INFORMATION AND STATISTICS

Andrew R. Barron

YALE UNIVERSITY
DEPARTMENT OF STATISTICS

Presentation, April 30, 2015

Information Theory Workshop, Jerusalem

Information and Statistics

Topics in the abstract from which I make a selection

- Information Theory and Inference:
  - Flexible high-dimensional function estimation
  - Neural nets: sigmoidal and sinusoidal activation functions
  - Approximation and estimation bounds
  - Minimum description length principle
  - Penalized likelihood risk bounds and minimax rates
  - Computational strategies
- Achieving Shannon Capacity:
  - Communication by regression
  - Sparse superposition coding
  - Adaptive successive decoding
  - Rate, reliability, and computational complexity
- Information Theory and Probability:
  - General entropy power inequalities
  - Entropic central limit theorem and its monotonicity
  - Monotonicity of relative entropy in Markov chains
  - Monotonicity of relative entropy in statistical mechanics

Information and Statistics

- Information Theory and Inference:
  - Flexible high-dimensional function estimation
  - Neural nets: sigmoidal and sinusoidal activation functions
  - Approximation and estimation bounds
  - Minimum description length principle
  - Penalized likelihood risk bounds and minimax rates
  - Computational strategies

Plan for Information and Inference

- Setting
  - Univariate & muntivariate polynomials, sinusoids, sigmoids
  - Fit to training data
  - statistical risk is the error of generalization to new data
- The challenge of high-dimensional function estimation
  - Estimation failure of rigid approximation models in high dim
  - Computation difficulities of flexible models in high dim
- Flexible approximation
  - by stepwise subset selection
  - by optimization of parameterized basis functions
- Approximation bounds
  - Relate error to number of terms
- Information-theoretic risk bounds
  - Relate error to number of terms and sample size
- Computational challenge
  - Constructing an optimization path

The Problem

From observational or experimental data, relate a response variable Y to several explanatory variables $X_1, X_2, \ldots, X_d$

- Common task throughout science and engineering
- Central to the "Scientific Method"

Aspects of this problem are variously called:

Statistical regression, prediction, response surface estimation, analysis of variance, function fitting, function approximation, nonparametric estimation, high-dimensional statistics, data mining, machine learning, computational learning, pattern recognition, artificial intelligence, cybernetics, artificial neural networks, deep learning

Dimensionality

The blessing and the curse of dimensionality

- With increasing number of variables d there is an exponential growth in the number of distinct terms that can be combined in modeling the function
- Larger number of relevant variables d allows in principle for better approximation to the response
- Large d might lead to a need for exponentially large number of observations n or to a need for exponentially large computation time
- Under what conditions can we take advantage of the blessing and overcome the curse.

Example papers for some of what is to follow

Papers illustrating my background addressing these questions of high dimensionality (available from www.stat.yale.edu)

- A. R. Barron, R. L. Barron (1988). Statistical learning networks: a unifying view. *Computing Science & Statistics: Proc. 20th Symp on the Interface*, ASA, p.192-203.
- A. R. Barron (1993). Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory*, Vol.39, p.930-944.
- A. R. Barron, A. Cohen, W. Dahmen, R. DeVore (2008). Approximation and learning by greedy algorithms. *Annals of Statistics*, Vol.36, p.64-94.
- A.R. Barron, C. Huang, J. Q. Li and Xi Luo (2008). MDL principle, penalized Likelihood, and statistical risk. *Proc. IEEE Information Theory Workshop*, Porto, Portugal, p.247-257. Also *Feschrift for Jorma Rissanen*. Tampere Univ. Press, Finland.

Data Setting

- Data: $(\underline{X}_i, Y_i)$, $i = 1, 2, \ldots, n$
- Inputs: explanatory variable vectors

$$\underline{X}_i = (X_{i,1}, X_{i,2}, \ldots, X_{i,d})$$

- Domain: Either a unit cube in R^d or all of R^d
- Random design: independent $\underline{X}_i \sim P$
- Output: response variable Y_i in R
  - Moment conditions, with Bernstein constant c
- Relationship: $E[Y_i|\underline{X}_i] = f(\underline{X}_i)$ as in:
  - Perfect observation: $Y_i = f(\underline{X}_i)$
  - Noisy observation: $Y_i = f(X_i) + \epsilon_i$ with ϵ_i indep $N(0, \sigma^2)$
  - Classification: $Y \in \{0, 1\}$ with $f(\underline{X}) = P[Y = 1|\underline{X}]$
- Function: $f(\underline{x})$ unknown

Univariate function approximation: $d = 1$

Basis functions for series expansion

$$\phi_0(x), \phi_1(x), \ldots, \phi_K(x), \ldots$$

Polynomial basis (with degree K)

$$1, \quad x, \quad x^2, \ldots, x^K$$

Sinusoidal basis (with period L, and with $K = 2k$),

$$1, \cos(2\pi(1/L)x), \sin(2\pi(1/L)x), \ldots, \cos(2\pi(k/L)x), \sin(2\pi(k/L)x)$$

Piecewise constant on $[0, 1]$

$$1_{\{x \geq 0\}}, 1_{\{x \geq 1/K\}}, 1_{\{x \geq 2/K\}}, \ldots, 1_{\{x \geq 1\}}$$

Other spline bases and wavelet bases

Univariate function approximation: $d = 1$

Standard 1-dim approximation models

Project to the linear span of the basis

- Rigid form (not flexible), with coefficients c_k adjusted to fit the response,

$$f_K(x) = \sum_{k=0}^{K} c_k \, \phi_k(x).$$

- Flexible form, with a subset $k_1 \ldots k_m$ chosen to best fit the response, for a given number of terms m

$$\sum_{j=1}^{m} c_j \, \phi_{k_j}(x).$$

Fit by all-subset regression (if m and K are not too large) or by forward stepwise regression, selecting from the dictionary $\Phi = \{\phi_0, \phi_1, \ldots, \phi_K\}$

Multivariate function approximation: $d > 1$

- Multivariate product bases:

$$\phi_{\underline{k}}(\underline{x}) = \phi_{k_1,k_2,\ldots,k_d}(x_1, x_2, \ldots, x_d)$$

$$= \phi_{k_1}(x_1)\phi_{k_2}(x_2)\cdots\phi_{k_d}(x_d)$$

- Rigid approximation model

$$\sum_{k_1=0}^{K}\sum_{k_2=0}^{K}\cdots\sum_{k_d=0}^{K} c_{\underline{k}}\,\phi_{\underline{k}}(\underline{x})$$

- Exponential size: $(K+1)^d$ terms in the sum
- Requires exponentially large sample size $n >> (K+1)^d$ for accurate estimation
- Statistically and computationally problematic

Flexible multivariate function approximation: $d > 1$

BY SUBSET SELECTION:

- A subset $\underline{k}_1 \ldots \underline{k}_m$ is chosen to fit the response, with a given number of terms m

$$\sum_{j=1}^{m} c_j \, \phi_{\underline{k}_j}(\underline{x})$$

- Full forward stepwise selection:
  - computationally infeasible for large d because the dictionary is exponentially large, of size $(K+1)^d$.
- Adhoc stepwise selection:
  - SAS stepwise polynomials.
  - Friedman MARS, Barron-Xiao MAPS, *Ann. Statist.* 1991.
  - Each step search only incremental modification of terms.
  - Manageable number of choices mKd each step.
  - Computationally fast, not known if it approximates well.

Flexible multivariate function approximation: $d > 1$

By internally parameterized models & nonlinear least squares

- Fit functions $f_m(x) = \sum_{j=1}^{m} c_j \phi(\underline{x}, \underline{\theta})$ in the span of a parameterized dictionary $\Phi = \{\phi(\cdot, \underline{\theta}) : \underline{\theta} \in \Theta\}$
- Product bases:
  using continuous powers, frequencies or thresholds

$$\phi(\underline{x}, \underline{\theta}) = \phi_1(x_1, \theta_1)\, \phi_1(x_2, \theta_2) \cdots \phi_1(x_d, \theta_d)$$

- Ridge bases: as in projection pursuit regression models, sinusoidal models, and single-hidden-layer neural nets:

$$\phi(\underline{x}, \underline{\theta}) = \phi_1(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_d x_d)$$

- Internal parameter vector $\underline{\theta}$ of dimension $d+1$.
- Univariate function $\phi(z) = \phi_1(z)$ is the activation function

Building Non-linear Dictionaries

Examples of activation functions $\phi(z)$

- Perceptron networks: $1_{\{z>0\}}$ or $\mathrm{sgn}(z)$
- Sigmoidal networks: $e^z/(1+e^z)$ or $\tanh(z)$
- Sinusoidal models: $cos(z)$
- Hinging hyperplanes: $(z)_+$
- Quadratic splines: 1, z, $(z)_+^2$
- Cubic splines: 1, z, z^2, $(z)_+^3$
- Polynomials: $(z)^q$

Notation

- Response vector: $Y = (Y_i)_{i=1}^n$ in R^n
- Dictionary vectors: $\Phi_{(n)} = \left\{(\phi(\underline{X}_i, \underline{\theta}))_{i=1}^n \ : \ \underline{\theta} \in \Theta\right\} \subset R^n$
- Sample squared norm: $\|f\|_{(n)}^2 = \frac{1}{n}\sum_{i=1}^n f^2(\underline{X}_i)$
- Population squared norm: $\|f\|^2 = \int f^2(\underline{x}) P(d\underline{x})$
- Normalized dictionary condition: $\|\phi\| \leq 1$ for $\phi \in \Phi$

Flexible $m-$term nonlinear optimization

Impractical one-shot optimization

- Sample version

$$\hat{f}_m \quad \text{achieves} \quad \min_{(\underline{\theta}_j, c_j)_{j=1}^m} \|Y - \sum_{j=1}^m c_j \, \phi_{\underline{\theta}_j}\|_{(n)}^2$$

- Population version

$$f_m \quad \text{achieves} \quad \min_{(\underline{\theta}_j, c_j)_{j=1}^m} \|f - \sum_{j=1}^m c_j \, \phi_{\underline{\theta}_j}\|^2$$

- Optimization of $(\underline{\theta}_j, c_j)_{j=1}^m$ in $R^{(d+2)m}$.

Flexible $m-$term nonlinear optimization

GREEDY OPTIMIZATIONS

- Step 1: Choose c_1, $\underline{\theta}_1$ to achieve $\min \| Y - c\phi_{\underline{\theta}}\|^2_{(n)}$
- Step $m > 1$: Arrange

$$\hat{f}_m = \alpha \, \hat{f}_{m-1} + c\,\phi(\underline{x}, \underline{\theta}_m)$$

with α_m, c_m, $\underline{\theta}_m$ chosen to achieve

$$\min_{\alpha, c, \underline{\theta}} \| Y - \alpha \, \hat{f}_{m-1} - c\,\phi_{\underline{\theta}}\|^2_{(n)}.$$

- Also acceptable, with $res_i = Y_i - \hat{f}_{m-1}(\underline{X}_i)$,
  - Choose $\underline{\theta}_m$ to achieve $\max_{\underline{\theta}} \sum_{i=1}^n res_i \, \phi(\underline{X}_i, \underline{\theta})$
  - Reduced dimension of the search space (still problematic?)
  - Foward stepwise selection of $S_m = \{\phi_{\underline{\theta}_1}, \ldots, \phi_{\underline{\theta}_m}\}$. Given S_{m-1}, combine the terms to achieve

$$\min_{\underline{\theta}} d(Y, span\{\phi_{\underline{\theta}_1}, \ldots, \phi_{\underline{\theta}_{m-1}}, \phi_{\underline{\theta}}\})$$

Basic $m-$term approximation and computation bounds

For either one-shot or greedy approximation
(B. *IT* 1993, Lee et al *IT* 1995)

- Population version:

$$\|f - f_m\| \leq \frac{\|f\|_\Phi}{\sqrt{m}}$$

  and moreover

$$\|f - f_m\|^2 \leq \inf_g \left\{ \|f - g\|^2 + \frac{2\|g\|_\Phi^2}{m} \right\}$$

- Sample version:

$$\|Y - \hat{f}_m\|_{(n)}^2 \leq \|Y - f\|_{(n)}^2 + \frac{2\|f\|_\Phi^2}{m}$$

  where $\|f\|_\Phi$ is the variation of f with respect to Φ (as will be defined on the next slide).

ℓ_1 norm on coefficients in representation of f

- Consider the range of a neural net, expressed via the bound,

$$\Big|\sum_j c_j \,\mathrm{sgn}(\theta_{0,j} + \theta_{1,j}x_1 + \ldots + \theta_{d,j}x_d)\Big| \le \sum_j |c_j|$$

equality if $\underline{x}$ is in polygon where $\mathrm{sgn}(\underline{\theta}_j \cdot \underline{x}) = \mathrm{sgn}(c_j)$ for all j

- Motivates the norm

$$\|f\|_\Phi = \lim_{\epsilon \to 0} \inf \Big\{ \sum_j |c_j| \;:\; \|\sum_j c_j \phi_{\underline{\theta}_j} - f\| \le \epsilon \Big\}$$

called the variation of f with respect to Φ (B. 1991)

$$\|f\|_\Phi = V_\Phi(f) = \inf\{ V \;:\; f/V \in closure(conv(\pm\Phi))\}$$

- It appears in the bound $\|f - f_m\| \le \frac{\|f\|_\Phi}{\sqrt{m}}$

ℓ_1 norm on coefficients in representation of f

- Finite sum representations, $f(\underline{x}) = \sum_j c_j \phi(\underline{x}, \underline{\theta}_j)$. Variation $\|f\|_\Phi = \sum_j |c_j|$, which is the ℓ_1 norm of the coefficients in representation of f in the span of Φ
- Infinite integral representation $f(\underline{x}) = \int e^{i\,\underline{\theta}\cdot\underline{x}}\,\tilde{f}(\underline{\theta})\,d\theta$ (Fourier representation), for $\underline{x}$ in a unit cube. The variation $\|f\|_\Phi$ is bounded by an L_1 spectral norm:

$$\|f\|_{cos} = \int_{R^d} |\tilde{f}(\underline{\theta})|\,d\underline{\theta}$$

$$\|f\|_{step} \le \int |\tilde{f}(\underline{\theta})|\ \|\underline{\theta}\|_1\,d\underline{\theta}$$

$$\|f\|_{q-spline} \le \int |\tilde{f}(\underline{\theta})|\ \|\underline{\theta}\|_1^{q+1}\,d\underline{\theta}$$

- As we said, this $\|f\|_\Phi$ appears in the numerator of the approximation bound.

Statistical Risk

- The population accuracy of function estimated from sample
- Statistical risk $E\|\hat{f}_m - f\|^2 = E(\hat{f}_m(\underline{X}) - f(\underline{X}))^2$
- Expected squared generalization error on new $\underline{X} \sim P$ of the estimator trained on the data $(\underline{X}_i, Y_i)_{i=1}^n$
- Minimax optimal risk bound, via information theory

$$E\|\hat{f}_m - f\|^2 \;\leq\; \|f_m - f\|^2 + c\frac{m}{n}\log N(\Phi, \delta_n).$$

Here $\log N(\Phi, \delta_n)$ is the metric entropy of Φ at $\delta_n = 1/n$; with Φ of metric dimension d, it is of order $d\log(1/\delta_n)$, so

$$E\|\hat{f}_m - f\|^2 \;\leq\; \frac{\|f\|_\Phi^2}{m} \;+\; \frac{cmd}{n}\log n$$

- Need only $n >> md$ rather than $n >> (K+1)^d$.
- Best bound is $2\|f\|_\Phi\sqrt{\frac{cd}{n}\log n}$ at $m^* = \|f\|_\Phi\sqrt{n/cd\log n}$

Adaptation

- Adapt network size m and choice of internal parameters
- Minimum Description Length Principle leads to Complexity penalized least squares criterion. Let $\hat{m}$ achieve

$$\min_m \left\{ \|Y - \hat{f}_m\|^2_{(n)} + 2c\frac{m}{n}\log N(\Phi, \delta_n) \right\}$$

- Information-theoretic risk bound

$$E\|\hat{f}_{\hat{m}} - f\|^2 \leq \min_m \left\{ \|f_m - f\|^2 + 2c\frac{m}{n}\log N(\Phi, \delta_n) \right\}$$

  - Performs as well as if the best m^* were known in advance.
  - $\|f\|^2_\Phi / m$ replaces $\|f_m - f\|^2$ in the greedy case.
- ℓ_1 penalized least squares
  - Achieves the same risk bound
  - Retains the MDL interpretation (B, Huang,Li,Luo,2008)

Confronting the computational challenge

- Greedy search
  - Reduces dimensionality of optimization from *md* to just *d*
  - Obtain a current $\underline{\theta}_m$ achieving within a constant factor of the maximum of

$$J_n(\theta) \;=\; \frac{1}{n}\sum_{i=1}^{n} res_i\,\phi(\underline{X}_i, \underline{\theta}).$$

- This surface can still have many maxima.
  - We might get stuck at an undesirably low local maximum.
- New computational strategies:
  1. A special case in which the set of maxima can be identified.
  2. Optimization path via solution to a pde for ridge bases.

A special case in which the maxima can be identified

- Insight from a special case:
  - Sinusoidal dictionary: $\phi(\underline{x}, \underline{\theta}) = e^{-i\underline{\theta}\cdot\underline{x}}$
  - Gaussian design: $\underline{X}_i \sim \text{Normal}(0, \tau I)$
  - Target function: $f(\underline{x}) = \sum_{j=1}^{m_o} c_j \, e^{i\underline{\alpha}_j \cdot \underline{x}}$
- For step 1, with large n, the objective function becomes near its population counterpart

$$J(\theta) = E\left[f(\underline{X}) e^{-i\underline{\theta}\cdot\underline{X}}\right] = \sum_{j=1}^{m_o} c_j \, E\left[e^{i\underline{\alpha}_j\cdot\underline{X}} e^{-i\underline{\theta}\cdot\underline{X}}\right]$$

  which simplifies to

$$\sum_{j=1}^{m_o} c_j \, e^{-(\tau/2)\|\underline{\alpha}_j - \underline{\theta}\|^2}.$$

- For large τ it has precisely m_o maxima, one at each of the $\underline{\alpha}_j$ in the target function.

Optimization path for bounded ridge bases

More general approach to seek approximation optimization of

$$J(\underline{\theta}) = \sum_{i=1}^{n} r_i \, \phi(\underline{\theta}^T \underline{X}_i)$$

Adaptive Annealing:

- recent & current work with Luo, Chatterjee, Klusowski
- Sample $\underline{\theta}_t$ from the evolving density

$$p_t(\underline{\theta}) \; = \; e^{\, t\, J(\underline{\theta}) - c_t} \, p_0(\underline{\theta})$$

  along a sequence of values of t from 0 to t_{final}
- use t_{final} of order $(d \log d)/n$
- Initialize with θ_0 drawn from a product prior $p_0(\underline{\theta})$, such as normal$(0, I)$ or a product of standard Cauchy
- Starting from the random θ_0 define the optimization path θ_t such that its distribution tracks the target density p_t

Optimization path

- Adaptive Annealing: Arrange θ_t from the evolving density

$$p_t(\theta) = e^{tJ(\theta) - c_t} p_0(\theta)$$

with θ_0 drawn from $p_0(\theta)$

- State evolution with vector-valued change function $G_t(\theta)$:

$$\theta_{t+h} = \theta_t \; - \; h\, G_t(\theta_t)$$

or better: θ_{t+h} is the solution to

$$\theta_t = \theta_{t+h} \; + \; h\, G_t(\theta_{t+h}),$$

with small step-size h, such that $\underline{\theta} + h\, G_t(\underline{\theta})$ is invertible with a positive definite Jacobian, and solves equations for the evolution of $p_t(\theta)$.

- As we will see there are many such change functions $G_t(\theta)$, though not all are nice.

Nice change functions G_t

- A function on R^d is said to be nice if *the logarithm of its magnitute is bounded by an expression of order logarithmic in d and in* $1 + \|\theta\|^2$.
- A vector-valued function is said to be nice if its norm is nice.
- For computationally feasibility and distributional validity, seek a nice change function G_t satifying the upcoming density evolution rule.

Solve for the change G_t to track the density p_t

- Density evolution: by the Jacobian rule

$$p_{t+h}(\theta) = p_t\big(\theta + h\,G_t(\theta)\big)\,det\big(I + h\nabla G_t^T(\theta)\big)$$

Up to terms of order h

$$p_{t+h}(\theta) = p_t(\theta) + h\left[(G_t(\theta))^T\,\nabla p_t(\theta) + p_t(\theta)\,\nabla^T G_t(\theta)\right]$$

- In agreement for small h with the partial diff equation

$$\frac{\partial}{\partial t}\,p_t(\theta) \;=\; \nabla^T\big[G_t(\theta)p_t(\theta)\big]$$

- The right side is $G_t^T(\theta)\nabla p_t(\theta) + p_t(\theta)\nabla^T G_t(\theta)$. Dividing by $p_t(\theta)$ it is expressed in the log density form

$$\frac{\partial}{\partial t}\log p_t(\theta) \;=\; \nabla^T G_t(\theta) \;+\; G_t^T(\theta)\,\nabla \log p_t(\theta)$$

Candidate solutions

Solution of smallest L_2 norm of $G_t(\theta)p_t(\theta)$ at a specific t.

- Let $G_t(\theta)p_t(\theta) = \nabla b(\theta)$, gradient of a function $b(\theta)$
- Let $f(\theta) = \frac{\partial}{\partial t} p_t(\theta)$
- Set $green(\theta)$ proportional to $1/\|\theta\|^{d-2}$, harmonic for $\theta \neq 0$.
- The partial diff equation becomes the Poisson equation:

$$\nabla^T \nabla b(\theta) = f(\theta)$$

- Solution

$$b(\theta) \quad = \quad (f * green)(\theta)$$

Candidate solutions

Solution of smallest L_2 norm of $G_t(\theta)p_t(\theta)$ at a specific t

- Let $G_t(\theta)p_t(\theta) = \nabla b(\theta)$, gradient of a function $b(\theta)$
- Let $f(\theta) = \frac{\partial}{\partial t} p_t(\theta)$
- Set $green(\theta)$ proportional to $1/\|\theta\|^{d-2}$, harmonic for $\theta \neq 0$.
- The partial diff equation becomes the Poisson equation:

$$\nabla^T \nabla b(\theta) = f(\theta)$$

- Solution, using $\nabla green(\theta) = c_d\, \theta/\|\theta\|^d$

$$\nabla b(\theta) \;=\; (f * \nabla green)(\theta)$$

Candidate solutions

Solution of smallest L_2 norm of $G_t(\theta)p_t(\theta)$ at a specific t

- Let $G_t(\theta)p_t(\theta) = \nabla b(\theta)$, gradient of a function $b(\theta)$
- Let $f(\theta) = \frac{\partial}{\partial t} p_t(\theta)$
- Set $green(\theta)$ proportional to $1/\|\theta\|^{d-2}$, harmonic for $\theta \neq 0$.
- The partial diff equation becomes the Poisson equation:

$$\nabla^T [G_t(\theta)p_t(\theta)] = f(\theta)$$

- Solution, using $\nabla green(\theta) = c_d\, \theta/\|\theta\|^d$

$$G_t(\theta)p_t(\theta) = \ (f * \nabla green)(\theta)$$

Candidate solutions

Solution of smallest L_2 norm of $G_t(\theta)p_t(\theta)$ at a specific t

- Let $G_t(\theta)p_t(\theta) = \nabla b(\theta)$, gradient of a function $b(\theta)$
- Let $f(\theta) = \frac{\partial}{\partial t} p_t(\theta)$
- Set $green(\theta)$ proportional to $1/\|\theta\|^{d-2}$, harmonic for $\theta \neq 0$.
- The partial diff equation becomes the Poisson equation:

$$\nabla^T [G_t(\theta)p_t(\theta)] = f(\theta)$$

- Solution, using $\nabla green(\theta) = c_d\, \theta/\|\theta\|^d$

$$G_t(\theta) \;=\; \frac{(f * \nabla green)(\theta)}{p_t(\theta)}$$

- Not nice !

Candidate solutions

Perhaps the ideal solution is one of smallest L_2 norm of $G_t(\theta)$

- It has $G_t(\theta) = \nabla b_t(\theta)$ equal to the gradient of a function
- The pde in log density form

$$\nabla^T G_t(\theta) \,+\, G_t^T(\theta)\,\nabla \log p_t(\theta) \,=\, \frac{\partial}{\partial t} \log p_t(\theta)$$

  then becomes an elliptic pde in $b_t(\theta)$ for fixed t.
- With $\nabla \log p_t(\theta)$ and $\frac{\partial}{\partial t} \log p_t(\theta)$ arranged to be bounded, the solution may exist and be nice.
- But explicit solution to this elliptic pde is not available (except perhaps numerically in low dim cases).

Candidate solutions

Ideal solution of smallest L_2 norm of $G_t(\theta)$

- It has $G_t(\theta) = \nabla b_t(\theta)$ equal to the gradient of a function
- The pde in log density form

$$\nabla^T G_t(\theta) + G_t^T(\theta)\,\nabla \log p_t(\theta) = \frac{\partial}{\partial t} \log p_t(\theta)$$

  then becomes an elliptic pde in $b_t(\theta)$ for fixed t.
- With $\nabla \log p_t(\theta)$ and $\frac{\partial}{\partial t} \log p_t(\theta)$ arranged to be bounded, the solution may exist and be nice.
- But explicit solution to this elliptic pde is not available (except perhaps numerically in low dim cases)
- To achieve explicit solution give up $G_t(\theta)$ being a gradient
- For ridge bases, we decompose into a system of first order differential equations and integrate

Candidate solution by decomposition of ridge sum

- Optimize $J(\theta) = \sum_{i=1}^{n} r_i\, \phi(X_i^T\theta)$
- Target density $p_t(\theta) = e^{tJ(\theta) - c_t}\, p_0(\theta)$ with $c_t' = E_{p_t}[J]$
- The time score is $\frac{\partial}{\partial t} \log p_t(\theta) = J(\theta) - E_{p_t}[J]$
- Specialize the pde in log density form

$$\nabla^T G_t(\theta) \;+\; G_t^T(\theta)\, \nabla \log p_t(\theta) \;=\; J(\theta) - E_{p_t}[J]$$

- The right side takes the form of a sum

$$\sum r_i \,[\phi(X_i^T\theta) - a_i].$$

- Likewise $\nabla \log p_t(\theta) = t\, \nabla J(\theta) + \nabla \log p_0(\theta)$ is a sum

$$t \sum r_i\, X_i\, \phi'(X_i^T\theta).$$

- Here we surpress the role of the prior. It can be accounted by appending d prior observations with columns of the identity as extra input vectors along with a multiple of the score of the marginal of the prior in place of ϕ'.

Approximate solution for ridge sums

- Seek approximate solution of the form

$$G_t(\theta) = \sum \frac{x_i}{\|x_i\|^2} g_i(\underline{u})$$

with $\underline{u} = (u_1, \ldots, u_n)$ evaluated at $u_i = X_i^T \theta$, for which

$$\nabla^T G_t(\theta) = \sum_i \frac{\partial}{\partial u_i} g_i(\underline{u}) + \sum_{i,j:i \neq j} \frac{x_i^T x_j}{\|x_i\|^2} \frac{\partial}{\partial u_j} g_i(\underline{u})$$

- Can we ignore the coupling in the derivative terms?
- $x_j^T x_i / \|x_i\|^2$ are small for uncorrelated designs, large d.
- Match the remaining terms in the sums to solve for $g_i(\underline{u})$
- Arrange $g_i(\underline{u})$ to solve the differential equations

$$\frac{\partial}{\partial u_i} g_i(\underline{u}) \; + \; t\, g_i(\underline{u}) \big[r_i \phi'(u_i) + rest_i \big] \; = \; r_i \big[\phi(u_i) - a_i \big]$$

where $rest_i = \sum_{j \neq i} r_j \, \phi'(u_j) x_j^T x_i / \|x_i\|^2$.

Integral form of solution

- Differential equation for $g_i(u_i)$, suppressing dependence on the coordinates other than i

$$\frac{\partial}{\partial u_i} g_i(u_i) \;+\; t\, g_i(u_i)\left[r_i \phi'(u_i) + rest_i\right] \;=\; r_i\left[\phi(u_i) - a_i\right]$$

- Define the density factor

$$m_i(u_i) = e^{t\, r_i\, \phi(u_i) + t\, u_i\, rest_i}$$

- Allows the above diff equation to be put back in the form

$$\frac{\partial}{\partial u_i}\left[g_i(u_i)\, m_i(u_i)\right] \;=\; r_i\left[\phi(u_i) - a_i\right] m_i(u_i)$$

- An explicit solution, evaluated at $u_i = x_i^T \theta$, is

$$g_i(u_i) \;=\; r_i\, \frac{\int_{c_i}^{u_i} m_i(\tilde{u}_i)\left[\phi(\tilde{u}_i) - a_i\right] d\tilde{u}_i}{m_i(u_i)}$$

where c_i is such that $\phi(c_i) = a_i$.

The derived change function G_t for evolution of θ_t

- Iinclude the u_j for $j \neq i$ upon which $rest_i$ depends. Our solution is
$$g_{i,t}(\underline{u}) = r_i \int_{c_i}^{u_i} e^{t\, r_i\, (\phi(\tilde{u}_i) - \phi(u_i)) + t(\tilde{u}_i - u_i) rest_i(\underline{u})} \left[\phi(\tilde{u}_i) - a_i\right] d\tilde{u}_i$$
- Evaluating at $\underline{u} = X\theta$ we have the change function
$$G_t(\theta) = \sum \frac{x_i}{\|x_i\|^2} g_{i,t}(X\theta)$$
for which θ_t evolves according to
$$\theta_{t+h} = \theta_t + h\, G_t(\theta_t)$$
- For showing $g_{i,t}$, G_t and ∇G_t are nice, assume the activation function ϕ and its derivative is bounded (e.g. a logistic sigmoid or a sinusoid).
- Run several optimization paths in parallel, starting from independent choices of θ_0. Allows access to empirical computation of $a_{i,t} = E_{p_t}[\phi(x_i^T \theta_t)]$

Conjectured conclusion

Derived the desired optimization procedure and the following.

Conjecture: With step size h of order $1/n^2$ and a number of steps of order $n\, d \log d$ and $X_1, X_2, \ldots, X_n$ i.i.d. Normal$(0, I)$ in R^d, and a product of independent standard Cauchy prior $p_0(\theta)$. *With high probability on the design X, the above procedure produces optimization paths θ_t whose distribution closely tracks the target*

$$p_t(\theta) = e^{t\, J(\theta) - c_t}\, p_0(\theta)$$

such that, with high probability, the solutions paths have instances of $J(\theta_t)$ which are at least $1/2$ the maximum.

Consequently, the relaxed greedy procedure is computationally feasible and achieves the indicated bounds for sparse linear combinations from the dictionary $\Phi = \{\phi(\theta^T x) : \theta \in R^d\}$

summary

- Flexible approximation models
  - Subset selection
  - Nonlinearly parameterized bases as with neural nets
  - ℓ_1 control on coefficients of combination
- Accurate approximation with moderate number of terms
  - Proof analogous to random coding
- Information theoretic risk bounds
  - Based on the minimum description length principle
  - Shows accurate estimation with a moderate sample size
- Computational challenges are being addressed
  - Adaptive annealing strategy appears to be promising

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):

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

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 $\theta, \theta' \in S$:

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

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)

Information Thy Formulation of Statistical Principle

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

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

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

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 $\sum_q 2^{-L(q)} \le 1$.
- 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 $\mathcal{L}(q) = 2L(q)$:

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

- 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?

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

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 $\text{Prob}\{\hat{u} \neq u\}$
  and small $\text{Prob}\{\textit{Fraction mistakes} \geq \alpha\}$

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 $\frac{1}{n}\sum_{i=1}^{n} x_i^2 \leq P$

$$\downarrow$$

- 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$\{\textit{Fraction mistakes} \geq \alpha\}$

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: $\beta = (00*0000000000*00\ldots0*000000)^T$
- 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: $\beta = (00*0000000000*00\ldots0*000000)^T$
- 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$\{\mathit{Fraction}\ \hat{\beta}\ \mathit{mistakes} \geq \alpha\}$, 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?

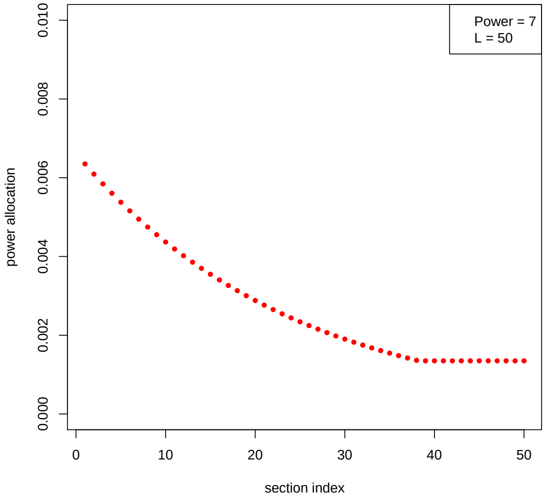
Partitioned Superposition Code

- 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 $\text{Prob}\{\textit{Fraction } \hat{\beta} \textit{ mistakes} \geq \alpha\}$
- 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

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

- 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

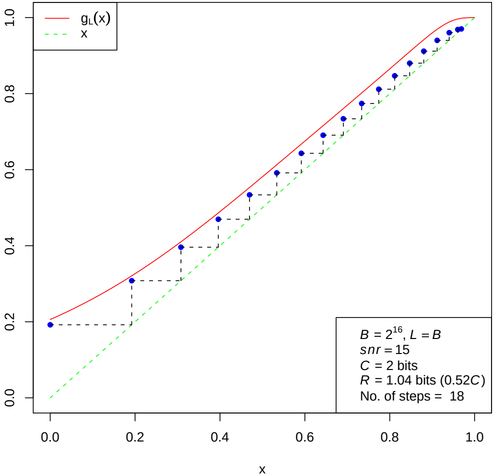


Figure : Plot of likely progression of weighted fraction of correct detections $\hat{q}_{1,k}$, for $snr = 15$.

Decoding Progression

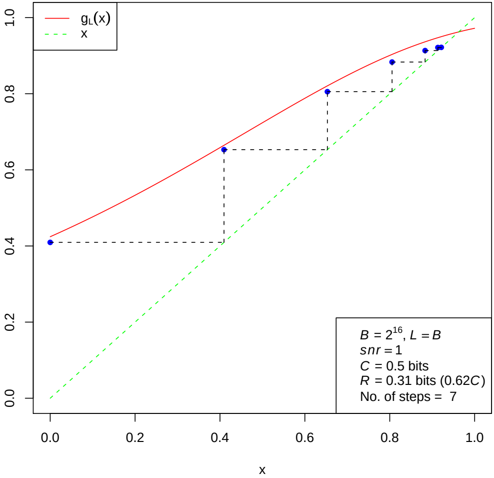


Figure : Plot of of likely progression of weighted fraction of correct detections $\hat{q}_{1,k}$, for $snr = 1$.

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 $R \leq C_B$

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

- 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

Outline

Information and Probability:

- Monotonicity of Information
- Markov Chains
- Martingales
- Large Deviation Exponents
- Information Stability (AEP)
- Central Limit Theorem
- Monotonicity of Information
- Entropy Power Inequalities

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 $\quad X \to X'$

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

- Chain Rule

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

- 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 $\quad X \to X'$

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

- Chain Rule

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

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

- 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 $\quad X \to X'$

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

- Chain Rule

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

- 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 $\quad X \to X'$

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

- Chain Rule

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

- 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 $\quad X \to X'$

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

- Chain Rule

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

- 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

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

- 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

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

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

  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

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

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

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

- 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

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

- 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)

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

- 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.
- (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
- 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)