# Provably Fast and Accurate Estimation of Neural Nets:

## Sampling of Neural Net Posterior Distributions

*The Blessing of Dimensionality*

Andrew R. Barron

YALE UNIVERSITY
Department of Statistics and Data Science

Joint work with
Curtis McDonald (Yale) and Jason Klusowski (Princeton)

International Symposium on Nonparametric Statistics

Braga, Portugal
26 June 2024

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

# Essentials of High-Dimensional Statistical Learning

A. Approximation

B. Estimation

C. Computation

# 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)

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

# Essentials of Sampling of a Neural Net Posterior

## C. Log Concave Coupling for Bayesian Computation

- Focus on single hidden-layer network models
- Prior density $p_0(w)$: Uniform on an $\ell_1$ 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 $\sum_j |c_j| \leq V$
  - $\mathcal{F}_V =$ 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: $||f - f_K||^2 \leq \frac{V(f)^2}{K}$
  - Relative Approximation error: $||f - f_K||^2 - ||f - f^*||^2 \leq \frac{V(f^*)^2}{K}$
  - 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 $||f - f_K||^2 \leq \frac{V_G^2(f)}{K}$
  There are similar bounds for empirical average squares
- Various Algorithmic Terminology
  Sparse term selection, variable selection, forward stepwise regression, relaxed greedy algorithm, orthogonal matching pursuit, Frank Wolf alg, $L_2$ 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 reg, neural networks, ridgelets, sinusoids (paramerized frequencies)
- Neural Network Models
  Single hidden-layer networks, multi-layer networks, deep networks, adaptive learning networks, polynomial networks, residual networks
- Network Units (neurons)
  Sigmoids, Rectified Linear Units (ReLU), low-order polynomials, compositions thereof

# Optional: 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: $\underline{j} = j_1, j_2, ..., j_L$.
- Path weight: $W_{\underline{j}} = w_{j_1,j_2} w_{j_2,j_3} \cdots w_{j_{L-1},j_L}$.
- Function representation:
  $f(x,c,w) = \sum_{j_L} c_{j_L} \psi\big(\sum_{j_{L-1}} w_{j_{L-1},j_L} \psi(...\psi(\sum_{j_1} w_{j_1,j_2} x_{j_1})...)\big)$
- 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, Stat. Risk, Computation

# B. Methods of Bounding Statistical Risk

- 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
    - Gaussian complexity: Ba. Klusowski 19
    - Rademacher complexity: Neshabur et al 15, Golowich et al 18
    - Metric entropy
  - Penalized least squares risk control via relation to MDL

    Adaptive bounds via an index of resolvability: Ba et al 90, 94, 99, 08
  - Concentration of posterior distributions

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

    Clean Information Theoretic bounds: Ba 87,98, Clarke, Ba 90, 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

## 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 $\psi$ Lipshitz 1

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

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:

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

  - Stat Risk of Constrained Least Squares:

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

# Optional: Minimum Description Length and Penalized Likelihood

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

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

- Minimum description-length interpretation when it is at least

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

  for Kraft valid codelengths $L(\omega)$, such that $\sum_\omega 2^{-L(\omega)} \leq 1$
- $\ell_1$ penalities with suitable multipliers are valid
- Battacharya-Renyi risk control via Index of Resolvability

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

  (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)

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

  Also, via the metric entropy bound, with $\ell_1$ weight control

$$E[||\hat{f} - f||^2] \leq cV(f) \Big( \frac{2log(4d)}{N} \Big)^{1/2}$$

- Computationally feasible?

# Optional: Predictive Bayes and its Cumulative Risk Control

- Predictive density $\hat{p}_n(y|x) = \int p(y|x,w)p(w|x^n,y^n)dw$

  Predictive mean $\hat{f}_n(x) = \int f(x,w)p(w|x^n,y^n)dw$

  Predictive evaluations for $Y_{n+1} = y$ when $X_{n+1} = x$
- Information theory chain rule for cumulative Kullback risk: Ba. 87,98

  $$\frac{1}{N}\sum_{n=0}^{N-1} ED(P^*_{Y|X}||\hat{P}^n_{Y|X}) = \frac{1}{N} D(P^*_{Y^N,X^N}||P_{Y^N,X^N})$$

  Controls data compression redundancy and the risk of $\hat{\hat{f}}(x) = \frac{1}{N}\sum_{n=0}^{N-1}\hat{f}_n(x)$

  $$E\left[||\hat{\hat{f}} - f||^2\right] \le \frac{1}{N}\sum_{n=0}^{N-1} E\left[||f - \hat{f}_n||^2\right]$$
- Total Kullback risk controlled by index of resolvability, Ba. 87,98

  $$\begin{aligned}\frac{1}{N} D(P^*_{Y^N,X^N}||P_{Y^N,X^N}) &= \frac{1}{N} E \log \frac{p^*(Y^N,X^N)}{\int p(Y^N,X^N|w)p_0(w)dw} \\ &\le \frac{1}{N} E \log \frac{p^*(Y^N,X^N)}{\int_A p(Y^N,X^N|w)p_0(w)dw} \\ &\le D_A + \frac{1}{N}\log\frac{1}{P_0(A)}\end{aligned}$$

  where $D_A = \max_{w\in A} D(P^*_{Y|X}||P_{Y|X,w})$ is Kullback approximation error
- Predictive risk for neural net estimators with priors uniform on optimal covers

  $$E[||\hat{\hat{f}} - f||^2] \le cV(f)\left(\frac{d\log N}{N}\right)^{1/2}$$ Yang, Ba. 98

  $$E[||\hat{\hat{f}} - f||^2] \le cV(f)\left(\frac{2\log(4d)}{N}\right)^{1/2}$$ Ba., Klusowski 19

  with practical priors and feasibly computable estimates for sufficiently large $d$

  $$E[||\hat{\hat{f}} - f||^2] \le cV(f)^{2/3}\left(\frac{\log(2d)}{N}\right)^{1/3}$$ Ba., McDonald 24, now

# Optional: Arbitrary Sequence Predictive Bayes Regret

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 $\frac{1}{N}\Big\{\log\frac{1}{p(Y^N,X^N)} - \log\frac{1}{p(Y^N,X^N|f)}\Big\}$
  - Corresponds to pointwise regret of an arithmetic code
- Amenable to Laplace approximation and resolvablity bound
- Bounds of the same form

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

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

$$Regret_N \le cV^{2/3}\Big(\frac{\log d}{N}\Big)^{1/3}$$

- Taking expectation controls

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

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

$$E\big[||\hat{\bar{f}} - f||^2\big] \le cV^{2/3}\Big(\frac{\log d}{N}\Big)^{1/3}$$

# C. Bayesian Computation for Neural Nets

- Data: $(X_i, Y_i)$ for $i = 1, 2, \ldots, n$, with $X_i$ in $[-1, 1]^d$ 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 $\psi$ 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 $a_k$

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

- $p(w)$ 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 $\xi$ 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 $\xi$
- 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 $\xi$ 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

$$\tilde{H}(\xi) = \nabla\nabla' \log 1/p(\xi) = \rho \left\{ I - \rho \, Cov \begin{bmatrix} Xw_1 \\ \cdots \\ Xw_K \end{bmatrix} \Big| \, \xi \right\}$$

- 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 $||\tilde{a}||^2 \leq n d$

- Requires 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 $Cov_0(w_m) = \frac{2}{(d+2)(d+1)} I$ and $\rho = \beta c V / K$ to see its at least

$$\rho \left\{ 1 - \frac{2\beta c V n}{K(d+2)} \right\}$$

- Constant $\beta$ chosen such that $\beta c V \leq 1/4$
- Strictly positive when number param $Kd$ exceeds sample size $n$
- Hessian $\geq (\rho/2) I$. Strictly log concave

# Rapid Convergence of Stochastic Diffusion

- Recall the Langevin diffusion

$$d\,\xi(t) = \tfrac{1}{2}\nabla \log p(\xi(t))\,dt \;+\; d\,B(t)$$

- There are time-discretizations (e.g. Metropolis adjusted)
- A natural initialization choice is $\xi(0)$ distributed $N(0, (1/\rho)I)$
- Bakry-Emery theory (initiated in 85)
- Strong log concavity yields rapid Markov process convergence
- In particular, in the stochastic diffusion setting

$$\nabla\nabla' \log 1/p(\xi) \geq (\rho/2)I$$

yields exponential conv. of relative entropy (Kullback distance)

$$D(p_t||p) \leq e^{-t\,\rho/2} D_0$$

- In particular, the time required for small relative entropy is controlled by $\tau = 2/\rho$, here equal to $2K/(\beta c V)$
- Note: with time discretization, one also has a number of draws of $w$ at given $\xi(t)$ to compute the score $\nabla \log p(\xi(t))$, and each such draw requires a number of MCMC steps, with order $nKd$ computation time for each $g_\xi(w)$ evaluation

# 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

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

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

- Hölder's inequality with $r \geq 1$

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

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

  $\leq \frac{c\beta V}{K} \frac{4nr}{de} \exp\{\beta C_V n / r\}$

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

  $= 4c \, V \, C_V \, \frac{\beta^2 n^2}{Kd}$

- Less than $1/2$ when num param $Kd$ exceeds a multiple of $(\beta n)^2$
- Then indeed Hessian $\geq (\rho/2) I$. Strictly log concave

# Optional: 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

# Optional: 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

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

  which is convex in $w$ for each $\xi$, 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 $p(w|\xi)$

# Optional: Variance control 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
  $$[E_0[(\tilde{a} \cdot w)^{2r}]]^{1/r} \exp\{\tfrac{r-1}{r}\Gamma_\xi(\tfrac{r}{r-1}\beta) - \Gamma_\xi(\beta)\}$$
- For the first factor, with integer $r \geq 1$
  $$E_0[(x_i \cdot w)^{2r}] \leq \binom{d+r-1}{r} \frac{(2r)!}{(d+2r)\cdots(d+1)}$$
- Implication
  $$[E_0[(\tilde{a} \cdot w)^{2r}]]^{1/r} \leq n \frac{4r}{ed}$$

# Optional: On the second factor from Hölders inequality

- The exponent of the second factor is
$$\frac{r-1}{r}\Gamma_\xi\left(\frac{r}{r-1}\beta\right) - \Gamma_\xi(\beta)$$
- Not more than $\frac{\beta}{r-1}\max_w \tilde{g}_\xi(w)$ where
$$\tilde{g}_\xi(w) = g_\xi(w) - E_0[g_\xi(w_0)]$$
- It has the bound $\beta \max_{w,w_0}(g_\xi(w) - g_\xi(w_0))/(r-1)$
- 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 $\xi$
- Hence exponent of second factor not more than value near
$$5\,\beta\lambda\,c\,n/r$$

# Optional: Verifying 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 $\xi$ 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

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

# Optional: Combining the two factors

- 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

  $$c\lambda\beta\, 4n\, r/(ed)\, exp\{5\beta\lambda c\, n/r\}$$
- The optimal $r = 5\beta\lambda c\, n$ 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

- Multimodal neural net posteriors can be efficiently sampled
- Log concave coupling provides the key trick
- Requires number of parameters $K\,d$ large compared to the sample size $N$
- Statistically accurate provided $\ell_1$ controls are maintained on the parameters
- 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 the 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

# 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*, Presentation available July 11. Paper soon after.

Additional topically-arranged references 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 $\ell_1$ and $\ell_0$ 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: 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" *Advances in Neural Inform Processing Systems*, 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 $\ell_1$ Penalization for Flexible Function Libraries" Yale Stat

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 $\ell_1$ Penalty and their Statistical Risk" *Workshop on Info Theoretic Methods in Sci and Eng*

# 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

# Transition

The following pages contain draft material for the the first part of the upcoming Shannon lecture in Athens

These pages were not presented at the ISNPS in Braga, Portugal, but they are available answer to certain questions that may arise

# Information Theory and High-Dimensional Bayesian Computation

*The Blessing of Dimensionality*

**Provably Fast & Accurate Neural Net Estimation**

Andrew R. Barron

YALE UNIVERSITY

Department of Statistics and Data Science

Joint work with Curtis McDonald (Yale)

*Shannon Lecture*

International Symposium on Information Theory

Athens, Greece

11 July 2024

You will soon be able to access these slides at

stat.yale.edu/∼arb4/ShannonLecture

# Outline

- The Blessing of Dimensionality

  Trap of Optimization of Multi-Modal Landscapes versus

  Freedom of Posterior Sampling in High Dimensions

- Historical Roots of Bayesian Computation: 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 Risk and Regret

- Inform Theory of Sampling Log-Concave Posterior Distributions

- Beyond Log-Concavity

  - Provably Fast Regression Codes, Achieving Shannon Capacity
  - Provably Computationally-Feasible Posterior Sampling

    for Neural Net Posterior Distributions

    in Sufficiently-High Dimensions

# Historical Highlight of Bayesian Computation: Laplace

- Bayes (1763)
  - Ideas of a prior and posterior distribution
  - Discusses the uniform as an appropriate prior choice
  - Posterior computations not available except in simple cases
- Laplace (1774) Commentary and translation by Stigler (1986)
  - Influential. Dominates statistical science perspective for over a century
  - Uniform prior (without discussion of any other choice)
  - Exact computation in discrete conditionally independent cases of
    - The predictive distrib $p(y_{n+1}|y_1, ..., y_n)$ (a rule of succession)
    - The joint distribution $p(y_1, ..., y_N) = \int p(y_1, ..., y_N|\theta)p(\theta)d\theta$
  - Approximate computation by integration using a normal approx
    - Central limit theory for posterior distributions
    - First appearance of the normal distribution, and $\sqrt{2\pi}$ normalization
  - Optimal estimation of location:
    Median of posterior minimizes posterior expected absolute deviation
  - Precursor to a result of Gauss:
    Posterior mean minimizes the posterior expected square
- Laplace (1812) Included in later *Essays on Probability*
  - Central limit theory for sums of independent random variables
  - Regarded by Gauss as infinite-causes justification of least squares

# 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 $n \leq N$: $p(y_1, ..., y_n) = \int p(y_1, ..., y_n|\theta)\, p(\theta) d\theta$
  - 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
    Average of Kullback loss of predictive distribution is $I(\theta; Y_{n+1}|Y_1, ..., Y_n)$
  - Code redundancy is the total Kullback divergence $D(P_{Y^N|\theta}||P_{Y^N})$
    Code with respect to Laplace joint distribution 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/arithmetic codes par
- Minimax tot Kullback risk = Minimax redundancy = Shannon capacity of $Y^N|\theta$

Laplace approximation in general smooth families with empirical Fisher info $\hat{I}$

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

- Provides approximate computation for prediction, model selection, and codes
- Yields the Bayes factor, and the pointwise regret of MDL, stochastic complexity Ba 85, Clarke, Ba 90,94, Rissanen 96, Takeuchi, Ba 24

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

  for posterior mode $\hat{\theta}$ in the interior of $\Theta$, where $d$ is the parameter dimension

- Expected total divergence rate (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 $p_0(\theta)$ prop to $|I(\theta)|^{1/2}$ is the approx equalizer rule, reference prior
- Approximately mimimax for total Kulback risk and redundancy, Clarke, Ba 94
- Approximately capacity-achieving optimizing asymptotic $I(\theta; Y^N)$, Bernardo 79, Ibragimov, Hasminskii 73, Clarke, Ba 94
- Hartigan 64: Jeffreys prior equalizes prob of small Kullback balls of given radius
- Hartigan 98: Asymptotics of individual Kullback risk

# From Laplace's Rule to Modern Bayesian Computation

Priors on probabilities $\theta$ that permit exact predictive distribution computation and corresponding exact joint distribution computation for arithmetic coding

For discrete memoryless sources with $m$ symbols

- Laplace uniform prior yields computation by Laplace rule of succession
  $$p(y_{n+1} = y|y_1, ..., y_n) = \frac{n_y+1}{n+m} \quad \text{from counts } n_y = \sum_{i=1}^{n} 1_{\{y_i=y\}}$$
  and computation of corresponding Laplace joint distribution
- Dirichlet$(\lambda, ..., \lambda)$ prior produces the prediction rule $\frac{n_y+\lambda}{n+m\lambda}$

  The case $\lambda = 1/2$ has distinguished properties identified by
  - Jeffreys 61: Specialization of his general prior, proportional to $|I(\theta)|^{1/2}$
  - Krichevski, Trofimov 81: Redundancy rate
    $$\frac{m-1}{2N} \log N + O(\tfrac{1}{N})$$
  - Xie, B.97,00: Minimax redundancy and minimax regret
    $$\frac{m-1}{2N} \log \frac{N}{2\pi} + \frac{1}{N} \log \int |I(\theta)|^{1/2} d\theta + o(\tfrac{1}{N})$$

For sources with memory

- Takeuchi, Kawabata, Ba. 02: Jeffreys prior and redundancy for Markov sources
- Willems, Shtarkov, Tjalkens 95: for variable order Markov models, e.g., for text,

  With a flexible prior and a recursive Context Tree Weighting (CTW) algorithm
  - Optimal prediction, compression, text generation for their prior & posterior
  - Scale up CTW at word level should yield competitive large language model

# 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 minimax-regret solution: $q(y^n) = \max_\theta p(y^n|\theta)/c_n$
  This is the normalized maximum likelihood championed by Rissanen
- Apparently, it is a not Bayes-Laplace mixture
- So how to compute predictive distributions for arithmetic coding?
- 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 for this $q(y^n)$!
  May evaluate its positive marginals and predictive distributions
- Negative prior probabilities!
  *These priors yield computation of positive-valued quantities*
  *for optimal prediction & compression quantities.*
  *They are not for prior subjective assessment*
- Here $y^n$ has an exponentially numerous domain. Fortunately, the set of values of sufficient statistics (e.g. counts) may be more moderate-sized, and the number of $\theta_j$ can be arranged accordingly
- Practical minimax-regret optimal data compression

# 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
  Also recognized in a paper by Legendre (1805)
- Gauss elimination solution iterative in the number of parameter

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
  - For independent random variables the variance of a sum is the sum of the variances. Leads to evaluation of $var(\hat{w}_j)$ and the standard error

# From Gauss to Modern Bayesian Computation

Later work: Gauss (1823) also noted

- The least squares solution is unbiased
- Least squares solution has smallest variance among linear unbiased estimators

Other Linear Model Conclusions credited to Gauss & Laplace, with normal $\phi$

- Least square solution is also the post mean, optimizes posterior expected square
- Least squares methods provide the 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 $E[Y|x, Data] = \int w \cdot x\, p(w|Data)\, dw = \hat{w} \cdot x$
- Gauss' recursive least squares yields solution iterating one observation at a time

Linear Filtering and Prediction

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

Model Selection and Data Compression: computation of Bayes factors

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

Minimax Estimation and Compression for linear models, general $\phi$

- The Uniform prior yields the minimax optimal procedure for
  - param estim with squared error loss (Hunt-Stein xx, 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)

# Optional Page: 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

$$\frac{1}{2}x^2 + c \qquad \text{and} \qquad x$$

- 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): $D(X) \le \frac{1}{2} J(X)$ Stam 57, Gross 75
  - Stochastic diffusion distribution properties with Gaussian limit

# From Gaussian to Log-Concave Distributions

- Summary thus far:
  Laplace and Gauss performed required normal distribution integrations
  in their linear models to compute the posterior optimal procedures
- What is the right extension
  to preserve rapid computation of high-dimensional posterior integrals?
- Main Generalized Setting of the Last Forty Years: Log-Concavity
  MCMC samplers: Accurate and Mmx rapidly for log concave posteriors
- Implication: Rapid computation of Minimax Optimal Procedures
  Minimax optimal location estimation, linear regression and minimax redundancy codes in lin predictive setting are low-order polynomial-time computable for any log-concave error distribution
- What about regressions with non-convex domains?
- What about non-linear regressions, such as neural networks?

# Information Theory of Rapid MCMC with Log Concavity

- Langevin Diffusion Path for sample parameter values $w(t)$

$$d\,w(t) = \frac{1}{2}\nabla \log p(w(t))\,dt \;+\; d\,B(t)$$

- Remarks:
  - Score $\nabla \log p(w)$ is non-linear in general
  - There are time-discretizations (e.g. Metropolis adjusted)
  - A natural initialization choice is $w(0)$ distributed $N(0, (1/\rho)I)$
- 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

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

  yields exponential conv. of relative entropy (Kullback distance)

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

- Time required for small relative entropy is controlled by $\tau = 1/\rho$
- Proof uses $D(p_t||p) = \frac{1}{2}\int_{\tau \geq t} J(p_\tau||p)\,d\tau$
- And demonstrates the Log Sobolev Ineq: $D(p_t||p) \leq \frac{1}{2\,\rho} J(p_\tau||p)$
  where $J$ is the mean square norm between the scores

# 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

the additive Gaussian noise channel with a specified power control

Capacity-achieving Sparse Superposition 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 $w_k$ concentrating on columns sent with high prob, after a logarithmic number of steps

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

# 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*, 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 $\ell_1$ and $\ell_0$ 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: 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" *Advances in Neural Inform Processing Systems*, 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 $\ell_1$ Penalization for Flexible Function Libraries" Yale Stat

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 $\ell_1$ Penalty and their Statistical Risk" *Workshop on Info Theoretic Methods in Sci and Eng*

# 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. M. Laplace 1774 *Mémoire sur la Probabilité des Causes par les Évènmens*, with commentary and translation by Stigler 1986 *Statistical Science*

P. S. M. de 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*

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

S. M. Stigler 1990 *The History of Statistics: The Measurement of Uncertainty before 1900*, Belknap Press

# 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*

# Refs: MDL, Redundancy & Total Kullback Risk of Bayes Proc


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*

# Reference: 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” Currently being put on ArXiv

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

# 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*

# Refs: 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

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