The Interplay of Information Theory, Probability, and Statistics

Andrew Barron

Yale University, Department of Statistics

Presentation at Purdue University, February 26, 2007

Outline

- Information Theory Quantities and Tools *

  Entropy, relative entropy

  Shannon and Fisher information

  Information capacity

- Interplay with Statistics **

  Information capacity determines fundamental rates

  for parameter estimation and function estimation

- Interplay with Probability Theory

  Central limit theorem ***

  Large deviation probability exponents ****

  for Markov chain Monte Carlo and optimization

* Cover & Thomas, Elements of Information Theory, 1990
** Hengartner & Barron 1998 Ann.Stat; Yang & Barron 1999 Ann.Stat.
*** Barron 1986 Ann.Prob; Johnson & B. 2004 Ann.Prob; Madiman & B. 2006 ISIT
**** Csiszar 1984 Ann.Prob.

Outline for Information and Probability

- Central Limit Theorem

  If $X_1, X_2, \ldots, X_n$ are i.i.d. with mean zero and variance 1 and f_n is the density function of $(X_1 + X_2 + \ldots + X_n)/\sqrt{n}$ and ϕ is the standard normal density, then

$$D(f_n || \phi) \searrow 0$$

  if and only if this entropy distance is ever finite

- Large Deviations and Markov Chains

  If $\{X_t\}$ is i.i.d. or reversible Markov and f is bounded then there is an exponent D_ϵ characterized as a relative entropy with which

$$P\{\frac{1}{n}\sum_{t=1}^{n} f(X_t) \geq E[f] + \epsilon\} \leq e^{-nD_\epsilon}$$

  Markov chains based on local moves permit a differential equation which when solved determines the exponent D_ϵ
  Should permit determination of which chains provide accurate Monte Carlo estimates.

Entropy

- For a random variable Y or sequence $\underline{Y} = (Y_1, Y_2, \ldots, Y_N)$ with probability mass or density function $p(\underline{y})$, the Shannon entropy is

$$H(\underline{Y}) = E \log \frac{1}{p(\underline{Y})}$$

- It is the shortest expected codelength for $\underline{Y}$

- It is the exponent of the size of the smallest set that has most of the probability

Relative Entropy

- For distributions P_Y, Q_Y the relative entropy or information divergence is
$$D(P_Y||Q_Y) = E_P\left[\log \frac{p(Y)}{q(Y)}\right]$$
- It is non-negative: $D(P||Q) \geq 0$ with equality iff $P = Q$
- It is the redundancy, the expected excess of the codelength $\log 1/q(Y)$ beyond the optimal $\log 1/p(Y)$ when $Y \sim P$
- It is the drop in wealth exponent when gambling according to Q on outcomes distributed according to P
- It is the exponent of the smallest Q measure set that has most of the P probability (the exponent of probability of error of the best test): Chernoff
- It is a standard measure of statistical loss for function estimation with normal errors and other statistical models (Kullback, Stein)

$$D(\theta^*||\theta) = D(P_{Y|\theta^*}||P_{Y|\theta})$$

Statistics Basics

- Data: $\underline{Y} = (Y_1, Y_2, \ldots, Y_n)$
- Likelihood: $p(\underline{Y}|\theta) = p(Y_1|\theta) \cdot p(Y_2|\theta) \cdots p(Y_n|\theta)$
- Maximum Likelihood Estimator (MLE):

$$\hat{\theta} = \arg\max_{\theta} p(\underline{Y}|\theta)$$

- Same as $\arg\min_\theta \; \log \frac{1}{p(\underline{Y}|\theta)}$
- MLE Consistency Wald 1948

$$\hat{\theta} = \arg\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} \log \frac{p(Y_i|\theta^*)}{p(Y_i|\theta)} = \arg\min_{\theta} \hat{D}_n(\theta^*||\theta)$$

Now

$$\hat{D}_n(\theta^*||\theta) \longrightarrow D(\theta^*||\theta) \quad \text{as} \quad n \longrightarrow \infty$$

and

$$D(\theta^*||\hat{\theta}_n) \longrightarrow 0$$

- Efficiency in smooth families: $\hat{\theta}_n$ is asymptotically Normal$(\theta, (nI(\theta))^{-1})$
- Fisher information: $I(\theta) = E[\nabla \log p(\underline{Y}|\theta) \; \nabla^T \log p(\underline{Y}|\theta)]$

Statistics Basics

- Data: $\underline{Y} = Y^n = (Y_1, Y_2, \ldots, Y_n)$
- Likelihood: $p(\underline{Y}|\theta), \quad \theta \in \Theta$
- Prior: $p(\theta) = w(\theta)$
- Marginal: $p(\underline{Y}) = \int p(\underline{Y}|\theta) w(\theta) d\theta$ Bayes mixture
- Posterior: $p(\theta|\underline{Y}) = w(\theta) p(\underline{Y}|\theta) / p(\underline{Y})$
- Parameter loss function: $\ell(\theta, \hat{\theta})$, for instance squared error $(\theta - \hat{\theta})^2$
- Bayes parameter estimator: $\hat{\theta}$ achieves $\min_{\hat{\theta}} E[\ell(\theta, \hat{\theta})|\underline{Y}]$

$$\hat{\theta} = E[\theta|\underline{Y}] = \int \theta p(\theta|\underline{Y}) d\theta$$

- Density loss function $\ell(P, Q)$, for instance $D(P, Q)$
- Bayes density estimator: $\hat{p}(y) = p(y|\underline{Y})$ achives $\min_Q E[\ell(P, Q)|\underline{Y}]$

$$\hat{p}(y) = \int p(y|\theta) p(\theta|Y^n) d\theta$$

- Predictive coherence: Bayes estimator is the predictive density $p(Y_{n+1}|Y^n)$ evaluated at $Y_{n+1} = y$
- Other loss functions do not share this property

Chain Rules for Entropy and Relative Entropy

- For joint densities

$$p(Y_1, Y_2, \ldots, Y_N) \ = \ p(Y_1)\, p(Y_2|Y_1) \cdots p(Y_N|Y_{N-1}, \ldots, Y_1)$$

- Taking the expectation this is

$$H(Y_1, Y_2, \ldots Y_N) = H(Y_1) + H(Y_2|Y_1) + \ldots + H(Y_N|Y_{N-1}, \ldots, Y_1)$$

- The joint entropy grows like $\mathcal{H}N$ for stationary processes
- For the relative entropy between distributions for a string $\underline{Y} = Y^N = (Y_1, \ldots, Y_N)$ we have the chain rule

$$D(P_{\underline{Y}}||Q_{\underline{Y}}) = \sum_n E_P D(P_{Y_{n+1}|Y^n}||Q_{Y_{n+1}|Y^n})$$

- Thus the total divergence is a sum of contributions in which the predictive distributions $Q_{Y_{n+1}|Y^n}$ based on the previous n data points is measured for their quality of fit to $P_{Y_{n+1}|Y^n}$ for each n less than N
- With good predictive distributions we can arrange $D(P_{Y^N}||Q_{Y^N})$ to grow at rates slower than N simultaneously for various P

Tying data compression to statistical learning

- Various plug-in $\hat{p}_n(y) = p(y|\hat{\theta}_n)$ and Bayes predictive estimators

$$\hat{p}_n(y) = q(y|Y^n) = \int p(y|\theta)p(\theta|Y^n)d\theta$$

achieve individual risk

$$D(P_{Y|\theta}||\hat{P}_n) \sim \frac{c}{n}$$

ideally with asymptotic constant $c = d/2$ where d is the parameter dimension (more on that ideal constant later)

- Successively evaluating the predictive densities $q(Y_{n+1}|Y^n)$ these piece fit together to give a joint density $q(Y^N)$ with total divergence

$$D(P_{Y^N|\theta}||Q_{Y^N}) \sim c \log N$$

- Conversely from any coding distribution Q_{Y^N} with good redundancy $D(P_{Y^N}|\theta)||Q_{Y^N})$ a succession of predictive estimators can be obtained

- Similar conclusions hold for nonparametric function estimation problems

Local Information, Estimation, and Efficiency

- The Fisher information $I(\theta) = I_{Fisher}(\theta)$ arises naturally in local analysis of Shannon information and related statistics problems.

- In smooth families the relative entropy loss is locally a squared error

$$D(\theta||\hat{\theta}) \sim \frac{1}{2}(\theta - \hat{\theta})^T I(\theta)(\theta - \hat{\theta})$$

- Efficient estimates have asymptotic covariance not more than $I(\theta)^{-1}$

- If smaller than that at some θ the estimator is said to be superefficient

- The expectation of the asymptotic distribution for the right side above is

$$\frac{d}{2n}$$

- The set of parameter values with smaller asymptotic covariance is negligible, in the sense that it has zero measure

Efficiency of Estimation via Info Theory Analysis

- LeCam 1950s: Efficiency of Bayes and maximum likelihood estimators. Negligibility of superefficiency for bounded loss and any efficient estimator
- Hengartner and B. 1998: Negligibility of superefficiency for any parameter estimator using $ED(\theta||\hat{\theta})$ and any density estimator using $ED(P||\hat{P}_n)$
- The set of parameter values for which $nED(P_{Y|\theta}||\hat{P}_n)$ has limit not smaller than $d/2$ includes all but a negligible set of θ

- The proof does not require a Fisher information, yet correspond to the classical conclusion when there is such
- The efficient level is from coarse covering properties of Euclidean space
- The core of the proof is the chain rule plus a result of Rissanen
- Rissanen 1986: no choice of joint distribution achieves $D(P_{Y^N|\theta}||Q_{Y_N})$ better than $(d/2)\log N$ except in a negligible set of θ
- The proof works also for nonparametric problems
- Negligibility of superefficiency determined by sparsity of its cover

Mutual Information and Information Capacity

- We shall need two additional quantities in our discussion of information theory and statistics. These are:

  the Shannon mutual information $\quad I$

  and the information capacity $\quad C$

Shannon Mutual Information

- For a family of distributions $P_{Y|U}$ of a random variable Y given an input U distributed according to P_U, the Shannon mutual information is

$$I(Y;U) = D(P_{U,Y}||P_U P_Y) = E_U D(P_{Y|U}||P_Y)$$

- In communications, it is the rate, the exponent of the number of input strings $\underline{U}$ that can be reliably communicated across a channel $P_{\underline{Y}|\underline{U}}$

- It is the error probability exponent with which a random $\underline{U}$ erroneously passes the test of being jointly distributed with a received string $\underline{Y}$

- In data compression, $I(\underline{Y};\theta)$ is the Bayes average redundancy of the code based on the mixture P_Y when $\theta = U$ is unknown

- In a game with relative entropy loss, it is the Bayes optimal value corresponding to the the Bayes mixture P_Y being the choice of Q_Y achieving

$$I(Y;\theta) = \min_{Q_Y} E_\theta D(P_{Y|\theta}||Q_Y)$$

- Thus it is the average divergence from the centroid P_Y

Information Capacity

- For a family of distributions $P_{Y|U}$ the Shannon information capacity is

$$C \;=\; \max_{P_U} I(Y;U)$$

- It is the communications capacity, the maximum rate that can be reliably communicated across the channel

- In the relative entropy game it is the maximin value

$$C = \max_{P_\theta} \min_{Q_Y} E_{P_\theta} D(P_{Y|\theta} || Q_Y)$$

- Accordingly it is also the minimax value

$$C = \min_{Q_Y} \max_{\theta} D(P_{Y|\theta} || Q_Y)$$

- Also known as the information radius of the family $P_{Y|\theta}$

- In data compression, this means that $C = \max_{P_\theta} I(\underline{Y};\theta)$ is also the minimax redundancy for the family $P_{Y|\theta}$ (Gallager; Ryabko; Davisson)

- In recent years the information capacity has been shown to also answer questions in statistics as we shall discuss

Information Asymptotics for Bayes Procedures

- The Bayes mixture density $p(\underline{Y}) = \int p(\underline{Y}|\theta)w(\theta)d\theta$ satisfies in smooth parametric families the Laplace approximation

$$\log \frac{1}{p(\underline{Y})} \;=\; \log \frac{1}{p(\underline{Y}|\hat{\theta})} \;+\; \frac{d}{2}\log\frac{N}{2\pi} \;+\; \log\frac{|I(\hat{\theta})|^{1/2}}{w(\hat{\theta})} \;+\; o_p(1)$$

- Underlies Bayes and description length criteria for model selection

- Clarke & B. 1990 show for θ in the interior of the parameter space that

$$D(P_{\underline{Y}|\theta}||P_{\underline{Y}}) = \frac{d}{2}\log\frac{N}{2\pi e} \;+\; \log\frac{|I(\theta)|^{1/2}}{w(\theta)} \;+\; o(1)$$

- Likewise, via Clarke & B. 1994, the average with respect to the prior has

$$I_{Shannon}(\underline{Y};\theta) \;=\; \frac{d}{2}\log\frac{N}{2\pi e} \;+\; \int w(\theta)\log\frac{|I_{Fisher}(\theta)|^{1/2}}{w(\theta)} \;+\; o(1)$$

- Provides capacity of multi-antenna systems (d input, N output) as well as minimax asymptotics for data compression and statistical estimation

Minimax Asymptotics in Parametric Families

- We identify the form of prior $w(\theta)$ that equalizes the risk $D(P_{\underline{Y}|\theta}||P_{\underline{Y}})$ and maximizes the Bayes risk $I(\underline{Y};\theta)$. This prior should be proportional to $|I_{Fisher}(\theta)|^{1/2}$, known in statistics and physics as Jeffreys' prior.

- This prior gives equal weight to small equal-radius relative entropy balls

- Clark and B. 1994: on any compact K in the interior of Θ, the information capacity C_N (and minimax redundancy) satisfies

$$C_N \;=\; \frac{d}{2}\log\frac{N}{2\pi e} \;+\; \log\int_K |I_{Fisher}(\theta)|^{1/2}d\theta \;+\; o(1)$$

- Asymptotically maximin priors and corresponding asymptotically minimax procedure are obtained by using boundary modifications of Jeffreys' prior

- Xie and B. 1998, 1999: refinement applicable to the whole probability simplex in the case of finite alphabet distributions

- Liang and B. 2004 show exact minimaxity for finite sample size in families with group structure such as location & scale problems, conditional on initial observations to make the minimax answer finite

Minimax Asymptotics for Function Estimation

- Let $\mathcal{F}$ be a function class and let data $\underline{Y}$ with sample size n come independently from a distribution $P_{Y|f}$ with $f \in \mathcal{F}$
- Thus f can be a density function, a regression function, a discriminant function or an intensity function depending in the nature of the model
- Let $\mathcal{F}$ be endowed with a metric $d(f,g)$ such as L_2 or Hellinger distance
- The Kolmogorov metric entropy or $\epsilon-$entropy, denoted $H(\epsilon)$ is the log of the size of the smallest cover of $\mathcal{F}$ by finitely many functions, such that every f in $\mathcal{F}$ is within ϵ of one of the functions in the cover
- The metric entropy rate is obtained by matching

$$\frac{H(\epsilon_n)}{n} = \epsilon_n^2$$

- The minimax rate of function estimation is

$$r_n = \min_{\hat{f}_n} \max_{f \in \mathcal{F}} Ed^2(f, \hat{f}_n)$$

- The information capacity rate of $\{P_{\underline{Y}|f},\ f \in \mathcal{F}\}$ is

$$C_n = \frac{1}{n} \sup_{P_f} I(\underline{Y}; f)$$

Minimax Asymptotics for Function Estimation

- Suppose $D(P_{Y|f}||P_{Y|g})$ is equivalent to the squared metric $d^2(f,g)$ in $\mathcal{F}$ in that their ratio is bounded above and below by positive constants
- Theorem: (Yang & B. 1998) The minimax rate of function estimation, the metric entropy rate, and the information capacity rate are the same

$$r_n \;\sim\; C_n \;\sim\; \epsilon_n^2$$

- The proof in one direction uses the chain rule and bounds the cumulative risk of a Bayes procedure using the uniform prior on an optimal cover
- The other direction is based on use of Fano's inequality
- Typical function classes constrain the smoothness s of the function, e.g. s may be number of bounded derivatives, and have

$$H(\epsilon) \;\sim\; (1/\epsilon)^{1/s}$$

- Accordingly $\qquad r_n \sim \epsilon_n^2 \sim n^{-2s/(2s+1)}$
- Analogous results in Haussler and Opper 1997.
- Precursors were in work by Pinsker, by Hasminskii, and by Birge

Outline for Information and Probability

- Central Limit Theorem

  If $X_1, X_2, \ldots, X_n$ are i.i.d. with mean zero and variance 1 and f_n is the density function of $(X_1 + X_2 + \ldots + X_n)/\sqrt{n}$ and ϕ is the standard normal density, then

  $$D(f_n || \phi) \searrow 0$$

  if and only if this entropy distance is ever finite

- Large Deviations and Markov Chains

  If $\{X_t\}$ is i.i.d. or reversible Markov and f is bounded then there is an exponent D_ϵ characterized as a relative entropy with which

  $$P\{\frac{1}{n}\sum_{t=1}^{n} f(X_t) \geq E[f] + \epsilon\} \leq e^{-nD_\epsilon}$$

  Markov chains based on local moves permit a differential equation which when solved provides approximately the exponent D_ϵ. Should permit determination of which chains provide accurate Monte Carlo estimates.

Outline for Information and CLT

- Entropy and the Central Limit Problem
- Entropy Power Inequality (EPI)
- Monotonicity of Entropy and new subset sum EPI
- Variance Drop Lemma
- Projection and Fisher Information
- Rates of Convergence in the CLT

Entropy Basics

- For a mean zero random variable X with density $f(x)$ and finite variance $\sigma^2 = 1$,

  the differential entropy is $H(X) = E[\log \frac{1}{f(X)}]$

  the entropy power of X is $e^{2H(X)}/2\pi e$

- For a Normal$(0, \sigma^2)$ random variable Z, with density function ϕ,

  the differential entropy is $H(Z) = (1/2)\log(2\pi e \sigma^2)$

  the entropy power of Z is σ^2

- The relative entropy is $D(f||\phi) = \int f(x) \log \frac{f(x)}{\phi(x)} dx$

  it is non-negative: $D(f||\phi) \geq 0$ with equality iff $f = \phi$

  it is larger than $(1/2)||f - \phi||_1^2$

Maximum entropy property

Boltzmann, Jaynes, Shannon

Let Z be a normal random variable with the same mean and variance as a random variable X, then $H(X) \leq H(Z)$ with equality iff X is normal

The relative entropy quantifies the entropy gap

$$H(Z) - H(X) = D(f||\phi)$$

Maximum entropy property

Boltzmann, Jaynes, Shannon

Let Z be a normal random variable with the same mean and variance as a random variable X, then $H(X) \leq H(Z)$ with equality iff X is normal.

The relative entropy quantifies the entropy gap. Indeed, this is Kullback's proof of the maximum entropy property

$$\begin{aligned} H(Z) - H(X) &= \int \phi(x) \log \frac{1}{\phi(x)} dx - \int f(x) \log \frac{1}{f(x)} dx \\ &= \int f(x) \log \frac{1}{\phi(x)} dx - \int f(x) \log \frac{1}{f(x)} dx \\ &= \int f(x) \log \frac{f(x)}{\phi(x)} dx \\ &= D(f||\phi) \\ &\geq 0 \end{aligned}$$

Here $\log \frac{1}{\phi(x)} = \frac{x^2}{2\sigma^2} \log e + \frac{1}{2} \log 2\pi\sigma^2$ is quadratic in x, so both f and ϕ give it the same expectation, which is $\frac{1}{2} \log 2\pi e \sigma^2$.

Fisher Information Basics

- For a mean zero random variable X with differentiable density $f(x)$ and finite variance $\sigma^2 = 1$,

  the score function is $score(X) = \frac{d}{dx} \log f(x)$

  the Fisher information is $I(X) = E[score^2(X)]$.

- For a Normal$(0, \sigma^2)$ random variable Z, with density function ϕ,

  the score function is linear $score(Z) = -Z/\sigma^2$

  the Fisher information is $I(Z) = 1/\sigma^2$

- The relative Fisher information is $J(f||\phi) = \int f(x) \left(\frac{d}{dx} \log \frac{f(x)}{\phi(x)} \right)^2 dx$

  it is non-negative

  it is larger than $D(f||\phi)$

- Minimum Fisher info property (Cramer-Rao ineq): $I(X) \geq 1/\sigma^2$

  equality iff Normal

- The information gap satisfies: $I(X) - I(Z) = J(f||\phi)$

The Central Limit Problem

For independent identically distributed random variables $X_1, X_2, \ldots, X_n$, with $E[X] = 0$ and $VAR[X] = \sigma^2 = 1$, consider the standardized sum

$$\frac{X_1 + X_2 + \ldots + X_n}{\sqrt{n}}.$$

Let its density function be f_n and its distribution function F_n.

Let the standard normal density be ϕ and its distribution function Φ.

Natural questions:

- In what sense do we have convergence to the normal?
- Do we come closer to the normal with each step?
- Can we give clean bounds on the "distance" from the normal and a corresponding rate of convergence?

Convergence

- **In distribution:** $F_n(x) \to \Phi(x)$

  Classical via Fourier methods or expansions of expectations of smooth functions.

  Linnick 59, Brown 82 via info measures applied to smoothed distributions.

- **In density:** $f_n(x) \to \phi(x)$

  Prohorov 52 showed $||f_n - \phi||_1 \to 0$ iff f_n exists eventually.

  Kolmogorov & Gnedenko 54 $||f_n - \phi||_\infty \to 0$ iff f_n bounded eventually.

- **In Shannon Information:** $H(\frac{1}{\sqrt{n}} \sum_{i=1}^n X_i) \to H(Z)$

  Barron 86 shows $D(f_n||\phi) \to 0$ iff it is eventually finite.

- **In Fisher Information:** $I(\frac{1}{\sqrt{n}} \sum_{i=1}^n X_i) \to 1/\sigma^2$

  Johnson & Barron 04 shows $J(f_n||\phi) \to 0$ iff it is eventually finite.

Original Entropy Power Inequality

Shannon 48, Stam 59: For independent random variables with densities,

$$e^{2H(X_1+X_2)} \geq e^{2H(X_1)} + e^{2H(X_2)}$$

where equality holds if and only if the X_i are normal.

Also

$$e^{2H(X_1+\ldots+X_n)} \geq \sum_{j=1}^{n} e^{2H(X_j)}$$

Original Entropy Power Inequality

Shannon 48, Stam 59: For independent random variables with densities,

$$e^{2H(X_1+X_2)} \geq e^{2H(X_1)} + e^{2H(X_2)}$$

where equality holds if and only if the X_i are normal.

Central Limit Theorem Implication

For X_i i.i.d., let $H_n = H\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i\right)$

- nH_n is superadditive

$$H_{n_1+n_2} \geq \frac{n_1}{n_1+n_2} H_{n_1} + \frac{n_2}{n_1+n_2} H_{n_2}$$

- monotonicity for doubling sample size

$$H_{2n} \geq H_n$$

- The superadditivity of nH_n and the monotonicity for the powers of two subsequence are key in the proof of entropy convergence [Barron '86]

Leave-one-out Entropy Power Inequality

Artstein, Ball, Barthe and Naor 2004 (ABBN): For independent X_i

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

Remarks

- This strengthens the original EPI of Shannon and Stam.
- ABBN's proof is elaborate.
- Our proof (Madiman & Barron 2006) uses familiar and simple tools and proves a more general result, that we present.
- The leave-one-out EPI implies in the iid case that entropy is increasing:

$$H_n \geq H_{n-1}$$

- A related proof of monotonicity is developed contemporaneously in Tulino & Verdú 2006.
- Combining with Barron 1986 the monotonicity implies

$$H_n \nearrow H(\text{Normal}) \quad \text{and} \quad D_n = \int f_n \log \frac{f_n}{\phi} \searrow 0$$

New Entropy Power Inequality

Subset-sum EPI (Madiman and Barron)

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

$$e^{2H(X_1+\ldots+X_n)} \geq \frac{1}{r(\mathcal{S})} \sum_{s \in \mathcal{S}} e^{2H(\text{sum}_s)}$$

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

$r(\mathcal{S})$ is the *prevalence*, the maximum number of subsets in $\mathcal{S}$ in which any index i can appear

Examples

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

New Entropy Power Inequality

Subset-sum EPI

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

$$e^{2H(X_1+\ldots+X_n)} \geq \frac{1}{r(\mathcal{S})} \sum_{s \in \mathcal{S}} e^{2H(\mathsf{sum}_s)}$$

Discriminating and balanced collections $\mathcal{S}$

- *Discriminating* if for any i, j, there is a set in $\mathcal{S}$ containing i but not j
- *Balanced* if each index i appears in the same number $r(\mathcal{S})$ of sets in $\mathcal{S}$

Equality in the Subset-sum EPI

For discriminating and balanced $\mathcal{S}$, equality holds in the subset-sum EPI
if and only if the X_i are normal

In this case, it becomes $\displaystyle\sum_{i=1}^{n} a_i = \frac{1}{r(\mathcal{S})} \sum_{s \in \mathcal{S}} \sum_{i \in s} a_i$ with $a_i = \mathsf{Var}(X_i)$

New Entropy Power Inequality

Subset-sum EPI

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

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

CLT Implication

Let X_i be independent, but not necessarily identically distributed.

The entropy of variance-standardized sums increases "on average":

$$H\left(\frac{\mathrm{sum}_{\mathrm{total}}}{\sigma_{\mathrm{total}}}\right) \geq \sum_{s \in \mathcal{S}} \lambda_s \, H\left(\frac{\mathrm{sum}_s}{\sigma_s}\right)$$

where

- $\sigma^2_{\mathrm{total}}$ is the variance of $\mathrm{sum}_{\mathrm{total}} = \sum_{i=1}^n X_i$ and σ_s^2 is the variance of $\mathrm{sum}_s = \sum_{j \in s} X_j$
- The weights $\lambda_s = \frac{\sigma_s^2}{r(\mathcal{S})\sigma^2_{\mathrm{total}}}$ are proportional to σ_s^2
- The weights add to 1 for balanced collections $\mathcal{S}$

New Fisher Information Inequality

For independent X_1, $X_2, \ldots, X_n$ with differentiable densities,

$$\frac{1}{I(\text{sum}_{\text{total}})} \geq \frac{1}{r(\mathcal{S})} \sum_{s \in \mathcal{S}} \frac{1}{I(\text{sum}_s)}$$

Remarks

- This extends Fisher information inequalities of Stam and ABBN
- Recall from Stam '59

$$\frac{1}{I(X_1 + \ldots + X_n)} \geq \frac{1}{I(X_1)} + \ldots + \frac{1}{I(X_n)}$$

- For discriminating and balanced $\mathcal{S}$, equality holds iff the X_i are normal

New Fisher Information Inequality

For independent $X_1, X_2, \ldots, X_n$ with differentiable densities,

$$\frac{1}{I(\text{sum}_{\text{total}})} \geq \frac{1}{r(\mathcal{S})} \sum_{s \in \mathcal{S}} \frac{1}{I(\text{sum}_s)}$$

CLT Implication

- For i.i.d. X_i, let $I_n = I\left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i\right)$

  The Fisher information I_n is a decreasing sequence:

  $$I_n \leq I_{n-1} \qquad \text{[ABBN '04]}$$

  Combining with Johnson and Barron '04 implies $I_n \searrow I(\text{Normal})$ and

  $$J(f_n || \phi) \searrow 0$$

- For i.n.i.d. X_i, the Fisher info. of standardized sums decreases on average

  $$I\left(\frac{\text{sum}_{\text{total}}}{\sigma_{\text{total}}}\right) \leq \sum_{s \in \mathcal{S}} \lambda_s \, I\left(\frac{\text{sum}_s}{\sigma_s}\right)$$

The Link between H and I

Definitions

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

Relationship

For a standard normal Z independent of X,

- Differential version:

$$\frac{d}{dt} H(X + \sqrt{t} Z) = \frac{1}{2} I(X + \sqrt{t} Z)$$ [de Bruijn, see Stam '59]

- Integrated version:

$$H(X) = \frac{1}{2}\log(2\pi e) - \frac{1}{2}\int_0^\infty \left[I(X + \sqrt{t} Z) - \frac{1}{1+t} \right] dt$$ [Barron '86]

The Projection Tool

For each subset s,

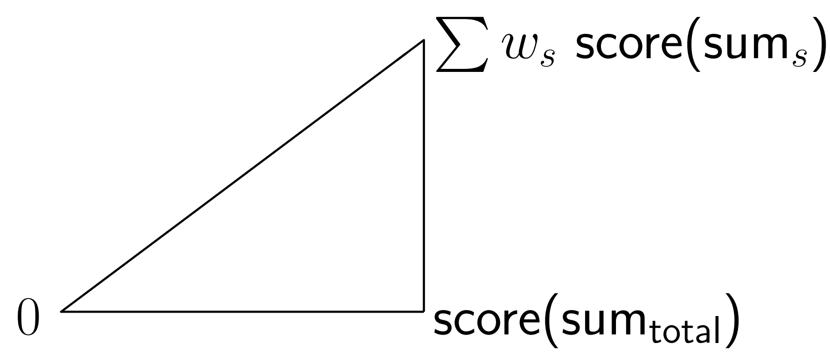
$$\text{score}(\text{sum}_{\text{total}}) = E\left[\text{score}(\text{sum}_s) \,\middle|\, \text{sum}_{\text{total}}\right]$$

Hence, for weights w_s that sum to 1,

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

Pythagorean inequality

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



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

The Heart of the Matter

Recall the Pythagorean inequality

$$I(\mathsf{sum}_{\mathsf{total}}) \leq E\left[\sum_{s\in\mathcal{S}} w_s \ \mathsf{score}(\mathsf{sum}_s)\right]^2$$

and apply the variance drop lemma to get

$$I(\mathsf{sum}_{\mathsf{total}}) \leq r(\mathcal{S}) \sum_{s\in\mathcal{S}} w_s^2 I(\mathsf{sum}_s)$$

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

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

have the variance bound

$$Eg^2 \leq r(\mathcal{S}) \sum_{s \in \mathcal{S}} Eg_s^2(\underline{X}_s)$$

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

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

have the variance bound

$$Eg^2 \le r(\mathcal{S}) \sum_{s \in \mathcal{S}} Eg_s^2(\underline{X}_s)$$

Remarks

- Note that $r(\mathcal{S}) \le |\mathcal{S}|$, hence the "variance drop"
- Examples:

  $\mathcal{S}$=singletons has $r = 1$: additivity of variance with independent summands

  $\mathcal{S}$=leave-one-out sets has $r = n-1$ as in the study of the jackknife and U-statistics
- Proof is based on ANOVA decomposition [Hoeffding '48, Efron and Stein '81]
- Introduced in leave-one-out case to info. inequality analysis by ABBN '04

Optimized Form for I

We have, for all weights w_s that sum to 1,

$$I(\mathsf{sum}_{\mathsf{total}}) \leq r(\mathcal{S}) \sum_{s \in \mathcal{S}} w_s^2 I(\mathsf{sum}_s)$$

Optimizing over w yields the new Fisher information inequality

$$\frac{1}{I(\mathsf{sum}_{\mathsf{total}})} \geq \frac{1}{r(\mathcal{S})} \sum_{s \in \mathcal{S}} \frac{1}{I(\mathsf{sum}_s)}$$

Optimized Form for H

We have (again)

$$I(\mathsf{sum}_{\mathsf{total}}) \le r(\mathcal{S}) \sum_{s\in\mathcal{S}} w_s^2 I(\mathsf{sum}_s)$$

Equivalently,

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

Adding independent normals and integrating,

$$H(\mathsf{sum}_{\mathsf{total}}) \ge \sum_{s\in\mathcal{S}} w_s H\left(\frac{\mathsf{sum}_s}{\sqrt{r(\mathcal{S})w_s}}\right)$$

Optimizing over w yields the new Entropy Power Inequality

$$e^{2H(\mathsf{sum}_{\mathsf{total}})} \ge \frac{1}{r(\mathcal{S})} \sum_{s\in\mathcal{S}} e^{2H(\mathsf{sum}_s)}$$

Fisher information and M.M.S.E. Estimation

Model: $Y = X + Z$

where $Z \sim N(0,1)$ and X is to be estimated

- Optimal estimate: $\hat{X} = E[X|Y]$

  Fact: score$(Y) = \hat{X} - Y$

  Note: $X - \hat{X}$ and $\hat{X} - Y$ are orthogonal, and sum to $-Z$

  Hence: $I(Y) = E(\hat{X} - Y)^2 = 1 - E(X - \hat{X})^2$

  $= 1 -$ Minimal M.S.E.

  From L.D. Brown '70's [c.f. the text of Lehmann and Casella '98]

- Thus derivative of entropy can be expressed equivalently in terms of either $I(Y)$ or minimal M.S.E.

- Guo, Shamai and Verdú, 2005 use the minimal M.S.E. interpretation to give a related proof of the EPI and Tulino and Verdú 2006 use this M.S.E. interpretation to give a related proof of monotonicity in the CLT

Recap: Subset-sum EPI

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

$$e^{2H(\mathsf{sum}_{\mathsf{total}})} \geq \frac{1}{r(\mathcal{S})} \sum_{s \in \mathcal{S}} e^{2H(\mathsf{sum}_s)}$$

- Generalizes original EPI and ABBN's EPI
- Simple proof using familiar tools
- Equality holds for normal random variables

Comment on CLT rate bounds

For iid X_i let

$$J_n = J(f_n||\phi)$$

and

$$D_n = D(f_n||\phi)$$

Suppose the distribution of the X_i has a finite Poincaré constant R.

Using the pythagorean identity for score projection, Johnson & Barron '04 show:

$$J_n \le \frac{2R}{n} J_1$$

$$D_n \le \frac{2R}{n} D_1$$

- Implies a $1/\sqrt{n}$ rate of convergence in distribution, known to hold for random variables with non-zero finite third moment.
- Our finite Poincaré assumption implies finite moments of all orders.
- Do similar bounds on information distance hold assuming only finite initial information distance and finite third moment?

Summary

Two ingredients

- score of sum $=$ projection of scores of subset-sums
- variance drop lemma

yield the conclusions

- existing Fisher information and entropy power inequalities
- new such inequalities for arbitrary collections of subset-sums
- monotonicity of I and H in central limit theorems

refinements using the pythagorean identity for the score projection yield

- convergence in information to the Normal
- order $1/n$ bounds on information distance from the Normal

○ — ○ — ○