Information and Statistics

Andrew Barron
Department of Statistics
Yale University

IMA Workshop

On Information Theory and Concentration Phenomena

Minneapolis, April 13, 2015

Outline

- Information and Probability:
  - Monotonicity of Information
  - Large Deviation Exponents
  - Central Limit Theorem
- Information and Statistics:
  - Nonparametric Rates of Estimation
  - Minimum Description Length Principle
  - Penalized Likelihood (one-sided concentration)
  - Implications for Greedy Term Selection
- Achieving Shannon Capacity:
  - Sparse Superposition Coding
  - Adaptive Successive Decoding
  - Rate, Reliability, and Computational Complexity

Probability Limits and Monotonicity

- Information and Probability:
  - Monotonicity of Information
  - Markov chains, martingales
  - Central Limit Theorem
  - Entropy and Fisher Information Inequalities
  - Information Stability (asymptotic equipartition property)
  - Large Deviation Exponents (law of large numbers)

Monotonicity of Information Divergence

- Information Inequality $X \to X'$

$$D(P_{X'} \| P^*_{X'}) \le D(P_X \| P^*_X)$$

- Chain Rule

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

- Markov Chain $\{X_n\}$ with P^* invariant

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

- Convergence

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

Monotonicity of Information Divergence

- Information Inequality $\quad X \to X'$

$$D(P_{X'} \| P^*_{X'}) \leq D(P_X \| P^*_X)$$

- Chain Rule

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

- Markov Chain $\{X_n\}$ with P^* invariant

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

- Convergence

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

Monotonicity of Information Divergence

- Information Inequality $\quad X \to X'$

$$D(P_{X'} \| P^*_{X'}) \leq D(P_X \| P^*_X)$$

- 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 Chain $\{X_n\}$ with P^* invariant

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

- Convergence

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

Monotonicity of Information Divergence

- Information Inequality $\quad X \to X'$

$$D(P_{X'} \| P^*_{X'}) \leq D(P_X \| P^*_X)$$

- Chain Rule

$$\begin{aligned} D(P_{X,X'} \| P^*_{X,X'}) &= D(P_{X'} \| P^*_{X'}) + E\, D(P_{X|X'} \| P^*_{X|X'}) \\ &= D(P_X \| P^*_X) + 0 \end{aligned}$$

- Markov Chain $\{X_n\}$ with P^* invariant

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

- Convergence

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

Monotonicity of Information Divergence

- Information Inequality $\quad X \to X'$

$$D(P_{X'} \| P^*_{X'}) \le D(P_X \| P^*_X)$$

- Chain Rule

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

- Markov Chain $\{X_n\}$ with P^* invariant

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

- Convergence

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

Monotonicity of Information Divergence

- Information Inequality $X \to X'$

$$D(P_{X'} \| P^*_{X'}) \leq D(P_X \| P^*_X)$$

- Chain Rule

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

- Markov Chain $\{X_n\}$ with P^* invariant

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

- Convergence

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

Monotonicity of Information Divergence

- Information Inequality $\quad X \to X'$

$$D(P_{X'} \| P^*_{X'}) \leq D(P_{X'} \| P^*_{X'})$$

- Chain Rule

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

- Markov Chain $\{X_n\}$ with P^* invariant

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

- Convergence

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

- Pinsker-Kullback-Csiszar inequalities

$$A \leq D + \sqrt{2D} \qquad V \leq \sqrt{2D}$$

Martingale Convergence and Limits of Information

- Nonnegative Martingales ρ_n correspond to the density of a measure Q_n given by $Q_n(A) = E[\rho_n 1_A]$.
- Limits can be established 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 $D_n = D(Q_n \| P)$ is an increasing sequence. Suppose it is bounded.
- Then 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

- For $n > m$

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

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

  P^* is the corresponding normal distribution

- For $n > m$

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

- Chain Rule for $n > m$: not clear how to use in this case

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

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

- 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

- For $n > m$

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

- Chain Rule for $n > m$: not clear how to use in this case

$$\begin{aligned} D(P_{Y_m,Y_n} \| P^*_{Y_m,Y_n}) &= D(P_n \| P^*) + ED(P_{Y_m|Y_n} \| P^*_{Y_m|Y_n}) \\ &= D(P_m \| P^*) + ED(P_{Y_n|Y_m} \| P^*_{Y_n|Y_m}) \\ &= D(P_m \| P^*) + D(P_{n-m} \| P^*) \end{aligned}$$

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

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

  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

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

yields

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

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

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

- (Johnson and B. 2004) with Poincare constant R

$$D(P_n \| P^*) \leq \frac{2R}{n-1+2R} D(P_1 \| P^*)$$

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):

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

- (Johnson and B. 2004) with Poincare constant R

$$D(P_n \| P^*) \leq \frac{2R}{n-1+2R} D(P_1 \| P^*)$$

- (Bobkov, Chirstyakov, Gotze 2013) Moment conditions and finite $D(P_1 \| | P^*)$ suffice for this $1/n$ rate

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)

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

  where r is max number of sets in $\mathcal{S}$ in which an index appears
- Proof:
  - simple L_2 projection property of entropy derivative
  - concentration inequality for sums of functions of subsets of independent variables

$$VAR(\sum_{s \in \mathcal{S}} g_s(X_s)) \leq r \sum_{s \in \mathcal{S}} VAR(g_s(X_s))$$

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)

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

where r is max number of sets in $\mathcal{S}$ in which an index appears

- Consequence, for all $n > m$,

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

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

Information-Stability and Error Probability of Tests

- Stability of log-likelihood ratios (AEP)
  (B. 1985, Orey 1985, Cover and Algoet 1986)

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

  where $\mathcal{D}(P\|Q)$ is the relative entropy rate.

- Optimal statistical test: critical region 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

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

- General form of the Chernoff-Stein Lemma.
- Relative entropy rate

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

Information-Stability and Error Probability of Tests

- Stability of log-likelihood ratios (AEP)
  (B. 1985, Orey 1985, Cover and Algoet 1986)

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

  where $\mathcal{D}(P\|Q)$ is the relative entropy rate.

- Optimal statistical test: critical region 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

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

- General form of the Chernoff-Stein Lemma.
- Relative entropy rate

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

Optimality of the Relative Entropy Exponent

- Information Inequality, for any set A_n,

$$D(P_{\underline{Y}^n} \| Q_{\underline{Y}^n}) \geq P(A_n) \log \frac{P(A_n)}{Q(A_n)} + P(A_n^c) \log \frac{P(A_n^c)}{Q(A_n^c)}$$

- Consequence

$$D(P_{\underline{Y}^n} \| Q_{\underline{Y}^n}) \geq P(A_n) \log \frac{1}{Q(A_n)} - H_2(P(A_n))$$

- Equivalently

$$Q(A_n) \geq \exp\left\{ -\frac{D(P_{\underline{Y}^n} \| Q_{\underline{Y}^n}) - H_2(P(A_n))}{P(A_n)} \right\}$$

- For any sequence of pairs of joint distributions, no sequence of tests with $P(A_n)$ approaching 1 can have better $Q(A_n)$ exponent than $D(P_{\underline{Y}^n} \| Q_{\underline{Y}^n})$.

Large Deviations, I-Projection, 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, from i.i.d. sample.
- (Csiszar 1985)

$$Q\{P_n \in C\} \;\leq\; \exp\left\{ -n\, D(C\|Q) \right\}$$

- Information-theoretic representation of Chernoff bound (when C is a half-space)

Large Deviations, I-Projection, 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, from i.i.d. sample.
- (Csiszar 1985)

$$Q\{P_n \in C\} \leq \exp\{ -n\, D(C\|Q) \}$$

- Information-theoretic representation of Chernoff bound (when C is a half-space)

Large Deviations, I-Projection, 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, from i.i.d. sample
- 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)

Information and Statistics:

- Nonparametric Rates of Estimation
- Minimum Description Length Principle
- Penalized Likelihood (one-sided concentration)
- Implications for Greedy Term Selection

Shannon Capacity

- Capacity
  - A Channel $\theta \to \underline{Y}$ is a family of distributions $\{P_{\underline{Y}|\theta} : \theta \in \Theta\}$
  - Information Capacity: $C = \max_{P_\theta} I(\theta; \underline{Y})$
- Communications Capacity
  - Thm: $C_{com} = C$ (Shannon 1948)
- 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)

Setting for Statistical Capacity

Statistical Risk Setting

- Loss function

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

- 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 risk equals expected loss

$$Risk = E[\ell(\theta, \hat{\theta})]$$

Statistical Capacity

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

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

Metric Entropy

Ingredients in Determining Minimax Rates of Statistical Risk

- 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

Information-theoretic Determination of Minimax Rates

- 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)

Minimum Description-Length (Rissanen78,83,B.85, B.&Cover 91...)

- Statistical measure of complexity of $\underline{Y}$

$$L(\underline{Y}) = \min_q \Big[\quad \log 1/q(\underline{Y}) \quad + \quad L(q) \quad \Big]$$

bits for $\underline{Y}$ given q $+$ bits for q

- It is an information-theoretically valid codelength for $\underline{Y}$ for any $L(q)$ satisfying Kraft summability $\sum_q 2^{-L(q)} \leq 1$.
- The minimization is for q in a family indexed by parameters $\{p_\theta(\underline{Y}) : \theta \in \Theta\}$ or by functions $\{p_f(\underline{Y}) : f \in \mathcal{F}\}$
- The estimator $\hat{p}$ is then $p_{\hat{\theta}}$ or $p_{\hat{f}}$.

Statistical Aim

- From training data $\underline{x}$ $\Rightarrow$ estimator $\hat{p}$
- Generalize to subsequent data $\underline{x}'$
- Want $\log 1/\hat{p}(\underline{x}')$ to compare favorably to $\log 1/p(\underline{x}')$
- For targets p close to or in the families
- With $\underline{X}'$ expectation, loss becomes Kullback divergence
- Bhattacharyya, Hellinger, Rényi loss also relevant

Loss

- Kullback Information-divergence:

$$D(P_{\underline{X}'} \| Q_{\underline{X}'}) = E\big[\log p(\underline{X}')/q(\underline{X}')\big]$$

- Bhattacharyya, Hellinger, Rényi divergence:

$$d^2(P_{\underline{X}'}, Q_{\underline{X}'}) = 2 \log 1/E[q(\underline{X}')/p(\underline{X}')]^{1/2}$$

- Product model case: $D(P_{\underline{X}'} \| Q_{\underline{X}'}) = n\, D(P \| Q)$

$$d^2(P_{\underline{X}'}, Q_{\underline{X}'}) = n\, d^2(P, Q)$$

- Relationship:

  $d^2 \leq D \leq (2+b)\, d^2$ if the log density ratio $\leq b$.

MDL Analysis

- Redundancy of Two-stage Code:

$$Red_n = \frac{1}{n} E \left\{ \min_q \left[\log \frac{1}{q(\underline{Y})} \; + \; L(q) \right] - \log \frac{1}{p(\underline{Y})} \right\}$$

- bounded by Index of Resolvability:

$$Res_n(p) = \min_q \left\{ D(p||q) \; + \; \frac{L(q)}{n} \right\}$$

- Statistical Risk Analysis in i.i.d. case with $\mathcal{L}(q) = 2L(q)$:

$$E\, d^2(p, \hat{p}) \leq \min_q \left\{ D(p||q) \; + \; \frac{\mathcal{L}(q)}{n} \right\}$$

- B.85, B.&Cover 91, B., Rissanen, Yu 98, Li 99, Grunwald 07

MDL Analysis: Key to risk consideration

- Discrepancy between training sample and future

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

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

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

- From which the risk bound follows:

$$\text{Risk} \leq \text{Redundancy} \leq \text{Resolvability}$$

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

Statistically valid penalized likelihood

- Likelihood penalties arise via
  - number parameters: $pen(p_\theta) = \lambda\, dim(\theta)$
  - roughness penalties: $pen(p_f) = \lambda\, \|f^s\|^2$
  - coefficient penalties: $pen(\theta) = \lambda \|\theta\|_1$
  - Bayes estimators: $pen(\theta) = \log 1/w(\theta)$
  - Maximum likelihood: $pen(\theta) = constant$
  - MDL:
- Penalized likelihood:

$$\hat{p} = \arg\min_q \{\log 1/q(\underline{Y}) \; + \; pen(q)\}$$

- Under what condition on the penalty will it be true that the sample based estimate $\hat{p}$ has risk controlled by the population counterpart?

$$Ed^2(p, \hat{p}) \leq \inf_q \{D(p\|q) \; + \; \frac{pen(q)}{n}\}$$

Statistically valid penalized likelihood

- Result with J. Li, C. Huang, X. Luo (Festschrift for J. Rissanen 2008)
- Penalized Likelihood:

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

- Penalty condition:

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

where the distortion $\Delta_n(q, \tilde{q})$ is the difference in discrepancies at q and a representer $\tilde{q}$

- Risk conclusion:

$$Ed^2(p, \hat{q}) \leq \inf_q \{D(p\|q) + pen_n(q)\}$$

Information-theoretic valid penalties

- Penalized likelihood

$$\min_{\theta \in \Theta} \left\{ \log \frac{1}{p_\theta(\underline{x})} + Pen(\theta) \right\}$$

- Possibly uncountable Θ
- Valid codelength interpretation if there exists a countable $\tilde{\Theta}$ and L satisfying Kraft such that the above is not less than

$$\min_{\tilde{\theta} \in \tilde{\Theta}} \left\{ \log \frac{1}{p_{\tilde{\theta}}(\underline{x})} + L(\tilde{\theta}) \right\}$$

A variable complexity, variable distortion cover

Equivalently:

- Penalized likelihood with a penalty $Pen(\theta)$ is information-theoretically valid with uncountable Θ, if there is a countable $\tilde{\Theta}$ and Kraft summable $L(\tilde{\theta})$, such that, for every θ in Θ, there is a representor $\tilde{\theta}$ in $\tilde{\Theta}$ such that

$$Pen(\theta) \;\geq\; L(\tilde{\theta}) + \log \frac{p_\theta(\underline{x})}{p_{\tilde{\theta}}(\underline{x})}$$

- This is the link between uncountable and countable cases

Statistical-Risk Valid Penalty

- For an uncountable Θ and a penalty $Pen(\theta)$, $\theta \in \Theta$, suppose there is a countable $\tilde{\Theta}$ and $\mathcal{L}(\tilde{\theta}) = 2L(\tilde{\theta})$ where $L(\tilde{\theta})$ satisfies Kraft, such that, for all $\underline{x}, \theta^*$,

$$\min_{\theta \in \Theta} \left\{ \left[\log \frac{p_{\theta^*}(\underline{x})}{p_\theta(\underline{x})} - d_n^2(\theta^*, \theta) \right] + Pen(\theta) \right\}$$

$$\geq \min_{\tilde{\theta} \in \tilde{\Theta}} \left\{ \left[\log \frac{p_{\theta^*}(\underline{x})}{p_{\tilde{\theta}}(\underline{x})} - d_n^2(\theta^*, \tilde{\theta}) \right] + \mathcal{L}(\tilde{\theta}) \right\}$$

- Proof of the risk conclusion:
  The second expression has expectation ≥ 0,
  so the first expression does too.

- B., Li,& Luo (Rissanen Festschrift 2008, Proc. Porto Info Theory Workshop 2008)

ℓ_1 Penalties are codelength and risk valid

Regression Setting: Linear Span of a Dictionary

- $\mathcal{G}$ is a dictionary of candidate basis functions
  E.g. wavelets, splines, polynomials, trigonometric terms, sigmoids, explanatory variables and their interactions
- Candidate functions in the linear span
  $f_\theta(x) = \sum_{g \in \mathcal{G}} \theta_g \, g(x)$
- weighted ℓ_1 norm of coefficients $\|\theta\|_1 = \sum_g a_g |\theta_g|$
- weights $a_g = \|g\|_n$ where $\|g\|_n^2 = \frac{1}{n} \sum_{i=1}^n g^2(x_i)$
- Regression $p_\theta(y|x) = \text{Normal}(f_\theta(x), \sigma^2)$
- ℓ_1 Penalty (Lasso, Basis Pursuit)

$$pen(\theta) = \lambda \|\theta\|_1$$

Regression with ℓ_1 penalty

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

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

- Regression with Gaussian model

$$\min_\theta \left\{ \frac{1}{2\sigma^2} \frac{1}{n} \sum_{i=1}^{n} (Y_i - f_\theta(x_i))^2 + \frac{1}{2} \log 2\pi\sigma^2 + \frac{\lambda_n}{\sigma} \|\theta\|_1 \right\}$$

- Codelength Valid and Risk Valid for

$$\lambda_n \geq \sqrt{\frac{2\log(2p)}{n}} \quad \text{with } p = Card(\mathcal{G})$$

Adaptive risk bound specialized to regression

- Again for fixed design and $\lambda_n = \sqrt{\frac{2 \log 2p}{n}}$, multiplying through by $4\sigma^2$,

$$E\|f^* - f_{\hat{\theta}}\|_n^2 \leq \inf_{\theta} \left\{ 2\|f^* - f_\theta\|_n^2 + 4\sigma\lambda_n \|\theta\|_1 \right\}$$

- In particular for all targets $f^* = f_{\theta^*}$ with finite $\|\theta^*\|$ the risk bound $4\sigma\lambda_n\|\theta^*\|$ is of order $\sqrt{\frac{\log M}{n}}$
- Details in Barron, Luo (proceedings Workshop on Information Theory Methods in Science & Eng. 2008), Tampere, Finland

Comment on proof

- The variable complexity cover property is demonstrated by choosing the 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 $|\theta_g|$

Practical Communication by Regression

- Achieving Shannon Capacity: (with A. Joseph, S. Cho)
  - Gaussian Channel with Power Constraints
  - History of Methods
  - Communication by Regression
  - Sparse Superposition Coding
  - Adaptive Successive Decoding
  - Rate, Reliability, and Computational Complexity

Shannon Formulation

- Input bits: $u = (u_1, u_2, \ldots\ldots, u_K)$

$$\downarrow$$

- Encoded: $x = (x_1, x_2, \ldots, x_n)$

$$\downarrow$$

- Channel: $p(y|x)$

$$\downarrow$$

- Received: $y = (y_1, y_2, \ldots, y_n)$

$$\downarrow$$

- Decoded: $\hat{u} = (\hat{u}_1, \hat{u}_2, \ldots\ldots, \hat{u}_K)$
- Rate: $R = \frac{K}{n}$ Capacity $C = \max I(X; Y)$
- Reliability: Want small Prob$\{\hat{u} \neq u\}$
  and small Prob$\{\textit{Fraction mistakes} \geq \alpha\}$

Gaussian Noise Channel

- Input bits: $u = (u_1, u_2, \ldots\ldots, u_K)$

$$\downarrow$$

- Encoded: $x = (x_1, x_2, \ldots, x_n)$ ave $\frac{1}{n}\sum_{i=1}^{n} x_i^2 \le P$

$$\downarrow$$

- Channel: $p(y|x)$ $y = x + \varepsilon \quad \varepsilon \sim N(0, \sigma^2 I)$

$$\downarrow$$

- Received: $y = (y_1, y_2, \ldots, y_n)$

$$\downarrow$$

- Decoded: $\hat{u} = (\hat{u}_1, \hat{u}_2, \ldots\ldots, \hat{u}_K)$
- Rate: $R = \frac{K}{n}$ Capacity $C = \frac{1}{2}\log(1 + P/\sigma^2)$
- Reliability: Want small Prob$\{\hat{u} \neq u\}$
  and small Prob$\{\textit{Fraction mistakes} \ge \alpha\}$

Shannon Theory meets Coding Practice

- The Gaussian noise channel is the basic model for
  - wireless communication
    radio, cell phones, television, satellite, space
  - wired communication
    internet, telephone, cable
- Forney and Ungerboeck 1998 review
  - modulation, coding, and shaping for the Gaussian channel
- Richardson and Urbanke 2008 cover much of the state of the art in the analysis of coding
  - There are fast encoding and decoding algorithms, with empirically good performance for LDPC and turbo codes
  - Some tools for their theoretical analysis, but obstacles remain for mathematical proof of these schemes achieving rates up to capacity for the Gaussian channel
- Arikan 2009, Arikan and Teletar 2009 polar codes
  - Adapting polar codes to Gaussian channel (Abbe and B. 2011)
- Method here is different. Prior knowledge of the above is not necessary to follow what we present.

Sparse Superposition Code

- Input bits: $u = (u_1 \ldots\ldots\ldots\ldots u_K)$
- Coefficients: $\beta = (00 * 0000000000 * 00 \ldots 0 * 000000)^T$
- Sparsity: L entries non-zero out of N
- Matrix: X, n by N, all entries indep Normal$(0, 1)$
- Codeword: $X\beta$, superposition of a subset of columns
- Receive: $y = X\beta + \varepsilon$, a statistical linear model
- Decode: $\hat{\beta}$ and $\hat{u}$ from X,y

Sparse Superposition Code

- Input bits: $u = (u_1 \ldots\ldots\ldots\ldots u_K)$
- Coefficients: $\beta = (00 * 0000000000 * 00 \ldots 0 * 000000)^T$
- Sparsity: L entries non-zero out of N
- Matrix: X, n by N, all entries indep Normal$(0,1)$
- Codeword: $X\beta$
- Receive: $y = X\beta + \varepsilon$
- Decode: $\hat{\beta}$ and $\hat{u}$ from X,y
- Rate: $R = \frac{K}{n}$ from $K = \log \binom{N}{L}$, near $L \log \left(\frac{N}{L} e\right)$

Sparse Superposition Code

- Input bits: $u = (u_1 \ldots\ldots\ldots\ldots u_K)$
- Coefficients: $\beta = (00 * 0000000000 * 00 \ldots 0 * 000000)^T$
- Sparsity: L entries non-zero out of N
- Matrix: X, n by N, all entries indep Normal$(0, 1)$
- Codeword: $X\beta$
- Receive: $y = X\beta + \varepsilon$
- Decode: $\hat{\beta}$ and $\hat{u}$ from X,y
- Rate: $R = \frac{K}{n}$ from $K = \log \binom{N}{L}$
- Reliability: small Prob$\{$*Fraction* $\hat{\beta}$ *mistakes* $\geq \alpha\}$, small α

Sparse Superposition Code

- Input bits: $u = (u_1 \ldots\ldots\ldots\ldots u_K)$
- Coefficients: $\beta = (00 * 0000000000 * 00 \ldots 0 * 000000)^T$
- Sparsity: L entries non-zero out of N
- Matrix: X, n by N, all entries indep Normal$(0, 1)$
- Codeword: $X\beta$
- Receive: $y = X\beta + \varepsilon$
- Decode: $\hat{\beta}$ and $\hat{u}$ from X,y
- Rate: $R = \frac{K}{n}$ from $K = \log \binom{N}{L}$
- Reliability: small Prob$\{Fraction\ \hat{\beta}\ mistakes \geq \alpha\}$, small α
- Outer RS code: rate $1-2\alpha$, corrects remaining mistakes
- Overall rate: $R_{tot} = (1-2\alpha)R$

Sparse Superposition Code

- Input bits: $u = (u_1 \ldots\ldots\ldots\ldots u_K)$
- Coefficients: $\beta = (00 * 0000000000 * 00 \ldots 0 * 000000)^T$
- Sparsity: L entries non-zero out of N
- Matrix: X, n by N, all entries indep Normal$(0,1)$
- Codeword: $X\beta$
- Receive: $y = X\beta + \varepsilon$
- Decode: $\hat{\beta}$ and $\hat{u}$ from X,y
- Rate: $R = \frac{K}{n}$ from $K = \log \binom{N}{L}$
- Reliability: small Prob$\{Fraction\ \hat{\beta}\ mistakes \geq \alpha\}$, small α
- Outer RS code: rate $1-2\alpha$, corrects remaining mistakes
- Overall rate: $R_{tot} = (1-2\alpha)R$.

Is it reliable with rate up to capacity?

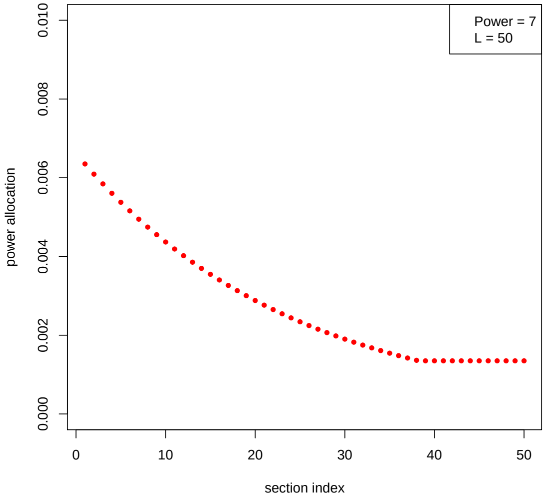
Partitioned Superposition Code

- Input bits: $u = (u_1 \ldots, \;\; \ldots, \;\; \ldots, \;\; \ldots u_K)$
- Coefficients: $\beta = (00*00000, \; 00000*00, \; \ldots, \; 0*000000)$
- Sparsity: L sections, each of size $B = N/L$, a power of 2.
  1 non-zero entry in each section
- Indices of nonzeros: $(j_1, j_2, \ldots, j_L)$ directly specified by u
- Matrix: X, n by N, splits into L sections
- Codeword: $X\beta$
- Receive: $y = X\beta + \varepsilon$
- Decode: $\hat{\beta}$ and $\hat{u}$
- Rate: $R = \frac{K}{n}$ from $K = L \log \frac{N}{L} = L \log B$
  may set $B = n$ and $L = nR/\log n$
- Reliability: small Prob$\{$*Fraction* $\hat{\beta}$ *mistakes* $\geq \alpha\}$
- Outer RS code: Corrects remaining mistakes
- Overall rate: up to capacity?

Power Allocation

- Coefficients: $\beta = (00{*}00000,\ 00000{*}00, \ldots, 0{*}000000)$
- Indices of nonzeros: $sent = (j_1, j_2, \ldots, j_L)$
- Coeff. values: $\beta_{j_\ell} = \sqrt{P_\ell}$ for $\ell = 1, 2, \ldots, L$
- Power control: $\sum_{\ell=1}^{L} P_\ell = P$
- Codewords: $X\beta$, have average power P
- Power Allocations
  - Constant power: $P_\ell = P/L$
  - Variable power: P_ℓ proportional to $u_\ell = e^{-2C\,\ell/L}$
  - Variable with leveling: P_ℓ proportional to $\max\{u_\ell, cut\}$

Power Allocation



Contrast Two Decoders

Decoders using received $y = X\beta + \varepsilon$

Optimal: Least Squares Decoder

$$\hat{\beta} = \text{argmin} \| Y - X\beta \|^2$$

- minimizes probability of error with uniform input distribution
- reliable for all $R < C$, with best form of error exponent

Practical: Adaptive Successive Decoder

- fast decoder
- reliable using variable power allocation for all $R < C$

Adaptive Successive Decoder

Decoding Steps

- Start: [Step 1]
  - Compute the inner product of Y with each column of X
  - See which are above a threshold
  - Form initial fit as weighted sum of columns above threshold
- Iterate: [Step $k \geq 2$]
  - Compute the inner product of residuals $Y - Fit_{k-1}$ with each remaining column of X
  - See which are above threshold
  - Add these columns to the fit
- Stop:
  - At Step $k = \log B$, or
  - if there are no inner products above threshold

Decoding Progression

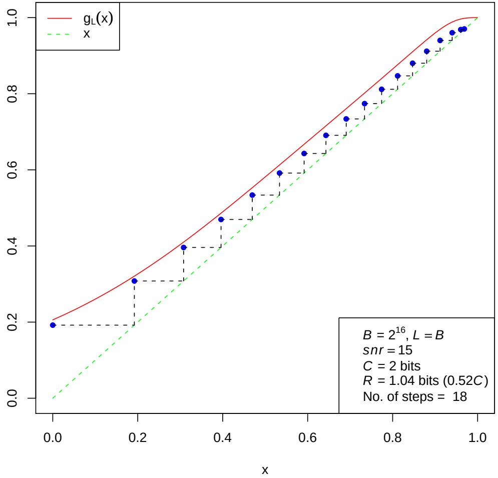


Figure : Plot of likely progression of weighted fraction of correct detections $\hat{q}_{1,k}$, for $snr = 15$.

Decoding Progression

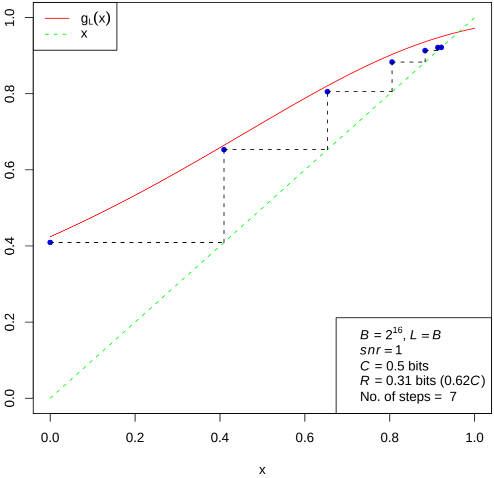


Figure : Plot of of likely progression of weighted fraction of correct detections $\hat{q}_{1,k}$, for $snr = 1$.

Rate and Reliability

Optimal: Least squares decoder of sparse superposition code

- Prob error exponentially small in n for small $\Delta = C - R > 0$

$$\text{Prob}\{Error\} \leq e^{-n(C-R)^2/2V}$$

- In agreement with the Shannon-Gallager optimal exponent, though with possibly suboptimal V depending on the *snr*

Practical: Adaptive Successive Decoder, with outer RS code.

- achieves rates up to C_B approaching capacity

$$C_B = \frac{C}{1 + c_1 / \log B}$$

- Probability exponentially small in L for $R \leq C_B$

$$\text{Prob}\Big\{Error\Big\} \leq e^{-L(C_B - R)^2 c_2}$$

- Improves to $e^{-c_3 L (C_B - R)^2 (\log B)^{0.5}}$ using a Bernstein bound.
- Nearly optimal when $C_B - R$ is of the same order as $C - C_B$.
- Our c_1 is near $(2.5 + 1/snr) \log\log B + 4C$

Summary

- Sparse superposition coding is fast and reliable at rates up to channel capacity
- Formulation and analysis blends modern statistical regression and information theory