Rapid Bayesian Computation & Estimation for Neural Nets via Log Concave Coupling

The Blessing of Dimensionality

Andrew R. Barron and Curtis McDonald

YALE UNIVERSITY
Department of Statistics and Data Science

Joint Colloquium of the Princeton University Departments of
Operations Research & Financial Engineering
and Electrical & Computer Engineering
18 February 2025

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

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

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

  Rate $r = 1/4$ with K neuron posterior; $r = 1/3$ with greedy Bayes

  Number of neurons K of order $[N/\log(2d)]^r$

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 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 $||f - f_K||^2 \leq \frac{V_G^2(f)}{K}$
  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 bound: Ba 87,98, Clarke,Ba 90, Yang,Ba 98, Ba, Klusowski 19, Ba, McDonald 24,25
  - 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 ψ 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:

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

  - 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)

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

- 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

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

- 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 $\hat{\hat{f}}(x) = \frac{1}{N}\sum_{n=0}^{N-1}\hat{f}_n(x)$

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

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

  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

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

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

  with practical priors and feasibly computable estimates for sufficiently large d

  $$E[||\hat{\hat{f}} - f||^2] \leq cV(f)^{2/3}\left(\frac{\log(2d)}{N}\right)^{1/3} \qquad \text{Ba., McDonald 24, 25}$$

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 $\frac{1}{N}\left\{\log\frac{1}{p(Y^N,X^N)} - \log\frac{1}{p(Y^N,X^N|f)}\right\}$
  - 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 $F_{1,V}$

$$Regret_N \le cV^{3/4}\left(\frac{\log d}{N}\right)^{1/4}$$

- 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

$$E\left[||\hat{\hat{f}} - f||^2\right] \le cV^{3/4}\left(\frac{\log d}{N}\right)^{1/4}$$

Rate becomes 1/3 with greedy predictive Bayes

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) indep $P_{X,Y}$, target $f(x) = E[Y \mid X = x]$, variance $\sigma^2_{Y|x} \leq \sigma^2$
  - 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:
  $p(\xi|w) = \big(\rho/2\pi\big)^{Kn/2} exp\{-\frac{\rho}{2}\sum_{i=1}^{n}\sum_{k=1}^{K}(\xi_{i,k} - x_i \cdot w_k)^2\}$
- 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
  $g_\xi(w) = g(w) + \frac{1}{2}\frac{V}{K} c \sum_{i=1}^{n}\sum_{k=1}^{K}\big(x_i \cdot w_k - \xi_{i,k}\big)^2$
- 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

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

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

- 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 β 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
  $$\rho Var[\tilde{a}\cdot w|\xi] \le \rho \int (\tilde{a}\cdot w)^2 \exp\{-\beta\tilde{g}_\xi(w) - \Gamma_\xi(\beta)\} p_0(w)dw$$
  using $\tilde{g}_\xi(w) = g_\xi(w) - E_0[g_\xi(w)]$
- Hölder's inequality with $r \ge 1$
  $$\le \rho\,[E_0[(\tilde{a}\cdot w)^{2r}]]^{1/r} \exp\{\tfrac{r-1}{r}\Gamma_\xi(\tfrac{r}{r-1}\beta) - \Gamma_\xi(\beta)\}$$
  which is, using a bound $C_V n$ on $g_\xi(w)$ with $C_V = 9V^2 + 7V \max_i |Y_i|$,
  $$\le \frac{c\beta V}{K}\frac{4nr}{de}\exp\{\beta C_V n/r\}$$
  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 $\ge (\rho/2)I$. 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 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 $p(w|\xi)$

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 ξ
- 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 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
$$5\beta\lambda c\, n/r$$

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 ℓ_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 2025 "Rapid Bayesian Computation and Estimation of Neural Networks by Log-Concave Coupling," ArXiv, Feb 2025, and prepared for submission to *Mathematical Statistics and Learning*

A.R. Barron 2024 "Information Theory and High-Dimensional Bayesian Computation", Shannon Lecture, *IEEE International Symposium on Information Theory*

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 and C. McDonald 2024 "Log Concave Coupling for Sampling From Neural Net Posterior Distributions," *Proc. IMS-NUS Singapore Workshop on Statistical Machine Learning for High Dimensional Data*

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 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: 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 ℓ_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 ℓ_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