Celebrations for
three influential scholars in Information Theory and Statistics:

- Tom Cover: On the occasion of his 70th birthday
  Coverfest.stanford.edu
  Elements of Information Theory Workshop
  Stanford University, May 16, 2008: TODAY!
- Imre Csiszár: On the occasion of his 70th birthday
  www.renyi.hu/∼infocom
  Information and Communication Conference
  Renyi Institute, Budapest, August 25-28, 2008
- Jorma Rissanen: On the occasion of his 75th birthday
  Festschrift at www.cs.tut.fi/∼tabus/
  presented at the *IEEE Information Theory Workshop*
  Porto, Portugal, May 8, 2008

Principles of Information Theory in Probability and Statistics

Andrew Barron
Department of Statistics
Yale University

May 16, 2008
Elements of Information Theory Workshop

On the Occasion of the Festival for Tom Cover

Outline of Principles

Outline of Principles

Monotonicity of Information Divergence

- Chain Rule

$$D(P_{X,X'}\|P^*_{X,X'}) = D(P_X\|P^*_X) + E\,D(P_{X'|X}\|P^*_{X'|X})$$

$$= D(P_{X'}\|P^*_{X'}) + E\,D(P_{X|X'}\|P^*_{X|X'})$$

- Markov Chains

$$D(P_{X_n}\|P^*) \le D(P_{X_m}\|P^*) \quad \text{for } n > m$$

and $\log p_n(X_n)/p^*(X_n)$ is a Cauchy sequence in $L_1(P)$

Monotonicity of Information Divergence

- Nonnegative Martingales ρ_n equal the density of a measure Q_n and can be examined in the same way by the chain rule for $n > m$

$$D(Q_n \| P) = D(Q_m \| P) + \int \left(\rho_n \log \frac{\rho_n}{\rho_m} \right) dP$$

- Thus $D(Q_n \| P)$ is an increasing sequence. When it is bounded ρ_n is a Cauchy sequences in $L_1(P)$ with limit ρ defining a measure Q, also, $\log \rho_n$ is a Cauchy sequence in $L_1(Q)$ and

$$D(Q_n \| P) \nearrow D(Q \| P)$$

Monotonicity of Information Divergence: CLT

- Central Limit Theorem Setting:

  $\{X_i\}$ i.i.d. mean zero, finite variance

  $P_n = P_{Y_n}$ is distribution of $Y_n = \frac{X_1 + X_2 + \ldots X_n}{\sqrt{n}}$

  P^* is the corresponding normal distribution

- Chain Rule: Action is mysterious in this case

$$D(P_{Y_m,Y_n} \| P^*_{Y_m,Y_n}) = D(P_{Y_m} \| P^*) + ED(P_{Y_n|Y_m} \| P^*_{Y_n|Y_m})$$

$$= D(P_{Y_n} \| P^*) + ED(P_{Y_m|Y_n} \| P^*_{Y_m|Y_n})$$

Monotonicity of Information Divergence: CLT

- Entropy Power Inequality

$$e^{2H(X+X')} \geq e^{2H(X)} + e^{2H(X')}$$

yields

$$D(P_{2n} \| P^*) \leq D(P_n \| P^*)$$

- Information Theoretic proof of CLT (B. 1986):

$$D(P_n \| P^*) \to 0 \text{ iff finite}$$

Monotonicity of Information Divergence: CLT

- Entropy Power Inequality

$$e^{2H(X+X')} \geq e^{2H(X)} + e^{2H(X')}$$

- Generalized Entropy Power Inequality (Madiman&B.2006)

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

- Proof: simple L_2 projection properties of entropy derivative.
- Consequence, for all $n > m$,

$$D(P_n \| P^*) \leq D(P_m \| P^*)$$

B. and Madiman 2006, Tolino and Verdú 2006
Earlier elaborate proof by Artstein, Ball, Barthe, Naor 2004.

Outline of Principles

AEP and behavior of optimal tests

- Stability of log-likelihood ratios,

$$\frac{1}{n}\log\frac{p(Y_1, Y_2, \ldots Y_n)}{q(Y_1, Y_2, \ldots, Y_n)} \to \mathcal{D}(P\|Q) \text{ with } P - prob\, 1$$

  where $\mathcal{D}(P\|Q)$ is the relative entropy (I-divergence) rate.
- Implication: Associated log-likelihood ratio test A_n has asymptotic P-power 1 (at most finitely many mistakes $P(A_n^c\ i.o.) = 0$) and has optimal Q-prob of error

$$Q(A_n) = \exp\{-n[\mathcal{D} + o(1)]\}$$

- Most general known form of the Chernoff-Stein Lemma.
- Intrinsic role for information divergence rate

$$\mathcal{D}(P\|Q) = \lim \frac{1}{n} D(P_{\underline{Y}^n}\|Q_{\underline{Y}^n})$$

Large Deviations for Empirical Prob and Conditional Limit

- P^*: Information Projection of Q onto convex C
- Pythagorean Identity (Csiszar 75, Topsoe 79): For P in C

$$D(P\|Q) \geq D(C\|Q) + D(P\|P^*)$$

where

$$D(C\|Q) = \inf_{P \in C} D(P\|Q)$$

- Empirical Distribution P_n
- If $D(\text{interior} C\|Q) = D(C\|Q)$ then

$$Q\{P_n \in C\} = \exp\{-n[D(C\|Q) + o(1)]\}$$

and the conditional distribution $P_{Y_1, Y_2, \ldots, Y_n | \{P_n \in C\}}$ converges to $P^*_{Y_1, Y_2, \ldots, Y_n}$ in the I-divergence rate sense (Csiszar 1985)

Outline of Principles

Information Capacity

- A Channel $\theta \rightarrow \underline{Y}$ is a family of probability distributions

$$\{P_{\underline{Y}|\theta} : \theta \in \Theta\}$$

- Information Capacity

$$C = \max_{P_\theta} I(\theta; \underline{Y})$$

Communications Capacity

- C_{com} = maximum rate of reliable communication, measured in message bits per use of a channel
- Shannon Channel Capacity Theorem (Shannon 1948)

$$C_{com} = C$$

Data Compression Capacity

- Minimax Redundancy

$$Red = \min_{Q_{\underline{Y}}} \max_{\theta \in \Theta} D(P_{\underline{Y}|\theta} \| Q_{\underline{Y}})$$

- Data Compression Capacity Theorem

$$Red = C$$

(Gallager, Davisson & Leon-Garcia, Ryabko)

Statistical Risk Setting

- Loss function

$$\ell(\theta, \theta')$$

- Examples:
- Kullback Loss

$$\ell(\theta, \theta') = D(P_{Y|\theta} \| P_{Y|\theta'})$$

- Squared metric loss, e.g. squared Hellinger loss:

$$\ell(\theta, \theta') = d^2(\theta, \theta')$$

Statistical Capacity

- Estimators: $\hat{\theta}_n$
- Based on sample $\underline{Y}$ of size n
- Minimax Risk (Wald):

$$r_n = \min_{\hat{\theta}_n} \max_{\theta} E\ell(\theta, \hat{\theta}_n)$$

Ingredients for determination of Statistical Capacity

- Kolmogorov Metric Entropy of $S \subset \Theta$:

$$H(\epsilon) = \max\{\log Card(\Theta_\epsilon) \ : \ d(\theta, \theta') > \epsilon \text{ for } \theta, \theta' \in \Theta_\epsilon \subset S\}$$

- Loss Assumption, for $\theta, \theta' \in S$:

$$\ell(\theta, \theta') \sim D(P_{Y|\theta} \| P_{Y|\theta'}) \sim d^2(\theta, \theta')$$

Statistical Capacity Theorem

- For infinite-dimensional Θ
- With metric entropy evaluated a critical separation ϵ_n
- Statistical Capacity Theorem

  Minimax Risk $\sim$ Info Capacity Rate $\sim$ Metric Entropy rate

$$r_n \quad \sim \quad \frac{C_n}{n} \quad \sim \quad \frac{H(\epsilon_n)}{n} \quad \sim \quad \epsilon_n^2$$

  Yang 1997, Yang and B. 1999, Haussler and Opper 1997

Outline of Principles

Shannon Codes

- Kraft-McMillan characterization:
  Uniquely decodeable codelengths

$$L(\underline{x}), \quad \underline{x} \in \underline{\mathcal{X}}, \qquad \sum_{\underline{x}} 2^{-L(\underline{x})} \leq 1$$

$$L(\underline{x}) = \log 1/p(\underline{x}) \qquad p(\underline{x}) = 2^{-L(\underline{x})}$$

- Operational meaning of probability:
  A probability distribution p is given by a choice of code

Complexity

- Kolmogorov Idealized Compression:

$$K(\underline{Y}) = \text{Length of shortest computer code for } \underline{Y} \text{ on a given universal computer}$$

- Shannon Idealized Codelength (expectation optimal):

$$\log 1/p^*(\underline{Y})$$

- But p^* is not generally known
- Hybrid: Statistical measure of Complexity of $\underline{Y}$

$$\min_p \left[\log 1/p(\underline{Y}) \; + \; L(p) \right]$$

- Valid for any $L(p)$ satisfying Kraft summability.

Complexity

- Minimum Description Length Principle (MDL)
- Special cases: Rissanen 1978,1983,...
- Two-stage codelength formulation:
  (Cover Scratch pad 1981, B. 1985, B. and Cover 1991)

$$L(\underline{Y}) = \min_p \left[\quad \log 1/p(\underline{Y}) \quad + \quad L(p) \quad \right]$$

bits for $\underline{x}$ given p $+$ bits for p

- Corresponding statistical estimator $\hat{p}$ achieves the above minimization in a family of distributions for a specified $L(p)$

MDL Analysis

- Redundancy of Two-stage Code:

$$Red = \frac{1}{n} E \left\{ \min_p \left[\log \frac{1}{p(\underline{Y})} + L(p) \right] - \log \frac{1}{p^*(\underline{Y})} \right\}$$

- bounded by Index of Resolvability:

$$Res_n(p^*) = \min_p \left\{ D(p^* \| p) + \frac{L(p)}{n} \right\}$$

- Statistical Risk Analysis in i.i.d. case:

$$E\, d^2(p^*, \hat{p}) \le \min_p \left\{ D(p^* \| p) + \frac{L(p)}{n} \right\}$$

- B. 1985, B.&Cover 1991, B., Rissanen, Yu 1998, Li 1999, Grunwald 2007

MDL Analysis

- Statistical Risk Analysis:

$$E\, d^2(p^*, \hat{p}) \leq \min_p \{D(p^* \| p) \;+\; L(p)/n\}$$

- Special Cases:
  Traditional parametric: $L(\theta) = (dim/2) \log n + C$
  Nonparametric: $L(p) =$ Metric entropy
  (log cardinality of optimal net)
  Idealized: $L(p) =$ Kolmogorov complexity
- Adaptation:
  Achieves minimax optimal rates simultaneously in every computable subfamily of distributions

MDL Analysis: Key to risk consideration

- Discrepancy between training sample and future

$$Disc(p) = \log \frac{p^*(\underline{Y})}{p(\underline{Y})} - \log \frac{p^*(\underline{Y}')}{p(\underline{Y}')}$$

- Future term may be replaced by population counterpart
- Discrepancy control: If $L(p)$ satisfies the Kraft sum then

$$E\left[\sup_p \{Disc(p) + 2L(p)\}\right] \geq 0$$

- From which the risk bound follows:

  Risk $\leq$ Redundancy $\leq$ Resolvability

$$E\, d^2(p^*, \hat{p}) \leq Red \leq Res_n(p^*)$$

Statistically valid penalized likelihood

- New result
- (B., Li, Huang, Luo 2008, Festschrift for Jorma Rissanen)
- Penalized Likelihood:

$$\hat{p} = \arg\min_{p} \left\{ \frac{1}{n} \log \frac{1}{p(\underline{Y})} \; + \; pen_n(p) \right\}$$

- Penalty condition:

$$pen_n(p) \geq \frac{1}{n} \min_{\tilde{p}} \{ 2L(\tilde{p}) \; + \; \Delta_n(p, \tilde{p}) \}$$

  where the distortion $\Delta_n(p, \tilde{p})$ is the difference in discrepancies at p and a representer $\tilde{p}$
- Risk conclusion:

$$Ed^2(p, \hat{p}) \leq \inf_{p} \{ D(p^* \| p) \; + \; pen_n(p) \}$$

Example: ℓ_1 penalties on coefficients in gen. linear models

- $\mathcal{G}$ is a dictionary of candidate basis functions
- Wavelets, splines, polynomials, trigonometric terms, sigmoids, explanatory variables and their interactions
- Candidate functions in the linear span

$$f(x) = f_\theta(x) = \sum_{g \in \mathcal{G}} \theta_g \, g(x)$$

- ℓ_1 norm of coefficients

$$\|\theta\|_1 = \sum_g |\theta_g|$$

Example Models

- $p = p_f$ specified through a family of function f
- Regression

$$p_f(y|x) = \text{Normal}(f(x), \sigma^2)$$

- Logistic regression with $y \in \{0, 1\}$

$$p_f(y|x) = \text{Logistic}(f(x)) \quad \text{for } y = 1$$

- Log-density estimation

$$p_f(x) = \frac{p_0(x) \exp\{f(x)\}}{c_f}$$

ℓ_1 Penalty

- $pen_n(f_\theta) = \lambda_n \|\theta\|_1$ where $f_\theta(x) = \sum_{g \in \mathcal{G}} \theta_g \, g(x)$
- Popular penalty: Chen & Donoho (96) Basis Pursuit; Tibshirani (96) LASSO; Efron et al (04) LARS; Precursors: Jones (92), B.(90,93,94) greedy algorithm and analysis of combined ℓ_1 and ℓ_0 penalty
- Risk analysis: specify valid λ_n for risk $\leq$ resolvability
- Computation analysis: bound accuracy of new ℓ_1-penalized greedy pursuit algorithm

ℓ_1 penalty is valid for λ_n of order $1/\sqrt{n}$

- Example: ℓ_1 penalized log-density estimation, i.i.d. case

$$\hat{\theta} = \mathrm{argmin}_{\theta} \left\{ \frac{1}{n} \log \frac{1}{p_{f_\theta}(\underline{x})} + \lambda_n \|\theta\|_1 \right\}$$

- Risk bound:

$$Ed(f^*, f_{\hat{\theta}}) \le \inf_{\theta} \left\{ D(f^* \| f_\theta) + \lambda_n \|\theta\|_1 \right\}$$

- Valid for $\lambda_n \ge B\sqrt{\frac{H}{n}}$ with $H = \log Card(\mathcal{G})$
- $B =$ bound on the range of functions in $\mathcal{G}$
- For infinite cardinality $\mathcal{G}$ use metric entropy in place of H
- Results for regression shown in a companion paper

ℓ_1 penalty is valid for λ_n of order $1/\sqrt{n}$

- Example: ℓ_1 penalized log-density estimation, i.i.d. case

$$\hat{\theta} = \mathrm{argmin}_\theta \left\{ \frac{1}{n} \log \frac{1}{p_{f_\theta}(\underline{x})} + \lambda_n \|\theta\|_1 \right\}$$

- Risk bound:

$$Ed(f^*, f_{\hat{\theta}}) \leq \inf_\theta \left\{ D(f^* \| f_\theta) + \lambda_n \|\theta\|_1 \right\}$$

- True with

$$\lambda_n = B\sqrt{\frac{H}{n}} \quad \text{with } H = \log Card(\mathcal{G})$$

- Risk of order λ_n when the target has finite ℓ_1 norm

Comment on proof

- Shannon-like demonstration of the existence of the variable complexity cover property
- Inspiration from technique originating with Lee Jones (92)
- Representer $\tilde{f}$ of f_θ of the form

$$\tilde{f}(x) = \frac{v}{m} \sum_{k=1}^{m} g_k(x)$$

- $g_1, \ldots g_m$ picked at random from $\mathcal{G}$, independently, where g arises with probability proportional to $|\theta_g|$
- May pick them in greedy fashion as in demonstration of fast computation properties
- In the paper in the Festschrift for Rissanen with summary in the ITW - Porto proceedings

Outline of Principles

Summary

- Information divergence is monotone
- Info Capacity $\sim$ Minimax Redundancy $\sim$ Minimax Risk
- Adaptivity of MDL: simultaneously minimax optimal
- Penalized Likelihood risk analysis
- ℓ_0 penalty $\frac{dim}{2} \log n$ classically analyzed
- ℓ_1 penalty $\lambda_n \|\theta\|_1$ valid for $\lambda_n \geq \sqrt{H/n}$.