Information Theory and High-Dimensional Bayesian Computation

The Blessing of Dimensionality

Andrew R. Barron

Yale University

Department of Statistics and Data Science

Joint work with Curtis McDonald (Yale)

Shannon Lecture

IEEE International Symposium on Information Theory

Athens, Greece, 11 July 2024

You may access these slides now at

stat.yale.edu/∼arb4/ShannonLecture.pdf

Outline

- Some of the computational core of Information Theory
  - Shannon-arithmetic codes for univ. data compression & algorithms for predictive distributions
  - Encoding and decoding for reliable communication, at rate near the Shannon capacity
- Average-case optimality or minimax optimality requires Bayes computation
- Historical roots of Laplace and Gauss

  From Laplace to modern prediction and compression: discrete data

  From Gauss to modern prediction and learning: continuous data
- Information-theoretic determination of performance

  Essential ingredients: Approximation, Estimation, and Computation
- Information theory of sampling log-concave posterior densities
- Beyond Log-Concavity
  - Provably Fast Sparse Regression Codes achieving Shannon capacity for the Gaussian channel
  - Provably Fast Posterior Sampling for neural net posterior distributions in sufficiently high dimensions
- Log concave coupling for sampling neural net posteriors

Shannon Optimal-length Arithmetic Codes for Data Compression

- For alphabetical or numerical $Y^N = (Y_1, ..., Y_N)$ modeled with a distribution $p(Y^N)$ with access to the predictive distributions $p(Y_{n+1}|Y^n)$ for $n < N$
- Shannon codelength: $\log 1/p(Y^N)$ (rounded up to an integer)
- Practical arithmetic coding achieving within 1 bit of Shannon codelength
  Shannon-Fano-Elias, Gilbert-Moore 59, Jelinek 68, Pasco 76, Rissanen 76
- The code-bits equal the binary-represented cumulative distribution function using the half-way point at its jump at Y^N to $\lceil \log 1/p(Y^N) \rceil + 1$ bits of accuracy
- The code-bits are computed recursively updating the cumulative distribution, from n to $n + 1$, using the predictive distributions

Shannon Optimal-length Arithmetic Codes for Data Compression

- For alphabetical or numerical $Y^N = (Y_1, ..., Y_N)$ modeled with a distribution $p(Y^N)$ with access to the predictive distributions $p(Y_{n+1}|Y^n)$ for $n < N$
- Shannon codelength: $\log 1/p(Y^N)$ (rounded up to an integer)
- Practical arithmetic coding achieving within 1 bit of Shannon codelength
  Shannon-Fano-Elias, Gilbert-Moore 59, Jelinek 68, Pasco 76, Rissanen 76
- The code-bits equal the binary-represented cumulative distribution function using the half-way point at its jump at Y^N to $\lceil \log 1/p(Y^N) \rceil + 1$ bits of accuracy
- The code-bits are computed recursively updating the cumulative distribution, from n to $n+1$, using the predictive distributions
- Coding for dependence of Y^N on given inputs $X^N = (X_1, ..., X_N)$: The code of length $\log 1/p(Y^N|X^N)$ uses predictive distributions $p(y|x, Y^n, X^n)$ evaluated at $X_{n+1} = x$ and $Y_{n+1} = y$
- Want predictive density estimates and compression for wide range of linear & nonlinear models

Shannon-Arithmetic Codes for Universal Data Compression

Realistic and practical data compression arises in the universal source coding context

- Parameters θ of the distribution $p(Y^N|\theta)$ not known, but can be modeled
- Redundancy is the difference in expected codelength with and without knowledge of the parameters, divided by N to get redundancy as a rate
- The one or two bits of difference from $\log 1/p(Y^N)$ are ignored, as they contribute negligibly to the redundancy rate
- For parameters modeled probabilistically, the average-case optimal codes use
$$p(Y^N) = \int p(Y^N|\theta)\, p_0(\theta)\, d\theta$$
to construct the Huffman code, or, preferably, the Shannon-arithmetic code

Shannon-Arithmetic Codes for Universal Data Compression

Realistic and practical data compression arises in the universal source coding context

- Parameters θ of the distribution $p(Y^N|\theta)$ not known, but can be modeled
- Redundancy is the difference in expected codelength with and without knowledge of the parameters, divided by N to get redundancy as a rate
- The one or two bits of difference from $\log 1/p(Y^N)$ are ignored, as they contribute negligibly to the redundancy rate
- For parameters modeled probabilistically, the average-case optimal codes use

  $$p(Y^N) = \int p(Y^N|\theta)\, p_0(\theta)\, d\theta$$

  to construct the Huffman code, or, preferably, the Shannon-arithmetic code
- The average redundancy is the Shannon mutual information $I(\theta; Y^N)$
- And the minimax redundancy is the capacity of the channel $\theta \to Y^N$
- Practical optimal-redundancy codes require computation of predictive distributions

  $$p(Y_{n+1}|Y^n) = \int p(Y_{n+1}|Y^n, \theta)\, p(\theta|Y^n)\, d\theta$$

Equivalence of Statistical Learning and Universal Data Compression

- The redundancy of a code takes the form of the Kullback divergence

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

- Chain rule of probability $p(Y^N) = \prod_{n=0}^{N-1} p(Y_{n+1}|Y^n)$ yields the chain rule of information theory

$$D(P_{Y^N|\theta}||P_{Y^N}) = \sum_{n=0}^{N-1} E_{Y^n|\theta}\left[D(P_{Y_{n+1}|Y^n,\theta}||P_{Y_{n+1}|Y^n})\right]$$

- Consider the case that the model makes $Y_1, \ldots, Y_N$ conditionally i.i.d. given θ
- Predictive $p(y|Y^n)$ at $Y_{n+1}=y$ is average-case optimal estimator of $p(y|\theta)$ with Kullback loss

Equivalence of Statistical Learning and Universal Data Compression

- The redundancy of a code takes the form of the Kullback divergence

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

- Chain rule of probability $p(Y^N) = \prod_{n=0}^{N-1} p(Y_{n+1}|Y^n)$ yields the chain rule of information theory

$$D(P_{Y^N|\theta}||P_{Y^N}) = \sum_{n=0}^{N-1} E_{Y^n|\theta}\left[D(P_{Y_{n+1}|Y^n,\theta}||P_{Y_{n+1}|Y^n})\right]$$

- Consider the case that the model makes $Y_1, \ldots, Y_N$ conditionally i.i.d. given θ
- Predictive $p(y|Y^n)$ at $Y_{n+1} = y$ is average-case optimal estimator of $p(y|\theta)$ with Kullback loss
- The Cesàro average of the risk with Kullback loss equals the redundancy rate

$$\frac{1}{N}\sum_{n=0}^{N-1} E_{Y^n|\theta}\left[D(P_{Y|\theta}||P_{Y|Y^n})\right] = \frac{1}{N}D(P_{Y^N|\theta}||P_{Y^N})$$

- Statistical learning and universal data compression have the same computational challenge: For suitable models $p(Y^n|\theta)$ and $p(\theta)$, find a procedure to compute the predictive distributions

$$p(y|Y^n) = \int p(y|\theta)\, p(\theta|Y^n)\, d\theta$$

Index of Resolvability

A simple tool for exploring the quality of a mixture $p(Y^N) = \int p(Y^N|\theta)p_0(\theta)d\theta$.

Examine the redundancy rate (i.e. Cesàro average of the risk with Kullback loss) as follows

$$R_N(\theta^*) = \frac{1}{N}D(P_{Y^N|\theta^*}||P_{Y^N}) = \frac{1}{N}E_{Y^N|\theta^*}\left[\log \frac{p(Y^N|\theta^*)}{\int p(Y^N|\theta)p_0(\theta)d\theta}\right]$$

Get the index of resolvability by restricting the integral to a Kullback neighborhood of θ^*

$$\leq \frac{1}{N}E_{Y^N|\theta^*}\left[\log \frac{p(Y^N|\theta^*)}{\int_A p(Y^N|\theta)p_0(\theta)d\theta}\right]$$

$$= D_A + \frac{1}{N}\log \frac{1}{P_0(A)}$$

where the $A = A_r = \{\theta : D(\theta^*||\theta) \leq r\}$ is the neighborhood of Kullback radius r, the $P_0(A)$ is its prior probability, the $D_A \leq r$ is the Kullback divergence from the mixture conditional on A and r is adjusted to suitably balance or optimize it.

Implications (Ba 87,98, Yang, Ba 99):

- Consistency: $R_N(\theta^*) \to 0$ for any θ^* whose Kullback neighborhoods are given positive prior probability
- Parametric rate: $R_N(\theta^*) \sim \frac{d}{2N}\log N$ in any smooth finite-dim family with positive prior density at θ^*
- Non-parametric rates: Information theory determines the minimax rates (Yang, Ba 99)
- Applicable to flexibly high-dimensional models such as neural nets (as we shall see)

Settings with Practical Predictive Distributions

For suitable models $p(Y^n|\theta)$ and $p(\theta)$, find practical procedures to compute the predictive distributions $p(y|Y^n) = \int p(y|\theta)\, p(\theta|Y^n)\, d\theta$

We discuss several settings:

- Discrete memoryless sources
- Markov models and variable order (context tree) models
- General smooth parametric families
- Location families for the normal and other log-concave error distributions
- Linear models with the normal and other log-concave error distributions
- Regression codes for achieving capacity in additive Gaussian noise channels
- Nonlinear models such as single hidden-layer neural networks

Computation of optimal procedures in such models has roots in work of Laplace & Gauss

New computational innovations are based on log-concave sampling and beyond

Aside: Remark Concerning Continuous-Valued Models

Models based on probability density functions allow nearly continuous-valued data:

- Numerical data is often modeled as discretized real data to accuracy 2^{-b} (that is to b bits accuracy, with large b)
- When large, b has little effect on the discretized redundancy, because the redundancy depends on the ratio of probabilities, near the density ratio
- Also, the supremum of redundancies over discretizations equals the Kullback divergence between the densities

The Kullback divergence for densities remains an appropriate redundancy measure

Historical Highlight: Bayes

Bayes (1763):

- Rule for reversing conditional probability: $P(A|B) = P(A)P(B|A)/P(B)$
- Provided notions of prior and posterior probability

Examined Binomial counts with uniform prior

- Found that the resulting marginal distribution on the counts is uniform on $\{0, 1, 2, ..., n\}$
- However, he was not able to compute the posterior predictive distribution.
  He did not see the solution by a rule of succession

Historical Computational Highlights: Laplace

Laplace (1774) Calculus of Probability. Commentary and translation by Stigler (1986)

- Exact computation, for discrete memoryless sources, of the key ingredients
  - The predictive distrib $p(y_{n+1}|y_1, ..., y_n)$
  - The joint distribution $p(y_1, ..., y_n) = \int p(y_1, ..., y_n|\theta)p(\theta)d\theta$
  - The posterior density $p(\theta|y^n) = p(y^n|\theta)p(\theta)/p(y^n)$
- Approximate computation, for general smooth families, by integration using a normal
  - Central limit theory for posterior distributions

Historical Computational Highlights: Laplace

Laplace (1774) Calculus of Probability. Commentary and translation by Stigler (1986)

- Exact computation, for discrete memoryless sources, of the key ingredients
  - The predictive distrib $p(y_{n+1}|y_1, ..., y_n)$
  - The joint distribution $p(y_1, ..., y_n) = \int p(y_1, ..., y_n|\theta)p(\theta)d\theta$
  - The posterior density $p(\theta|y^n) = p(y^n|\theta)p(\theta)/p(y^n)$
- Approximate computation, for general smooth families, by integration using a normal
  - Central limit theory for posterior distributions
- Decision Theory for location models and linear models
  - Median of posterior minimizes expected absolute deviation
  - Two-sided exponential error distribution
  - Could not compute posterior median except when $n \leq 3$
  - Fall-back choice of sample median recognized as suboptimal

Laplace (1810, 1812)

- Central limit theory for sums of independent random variables
- A many-causes justification of least squares for linear models
- Normal error distrib. allows computation of posterior mean, optimizes expected posterior loss

Laplace's Prediction Rule based on Count Data

Certain priors on probabilities θ in the simplex $\{\theta : \theta_j \geq 0, \sum_{j=1}^m \theta_j = 1\}$

- permit exact predictive distribution computation
- allowing computation for arithmetic codes

For discrete memoryless sources with m symbols (Laplace 1774 used $m = 2$)

- Laplace 1774. Uniform prior yields computation by Laplace's rule of succession

  $\hat{p}_n(y) = p(y_{n+1} = y | y_1, \ldots, y_n) = \frac{n_y + 1}{n + m}$ from counts $n_y = \sum_{i=1}^n 1_{\{y_i = y\}}$

  Laplace joint distribution $p(y_1, \ldots, y_N) = \frac{1}{\binom{N+m-1}{m-1}} \frac{1}{\binom{N}{N_1 \ldots N_m}}$

  It gives the average-case optimal code for uniform prior (Gilbert 71, Cover 72, 73)

  Risk bound for Kullback loss (Ba 86): $E\left[D(p||\hat{p}_n)\right] \leq \log\left(1 + \frac{m}{n}\right) \leq \frac{m}{n}$

Laplace's Prediction Rule based on Count Data

Certain priors on probabilities θ in the simplex $\{\theta : \theta_j \geq 0, \sum_{j=1}^m \theta_j = 1\}$

- permit exact predictive distribution computation
- allowing computation for arithmetic codes

For discrete memoryless sources with m symbols (Laplace 1774 used $m{=}2$)

- Laplace 1774. Uniform prior yields computation by Laplace's rule of succession

  $\hat{p}_n(y) = p(y_{n+1} = y|y_1, ..., y_n) = \frac{n_y+1}{n+m}$ from counts $n_y = \sum_{i=1}^n 1_{\{y_i=y\}}$

  Laplace joint distribution $p(y_1, ..., y_N) = \frac{1}{\binom{N+m-1}{m-1}} \frac{1}{\binom{N}{N_1 \ldots N_m}}$

  It gives the average-case optimal code for uniform prior (Gilbert 71, Cover 72, 73)

  Risk bound for Kullback loss (Ba 86): $E\left[D(p||\hat{p}_n)\right] \leq \log\left(1 + \frac{m}{n}\right) \leq \frac{m}{n}$

- Dirichlet($\lambda, ..., \lambda$) prior (originally in Laplace 1781) produces the prediction rule $\frac{n_y+\lambda}{n+m\lambda}$

  Distinguished choice $\lambda = 1/2$

  - Asymtotically capacity-achieving, providing minimax redundancy
  - Krichevski, Trofimov 81: Redundancy rate $\frac{m-1}{2N} \log N + O(\frac{1}{N})$
  - Xie, Ba 97,00: Minimax redundancy & regret $\frac{m-1}{2N} \log \frac{N}{2\pi} + \frac{1}{N} \log \int |I(\theta)|^{1/2} d\theta + o(\frac{1}{N})$

Prediction and Compression for Sources with Memory

For discrete Markov sources: Takeuchi, Kawabata, Ba 02

- Evaluates the asymtotically capacity-achieving Jeffreys prior achieving minimax redundancy
- again redundancy rate equals $\frac{d}{2N} \log N + \frac{C}{N} + o\left(\frac{1}{N}\right)$ where $d =$ parameter dimension

Prediction and Compression for Sources with Memory

For discrete Markov sources: Takeuchi, Kawabata, Ba 02

- Evaluates the asymtotically capacity-achieving Jeffreys prior achieving minimax redundancy
- again redundancy rate equals $\frac{d}{2N} \log N + \frac{C}{N} + o\left(\frac{1}{N}\right)$ where $d =$ parameter dimension

For variable order Markov sources: Willems, Shtarkov, Tjalkens 95

- recursive Context Tree Weighting (CTW) algorithm
- Optimal prediction, compression, text generation for their prior & posterior

Scaling-up CTW at the word level, with access to massive amounts of text data, should yield a competitive, stochastically-optimal, large language model

Laplace Approximation for Posterior and Bayes Factor

For general smooth parametric families

- Laplace Approximation of the Posterior

  from second order Taylor expansion of log density with empirical Fisher information $\hat{I}$

  $$p(Y^n|\theta)\, p_0(\theta) \quad \sim \quad p(Y^n|\hat{\theta})\, p_0(\hat{\theta}) \quad \exp\{-\tfrac{1}{2} n \hat{I}\, (\theta - \hat{\theta})^2\}$$

  yields approximate normality of the posterior

Laplace Approximation for Posterior and Bayes Factor

For general smooth parametric families

- Laplace Approximation of the Posterior

  from second order Taylor expansion of log density with empirical Fisher information $\hat{I}$

$$p(Y^n|\theta)\,p_0(\theta) \quad \sim \; p(Y^n|\hat{\theta})\,p_0(\hat{\theta}) \quad \exp\{-\tfrac{1}{2} n \hat{I}\,(\theta-\hat{\theta})^2\}$$

  yields approximate normality of the posterior

- Integrating it yields the Laplace Approximation of the Joint Distribution, Bayes factor

$$\int p(Y^n|\theta)\,p_0(\theta)\,d\theta \;\sim\; p(Y^n|\hat{\theta})\,p_0(\hat{\theta}) \int \exp\{-\tfrac{1}{2} n \hat{I}\,(\theta-\hat{\theta})^2\}\,d\theta$$
$$= p(Y^n|\hat{\theta})\,p_0(\hat{\theta}) \left(\frac{2\pi}{n^d|\hat{I}|}\right)^{1/2}$$

Laplace Approximation for Posterior and Bayes Factor

For general smooth parametric families

- Laplace Approximation of the Posterior

  from second order Taylor expansion of log density with empirical Fisher information $\hat{I}$

  $$p(Y^n|\theta)\,p_0(\theta) \quad \sim \; p(Y^n|\hat{\theta})\,p_0(\hat{\theta}) \quad \exp\{-\tfrac{1}{2}n\,\hat{I}\,(\theta-\hat{\theta})^2\}$$

  yields approximate normality of the posterior

- Integrating it yields the Laplace Approximation of the Joint Distribution, Bayes factor

  $$\int p(Y^n|\theta)\,p_0(\theta)\,d\theta \;\sim\; p(Y^n|\hat{\theta})\,p_0(\hat{\theta})\int \exp\{-\tfrac{1}{2}n\,\hat{I}\,(\theta-\hat{\theta})^2\}\,d\theta$$

  $$= p(Y^n|\hat{\theta})\,p_0(\hat{\theta})\left(\frac{2\pi}{n^d|\hat{I}|}\right)^{1/2}$$

- Taking logs yields the pointwise regret of stochastic complexity, MDL

  Ba 85, Clarke, Ba 90,94, Rissanen 96, Takeuchi, Ba 24

  $$\frac{1}{N}\log\frac{p(Y^n|\hat{\theta})}{\int p(Y^n|\theta)p_0(\theta)d\theta} = \frac{d}{2n}\log\frac{n}{2\pi} + \frac{1}{n}\log\frac{|\hat{I}(\hat{\theta})|^{1/2}}{p_0(\hat{\theta})} + o\left(\frac{1}{n}\right)$$

Kullback Risk and Data Compression

Continuing for general smooth parametric families with i.i.d. observations

Taking the expected value yields the redundancy of data compression, equivalently, it is the cumulative Kullback risk for sample sizes $n \leq N$ (Clarke, Ba 90,94)

$$\frac{1}{N} D(P_{Y^N|\theta} || P_{Y^N}) = \frac{d}{2N} \log \frac{N}{2\pi e} + \frac{1}{N} \log \frac{|I(\theta)|^{1/2}}{p_0(\theta)} + o\left(\frac{1}{N}\right)$$

Jeffreys prior $p_0(\theta)$ proportional to $|I(\theta)|^{1/2}$

- Approximately mimimax for total Kullback risk and redundancy, (Clarke, Ba 94)
- Approximately capacity-achieving, maximizing $I(\theta; Y^N)$ asymptotically (Bernardo 79, Ibragimov, Hasminskii 73, Clarke, Ba 94)
- Hartigan 64: Jeffreys prior equalizes probability of small Kullback balls of given radius

Kullback Risk and Data Compression

Continuing for general smooth parametric families with i.i.d. observations

Taking the expected value yields the redundancy of data compression, equivalently, it is the cumulative Kullback risk for sample sizes $n \le N$ (Clarke, Ba 90,94)

$$\frac{1}{N} D(P_{Y^N|\theta} || P_{Y^N}) = \frac{d}{2N} \log \frac{N}{2\pi e} + \frac{1}{N} \log \frac{|I(\theta)|^{1/2}}{p_0(\theta)} + o\left(\frac{1}{N}\right)$$

Jeffreys prior $p_0(\theta)$ proportional to $|I(\theta)|^{1/2}$

- Approximately mimimax for total Kullback risk and redundancy, (Clarke, Ba 94)
- Approximately capacity-achieving, maximizing $I(\theta; Y^N)$ asymptotically (Bernardo 79, Ibragimov, Hasminskii 73, Clarke, Ba 94)
- Hartigan 64: Jeffreys prior equalizes probability of small Kullback balls of given radius

Individual Kullback risk based on a sample of size n

- Parametric settings: (Cencov 72, Akaike 73, Yang, Ba 98, Hartigan 99), in i.i.d. case

$$E\left[D(P_{Y|\theta} || P_{Y|Y^n})\right] \sim \frac{d}{2n}$$

  Dependence on θ and on the choice of prior arise only in terms of order $(1/n)^2$
- Nonparametric settings: approximation and estimation tradeoff (Ba, Sheu 91)

$$D(P||\hat{P}_n) \sim \min_K \left\{ \left(\frac{1}{K}\right)^{2/d_0} + \frac{K}{n} \right\} \sim \left(\frac{1}{n}\right)^{2/(2+d_0)} \quad \text{in the one derivative case}$$

Historical Highlight of Bayesian Computation: Gauss

Gauss (1806 German, 1809 Latin) Treatise on Planetary Motion. English Transl. Davis (1857)

- Investigates orbit determination when there are multiple observations
- Linearizes smooth nonlinear dependence on parameters (per Newton)
- Linear system of equations characterizing least squares solution
  Recognized in a paper by Legendre (1805)
- Gauss elimination solution

Gauss justification of least squares as a Bayesian Computation

- For linear models $f(x_i, w) = w \cdot x_i$ with observed responses y_i, including location families, corresponding to constant $x_i = 1$

Historical Highlight of Bayesian Computation: Gauss

Gauss (1806 German, 1809 Latin) Treatise on Planetary Motion. English Transl. Davis (1857)

- Investigates orbit determination when there are multiple observations
- Linearizes smooth nonlinear dependence on parameters (per Newton)
- Linear system of equations characterizing least squares solution
  Recognized in a paper by Legendre (1805)
- Gauss elimination solution

Gauss justification of least squares as a Bayesian Computation

- For linear models $f(x_i, w) = w \cdot x_i$ with observed responses y_i, including location families, corresponding to constant $x_i = 1$
- Given a density $\phi(z)$ for deviations with score $s(z) = -\phi'(z)/\phi(z)$
- The posterior density $p(w|Data)$ is proportional to the joint density function
  $$\phi(y_1 - w \cdot x_1) \ldots \phi(y_n - w \cdot x_n)$$
- The mode $\hat{w}$ of the posterior distribution is found by solving the system of equations
  $$\sum_{i=1}^{n} s(y_i - w \cdot x_i)\, x_i = 0$$
- Gauss' density $\phi(z)$ with linear score provides the desired linear system of equations
- Accordingly the least squares solution is the posterior mode
- Moreover, if the posterior mode is a linear function of y, then $\phi(z)$ must be the Gaussian

From Laplace and Gauss to Modern Bayesian Computation

Laplace & Gauss work for linear models and location families is celebrated

- for providing computation of the posterior optimal solutions for Gaussian ϕ
- for providing the predictive densities $p(y|x, Data)$, the predictive means, and the Bayes factors
- and Gauss' recursive least squares solution, which iterates one observation at a time

From Laplace and Gauss to Modern Bayesian Computation

Laplace & Gauss work for linear models and location families is celebrated

- for providing computation of the posterior optimal solutions for Gaussian ϕ
- for providing the predictive densities $p(y|x, Data)$, the predictive means, and the Bayes factors
- and Gauss' recursive least squares solution, which iterates one observation at a time

Linear Filtering and Prediction

- Kalman (1960) theory extends recursive posterior predictive computation to the setting of linear difference equation evolution of the states x_n

Model Selection & Data Compression: for Gaussian ϕ, compute Bayes factors & MDL stochastic complexity

- Evaluating $p(Y^n|X^n) = \int p(Y^n|X^n, w)p(w)dw$ and associated predictive densities
- Permits optimal arithmetic coding of finely discretized observations
- Related to linear predictive coding

From Laplace and Gauss to Modern Bayesian Computation

Laplace & Gauss work for linear models and location families is celebrated

- for providing computation of the posterior optimal solutions for Gaussian ϕ
- for providing the predictive densities $p(y|x, Data)$, the predictive means, and the Bayes factors
- and Gauss' recursive least squares solution, which iterates one observation at a time

Linear Filtering and Prediction

- Kalman (1960) theory extends recursive posterior predictive computation to the setting of linear difference equation evolution of the states x_n

Model Selection & Data Compression: for Gaussian ϕ, compute Bayes factors & MDL stochastic complexity

- Evaluating $p(Y^n|X^n) = \int p(Y^n|X^n, w)p(w)dw$ and associated predictive densities
- Permits optimal arithmetic coding of finely discretized observations
- Related to linear predictive coding

Minimax Estimation and Compression for linear models, for general ϕ

The uniform prior yields minimax optimality for

- parameter estimation with squared error loss (Girshick, Savage 51)
- predictive density estimation with Kullback risk (Liang, Ba 02)
- data compression with minimax redundancy (Liang, Ba 02)

Is there a class of ϕ for which we have feasible Bayes computation in these settings?

From Gaussian to Log-Concave Distributions

- Summary thus far: Laplace & Gauss performed the required normal integrations in linear models with normal errors to compute the posterior-optimal procedures
- What is the right extension to non-normal error distributions to preserve rapid computation of high-dimensional posterior integrals?

From Gaussian to Log-Concave Distributions

- Summary thus far: Laplace & Gauss performed the required normal integrations in linear models with normal errors to compute the posterior-optimal procedures
- What is the right extension to non-normal error distributions to preserve rapid computation of high-dimensional posterior integrals?
- Answer emerging in the last forty years: Log-Concavity
  Permits MCMC samplers of the posterior: Accurate and mix rapidly for log concave posteriors
- Rapid computation of minimax optimal procedures in settings with log-concave error distributions for:
  - location estimation
  - linear regression
  - minimax redundancy compression in linear predictive models

  The optimal procedures in these settings become polynomial-time computable
- Important settings that are not log-concave:
  - regressions with non-convex domains
  - non-linear regressions, such as neural networks

Information Theory of Rapid MCMC with Log Concavity

- Langevin Diffusion Path for sample parameter values w_t

$$d\,w_t = \frac{1}{2}\nabla \log p(w_t)\,dt + d\,B_t$$

  - Score $\nabla \log p(w)$ is non-linear in general
  - There are time-discretizations (e.g. Metropolis adjusted Langevin) with similar mixing processing
  - Initialize with w_0 distributed $N(0,(1/\rho);I)$ or initialize using the Laplace approximation
- Theory of Bakry-Emery 85, see Bakry, Gentil, Ledoux 14
  Strong log concavity yields rapid Markov process convergence

Information Theory of Rapid MCMC with Log Concavity

- Langevin Diffusion Path for sample parameter values w_t

$$d\,w_t = \tfrac{1}{2}\nabla \log p(w_t)\,dt \,+\, d\,B_t$$

  - Score $\nabla \log p(w)$ is non-linear in general
  - There are time-discretizations (e.g. Metropolis adjusted Langevin) with similar mixing processing
  - Initialize with w_0 distributed $N(0,(1/\rho);I)$ or initialize using the Laplace approximation
- Theory of Bakry-Emery 85, see Bakry, Gentil, Ledoux 14
  Strong log concavity yields rapid Markov process convergence
- In particular, in the stochastic diffusion setting, if for $\rho > 0$

$$\nabla\nabla' \log 1/p(w) \geq \rho\, I$$

  yields exponential convergence of relative entropy (Kullback divergence)

$$D(p_t||p) \leq e^{-t\,\rho}\, D(p_0||p)$$

- The time required for small relative entropy is controlled by $\tau = 1/\rho$

Information Theory of Rapid MCMC with Log Concavity

- Langevin Diffusion Path for sample parameter values w_t

  $$d\,w_t = \frac{1}{2}\nabla \log p(w_t)\,dt \,+\, d\,B_t$$

  - Score $\nabla \log p(w)$ is non-linear in general
  - There are time-discretizations (e.g. Metropolis adjusted Langevin) with similar mixing processing
  - Initialize with w_0 distributed $N(0,(1/\rho);I)$ or initialize using the Laplace approximation
- Theory of Bakry-Emery 85, see Bakry, Gentil, Ledoux 14
  Strong log concavity yields rapid Markov process convergence
- In particular, in the stochastic diffusion setting, if for $\rho > 0$

  $$\nabla\nabla' \log 1/p(w) \geq \rho\, I$$

  yields exponential convergence of relative entropy (Kullback divergence)

  $$D(p_t||p) \leq e^{-t\,\rho}\, D(p_0||p)$$

- The time required for small relative entropy is controlled by $\tau = 1/\rho$
- Proof uses $D(p_t||p) = \frac{1}{2}\int_{\tau\geq t} J(p_\tau||p)\,d\tau$ associated with $\frac{d}{dt}D(p_t||p) = -\frac{1}{2}J(p_t||p)$ and establishes Log Sobolev Ineq: $D(p_t||p) \leq \frac{1}{2\,\rho}J(p_t||p)$ where J is mean square norm between the scores
- Similar identities in Stam (59) for entropy power inequality & log Sobolev ineq for the normal, and in Ba 86
- Central Limit Theorem of Ba 86, shows relative entropy convergence to the normal for standardized sums of i.i.d. random variables, using similar tools and the linear score target

Beyond Log-Concavity

Some important posteriors are not log-concave

- Bayes Computation for Communications
  - Capacity-achieving sparse regression codes
  - For a Gaussian noise channel
  - Codes are in a linear model Xw but with a non-convex constraint on w
  - Rapid decoders developed with Joseph, Cho and Rush
- Bayes Computation for Non-linear Models, including Neural Nets
  - Applies to neural nets with smooth activation functions
  - Posterior density has many peaks. It is not log-concave
  - Introducing many auxiliary random variables simplifies the sampling landscape

Bayes Computation for Communication

Communication strategy for additive Gaussian noise channel with specified power control

Capacity-achieving Sparse Regression Codes Joseph, Ba 12

- Gaussian design matrix X
- Codewords of form $X\,w$
- Non-convex constraint set W of size 2^{nC} for the weights w
  specified by a sparsity requirement of one non-zero in each of several sections and by a power allocation
- Bayes optimal decoder seeks $\min_{w \in W} ||Y - Xw||^2$

Communication strategy for additive Gaussian noise channel with specified power control

Capacity-achieving Sparse Regression Codes Joseph, Ba 12

- Gaussian design matrix X
- Codewords of form $X\,w$
- Non-convex constraint set W of size 2^{nC} for the weights w
  specified by a sparsity requirement of one non-zero in each of several sections and by a power allocation
- Bayes optimal decoder seeks $\min_{w \in W} ||Y - Xw||^2$

Computationally-feasible capacity-achieving iterative decoders

Compute weight estimates w_k iteratively, for a small (logarithmic) number of steps.
After which the estimates concentrate on the columns sent with high probability

- Adaptive Successive Hard-Decision Decoder (Joseph, Ba 14)
- Adaptive Successive Soft-Decision Decoder (Ba, Cho, 12)
  Compute w_k as posterior mean of indicators, given approx normal distributions
  of the inner products of the columns of X with residuals $Y - Xw_{k-1}$, normalized
- Approx Message Passing Decoder (Rush, Greig, Venkataramanan 17)

Sparse Regression Codes Monograph: Venkataramanan, Tatikonda, Ba 19

Artificial Neural Network Learning

A. Approximation

Squared approx error is of order $\frac{1}{K}$ with K neurons combined on last layer

B. Estimation

Squared estimation error is of order $\frac{K \log d}{N}$ with sample of size N, input dimension d

C. Computation

Computation time is a low order polynomial in N, K, d, when Kd is larger than N

A. Approximation, Neural Nets, and Function Variation with a Dictionary G

- Neural Nets and Variation with respect to a Dictionary
  - Dictionary G of functions $g(x, w)$, each bounded by 1
  - Consider linear combinations $\sum_j c_j\, g(x, w_j)$
  - G may be the class of depth $L-1$ subnetworks with control on their path weights
  - Single hidden-layer case $\sum_j c_j\, \psi(w_j \cdot x)$ with a bounded scalar activation function ψ
  - Control the sum of abs values of weights $\sum_j |c_j| \le V$
  - $\mathcal{F}_V =$ closure of signed convex hull of functions $V\, g(x, w)$
  - Variation $V(f) = V_G(f) =$ the infimum of V such that $f \in \mathcal{F}_V$.
- Approximation accuracy
  - K term approximation: $f_K(x) = \sum_{k=1}^K c_k\, g(x, w_k)$
  - Approximation error: $\|f - f_K\|^2 \le \frac{V(f)^2}{K}$ using the $L_2(P_X)$ norm squared
  - An existence proof and a Greedy approximation proof, Ba 93
  - Outer weights c_k may equal $\pm\frac{V}{K}$
  - Error better than order $\left(\frac{1}{K}\right)^{1.5}$ is NP–hard (Vu 97)
  - Rate $\frac{1}{K}$ is dimension independent

B. Neural Net Estimation and its Statistical Risk

- Via constrained least squares, penalized least squares or Bayes predictions $\hat{f}$,

  risk $\quad E[||\hat{f} - f||^2] \leq \; c\, V(f) \left(\frac{\log(2d)+L}{N}\right)^{1/2}$

  There are also lower bounds of such order (Klusowski, Ba 17)

- Computationally-feasible Bayes prediction accuracy (in the single hidden layer case)

$$E[||\hat{f} - f||^2] \leq \; c\, V(f)^{2/3} \left(\frac{\log(2d)}{N}\right)^{1/3}$$

- Both rates can be obtained by the Index of Resolvability:

$$\text{ApproxError} + \frac{1}{N} \log[\, 1 \,/\, \text{PriorProb(ApproxSet)}]$$

B. Methods of Obtaining such Statistical Risk Control

- Statistical risk or generalization squared error: $E[||\hat{f} - f||^2]$
- Five methods of controlling such statistical risk
  - Empirical process control of constrained least squares via metric entropy
    - Gaussian complexity: Ba, Klusowski 19
    - Rademacher complexity: Neshabur et al 15, Golowich et al 18
  - Penalized least squares risk control via relationship to MDL

    Adaptive bounds via an index of resolvability: Ba, Cover 90, Ba, Li et al 99, 08
  - Concentration of posterior distributions

    Necessary and sufficient conditions for posterior concentration Ba 88, 98, Ba, Shervish, Wasserman 98, Ghoshal, Ghosh, Van der Vaart 00
  - Cumulative Kullback risk of Bayes predictive distributions

    Clean information-theoretic bounds, again by an index of resolvability: Ba 87, 98, Yang, Ba 98, Ba, Klusowski 19, Ba, McDonald 24
  - Online learning regret bounds for squared error & log-loss

    Provides bounds for arbitrary data sequences
- All five have connections to information theory
- The posterior predictive procedures allow rapid computation

C. Log Concave Coupling for Bayesian Computation

- Focus on single hidden-layer network models
- Prior density $p_0(w)$: Uniform on an ℓ_1 constrained set
- Posterior $p(w)$: Multimodal. No known direct rapid sampler
- Coupling $p(\xi|w)$: conditionally independent Gaussian auxiliary variables $\xi_{i,k}$ with mean $x_i \cdot w_k$ for each observation i and neuron k
- The reverse conditional $p(w|\xi)$ is always log-concave
- The marginal $p(\xi)$ and its score $\nabla \log p(\xi)$ are rapidly computable

C. Log Concave Coupling for Bayesian Computation

- Focus on single hidden-layer network models
- Prior density $p_0(w)$: Uniform on an ℓ_1 constrained set
- Posterior $p(w)$: Multimodal. No known direct rapid sampler
- Coupling $p(\xi|w)$: conditionally independent Gaussian auxiliary variables $\xi_{i,k}$ with mean $x_i \cdot w_k$ for each observation i and neuron k
- The reverse conditional $p(w|\xi)$ is always log-concave
- The marginal $p(\xi)$ and its score $\nabla \log p(\xi)$ are rapidly computable
- $p(\xi)$ is log concave when the number of parameters $K\,d$ is large compared to N
- Langevin diffusion and other samplers are rapidly mixing
- A draw from $p(\xi)$ followed by a draw from $p(w|\xi)$ yields a draw from the desired posterior $p(w)$

C. Bayesian Computation for Neural Nets

- Data: (X_i, Y_i) for $i = 1, 2, \ldots, n$, with each X_i in $[-1, 1]^d$ and sample sizes $n \le N$
- Natural yet optional statistical assumption:
  (X_i, Y_i) independent $P_{X,Y}$, target $f(x) = E[Y|X{=}x]$, variance $\sigma_Y^2 = \sigma^2$, sub-Gaussian Y
  - Useful for motivation and for risk bounds
  - Not needed for Bayesian computation statements
  - Not needed for online learning bounds
- Single hidden-layer network model: $f(x, \underline{w})$
  $f_K(x, \underline{w}_1, \ldots \underline{w}_K) = \frac{V}{K} \sum_{k=1}^{K} \psi(\underline{w}_k \cdot x_i)$
  with each $\underline{w}_k$ in the symmetric simplex $S_1^d = \{w : \sum_{j=1}^{d} |w_j| \le 1\}$
- Prior: $p_0(\underline{w})$ makes $\underline{w}_k$ independent uniform on S_1^d
- Likelihood: $\exp\{-\beta g(w)\}$ with gain $0 < \beta \le 1/\sigma^2$
  where $g(w) = \frac{1}{2} \sum_{i=1}^{n} \left(Y_i - \frac{V}{K} \sum_{k=1}^{K} \psi(x_i \cdot w_k)\right)^2$

C. Bayesian Computation for Neural Nets

- Data: (X_i, Y_i) for $i = 1, 2, \ldots, n$, with each X_i in $[-1, 1]^d$ and sample sizes $n \le N$
- Natural yet optional statistical assumption:
  (X_i, Y_i) independent $P_{X,Y}$, target $f(x) = E[Y|X = x]$, variance $\sigma_Y^2 = \sigma^2$, sub-Gaussian Y
  - Useful for motivation and for risk bounds
  - Not needed for Bayesian computation statements
  - Not needed for online learning bounds
- Single hidden-layer network model: $f(x, \underline{w})$
  $f_K(x, \underline{w}_1, \ldots \underline{w}_K) = \frac{V}{K} \sum_{k=1}^{K} \psi(\underline{w}_k \cdot x_i)$
  with each $\underline{w}_k$ in the symmetric simplex $S_1^d = \{w : \sum_{j=1}^{d} |w_j| \le 1\}$
- Prior: $p_0(\underline{w})$ makes $\underline{w}_k$ independent uniform on S_1^d
- Likelihood: $\exp\{-\beta g(w)\}$ with gain $0 < \beta \le 1/\sigma^2$
  where $g(w) = \frac{1}{2} \sum_{i=1}^{n} \left(Y_i - \frac{V}{K} \sum_{k=1}^{K} \psi(x_i \cdot w_k)\right)^2$
- Posterior: $p(w) = p_0(w) \exp\{-\beta g(w) - \Gamma(\beta)\}$
- Bayesian Computation: Estimate $\hat{f}(x) = \int f(x, w) p(w) dw$
  by drawing independent samples from $p(w)$ and averaging $f(x, w)$

Hessian of the Minus Log Likelihood

- Log 1/Likelihood $= \beta\, g(w)$

  $$\text{Hessian} = \beta\, H(w) = \beta\, \nabla\nabla' g(w)$$

- Squared error loss: $g(w) = \frac{1}{2}\sum_{i=1}^{n}(res_i(w))^2$ where

  $$res_i(w) = Y_i - \frac{V}{K}\sum_{k=1}^{K}\psi(x_i \cdot w_k)$$

- Hessian Quadratic form: $a' H(w) a$, where a has blocks a_k

  $$\frac{V^2}{K^2}\sum_{i=1}^{n}\left(\sum_{k=1}^{K}\psi'(x_i \cdot w_k)\, a_k \cdot x_i\right)^2$$
  $$-\frac{V}{K}\sum_{i=1}^{n} res_i(w)\sum_{k=1}^{K}\psi''(x_i \cdot w_k)(a_k \cdot x_i)^2$$

Hessian of the Minus Log Likelihood

- Log 1/Likelihood $= \beta\, g(w)$

  $$\text{Hessian} = \beta\, H(w) = \beta\, \nabla\nabla' g(w)$$

- Squared error loss: $g(w) = \frac{1}{2}\sum_{i=1}^{n}(res_i(w))^2$ where

  $$res_i(w) = Y_i - \frac{V}{K}\sum_{k=1}^{K}\psi(x_i \cdot w_k)$$

- Hessian Quadratic form: $a' H(w) a$, where a has blocks a_k

  $$\frac{V^2}{K^2}\sum_{i=1}^{n}\left(\sum_{k=1}^{K}\psi'(x_i \cdot w_k)\, a_k \cdot x_i\right)^2$$
  $$-\frac{V}{K}\sum_{i=1}^{n} res_i(w)\sum_{k=1}^{K}\psi''(x_i \cdot w_k)(a_k \cdot x_i)^2$$

- This $g(w)$ is not convex, that is, $p(w)$ is not log-concave

  The first term is positive definite, the second term is not

- No clear reason for direct gradient methods to be effective

Log Concave Coupling

- Auxiliary Random Variables $\xi_{i,k}$ chosen conditionally independent Normal with mean $x_i \cdot w_k$, variance $1/\rho$, with $\rho = \beta c V/K$ restricted to ξ with each $\sum_{i=1}^{n} \xi_{i,k} x_{i,j}$ in a high probability interval
- Conditional density:

  $p(\xi|w) = (\rho/2\pi)^{Kn/2} exp\{-\frac{\rho}{2}\sum_{i=1}^{n}\sum_{k=1}^{K}(\xi_{i,k} - x_i \cdot w_k)^2\}$
- Multiplier $c = c_{Y,V} = \max_i |Y_i| + V$ bounds $|res_i(w)|$ for all w
- Activation second derivative: $|\psi''(z)| \leq 1$ for $|z| \leq 1$
- Joint density: $p(w, \xi) = p(w)p(\xi|w)$

Log Concave Coupling

- Auxiliary Random Variables $\xi_{i,k}$ chosen conditionally independent Normal with mean $x_i \cdot w_k$, variance $1/\rho$, with $\rho = \beta cV/K$ restricted to ξ with each $\sum_{i=1}^n \xi_{i,k} x_{i,j}$ in a high probability interval
- Conditional density:
  $p(\xi|w) = (\rho/2\pi)^{Kn/2} exp\{-\frac{\rho}{2}\sum_{i=1}^n \sum_{k=1}^K (\xi_{i,k} - x_i \cdot w_k)^2\}$
- Multiplier $c = c_{Y,V} = \max_i |Y_i| + V$ bounds $|res_i(w)|$ for all w
- Activation second derivative: $|\psi''(z)| \le 1$ for $|z| \le 1$
- Joint density: $p(w,\xi) = p(w)p(\xi|w)$
- Reverse conditional density: $p(w|\xi) = p_0(w)\exp\{-\beta g_\xi(w) - \Gamma_\xi(\beta)\}$
- Conditional log 1/Likelihood $= \beta g_\xi(w)$ with
  $g_\xi(w) = g(w) + \frac{1}{2}\frac{V}{K} c \sum_{i=1}^n \sum_{k=1}^K \left(x_i \cdot w_k - \xi_{i,k}\right)^2$
- Modifies Hessian $a' H_\xi(w) a$ with new positive def second term
  $\frac{V}{K}\sum_i \sum_k [c - res_i(w)\psi''(x_i \cdot w_k)](a_k \cdot x_i)^2$
- $p(w|\xi)$ is log concave in w for each ξ
- MCMC Efficient sample Applegate, Kannan 91, Lovász, Vempala 07

Marginal Density and Score of the Auxiliary Variables

- Auxiliary variable density function:

  $$p(\xi) = \int p(w,\xi)dw$$

  Integral of a log concave function of w

- Rule for Marginal Score:

  $$\nabla \log 1/p(\xi) = E[\nabla \log 1/p(\xi|w) \,|\, \xi\,]$$

- Normal Score: linear

  $$\partial_{\xi_{i,k}} \log 1/p(\xi|w) = \rho\,\xi_{i,k} \;-\; \rho\, x_i \cdot w_k$$

- Marginal Score:

  $$\partial_{\xi_{i,k}} \log 1/p(\xi) = \rho\,\xi_{i,k} \;-\; \rho\, x_i \cdot E[w_k \,|\, \xi\,]$$

Marginal Density and Score of the Auxiliary Variables

- Auxiliary variable density function:

  $p(\xi) = \int p(w,\xi)dw$

  Integral of a log concave function of w
- Rule for Marginal Score:

  $\nabla \log 1/p(\xi) = E[\nabla \log 1/p(\xi|w) \,|\, \xi\,]$
- Normal Score: linear

  $\partial_{\xi_{i,k}} \log 1/p(\xi|w) = \rho\,\xi_{i,k} \;-\; \rho\, x_i \cdot w_k$
- Marginal Score:

  $\partial_{\xi_{i,k}} \log 1/p(\xi) = \rho\,\xi_{i,k} \;-\; \rho\, x_i \cdot E[w_k \,|\, \xi\,]$
- Efficiently compute ξ score by Monte Carlo sampling of $w|\xi$
- Permits Langevin stochastic diffusion: with gradient drift

  $d\,\xi_t = \frac{1}{2}\nabla \log p(\xi_t)\, dt \;+\; d\,B_t$

  converging to a draw from the invariant density $p(\xi)$

Hessian of $\log 1/p(\xi)$. Is $p(\xi)$ log concave?

- Hessian of $\log 1/p(\xi)$, an nK by nK matrix
  $$\tilde{H}(\xi) = \nabla\nabla' \log 1/p(\xi) = \rho\left\{ I - \rho\, Cov \left[\begin{bmatrix} Xw_1 \\ \cdots \\ Xw_K \end{bmatrix} \Big| \, \xi \right] \right\}$$
- Hessian quadratic form for unit vectors a in R^{nK} with blocks a_k
  $$a' \tilde{H}(\xi) a = \rho\,\{1 - \rho\, Var[\tilde{a} \cdot w|\xi]\} \quad \text{where } \tilde{a} = \begin{bmatrix} X'a_1 \\ \cdot \\ X'a_K \end{bmatrix} \text{ has } ||\tilde{a}||^2 \leq n\, d$$
- Role for variance of $\tilde{a} \cdot w$ using the log-concave $p_\beta(w|\xi)$
- More concentrated, having smaller variance than with the prior?

Hessian of $\log 1/p(\xi)$. Is $p(\xi)$ log concave?

- Hessian of $\log 1/p(\xi)$, an nK by nK matrix
  $$\tilde{H}(\xi) = \nabla\nabla' \log 1/p(\xi) = \rho \left\{ I - \rho \, Cov \left[\begin{matrix} Xw_1 \\ \cdots \\ Xw_K \end{matrix} \Big| \xi \right] \right\}$$
- Hessian quadratic form for unit vectors a in R^{nK} with blocks a_k
  $$a' \tilde{H}(\xi) a = \rho \{ 1 - \rho \, Var[\tilde{a} \cdot w | \xi] \} \quad \text{where } \tilde{a} = \begin{bmatrix} X' a_1 \\ \cdot \\ X' a_K \end{bmatrix} \text{ has } ||\tilde{a}||^2 \le n\, d$$
- Role for variance of $\tilde{a} \cdot w$ using the log-concave $p_\beta(w|\xi)$
- More concentrated, having smaller variance than with the prior?
- Counterpart using the prior $\rho \{ 1 - \rho \, Var_0[\tilde{a} \cdot w] \}$
- Use $Cov_0(w_k) = \frac{2}{(d+2)(d+1)} I$ and $\rho = \beta c V / K$ to see its at least $\rho \{ 1 - \frac{2 \beta c V n}{K(d+2)} \}$
- Constant β chosen such that, say, $\beta c V \le 1/4$
- Strictly positive when the number of parameters Kd exceeds the sample size n
- Hessian $\ge (\rho/2) I$. Strictly log concave

Is $p(\xi)$ log concave?

- Recap: The quadratic form of the Hessian of $\log 1/p(\xi)$ is

$$a' \tilde{H}(\xi) a = \rho\, \{1 - \rho\; Var[\tilde{a} \cdot w|\xi]\}$$

- Control of the variance, dropping the mean from inside the square,

$$\rho\, Var[\tilde{a} \cdot w|\xi] \le \rho \int (\tilde{a} \cdot w)^2 \exp\{-\beta \tilde{g}_\xi(w) - \Gamma_\xi(\beta)\} p_0(w) dw$$

$$\text{using } \tilde{g}_\xi(w) = g_\xi(w) - E_0[g_\xi(w)]$$

- Hölder's inequality with $\ell \ge 1$

$$\le \rho\, [E_0[(\tilde{a} \cdot w)^{2\ell}]]^{1/\ell} \exp\{\tfrac{\ell-1}{\ell} \Gamma_\xi(\tfrac{\ell}{\ell-1}\beta) - \Gamma_\xi(\beta)\}$$

Is $p(\xi)$ log concave?

- Recap: The quadratic form of the Hessian of $\log 1/p(\xi)$ is
  $$a'\tilde{H}(\xi)a = \rho\{1 - \rho\, Var[\tilde{a}\cdot w|\xi]\}$$
- Control of the variance, dropping the mean from inside the square,
  $$\rho\, Var[\tilde{a}\cdot w|\xi] \le \rho \int (\tilde{a}\cdot w)^2 \exp\{-\beta \tilde{g}_\xi(w) - \Gamma_\xi(\beta)\} p_0(w)dw$$
  using $\tilde{g}_\xi(w) = g_\xi(w) - E_0[g_\xi(w)]$
- Hölder's inequality with $\ell \ge 1$
  $$\le \rho\, [E_0[(\tilde{a}\cdot w)^{2\ell}]]^{1/\ell} \exp\{\tfrac{\ell-1}{\ell}\Gamma_\xi(\tfrac{\ell}{\ell-1}\beta) - \Gamma_\xi(\beta)\}$$
  which is, using $|\tilde{g}_\xi(w)| \le C_V n$ with $C_V = 9V^2 + 7V \max_i |Y_i|$,
  $$\le \frac{c\beta V}{K}\frac{4n\ell}{d_0 e}\exp\{\beta C_V n/\ell\}$$
  $$= 4c\, V\, C_V \frac{\beta^2 n^2}{Kd}, \text{ with the optimal } \ell = \beta C_V n$$
- Less than $1/2$ when the number of parameters Kd exceeds a multiple of $(\beta n)^2$
- Then indeed Hessian $\ge (\rho/2)I$. Strictly log concave
- Hence the posterior sampler is rapidly mixing

Summary

- Information Theory provides keys to the study of Bayes predictive distributions
- Multi-modal neural net posteriors can be efficiently sampled
- Log concave coupling provides the key trick
- Requires a number parameters $K\,d$ large compared to the sample size N
- Statistically accurate provided ℓ_1 controls on parameters are maintained
- Provides the first demonstration that the class $\mathcal{F}_{1,V}$ associated with single hidden-layer networks is both computationally and statistically learnable
- A polynomial number of computations in size of the problem is sufficient
- The approximation rate $1/K$ and statistical learning rate $1/\sqrt{N}$ are independent of dimension for this class of functions

Pages with additional details as well as topically arranged references can be accessed at

stat.yale.edu/∼arb4/ShannonLecture

The following pages contain essentially the same presentation but with some more details, some more material, a few more citations, and a topically arranged bibliography of references

In this expanded version, an asterisk * in the upper right corner means that it is similar to an included page but has added detail, a double asterisk ** means that it is a new page that explains material that is only briefly alluded to in the original presentation

Information Theory and High-Dimensional Bayesian Computation

The Blessing of Dimensionality

Expanded Version

Andrew R. Barron

YALE UNIVERSITY

Department of Statistics and Data Science

Joint work with Curtis McDonald (Yale)

Shannon Lecture

IEEE International Symposium on Information Theory

Athens, Greece, 11 July 2024

You may access these slides now at
stat.yale.edu/∼arb4/ShannonLecture.pdf

Outline

- Some of the computational core of Information Theory
  - Shannon-arithmetic codes for univ. data compression & algorithms for predictive distributions
  - Encoding and decoding for reliable communication, at rate near the Shannon capacity
- Average-case optimality or minimax optimality requires Bayes computation
- Historical roots of Laplace and Gauss

  From Laplace to modern prediction and compression: discrete data

  From Gauss to modern prediction and learning: continuous data
- Information-Theoretic determination of performance

  Essential ingredients: Approximation, Estimation, and Computation
- Information Theory of sampling log-concave posterior densities
- Beyond Log-Concavity
  - Provably Fast Sparse Regression Codes achieving Shannon capacity for the Gaussian channel
  - Provably Fast Posterior Sampling for neural net posterior distributions in sufficiently high dimensions
- Log concave coupling for sampling neural net posteriors

Outline of Conclusions for Mean Squared Error and Kullback Risk **

- Approximation, Estimation and Computation
  Can we meet all three objectives in flexible high-dimensional models?
- Function Models: $f(x, w)$, inputs $x \in R^{d_0}$, weights $w \in R^d$
  - Linear and non-linear models
  - Unconstrained versus constrained parameters
  - Traditional models versus modern neural networks
  - K term approx squared error $1/K^{2/d_0}$ versus $1/K$
- Mean Squared Prediction Error or Kullback Risk, with sample size N

$$\frac{d}{2N} \quad \text{or} \quad \left(\frac{1}{N}\right)^{2/(2+d_0)} \quad \text{or} \quad \left(\frac{\log d_0}{N}\right)^{1/2}$$

  and $\frac{1}{N}$ times Cumulative Risk or Data Compression Redundancy

$$\frac{d}{2N}\log N \quad \text{or} \quad \left(\frac{1}{N}\right)^{2/(2+d_0)} \quad \text{or} \quad \left(\frac{\log d_0}{N}\right)^{1/2}$$

  where the appropriate number of terms or neurons K grows with N
- Arrange a large number of variables d_0 and number of parameters $d = Kd_0 >> N$
- A Computational Success of Predictive Bayes
  Log concave coupling for sampling neural net posteriors
- The Blessing of Dimensionality
  Posterior sampling in high dim avoids traps of multi-modal optimization

Shannon-Arithmetic Codes for Universal Data Compression

Realistic and practical data compression arises in the universal source coding context

- Parameters θ of the distribution $p(Y^N|\theta)$ not known, but can be modeled
- Redundancy is the difference in expected codelength with and without knowledge of the parameters, divided by N to get redundancy as a rate
- The one or two bits of difference from $\log 1/p(Y^N)$ are ignored, as they contribute negligibly to the redundancy rate
- For parameters modeled probabilistically, the average-case optimal codes use
  $$p(Y^N) = \int p(Y^N|\theta)\, p(\theta)\, d\theta$$
  to construct the Huffman code, or, preferably, the Shannon-arithmetic code
- The average redundancy is the Shannon mutual information $I(\theta; Y^N)$
- And the minimax redundancy is the capacity of the channel $\theta \to Y^N$
- Practical optimal-redundancy codes require computation of predictive distributions
  $$p(Y_{n+1}|Y^n) = \int p(Y_{n+1}|Y^n, \theta)\, p(\theta|Y^n)\, d\theta$$

Equivalence of Statistical Learning and Universal Data Compression

- The redundancy of a code takes the form of the Kullback divergence

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

- Chain rule of probability $p(Y^N) = \prod_{n=0}^{N-1} p(Y_{n+1}|Y^n)$ yields the chain rule of information theory

$$D(P_{Y^N|\theta}||P_{Y^N}) = \sum_{n=0}^{N-1} E_{Y^n|\theta}\left[D(P_{Y_{n+1}|Y^n,\theta}||P_{Y_{n+1}|Y^n})\right]$$

- Consider the case that the model makes $Y_1, \ldots, Y_N$ conditionally i.i.d. given θ
- Predictive $p(y|Y^n)$ at $Y_{n+1} = y$ is average-case optimal estimator of $p(y|\theta)$ with Kullback loss
- The Cesàro average of the risk with Kullback loss equals the redundancy rate

$$\frac{1}{N}\sum_{n=0}^{N-1} E_{Y^n|\theta}\left[D(P_{Y|\theta}||P_{Y|Y^n})\right] = \frac{1}{N} D(P_{Y^N|\theta}||P_{Y^N})$$

- Statistical learning and universal data compression have the same computational challenge: For suitable models $p(Y^n|\theta)$ and $p(\theta)$, find a procedure to compute the predictive distributions

$$p(y|Y^n) = \int p(y|\theta)\, p(\theta|Y^n)\, d\theta$$

Settings with Practical Predictive Distributions

For suitable models $p(Y^n|\theta)$ and $p(\theta)$, find practical procedures to compute the predictive distributions $p(y|Y^n) = \int p(y|\theta)\, p(\theta|Y^n)\, d\theta$

We discuss several settings:

- Discrete memoryless sources
- Markov models and variable order (context tree) models
- General smooth parametric families
- Location families for the normal and other log-concave error distributions
- Linear models with the normal and other log-concave error distributions
- Regression codes for achieving capacity in additive Gaussian noise channels
- Nonlinear models such as single hidden-layer neural networks

Computation of optimal procedures in such models has roots in work of Laplace & Gauss

New computational innovations are based on log-concave sampling and beyond

A Remark Concerning Continuous-Valued Models

Models based on probability density functions

allow nearly continuous-valued data:

- Numerical data is often modeled as discretized real data to accuracy 2^{-b} (that is to b bits accuracy, with large b)
- When large, b has little effect on the discretized redundancy, because the redundancy depends on the ratio of probabilities, near the density ratio
- The supremum of redundancies over discretizations equals the Kullback divergence between the densities
- Thus the Kullback divergence for densities is still an appropriate redundancy measure

Historical Highlight: Bayes

Bayes (1763):

- Rule for reversing conditional probability: $P(A|B) = P(A)P(B|A)/P(B)$
- Provided notions of prior and posterior probability

Examined Binomial counts with uniform prior

- Found that the resulting marginal distribution on the counts is uniform on $\{0, 1, 2, ..., n\}$
- However, he was not able to compute the posterior predictive distribution.
  He did not see the solution by a rule of succession
- Also, the posterior probability of intervals was not computationally available to him
- He did not submit his work for publication. It was submitted and read before the Philosophical Society posthumously by Price.

Historical Computational Highlights: Laplace

Laplace (1774) Calculus of Probability. Commentary and translation by Stigler (1986)

- Also chooses the uniform prior
- Exact computation, for discrete memoryless sources, of the key ingredients
  - The predictive distrib $p(y_{n+1}|y_1,\ldots,y_n)$
  - The joint distribution $p(y_1,\ldots,y_n) = \int p(y_1,\ldots,y_n|\theta)p(\theta)d\theta$
  - The posterior density $p(\theta|y^n) = p(y^n|\theta)p(\theta)/p(y^n)$
- Approximate computation, for general smooth families, by integration using a normal
  - Central limit theory for posterior distributions
  - First appearance of the normal distribution, and $\sqrt{2\pi}$ normalization
- Decision Theory for location models and linear models
  - Median of posterior minimizes expected absolute deviation
  - Two-sided exponential error distribution
  - Could not compute posterior median except when $n \le 3$
  - Fall-back choice of sample median recognized as suboptimal

Laplace (1810, 1812)

- Central limit theory for sums of independent random variables
- A many-causes justification of least squares for linear models
- Normal error distrib. allows computation of posterior mean, optimizes expected posterior loss

From Laplace to Inform Theory of Prediction & Data Compression

- The Computational Heart of Laplace's Calculus of Probability
  - Joint distribution: $p(y_1, ..., y_N) = \int p(y_1, ..., y_N|\theta)\, p(\theta) d\theta$
  - Reduction for $n \le N$: $p(y_1, ..., y_n) = \int p(y_1, ..., y_n|\theta)\, p(\theta) d\theta$
  - Predictive distributions $p(y_{n+1}|y_1, ..., y_n)$
    - Ratios of joint at $n+1$ and n
    - Interpretable as posterior mean distribution estimator at $y_{n+1} = y$

      $$p(y_{n+1}|y_1, ..., y_n) = \int p(y|\theta)\, p(\theta|y^n) d\theta$$
  - Chain rule of probability

    $$p(y_1, ..., y_N) = \prod_{n=0}^{N-1} p(y_{n+1}|y_1, ..., y_n)$$
  - Also heart of AEP: Shannon 48, McMillan 53, Breiman 57, Ba. 85, Orey 85
- Decision Theory of Compression and Prediction with Kullback loss
  - Predictive distribution minimizes posterior mean of Kullback divergence
  - Code redundancy is the total Kullback divergence $D(P_{Y^N|\theta}||P_{Y^N})$
    Code with respect to P_{Y^N} is average case optimal
    Average redundancy is the mutual information $I(\theta; Y^N)$
  - Information theory chain rule for cumulative Kullback risk

    $$\frac{1}{N} \sum_{n=0}^{N-1} E_{Y^n|\theta}\, D(P_{Y_{n+1}|Y^n,\theta}||P_{Y_{n+1}|Y^n}) = \frac{1}{N}\, D(P_{Y^N|\theta}||P_{Y^N})$$
  - Joint and predictive distributions permit Shannon and arithmetic codes

Minimax total Kullback risk = Minimax redundancy = Shannon capacity of $Y^N|\theta$

Laplace's Prediction Rule based on Count Data

Certain priors on probabilities θ in the simplex $\{\theta : \theta_j \geq 0, \sum_{j=1}^m \theta_j = 1\}$

- permit exact predictive distribution computation
- allowing computation for arithmetic codes

For discrete memoryless sources with m symbols (Laplace 1774 used $m{=}2$)

- Laplace 1774. Uniform prior yields computation by Laplace's rule of succession

  $\hat{p}_n(y) = p(y_{n+1} = y | y_1, \ldots, y_n) = \frac{n_y+1}{n+m}$ from counts $n_y = \sum_{i=1}^n 1_{\{y_i = y\}}$

  Laplace joint distribution $p(y_1, \ldots, y_N) = \frac{1}{\binom{N+m-1}{m-1}} \frac{1}{\binom{N}{N_1 \ldots N_m}}$

  It gives the average-case optimal code for uniform prior (Gilbert 71, Cover 72, 73)

  Risk bound for Kullback loss (Ba 86): $E\left[D(p||\hat{p}_n)\right] \leq \log\left(1 + \frac{m}{n}\right) \leq \frac{m}{n}$

- Dirichlet($\lambda, \ldots, \lambda$) prior (originally in Laplace 1781) produces the prediction rule $\frac{n_y+\lambda}{n+m\lambda}$

  Distinguished choice $\lambda = 1/2$

  - Asymtotically capacity-achieving, providing minimax redundancy
  - Krichevski, Trofimov 81: Redundancy rate $\frac{m-1}{2N} \log N + O(\frac{1}{N})$
  - Xie, Ba 97,00: Minimax redundancy & regret $\frac{m-1}{2N} \log \frac{N}{2\pi} + \frac{1}{N} \log \int |I(\theta)|^{1/2} d\theta + o(\frac{1}{N})$

Prediction and Compression for Sources with Memory

For discrete Markov sources: Takeuchi, Kawabata, Ba 02

- Evaluates the asymtotically capacity-achieving Jeffreys prior achieving minimax redundancy
- again redundancy rate equals $\frac{d}{2N} \log N + \frac{C}{N} + o(\frac{1}{N})$ where $d =$ parameter dimension

For variable order Markov sources: Willems, Shtarkov, Tjalkens 95

- recursive Context Tree Weighting (CTW) algorithm
- Optimal prediction, compression, text generation for their prior & posterior

Scaling-up CTW at the word level, with access to massive amounts of text data, should yield a competitive, stochastically-optimal, large language model

Laplace Approximation for Posterior and Bayes Factor

For general smooth parametric families

- Laplace Approximation of the Posterior

  from second order Taylor expansion of log density with empirical Fisher information $\hat{I}$

$$p(Y^n|\theta)\, p_0(\theta) \quad \sim \; p(Y^n|\hat{\theta})\, p_0(\hat{\theta}) \quad \exp\{-\tfrac{1}{2} n \hat{I}\, (\theta-\hat{\theta})^2\}$$

  yields approximate normality of the posterior

- Integrating it yields the Laplace Approximation of the Joint Distribution, Bayes factor

$$\int p(Y^n|\theta)\, p_0(\theta)\, d\theta \;\sim\; p(Y^n|\hat{\theta})\, p_0(\hat{\theta}) \int \exp\{-\tfrac{1}{2} n \hat{I}\, (\theta-\hat{\theta})^2\}\, d\theta$$

$$= p(Y^n|\hat{\theta})\, p_0(\hat{\theta}) \left(\frac{2\pi}{n^d |\hat{I}|}\right)^{1/2}$$

- Taking logs yields the pointwise regret of stochastic complexity, MDL

  Ba 85, Clarke, Ba 90,94, Rissanen 96, Takeuchi, Ba 24

$$\frac{1}{N} \log \frac{p(Y^n|\hat{\theta})}{\int p(Y^n|\theta) p_0(\theta) d\theta} = \frac{d}{2n} \log \frac{n}{2\pi} + \frac{1}{n} \log \frac{|\hat{I}(\hat{\theta})|^{1/2}}{p_0(\hat{\theta})} + o\left(\frac{1}{n}\right)$$

Kullback Risk and Data Compression

Continuing for general smooth parametric families with i.i.d. observations

Taking the expected value yields the redundancy of data compression, equivalently, it is the cumulative Kullback risk for sample sizes $n \leq N$ (Clarke, Ba 90,94)

$$\frac{1}{N} D(P_{Y^N|\theta} || P_{Y^N}) = \frac{d}{2N} \log \frac{N}{2\pi e} + \frac{1}{N} \log \frac{|I(\theta)|^{1/2}}{p_0(\theta)} + o\left(\frac{1}{N}\right)$$

Jeffreys prior $p_0(\theta)$ proportional to $|I(\theta)|^{1/2}$

- Approximately mimimax for total Kullback risk and redundancy, (Clarke, Ba 94)
- Approximately capacity-achieving, maximizing $I(\theta; Y^N)$ asymptotically (Bernardo 79, Ibragimov, Hasminskii 73, Clarke, Ba 94)
- Hartigan 64: Jeffreys prior equalizes probability of small Kullback balls of given radius

Individual Kullback risk based on a sample of size n

- Parametric settings: (Cencov 72, Akaike 73, Yang, Ba 98, Hartigan 99), in i.i.d. case

$$E\left[D(P_{Y|\theta} || P_{Y|Y^n})\right] \sim \frac{d}{2n}$$

Dependence on θ and on the choice of prior arise only in terms of order $(1/n)^2$

- Nonparametric settings: approximation and estimation tradeoff (Ba, Sheu 91)

$$D(P||\hat{P}_n) \sim \min_K \left\{ \left(\frac{1}{K}\right)^{2/d_0} + \frac{K}{n} \right\} \sim \left(\frac{1}{n}\right)^{2/(2+d_0)} \quad \text{in the one derivative case}$$

Optional: From Laplace to Large Deviations **

Laplace (1785) approximation with series expansion

- to compute integrals of products of functions raised to high powers
- in particular to compute posterior probabilities of intervals
- for the Beta distribution (posterior for Binomial)
- for the normal, in particular

Large deviation probability:

$$\int_T^\infty e^{-t^2} dt = \frac{e^{-T^2}}{2T}\left(1 - \frac{1}{2T^2} + \frac{1\cdot 3}{2^2 T^4} - \frac{1\cdot 3\cdot 5}{2^3 T^6} + \ldots\right)$$

Leading term e^{-T^2} provides the large deviations exponent for the normal

Refinements for sums of i.i.d. random variables:

- Similar infinite series: Bahadur, Ranga-Rao 1960
  Coefficients of expansion related to moments
- Focus on the leading term
  - Cramèr 37, Chernoff 52 large deviations exponents for other distributions
  - Sanov 57, Hoeffding 65, Csiszár 75, 84 Information theory characterization
  - Kullback 59, v. Campenhout,Cover 81, Csiszár84,91 Information projection & conditional limit theory. Presented as an alternative to inverse probability

Optional Page: A Surprising Application of Bayes-Laplace Computation **

Contrast minimax redundancy $\min_Q \max_\theta D(P_{Y^n|\theta}||Q_{Y^n})$

with minimax pointwise regret $\min_q \max_{\theta,y^n} \log p(y^n|\theta)/q(y^n)$

- Shtarkov (88) minimax-regret solution: $q(y^n) = \max_\theta p(y^n|\theta)/c_n$
  This is the normalized maximum likelihood championed by Rissanen 96
  Detailed asymptotics: Szpankowski 95, Takeuchi, Ba 24
- It is not a Bayes-Laplace mixture
- So how can one compute its predictive distributions needed for its arithmetic code?
- Ba, Roos, Watanabe 14, solution in discrete settings by linear algebra:
  Represent $q(y^n) = \sum_j w_j\, p(y^n|\theta_j)$ with weights w_j possibly negative. Then
  Laplace's calculus still applies! May evaluate its positive marginals and predictive distributions
- Negative prior probabilities!
  These priors yield computation of positive-valued quantities for optimal prediction & compression.
  They are not for prior subjective assessment
- Here y^n has an exponentially large domain. Fortunately, the set of values of sufficient statistics (e.g. counts) is more moderate-sized, and the number of θ_j can be arranged accordingly
- Practical exactly minimax regret data compression for arbitrary sequences

Historical Highlight of Bayesian Computation: Gauss

Gauss (1806 German, 1809 Latin) English Transl. Davis (1857)

- Treatise on planetary motion (describing work developed 1794 -1805)
- Improves orbit determination when there are more than three observations
- Linearizes smooth nonlinear dependence on parameters (per Newton)
- Linear system of equations characterizing least squares solution
  Recognized in a paper by Legendre (1805)
- Gauss elimination solution

Gauss justification of least squares as a Bayesian Computation

- For linear models $f(x_i, w) = w \cdot x_i$ with observed responses y_i
- Given a density $\phi(z)$ for deviations with score $s(z) = -\phi'(z)/\phi(z)$
- The posterior density $p(w|Data)$ is proportional to the joint density function
  $$\phi(y_1 - w \cdot x_1) \ldots \phi(y_n - w \cdot x_n)$$
- Mode $\hat{w}$ of the posterior distribution is found by solving the system of equations
  $$\sum_{i=1}^{n} s(y_i - w \cdot x_i)\, x_i = 0$$
- Gauss' density $\phi(z)$ with linear score provides the linear system of equations
- Accordingly the least squares solution is the posterior mode
- Moreover Gauss showed:
  - The least squares solution is a linear combination of the observed y_i
  - Moreover, if posterior modes are linear for location and regression problems then the density $\phi(z)$ must be the Gaussian

Further Probability and Computation Conclusions of Laplace and Gauss **

Further linear model work, Laplace 1820, Gauss 1823, see Stigler (1986)

- With independence, the variance of a sum is the sum of the variances
- Provides valuation of $var(\hat{w}_j)$ and the standard error
- The least squares solution is unbiased
- Least squares solution has smallest variance among linear unbiased estimators
- Its variance is the same for all w
- The parameter estimates $\hat{w}_j$ and predictions $\hat{w} \cdot x$ are asymptotically normal
- Interval widths of given prob are asymp smallest with least squares estimates

Moreover, if the error density ϕ is normal, then

- The least square solution is the post mean, optimizing posterior expected square
- Normal integration explicitly provides predictive densities for $y_{n+1} = y$ at $x_{n+1} = x$

$$p(y|x, Data) = \int \phi(y - w \cdot x)\, p(w|Data)\, dw$$

as well as their predictive means $E[Y|x, Data] = \int w \cdot x\, p(w|Data)\, dw = \hat{w} \cdot x$

From Laplace and Gauss to Modern Bayesian Computation *

Laplace and Gauss least squares work celebrated

- for appropriate setting providing computation of the posterior optimal solutions
- for providing predictive densities $p(y|x, Data)$, predictive means, and Bayes factors
- Gauss' recursive least squares yields solution iterating one observation at a time

Linear Filtering and Prediction

- Kalman (1960) theory extends recursive Bayes computation to the setting of linear difference equation evolution of the states x_n

Model Selection and Data Compression: compute Bayes factors and MDL stochastic complexity

- Evaluating $p(Y^N|X^N) = \int p(Y^N|X^N, w)p(w)dw$ and associated predictive densities
- Permits optimal arithmetic coding of finely discretized observations
- Related to linear predictive coding

Minimax Estimation and Compression for linear models, general ϕ

- The Uniform prior yields minimax optimality per Hunt-Stein theory for
  - parameter estimation with squared error loss (Girshick, Savage 51)
  - predictive density estimation with Kullback risk (Liang, Ba.02)
  - data compression with minimax redundancy (Liang, Ba.02)
- Gaussian model continues providing ease of Bayes computation in these settings
- Proper Bayes minimax rules found for $d \geq 5$ (Strawderman 72, Liang 00)

From Gaussian to Log-Concave Distributions

- Summary thus far:
  Laplace and Gauss performed the required normal integrations
  in linear models to compute the posterior optimal procedures
- What is the right extension
  to preserve rapid computation of high-dimensional posterior integrals?
- Main approach emerging in the last forty years: Log-Concavity
  MCMC samplers: Accurate and mix rapidly for log concave posteriors
- Implication: Rapid computation of minimax optimal procedures
  for location estimation, linear regression and for minimax redundancy compression in linear predictive setting are polynomial-time computable for any log-concave error distribution
- Important settings that are not log-concave:
  - regressions with non-convex domains
  - non-linear regressions, such as neural networks

Optional: Entropic Central Limit Theorem **

- Random variable X centered and scaled to have mean 0 and variance 1
- - log density $\log 1/p(x)$ and score $s(x) = \frac{d}{dx} \log 1/p(x)$
- For the standard normal density $\phi(x)$ these are, respectively

  $\frac{1}{2}x^2 + c$ and x
- Closeness of the score to linear: $J(X) = E[(s(X) - X)^2]$
  to assess statistical efficiency of Gauss likelihood equation solution
- Closeness of log densities to quadratic: $D(X) = D(p||\phi)$
  to assess redundancy of descriptions based on the normal
- Score representation of divergence: Ba 86, with $\tau_t = e^{-2t}$, indep $Z \sim \phi$

  $$D(X) = \frac{1}{2} \int_0^\infty J(\sqrt{\tau_t}\, X + \sqrt{1-\tau_t}\, Z)\, dt$$

  Remark: Score of $Y = X+Z$ relates best nonlinear and linear estimates of X given Y, Brown 71, 82, Ba 86, so its an integrated mmse representation
- For $S_n = \frac{X_1 + \ldots X_n}{\sqrt{n}}$ with X_i i.i.d. Precursor results: Linnik 59, Brown 82
- Entropic CLT: $D(S_n) \to 0$ iff eventually finite, Ba 86
- Score CLT: $J(S_n) \to 0$ iff eventually finite, Johnson, Ba 04
- Monotone: Artstein, Ball, Barthe, Naor 04, Tulino, Verdú 06, Madiman, Ba 06
- Related results:
  - Subset Sum Entropy Power Inequality, Madiman, Ba 07
  - Log Sobolev Inequality (LSI): $D(X) \leq \frac{1}{2}J(X)$ Stam 57, Gross 75
  - Stochastic diffusion distribution properties with Gaussian limit

Information Theory of Rapid MCMC with Log Concavity

- Langevin Diffusion Path for sample parameter values w_t
  $$d\,w_t = \frac{1}{2}\nabla \log p(w_t)\,dt + d\,B_t$$
  - Score $\nabla \log p(w)$ is non-linear in general
  - There are time-discretizations (e.g. Metropolis adjusted Langevin) with similar mixing processing
  - Initialize with w_0 distributed $N(0,(1/\rho);I)$ or initialize using the Laplace approximation
- Theory of Bakry-Emery 85, see Bakry, Gentil, Ledoux 14
  Strong log concavity yields rapid Markov process convergence
- In particular, in the stochastic diffusion setting, if for $\rho > 0$
  $$\nabla\nabla' \log 1/p(w) \geq \rho\, I$$
  yields exponential convergence of relative entropy (Kullback divergence)
  $$D(p_t||p) \leq e^{-t\,\rho}\, D(p_0||p)$$
- The time required for small relative entropy is controlled by $\tau = 1/\rho$
- Proof uses $D(p_t||p) = \frac{1}{2}\int_{\tau\geq t} J(p_\tau||p)\,d\tau$ associated with $\frac{d}{dt}D(p_t||p) = -\frac{1}{2}J(p_t||p)$ and establishes Log Sobolev Ineq: $D(p_t||p) \leq \frac{1}{2\,\rho}J(p_t||p)$ where J is mean square norm between the scores
- Similar identities in Stam (59) for entropy power inequality & log Sobolev ineq for the normal, and in Ba 86
- Central Limit Theorem of Ba 86, showing relative entropy convergence to the normal for standardized sums of i.i.d. random variables, uses similar tools and the linear score target

Beyond Log-Concavity

Some important posterior are not log-concave

Examples with computationally feasible and accurate procedures in high-dimensions

- Bayes Computation for Communications
  - Capacity-achieving sparse regression codes
  - For a Gaussian noise channel
  - Codes are in a linear model Xw but with a non-convex constraint on w
- Bayes Computation for Non-linear Regression
  - Applies to neural nets with smooth activation functions
  - Posterior density has many peaks. It is not log-concave
  - Introduce of sufficiently many auxiliary random variable to simplify the sampling landscape

Bayes Computation for Communication

Communication strategy for additive Gaussian noise channel with specified power control

Capacity-achieving Sparse Regression Codes Joseph, Ba 12

- Gaussian design matrix X
- Codewords of form $X\,w$
- Non-convex constraint set W of size 2^{nC} for the weights w
  specified by a sparsity requirement of one non-zero in each of several sections and by a power allocation
- Bayes optimal decoder seeks $\min_{w \in W} ||Y - Xw||^2$

Computationally-feasible capacity-achieving iterative decoders

Compute weight estimates w_k iteratively, for a small (logarithmic) number of steps.
After which the estimates concentrate on the columns sent with high probability

- Adaptive Successive Hard-Decision Decoder (Joseph, Ba 14)
- Adaptive Successive Soft-Decision Decoder (Ba, Cho, 12)
  Compute w_k as posterior mean of indicators, given approx normal distributions of the inner products of the columns of X with residuals $Y - Xw_{k-1}$, normalized
- Approx Message Passing Decoder (Rush, Greig, Venkataramanan 17)

Sparse Regression Codes Monograph: Venkataramanan, Tatikonda, Ba 19

Artificial Neural Network Learning

A. Approximation

Squared approx error is of order $\frac{1}{K}$ with K neurons combined on last layer

B. Estimation

Squared estimation error is of order $\frac{K \log d}{N}$ with sample of size N, input dimension d

C. Computation

Computation time is a low order polynomial in N, K, d, when Kd is larger than N

Approximation and Estimation Essentials

A. Neural Net Model and Approximation Error

- Target function f, Variation $V(f) = V_L(f)$ with L hidden-layers
- Approximation $f_{K,L}$ with K subnetworks
- Single hidden-layer case ($L = 1$)

$$f_K(x) = \sum_{k=1}^{K} c_k \psi(w_k \cdot x)$$

- Approximation Accuracy

$$||f - f_{K,L}||^2 \leq \frac{V^2(f)}{K}$$

B. Neural Net Estimation and Risk

- Via constrained least squares, penalized least squares or Bayes predictions $\hat{f}$, with sample size N, input dimension d
- Risk $E[||\hat{f} - f||^2] \leq \; c\, V(f) \left(\frac{\log(2d)+L}{N}\right)^{1/2}$

  There are also lower bounds of such order (Klusowski, Ba 17)
- We provide computationally-feasible Bayes predictions with accuracy (in the single hidden layer case)

$$E[||\hat{f} - f||^2] \leq \; c\, V(f)^{2/3} \left(\frac{\log(2d)}{N}\right)^{1/3}$$

C. Log Concave Coupling for Bayesian Computation

- Focus on single hidden-layer network models
- Prior density $p_0(w)$: Uniform on an ℓ_1 constrained set
- Posterior $p(w)$: Multimodal. No known direct rapid sampler
- Coupling $p(\xi|w)$: cond indep Gaussian auxiliary variables $\xi_{i,k}$ with mean $x_i \cdot w_k$ for each observation i and neuron k
- Conditional $p(w|\xi)$ always log-concave
- Marginal $p(\xi)$ and its score $\nabla \log p(\xi)$ rapidly computable
- $p(\xi)$ is log concave when the number of parameters $K\,d$ is large compared to the sample size N
- Langevin diffusion and other samplers are rapidly mixing
- A draw from $p(\xi)$ followed by a draw from $p(w|\xi)$ yields a draw from the desired posterior $p(w)$

A. Variation and Approximation with a Dictionary G

- Variation with respect to a dictionary
  - Dictionary G of functions $g(x, w)$, each bounded by 1
  - Linear combinations $\sum_j c_j\, g(x, w_j)$
  - Control the sum of abs values of weights $\sum_j |c_j| \le V$
  - $\mathcal{F}_V =$ closure of signed convex hull of functions $V\, g(x, w)$
  - Variation $V(f) = V_G(f) =$ the infimum of V such that $f \in \mathcal{F}_V$.
- Approximation accuracy
  - Function norm square $||f - g||^2$ in $L_2(P_X)$
  - K term approximation: $f_K(x) = \sum_{k=1}^{K} c_k\, g(x, w_k)$
  - Approximation error: $||f - f_K||^2 \le \frac{V(f)^2}{K}$
  - Relative Approximation error: $||f - f_K||^2 - ||f - f^*||^2 \le \frac{V(f^*)^2}{K}$
  - Existence proof: Ba 93 Precursors: Gauss, Hilbert, Pisier
  - Greedy approximation proof: Jones, Ba 93
  - Outer weights c_k may equal $\pm \frac{V}{K}$
  - Relative approx error better than order $\left(\frac{1}{K}\right)^{1.5}$ is NP–hard (Vu 97)
  - Rate $\frac{1}{K}$ is dimension independent

Models **

- Models $f_K(x) = \sum_{k=1}^{K} c_k\, g(x, w_k)$ with error $||f - f_K||^2 \le \frac{V_G^2(f)}{K}$
  There are similar bounds for empirical average squares
- Various Algorithmic Terminology
  Sparse term selection, variable selection, forward stepwise regression, relaxed greedy alg, orthogonal matching pursuit, Frank Wolf algorithm, L_2 boosting, greedy Bayes
- Dictionary
  - Finite set of terms: Original predictors, products, polynomials, wavelets, sinusoids (grid of frequencies)
  - Product-type models: Parameterized bases, MARS (splines), CART regression trees, random forests
  - Ridge-type models: Multiple-index models, projection pursuit regression, neural networks, ridgelets, sinusoids (paramerized frequencies)
- Neural Network Models
  Single hidden-layer networks, multi-layer networks, deep networks, residual networks, adaptive learning networks, polynomial networks
- Network Units (neurons)
  Sigmoids, Rectified Linear Units (ReLU), low-order polynomials, compositions thereof

B. Neural Net Estimation and its Statistical Risk

- Via constrained least squares, penalized least squares or Bayes predictions $\hat{f}$,

  risk $\quad E[||\hat{f} - f||^2] \leq \; c\, V(f) \left(\frac{\log(2d)+L}{N}\right)^{1/2}$

  There are also lower bounds of such order (Klusowski, Ba 17)

- Computationally-feasible Bayes prediction accuracy (in the single hidden layer case)

$$E[||\hat{f} - f||^2] \leq \; c\, V(f)^{2/3} \left(\frac{\log(2d)}{N}\right)^{1/3}$$

- Both rates can be obtained by the Index of Resolvability:

$$\text{ApproxError} + \frac{1}{N} \log[\, 1 \,/\, \text{PriorProb(ApproxSet)}]$$

B. Methods of Obtaining such Statistical Risk Control

- Statistical risk or generalization squared error: $E[||\hat{f} - f||^2]$
- Five methods of controlling such statistical risk
  - Empirical process control of constrained least squares via metric entropy
    - Gaussian complexity: Ba, Klusowski 19
    - Rademacher complexity: Neshabur et al 15, Golowich et al 18
  - Penalized least squares risk control via relationship to MDL

    Adaptive bounds via an index of resolvability: Ba, Cover 90, Ba, Li et al 99, 08
  - Concentration of posterior distributions

    Necessary and sufficient conditions for posterior concentration Ba 88, 98, Ba, Shervish, Wasserman 98, Ghoshal, Ghosh, Van der Vaart 00
  - Cumulative Kullback risk of Bayes predictive distributions

    Clean information-theoretic bounds, again by an index of resolvability: Ba 87, 98, Yang, Ba 98, Ba, Klusowski 19, Ba, McDonald 24
  - Online learning regret bounds for squared error & log-loss

    Provides bounds for arbitrary data sequences
- All five have connections to information theory
- The posterior predictive procedures allow rapid computation

Multi-Layer Neural Network Model **

- Multi-Layer Net: Layers L, input x in $[-1,1]^d$, weights w
- Activation function: $\psi(z)$.
  - Rectified linear unit (ReLU): $\psi(z) = (z)_+$
  - Twice differentiable unit: sigmoid, smoothed ReLU, squared ReLU
- Paths of linked nodes: $\underline{j} = j_1, j_2, \ldots, j_L$.
- Path weight: $W_{\underline{j}} = w_{j_1,j_2} w_{j_2,j_3} \cdots w_{j_{L-1},j_L}$.
- Function representation:

$$f(x,c,w) = \sum_{j_L} c_{j_L} \psi\Big(\sum_{j_{L-1}} w_{j_{L-1},j_L} \psi(\ldots \psi(\sum_{j_1} w_{j_1,j_2} x_{j_1}) \ldots)\Big)$$

- Network Variation:
  - Internal: Sum abs. values of path weights set to 1.
  - External: $\sum_j |c_j| \le V$
  - Variation: $V_L(f) =$ infimum of such V to represent f
  - Single Hidden-Layer Case: $V_1(f) \le \int |\omega|_1^2 |\tilde{f}(\omega)| d\omega$ spectral norm
  - Class $\mathcal{F}_{L,V}$ of functions f with $V_L(f) \le V$
- Interests: Approx, Metric Entropy, Statistical Risk, Computation

Gaussian complexity approach to bounding risk

- Function class restricted to data $\mathcal{F}^n = \{f(x_1), f(x_2), \ldots, f(x_n) : f \in \mathcal{F}\}$
- Gaussian Complexity of $A \subset R^n$

$$C(A) = \frac{1}{\sqrt{n}} E_Z[\sup_{a \in A} a \cdot Z] \text{ for } Z \sim N(0, I),$$

- Complexity of Neural Nets: for ψ Lipshitz 1

$$C(\mathcal{F}^n_{L,V}) \le V \sqrt{2 \log 2d + 2L \log 2}$$

Via Sudakov-Fernique 75 comparison ineq. (Ba, Klusowski, 19)

(cf Neshabur, Tomioka, Srebro 15, Golowich, Rakhlin, Shamir 18)

- Gaussian complexity provides control of
  - Metric Entropy:

$$\log |\mathrm{Cover}(\mathcal{F}_{L,V}, \delta)| \le \frac{16 C^2(\mathcal{F}_{L,V})}{\delta^2}$$

  - Statistical Risk of Constrained Least Squares:

$$E[||\hat{f} - f||^2]| \le c \, \frac{C(\mathcal{F}_{L,V})}{\sqrt{n}} \le c\, V \left(\frac{2 \log 2d + 2L \log 2}{n}\right)^{1/2}$$

Minimum Description Length and Penalized Likelihood

- minus log likelihood plus penalty (e.g. penalized least squares)

$$\min_{w,K,V\in\Omega}\left\{\log\frac{1}{p(Y^N|X^N,f_{w,K,V})}+pen_N(w,K,V)\right\}$$

- Minimum description-length interpretation when it is at least

$$\min_{w,K,V\in\tilde{\Omega}}\left\{\log\frac{1}{p(Y^N|X^N,f_{w,K,V})}+L(w,K,V)\right\}$$

  for Kraft valid codelengths $L(\omega)$, such that $\sum_\omega 2^{-L(\omega)}\leq 1$
- ℓ_1 penalities with suitable multipliers are valid
- Battacharya-Renyi risk control via Index of Resolvability

$$E[d^2(p_f,p_{f_{\hat{\omega}}})]\leq\min_{\omega\in\Omega}\left\{D(p_f||p_{f_\omega})+\frac{pen_N(\omega)}{N}\right\}$$

  (Ba., Cover 90, Li, Ba. 99, Grünwald 07, Li, Huang, Luo, Ba. 08)
- Index of Resolvability: ApproxError $+$ Complexity$/N$
- Bounds for neural net risk $E[||\hat{f}-f||^2]$ in the $L=1$ case
  (Ba. 94, Ba., Birge, Massart 99, Huang, Cheang, Ba. 08, Ba., Luo 08)

$$\min_K\left\{\frac{V^2(f)}{K}+\frac{Kd}{N}\log N\right\}=V(f)\left(\frac{d\ \log\ N}{N}\right)^{1/2}$$

  Also, via the metric entropy bound, with ℓ_1 weight control

$$E[||\hat{f}-f||^2]\leq cV(f)\left(\frac{\log d}{N}\right)^{1/2}$$

- Computationally feasible?

B. Optional: Predictive Bayes and its Cumulative Risk Control

- Predictive density $\hat{p}_n(y|x) = \int p(y|x,w)p(w|x^n, y^n)dw$ evaluated at $Y_{n+1} = y$ with $X_{n+1} = x$
  Predictive mean $\hat{f}_n(x) = \int f(x,w)p(w|x^n, y^n)dw$
- The information theory chain rule for cumulative Kullback risk, in Gaussian noise case, controls data compression redundancy and the risk of $\hat{\hat{f}}(x) = \frac{1}{N}\sum_{n=0}^{N-1} \hat{f}_n(x)$ (Ba 87,98, Yang, Ba 99)
  $$E[||\hat{\hat{f}} - f||^2] \le \frac{1}{N}\sum_{n=0}^{N-1} E[||\hat{f}_n - f||^2]$$
- Indeed, the risk is controlled by the index of resolvability, Ba 87,98
  $$\frac{1}{N} D(P^*_{Y^N,X^N}||P_{Y^N,X^N}) = \frac{1}{N} E \log \frac{p^*(Y^N,X^N)}{\int p(Y^N,X^N|w)p_0(w)dw}$$
  $$\le \frac{1}{N} E \log \frac{p^*(Y^N,X^N)}{\int_A p(Y^N,X^N|w)p_0(w)dw}$$
  $$\le D_A + \frac{1}{N} \log \frac{1}{P_0(A)}$$
  where $D_A = \max_{w\in A} D(P^*_{Y|X}||P_{Y|X,w})$ is Kullback approx error. Best for a Kullback ball of optimized radius
- Predictive risk for neural net estimators with priors uniform on optimal covers
  $$E[||\hat{\hat{f}} - f||^2] \le cV(f)\left(\frac{d \log N}{N}\right)^{1/2} \qquad \text{Yang, Ba 98}$$
  $$E[||\hat{\hat{f}} - f||^2] \le cV(f)\left(\frac{\log d}{N}\right)^{1/2} \qquad \text{Ba, Klusowski 19}$$
  with practical priors and feasibly computable estimates for sufficiently large d
  $$E[||\hat{\hat{f}} - f||^2] \le cV(f)^{2/3}\left(\frac{\log(d_0)}{N}\right)^{1/3} \qquad \text{Ba, McDonald 24, now}$$

On-line learning

- Arbitrary-sequence regret for predictive Bayes
  - Squared error $\frac{1}{N}\sum_{n=1}^{N}(Y_n - \hat{f}_{n-1}(X_n))^2 - \frac{1}{N}\sum_{n=1}^{N}(Y_n - f(X_n))^2$
  - Log-loss case $\frac{1}{N}\sum_{n=1}^{N}\log\frac{1}{p(Y_n|f_{n-1}(X_n))} - \frac{1}{N}\sum_{n=1}^{N}\log\frac{1}{p(Y_n|f(X_n))}$
  - Simplification $\frac{1}{N}\{\log\frac{1}{p(Y^N,X^N)} - \log\frac{1}{p(Y^N,X^N|f)}\}$
  - Corresponds to pointwise regret of an arithmetic code
- Amenable to Laplace approximation and resolvablity bound
- Bounds of the same form

$$Regret_N \leq Approx\ Error + \frac{1}{N}\log\frac{1}{PriorProb(Approx\ Set)}$$

- Specialization to the case of functions f in $F_{1,V}$

$$Regret_N \leq cV^{2/3}\left(\frac{\log d}{N}\right)^{1/3}$$

- Taking expectation controls

$$\frac{1}{N}\sum_{n=1}^{N} E\left[||f - \hat{f}_{n-1}||^2\right]$$

- The estimator $\hat{\hat{f}}(x) = \frac{1}{N}\sum_{n=1}^{N}\hat{f}_{n-1}(x)$ also has this bound

C. Bayesian Computation for Neural Nets

- Data: (X_i, Y_i) for $i = 1, 2, \ldots, n$, with X_i in $[-1, 1]^{d_0}$ and $n \leq N$
- Natural yet optional statistical assumption:
  (X_i, Y_i) independent $P_{X,Y}$, target $f(x) = E[Y \mid X = x]$, variance $\sigma_Y^2 = \sigma^2$
  - Not needed for Bayesian computation statements
  - Not needed for online learning bounds
- Single hidden-layer network model: $f(x, \underline{w})$
  $f_K(x, \underline{w}_1, \ldots \underline{w}_K) = \frac{V}{K} \sum_{k=1}^{K} \psi(\underline{w}_k \cdot x_i)$
  One coordinate of each x_i always -1 to allow shifts
  Odd symmetry of ψ provides sign freedom
  Each $\underline{w}_k$ in the symmetric simplex $S_1^d = \{w : \sum_{j=1}^{d} |w_j| \leq 1\}$
- Prior: $p_0(\underline{w})$ makes $\underline{w}_k$ independent uniform on S_1^d
- Likelihood: $\exp\{-\beta g(w)\}$ with gain $0 < \beta \leq 1/\sigma^2$
  where $g(w) = \frac{1}{2} \sum_{i=1}^{n} \left(Y_i - \frac{V}{K} \sum_{k=1}^{K} \psi(x_i \cdot w_k)\right)^2$
- Posterior: $p(w) = p_0(w) \exp\{-\beta g(w) - \Gamma(\beta)\}$
- Bayesian Computation: Estimate $\hat{f}(x) = \int f(x, w) p(w) dw$
  by drawing independent samples from $p(w)$ and averaging $f(x, w)$

Hessian of the Minus Log Likelihood

- Log 1/Likelihood $= \beta\, g(w)$

$$\text{Hessian} = \beta H(w) = \beta\, \nabla\nabla' g(w)$$

- Squared error loss: $g(w) = \frac{1}{2}\sum_{i=1}^{n}(res_i(w))^2$ where

$$res_i(w) = Y_i - \frac{V}{K}\sum_{k=1}^{K}\psi(x_i \cdot w_k)$$

- Hessian Quadratic form: $a' H(w) a$, where a has blocks a_k

$$\frac{V^2}{K^2}\sum_{i=1}^{n}\left(\sum_{k=1}^{K}\psi'(x_i \cdot w_k)\, a_k \cdot x_i\right)^2$$
$$-\frac{V}{K}\sum_{i=1}^{n} res_i(w)\sum_{k=1}^{K}\psi''(x_i \cdot w_k)(a_k \cdot x_i)^2$$

- $p(w)$ is not log-concave; that is, $g(w)$ is not convex

  The first term is positive definite, the second term is not

- No clear reason for gradient methods to be effective

Log Concave Coupling

- Auxiliary Random Variables $\xi_{i,k}$ chosen conditionally indep
- Normal with mean $x_i \cdot w_k$, variance $1/\rho$, with $\rho = \beta c V/K$ restricted to ξ with each $\sum_{i=1}^n \xi_{i,k} x_{i,j}$ in a high probability interval
- Conditional density:

  $p(\xi|w) = (\rho/2\pi)^{Kn/2} exp\{-\frac{\rho}{2}\sum_{i=1}^n \sum_{k=1}^K (\xi_{i,k} - x_i \cdot w_k)^2\}$
- Multiplier $c = c_{Y,V} = \max_i |Y_i| + V$ bounds $|res_i(w)|$ for all w
- Activation second derivative: $|\psi''(z)| \leq 1$ for $|z| \leq 1$
- Joint density: $p(w,\xi) = p(w)p(\xi|w)$
- Reverse conditional density: $p(w|\xi) = p_0(w)\exp\{-\beta g_\xi(w) - \Gamma_\xi(\beta)\}$
- Conditional log 1/Likelihood $= \beta g_\xi(w)$ with

  $g_\xi(w) = g(w) + \frac{1}{2}\frac{V}{K} c \sum_{i=1}^n \sum_{k=1}^K \left(x_i \cdot w_k - \xi_{i,k}\right)^2$
- Modifies Hessian $a' H_\xi(w) a$ with new positive def second term

  $\frac{V}{K}\sum_i \sum_k [c - res_i(w)\psi''(x_i \cdot w_k)](a_k \cdot x_i)^2$
- $p(w|\xi)$ is log concave in w for each ξ
- MCMC Efficient sample Applegate, Kannan 91, Lovász, Vempala 07

Marginal Density and Score of the Auxiliary Variables

- Auxiliary variable density function:

  $p(\xi) = \int p(w,\xi) dw$

  Integral of a log concave function of w
- Rule for Marginal Score:

  $\nabla \log 1/p(\xi) = E[\nabla \log 1/p(\xi|w) \,|\, \xi\,]$
- Normal Score: linear

  $\partial_{\xi_{i,k}} \log 1/p(\xi|w) = \rho\,\xi_{i,k} \,-\, \rho\, x_i \cdot w_k$
- Marginal Score:

  $\partial_{\xi_{i,k}} \log 1/p(\xi) = \rho\,\xi_{i,k} \,-\, \rho\, x_i \cdot E[w_k \,|\, \xi\,]$
- Efficiently compute ξ score by Monte Carlo sampling of $w|\xi$
- Permits Langevin stochastic diffusion: with gradient drift

  $d\,\xi_t = \frac{1}{2} \nabla \log p(\xi_t)\, dt \,+\, d\,B_t$

  converging to a draw from the invariant density $p(\xi)$

Hessian of $\log 1/p(\xi)$. Is $p(\xi)$ log concave?

- Hessian of $\log 1/p(\xi)$, an nK by nK matrix
$$\tilde{H}(\xi) = \nabla\nabla' \log 1/p(\xi) = \rho\left\{ I - \rho\, Cov\left[\begin{matrix} Xw_1 \\ \cdots \\ Xw_K \end{matrix} \Big| \xi\right]\right\}$$
- Hessian quadratic form for unit vectors a in R^{nK} with blocks a_k
$$a'\tilde{H}(\xi)a = \rho\,\{1 - \rho\, Var[\tilde{a}\cdot w|\xi]\}$$
where $\tilde{a} = \begin{bmatrix} X'a_1 \\ \cdot \\ X'a_K \end{bmatrix}$ has $||\tilde{a}||^2 \le n\, d_0$
- Role for variance of $\tilde{a}\cdot w$ using the log-concave $p_\beta(w|\xi)$
- More concentrated, smaller variance, than with the prior?
- Counterpart using the prior
$$\rho\,\{1 - \rho\, Var_0[\tilde{a}\cdot w]\}$$
- Use $Cov_0(w_m) = \frac{2}{(d_0+2)(d_0+1)} I$ and $\rho = \beta c V/K$ to see its at least
$$\rho\left\{1 - \frac{2\beta c V n}{K(d_0+2)}\right\}$$
- Constant β chosen such that, say, $\beta c V \le 1/4$
- Strictly positive when number param Kd_0 exceeds sample size n

Is $p(\xi)$ log concave?

- Recap: quadratic form in Hessian of $\log 1/p(\xi)$

  $$a' \tilde{H}(\xi) a = \rho\{1 - \rho\, Var[\tilde{a} \cdot w|\xi]\}$$
- Another control on the variance

  $$\rho\, Var[\tilde{a} \cdot w|\xi] \le \rho \int (\tilde{a} \cdot w)^2 \exp\{-\beta \tilde{g}_\xi(w) - \Gamma_\xi(\beta)\} p_0(w) dw$$

  $$\text{using } \tilde{g}_\xi(w) = g_\xi(w) - E_0[g_\xi(w)]$$
- Hölder's inequality with $\ell \ge 1$

  $$\le \rho\, [E_0[(\tilde{a} \cdot w)^{2\ell}]]^{1/\ell} \exp\{\tfrac{\ell-1}{\ell} \Gamma_\xi(\tfrac{\ell}{\ell-1}\beta) - \Gamma_\xi(\beta)\}$$

  which is, using a bound $C_V n$ on $g_\xi(w)$ with $C_V = 9V^2 + 7V \max_i |Y_i|$,

  $$\le \frac{c\beta V}{K} \frac{4n\ell}{d_0 e} \exp\{\beta C_V n/\ell\}$$

  which is, with the optimal $\ell = \beta C_V n$,

  $$= 4c\, V\, C_V\, \frac{\beta^2 n^2}{K d_0}$$
- Less than $1/2$ when num param Kd exceeds a multiple of $(\beta n)^2$
- Then indeed Hessian $\ge (\rho/2) I$. Strictly log concave
- Hence the posterior sampler is rapidly mixing

Greedy Bayes **

- Initialize $\hat{f}_{n,0}(x) = 0$
- Given previous neuron fits, iterate k, for each n

$$f_{n,k}(x,w) = (1-\alpha) f_{n,k-1}(x) + \lambda \psi(w \cdot x)$$

- $\alpha = 1/\sqrt{n}$ and $\lambda = V\alpha$ are suitable.
- Form the iterative squared error $g(w)$

$$g_{n,k}(w) = \frac{1}{2} \sum_{i=1}^{n-1} \left(y_i - f_{i,k}(x_i, w) \right)^2$$

  Again Hessian has a not necessarily positive definite part

$$-\lambda \sum_{i=1}^{n-1} r_{i,k-1}\, \psi''(w \cdot x_i)\, x_i x_i'$$

  where $r_{i,k-1}$ are the previous residuals
- Associated greedy posterior $p_{n,k}(w)$ proportional to

$$p_0(w) \exp\{-\beta g_{n,k}(w)\}$$

- Update $f_{n,k}$ replacing $\psi(w \cdot x)$ with its posterior mean
- Estimate by sampling from the greedy posterior

Log Concave Coupling for Greedy Bayes **

- For the moment, fix n, k
- Again $p(w) = p_0(w)\exp\{-\beta g(w)\}$
- Coupling random variables $\xi_i \sim N(x_i \cdot w, 1/\rho)$ with $\rho = c\lambda\beta$ where c bounds the absolute values of the residuals $r_{i,k}$
- Joint density $p(w,\xi)$ with logarithm $-\beta\, g_\xi(w)$ built from

  $$g_\xi(w) = g(w) + \frac{1}{2}c\lambda\sum_{i=1}^{n-1}(\xi_i - w\cdot x_i)^2$$

  which is convex in w for each ξ, so $p(w|\xi)$ is log concave
- The associated marginal is $p(\xi)$
- Hessian quadratic form $a'\,\nabla\nabla'\log(1/p(\xi))\,a$

  $$\rho\{1-\rho Var[\,\tilde{a}\cdot w\,|\xi\,]\}$$

  for a with $||a|| = 1$ and $\tilde{a} = X'a$
- Deduce $p(\xi)$ is log concave for sufficiently large d
- From which get w by a draw from $p(w|\xi)$

Variance control for Greedy Bayes using Hölder's inequality **

- As before $Var[\tilde{a}\cdot w|\xi]$ is not more than

  $$\int (\tilde{a}\cdot w)^2 \exp\{-\beta \tilde{g}_\xi(w) - \Gamma_\xi(\beta)\}\, p_0(w)\, dw$$

  where $\tilde{g}_\xi(w)$ is $g_\xi(w)$ minus its mean value at $\beta = 0$
- $\Gamma_\xi(w)$ is the cumulant generating function of $-\tilde{g}_\xi(w)$
- By Hölders inequality that variance is not more than

  $$[E_0[(\tilde{a}\cdot w)^{2\ell}]]^{1/\ell} \exp\{\tfrac{\ell-1}{\ell}\Gamma_\xi(\tfrac{\ell}{\ell-1}\beta) - \Gamma_\xi(\beta)\}$$
- For the first factor, with integer $\ell \geq 1$

  $$E_0[(x_i\cdot w)^{2\ell}] \leq \binom{d+\ell-1}{\ell} \frac{(2\ell)!}{(d+2\ell)\cdots(d+1)}$$

  hence

  $$[E_0[(\tilde{a}\cdot w)^{2\ell}]]^{1/\ell} \leq n\,\frac{4\ell}{ed}$$

On the second factor from Hölders inequality **

- The exponent of the second factor is

$$\frac{\ell-1}{\ell}\Gamma_\xi\left(\frac{\ell}{\ell-1}\beta\right) - \Gamma_\xi(\beta)$$

- Not more than $\frac{\beta}{\ell-1}\max_w \tilde{g}_\xi(w)$ where

$$\tilde{g}_\xi(w) = g_\xi(w) - E_0[g_\xi(w_0)]$$

- It has the bound $\beta \max_{w,w_0}(g_\xi(w) - g_\xi(w_0))/(\ell - 1)$
- Indeed a value near $5c\lambda n$ bounds $\max_{w,w_0}(g_\xi(w) - g_\xi(w_0))$
- Optional page verifies this for a suitable set of ξ
- Hence exponent of second factor not more than value near

$$5\,\beta\lambda\,c\,n/\ell$$

Optional Page: Verifying the Bound on $\tilde{g}_\xi(w)$ **

- The $g_\xi(w) - g_\xi(w_0) = (w - w_0) \cdot \nabla g_\xi(\tilde{w})$.
- Concerning $\nabla g_\xi(\tilde{w})$ it is

$$-\lambda \left\{ \sum_{i=1}^{n-1} \left[res_{i,k-1} \psi'(\tilde{w} \cdot x_i) - c\tilde{w} \cdot x_i \right] x_i + \sum_{i=1}^{n-1} \xi_i x_i \right\}$$

- Hit with $w - w_0$, the result has magnitude not more than

$$4c\lambda n + \lambda \max_j \left| \sum_{i=1}^{n-1} \xi_i x_{i,j} \right|$$

- With high probability, the max is $\leq n + \kappa\sqrt{n/\rho}$ where $\kappa \geq \sqrt{2 \log 2d}$
- Conditioning on ξ which have this bound, the conditional density remains log concave when $\kappa = \sqrt{2 \log 6d^4}$
- With $\rho = c\lambda\beta$ and $\lambda = V/\sqrt{n}$, the max is $\leq n + \tilde{O}(n^{3/4})$
- Then exponent of second factor not more than value near

$$5\beta\lambda c\, n/\ell$$

Combining the two factors for Greedy Bayes log concavity **

- Use $\tilde{a} = \sum_i a_i x_i$ with $||\tilde{a}||^2 \leq nd$ and $\rho = c\lambda\beta$
- Combine the two factors
- Obtain $\rho Var[\tilde{a} \cdot w|\xi]$ not more than a value near

$$c\lambda\beta\, 4n\,\ell/(ed)\; exp\{5\beta\lambda c\, n/\ell\}$$

- The optimal $\ell = 5\beta\lambda c\, n$ yielding not more than

$$20(c\lambda\beta n)^2/d$$

- Recall $\lambda = V\alpha = V/\sqrt{n}$
- Choose $\beta = 1/(5cV)$, choose $d \geq n$.
- $\rho Var[\tilde{a} \cdot w|\xi]$ is strictly less than 1 (indeed less than $4/5$)
- Hence $p(\xi)$ is strictly log concave, for d exceeding n

Summary

- Information Theory provides keys to the study of Bayes predictive distributions
- Multi-modal neural net posteriors can be efficiently sampled
- Log concave coupling provides the key trick
- Requires a number parameters $K\,d$ large compared to the sample size N
- Statistically accurate provided ℓ_1 controls on parameters are maintained
- Provides the first demonstration that the class $\mathcal{F}_{1,V}$ associated with single hidden-layer networks is both computationally and statistically learnable
- A polynomial number of computations in size of the problem is sufficient
- The approximation rate $1/K$ and statistical learning rate $1/\sqrt{N}$ are independent of dimension for this class of functions

Pages with topically arranged references can be accessed next

References: Current

C. McDonald and A.R. Barron 2024 "Log Concave Coupling for Sampling Neural Net Posteriors," *Proc. IEEE Int Symposium on Information Theory*

A.R. Barron 2024 "Information Theory and High-Dimensional Bayesian Computation", Shannon Lecture, *IEEE International Symposium on Information Theory*, This presentation available July 11. Paper soon after.

Additional topically-arranged references are on the following pages

Many of these papers can be viewed at stat.yale.edu/∼arb4

References: Neural Nets and Greedy Approximation

A.R. Barron 1993 “Universal Approximation Bounds for Superpositions of Sigmoidal Function” *IEEE Trans Inform Theory*

A.R. Barron 1994 “Approximation and Estimation Bounds for Artificial Neural Networks” *Machine Learning*

A.R. Barron, L. Birge and P. Massart 1999 “Risk Bounds for Model Selection by Penalization” *Probability Theory and Related Fields*

A.R. Barron, A. Cohen, W. Dahmen and R. DeVore 1008 “Approximation and Learning by Greedy Methods” *Annals of Statistics*

L. Jones 1992 “A Simple Lemma on Greedy Approximation in Hilbert Space and Convergence Rates for Projection Pursuit Regression and Neural Network Training” *Annals of Statistics*

V. Vu 1997 “On the Infeasibility of Training Neural Networks with Small Squared Errors” *Adv in Neural Information Processing Systems*

G. Pisier 1980 “Remarques sur un Resultat Non-Publie de B. Maurey”, Presention at Seminaire d’Analyse Fonctionelle, Ecole Poly, Math, Paiseau

J.M. Klusowski and A.R. Barron 2017 “Minimax Lower Bounds for Ridge Combinations including Neural Networks” *Intern Symp Inform Theory*

J.M. Klusowski and A.R. Barron 2018 “Approximation by Combinations of ReLU and Squared ReLU with ℓ_1 and ℓ_0 Controls” *IEEE Trans Inform Theory*

A.R. Barron and J.M. Klusowski 2018 “Approximation and Estimation for High-Dimensional Deep Learning Networks,” ArXiv:1809.03090v2

A.R. Barron and J.M. Klusowski 2019 “Complexity, Statistical Risk and Metric Entropy of Deep Nets using Total Path Variation,” ArXiv:1902.00800v2

Neural Nets and Gaussian & Rademacher Compexity

A.R. Barron and J.M. Klusowski 2019 “Complexity, Statistical Risk and Metric Entropy of Deep Nets using Total Path Variation” ArXiv:1902.00800v2

B. Neshabur, R. Tomioka, N. Srebro 2015 “Norm-Based Capacity Control in Neural Networks”, *Conference on Learning Theory*

N. Golowich, A. Rakhlin, O. Shamir 2018 “Size-Independent Sample Complexity of Neural Networks” *Proc. Machine Learning Research*

X. Fernique 1975 “Regularité des Trejectoires de Fonctions Aleéatoires Gaussiennes” Lecture Notes in Mathematics Springer

V. N. Sudakov 1971 “Gaussian Random Processes and Measures of Solid Angles in Hilbert Space”, Translation in *Soviet Math. Dokl.*

V. N. Sudakov 1976 “Geometric Problems in the Theory of Infinite Dimensional Probability Distributions, *Proc. Steklov Inst. Math*, Translation 1979 by H.H. McFadden, American Mathematics Society

References: Consistency, Rates for Bayes Procedures

A.R. Barron 1987 “Are Bayes Rules Consistent in Information?” *Open Problems in Communication and Computation*, Springer

A.R. Barron 1988 “The Exponential Convergence of Posterior Probabilities with Implications for Bayes Estimators of Density Functions”, UIUC Dept Stat, Tech Rept #7.

A.R. Barron 1998 “Information-Theoretic Characterization of Bayes Performance and the Choice of Priors in Parametric and Nonparametric Problems”, *Bayesian Statistics 6*

A.R. Barron, M. Shervish, L. Wasserman 1998 “The Consistency of Posterior Distributions in Nonparametric Problems,” *Annals of Statistics*

Y. Yang and A.R. Barron 1999 “Information-Theoretic Determination of Minimax Rates of Convergence,” *Annals of Statistics*

S. Ghosal, J.K. Ghosh, A.W. Van Der Vaart 2000 “Convergence Rates of Posterior Distributions,” *Annals of Statistics*

References: d/n contrib to Kullback risk and information criteria

N. N. Cencov 1972,1982 *Statistical Decision Rules and Optimal Inference* American Mathematics Society, Translations of Mathematical Monographs

H. Akaike 1973 “Information Theory and an Extension of the Maximum Likelihood Principle” *Proc. Int. Symposium Information Theory*

A.R. Barron, C. Sheu 1991 “Approximation of density functions by sequences of exponential families” *Annals of Statistics*

Y. Yang, A.R. Barron 1998 “An Asymptotic Property of Model Selection Criteria” *IEEE Transactions Information Theory*

J. A. Hartigan 1998 “The Maximum Likelihood Prior” *Annals of Statistics*

References: Penalized Likelihood, MDL, Resolvability

A.R. Barron 1985 *Logical Smoothing*, Stanford Univ, PhD Dissertation

A.R. Barron and T.M. Cover 1991 "Minimum Complexity Density Estimation" *IEEE Trans Inform Theory*

A.R. Barron 1990 "Complexity Regularization with Application to Artificial Neural Networks" *Nonparametric Estimation and Related Topics*, Kluwer

A.R. Barron, L. Birge, P. Massart 1999 "Risk Bounds for Model Selection by Penalization," *Probability Theory and Related Fields*

J. Qiang Li 1999 *Estimation of Mixture Models* Yale Stat, PhD Dissertation

J.Q. Li and A.R. Barron 2000 "Mixture Density Estimation" *Adv Neural Inform Proc. Sys*, MIT Press

P.D. Grünwald 2005 *The Minimum Description-Length Principle* MIT Press

C. Huang, G. Cheang, A.R. Barron 2008 "Risk of Penalized Least Squares, Greedy Selection and ℓ_1 Penalization for Flexible Function Libraries," Yale Statistics publ list at www.stat.yale.edu/∼arb4

A.R. Barron, C. Huang, J. Li, X. Luo 2008 "MDL Principle, Penalized Likelihood, and Statistical Risk" *Festschrift for Jorma Rissanen* Tampere Univ Press

A.R. Barron and X. Luo 2008 "MDL Procedures with ℓ_1 Penalty and their Statistical Risk" *Workshop on Information Theoretic Methods in Science and Engineering*

References: Log Concave Sampling

D. Bakry, M. Émery 1985 “Diffusions Hypercontractives,” *Séminaire de Probabilités XIX*, Springer

D. Bakry, I. Gentil, M. Ledoux 2014 *Analysis and Geometry of Markov Diffusion Operators*, Springer

D Applegate and R Kannan, 1991 “Sampling and Integration of Near Log-Concave Functions,” *Proc. ACM Symposium on Theory of Computing*

L. Lovász and S. Vempala 2007 “The Geometry of Log Concave Functions and Sampling Algorithms,” *Random Structures & Algorithms*

Y. Kook, Y.T. Lee, R. Shen, S. Vempala 2023 “Condition-Number-Independent Convergence Rate of Reimannian Hamiltonian Monte Carlo with Numerical Integrators,” ArXiv 2210.07219v2

References: Era of Laplace and Gauss

T. Bayes 1763 “An Essay toward Solving a Problem in the Doctrine of Chances” *Philosophical Transactions*

P. S. Laplace 1774 *Mémoire sur la Probabilité des Causes par les Évènmens* [Commentary and translation by Stigler 1986 *Statistical Science*]

P. S. Laplace 1781 *Mémoire sur la Probabilité*, *Mémoirs de l’Académie Royal des Sciences de Paris* [As discussed in Hald 1998, Gupta, Richards 2001]

P. S. Laplace 1785 “Mémoire sur let approximations des formules qui sont fonctions de très grands nombres. *Mémoirs de l’Académie Royal des Sciences de Paris.* [As described in Stigler 1986]

P. S. Laplace 1812 “Théorie Analytique des Probabilités” Courcier. In 1825 *Essai Philosophique sur les Probabilités*. [Translated by Truscott and Emory 1902, Wiley, and by Dale 1995, Springer]

A-M. LeGendre 1805 *Nouvelles Méthodes pour la Détermination des Orbites de Comètes*, Firmin Didot

C. F. Gauss 1809 *Theoria Motus Corporum Coelestium in Sectionibus Conicis Solem Ambientium*, translated by Davis 1857.

C. F. Gauss 1823 “Theoria Combinationis Observationum Erroribus Minimis Obnoxiae” Parts I and II, *Proc. Roy. Soc. Gottingen*

References: Histories of Statistics

S. M. Stigler 1981 “Gauss and the Invention of Least Squares” *Annals of Statistics*

S. M. Stigler 1986 “Laplace’s 1774 Memoir on Inverse Probability” *Statistical Science*

S. M. Stigler 1986 *The History of Statistics: The Measurement of Uncertainty before 1900*, Belknap Press

A. Hald 1998 *A History of Mathematical Statistics from 1750 to 1930* Wiley

R. D. Gupta, D. St. P. Richards 2001 “The History of the Dirichlet and Liouville Distributions,” *International Statistical Review*

Reference Priors

H. Jeffreys 1961 *Theory of Probability*, Oxford Univ. Press

J. A. Hartigan 1964 “Invariant Prior Distributions” *Annals of Math Statistics*

I. A. Ibragimov, R.Z. Hasminski 1973 “On the Information in a Sample about a Parameter”, *Proc Intern Symp Inform Theory*

J. M. Bernardo 1979 “Reference Posterior Distributions for Bayesian Inference,” *J. Royal Statistics Society B*

B.S. Clarke and A.R. Barron 1994 Jeffreys’ Prior is Asymptotically Least Favorable Under Entropy Risk *J. Statistical Planning and Inference*

J. A. Hartigan 1998 “The Maximum Likelihood Prior” *Annals of Statistics*

References: Early Universal Source Codes and Arithmetic Coding

E. N. Gilbert, E. F. Moore 1959 “Variable Length Binary Encodings,” *Bell Systems Technical J.*

F. Jelinek 1968 *Probabilistic Information Theory: Discrete and Memoryless Models* McGraw-Hill

E. N. Gilbert 1971 “Codes based on Inaccurate Source Probabilities” *IEEE Trans Inform Theory*

T. M. Cover 1972 “Admissibility Properties of Gilbert’s Encoding for Unknown Source Probabilities” *IEEE Trans Inform Theory*

T. M. Cover 1973 “Enumerative Source Encoding” *IEEE Trans Inform Theory*

L. Davison 1973 “Universal noiseless coding,” *IEEE Trans Inform Theory*

L. Davison, A. Leon-Garcia 1980 “A Source Matching Approach to Finding Minimax Codes,” *IEEE Trans Inform Theory*

B. Ryabko 1979 “Encoding of a Source with Unknown but Ordered Probabilities,” *Problems in Information Transmission*

R. C. Pasco (1976) *Source Coding Algorithms for Fast Data Compression,* PhD Dissertation, Stanford Univ

J. Rissanen (1976) “Generalized Kraft Inequality and Arithmetic Encoding,” *IBM J. Research and Development*

Refs: MDL, Redundancy and Total Kullback Risk of Bayes Procedures

R. E. Krichevski, V. K. Trofimov 1981 “The Performance of Universal Coding” *IEEE Trans Inform Theory*

J. Rissenen 1984 “Universal Coding, Information, Prediction and Estimation” *IEEE Trans Inform Theory*

A.R. Barron 1985 “Logical Smoothing” PhD Dissertation, Stanford University

B.S. Clarke and A.R. Barron 1990 “Information-Theoretic Asymptotics of Bayes Methods”, *IEEE Trans Inform Theory*

B.S. Clarke and A.R. Barron 1994 Jeffreys’ Prior is Asymptotically Least Favorable Under Entropy Risk *J. Statistical Planning and Inference*

F. Willems, Y. Shtarkov, T. Tjalkens 1995 “The Context Tree Weighting Method: Basic Properties,” *IEEE Trans Inform Theory*

Q. Xie and A.R. Barron 1997 “Minimax Redundancy for the Class of Memoryless Sources,” *IEEE Trans Inform Theory*

A.R. Barron, J. Rissanen, B. Yu 1998 “The Minimum Description Length Principle in Coding and Modeling” *IEEE Trans Inform Theory*

A.R. Barron 1998 “Information-Theoretic Characterization of Bayes Performance and the Choice of Priors in Parametric and Nonparametric Problems”, *Bayesian Statistics 6*

J. Takeuchi, T. Kawabata, A.R. Barron 2013 “Properties of Jeffreys Mixture for Markov Sources,” *IEEE Trans Inform Theory*

F. Liang and A.R. Barron 2004 Exact Minimax Strategies for Predictive Density Estimation, Data Compression and Model Selection, *IEEE Trans Inform Theory*

References: Regret of Compression and Online Learning

Y. M. Shtarkov 1988 “Universal Sequential Coding of Single Messages” *Probl Inform Transm*

W. Szpankowski 1995 “On Asymptotics of Certain Sums Arising in Coding Theory” *IEEE Trans Inform Theory*

J. Rissanen 1996 “Fisher Information and Stochastic Complexity,” *IEEE Inform Theory*

A.R. Barron, J. Rissanen, B. Yu 1998 “The Minimum Description Length Principle in Coding and Modeling” *IEEE Trans Inform Theory*

J. Takeuchi, A.R. Barron 1998 “Asymptotically Minimax Regret by Bayes Mixtures” *Proc IEEE Intern Sym Inform Theory*

Q. Xie and A.R. Barron 2000 “Asymptotic Minimax Regret for Data Compression, Gambling and Prediction,” *IEEE Trans Inform Theory*

J. Takeuchi, T. Kawabata, A.R. Barron 2013 “Properties of Jeffreys Mixture for Markov Sources,” *IEEE Trans Inform Theory*

J. Takeuchi, A.R. Barron 2014 “Stochastic Complexity for Tree Models” *Proc IEEE Intern Sym Inform Theory*

J. Takeuchi, A.R. Barron 2024 “Asymptotically Minimax Regret by Bayes Mixtures” arXiv:2406.17929

A.R. Barron, T. Roos, K. Watanabe 2014 “Bayesian Properties of Normalized Maximum Likelihood Computation,” *Proc Intern Sym Inform Theory*

References: Decision Theory of Location Estimation and Linear Regression

E.J.G. Pitman 1939 “The Estimation of Location and Scale Parameters of a Continuous Population of any Given Form” *Biometrica*

M. A. Girshick, L. J. Savage 1951 “Bayes and Minimax Estimates for Quadratic loss functions” *Proc. Berkeley Symp. Math. Stat. and Prob.*

C. Stein 1956 “Inadmissibility of the Usual Estimator for the Mean of a Multivariate Normal Distribution” *Proc. Berkeley Symp. Math. Stat. and Prob.*

W. E. Strawderman 1971 “Proper Bayes Minimax Estimators of the Multivariate Normal Mean” *Annals of Mathematical Statistics*

F. Liang 2002 “Exact Minimax Procedures for Predictive Density Estimation and Data Compression,” Yale Department of Statistics, PhD Dissertation

F. Liang, A.R. Barron 2002 “Exact Minimax Strategies for Predictive Density Estimation, Data Compression and Model Selection” *IEEE Inform Theory*

G. Leung, A.R. Barron 2006 “Information Theory and Mixing Least-Squares Regressions,” *IEEE Trans Inform Theory*

M. L. Eaton, E. I. George 2021 “Charles Stein and Invariance: Beginning with the Hunt-Stein Theorem *Annals of Statistics*

References: Information Stability, Asymptotic Equipartition Property

C. Shannon 1984 “A Mathematical Theory of Communication,” *Bell Systems Tech J*

A. Wald 1949 “Note on the Consistency of the Maximum Likelihood Estimate,” *Annals of Math Statistics*

B. McMillan 1953 “The Basic Theorems of Information Theory,” *Annals of Math Statistics*

L. Breiman 1957 “The Individual Ergodic Theorem of Information Theory,” *Annals of Math Statistics* Correction 1960

A.R. Barron 1985 “The Strong Ergodic Theorem for Densities: Generalized Shanon-McMillan-Breiman Theorem” *Ann Probability*

S. Orey 1985 “On the Shannon-Perez-Moy Theorem,” *Contemporary Mathematics*

References: Additional Highlights of Bayes Computation

R. E. Kalman 1960 “A New Approach to Linear Filtering and Prediction Problems” *Journal of Basic Engineering*

M. West, J. Harrison 1989, 1997 *Bayesian Forecasting and Dynamic Models*, Springer

A. P. Dempster 2001 “Normal belief functions and the Kalman filter,” *Data Analysis from Statistical Foundations*, Nova Science Publ

J. Pearl 1982 “Reverend Bayes on Inference Engines: A Distributed Hierarchical Approach” *Proc National Conf Artificial Intelligence.* AAAI Press

J. Pearl 1988 *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference* Morgan Kauffmann

M. J. Wainwright, M. I. Jordan 2008 *Graphical Models, Exponential Families and Variational Inference* Foundations and Trends in Machine Learning, Now

References: Large Deviations and Information Projection

H. Cramér 1937 *Random Variables and Probability Distributions.* Cambridge

H. Chernoff 1952 "A Measure of Asymptotic Efficiency for Tests of a Hypothesis based on the Sum of Observations" *Ann Mathematical Statistics*

I. N. Sanov 1957 "On the Probability of Large Deviations of Random Variables," *Mat. Sb.* In *Selected Translations in Math Stat Prob* 1961

S. Kullback 1959 *Information Theory and Statistics* Wiley

R. R. Bahadur, R. Ranga Rao 1960 "On the Deviations of the Sample Mean"

Hoeffding 1965 "Asymptotically Optimal Tests for Multinomial Distributions," *Annals Math Statistics*

Jaynes 1982 "On the Rational of Maximum Entropy Methods," *Proc IEEE*

Csiszár 1975 "I-Divergence Geometry of Probability Distributions and Minimization Problems" *Annals of Probability*

Csiszár 1984 "Sanov Property, Generalized I-Projection and a Conditional Limit Theorem," *Annals of Probability*

Csiszár 1991 "Why Least Squares and Maximum Entropy? An Axiomatic Approach to Inference for Linear Inverse Problems" *Annals of Statistics*

J. M. Van Campenhout, T. M. Cover 1981 "Maximum Entropy and Conditional Probability" *IEEE Trans Information Theory*

Reference: Information-Theoretic CLTs and Related Inequalities

A. J. Stam 1959 “Some Inequalities Satisfied by the Quantities of Information of Fisher and Shannon,” *Information and Control*

Y. V. Linnik 1959 “An Information-Theoretic Proof of the Central Limit Theorem with the Lindeberg Condition,” *Theory Probability and its Applications*

L. Gross 1975 “Logarithmic Sobolev Inequalities” *American J. Math*

L. D. Brown 1971 “Admissible Estimators, Recurrent Diffusions, and Insoluble Boundary Value Problems,” *Annals of Mathematical Statistics*

L. D. Brown 1982 “A Proof of the Central Limit Theorem Motivated by the Cramér-Rao Inequality” *Statistics and Probability: Essays in Honor of C. R. Rao*

A. R. Barron 1986 “Entropy and the Central Limit Theorem” *Annals Probability*

O. Johnson, A.R. Barron 2004 “Fisher Information Inequalities and the Central Limit Theorem” *Probability Theory and Related Fields*

S. Arstein, K. M. Ball, F. Barthe, A. Naor “Solution of Shannon’s Problem on the monotonicity of Entropy,” *J. American Math Society*

A. M. Tulino, S. Verd‘u 2006 “Monotonic Decrease of the non-Gaussian-ness of the sum of indpendent random variables: A Simple Proof’,” *IEEE Trans Inform Theory*

M. Madiman, A.R. Barron 2006, “The Monontonicity of Information in the Central Limit Theorem and Entropy Power Inequalities” *Proc Int Symp Inform Theory*

M. Madiman, A.R. Barron 2007 “Generalized Entropy Power Inequalities and Monotonicity Properties of Information” *IEEE Trans Inform Theory*

References: Sparse Regression Codes

A. Joseph, A.R. Barron 2012 "Least Squares Superposition Codes of Moderate Dictionary Size are Reliable at Rates up to Capacity," *IEEE Trans Inform Theory*

A. Joseph, A.R. Barron 2014 "Fast Sparse Superposition Codes have Exponentially Small Error Porbability for $R < C$," *IEEE Trans Inform Theory*

A.R. Barron, S. Cho 2012 "High-Rate Sparse Superposition Codes with Iteratively Optimal Estimates," *Proc IEEE Int Symp Inform Theory*

C. Rush, A. Greig, R. Venkataramanan 2017 "Capacity-Achieiving Sparse Superposition Codes via Approximate Message Passing Decoding," *IEEE Trans Inform Theory*

R. Venkataramanan, S. Tatikonda, A.R. Barron 2019 "Sparse Regression Codes" *Foundations and Trends in Communications and Information Theory*