An equipartition property for high-dimensional log-concave distributions

(Invited Paper)

Sergey Bobkov[2]
School of Mathematics
University of Minnesota
206 Church St. S.E.
Minneapolis, MN 55455 USA.
Email: bobkov@math.umn.edu

Mokshay Madiman[1,2]
Department of Statistics
Yale University
24 Hillhouse Avenue
New Haven, CT 06511, USA.
Email: mokshay.madiman@yale.edu

***Abstract*—A new effective equipartition property for log-concave distributions on high-dimensional Euclidean spaces is described, and some applications are sketched.**

I. Introduction

Let $\mathbb{X} = (X_1, X_2, \ldots)$ be a stochastic process with each X_i taking values on the real line $\mathbb{R}$. Suppose that the joint distribution of $X^n = (X_1, \ldots, X_n)$ has a density f with respect to either Lebesgue measure on $\mathbb{R}^n$. We are interested in the random variable

$$\widetilde{h}(X^n) = -\log f(X^n).$$

In the discrete case, the quantity $\widetilde{h}(X^n)$ (using f for the probability mass function in this case, thought of as the density with respect to counting measure on some discrete subset of $\mathbb{R}^n$) is essentially the number of bits needed to represent X by a coding scheme that minimizes average code length (cf. [24]), and therefore may be thought of as the (random) *information content* of X^n. Such an interpretation is not justified in the continuous case, but the quantity $\widetilde{h}(X^n)$ remains of central interest in information theory, statistical physics, and statistics, and so we will with some abuse of terminology continue to call it the information content. Its importance in information theory comes from the fact that it is the building block for Pinsker's information density; its importance in statistical physics comes from the fact that it represents (up to an additive constant involving the logarithm of the partition function) the Hamiltonian or energy of a physical system under a Gibbs measure; and its importance in statistics comes from the fact that it represents the log-likelihood function in the nonparametric inference problem of density estimation.

The average value of the information content of X is the differential entropy. Indeed, the entropy of X^n is defined by

$$h(X^n) = -\int f(x) \log f(x)\, dx = -\,\mathbf{E} \log f(X^n),$$

when it exists. If the limit

$$h(\mathbb{X}) := \lim_{n\to\infty} \frac{h(X^n)}{n}$$

exists, it is called the entropy rate of the process $\mathbb{X}$.

The Shannon-McMillan-Breiman theorem [24], [16], [8] is a central result of information theory; indeed, an early form of this result was called by McMillan the "fundamental theorem of information theory" [16]. (McMillan also gave it the pithy and expressive title of the "Asymptotic Equipartition Property".) It asserts that for any stationary, ergodic process whose entropy rate exists, the information content per coordinate converges almost surely to the entropy rate. This version, and a generalization involving log likelihood ratios with respect to a Markov process, is due independently to Barron [2] and Orey [20]; the definitive version for asymptotically mean stationary processes is due to [2], and Algoet and Cover [1] give an elementary approach to it. The theorem implies in particular that for purposes of coding discrete data from ergodic sources, it is sufficient to consider "typical sequences", and that the entropy rate of the process plays an essential role in characterizing fundamental limits of compressing data from such sources.

Our main goal in this note is to present a result akin to the Shannon-McMillan-Breiman theorem for log-concave distributions, which need not arise as marginals of an asymptotically mean stationary process (the most general condition under which such a theorem is known). Log-concavity is a global restriction on the joint distribution of the process, just like stationarity or ergodicity; however it is a shape restriction as opposed to a restriction requiring certain symmetries. In fact, instead of establishing a limiting result involving convergence to an entropy rate (which may not necessarily exist for a "log-concave process")– we give, in the finite-dimensional (i.e., finite sample size) setting, exponential

[1]Corresponding author.
[2]M.M. was supported by the NSF grants DMS-1056996 (CAREER) and CCF-1065494. S.B. was supported by NSF grant DMS-1106530.

probability bounds quantifying how close the information content per coordinate must be to its mean. In other words, one may think of our result as a "non-asymptotic equipartition principle" (NEP). It is well known that for stationary, ergodic processes, such probability bounds are impossible to obtain in general (they can only be usually obtained in the simplest of models such as processes with independent and identically distributed components, Gaussian processes, or certain classes of Markov processes).

For the convenience of the reader, let us recall the notion of a log-concave density. A probability density function (or simply "density") f defined on the linear space $\mathbb{R}^n$ is said to be log-concave if

$$f(\alpha x + (1-\alpha)y) \geq f(x)^{\alpha} f(y)^{1-\alpha}, \tag{1}$$

for each $x, y \in \mathbb{R}^n$ and each $0 \leq \alpha \leq 1$. If f is log-concave, we will also use the adjective "log-concave" for a random variable X distributed according to f, and for the probability distribution induced by it. Log concavity has been deeply studied in probability, statistics, optimization and geometry. Log-concave probability measures include a large variety of distributions including the uniform distribution on any compact, convex set, the exponential and Laplacian (double-exponential) distributions, and of course any Gaussian.

Section II contains the main effective equipartition result. This is interpreted in terms of typical sets in Section III. Sections IV and V describe how our main result can be used as a step in proving a reverse entropy power inequality, which is a result at the interface between convex geometry and information theory. Finally, in Section VI, we develop a version of the equipartition result for the mutual information density, and discuss some possible applications of this to communications. Although several of the key technical results described in this note have recently appeared in [3], [4], [6], our purpose here is to provide an exposition and elaboration of those results, and to discuss their potential consequences for the communication and control communities.

II. The equipartition property

Suppose X is a random vector taking values in $\mathbb{R}^n$. The quantity $\widetilde{h}(X) - h(X)$ is the centered information content, and is a central object of interest in information theory. Note that the distribution of the difference $\widetilde{h}(X) - h(X)$ is stable under all affine transformations of the space, in fact,

$$\widetilde{h}(TX) - h(TX) = \widetilde{h}(X) - h(X)$$

for all invertible affine maps $T : \mathbb{R}^n \to \mathbb{R}^n$.

If $X = (X_1, \ldots, X_n)$ has i.i.d. components, then

$$\widetilde{h}(X) - h(X) = \sum_{i=1}^{n} \widetilde{h}(X_i) - h(X_i)$$

is a sum of i.i.d. random variables, so that a $\sqrt{n}$-normalization yields a central limit theorem (Gaussian approximation) for the centered information content. In particular, in this case, it would not be surprising to find that

$$\mathbf{P}\left\{ \frac{1}{\sqrt{n}} \left| \widetilde{h}(X) - h(X) \right| \geq t \right\}$$

decays exponentially or even has a Gaussian tail as t increases, and indeed, such a concentration inequality can be proved under appropriate assumptions by standard large deviation techniques. In particular, when X is standard (multivariate) normal, we have the explicit formula

$$\widetilde{h}(X) - h(X) = \sum_{i=1}^{n} \frac{X_i^2 - 1}{2},$$

whose distribution is related to the chi-squared distribution. Thus, for any Gaussian distribution, tight concentration inequalities for the normalized and centered information content may be derived by simple explicit calculation. This was done by Cover and Pombra [9] as an ingredient in studying the feedback capacity of time-varying additive Gaussian noise channels.

For general random vectors X, there is no reason to expect that the centered information content of X, when normalized, should be concentrated around 0 with high probability. Nonetheless, our main result is the somewhat surprising fact that not only does such a concentration property hold for any log-concave X, but it holds *uniformly* for *all* log-concave random vectors X.

Theorem 1: Suppose $X = (X_1, \ldots, X_n)$ is distributed according to a log-concave density f on $\mathbb{R}^n$. Then, for all $t > 0$,

$$\mathbf{P}\left\{ \frac{1}{n} \left| \widetilde{h}(X) - h(X) \right| \geq t \right\} \leq 2\, e^{-ct\sqrt{n}},$$

where $c \geq 1/16$.

Once again we emphasize that this a *uniform* bound– the decaying quantity on the right side depends only on t and n, and not on the density f (as long as it is log-concave). In particular, Theorem 1 can be considered as a generalization of the Cover-Pombra result from the finite-dimensional family of n-dimensional Gaussian distributions to the infinite-dimensional family of n-dimensional log-concave distributions. However, the constant 1/16 is not tight and can be improved.

Observe that Theorem 1 can be rewritten (by setting $s = t\sqrt{n}$) as

$$\mathbf{P}\left\{ \frac{1}{\sqrt{n}} \left| \widetilde{h}(X) - h(X) \right| \geq s \right\} \leq 2\, e^{-s/16}.$$

Comparing this with the i.i.d. and Gaussian cases mentioned earlier, we see that the normalization (with respect to n) in

Theorem 1 is chosen correctly and cannot be improved for the class of log-concave distributions.

The exponential decay with respect to t in Theorem 1 may be improved to Gaussian decay, but only on an interval $0 < t < O(1)$ rather than on the whole real line.

Theorem 2: Let X be a random vector in $\mathbb{R}^n$ with log-concave density f. For any $0 \leq t \leq 2$,

$$\mathbf{P}\left\{\frac{1}{n}\,|\log f(X) - \mathbf{E}\log f(X)| \geq t\right\} \leq 4e^{-ct^2 n},$$

where $c \geq 1/16$.

Observe that, together, Theorems 1 and 2 can be thought of in analogy to the Bernstein inequality, since they express Gaussian decay for small deviation parameter, and exponential decay when the deviation parameter gets large.

We outline the proof template for both Theorems 1 and 2 below (see [3] for details):

Step 1: First, a one-dimensional version is proved. If a random variable X has a log-concave density f, then

$$\mathbf{E}e^{\frac{1}{2}\,|\log f(X) - \mathbf{E}\log f(X)|} < 4. \tag{2}$$

The basic idea here is to use the explicit representation of X as $F^{-1}(U)$, where U is uniformly distributed on (0,1), and the concavity of $f \circ F^{-1}$ that follows from log-concavity of f. Observe that this first step already yields, by Markov's inequality, a strong concentration inequality for the information content of any one-dimensional log-concave random variable.

Step 2: A highly non-trivial dimension reduction technique, the Lovász-Simonovits localization lemma, can be used to reduce the desired statement for $\mathbb{R}^n$ to a one-dimensional statement. Unfortunately this one-dimensional statement is not the same as inequality (2); however it can be decomposed into a part that depends on Step 1, and a separate inequality that expresses the concentration of the function $\log Y$ for Y drawn from a class of "strongly" log-concave one-dimensional distributions.

Step 3: The concentration inequality for $\log Y$ mentioned at the end of Step 2 is proved.

III. Typical Sets for Log-Concave Distributions

As suggested in the Introduction, Theorems 1 and 2 may be interpreted as a non-asymptotic equipartition property (NEP) for log-concave distributions. Define the class of typical observables as consisting of all those $x \in \mathbb{R}^n$ such that $f(x)$ lies between

$$\exp\left[-n\left(\frac{h(X)}{n} + \epsilon\right)\right] \quad \text{and} \quad \exp\left[-n\left(\frac{h(X)}{n} - \epsilon\right)\right]$$

for some small fixed $\epsilon > 0$. This subset of $\mathbb{R}^n$ is called the "typical set" since the probability that X lies in this set is close to 1 (by Theorem 1) when the dimension n is large. Thus, for large but finite n, the distribution of X is effectively the uniform distribution on the typical set.

Our NEP also implies that among all sets of high probability, the typical set has the least volume (this can be seen in the standard way). However sometimes it is useful to consider one-sided relaxations of the typical set, and in the log-concave context, superlevel sets of the density are especially interesting.

Proposition 1: Let X have log-concave density f on $\mathbb{R}^n$, and consider

$$A(\lambda) = \{x \in \mathbb{R}^n : f(x) > \lambda\}$$

with $\lambda = e^{-(h(X)+n\epsilon)}$. Then $A(\lambda)$ is a bounded, convex set such that $\mathbf{P}\{X \in A(\lambda)\} = p \approx 1$ (quantified by Theorem 1). Furthermore, the volume of the set $A(\lambda)$ can be sandwiched as follows:

$$\frac{p}{\|f\|_\infty} \leq |A(\lambda)| \leq \frac{1}{\lambda},$$

where $\|f\|_\infty$ is the essential supremum of f (which is finite since f is log-concave).

Proof: The convexity and boundedness of $A(\lambda)$ follow immediately from the fact that f is a log-concave density. For the volume upper bound, note that

$$1 \geq \int_A f(x)\,dx \geq \lambda |A(\lambda)|,$$

while for the lower bound, note that

$$p \leq \mu(A(\lambda)) = \int_{A(\lambda)} f(x)\,dx \leq |A(\lambda)| \cdot \|f\|_\infty.$$

■

Proposition 1 is exploited in a later section as a step in proving a reverse entropy power inequality for log-concave distributions. It also implies that the typical set itself has nice properties– it is the "annulus" between two nested convex sets, and its volume can be upper and lower bounded explicitly in terms of simple parameters.

Observe that a well known concentration property of the standard Gaussian distribution implies that a random vector from this distribution lies with high probability in a thin shell around the sphere of radius $\sqrt{n}$. The concentration of random vectors with arbitrary log-concave distributions in an annulus between nested convex sets may be seen as a generalization of this fact.

So far, our discussion has focused on typical sets for a log-concave distribution in high but fixed dimension. Let us comment now on the case where the dimension is increasing. This could happen either because one is looking at higher-dimensional marginals of a discrete-time stochastic process, or because it is natural to consider triangular arrays of random vectors X^n (rather than just a single linear stream of random variables). Unlike the usual AEP for ergodic processes, where the typical set is determined by a constant

(the entropy rate), the typical set in our context is pegged to the possibly moving target $h(X^n)/n$. Indeed, recall that the typical set is

$$A_\varepsilon^{(n)} = \{x \in \mathbb{R}^n : \left| -\frac{1}{n} \log f(X^n) - \frac{h(X^n)}{n} \right| \le \varepsilon\},$$

and for fixed ε, Theorem 1 implies that $P(A_\varepsilon^{(n)}) \to 1$ as $n \to \infty$. In fact, it is easy to see that Theorem 1 actually answers the finer (and natural) question: how fast can we decrease ε with n while keeping $A_\varepsilon^{(n)}$ a typical set?

Corollary 1: Suppose that $\mathbb{X} = (X_1, X_2, \ldots)$ is a stochastic process with finite dimensional marginals that are absolutely continuous and log-concave. If $n\varepsilon_n^2 \to \infty$, and $\varepsilon_n \le 2$, then

$$P(A_{\varepsilon_n}^{(n)}) \to 1$$

as $n \to \infty$.

Note that the condition of Corollary 1 is satisfied if $\varepsilon_n = \Omega\left(\sqrt{\frac{\log\log n}{n}}\right)$.

IV. M-POSITION FOR LOG-CONCAVE DISTRIBUTIONS

The so-called M-position of convex bodies was introduced by Milman in connection with reverse forms of the Brunn-Minkowski inequality [17]. By now several equivalent definitions of this important concept are known, and for our purposes we choose one of them. We refer the interested reader to the book by Pisier [21], which also contains historical remarks.

For any convex body A in $\mathbb{R}^n$, define

$$M(A) = \sup_{|\mathcal{E}|=|A|} \frac{|A \cap \mathcal{E}|^{1/n}}{|A|^{1/n}},$$

where the supremum is over all ellipsoids $\mathcal{E}$ with volume $|\mathcal{E}| = |A|$. The main result of Milman may be stated as follows (it was originally stated for centrally symmetric convex bodies, but the symmetry assumption can be removed):

Proposition 2: If A is a convex body in $\mathbb{R}^n$, then with some universal constant $c > 0$

$$M(A) \ge c.$$

If $|A \cap \mathcal{E}|^{1/n} \ge c\,|A|^{1/n}$ with c being the universal constant of Proposition 2, then $\mathcal{E}$ is called an M-ellipsoid or Milman's ellipsoid.

It follows from the definition that, for any convex body A in $\mathbb{R}^n$, one can find an affine volume preserving map $u : \mathbb{R}^n \to \mathbb{R}^n$ such that $u(A)$ has a multiple of the unit centered Euclidean ball as an M-ellipsoid. In that case, one says that $u(A)$ is in M-position. Or equivalently, A is in M-position, if

$$|A \cap D|^{1/n} \ge c\,|A|^{1/n}, \tag{1}$$

where D is a Euclidean ball with center at the origin, such that $|D| = |A|$, and where $c > 0$ is universal.

The definition of M-position may naturally be extended to probability distributions.

Definition 1: Let μ be a distribution on $\mathbb{R}^n$. Then we say that μ is in M-position with constant $c > 0$, if

$$\mu(D)^{1/n} \ge c, \tag{2}$$

where D is a Euclidean ball with center at the origin of volume $|D| = 1$.

Correspondingly, Proposition 2 can be generalized to log-concave measures. In order to do this, we need some normalization of the density (corresponding to normalization by volume for the convex body case). A relevant normalization is to constrain the maximum of the density.

Proposition 3: Let μ be a probability measure on $\mathbb{R}^n$ with log-concave density f such that $\|f\|_\infty \ge 1$. Then there exists an affine volume-preserving map $u : \mathbb{R}^n \to \mathbb{R}^n$ such that the image $\tilde{\mu} == \mu u^{-1}$ of the measure μ under the map u satisfies

$$\tilde{\mu}(D)^{1/n} \ge c_M,$$

where D is the Euclidean ball of volume one, and $c_M \in (0, 1)$ is a universal constant.

We say that any log-concave probability measure μ on $\mathbb{R}^n$ with density f such that $\|f\|_\infty = 1$ may be "put in M-position" (i.e., by applying a linear transformation of determinant ± 1 to the random vector) with a universal constant.

Proof: We may assume that $\|f\|_\infty = 1$. By Proposition 1, for some constant $c_0 > 0$, the essential support of μ, i.e., the set $K_f = \{f(x) \ge c_0^n\}$ has measure $\mu(K_f) \ge 1/2$. Hence, again by Proposition 1, we have

$$\frac{1}{2} \le |K_f|^{1/n} \le c_0^{-1}.$$

Put $K_f' = \frac{1}{|K_f|^{1/n}} K_f$, which is a convex body with volume $|K_f'| = 1$. One may assume that K_f' contains the origin and is already in M-position (otherwise, apply to K_f' a linear, volume preserving map u to put it in M-position and consider the image $u(\mu)$ in place of μ). We claim that if K_f' is in M-position, then μ is also in M-position.

Indeed, if D is the Euclidean ball with center at the origin of volume $|D| = 1$, then (1) is satisfied for the set $A = K_f'$ with a universal constant $c > 0$. Since $K_f' \subset 2K_f$, we have $|K_f' \cap D| \le |2K_f \cap D| \le 2^n |K_f \cap D|$. Therefore,

$$\mu(D) \ge \int_{K_f \cap D} f(x)\,dx \ge c_0^n\,|K_f \cap D|$$
$$\ge c_0^n \cdot 2^{-n} |K_f' \cap D| \ge \left(\frac{c_0 c}{2}\right)^n.$$

Proposition 3 is proved. ∎

V. A Reverse EPI

Given a random vector X in $\mathbb{R}^n$ with density $f(x)$, the entropy power is defined by $\mathcal{N}(X) = e^{2h(X)/n}$. The entropy power inequality, due to Shannon and Stam [24], [25], asserts that

$$\mathcal{N}(X+Y) \geq \mathcal{N}(X) + \mathcal{N}(Y), \tag{3}$$

for any two independent random vectors X and Y in $\mathbb{R}^n$, for which the entropy is defined.

The entropy power inequality may be formally strengthened by using the invariance of entropy under affine transformations of determinant ± 1, i.e., $\mathcal{N}(u(X)) = \mathcal{N}(X)$ whenever $|\det(u)| = 1$. Specifically,

$$\inf_{u_1,u_2} \mathcal{N}(u_1(X) + u_2(Y)) \geq \mathcal{N}(X) + \mathcal{N}(Y), \tag{4}$$

where the maps $u_i : \mathbb{R}^n \to \mathbb{R}^n$ range over all affine entropy-preserving transformations. It turns out that in exact analogy to the so-called reverse Brunn-Minkowski inequality of Milman [17] (see also [18], [19], [21]), which is a celebrated result in convex geometry, the inequality (4) can be reversed with a constant not depending on dimension if we restrict to log-concave distributions.

Theorem 3: If X and Y are independent random vectors in $\mathbb{R}^n$ with log-concave densities, there exist affine entropy-preserving maps $u_i : \mathbb{R}^n \to \mathbb{R}^n$ such that

$$\mathcal{N}\big(\widetilde{X} + \widetilde{Y}\big) \leq C\,(\mathcal{N}(X) + \mathcal{N}(Y)), \tag{5}$$

where $\widetilde{X} = u_1(X)$, $\widetilde{Y} = u_2(Y)$, and where C is a universal constant.

Specializing to uniform distributions on convex bodies, we can show (details in [6]) that Theorem 3 recovers Milman's reverse Brunn-Minkowski inequality [17]. Thus one may think of Theorem 3 as completing in a reverse direction the usual analogy (see, e.g., [10]) between the Brunn-Minkowski and entropy power inequalities.

To prove Theorem 3, we need two lemmas of independent interest. The first is an entropy comparison inequality: in terms of the "amount of randomness" as measured by entropy per coordinate, any log-concave random vector of any dimension contains randomness that differs from that in the normal random variable with the same maximal density value by at most 1/2. For our purposes, it is convenient to write the lemma in the following form (see [5], where the lemma is proved and several applications of it are studied, for why it is equivalent to the previous statement).

Lemma 1: If a random vector X in $\mathbb{R}^n$ has log-concave density f, then

$$h(X) \leq n + \log \|f\|_\infty^{-1}.$$

The second lemma we need is the "submodularity" property of the differential entropy with respect to convolutions, which is rather elementary but was first explicitly observed (to our knowledge) in [15].

Lemma 2: Given independent random vectors X, Y, Z in $\mathbb{R}^n$ with absolutely continuous distributions,

$$h(X+Y+Z) + h(Z) \leq h(X+Z) + h(Y+Z)$$

provided that all entropies are well-defined.

This inequality is not hard to prove; indeed, it reduces to the data processing inequality for mutual information. However, although it is simple, it has some interesting applications. First, it can be used as an ingredient in proving the reverse entropy power inequality, which is why we discuss it here. Other applications are discussed in [15], [14], [6], [13].

Proposition 4: Let X be distributed according to a log-concave measure μ on $\mathbb{R}^n$, which is in M-position with the universal constant c_M, and has density f such that $\|f\|_\infty \geq 1$. Then

$$h(X+Z) \leq c_A n,$$

where Z is uniformly distributed on the Euclidean ball D of volume one, and $c_A = 1 - \log c_M \in (1, \infty)$ is an absolute constant.

Proof: The density p of $S = \widetilde{X} + Z$ is given by

$$p(x) = \int_D f(x-z)\,dz = \mu(D-x).$$

Since μ is in M-position, we have

$$p(0) = \widetilde{\mu}(D) \geq c_M^n.$$

with the universal constant $c_M \in (0,1)$. In particular, $\log \|p\|_\infty \geq \log p(0) \geq n \log c_M$. Applying Lemma 1 to the random vector S gives

$$h(S) \leq n + \log \|p\|_\infty^{-1} \leq (1 - \log c_M) n,$$

which completes the proof. ∎

We can now give the proof of Theorem 3.

Proof: Suppose X and Y have log-concave densities f and g respectively. Note that Theorem 3 is homogeneous with respect to scaling the random variables by a constant, i.e.,

$$\mathcal{N}\big(a\widetilde{X} + a\widetilde{Y}\big) \leq C\,(\mathcal{N}(aX) + \mathcal{N}(aY))$$

is the same as (5) since $\mathcal{N}(aX) = a^2 N(X)$. Thus by choosing a appropriately, we can assume without loss of generality that $\|f\|_\infty \geq 1$ and $\|g\|_\infty \geq 1$.

Let u_1 and u_2 be the affine volume-preserving transformations that put X and Y in M-position (that these exist is the content of Proposition 3). Since $h(Z) = 0$ for Z being

uniformly distributed on the Euclidean ball D of unit volume, Lemma 2 implies

$$h(\widetilde{X}+\widetilde{Y}) \le h(\widetilde{X}+\widetilde{Y}+Z) \le h(\widetilde{X}+Z) \; + \; h(\widetilde{Y}+Z),$$

when $\widetilde{X}, \widetilde{Y}, Z$ are independent random vectors. Now, by Proposition 4, since $\widetilde{X}$ and $\widetilde{Y}$ are in M-position and the maximum values of their density functions are at least 1, each of the terms on the right is bounded from above by $c_A n$. Thus $h(\widetilde{X}+\widetilde{Y}) \le 2c_A n$, so that $\mathcal{N}(\widetilde{X}+\widetilde{Y}) \le e^{4c_A}$. This completes the proof. ∎

The universal constant provided by the proof of Theorem 3 is not explicit, and it is not easy to even get bounds on it. However, in the special case where X and Y are not just independent but also identically distributed, an explicit constant can be obtained by other means; see [7] for details.

VI. On applications to communications

Although we have discussed implications of our NEP for typical sets, we have not explicitly discussed the applications to compression and communication. Since we are in a continuous setting, these arise through the implications for the information density. The (mutual) information density between X and Y is defined by

$$\iota(X^n, Y^n) = \frac{1}{n} \log \frac{f_{X^n,Y^n}(X^n, Y^n)}{f_{X^n}(X^n) \cdot f_{Y^n}(Y^n)},$$

and plays a key role in channel coding.

Theorem 4: Suppose that (X^n, Y^n) taking values in $\mathbb{R}^n \times \mathbb{R}^n$ has a log-concave density f_{X^n,Y^n}. Then, for any $0 < t \le 6$ and any n,

$$\mathbf{P}\{\, |\iota(X^n, Y^n) - \mathbf{E}\iota(X^n, Y^n)| \ge t\} \le 10e^{-c' n t^2}, \qquad (6)$$

where $c' \ge 1/144$ is a universal constant.

Proof: Let us first note that the triangle inequality allows us to reduce the event of interest to events that we know how to treat.

$$\begin{aligned}
&\mathbf{P}\{\, |\iota(X^n, Y^n) - \mathbf{E}\iota(X^n, Y^n)| \ge t\} \\
&= \mathbf{P}\{\, |\log f_{X^n,Y^n}(X^n, Y^n) - \mathbf{E}\log f_{X^n,Y^n}(X^n, Y^n) \\
&\quad - [\log f_{X^n}(X^n) - \mathbf{E}\log f_{X^n}(X^n)] \\
&\quad - [\log f_{Y^n}(Y^n) - \mathbf{E}\log f_{Y^n}(Y^n)]| \ge nt\} \\
&\le \mathbf{P}\{\, E_1 \cup E_2 \cup E_3\},
\end{aligned}$$

where

$$\begin{aligned}
E_1 &= \{|\log f_{X^n,Y^n}(X^n, Y^n) - \\
&\qquad \mathbf{E}\log f_{X^n,Y^n}(X^n, Y^n)| \ge nt/3\}, \\
E_2 &= \{|\log f_{X^n}(X^n) - \mathbf{E}\log f_{X^n}(X^n)| \ge nt/3\}, \\
E_3 &= \{|\log f_{Y^n}(Y^n) - \mathbf{E}\log f_{Y^n}(Y^n)| \ge nt/3\}.
\end{aligned}$$

Consequently, by the union bound,

$$\begin{aligned}
&\mathbf{P}\{\, |\iota(X^n, Y^n) - \mathbf{E}\iota(X^n, Y^n)| \ge t\} \\
&\le P(E_1) + P(E_2) + P(E_3) \\
&\le 4e^{-2cnt^2/9} + 4e^{-cnt^2/9} + 4e^{-cnt^2/9}.
\end{aligned}$$

Observe that the bounds in the last step hold by 3 uses of Theorem 1, each under the assumption that $t/3 \le 2$. Moreover, the first term $4e^{-2cnt^2/9} = (2e^{-cnt^2/9})^2 \le 2e^{-cnt^2/9}$ provided $2e^{-cnt^2/9} \le 1$, but the upper bound in (6) is in any case trivial (greater than 1) if this does not hold. The bound $c' \ge 1/144$ follows from the bound on c in Theorem 1. ∎

The constants in Theorem 4 can be significantly improved by refining the constants in Theorem 1 beyond what is proved in [3]; this, however, requires more work and we will develop it elsewhere.

Note that $\mathbf{E}\iota(X^n, Y^n) = \frac{1}{n} I(X^n; Y^n)$; so Theorem 4 expresses the concentration of the information density around the mutual information per symbol. Such concentration inequalities for the information density are exactly the kinds of inequalities required for non-asymptotic information theory (i.e., fundamental limits of communication for finite block lengths, without assumptions on memory etc.). Although questions about fundamental limits of communications in non-asymptotic settings have been studied for a long time (e.g., through reliability functions in Gallager's book [11]), classically these questions were studied for very special channels like the binary symmetric channel (with the notable exception of a paper of Strassen [26] focusing on general discrete memoryless channels) or the bounds obtained were not very satisfactory. However, non-asymptotic information theory has seen a rapid resurgence in recent years, spurred by work of Polyanskiy, Poor and Verdú [22], [23] and Hayashi [12] (see also Yang and Meng [28], [27]). We believe that Theorem 4 has implications for non-asymptotic information theory, and expect to develop these in future work.

Theorem 4 also generalizes a key technical result of Cover and Pombra [9] from additive Gaussian channels to additive channels with log-concave noise, *provided* we restrict to log-concave input distributions. Indeed, suppose the output sequence Y^n is given by $X^n + Z^n$, where X^n is an input sequence with log-concave joint distribution, and the noise Z^n has a log-concave joint density (it would be natural to assume that the noise distribution is also symmetric about 0, i.e., even, but we do not need this for the current discussion). Then the joint density of (X^n, Y^n) is given by

$$\begin{aligned}
f_{X^n,Y^n}(x^n, y^n) &= f_{X^n}(x^n) f_{Z^n}(y^n - x^n) \\
&= \exp\{-[V_1(x^n) + V_2(y^n - x^n)]\},
\end{aligned}$$

where V_1, V_2 are convex functions. Since $V_1(x^n) + V_2(y^n - x^n)$ is convex in the pair (x^n, y^n), the joint density f_{X^n,Y^n} is log-concave, and so Theorem 4 is applicable. Of course, in principle, one will not immediately get capacity results from this observation because of the restriction on the input

distribution (although note that considerable dependence both in the input and noise sequences is still allowed by the restriction). However, we expect that the above observations will still be useful in studying capacity of channels with time-varying log-concave additive noise with and without feedback, and expect to develop these in future work as well.

REFERENCES

[1] P.H. Algoet and T.M. Cover. A sandwich proof of the Shannon-Mcmillan-Breiman theorem. *Ann. Probab.*, 16:876–898, 1988.

[2] A.R. Barron. The strong ergodic theorem for densities: Generalized Shannon-Mcmillan-Breiman theorem. *Ann. Probab.*, 13:1292–1303, 1985.

[3] S. Bobkov and M. Madiman. Concentration of the information in data with log-concave distributions. *Ann. Probab.*, 39(4):1528–1543, 2011.

[4] S. Bobkov and M. Madiman. Dimensional behaviour of entropy and information. *C. R. Acad. Sci. Paris Sér. I Math.*, 349:201–204, Février 2011.

[5] S. Bobkov and M. Madiman. The entropy per coordinate of a random vector is highly constrained under convexity conditions. *IEEE Trans. Inform. Theory*, 57(8):4940–4954, August 2011.

[6] S. Bobkov and M. Madiman. Reverse Brunn-Minkowski and reverse entropy power inequalities for convex measures. *J. Funct. Anal.*, 262:3309–3339, 2012.

[7] S. G. Bobkov and M. M. Madiman. On the problem of reversibility of the entropy power inequality. *To appear in Festschrift on the occasion of F. Götze's 60th birthday*, 2012. Available online at `http://arxiv.org/abs/1111.6807`.

[8] L. Breiman. The individual ergodic theorem for information theory. *Ann. Math. Stat.*, 28:809–811, 1957. (See also the correction: *Ann. Math. Stat.*, Vol. 31, pp. 809-810, 1960).

[9] T. M. Cover and S. Pombra. Gaussian feedback capacity. *IEEE Trans. Inform. Theory*, 35(1):37–43, 1989.

[10] A. Dembo, T.M. Cover, and J.A. Thomas. Information-theoretic inequalities. *IEEE Trans. Inform. Theory*, 37(6):1501–1518, 1991.

[11] R. Gallager. *Information theory and reliable communication*. Wiley, New York, 1968.

[12] M. Hayashi. Information spectrum approach to second-order coding rate in channel coding. *IEEE Trans. Inform. Theory*, 55(11):4947–4966, 2009.

[13] I. Kontoyiannis and M. Madiman. Sumset and inverse sumset inequalities for differential entropy and mutual information. *Preprint*, 2012.

[14] M. Madiman. Determinant and trace inequalities for sums of positive-definite matrices. *Preprint*, 2008.

[15] M. Madiman. On the entropy of sums. In *Proc. IEEE Inform. Theory Workshop*, pages 303–307. Porto, Portugal, 2008.

[16] B. McMillan. The basic theorems of information theory. *Ann. Math. Stat.*, 24:196–219, 1953.

[17] V. D. Milman. Inégalité de Brunn-Minkowski inverse et applications à la théorie locale des espaces normés. *C. R. Acad. Sci. Paris Sér. I Math.*, 302(1):25–28, 1986.

[18] V. D. Milman. Entropy point of view on some geometric inequalities. *C. R. Acad. Sci. Paris Sér. I Math.*, 306(14):611–615, 1988.

[19] V. D. Milman. Isomorphic symmetrizations and geometric inequalities. In *Geometric aspects of functional analysis (1986/87)*, volume 1317 of *Lecture Notes in Math.*, pages 107–131. Springer, Berlin, 1988.

[20] S. Orey. On the Shannon-Perez-Moy theorem. In *Particle systems, random media and large deviations (Brunswick, Maine, 1984)*, pages 319–327. Amer. Math. Soc., Providence, R.I., 1985.

[21] G. Pisier. *The volume of convex bodies and Banach space geometry*, volume 94 of *Cambridge Tracts in Mathematics*. Cambridge University Press, Cambridge, 1989.

[22] Y. Polyanskiy, H. V. Poor, and S. Verdú. Channel coding rate in the finite blocklength regime. *IEEE Trans. Inform. Theory*, 56(5):2307–2359, 2010.

[23] Y. Polyanskiy, H. V. Poor, and S. Verdú. Feedback in the non-asymptotic regime. *IEEE Trans. Inform. Theory*, 57(8):4903–4925, 2011.

[24] C.E. Shannon. A mathematical theory of communication. *Bell System Tech. J.*, 27:379–423, 623–656, 1948.

[25] A.J. Stam. Some inequalities satisfied by the quantities of information of Fisher and Shannon. *Information and Control*, 2:101–112, 1959.

[26] V. Strassen. Asymptotische abschätzungen in Shannon's informationstheorie. In *Trans. 3rd Prague Conf. Inf. Theory*, pages 689–723, Prague, 1962.

[27] E.-H. Yang and J. Meng. Jar decoding: Non-asymptotic converse coding theorems, taylor-type expansion, and optimality. *Preprint,* `arXiv:1204:3658`, 2012.

[28] E.-H. Yang and J. Meng. Non-asymptotic equipartition properties for independent and identically distributed sources. *Preprint,* `arXiv:1204:3661`, 2012.