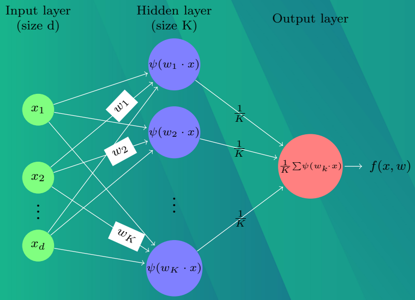
Log Concave Coupling for Sampling Neural Net Posteriors

Curtis McDonald, Andrew R. Barron

Department of Statistics and Data Science,
Yale University

International Symposium on Information Theory
Athens, Greece
July 11, 2024

Problem Definition



- Wide single hidden layer neural network
- Outer weights fixed, linear combination of neurons from infinite dictionary
- Train inner weights w not by optimization:

$$w = \text{argmax}_w L(\text{Data}, w)$$

  but via posterior sampling:

$$w \sim p(w|\text{Data}) \propto e^{L(\text{Data}, w)}$$

- Overall fit is posterior mean,

$$\hat{f}(x) = E[f(x, w)|\text{Data}]$$

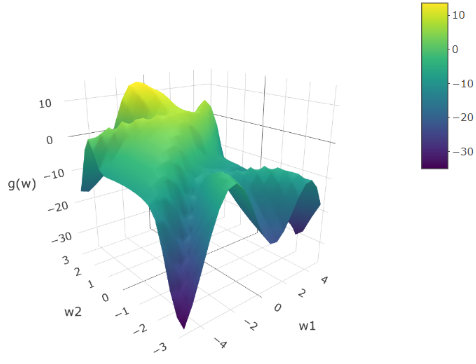
Target Density

- Posterior density priorities:
  i) Predictive risk control, $E[\|\hat{f} - f\|^2]$
  ii) Computational feasibility (i.e. rapid MCMC sampling)
- Train network in a Greedy fashion, add in one neuron at a time
- Previous fit $f_k(x)$ using k neurons, add in new neuron w_{k+1}
- Prior density $p_0(w)$ uniform over ℓ_1 ball, $\|w\|_1 \leq 1$
- For some $\alpha \in [0, 1]$, residuals $r_i = y_i - (1 - \alpha)f_k(x_i)$
- For some gain $\beta > 0$,

$$p_{k+1}(w) \propto e^{\beta \sum_{i=1}^{n} r_i \psi(x_i \cdot w)} p_0(w)$$
$$f_{k+1}(x) = (1 - \alpha)f_k(x) + \beta E_{p_{k+1}}[\psi(x \cdot w)]$$

- **Assume**:
  1) Bounded data entries $|x_{i,j}| \leq 1$ for all i, j
  2) Activation function ψ has bounded second derivative, $|\psi''(u)| \leq c$, example tanh or squared ReLU.

Difficulties Sampling and Multi-Modal Posterior



- Posterior log likelihood:

  $$g(w) = \sum_{i=1}^{n} r_i \, \psi(x_i \cdot w)$$

  $$p(w) \propto e^{\beta g(w)}$$

- (Jones, Ba. 93) greedy pursuit, maximize $g(w)$ to find best next neuron

- However, $g(w)$ is multi-modal and not concave

- $\nabla^2 g(w) = \sum_{i=1}^{n} r_i \, \psi''(x_i \cdot w) x_i x_i^T$ is sum of rank 1 matrices with $\pm$ scaling, not negative definite (i.e. not concave)

Auxiliary Random Variable and Joint Density

- Auxiliary r.v.'s $\xi_i \sim \text{Normal}(\sqrt{\beta |r_i| c}\, x_i \cdot w\ , 1),\ i = 1, \ldots, n$
- Joint density $p(w, \xi) = p(w)p(\xi | w)$
- Pair $(w, \xi) \sim p(w, \xi)$ has correct w marginal $w \sim p(w)$
- By Bayes rule, also have $p(w, \xi) = p(\xi)p(w|\xi)$
- $p(\xi)$: marginal of auxiliary r.v., convolution with normal
- $p(w|\xi)$: reverse conditional, has addition negative definite terms in log likelihood Hessian (will make log concave)
- Goal: first sample a $\xi \sim p(\xi)$ and then sample a $w \sim p(w|\xi)$. If both densities are log concave, MCMC methods will be rapidly mixing

Log Concavity of Reverse Conditional Density

$$
\begin{aligned}
p(w|\xi) &\propto p(w)p(\xi|w) \\
&\propto \exp\left(\beta \sum_{i=1}^{n} r_i \psi(x_i \cdot w) - \frac{1}{2}(\xi_i - \sqrt{\beta |r_i| c}\, x_i \cdot w)^2\right) p_0(w) \\
&\propto \exp\left(\beta \sum_{i=1}^{n} r_i \psi(x_i \cdot w) - \beta |r_i| c \frac{1}{2}(x_i \cdot w)^2 + \sqrt{\beta |r_i| c}\, \xi_i x_i \cdot w\right) p_0(w)
\end{aligned}
$$

$$
\nabla^2 \log p(w) = \beta \sum_{i=1}^{n} \left(r_i \psi''(x_i \cdot w)\right) x_i x_i^T
$$

$$
\nabla^2 \log p(w|\xi) = \beta \sum_{i=1}^{n} \left(r_i \psi''(x_i \cdot w) - |r_i| c\right) x_i x_i^T
$$

- Negative definite term introduced to Hessian
- By assumption $r_i \psi''(x_i \cdot w) - |r_i| c < 0$
- Reverse conditional now log-concave
- Log concave density over convex set rapidly sampled by existing MCMC methods (Applegate, Kannan 91, Lovász, Vempala 07)

Marginal Density of Auxiliary RV

- Marginal density $p(\xi)$ is convolution of target $p(w)$ with normal $p(\xi|w)$
$$p(\xi) = \int p(w)p(\xi|w)dw$$
- Denote $g(w) = \beta \sum_{i=1}^{n} r_i \psi(x_i \cdot w) - \beta |r_i| c \frac{1}{2}(x_i \cdot w)^2$ (concave modified target)
$$\log p(\xi) = -\frac{1}{2}\|\xi\|^2 + \log\left(\int p_0(w) e^{g(w)} e^{\sum_{i=1}^{n} \sqrt{\beta |r_i| c}\, \xi_i x_i w} dw\right)$$
- Negative quadratic (concave, neg def hessian)
- Cumulant generating function (convex, pos def hessian)
- Competition for overall log concavity

Score of Auxiliary RV

- Score, define $|R|$ as diagonal matrix of residuals:

$$\nabla \log p(\xi) = -\xi + E[\sqrt{\beta c}|R|^{\frac{1}{2}} Xw|\xi]$$

- MCMC algorithms need access to score (e.g. Langevin Diffusion)
- Score defined by expectation over log concave density $p(w|\xi)$
- Score estimated empirically via MCMC method on reverse conditional $p(w|\xi)$

Log Concavity of Auxiliary RV

- Hessian:

$$\nabla^2 \log p(\xi) = -I + \mathrm{Cov}[\sqrt{\beta c}|R|^{\frac{1}{2}} X w | \xi]$$

- Log concave if for all unit vector a, direction $v = \sqrt{\beta c}|R|^{\frac{1}{2}} X^T a$

$$\mathrm{Var}(v \cdot w | \xi) \leq 1$$

- Upper bound on maximum eigenvalue of $\mathrm{Cov}[w|\xi]$, i.e. max variance in any direction sufficiently small

Covariance Under Prior

- Want $\mathrm{Var}[v \cdot w | \xi]$
- Consider first prior density $\mathrm{Var}_{p_0}[v \cdot w]$
- Prior $w \sim \mathrm{Uniform}\{w \ s.t. \|w\|_1 \leq 1\}$
- $\mathrm{Cov}_{p_0}[w] \preceq \frac{1}{d^2} I$
- If w drawn from prior, $\mathrm{Cov}_{p_0}[\sqrt{\beta c}|R|^{\frac{1}{2}} X w] \preceq \frac{c\|r\|_\infty \beta n}{d} I$
- **Intuition from prior:** large dimension d can control covariance, d of order $c\|r\|_\infty \beta n$.

Two Proof Methods

1. Contraction Conjecture (not proven):

- Conditional covariance LESS than prior,

$$\mathrm{Cov}[w|\xi] \preceq \mathrm{Cov}_{p_0}[w]$$

- Condition for log concavity: $\frac{c\|r\|_\infty \beta n}{d} < 1$
- With $\beta = \frac{1}{\sqrt{n}}$ need dimension

$$(c\|r\|_\infty)\sqrt{n} < d$$

2. Hölder Inequality Lemma:

- Any direction $v = \sqrt{\beta c}|R|^{\frac{1}{2}} X^T a$,

$$\mathrm{Var}(v \cdot w|\xi) < 20\frac{(c\|r\|_\infty \beta n)^2}{d}$$

- Condition for log concavity, $20\frac{(c\|r\|_\infty \beta n)^2}{d} < 1$
- With $\beta = \frac{1}{\sqrt{n}}$ need dimension

$$20(c\|r\|_\infty)^2 n < d$$

Contraction Conjecture

- $p(w|\xi) \propto p_0(w)e^{\beta g_\xi(w)}$, $g_\xi(w)$ is concave in w
- Would seem to reduce variance in any direction
- When prior $p_0(w)$ is over all of $\mathbb{R}^d$, this is true
- One dimensional marginal in any direction $u = v \cdot w$

$$\frac{d^2}{du^2} \log p(u|\xi) < \frac{d^2}{du^2} \log p_0(u)$$

- One dimensional optimal transport map is a contraction
- Less variance in any direction

- Does NOT hold when $p_0(w)$ restricted to convex set (e.g. uniform over ℓ_1 ball)
- Examples where concave function in one direction increases variance in another
- Example,

$$p(w_1, w_2) \propto e^{-\beta w_2^2} 1\{|w_1| + |w_2| \leq 1\}$$

Increased β INCREASES variance in w_1 direction

Hölder Inequality Proof Sketch

- $\mathrm{Var}(v \cdot w|\xi)$ is not more than

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

where $\tilde{g}_\xi(w)$ is $g_\xi(w)$ minus it's mean under $p_0(w)$

- $\Gamma_\xi(w)$ is cumulant generating function of $\tilde{g}_\xi(w)$
- $\Gamma_\xi(w)$ concave, $\Gamma_\xi(0) = 0, \Gamma'_\xi(0) = 0$
- By Hölders inequality variance is not more than

$$(E_{p_0}[(v \cdot w)^{2\ell}])^{\frac{1}{\ell}} \exp\left\{ \frac{\ell}{\ell - 1} \Gamma_\xi\left(\frac{\ell}{\ell - 1}\beta\right) - \Gamma_\xi(\beta) \right\}$$

- Moments of prior analysis, $(E_{p_0}[(v \cdot w)^{2\ell}])^{\frac{1}{\ell}} < c\|r\|_\infty \beta n^{\frac{4\ell}{ed}}$
- CGF analysis, $\frac{\ell}{\ell-1}\Gamma_\xi(\frac{\ell}{\ell-1}\beta) - \Gamma_\xi(\beta) < c\|r\|_\infty \beta n^{\frac{5}{\ell}}$
- Optimize over ℓ to get bound $20\frac{(c\|r\|_\infty \beta n)^2}{d}$

Summary

- Train single hidden layer NN via sequential Greedy Bayes, residuals $r_i = y_i - (1-\alpha) f_k(x_i)$

$$p_{k+1}(w) \propto e^{\beta \sum_{i=1}^{n} r_i \psi(x_i \cdot w)}$$
$$f_{k+1}(x) = (1-\alpha) f_k(x) + \beta E_{p_k}[\psi(x \cdot w)]$$

- Define joint density $p(w, \xi)$ with correct w marginal
- $p(w|\xi)$ log concave and can be sampled via MCMC
- Marginal $p(\xi)$ has score which can be estimated via MCMC sampling over $p(w|\xi)$:

$$\nabla \log p(\xi) = -\xi + E[\sqrt{\beta c}|R|^{\frac{1}{2}} X w | \xi]$$

- $p(\xi)$ is log concave for large dimension d:
  - (Conjectured) Log-concave when $c\|r\|_\infty \beta n < d$
  - (Proven) Log-concave when $20(c\|r\|_\infty \beta n)^2 \leq d$

Future Work: Predictive Risk Control

- Statistical risk or generalization squared error: $E[\|\hat{f} - f\|^2]$
- Use multiple subsets of data $1 \leq n \leq N$
  - Predictive density: $p_n(y|x) = \int p(y|x, w)p(w|x^n, y^n)dw$
  - Predictive mean: $\hat{f}_n(x) = \int f(x, w)p(w|x^n, y^n)dw$
  - Cumulative estimator: $\hat{\hat{f}}(x) = \frac{1}{N}\sum_{n=1}^{N} \hat{f}_n(x)$
- Upper bound with Kullback divergence,

$$E[\|\hat{\hat{f}} - f\|^2] \leq \frac{1}{N} D(P^*_{Y^N,X^N} \| P_{Y^N,X^N})$$

- Index of resolvability (Ba. 87, 98):

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

- Prior probability control $P_0(A)$, Gaussian not sufficient, need uniform prior
- Best bounds to date: $E[\|\hat{\hat{f}} - f\|^2] \leq C\left(\frac{(\log 2d)}{N}\right)^{\frac{1}{3}}$

THE END

References

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

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*

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*

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

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