A Quick Primer on Entropic Limit Theorems

Mokshay Madiman

Yale University

NSF/CBMS Conference on Small Deviation Probabilities

University of Alabama, Huntsville

June 5, 2012

Themes in the intertwining of Information theory and Probability

Classical themes

- Stationary, ergodic processes (e.g., Shannon-McMillan-Breiman theorem)
- Large deviations (e.g., rate function in Sanov's theorem)
- Mathematical physics motivations (e.g., convergence to equilibrium)

Recent themes

- Entropic Limit Theorems (TODAY's FOCUS)
- Information theory and High-Dimensional Convex Geometry
- Information-theoretic inequalities in Combinatorics
- Information theory and Statistics

Entropy

- When random variable X has density $f(x)$ on $\mathbb{R}$, the entropy of X is

$$h(X) = h(f) := -\int_{\mathbb{R}} f(x) \log f(x) dx = E[-\log f(X)]$$

- The relative entropy between the distributions of $X \sim f$ and $Y \sim g$ is

$$D(f\|g) = \int f(x) \log \frac{f(x)}{g(x)} dx$$

  For any f, g, $D(f\|g) \geq 0$ with equality iff $f = g$

Why are they relevant?

- Entropy is a measure of randomness
- Relative Entropy is a very useful notion of "distance" between probability measures (non-negative, and dominates several of the usual distances, although non-symmetric)

Non-Gaussianity

For $X \sim f$ in $\mathbb{R}$, its relative entropy from Gaussianity is

$$D(X) = D(f) := D(f \| f^G),$$

where f^G is the Gaussian with the same mean and variance as X

Observe:

- For any density f, its non-Gaussianity $D(f) = h(f^G) - h(f)$

  Proof: Gaussian density is exponential in first two moments

- Thus Gaussian is MaxEnt: $N(0, \sigma^2)$ has maximum entropy among all densities on $\mathbb{R}$ with variance $\leq \sigma^2$

  Proof: $D(f) \geq 0$

Entropic Central Limit Theorem

Two observations . . .

- Gaussian is MaxEnt: $N(0, \sigma^2)$ has maximum entropy among all densities on $\mathbb{R}$ with variance $\leq \sigma^2$
- Let X_i be i.i.d. with $EX_1 = 0$ and $EX_1^2 = \sigma^2$.

  For the CLT, we are interested in $S_M := \frac{1}{\sqrt{M}} \sum_{i=1}^{M} X_i$

  The CLT scaling preserves variance

suggest . . .

Question: Is it possible that the CLT may be interpreted like the 2nd law of thermodynamics, in the sense that $h(S_M)$ monotonically increases in M until it hits the maximum entropy possible (namely, the entropy of the Gaussian)?

The Entropic Central Limit Theorem

If $D(S_M) < \infty$ for some M, then as $M \to \infty$,

$$D(S_M) \downarrow 0 \qquad \text{or equivalently,} \qquad h(S_M) \uparrow h(N(0, \sigma^2))$$

Remarks

- Convergence shown by Barron '86
- Monotonicity shown by Artstein-Ball-Barthe-Naor '04 with simple proofs by Barron–M. '06-'07, Tulino–Verdú '06
- Monotonicity in n indicates that the entropy is a $natural\ measure$ for CLT convergence (cf. second law of thermodynamics)

Original Entropy Power Inequality (EPI)

For independent random variables with densities,

$$e^{2h(X_1+X_2)} \geq e^{2h(X_1)} + e^{2h(X_2)} \qquad \text{[Shannon '48, Stam '59]}$$

Remarks

- The non-negative number $e^{2h(X)}$ is called the entropy power of X
- The EPI is quite powerful: it implies both the Gaussian logarithmic Sobolev inequality and the Heisenberg-Pauli-Weyl uncertainty principle
- Equality holds if and only if both X_1 and X_2 are normal
- Since $h(aX) = h(X) + \log|a|$, implies for i.i.d. X_i,

$$h\left(\frac{X_1 + X_2}{\sqrt{2}}\right) \geq h(X_1)$$

For X_i i.i.d., if $S_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i$, then $h(S_{2n}) \geq h(S_n)$

Barron '86 used this to prove entropy convergence $h(S_n) \to h(Z_X)$

ABBN's Entropy Power Inequality

Leave-one-out Inequality for independent X_i

$$e^{2h(X_1+\ldots+X_n)} \geq \frac{1}{n-1}\sum_{i=1}^{n} e^{2h\left(\sum_{j\neq i} X_j\right)}$$

CLT Implication

For X_i i.i.d., let $S_n = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i$

- Entropy is an increasing sequence:

$$h(S_{n+1}) \geq h(S_n)$$

- Combining with Barron '86 implies an analogy with the 2nd law

$$h(S_n) \nearrow h(Z_X) \quad \text{and} \quad D(S_n \| Z_X) \searrow 0$$

- The original proof of Artstein–Ball–Barthe–Naor '04 is rather complicated and uses a variational characterization of Fisher information
- We follow Barron–M. '07, who gave a simple proof of a more general result

New Entropy Power Inequality

Subset-sum EPI

For any collection G of subsets s of indices $\{1, 2, \ldots, n\}$,

$$e^{2h(X_1+\ldots+X_n)} \geq \frac{1}{r} \sum_{s \in G} e^{2h(\mathrm{sum}_s)}$$

[Barron–M. '07]

where $\mathrm{sum}_s = \sum_{j \in s} X_j$ is the subset-sum

$r = r(G)$ is the *maximal degree*, the maximum number of subsets in G in which any index i can appear

Examples

- G=singletons, $r = 1$, original EPI
- G=leave-one-out sets, $r = n{-}1$, ABBN's EPI
- G=sets of size m, $r = \binom{n-1}{m-1}$, leave $n{-}m$ out EPI
- G=sets of m consecutive indices, $r = m$

mile-marker

- Entropy and the CLT
- New Entropy power inequalities
- New Fisher Information inequalities
- Simple proof ideas

The Link between h and I

Definitions

- Shannon entropy: $h(X) = E\big[\log \frac{1}{f(X)}\big]$
- Score function: $\mathsf{score}(X) = \frac{\partial}{\partial x} \log f(X)$
- Fisher information: $I(X) = E\left[\mathsf{score}^2(X)\right]$

Relationship

For a standard normal Z independent of X,

- Differential version:

$$\frac{d}{dt} h(X + \sqrt{t}Z) = \tfrac{1}{2} I(X + \sqrt{t}Z)$$

[de Bruijn, see Stam '59]

- Integrated version:

$$h(X) = \tfrac{1}{2}\log(2\pi e) - \tfrac{1}{2}\int_0^\infty \left[I(X + \sqrt{t}Z) - \frac{1}{1+t} \right] dt$$

[Barron '86]

New Fisher Information Inequality

For independent X_1, $X_2, \ldots, X_n$ with differentiable densities, and any collection G of subsets s of indices $\{1, 2, \ldots, n\}$,

$$\frac{1}{I(\text{sum}_{\text{tot}})} \geq \frac{1}{r} \sum_{s \in G} \frac{1}{I(\text{sum}_s)}$$ [Barron–M. '07]

where:

$\text{sum}_{\text{tot}} = \sum_{j=1}^{n} X_j$ is the total sum,

$\text{sum}_s = \sum_{j \in s} X_j$ is the subset-sum,

and $r = r(G)$ is the $maximal\ degree$, the maximum number of subsets in G in which any index i can appear.

Showing this would imply the new EPI, via the transference technique

Score of a sum

Lemma: Suppose

- V_1, V_2 independent random variables
- V_1 has a differentiable density f_1 and score score_1
- $V = V_1 + V_2$ has density f_V and score score

Then

$$\text{score}(V) = E[\text{score}_1(V_1)|V] \qquad \text{[Stam '59, Blachman '65]}$$

Proof

$$f'(v) = \frac{\partial}{\partial v}E[f_1(v - V_2)] = E[f_1'(v - V_2)] = E[f_1(v - V_2)\text{score}_1(v - V_2)]$$

so that

$$\rho(v) = \frac{f'(v)}{f(v)} = E\left[\frac{f_1(v - V_2)}{f(v)}\text{score}_1(v - V_2)\right] = E[\text{score}_1(V_1)|V_1 + V_2 = v].$$

Thus $V = V_1 + V_2$ has the score

$$\text{score}(V) = E[\text{score}_1(V_1)|V]$$

A Projection Inequality

For each subset s,

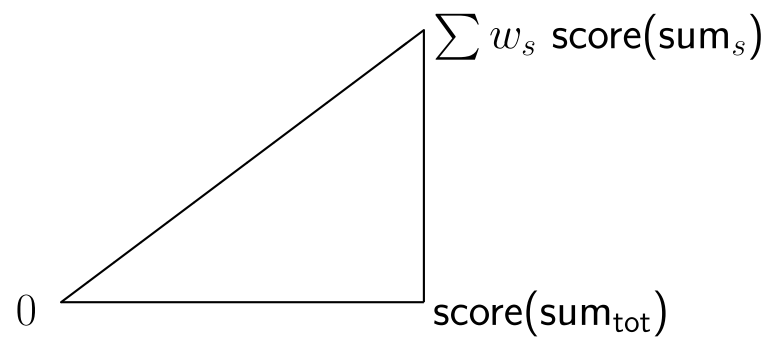
$$\text{score}(\text{sum}_{\text{tot}}) = E\left[\text{score}(\text{sum}_s) \,\middle|\, \text{sum}_{\text{tot}}\right]$$

Hence, for weights w_s that sum to 1,

$$\text{score}(\text{sum}_{\text{tot}}) = E\left[\sum_{s \in G} w_s \ \text{score}(\text{sum}_s) \,\middle|\, \text{sum}_{\text{tot}}\right]$$

Pythagorean inequality

The Fisher info. of the sum is the mean squared length of the projection



$$I(\text{sum}_{\text{tot}}) \le E\left[\sum_{s \in G} w_s \ \text{score}(\text{sum}_s)\right]^2$$

The Variance Drop Lemma

Let X_1, $X_2, \ldots, X_n$ be independent. Let $\underline{X}_s = (X_i : i \in s)$ and $g_s(\underline{X}_s)$ be some mean-zero function of $\underline{X}_s$. Then sums of such functions

$$g(X_1, X_2, \ldots, X_n) = \sum_{s \in G} g_s(\underline{X}_s)$$

have the variance bound

$$Eg^2 \leq r \sum_{s \in G} Eg_s^2(\underline{X}_s) \qquad \text{[Barron–M. '07]}$$

The Variance Drop Lemma

Let X_1, $X_2, \ldots, X_n$ be independent. Let $\underline{X}_s = (X_i : i \in s)$ and $g_s(\underline{X}_s)$ be some mean-zero function of $\underline{X}_s$. Then sums of such functions

$$g(X_1, X_2, \ldots, X_n) = \sum_{s \in G} g_s(\underline{X}_s)$$

have the variance bound

$$E g^2 \leq r \sum_{s \in G} E g_s^2(\underline{X}_s)$$ [Barron–M. '07]

Remarks

- Note that $r \leq |G|$, hence the "variance drop"
- Examples:

  G=singletons has $r = 1$: additivity of variance with independent summands

  G=leave-one-out sets has $r = n-1$ as in the study of the jackknife and U-statistics

- Proof is based on ANalysis Of VAriance decomposition [Hoeffding '48]

The Heart of the Matter

Recall the Pythagorean inequality

$$I(\mathsf{sum}_{\mathsf{tot}}) \le E\left[\sum_{s \in G} w_s \ \mathsf{score}(\mathsf{sum}_s)\right]^2$$

and apply the variance drop lemma to get

$$I(\mathsf{sum}_{\mathsf{tot}}) \le r \sum_{s \in G} w_s^2 I(\mathsf{sum}_s)$$

for all weights w_s that sum to 1.

Optimizing over w yields the new Fisher information inequality

$$\frac{1}{I(\mathsf{sum}_{\mathsf{tot}})} \ge \frac{1}{r} \sum_{s \in G} \frac{1}{I(\mathsf{sum}_s)}$$

Optimized Form for H

We have (again)

$$I(\mathsf{sum}_{\mathsf{tot}}) \le r \sum_{s \in G} w_s^2 I(\mathsf{sum}_s)$$

Equivalently,

$$I(\mathsf{sum}_{\mathsf{tot}}) \le \sum_{s \in G} w_s I\left(\frac{\mathsf{sum}_s}{\sqrt{r w_s}}\right)$$

Adding independent normals and integrating, [*not immediate that this is possible but can be justified*]

$$h(\mathsf{sum}_{\mathsf{tot}}) \ge \sum_{s \in G} w_s h\left(\frac{\mathsf{sum}_s}{\sqrt{r w_s}}\right)$$

Optimizing over w yields the new Entropy Power Inequality

$$e^{2h(\mathsf{sum}_{\mathsf{tot}})} \ge \frac{1}{r} \sum_{s \in G} e^{2h(\mathsf{sum}_s)}$$

Discrete Entropic Limit Theorems

Theorem 2: [Johnson '06]

$$H(\text{Po}(\lambda)) = \max\left\{H(P) \ : \ P \text{ ULC with mean } \lambda\right\}$$

Remarks

- A probability distribution P on $\mathbb{Z}_+$ is *ultra-log-concave* (ULC) if for each k,

$$P(k)^2 \geq \left(\frac{k+1}{k}\right) P(k-1)P(k+1)$$

- The ULC class is closed under convolution [Pemantle '99, Liggett '97]
- Theorem 2 was extended to the much more general *compound Poisson* case by [Johnson-Kontoyiannis-M.'09–'11]
- Latter has applications to combinatorics (random independent sets in matroids etc.)
- Related techniques also allow one to obtain optimal-order approximation bounds for independent summands

Summary

- New Fisher information and entropy power inequalities
- Variance drop lemma of independent interest
- Monotonicity of I and h in central limit theorems ("2nd law")
- A similar entropic view of discrete limit theorems is possible
- Bonus:
  - Statistical proofs with implications for distributed inference
  - Multivariate generalization holds and there are interesting dimension-independent reverse forms for log-concave measures
  - Applications: Capacity/rate regions in multi-user information theory

Thank you!

○ — ○ — ○

References

- "Generalized Entropy Power Inequalities and Monotonicity Properties of Information". *IEEE Transactions on Information Theory*, Vol. 53, no. 7, pp. 2317-2329, 2007. [with A. R. Barron]

- "Compound Poisson approximation via information functionals". *Electronic Journal of Probability*, 15, paper no. 42, pp. 1344-1368, 2010. [With A. Barbour, O. Johnson and I. Kontoyiannis]

- "Log-concavity, ultra-log-concavity, and a maximum entropy property of discrete compound Poisson measures". JCDM 2009 special issue edited by D. J. Kleitman, A. Shastri, V. T. Sós, *Discrete Applied Mathematics*, 2011. [With O. Johnson and I. Kontoyiannis]

Bonus References

These topics were not covered in the talk, but are related in some way.

- "Minimax risks for distributed estimation of the background in a field of noise sources". *Proceedings of the 2nd International Workshop on Information Theory for Sensor Networks* (WITS '08), Santorini Island, Greece, 2008. [With A. R. Barron, A. M. Kagan and T. Yu]

- "Dimensional behaviour of entropy and information". *Comptes Rendus de l'Académies des Sciences Paris*, Série I Mathematique, 349, pp. 201-204, 2011. [With S. Bobkov]

- "Reverse Brunn-Minkowski and reverse entropy power inequalities for convex measures". *Journal of Functional Analysis*, Vol. 262, no. 7, pp. 3309-3339, 2012. [With S. Bobkov]