Efron and Stein [18] and generalized by Steele [19]:

Lemma 3 (Efron–Stein–Steele)**.** *Let $Z = f(X^n)$ where $X_1, \ldots, X_n$ are independent, then*

$$\mathrm{Var}[Z] \leq \sum_{k=1}^{n} \mathbb{E}\left[\mathrm{Var}[Z|\overline{X}^k]\right]. \tag{5}$$

Proof: We exploit the fact that $\{\xi_k\}_{k=1}^n$ in (4) is a *martingale difference sequence* with respect to X^n, i.e.,

$$\mathbb{E}[\xi_k|X^{k-1}] = 0 \tag{6}$$

for all $k \in \{1, \ldots, n\}$. Hence, since $\mathbb{E}[\xi_k \xi_l] = 0$ for $k \neq l$,

$$\mathrm{Var}[Z] = \sum_{k=1}^{n} \mathbb{E}[\xi_k^2]. \tag{7}$$

The independence of $X_1, \ldots, X_n$ in (4) yields

$$\xi_k = \mathbb{E}\big[Z - \mathbb{E}[Z|\overline{X}^k] \,|\, X^k\big]$$

and, from Jensen's inequality,

$$\xi_k^2 \leq \mathbb{E}\big[\big(Z - \mathbb{E}[Z|\overline{X}^k]\big)^2 \,|\, X^k\big].$$

Due to the independence of $X_1, \ldots, X_n$, this in turn yields

$$\begin{aligned} \mathbb{E}[\xi_k^2] &\leq \mathbb{E}[\big(Z - \mathbb{E}[Z|\overline{X}^k]\big)^2] \\ &= \mathbb{E}[\mathrm{Var}[Z|\overline{X}^k]]. \end{aligned} \tag{8}$$

Substituting (8) into (7) yields (5). ∎

The Efron–Stein–Steele inequality is our first example of tensorization: it upper-bounds the variance of $Z = f(X_1, \ldots, X_n)$ by the sum of the expected values of the conditional variances of Z given all but one of the variables. In other words, we say that $\mathrm{Var}[f(X_1, \ldots, X_n)]$ tensorizes. This fact has immediate useful consequences. For example, we can use any convenient technique for upper-bounding variances to control each term on the right-hand side of (7), and thus obtain many useful variants of the Efron–Stein–Steele inequality:

1) For every random variable U with a finite second moment,

$$\operatorname{Var}[U] = \tfrac{1}{2}\,\mathbb{E}[(U-U')^2]$$

where U' is an i.i.d. copy of U. Thus, if we let

$$Z'_k = f(X_1,\dots,X_{k-1},X'_k,X_{k+1},\dots,X_n),$$

where X'_k is an i.i.d. copy of X_k, then Z and Z'_k are i.i.d. given $\overline{X}^k$. This implies that

$$\operatorname{Var}[Z|\overline{X}^k] = \tfrac{1}{2}\,\mathbb{E}\Big[(Z-Z'_k)^2\Big|\overline{X}^k\Big]$$

for $k \in \{1,\dots,n\}$, yielding the following variant of the Efron–Stein–Steele inequality:

$$\operatorname{Var}[Z] \le \tfrac{1}{2}\sum_{i=1}^{n}\mathbb{E}[(Z-Z'_k)^2]. \tag{9}$$

This inequality is sharp: if $Z=\sum_{k=1}^n X_k$, then

$$\mathbb{E}[(Z-Z'_k)^2] = 2\operatorname{Var}[X_k],$$

and (9) holds with equality. This shows that sums of independent random variables $X_1,\dots,X_n$ are the least concentrated among all functions of X^n.

2) For every random variable U with a finite second moment and for all $c\in\mathbb{R}$,

$$\operatorname{Var}[U] \le \mathbb{E}[(U-c)^2].$$

Thus, by conditioning on $\overline{X}^k$, we let $Z_k = f_k(\overline{X}^k)$ for arbitrary functions f_k ($k\in\{1,\dots,n\}$) of $n-1$ variables to obtain

$$\operatorname{Var}[Z|\overline{X}^k] \le \mathbb{E}[(Z-Z_k)^2|\overline{X}^k].$$

From (8), this yields another variant of the Efron–Stein–Steele inequality:

$$\operatorname{Var}[Z] \le \sum_{i=1}^{n}\mathbb{E}[(Z-Z_k)^2]. \tag{10}$$

3) Suppose we know that, by varying just one of the arguments of f while holding all others fixed, we cannot change the value of f by more than some bounded amount. More precisely, suppose that there exist finite constants $c_1,\dots,c_n\ge 0$, such that

$$\begin{aligned}&\sup_x f(x_1,\dots,x_{i-1},x,x_{i+1},\dots,x_n)\\ &-\inf_x f(x_1,\dots,x_{i-1},x,x_{i+1},\dots,x_n) \le c_i\end{aligned} \tag{11}$$

for all i and all $x_1,\dots,x_{i-1},x_{i+1},\dots,x_n$. Then, by Lemma 1,

$$\operatorname{Var}[Z|\overline{X}^k] \le \tfrac{1}{4}\,c_k^2$$

and therefore from (5), (8)

$$\operatorname{Var}[Z] \le \tfrac{1}{4}\sum_{k=1}^{n} c_k^2. \tag{12}$$

Example: Kernel Density Estimation

As an example of Efron–Stein–Steele inequalities in action, let us look at *kernel density estimation* (KDE), a nonparametric procedure for estimating an unknown pdf ϕ of a real-valued random variable X based on observing n i.i.d. samples $X_1, \ldots, X_n$ drawn from ϕ [20, Chap. 9]. A *kernel* is a function $K\colon \mathbb{R} \to \mathbb{R}^+$ satisfying the following conditions:

1) It is integrable and normalized: $\int_{-\infty}^{\infty} K(u)\mathrm{d}u = 1$.
2) It is even: $K(u) = K(-u)$ for all $u \in \mathbb{R}$.
3) $\lim_{h\downarrow 0} \frac{1}{h} K(\frac{x-u}{h}) = \delta(x-u)$, where δ is the Dirac function.

The KDE is given by

$$\phi_n(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right),$$

where $h > 0$ is a parameter called the *bandwidth*. From the properties of K, for each $x \in \mathbb{R}$ we have

$$\mathbb{E}[\phi_n(x)] = \frac{1}{h} \int_{-\infty}^{\infty} K\left(\frac{x-u}{h}\right) \phi(u)\mathrm{d}u \xrightarrow{h\downarrow 0} \phi(x).$$

Thus, we expect the KDE ϕ_n to concentrate around the true pdf ϕ; to quantify this, let us examine the L_1 error

$$Z_n = f(X_1, \ldots, X_n) = \int_{-\infty}^{\infty} |\phi_n(x) - \phi(x)|\mathrm{d}x.$$

A simple calculation shows that f satisfies (11) with

$$c_1 = \ldots = c_n = \frac{2}{n},$$

and therefore (12) yields

$$\mathrm{Var}[Z_n] \le \frac{1}{n}.$$

Now, to take full advantage of the martingale method, we need to combine the martingale decomposition (3) with the Chernoff bound. To proceed, we first note that the sequence of random variables $Z_k \triangleq \mathbb{E}[Z|X^k]$, for $k = 0, 1, \ldots, n$, is a martingale with respect to $X_1, \ldots, X_n$, i.e., $\mathbb{E}[Z_{k+1}|X^k] = Z_k$ for each k. Here is one frequently used concentration result:

Theorem 1 (Azuma–Hoeffding inequality [10], [11])**.** *Let $\{Z_k\}_{k=0}^n$ be a real-valued martingale sequence. Suppose that the martingale increments $\xi_k = Z_k - Z_{k-1}$, for $k = 1, \ldots, n$, are almost surely bounded, i.e., $|\xi_k| \le d_k$ a.s. for some constants $d_1, \ldots, d_n \ge 0$. Then*

$$\mathbb{P}\left[|Z_n - Z_0| \ge t\right] \le 2\exp\left(-\frac{t^2}{2\sum_{k=1}^n d_k^2}\right), \quad \forall t > 0. \tag{13}$$

The main idea behind the proof is to apply Hoeffding's lemma to each term ξ_k in the Doob martingale decomposition (3), conditionally on X^{k-1}: for all $\lambda > 0$

$$\begin{aligned}\mathbb{E}[e^{\lambda(Z_n - Z_0)}] &= \mathbb{E}\left[\prod_{k=1}^{n} e^{\lambda\xi_k}\right] \\ &= \mathbb{E}\left[\prod_{k=1}^{n-1} e^{\lambda\xi_k} \mathbb{E}[e^{\lambda\xi_n}|X^{n-1}]\right].\end{aligned}$$

Since $|\xi_n| \le d_n$, we have $\ln \mathbb{E}[e^{\lambda \xi_n}|X^{n-1}] \le \frac{\lambda^2 d_n^2}{2}$, by Hoeffding's lemma. Continuing in this manner and peeling off the terms ξ_k one by one, we can apply the Chernoff bound and obtain (13). However, the Azuma-Hoeffding inequality is not tight in general (e.g., if $t > \sum_{k=1}^n d_k$, then the probability in the left side of (13) is zero, due to the boundedness of the ξ_k's, whereas its bound in the right side of (13) is strictly positive). One way to tighten it is to make use of additional information on the conditional variances along the martingale sequence [21]:

Theorem 2 (McDiarmid)**.** *Let $\{Z_k\}_{k=0}^{\infty}$ be a martingale satisfying the following two conditions for some constants $d, \sigma > 0$:*

- *$|\xi_k| \le d$ for all k.*
- *$\mathrm{Var}[Z_k|X^{k-1}] = \mathbb{E}[|\xi_k|^2|X^{k-1}] \le \sigma^2$ for all k.*

Then, for every $\alpha \ge 0$,

$$\mathbb{P}\left[|Z_n - Z_0| \ge \alpha n\right] \le 2\exp\left(-nd\Big(\frac{\delta+\gamma}{1+\gamma}\Big\|\frac{\gamma}{1+\gamma}\Big)\right),$$

where $\gamma = \sigma^2/d^2$, $\delta = \alpha/d$, and $d(p\|q) \triangleq p\ln\frac{p}{q} + (1-p)\ln\frac{1-p}{1-q}$ is the binary relative entropy function.

Note that, in contrast to Theorem 1, the martingale increments $\{\xi_k\}$ in Theorem 2 should be bounded by a constant d which is independent of k.

A prominent application of the martingale method is a powerful inequality due to McDiarmid [21], also known as the bounded difference inequality:

Theorem 3 (McDiarmid's inequality)**.** *If f satisfies the bounded difference property* (11)*, and $X_1, \ldots, X_n$ are independent random variables, then for all $t > 0$*

$$\mathbb{P}\left[|f(X^n) - \mathbb{E}[f(X^n)]| \ge t\right] \le 2\exp\left(-\frac{2t^2}{\sum_{k=1}^n c_k^2}\right). \tag{14}$$

The strategy of the proof is similar to the one used to derive the Azuma–Hoeffding inequality. In fact, we could have used the Azuma–Hoeffding inequality to bound the tail probability in (14); however, McDiarmid's inequality provides a factor of 4 improvement in the exponent of the bound when f is a function of n independent random variables.

Here is a nice information-theoretic application of McDiarmid's inequality [22]. Consider a discrete memoryless channel (DMC) with input alphabet X, output alphabet Y, and strictly positive transition probabilities $T(y|x)$. Fix an arbitrary distribution P_{X^n} of the input n-block X^n, and let P_{Y^n} denote the resulting output distribution. Then, for every input n-block $x^n \in \mathsf{X}^n$,

$$\mathbb{P}_{Y^n|X^n=x^n}\Big[\log\frac{P_{Y^n|X^n=x^n}(Y^n)}{P_{Y^n}(Y^n)} \ge D(P_{Y^n|X^n=x^n}\|P_{Y^n}) + t\Big] \le \exp\left(-\frac{2t^2}{nc(T)}\right), \tag{15}$$

where

$$c(T) \triangleq 2\max_{x,x'\in\mathsf{X}}\max_{y\in\mathsf{Y}}\log\frac{T(y|x)}{T(y|x')}. \tag{16}$$

Proof: Let us consider the function

$$f(y_1, \ldots, y_n) \triangleq \log\frac{P_{Y^n|X^n=x^n}(y^n)}{P_{Y^n}(y^n)}$$

(recall that the input block x^n is fixed). A simple calculation shows that this f has bounded differences with

$$c_1 = \ldots = c_n = c(T).$$

Moreover, since the channel is memoryless, $Y_1, \ldots, Y_n$ are independent random variables under $P_{Y^n|X^n=x^n}$ (although not under P_{Y^n}, unless P_{X^n} is a product distribution). Applying McDiarmid's inequality, we get (15). ∎

The martingale method has also been used successfully to analyze concentration properties of random codes around their ensemble averages. The performance analysis of a particular code is usually difficult, especially for codes of large block lengths. Availability of a concentration result for the performance of capacity-approaching code ensembles under low-complexity decoding algorithms, as it is the case with low-density parity-check (LDPC) codes [14], validates the use of the density evolution technique as an analytical tool to assess the performance of individual codes from a code ensemble whose block length is sufficiently large, and to assess their asymptotic gap to capacity. However, it should be borne in mind that the current concentration results for codes defined on graphs, which mainly rely on the Azuma–Hoeffding inequality, are weak since in practice concentration is observed at much shorter block lengths.

Here are two illustrative examples of the use of martingale concentration inequalities in the analysis of code performance. The first result, due to Sipser and Spielman [23], is useful for assessing the performance of bit-flipping decoding algorithms for expander codes:

Theorem 4 (Sipser and Spielman)**.** *Let $\mathcal{G}$ be a bipartite graph that is chosen uniformly at random from the ensemble of bipartite graphs with n vertices on the left, a left degree l, and a right degree r. Let $\alpha \in (0,1)$ and $\delta > 0$ be fixed numbers. Then, with probability at least $1 - \exp(-\delta n)$, all sets of αn vertices on the left side of $\mathcal{G}$ are connected to at least*

$$n \left[\frac{l\big(1-(1-\alpha)^r\big)}{r} - \sqrt{2l\alpha\left(h(\alpha)+\delta\right)} \right]$$

vertices (neighbors) on the right side of $\mathcal{G}$, where h is the binary entropy function to base e (i.e., $h(x) = -x\ln(x) - (1-x)\ln(1-x)$, $x \in [0,1]$).

The proof revolves around the analysis of the so-called *neighbor exposure martingale* via the Azuma–Hoeffding inequality to bound the probability that the number of neighbors deviates significantly from its mean value.

Let $\mathrm{LDPC}(n, \lambda, \rho)$ denote an LDPC code ensemble of block length n, respectively, and with left and right degree distributions λ and ρ from the edge perspective (i.e., λ_i designates the fraction of edges which are connected to a variable node of degree i, and ρ_i designates the fraction of edges which are connected to parity-check nodes of degree i).

The second result, due to Richardson and Urbanke [24], concerns the performance of message-passing decoding algorithms for LDPC codes.

Theorem 5 (Richardson–Urbanke)**.** *Let $\mathcal{C}$, a code chosen uniformly at random from the ensemble $\mathrm{LDPC}(n, \lambda, \rho)$, be used for transmission over a memoryless binary-input output-symmetric (MBIOS) channel. Assume that the decoder performs ℓ iterations of message-passing decoding, and let $P_{\mathrm{b}}(\mathcal{C}, \ell)$ denote the resulting bit error probability. Then, for every $\delta > 0$, there exists some $\alpha = \alpha(\lambda, \rho, \delta, \ell) > 0$ (independent of the block length n), such that*

$$\mathbb{P}\left[\left|P_{\mathrm{b}}(\mathcal{C},\ell) - \mathbb{E}_{\mathrm{LDPC}(n,\lambda,\rho)}[P_{\mathrm{b}}(\mathcal{C},\ell)]\right| \geq \delta\right] \leq e^{-\alpha n}$$

The proof also applies the Azuma–Hoeffding inequality to a certain martingale sequence. Some additional references on the use of the martingale method in the context of codes include [14], [23]–[29]. For more details, we refer the reader to our monograph [9].

IV. The entropy method and logarithmic Sobolev inequalities

The entropy method, as its name suggests, relies on information-theoretic techniques to control the logarithmic moment-generating function ψ directly in terms of certain relative entropies. Recall our roadmap for proving a concentration inequality for $Z = f(X)$, where X is an arbitrary random variable:

- Derive a tight quadratic bound on ψ:
$$\psi(\lambda) = \log \mathbb{E}[e^{\lambda(Z-\mathbb{E}[Z])}] \leq \tfrac{1}{2}\,\lambda^2\sigma^2.$$
- Use the Chernoff bound to get
$$\mathbb{P}[Z \geq \mathbb{E}[Z] + t] \leq e^{-\frac{t^2}{2\sigma^2}}, \quad \forall\, t \geq 0.$$

Let $P = \mathcal{L}(X)$, and introduce the tilted distribution $P^{(\lambda f)}$:

$$\mathrm{d}P^{(\lambda f)} = \frac{e^{\lambda f}\mathrm{d}P}{\mathbb{E}_P[e^{\lambda f}]}.$$

The entropy method revolves around the relative entropy $D(P^{(\lambda f)}\|P)$, and has two ingredients: (1) the Herbst argument, and (2) tensorization.

We start with the Herbst argument (the name refers to an unpublished note by I. Herbst that proposed the use of such an argument in the context of mathematical physics of quantum fields). Let us examine the relative entropy:

$$\begin{aligned} D(P^{(\lambda f)}\|P) &= \int \mathrm{d}P^{(\lambda f)} \log \frac{\mathrm{d}P^{(\lambda f)}}{\mathrm{d}P} \\ &= \mathbb{E}^{(\lambda f)}\left[\lambda f(X) - \psi(\lambda)\right] \\ &= \lambda\psi'(\lambda) - \psi(\lambda), \end{aligned}$$

where $\mathbb{E}^{(\lambda f)}[\cdot]$ denotes expectation with respect to the tilted distribution $P^{(\lambda f)}$. Now, with a bit of foresight, we rewrite the last expression as

$$\lambda\psi'(\lambda) - \psi(\lambda) = \lambda^2 \frac{\mathrm{d}}{\mathrm{d}\lambda}\left(\frac{\psi(\lambda)}{\lambda}\right).$$

Thus, we end up with the identity

$$D(P^{(\lambda f)}\|P) = \lambda^2\,\frac{\mathrm{d}}{\mathrm{d}\lambda}\left(\frac{\psi(\lambda)}{\lambda}\right).$$

Integrating and using the fact that $\lim_{\lambda\to 0}\frac{\psi(\lambda)}{\lambda} = 0$ (which can be proved using l'Hopital's rule), we get

$$\psi(\lambda) = \lambda\int_0^\lambda \frac{D(P^{(tf)}\|P)}{t^2}\mathrm{d}t. \tag{17}$$

Appealing to the Chernoff bound, we end up with the following:

Lemma 4 (The Herbst argument)**.** *Suppose that $Z = f(X)$ is such that*

$$D(P^{(\lambda f)}\|P) \leq \tfrac{1}{2}\,\lambda^2\sigma^2, \qquad \forall\,\lambda \geq 0. \tag{18}$$

Then Z is σ^2-subgaussian, and therefore

$$\mathbb{P}\left[f(X) \geq \mathbb{E}[f(X)] + t\right] \leq e^{-\frac{t^2}{2\sigma^2}}, \qquad \forall\, t \geq 0. \tag{19}$$

In fact, it can be shown that the reverse implication holds as well, but with some loss in the constants [30]: if $Z = f(X)$ is $\sigma^2/4$-subgaussian, then

$$D(P^{(\lambda f)} \| P) \leq \tfrac{1}{2}\,\lambda^2 \sigma^2, \qquad \lambda \geq 0.$$

In other words, subgaussianity of $Z = f(X)$ is equivalent to $D(P^{(\lambda f)} \| P) = O(\lambda^2)$. It seems, therefore, that we have not really accomplished anything, apart from arriving at an equivalent characterization of subgaussianity. However, the relative entropy has one crucial property: it tensorizes. Recall that we are interested in the high-dimensional setting, where $X = (X_1, \ldots, X_n)$ is a tuple of n independent random variables. Thus, $P = \mathcal{L}(X)$ is a product distribution: $P_X = P_{X_1} \otimes \ldots \otimes P_{X_n}$. Using this fact together with the chain rule for relative entropy, we arrive at the following:

Lemma 5 (Tensorization of the relative entropy)**.** *Let P and Q be two probability distributions of a random n-tuple $X = (X_1, \ldots, X_n)$, such that the coordinates of X are independent under P. Then*

$$D(Q\|P) \leq \sum_{i=1}^{n} D(Q_{X_i|\overline{X}^i} \| P_{X_i} | Q_{\overline{X}^i}). \tag{20}$$

The quantity on the right-hand side of (20) is the *erasure divergence* between Q and P [31]. We now particularize this general bound to our problem, where Q is given by the tilted distribution $P^{(\lambda f)}$. In that case, using Bayes' rule and the fact that the X_i's are independent, we can express the conditional distributions $P^{(\lambda f)}_{X_i|\overline{X}^i}$ as follows: for each $\bar{x}^i$,

$$\mathrm{d}P^{(\lambda f)}_{X_i|\overline{X}^i = \bar{x}^i} = \frac{e^{\lambda f(x_1,\ldots,x_{i-1},\cdot,x_{i+1},\ldots,x_n)}}{\mathbb{E}\left[e^{\lambda f(x_1,\ldots,x_{i-1},X_i,x_{i+1},\ldots,x_n)}\right]}\,\mathrm{d}P_{X_i}.$$

This looks formidable; nevertheless, it reveals that the conditional distribution $P^{(\lambda f)}_{X_i|\overline{X}^i=\bar{x}^i}$ is the exponential tilting of the marginal distribution P_{X_i} with respect to the random variable $f_i(X_i) = f(x_1, \ldots, x_{i-1}, X_i, x_{i+1}, \ldots, x_n)$, which depends only on X_i because $\bar{x}^i$ is fixed. Thus, we arrive at the following bound:

$$D(P^{(\lambda f)}\|P) \leq \sum_{i=1}^{n} \widetilde{\mathbb{E}}\left[D(P^{(\lambda f_i)}_{X_i} \| P_{X_i})\right],$$

where the expectation on the right-hand side is with respect to the tilted distribution.

We can now distill the entropy method into a series of steps:

1) We wish to derive a subgaussian tail bound

$$\mathbb{P}\left[f(X^n) \geq \mathbb{E}[f(X^n)] + t\right] \leq e^{-\frac{t^2}{2\sigma^2}}, \qquad t \geq 0,$$

   where $X_1, \ldots, X_n$ are independent random variables.

2) Suppose that we can prove that there exist constants $c_1, \ldots, c_n \geq 0$, such that

$$D(P^{(\lambda f_i)}_{X_i} \| P_{X_i}) \leq \tfrac{1}{2}\,\lambda^2 c_i^2, \qquad \forall\, i \in \{1, \ldots, n\}. \tag{21}$$

3) Then, by the tensorization lemma,

$$D(P^{(\lambda f)}\|P) \leq \tfrac{1}{2}\,\lambda^2 \sum_{i=1}^{n} c_i^2,$$

   and therefore, by the Herbst argument, $Z = f(X^n)$ is σ^2-subgaussian with $\sigma^2 = \sum_{i=1}^{n} c_i^2$.

The main benefit of passing to the relative-entropy characterization of subgaussianity is that now, via tensorization, we have broken up a difficult n-dimensional problem into n presumably easier 1-dimensional problems, each of which boils down to analyzing the behavior of the function $f_i(X_i) \equiv f(x_1, \ldots, x_{i-1}, X_i, x_{i+1}, \ldots, x_n)$, where only the ith input coordinate is random, and the remaining ones are fixed at some arbitrary values.

Of course, the problem now reduces to showing that (21) holds. One route, which often yields tight constants, is via so-called *logarithmic Sobolev inequalities*. In a nutshell, a logarithmic Sobolev inequality (or LSI, for short) ties together a probability distribution P, some function class $\mathcal{A}$ that contains the function f of interest, and an "energy" functional $E : \mathcal{A} \to \mathbb{R}$ with the property

$$E(\alpha f) = \alpha E(f), \qquad \forall \alpha \geq 0, f \in \mathcal{A}.$$

With these ingredients in place, a log-Sobolev inequality takes the form

$$D(P^{(f)} \| P) \leq \tfrac{1}{2}\, c\, E^2(f), \qquad \forall f \in \mathcal{A}.$$

Now suppose that $E(f) \leq L$. Then we readily get the bound

$$D(P^{(\lambda f)} \| P) \leq \tfrac{1}{2}\, c\, E^2(\lambda f) = \tfrac{1}{2}\, \lambda^2 c\, E^2(f) \leq \tfrac{1}{2}\, \lambda^2 c L^2,$$

so $f(X)$, $X \sim P$, is σ^2-subgaussian with $\sigma^2 = cL^2$.

There is a vast literature on log-Sobolev inequalities, and an interested reader may consult our monograph for more details and additional references. Here we will give the two classic examples: the Bernoulli LSI and the Gaussian LSI, due to Gross [32].

Theorem 6 (Bernoulli LSI)**.** *Let $X_1, \ldots, X_n$ be i.i.d.* $\mathrm{Bern}(1/2)$ *random variables. Then, for every function $f \colon \{0,1\}^n \to \mathbb{R}$, we have*

$$D(P^{(f)} \| P) \leq \frac{1}{8} \frac{\mathbb{E}\left[|Df(X^n)|^2 e^{f(X^n)}\right]}{\mathbb{E}[e^{f(X^n)}]}, \tag{22}$$

where $P = \mathrm{Bern}(1/2)^{\otimes n}$,

$$Df(x^n) \triangleq \sqrt{\sum_{i=1}^n |f(x^n) - f(x^n \oplus e_i)|^2},$$

and $x^n \oplus e_i$ is the XOR of x^n with the bit string of all zeros, except for the ith bit. In other words, $x^n \oplus e_i$ is x^n with the ith bit flipped.

The proof, which we omit, is to first establish the $n = 1$ case via a straightforward if tedious exercise in calculus, and then to extend to an arbitrary n by tensorization. Note that the mapping $f \mapsto Df$ has the desired scaling property: $D(\alpha f) = \alpha D(f)$ for all $\alpha \geq 0$.

Theorem 7 (Gaussian LSI)**.** *Let $X_1, \ldots, X_n$ be i.i.d. $N(0,1)$ random variables. Then, for an arbitrary smooth function $f \colon \mathbb{R}^n \to \mathbb{R}$,*

$$D(P^{(f)} \| P) \leq \frac{1}{2} \frac{\mathbb{E}\left[\|\nabla f(X^n)\|_2^2 e^{f(X^n)}\right]}{\mathbb{E}[e^{f(X^n)}]}. \tag{23}$$

Note that the mapping $f \mapsto \|\nabla f\|_2$ has the scaling property: $\|\nabla(\alpha f)\|_2 = \alpha \|\nabla f\|_2$ for all $\alpha \geq 0$. By now, there are at least fifteen different ways in the literature for proving the Gaussian LSI. The original proof by Gross was to apply the Bernoulli LSI to the function

$$f\left(\frac{X_1 + \ldots + X_n - n/2}{\sqrt{n/4}}\right), \qquad X_i \overset{\text{i.i.d.}}{\sim} \mathrm{Bern}(1/2),$$

and then pass to the Gaussian limit by appealing to the Central Limit Theorem.

The Gaussian LSI can be used to give a short proof of the following concentration inequality for Lipschitz functions of Gaussians, which was originally obtained by Tsirelson, Ibragimov, and Sudakov [33] using different methods:

Theorem 8 (Tsirelson–Ibragimov–Sudakov)**.** *Let $X_1, \ldots, X_n$ be i.i.d. $N(0,1)$ random variables, and let $f\colon \mathbb{R}^n \to \mathbb{R}$ be a function which is L-Lipschitz:*

$$|f(x^n) - f(y^n)| \leq L \left\| x^n - y^n \right\|_2 .$$

Then, $f(X^n)$ is L^2-subgaussian, which yields

$$\mathbb{P}\left[f(X^n) \geq \mathbb{E}[f(X^n)] + t\right] \leq e^{-\frac{t^2}{2L^2}} \tag{24}$$

for all $t > 0$.

Proof: By a standard approximation argument, we may assume that f is differentiable. Since it is also L-Lipschitz, $\|\nabla f\|_2^2 \leq L^2$ everywhere. Substituting this bound into the Gaussian LSI for λf, we obtain

$$D(P^{(\lambda f)} \| f) \leq \tfrac{1}{2}\lambda^2 L^2.$$

By the Herbst argument, $Z = f(X^n)$, $X^n \sim N(0, I_n)$, is L^2-subgaussian, and we are done. ∎

This result is remarkable in two ways: It only assumes Lipschitz continuity of f, and gives *dimension-free* concentration (i.e., the exponent in (24) does not depend on n).

Deriving log-Sobolev inequalities, especially with tight constants, is a subtle art. A commonly used method is to realize P as an invariant distribution of some continuous-time reversible Markov process and to extract a suitable energy functional E from the structure of the infinitesimal generator of the process. In many cases, however, it is possible to derive a log-Sobolev inequality via tensorization and a nice and simple variance-based representation of the relative entropy due to A. Maurer [34]:

Theorem 9 (Maurer)**.** *Let X be a random variable with law P. Then, for every real-valued function f and all $\lambda \geq 0$*

$$D(P^{(\lambda f)} \| P) = \int_0^\lambda \int_t^\lambda \mathrm{Var}^{(sf)}[f(X)] \mathrm{d}s \, \mathrm{d}t,$$

where $\mathrm{Var}^{(sf)}[f(X)]$ is the variance of $f(X)$ under the tilted distribution $P^{(sf)}$.

Proof: As before, let $\psi(\lambda) = \log \mathbb{E}[e^{\lambda(f(X)) - \mathbb{E}[f(X)]}]$ be the logarithmic moment-generating function of $f(X)$. Then

$$\begin{aligned} D(P^{(\lambda f)} \| P) &= \lambda \psi'(\lambda) - \psi(\lambda) \\ &= \int_0^\lambda \left[\psi'(\lambda) - \psi'(t)\right] \mathrm{d}t \\ &= \int_0^\lambda \int_t^\lambda \psi''(s) \mathrm{d}s \, \mathrm{d}t, \end{aligned}$$

where we have used the fact that $\psi(0) = \psi'(0) = 0$ and the fundamental theorem of calculus. Recalling that $\psi''(s) = \mathrm{Var}^{(sf)}[f(X)]$, we are done. ∎

The following result is a direct consequence of Theorem 9:

Theorem 10. *Let $\mathcal{A}$ be a class of functions of X, and suppose that there is a mapping $\Gamma : \mathcal{A} \to \mathbb{R}$, such that:*

1) *For all* $f \in \mathcal{A}$ *and* $\alpha \geq 0$, $\Gamma(\alpha f) = \alpha \Gamma(f)$.
2) *There exists a constant* $c > 0$, *such that*

$$\mathrm{Var}^{(\lambda f)}[f(X)] \leq c|\Gamma(f)|^2, \qquad \forall f \in \mathcal{A},\, \lambda \geq 0.$$

Then

$$D(P^{(\lambda f)} \| P) \leq \tfrac{1}{2}\, \lambda^2 c\, |\Gamma(f)|^2, \qquad \forall f \in \mathcal{A}, \lambda \geq 0.$$

To illustrate Maurer's method, let's use it to derive the Bernoulli LSI. It suffices to prove the $n = 1$ case, and then to scale up to an arbitrary n by tensorization. Thus, let $P = \mathrm{Bern}(1/2)$, and for every function $f\colon \{0,1\} \to \mathbb{R}$ define $\Gamma(f) \triangleq |f(0) - f(1)|$. By Lemma 1,

$$\mathrm{Var}^{(\lambda f)}[f(X)] \leq \tfrac{1}{4}\, |f(0) - f(1)|^2 = \tfrac{1}{4}\, |\Gamma(f)|^2.$$

Thus, the conditions of Theorem 10 are satisfied with $c = 1/4$, and we get precisely the Bernoulli LSI. One can also use Maurer's method to prove McDiarmid's inequality (see Theorem 3).

V. Transportation-cost inequalities

At this point, we notice a common theme running through the above examples of concentration:

- Let $f\colon \mathbb{R}^n \to \mathbb{R}$ be 1-Lipschitz with respect to the Euclidean norm $\|\cdot\|_2$, and let $X_1, \ldots, X_n$ be i.i.d. $N(0,1)$ random variables. Then, for every $t \geq 0$,

$$\mathbb{P}[f(X^n) \geq \mathbb{E}[f(X^n)] + t] \leq e^{-t^2/2}.$$

- Let X be an arbitrary space, and consider a function $f\colon \mathsf{X}^n \to \mathbb{R}$, which is 1-Lipschitz with respect to the weighted Hamming metric

$$d_{\mathbf{c}}(x^n, y^n) \triangleq \sum_{i=1}^n c_i \mathbf{1}_{\{x_i \neq y_i\}},$$

  where $c_1, \ldots, c_n \geq 0$ are some fixed constants. It is easy to see that such a Lipschitz property is equivalent to the bounded difference property (11), and in that case McDiarmid's inequality tells us that

$$\mathbb{P}[f(X^n) \geq \mathbb{E}[f(X^n)] + t] \leq e^{-2t^2/\sum_{i=1}^n c_i^2}$$

  for every tuple $X_1, \ldots, X_n$ of independent X-valued random variables.

Thus, metric spaces and Lipschitz functions seem to be a natural setting to study concentration. To make this statement more precise, let (X, d) be a metric space. We say that a function $f\colon \mathsf{X} \to \mathbb{R}$ is L-Lipschitz (with respect to d) if

$$|f(x) - f(y)| \leq L d(x, y), \qquad \forall x, y \in \mathsf{X}.$$

Denoting by $\mathrm{Lip}_L(\mathsf{X}, d)$ the class of all L-Lipschitz functions, we can pose the following question: What conditions does a probability distribution P on X have to satisfy, so that $f(X)$ with $X \sim P$ is σ^2-subgaussian for every $f \in \mathrm{Lip}_1(\mathsf{X}, d)$?

Through the pioneering work of Katalin Marton [17], [35]–[39], the answer to the above question has deep links to information theory via the notion of so-called *transportation-cost inequalities* [40]. In order to introduce them, we first need some definitions. A *coupling* of two probability distributions P and Q on X is a probability distribution π on the Cartesian product

$\mathsf{X} \times \mathsf{X}$, such that for $(X, Y) \sim \pi$ we have $X \sim P$ and $Y \sim Q$. Let $\Pi(P, Q)$ denote the set of all couplings of P and Q. For $p \geq 1$, the L^p *Wasserstein distance* between P and Q is defined as

$$W_p(P, Q) \triangleq \inf_{\pi \in \Pi(P,Q)} (\mathbb{E}_\pi[d^p(X, Y)])^{1/p}.$$

The name "transportation cost" comes from the following interpretation: Let P (resp., Q) represent the initial (resp., desired) distribution of some matter (say, sand) in space, such that the total mass in both cases is normalized to one. Thus, both P and Q correspond to sand piles of some given shapes. The objective is to rearrange the initial sand pile with shape P into one with shape Q with minimum cost, where the cost of transporting a grain of sand from location x to location y is given by $d^p(x, y)$. If we allow randomized transportation policies, i.e., those that associate with each location x in the initial sand pile a conditional probability distribution $\pi(\mathrm{d}y|x)$ for its destination in the final sand pile, then the minimum transportation cost is given by $W_p(P, Q)$. We say that P satisfies an L^p *transportation-cost inequality* with constant c, or $T_p(c)$ for short, if

$$W_p(P, Q) \leq \sqrt{2cD(Q\|P)}, \qquad \forall Q.$$

The well-known Pinsker's inequality is, in fact, a transportation-cost inequality: If we take X to be an arbitrary space and equip it with the metric $d(x, y) = \mathbf{1}_{\{x \neq y\}}$, then the L^1 Wasserstein distance $W_1(P, Q)$ is simply the total variation distance

$$\|P - Q\|_{\mathrm{TV}} = \sup_A |P(A) - Q(A)|,$$

and Pinsker's inequality

$$\|P - Q\|_{\mathrm{TV}} \leq \sqrt{\tfrac{1}{2} D(Q\|P)}$$

(in nats) is then a $T_1(\frac{1}{4})$ inequality, which is satisfied by all probability measures P, Q where $Q \ll P$ (i.e., Q is absolutely continuous with respect to P). Various distribution-dependent refinements of Pinsker's inequality where the constant is optimized for a fixed P while varying only Q [41], [42] can be interpreted in the same vein as well. Another well-known transportation-cost (TC) inequality is due to Talagrand [43]: Let X be the Euclidean space $\mathbb{R}^n$, equipped with the Euclidean metric $d(x, y) = \|x - y\|_2$. Then $P = N(0, I_n)$ satisfies the $T_2(1)$ inequality: $W_2(P, Q) \leq \sqrt{2D(Q\|P)}$. The remarkable thing here is that the constant is independent of the dimension n.

With these preliminaries out of the way, we can now state the theorem, due to Bobkov and Götze [44], which provides an answer to the question posed above:

Theorem 11 (Bobkov–Götze)**.** *Let X be a random variable taking values in a metric space (X, d) according to a probability distribution P. Then, the following are equivalent:*

1) *$f(X)$ is σ^2-subgaussian for every $f \in \mathrm{Lip}_1(\mathsf{X}, d)$.*
2) *P satisfies $T_1(\sigma^2)$, i.e.,*

$$W_1(P, Q) \leq \sqrt{2\sigma^2 D(Q\|P)}$$

for all Q.

At this point, one may wonder what we have gained – verifying that a given P satisfies a TC inequality, let alone determining tight constants, is a formidable challenge. However, once again, tensorization comes to the rescue. Marton's insight was that TC inequalities tensorize [40]:

Theorem 12. *Let (X_i, P_i, d_i), $1 \le i \le n$, be probability metric spaces. If for some $1 \le p \le 2$ each P_i satisfies $T_p(c)$ on (X_i, d_i), then the product measure $P = P_1 \otimes \ldots \otimes P_n$ on $\mathsf{X} = \mathsf{X}_1 \times \ldots \times \mathsf{X}_n$ satisfies $T_p(cn^{2/p-1})$ w.r.t. the metric*

$$d_p(x^n, y^n) \triangleq \left(\sum_{i=1}^n d_i^p(x_i, y_i) \right)^{1/p}.$$

In particular, if each P_i satisfies $T_1(c)$, then $P = P_1 \otimes \ldots \otimes P_n$ satisfies $T_1(cn)$ with respect to the metric $\sum_i d_i$. Note that the constant deteriorates with n. On the other hand, if each P_i satisfies $T_2(c)$, then P satisfies $T_2(c)$ with respect to $\sqrt{\sum_i d_i^2}$. Note that the latter constant is independent of n.

To give a simple illustration of all these concepts, let us outline yet another proof of McDiarmid's inequality. Consider a product probability space $(\mathsf{X}_1 \times \ldots \times \mathsf{X}_n, P_1 \otimes \ldots \otimes P_n)$. For a fixed choice of constants $c_1, \ldots, c_n \ge 0$, equip X_i with the metric $d_i(x_i, y_i) = c_i \mathbf{1}_{\{x_i \neq y_i\}}$. Then, by rescaling Pinsker's inequality, we see that P_i satisfies a $T_1(c_i^2/4)$ inequality with respect to the metric d_i:

$$W_{1,d_i}(P_i, Q_i) \le \sqrt{\tfrac{1}{2} c_i^2 D(Q_i \| P_i)}, \qquad \forall Q_i. \tag{25}$$

By the tensorization theorem for TC inequalities, the product distribution P satisfies a $T_1(c)$ inequality with $c = (1/4)\sum_{i=1}^n c_i^2$ with respect to the weighted Hamming metric $d_{\mathbf{c}}$. By the Bobkov–Götze theorem, this is equivalent to the subgaussianity of all $f(X_1, \ldots, X_n)$ with $f \in \mathrm{Lip}_1(\mathsf{X}, d)$ and mutually independent $X_i \in \mathsf{X}_i$, $1 \le i \le n$. But this is precisely McDiarmid's inequality.

VI. Some applications in information theory

We end this survey by briefly describing some information-theoretic applications of concentration inequalities.

A. The Blowing-up Lemma and Information-Theoretic Consequences

The first explicit appeal to the concentration phenomenon in information theory dates back to the 1970s work by Ahlswede and collaborators, who used the so-called *blowing-up lemma* for deriving strong converses for a variety of communications and coding problems.

Consider a product space Y^n equipped with the Hamming metric $d(y^n, z^n) = \sum_{i=1}^n \mathbf{1}_{\{y_i \neq z_i\}}$. For $r \in \{0, 1, \ldots, n\}$, define the *r-blowup* of a set $A \subseteq \mathsf{Y}^n$ as

$$[A]_r \triangleq \left\{ z^n \in \mathsf{Y}^n : \min_{y^n \in A} d(z^n, y^n) \le r \right\}$$

The following result, in a different (asymptotic) form was first proved by Ahlswede, Gács, and Körner [45]; a simple proof, which we sketch below, was given by Marton [35]:

Lemma 6 (Blowing-up)**.** *Let $Y_1, \ldots, Y_n$ be independent random variables taking values in Y. Then for every set $A \subseteq \mathsf{Y}^n$ with $P_{Y^n}(A) > 0$*

$$P_{Y^n}\{[A]_r\} \ge 1 - \exp\left[-\frac{2}{n} \left(r - \sqrt{\frac{n}{2} \log \frac{1}{P_{Y^n}(A)}} \right)_+^2 \right],$$

where $(u)_+ \triangleq \max\{0, u\}$.

Proof: We sketch the proof in order to highlight the role of TC inequalities. For each $i \in \{1, \ldots, n\}$, let $P_i = \mathcal{L}(Y_i)$. By tensorization, the product distribution $P = P_{Y^n}$ satisfies the TC inequality

$$W_1(P, Q) \le \sqrt{\frac{n}{2} D(Q\|P)}, \qquad \forall Q, \tag{26}$$

where

$$W_1(P, Q) = \inf_{\pi \in \Pi(P,Q)} \mathbb{E}_\pi \left[\sum_{i=1}^n \mathbf{1}_{\{X_i \neq Y_i\}} \right].$$

Now, for an arbitrary $B \subseteq \mathsf{Y}^n$ with $P(B) > 0$, consider the conditional distribution $P_B(\cdot) \triangleq \frac{P(\cdot \cap B)}{P(B)}$. Then $D(P_B\|P) = \log \frac{1}{P(B)}$, and in that case using (26) with $Q = P_B$, we get

$$W_1(P, P_B) \le \sqrt{\frac{n}{2} \log \frac{1}{P(B)}}. \tag{27}$$

Applying (26) to $B = A$ and $B = [A]_r^c$, we get

$$W_1(P, P_A) \le \sqrt{\frac{n}{2} \log \frac{1}{P(A)}},$$

$$W_1(P, P_{[A]_r^c}) \le \sqrt{\frac{n}{2} \log \frac{1}{1 - P([A]_r)}}.$$

Adding up these two inequalities, we obtain

$$\begin{aligned} &\sqrt{\frac{n}{2} \log \frac{1}{P(A)}} + \sqrt{\frac{n}{2} \log \frac{1}{1 - P([A]_r)}} \\ &\quad \ge W_1(P_A, P) + W_1(P_{[A]_r^c}, P) \\ &\quad \ge W_1(P_A, P_{[A]_r^c}) \\ &\quad \ge \min_{x^n \in A, y^n \in [A]_r^c} d(x^n, y^n) \\ &\quad \ge r, \end{aligned}$$

where the first step holds due to (27), the second step is verified by the triangle inequality, and the remaining steps follow from definitions. Rearranging, we obtain the lemma. ∎

Informally, the lemma states that every set in a product space can be "blown up" to engulf most of the probability mass. Using this fact, one can prove strong converses for channel coding in single-terminal and multiterminal settings. Here is the simplest consequence of the blowing-up lemma in the context of channel codes: Consider a DMC with input alphabet X, output alphabet Y, and transition probabilities $T(y|x)$, $x \in \mathsf{X}, y \in \mathsf{Y}$. An (n, M, ε)-code for T consists of an encoder $f \colon \{1, \ldots, M\} \to \mathsf{X}^n$ and a decoder $g : \mathsf{Y}^n \to \{1, \ldots, M\}$, such that

$$\max_{1 \le j \le M} \mathbb{P}[g(Y^n) \neq j \,|\, f(X^n) = j] \le \varepsilon.$$

Lemma 7. *Let $u_j = f(j)$, $1 \le j \le M$, denote the M codewords of the code, and let $D_j \triangleq g^{-1}(j)$ be the corresponding decoding regions in Y^n. There exists some $\delta_n > 0$, such that*

$$T^n\Big([D_j]_{n\delta_n} \,\Big|\, X^n = u_j\Big) \ge 1 - \frac{1}{n}, \qquad \forall\, j \in \{1, \ldots, M\}.$$

Informally, this corollary of the blowing-up lemma says that "any bad code contains a good subcode." Using this result, Ahlswede and Dueck [46] established a strong converse for channel coding as follows: Consider an (n, M, ε)-code $\mathcal{C} = \{(u_j, D_j)\}_{j=1}^M$. Each decoding set D_j can be "blown up" to a set $\tilde{D}_j \subseteq \mathsf{Y}^n$ with

$$T^n(\tilde{D}_j|u_j) \geq 1 - \frac{1}{n}.$$

The object $\tilde{\mathcal{C}} = \{(u_j, \tilde{D}_j)\}_{j=1}^M$ is not a code (since the sets $\tilde{D}_j$ are no longer disjoint), but a random coding argument can be used to extract an (n, M', ε') "subcode" with M' slightly smaller than M and $\varepsilon' < \varepsilon$. Then one can apply the usual (weak) converse to the subcode. Similar ideas have found use in multiterminal settings, starting with the work of Ahlswede–Gács–Körner [45].

B. Empirical distribution of good channel codes with non-vanishing error probability

Another recent application of concentration inequalities to information theory has to do with characterizing stochastic behavior of output sequences of good channel codes. On a conceptual level, the random coding argument originally used by Shannon (and many times since) to show the existence of good channel codes suggests that the input/output sequence of such a code should resemble, as much as possible, a typical realization of a sequence of i.i.d. random variables sampled from a capacity-achieving input/output distribution. For capacity-achieving sequences of codes with asymptotically vanishing probability of error, this intuition has been analyzed rigorously by Shamai and Verdú [47], who have proved the following remarkable statement [47, Theorem 2]: given a DMC T, any capacity-achieving sequence of channel codes with asymptotically vanishing probability of error (maximal or average) has the property that

$$\lim_{n\to\infty} \frac{1}{n} D(P_{Y^n} \| P^*_{Y^n}) = 0, \tag{28}$$

where, for each n, P_{Y^n} denotes the output distribution on Y^n induced by the code (assuming that the messages are equiprobable), while $P^*_{Y^n}$ is the product of n copies of the single-letter capacity-achieving output distribution. In a recent paper [48], Polyanskiy and Verdú extended the results of [47] for codes with *nonvanishing* probability of error.

To keep things simple, we will only focus on channels with finite input and output alphabets. Thus, let X and Y be finite sets, and consider a DMC T with capacity C. Let $P^*_X \in \mathcal{P}(\mathsf{X})$ be a capacity-achieving input distribution (which may be nonunique). It can be shown [49] that the corresponding output distribution $P^*_Y \in \mathcal{P}(\mathsf{Y})$ is unique. Consider any (n, M)-code $\mathcal{C} = (f, g)$, let $P^{(\mathcal{C})}_{X^n}$ denote the distribution of $X^n = f(J)$, where J is uniformly distributed in $\{1, \ldots, M\}$, and let $P^{(\mathcal{C})}_{Y^n}$ denote the corresponding output distribution. The central result of [48] is that the output distribution $P^{(\mathcal{C})}_{Y^n}$ of any (n, M, ε)-code satisfies

$$D\big(P^{(\mathcal{C})}_{Y^n} \big\| P^*_{Y^n}\big) \leq nC - \log M + o(n); \tag{29}$$

moreover, the $o(n)$ term was refined in [48, Theorem 5] to $O(\sqrt{n})$ for any DMC, except those that have zeroes in their transition matrix. Using McDiarmid's inequality, this result is sharpened as follows [22]:

Theorem 13. *Consider a DMC T with positive transition probabilities. Then any (n, M, ε)-code $\mathcal{C}$ for T, with $\varepsilon \in (0, 1/2)$, satisfies*

$$D\Big(P^{(\mathcal{C})}_{Y^n} \Big\| P^*_{Y^n}\Big) \leq nC - \log M + \log\frac{1}{\varepsilon} + c(T)\sqrt{\frac{n}{2}\log\frac{1}{1-2\varepsilon}},$$

where $c(T)$ is defined in (16)*.*

Proof (Sketch): Using the inequality (15) with $P_{Y^n} = P_{Y^n}^{(\mathcal{C})}$ and $t = c(T)\sqrt{\frac{n}{2}\log\frac{1}{1-2\varepsilon}}$, we get

$$P_{Y^n|X^n=x^n}\left[\log\frac{P_{Y^n|X^n=x^n}(Y^n)}{P_{Y^n}^{(\mathcal{C})}(Y^n)} \geq D\Big(P_{Y^n|X^n=x^n}\Big\|P_{Y^n}^{(\mathcal{C})}\Big) + c(T)\sqrt{\frac{n}{2}\log\frac{1}{1-2\varepsilon}}\right] \leq 1-2\varepsilon$$

Now, just like Polyanskiy and Verdú, we can appeal to a strong converse result due to Augustin [50] to get

$$\log M \leq \log\frac{1}{\varepsilon} + D\Big(P_{Y^n|X^n}\Big\|P_{Y^n}^{(\mathcal{C})}\Big|P_{X^n}^{(\mathcal{C})}\Big) + c(T)\sqrt{\frac{n}{2}\log\frac{1}{1-2\varepsilon}}. \tag{30}$$

Therefore,

$$\begin{aligned} D\Big(P_{Y^n}^{(\mathcal{C})}\Big\|P_{Y^n}^*\Big) &= D\Big(P_{Y^n|X^n}\Big\|P_{Y^n}^*\Big|P_{X^n}^{(\mathcal{C})}\Big) - D\Big(P_{Y^n|X^n}\Big\|P_{Y^n}^{(\mathcal{C})}\Big|P_{X^n}^{(\mathcal{C})}\Big) \\ &\leq nC - \log M + \log\frac{1}{\varepsilon} + c(T)\sqrt{\frac{n}{2}\log\frac{1}{1-2\varepsilon}}, \end{aligned}$$

where the first step is by the chain rule, the second follows from the properties of the capacity-achieving output distribution, and the last step uses (30). ∎

A useful consequence of this result is that a broad class of functions evaluated on the output of a good code concentrate sharply around their expectations with respect to the capacity-achieving output distribution:

Theorem 14. *Consider a DMC T with $c(T) < \infty$. Let d be a metric on Y^n, and suppose that $P_{Y^n|X^n=x^n}$, $x^n \in \mathsf{X}^n$, as well as $P_{Y^n}^*$, satisfy $T_1(c)$ for some $c > 0$. Then, for every $\varepsilon \in (0, 1/2)$, every (n, M, ε)-code $\mathcal{C}$ for T, and every function $f\colon \mathsf{Y}^n \to \mathbb{R}$ which is L-Lipschitz on (Y^n, d), we have*

$$P_{Y^n}^{(\mathcal{C})}\Big(|f(Y^n) - \mathbb{E}[f(Y^{*n})]| \geq t\Big) \leq \frac{4}{\varepsilon}\exp\left(nC - \ln M + a\sqrt{n} - \frac{t^2}{8cL^2}\right),\ \forall\, r \geq 0 \tag{31}$$

where $Y^{*n} \sim P_{Y^n}^*$, and $a \triangleq c(T)\sqrt{\frac{1}{2}\ln\frac{1}{1-2\varepsilon}}$.

As pointed out in [48], concentration inequalities like (31) can be very useful for gaining insight into the performance characteristics of good channel codes without having to explicitly construct such codes: all one needs to do is to find the capacity-achieving output distribution P_Y^* and evaluate $\mathbb{E}[f(Y^{*n})]$ for an arbitrary f of interest. Consequently, the above theorem guarantees that $f(Y^n)$ concentrates tightly around $\mathbb{E}[f(Y^{*n})]$, which is relatively easy to compute since $P_{Y^n}^*$ is a product measure.

REFERENCES

[1] M. Talagrand, "A new look at independence," *Annals of Probability*, vol. 24, no. 1, pp. 1–34, January 1996.

[2] S. Boucheron, G. Lugosi, and P. Massart, *Concentration Inequalities - A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

[3] M. Ledoux, *The Concentration of Measure Phenomenon*, ser. Mathematical Surveys and Monographs. American Mathematical Society, 2001, vol. 89.

[4] G. Lugosi, "Concentration of measure inequalities - lecture notes," 2009, available at http://www.econ.upf.edu/~lugosi/anu.pdf.

[5] P. Massart, *The Concentration of Measure Phenomenon*, ser. Lecture Notes in Mathematics. Springer, 2007, vol. 1896.

[6] C. McDiarmid, "Concentration," in *Probabilistic Methods for Algorithmic Discrete Mathematics*. Springer, 1998, pp. 195–248.

[7] M. Talagrand, "Concentration of measure and isoperimteric inequalities in product space," *Publications Mathématiques de l'I.H.E.S*, vol. 81, pp. 73–205, 1995.

[8] N. Alon and J. H. Spencer, *The Probabilistic Method*, 3rd ed. Wiley Series in Discrete Mathematics and Optimization, 2008.

[9] M. Raginsky and I. Sason, *Concentration of Measure Inequalities in Information Theory, Communications, and Coding*, 2nd ed. Foundations and Trends in Communications and Information Theory, Now Publishers, 2014. [Online]. Available: http://arxiv.org/abs/1212.4663.

[10] K. Azuma, "Weighted sums of certain dependent random variables," *Tohoku Mathematical Journal*, vol. 19, pp. 357–367, 1967.

[11] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *Journal of the American Statistical Association*, vol. 58, no. 301, pp. 13–30, March 1963.

[12] F. Chung and L. Lu, *Complex Graphs and Networks*, ser. Regional Conference Series in Mathematics. Wiley, 2006, vol. 107.

[13] ——, "Concentration inequalities and martingale inequalities: a survey," *Internet Mathematics*, vol. 3, no. 1, pp. 79–127, March 2006, available at http://www.math.ucsd.edu/~fan/wp/concen.pdf.

[14] T. J. Richardson and R. Urbanke, *Modern Coding Theory*. Cambridge University Press, 2008.

[15] Y. Seldin, F. Laviolette, N. Cesa-Bianchi, J. Shawe-Taylor, and P. Auer, "PAC-Bayesian inequalities for martingales," *IEEE Trans. on Information Theory*, vol. 58, no. 12, pp. 7086–7093, December 2012.

[16] N. Gozlan and C. Leonard, "Transport inequalities: a survey," *Markov Processes and Related Fields*, vol. 16, no. 4, pp. 635–736, 2010.

[17] K. Marton, "Distance-divergence inequalities," *IEEE Information Theory Society Newsletter*, vol. 64, no. 1, pp. 9–13, March 2014.

[18] B. Efron and C. Stein, "The jackknife estimate of variance," *Annals of Statistics*, vol. 9, pp. 586–596, 1981.

[19] J. M. Steele, "An Efron–Stein inequality for nonsymmetric statistics," *Annals of Statistics*, vol. 14, pp. 753–758, 1986.

[20] L. Devroye and G. Lugosi, *Combinatorial Methods in Density Estimation*. Springer, 2001.

[21] C. McDiarmid, "On the method of bounded differences," in *Surveys in Combinatorics*. Cambridge University Press, 1989, vol. 141, pp. 148–188.

[22] M. Raginsky and I. Sason, "Refined bounds on the empirical distribution of good channel codes via concentration inequalities," in *Proceedings of the 2013 IEEE International Workshop on Information Theory*, Istanbul, Turkey, July 2013, pp. 221–225.

[23] M. Sipser and D. A. Spielman, "Expander codes," *IEEE Trans. on Information Theory*, vol. 42, no. 6, pp. 1710–1722, November 1996.

[24] T. J. Richardson and R. Urbanke, "The capacity of low-density parity-check codes under message-passing decoding," *IEEE Trans. on Information Theory*, vol. 47, no. 2, pp. 599–618, February 2001.

[25] M. G. Luby, Mitzenmacher, M. A. Shokrollahi, and D. A. Spielmann, "Efficient erasure-correcting codes," *IEEE Trans. on Information Theory*, vol. 47, no. 2, pp. 569–584, February 2001.

[26] A. Kavčić, X. Ma, and M. Mitzenmacher, "Binary intersymbol interference channels: Gallager bounds, density evolution, and code performance bounds," *IEEE Trans. on Information Theory*, vol. 49, no. 7, pp. 1636–1652, July 2003.

[27] A. Montanari, "Tight bounds for LDPC and LDGM codes under MAP decoding," *IEEE Trans. on Information Theory*, vol. 51, no. 9, pp. 3247–3261, September 2005.

[28] C. Méasson, A. Montanari, and R. Urbanke, "Maxwell construction: the hidden bridge between iterative and maximum apposteriori decoding," *IEEE Trans. on Information Theory*, vol. 54, no. 12, pp. 5277–5307, December 2008.

[29] I. Sason and R. Eshel, "On concentration of measures for LDPC code ensembles," in *Proceedings of the 2011 IEEE International Symposium on Information Theory*, Saint Petersburg, Russia, August 2011, pp. 1273–1277.

[30] R. van Handel, "Probability in high dimension," ORF 570 lecture notes, Princeton University, June 2014.

[31] S. Verdú and T. Weissman, "The information lost in erasures," *IEEE Trans. on Information Theory*, vol. 54, no. 11, pp. 5030–5058, November 2008.

[32] L. Gross, "Logarithmic Sobolev inequalities," *American Journal of Mathematics*, vol. 97, no. 4, pp. 1061–1083, 1975.

[33] B. S. Tsirelson, I. A. Ibragimov, and V. N. Sudakov, "Norms of Gaussian sample functions," in *Proceedings of the Third Japan-USSR Symposium on Probability Theory*, ser. Lecture Notes in Mathematics. Springer, 1976, vol. 550, pp. 20–41.

[34] A. Maurer, "Thermodynamics and concentration," *Bernoulli*, vol. 18, no. 2, pp. 434–454, 2012.

[35] K. Marton, "A simple proof of the blowing-up lemma," *IEEE Trans. on Information Theory*, vol. 32, no. 3, pp. 445–446, May 1986.

[36] ——, "A measure concentration inequality for contracting Markov chains," *Geometric and Functional Analysis*, vol. 6, pp. 556–571, 1996, see also erratum in *Geometric and Functional Analysis*, vol. 7, pp. 609–613, 1997.

[37] ——, "Bounding $\bar{d}$-distance by informational divergence: a method to prove measure concentration," *Annals of Probability*, vol. 24, no. 2, pp. 857–866, 1996.

[38] ——, "Measure concentration for Euclidean distance in the case of dependent random variables," *Annals of Probability*, vol. 32, no. 3B, pp. 2526–2544, 2004.

[39] ——, "Correction to 'Measure concentration for Euclidean distance in the case of dependent random variables'," *Annals of Probability*, vol. 38, no. 1, pp. 439–442, 2010.

[40] C. Villani, *Topics in Optimal Transportation*. Providence, RI: American Mathematical Society, 2003.

[41] E. Ordentlich and M. Weinberger, "A distribution dependent refinement of Pinsker's inequality," *IEEE Trans. on Information Theory*, vol. 51, no. 5, pp. 1836–1840, May 2005.

[42] D. Berend, P. Harremoës, and A. Kontorovich, "Minimum KL-divergence on complements of L_1 balls," *IEEE Trans. on Information Theory*, vol. 60, no. 6, pp. 3172–3177, June 2014.

[43] M. Talagrand, "Transportation cost for Gaussian and other product measures," *Geometry and Functional Analysis*, vol. 6, no. 3, pp. 587–600, 1996.

[44] S. G. Bobkov and F. Götze, "Exponential integrability and transportation cost related to logarithmic Sobolev inequalities," *Journal of Functional Analysis*, vol. 163, pp. 1–28, 1999.

[45] R. Ahlswede, P. Gács, and J. Körner, "Bounds on conditional probabilities with applications in multi-user communication," *Z. Wahrscheinlichkeitstheorie verw. Gebiete*, vol. 34, pp. 157–177, 1976, see correction in vol. 39, no. 4, pp. 353–354, 1977.

[46] R. Ahlswede and G. Dueck, "Every bad code has a good subcode: a local converse to the coding theorem," *Z. Wahrscheinlichkeitstheorie verw. Gebiete*, vol. 34, pp. 179–182, 1976.

[47] S. Shamai and S. Verdú, "The empirical distribution of good codes," *IEEE Trans. on Information Theory*, vol. 43, no. 3, pp. 836–846, May 1997.

[48] Y. Polyanskiy and S. Verdú, "Empirical distribution of good channel codes with non-vanishing error probability," *IEEE Trans. on Information Theory*, vol. 60, no. 1, pp. 5–21, January 2014.

[49] F. Topsøe, "An information theoretical identity and a problem involving capacity," *Studia Scientiarum Mathematicarum Hungarica*, vol. 2, pp. 291–292, 1967.

[50] U. Augustin, "Gedächtnisfreie Kanäle für diskrete Zeit," *Z. Wahrscheinlichkeitstheorie verw. Gebiete*, vol. 6, pp. 10–61, 1966.