Math. Stat. Learn. 9 (2026), 199–246
DOI 10.4171/MSL/59

Rapid Bayesian computation and estimation for neural networks via log-concave coupling

Curtis McDonald and Andrew R. Barron

Abstract. This paper presents the study of a Bayesian estimation procedure for single-hidden-layer neural networks using ℓ_1 controlled neuron weight vectors. We study the structure of the posterior density and provide a representation that makes it amenable to rapid sampling via Markov Chain Monte Carlo (MCMC). Let the neural network have K neurons with internal weights of dimension d and fix the outer weights. Thus, there are Kd parameters overall. With N data observations, use a gain parameter or inverse temperature of β in the posterior density for the internal weights.

The posterior is intrinsically multi-modal and not naturally suited to rapid mixing of direct MCMC algorithms. For a continuous uniform prior on the ℓ_1 ball, we demonstrate that the posterior density can be written as a mixture density with suitably defined auxiliary random variables, where the mixture components are log-concave. Furthermore, when the total number of model parameters Kd is large enough that $Kd \geq C(\beta N)^2$, the mixing distribution of the auxiliary random variables is also log-concave. Thus, neuron parameters can be sampled from the posterior by only sampling log-concave densities. The authors refer to the pairing of weights with such auxiliary random variables as a log-concave coupling.

1. Introduction

Single-hidden-layer artificial neural networks provide a flexible class of parameterized functions for data fitting applications. Specifically, denote a single-hidden-layer neural network as the parameterized function

$$f_w(x) = f(w, x) = \sum_{k=1}^{K} c_k \psi(w_k \cdot x), \tag{1.1}$$

with K neurons, activation function ψ, and interior weights $w_k \in \mathbb{R}^d$. Fix a positive scaling V and let the exterior weights c_k be positive or negative values $c_k \in \{-V/K, V/K\}$. Thus, $f_w(x)$ is a convex combination of K signed neurons scaled

Mathematics Subject Classification 2020: 62F15 (primary); 62M45, 65C05 (secondary).
Keywords: neural networks, Bayesian methods, sampling, statistical learning.

by V. Constant and linear terms $c_0 + w_0 \cdot x$ may be added in the definition of $f_w(x)$ to achieve additional flexibility, though we will not address that matter explicitly.

While the class of functions which can be modeled well by a shallow neuron network is well understood, computational guarantees for algorithms to fit specific neural networks have been lacking. The approximation ability of these networks has been studied for many years, which we briefly review here. Assume input vectors $x \in \mathbb{R}^d$ as having bounded entries, $x \in [-1, 1]^d$, and interior weights have bounded ℓ_1 norm $\|w_k\|_1 \leq 1$. Thus the input to each neuron is between $[-1, 1]$.

The early work of [6] showed that moderately wide single-hidden-layer networks with sigmoid activation functions can accurately approximate target functions with a condition on the Fourier components of the target function. For a sigmoid activation function and K neurons, the squared error with the target function was shown to be on the order of $O(1/K)$.

Denote the set of signed neurons with ℓ_1 controlled interior weights as the collection of functions $h: [-1, 1]^d \to \mathbb{R}$ as follows:

$$\Psi = \{h : h(x) = \pm\psi(w \cdot x), \|w\|_1 \leq 1\}.$$

The closed convex hull of Ψ includes functions f which can be written as a possibly infinite mixture of signed neurons, and functions which are the limit of a sequence of such mixtures. Specializing the results of [5,6,45], networks of the form (1.1) provide accurate approximation for functions f with f/V in the closure of the convex hull of Ψ. The infimum of such V is called the variation V_f of the function f with respect to the dictionary Ψ.

Assume a data set of size N of iid data $(x_i, y_i)_{i=1}^N$ and $E[y_i|x_i] = f(x_i)$ with $f(x)$ in the closure of the convex hull of Ψ. Results in [10] show, for network of width $K = O([N/\log(d)]^{1/2})$, the empirical risk minimizer (ERM) $f_{\hat{w}}$ yields a statistical risk control of the order

$$E\big[\|f_{\hat{w}} - f\|^2\big] = O\left(\left(\frac{\log(d)}{N}\right)^{1/2}\right),$$

provided there is sub-Gaussian control of the distribution of the response Y. The expectation here is with respect to the training data, while the norm square provides the expectation for the loss at an independent new input vector. Analogous deep net conclusions are also in [9, 10].

Given that the ERM can achieve such risk control, there is much interest to understand theoretically the optimization properties of neural networks by gradient-based properties. One line of investigation as in [16, 31, 44, 51, 66] compares the network to a certain infinite width limit under initialization and scaling assumptions (called the neural tangent kernel, NTK), where the network trained under gradient descent are

shown to approach a kernel regression solution linear in the response vector y. They also utilize a scaling of $1/\sqrt{K}$ on their outer weights rather than the $1/K$ scaling we use.

With finite $K < N$ and outer weights that scale as $c_k = 1/K$, even in the single-hidden-layer case no known optimization algorithm is able to solve this optimization problem in a polynomial number of iterations in N and d. Instead, in this work we move away from an optimization approach to choosing neuron parameters and use a Bayesian method of estimation placing a posterior distribution on neuron parameters.

Bayesian neural networks have been studied for many years [18,34,59], although specific mixing time bounds for Markov Chain Monte Carlo (MCMC) to guarantee polynomial time complexity have been a barrier to their implementation. Recent approaches have studied the simplification of the posterior in the NTK regime, resulting in the posterior being near the posterior associated with a Gaussian process prior [37,40]. These approaches require $K/N \to \infty$ to achieve that simplification of the posterior density. The bounded K/N setting is shown in [37,40] to be distinct with potentially more flexible non-Gaussian process behavior. Indeed, such flexibility arises in our model where the internal weights are adapted by the posterior.

We will quantify when sampling can be accomplished in polynomial time with MCMC. Say we have data consisting of N input and response pairs $(x_i, y_i)_{i=1}^N$. Define a prior distribution P_0 on neuron parameters and a gain or inverse temperature $\beta > 0$. Define the posterior density

$$p\big(w|x^N, y^N\big) \propto p_0(w) e^{-(\beta/2)\sum_{i=1}^N (y_i - f(w,x_i))^2},$$

and the associated posterior mean at a given x value $\mu(x) = E_P[f(x,w)|x^N, y^N]$.

The barrier to implementing the Bayesian approaches defined above is being able to sample from the density $p(w|x^N, y^N)$, and thus compute the posterior average $\mu(x)$. The densities $p(w|x^N, y^N)$ will be high-dimensional and multi-modal densities with no immediately evident structure that would make sampling possible.

The neural network model has Kd parameters and we have N data observations. A natural method to compute the required posterior averages would be a MCMC sampling algorithm. A key issue for an MCMC method is whether it provably gives accurate sampling in a low polynomial number of iterations in K, d, N. Any exponential dependence on the parameters of the problem is not practically useful.

The most common sufficient condition for proving rapid mixing of MCMC methods is log-concavity of the target density [3, 4, 26, 53]. As such, we want to find a representation of the problem built from log-concave densities, so that we may restrict our computation task to only require sampling from log-concave densities.

We show that with the use of an auxiliary random variable ξ, the posterior density $p(w|x^N, y^N)$ can be re-written as a mixture density (also called a measure decom-

position in the language of [58])

$$p(w|x^N, y^N) = \int p(w|\xi, x^N, y^N) p(\xi|x^N, y^N)\, d\xi,$$

using a reverse conditional density $p(w|\xi) = p(w|\xi, x^N, y^N)$ and an induced marginal density $p(\xi) = p(\xi|x^N, y^N)$. When considering a fixed input and response sequence, we will drop the x^N and y^N conditioning notation. For a certain choice of priors and relationships between d, K, N, β, we show the reverse conditional is a log-concave density, and the induced marginal for ξ is also a log-concave density. We call such a joint distribution $p(w, \xi)$ that preserves the target marginal $p(w)$ and has a log-concave marginal distribution $p(\xi)$ and a log-concave conditional distribution $p(w|\xi)$ a *log-concave coupling*. As such, samples for w from the posterior can be produced by merely sampling from log-concave densities: that is, we sample from the density of ξ followed by sampling from the density of w given ξ.

For a continuous uniform prior on the ℓ_1 ball, we demonstrate the mixture is a log-concave coupling when the total number of parameters Kd is large enough such that

$$Kd \geq C(\beta N)^2,$$

for a given constant C that depends only on the range of data values and the scaling V of the network.

We presume access to a sampling algorithm able to produce accurate samples from a log-concave density in a number of iterations proportional to a low polynomial power of the number of model parameters [26, 47, 48, 52, 53]. We leave the specific choice of this algorithm in our setting (e.g. Metropolis Adjusted Langevin Diffusion (MALA), Hamiltonian Monte Carlo, etc.) as well as the tuning of parameters as a technical study for future work, and treat the sampling algorithm as a black box method available to the user. Then one can sample a value for ξ from its marginal, and a value $w|\xi$ from its reverse conditional, resulting in a true draw from the posterior distribution for w.

2. Notation

Here we present the mathematical notation used in the paper.

- Capital P refers to a probability distribution, while lowercase p is its probability mass or density function.
- $f'(\cdot)$ refers to the derivative of a scalar function f.
- ∇ is the gradient operator and ∇^2 is the Hessian operator, producing a matrix of second derivatives.

- $\{1, \ldots, N\}$ is the set of whole numbers between 1 and N.
- $[a, b]$ is the interval of real values between a and b.
- $u \cdot v$ is the Euclidean inner product between two vectors.
- $u^{\mathrm{T}}, \mathbf{X}^{\mathrm{T}}$ refers to the transpose of a vector or matrix, so quadratic forms of a column vector u with a matrix $\mathbf{X}$ will be written as $u^{\mathrm{T}}\mathbf{X}u$.
- $\|w\|_p$ refers to the ℓ_p norm, $\|w\|_p = (\sum_j (w_j)^p)^{1/p}$.
- The ℓ_1 ball is denoted as $S_1^d = \{w \in \mathbb{R}^d : \|w\|_1 \leq 1\}$.
- The K fold Cartesian product of this set is $(S_1^d)^K$.
- For variables in a sequence, the set of variables $X^N = (X_i)_{i=1}^N$ is indicated by superscripts.
- Logarithms in the paper are natural logarithms.

3. Bayesian model and result

Consider input and response pairs $(x_i, y_i)_{i=1}^N$ where the x_i input vectors are d-dimensional and the response values y_i are real valued. Consider the x_i as being bounded by 1 in each coordinate, $|x_{i,j}| \leq 1$ for all $i \in \{1, \ldots, N\}$, $j \in \{1, \ldots, d\}$. Further, assume $x_{i,1} = 1$ for all $i \in \{1, \ldots, N\}$ so an intercept term is naturally included in the data definition. Accordingly, this requires $d \geq 2$. Denote $\mathbf{X}$ as the N by d data matrix which uses the x_i as its rows.

Recall the definition of a single-hidden-layer neural network in (1.1). We restrict the class of neuron activation functions we consider to have two bounded derivatives with $\psi(0) = 0$, $|\psi(z)| \leq a_0$, $|\psi'(z)| \leq a_1$ and $|\psi''(z)| \leq a_2$ for all $z \in [-1, 1]$. We assume $a_0, a_1, a_2 \geq 1$. This includes for example the squared ReLU activation function $\psi(z) = a(z_+)^2$ or the scaled tanh activation function $\psi(z) = a \tanh(cz)$.

Fix a positive V and let the exterior weights c_k be positive or negative values $c_k \in \{-V/K, V/K\}$. Thus, $f_w(x)$ is a convex combination of K signed neurons scaled by V. Note, if ψ is odd symmetric as in the case of the tanh activation function, the c_k can be all set to positive V/K. For non-symmetric activation functions, we can use twice the variation $\widetilde{V} = 2V$ and use twice the number of neurons $\widetilde{K} = 2K$. For the first K neurons set $c_k = V/K$ and for the second set of K neurons set $c_k = -V/K$. Under such a structure using $2K$ neurons, any size K network of variation V with any number of positive or negative signed neurons can be constructed from the wider network by setting K of the neurons to be active and the other K to be inactive with 0 for their interior weight vector. In either case, we consider the outer weights c_k to be fixed values, and it is only necessary to train the interior weights w_k of the network.

Define P_0 as a prior measure on $\mathbb{R}^{Kd}$, with density p_0 with respect to a reference measure η (e.g. Lebesgue or counting measure). For each index $i \in \{1, \ldots, N\}$, define the residual of a neural network as

$$\text{res}_i(w) = y_i - \sum_{k=1}^{K} c_k \psi(w_k \cdot x_i).$$

Define the loss function as half the sum of squares of residuals

$$\ell(w) = \frac{1}{2} \sum_{i=1}^{N} \big(\text{res}_i(w)\big)^2.$$

For any gain parameter $\beta > 0$, we define the posterior density as

$$p(w) = \frac{p_0(w) e^{-\beta \ell(w)}}{\int e^{-\beta \ell(w)} p_0(w)\, \eta(dw)}. \tag{3.1}$$

Denote the posterior mean with respect to this density as

$$\mu(x) = E_P\big[f(w, x)\big].$$

3.1. Choice of prior

The prior we will consider is the uniform on the set $(S_1^d)^K$. That is, independently each weight vector w_k is iid uniform on the set of weight vectors with ℓ_1 norm less than 1. This has the density function

$$p_0(w) = \prod_{k=1}^{K} \left(1\{\|w_k\|_1 \le 1\} \frac{1}{\text{Vol}(S_1^d)} \right),$$

with respect to Lebesgue measure. Note that for each k, the vector of absolute values of the coordinates of w_k is uniformly distributed on the simplex, which is also a symmetric Dirichlet distribution in $d + 1$ dimensions with the all 1's parameter vector.

The authors would like to provide some context as to why the above prior is chosen. The choice of the continuous uniform prior is influenced by previous work using a discrete uniform prior. Consider a discrete version of this density. For some positive integer $M \le d$, consider the discrete set which is the intersection of S_1^d with the lattice of points of equal spacing $1/M$. Define this set as $S_{1,M}^d$, as follows:

$$S_{1,M}^d = \big\{w : Mw \in \{-M, \ldots, M\}^d, \|w\|_1 \le 1\big\}.$$

That is, each coordinate $w_{k,j}$ can only be an integer multiple of the grid size $1/M$, and we force the ℓ_1 norm to be less than or equal to 1. We consider the prior under

which w_k is independent uniform on the discrete set $S^d_{1,M}$. This has probability mass function

$$p_0(w) = \prod_{k=1}^{K} \left(1\{w_k \in S^d_{1,M}\} \frac{1}{|S^d_{1,M}|} \right),$$

with respect to counting measure in $(S^d_{1,M})^K$.

In the author McDonald's PhD thesis [55], various mean-square statistical risk and arbitrary-sequence regret bounds are established for sampling from the posterior distribution with the discrete uniform prior. In particular, with $M, K = O(N/\log(d))^{1/4})$ and $\beta = (\log(d)/N)^{1/4}$, statistical risk control of the order $O(\log(d)/N)^{1/4}$ is shown with this β, and faster rates are established with Gaussian distributed y_i and constant β equal to the reciprocal of the variance. Somewhat larger M and K are allowed, but the risk analysis does not succeed unless MK is of smaller order than N. Results of this type are also summarized in [7, 11].

In this work, we show for the continuous uniform prior there is a log-concave coupling and thus MCMC can be accomplished in polynomial time when $Kd > C(\beta N)^2$. With sufficiently large M of larger order than N, these mixing time results can be shown for the discrete prior as well. However, this analysis requires M that it is incompatible with the statistical control results. We were not able to combine statistical risk control and polynomial-time sampling simultaneously. In the present work, we only present the polynomial-time sampling results for the continuous prior.

3.2. Log-concave coupling

The log-concave coupling result is as follows.

Theorem 3.1. *Let the neural network have inner weight dimension $d \geq 2$ and $K \geq 2$ neurons with N data observations $(x_i, y_i)_{i=1}^N$. Assume $\beta N \geq 2$. Define the values*

$$\begin{aligned} C_N &= \max_{i \in \{1,N\}} |y_i| + a_0 V, \\ A_1 &= 2a_1 + 8a_2, \\ A_2 &= 2\sqrt{a_2}, \\ A_3 &= 8\sqrt{e} a_2 (C_N V)^{3/2} \big[A_1 + A_2 (C_N V)^{1/2} \big]. \end{aligned}$$

Set a positive $\delta \leq 0.05$, and let d and K satisfy

$$K\big[\log(2Kd) + \log(1/\delta)\big] \leq \beta N$$

and

$$Kd \geq A_3 (\beta N)^2. \tag{3.2}$$

Using a continuous uniform prior on $(S_1^d)^K$ *the posterior distribution* $p(w)$ *can be written as a mixture distribution with an auxiliary random variable* ξ*,*

$$p(w) = \int p(w|\xi)p(\xi)\,d\xi,$$

where $p(w|\xi)$ *is a log-concave density for each* ξ *and* $p(\xi)$ *is a log-concave density. If equation* (3.2) *is a strict inequality,* $p(\xi)$ *is strictly log-concave.*

4. Posterior densities and log-concave coupling

4.1. Posterior density

Consider the log of the posterior density $p(w)$ defined in equation (3.1), with the continuous uniform prior on $(S_1^d)^K$. The log and score of the posterior within the constrained set are equal to the following (with some constant B that is just the log of the normalizing constant):

$$\log p(w) = -\beta\ell(w) + B,$$
$$\nabla_{w_k} \log p(w) = \beta \sum_{i=1}^{N} \mathrm{res}_i(w)\big(c_k\psi'(w_k \cdot x_i)x_i\big).$$

Denote the Hessian as $H(w) \equiv \nabla^2 \log p(w)$. The density $p(w)$ is log-concave if $H(w)$ is negative semi-definite for all choices of w. For any vector $u \in \mathbb{R}^{Kd}$ with blocks $u_k \in \mathbb{R}^d$, the quadratic form $u^{\mathrm{T}}H(w)u$ can be expressed as

$$-\beta \sum_{i=1}^{N} \left(\sum_{k=1}^{K} c_k\psi'(w_k \cdot x_i)u_k \cdot x_i \right)^2 \tag{4.1}$$
$$+\beta \sum_{i=1}^{N} \mathrm{res}_i(w) \sum_{k=1}^{K} c_k\psi''(w_k \cdot x_i)(u_k \cdot x_i)^2. \tag{4.2}$$

It is clear that for any vector u the first line (4.1) is a negative term, but term (4.2) may be positive. The scalar values $c_k\psi''(w_k \cdot x_i)$ could be either a positive or negative value for each k and i, while the residuals $\mathrm{res}_i(w)$ can also be positive or negative signed. Thus, the Hessian is not a negative definite matrix in general and $p(w)$ may not be a log-concave density.

The term (4.2) is capturing how the non-linearity of ψ, which provides the benefit of neural networks over linear regression, is complicating matters. If ψ were linear, $\psi''(z) = 0$ for all z and we would have a simple linear regression problem. However, since ψ has second derivative contributions, this term must be addressed.

For each data index $i \in \{1, \ldots, N\}$ and each neuron index $k \in \{1, \ldots, K\}$, we introduce a coupling with an auxiliary random variable $\xi_{i,k}$. The goal of this auxiliary random variable is to force the corresponding individual i, k terms in (4.2) to be negative. Define the values

$$C_N = \max_{i \le N} |y_i| + a_0 V,$$
$$\rho_{0,N,K} = a_2 \frac{\beta C_N V}{K}.$$

We denote this $\rho_{0,N,K}$ as ρ_0 when it is clear N and K are set. The role of ρ_0 is that of a lower bound on the values of an inverse variance parameter ρ for the conditional distribution of auxiliary random variables below.

Ultimately, we will use bounded auxiliary random variables to yield the desired log-concave coupling. But to motivate the construction, first consider tentatively a simpler unbounded construction.

Conditioning on a weight vector w, define the forward coupling as conditionally independent random variables $\xi_{i,k}$, which are normal with mean $w_k \cdot x_i$ and variance $1/\rho$, as follows:

$$\xi_{i,k} \sim \text{Normal}\Big(w_k \cdot x_i, \frac{1}{\rho}\Big).$$

This then defines a forward conditional density (or coupling)

$$p(\xi|w) \propto e^{-(\rho/2)\sum_{i,k}(\xi_{i,k} - x_i \cdot w_k)^2},$$

and a joint density for w, ξ,

$$p(w, \xi) = p(w)p(\xi|w).$$

Via Bayes' rule, this joint density also has an expression using the induced marginal on the auxiliary ξ random vector and the reverse conditional density on $w|\xi$,

$$p(w, \xi) = p(\xi)p(w|\xi).$$

As we will show, this choice of forward coupling provides a negative definite correction to the Hessian of the log of $p(w|\xi)$ compared to what we had with $p(w)$, resulting in a log-concave reverse conditional density.

4.2. Reverse conditional density $p(w|\xi)$

First, we allow for $\xi_{i,k}$ to take arbitrary real values arising from the conditional normal distribution.

Theorem 4.1. *With the continuous uniform prior and* $\xi_{i,k} \sim \text{Normal}(x_i \cdot w_k, 1/\rho)$ *with* $\rho \geq \rho_0$*, the reverse conditional density* $p(w|\xi)$ *is log-concave for the given* ξ *coupling.*

Proof. The log of the reverse conditional density $p(w|\xi)$ is given by

$$\log p(w|\xi) = -\beta \ell(w) + B(\xi) - \sum_{i=1}^{N} \sum_{k=1}^{K} \frac{\rho}{2} (\xi_{i,k} - w_k \cdot x_i)^2, \tag{4.3}$$

for some function $B(\xi)$, which does not depend on w and is only required to make the density integrate to 1. The term (4.3) is a negative quadratic in w, which treats each w_k as an independent normal random variable. Thus, the additional Hessian contribution will be a $(Kd) \times (Kd)$ negative definite block diagonal matrix with $d \times d$ blocks of the form $\rho \sum_{i=1}^{N} x_i x_i^{\mathrm{T}}$. Denote the Hessian as

$$H(w|\xi) = \nabla^2 \log p(w|\xi).$$

For any vector $u \in \mathbb{R}^{Kd}$, with blocks $u_k \in \mathbb{R}^d$, the quadratic form $u^{\mathrm{T}} H(w|\xi) u$ can be expressed as

$$-\beta \sum_{i=1}^{N} \Big(\sum_{k=1}^{K} c_k \psi'(w_k \cdot x_i) u_k \cdot x_i \Big)^2 + \sum_{k=1}^{K} \sum_{i=1}^{N} (u_k \cdot x_i)^2 \big[\beta \operatorname{res}_i(w) c_k \psi''(w_k \cdot x_i) - \rho \big]. \tag{4.4}$$

By the assumptions on the second derivative of ψ and the definition of ρ, we have

$$\max_{i,k} \big(\beta \operatorname{res}_i(w) c_k \psi''(w_k \cdot x_i) - \rho \big) \leq 0.$$

So all the terms in the sum in (4.4) are negative. Thus, the Hessian of the log-likelihood of $p(w|\xi)$ is negative semi-definite and $p(w|\xi)$ is a log-concave density. ∎

While this proof offers a simple way to make a conditional density $p(w|\xi)$ which is log-concave, we also wish to study if there is log-concavity of the induced marginal of $p(\xi)$. The joint log-likelihood for $p(w, \xi)$ contains a bilinear term in ξ, w from expanding the quadratic

$$\sum_{k=1}^{K} \sum_{j=1}^{d} w_{k,j} \sum_{i=1}^{N} \xi_{i,k} x_{i,j}. \tag{4.5}$$

We want some control on how large this term can become, so we restrict the allowed support of ξ. When we make this support restriction we will need the parameter ρ to be greater than ρ_0. We define a convenience choice of

$$\rho = 2a_2 \frac{\beta C_N V}{K},$$

which we recognize as $\rho = 2\rho_0$.

For $0 < \alpha \leq 1/2$, let $z_\alpha \geq 0$ denote the upper α quantile of the standard normal distribution function. This z_α is the value such that $1 - \Phi(z_\alpha) = \alpha$, with the familiar inequality $z_\alpha \leq \sqrt{2\log(1/\alpha)}$.

For a positive $\delta < 1$, we define a constrained set

$$B = \left\{ \xi : \max_{j,k} \left| \sum_{i=1}^{N} x_{i,j} \xi_{i,k} \right| \leq N + z_{\delta/2Kd} \sqrt{\frac{N}{\rho}} \right\}, \tag{4.6}$$

where we note that $z_{\delta/2Kd} \leq \sqrt{2\log(2Kd/\delta)}$. We then define our forward conditional distribution for $p^*(\xi|w) = p(\xi|w, B)$ as the normal distribution restricted to the set B,

$$\begin{aligned} p^*(\xi|w) = p(\xi|w, B) &= \frac{1_B(\xi) p(\xi|w)}{P(\xi \in B|w)} \\ &= 1_B(\xi) \frac{\prod_{i=1}^{N} \prod_{k=1}^{K} (\rho/(2\pi))^{1/2} e^{-(\rho/2)(\xi_{i,k} - x_i \cdot w_k)^2}}{\int_B \prod_{i=1}^{N} \prod_{k=1}^{K} (\rho/(2\pi))^{1/2} e^{-(\rho/2)(\xi_{i,k} - x_i \cdot w_k)^2} \, d\xi}. \end{aligned}$$

In accordance with this constrained density, the term (4.5) will be bounded for any choice of $\xi \in B$ and $w_k \in S_1^d$, which will be a useful property in later proofs.

The denominator of this fraction is the normalizing constant of the density as a result of the restriction to the set B. Denote the log normalizing constant as $Z(w) = \log[P(\xi \in B|w)]$, such that

$$Z(w) = \log \int_B \prod_{i=1}^{N} \prod_{k=1}^{K} \left(\frac{\rho}{2\pi} \right)^{1/2} e^{-(\rho/2)(\xi_{i,k} - x_i \cdot w_k)^2} \, d\xi. \tag{4.7}$$

An equivalent expression for the forward coupling is then

$$p^*(\xi|w) = 1_B(\xi) \left(\frac{\rho}{2\pi} \right)^{NK/2} e^{-(\rho/2) \sum_{i,k} (\xi_{i,k} - x_i \cdot w_k)^2 - Z(w)}.$$

This construction also yields, for $\xi \in B$, the induced marginal density $p^*(\xi)$ with respect to Lebesgue measure,

$$\begin{aligned} p^*(\xi) &= \int p(w) p^*(\xi|w) \, \eta(dw) \\ &= 1_B(\xi) \int p(w) e^{-(\rho/2) \sum_{i,k} (\xi_{i,k} - x_i \cdot w_k)^2 - Z(w)} \, \eta(dw), \end{aligned}$$

and the reverse conditional density $p^*(w|\xi)$ with respect to reference measure η,

$$p^*(w|\xi) = \frac{p(w)p^*(\xi|w)}{p^*(\xi)} = \frac{p(w)e^{-(\rho/2)\sum_{i,k}(\xi_{i,k}-x_i\cdot w_k)^2 - Z(w)}}{\int p(w)e^{-(\rho/2)\sum_{i,k}(\xi_{i,k}-x_i\cdot w_k)^2 - Z(w)}\,\eta(dw)}.$$

Note these densities differ from the $p(w|\xi)$ and $p(\xi)$ defined before without restricting to the set B due to the presence of the $Z(w)$ function.

We can then show two results. The first is that $Z(w)$ has bounded range, and does not modify the density that much for small δ. Via a Holley–Stroock perturbation argument (Section A.4), we show that $p^*(w|\xi)$ has a reasonably controlled log-Sobolev constant, less than $10K^2/(1-\delta)$, even with the original ρ. The second result is that the Hessian of $Z(w)$ is small, so that, with the somewhat larger ρ of the present section, we obtain the stronger conclusion that $p^*(w|\xi)$ is log-concave for each ξ in B.

This step allows us to construct a *log-concave coupling*, which is to say

$$p(w) = \int p^*(w|\xi)p^*(\xi)\,d\xi,$$

where both the conditional $p^*(w|\xi)$ and marginal $p^*(\xi)$ are log-concave. This conclusion provides a simplified story. Moreover, some results for rapid mixing of MCMC such as Ball Walk and Hit and Run for densities on a constrained set have been stated naturally for densities required to be log-concave, not just log-Sobolev [53].

The restriction of ξ to the set B is the restriction to a very likely set under the unconstrained coupling, in particular we have the following lemma.

Lemma 4.2. *For any weight vector w with $\|w_k\|_1 \le 1$, the set B in (4.6) has probability using $p(\xi|w)$ satisfying*

$$P(\xi \in B|w) \ge 1-\delta.$$

Proof. See Section A.1. ∎

Furthermore, the function $Z(w)$ is nearly constant, having small first and second derivative. Therefore, the function has little impact on the log-likelihood.

Lemma 4.3. *For any specified vector $u \in \mathbb{R}^{Kd}$, define the value*

$$\tilde{\sigma}^2 = \frac{\sum_{i=1}^N \sum_{k=1}^K (u_k \cdot x_i)^2}{\rho}.$$

For positive values $\delta \le 0.158$, we then have upper bounds

$$|u \cdot \nabla Z(w)| \le \rho\tilde{\sigma}\frac{\delta}{1-\delta}\left(\sqrt{2\log(1/\delta)}+1\right)$$

and

$$\left|u^{\mathrm{T}}\left(\nabla^2 Z(w)\right)u\right| \leq \rho^2\widetilde{\sigma}^2\frac{\delta}{1-\delta}\left[2\log(2/\delta)+1+\frac{\delta}{1-\delta}\left(2\log(1/\delta)+3\right)\right].$$

Note both bounds go to 0 *as* $\delta \to 0$*, and thus can be made arbitrarily small by choice of* δ*.*

Proof. See Section A.1. ∎

The restriction to $\delta \leq 0.158$ implies $\delta \leq 1-\Phi(1)$ so that the upper quantile z_δ is at least 1. More generally, the $+1$ in the gradient bound may be replaced by $1/z_\delta$ and the $+1$ and $+3$ in the Hessian bound may be replaced by $1/z_{\delta/2}$ and $2+1/z_{\delta/2}^2$, respectively. Further refinements are specified in the proof in the appendix.

Thus, with restriction to the set B, whose size is determined by δ, and a slightly larger ρ, we can give a result that is similar to Theorem 4.1. Note this result is for $p^*(w|\xi)$, which is distinct from $p(w|\xi)$ due to the presence of the $Z(w)$ function in the log-likelihood and the restriction to $\xi \in B$.

Theorem 4.4. *Define the notation*

$$H_1(\delta) = \frac{\delta}{1-\delta}\left(2\log(2/\delta)+1\right),$$
$$H_2(\delta) = \frac{\delta^2}{(1-\delta)^2}\left(2\log(1/\delta)+3\right),$$

and let $H(\delta)$ *be defined as* $H_1(\delta)+H_2(\delta)$*. For* $\rho = 2\rho_0$*, choose a positive* $\delta \leq 0.054$ *satisfying* $H(\delta) \leq 1/2$*. Then, for the continuous uniform prior, with* ξ *restricted to the set* B *defined by* δ*, the reverse conditional density* $p^*(w|\xi)$ *is a log-concave density in* w *for each* ξ *in* B*.*

Proof. See Section A.2. ∎

The condition $H(\delta) \leq 1/2$ is convenient though not essential. It is tailored to the choice $\rho = 2\rho_0$. The proof of Theorem 4.4 will reveal refined choices of ρ, depending on δ, that allows for $H(\delta) < 1$, permitting $0 < \delta < 0.115$, as well as permitting ρ arbitrarily close to ρ_0 when δ is sufficiently small.

By itself, the pairing of a target density $p(w)$ with a suitable normal forward coupling $p^*(\xi|w)$ to produce a reverse conditional $p^*(w|\xi)$ which is log-concave is not new. As later discussed, the same concept is used in proximal sampling methods and diffusion models. In this work, we go further for our model, obtaining that the induced marginal $p^*(\xi)$ is itself log-concave, which we call a log-concave coupling.

4.3. Marginal density $p^*(\xi)$

Lemma 4.5. *The score and Hessian of the induced marginal density for $p^*(\xi)$, where $\xi \in B$, are expressed as*

$$\begin{aligned} \partial_{\xi_{i,k}} \log p^*(\xi) &= -\rho\, \xi_{i,k} + \rho\, x_i \cdot E_{p^*}[w_k|\xi], \\ \partial_{\xi_{i_1,k_1},\xi_{i_2,k_2}} \log p^*(\xi) &= -\rho 1\{(i_1,k_1) = (i_2,k_2)\} \\ &\quad + \rho^2 \operatorname{Cov}_{p^*}[w_{k_1} \cdot x_{i_1}, w_{k_2} \cdot x_{i_2}|\xi]. \end{aligned} \tag{4.8}$$

Equivalently, in vector form using the N by d data matrix $\mathbf{X}$, we have

$$\nabla \log p^*(\xi) = \rho\Big(-\xi + E_{p^*}\begin{bmatrix} \mathbf{X}w_1 \\ \cdots \\ \mathbf{X}w_K \end{bmatrix}|\xi\Big), \tag{4.9}$$

$$\nabla^2 \log p^*(\xi) = \rho\Big(-I + \rho\, \operatorname{Cov}_{p^*}\begin{bmatrix} \mathbf{X}w_1 \\ \cdots \\ \mathbf{X}w_K \end{bmatrix}|\xi\Big). \tag{4.10}$$

Proof. The stated results are a consequence of simple calculus, but we will derive them using a statistical interpretation that avoids tedious calculations.

Consider ξ in B. The log of the induced marginal for $p^*(\xi)$ is equal to the log of the joint density with w integrated out,

$$\log p^*(\xi) = \log\Big(\int p(w) p^*(\xi|w)\, \eta(dw)\Big).$$

Rearranging the Gaussian forward conditional, this can be expressed as a quadratic term in ξ and a term which represents a cumulant-generating function plus a constant. Recall $Z(w)$ as defined in equation (4.7). Denote the function

$$h(w) = -\beta \ell(w) - \frac{\rho}{2} \sum_{i=1}^{N} \sum_{k=1}^{K} (w_k \cdot x_i)^2 - Z(w),$$

which is the part of the log of the joint density which does not depend on ξ. The marginal density can then be expressed for ξ in B as

$$\begin{aligned} \log p^*(\xi) = &-\frac{\rho}{2}\|\xi\|_2^2 \\ &+ \log\Big(\int p_0(w) e^{h(w)} e^{\rho \sum_{i=1}^{N} \sum_{k=1}^{K} \xi_{i,k} w_k \cdot x_i}\, \eta(dw)\Big) + C \end{aligned} \tag{4.11}$$

for some constant C which makes the density integrate to 1.

It is clear that, with suitable constant adjustment, the term (4.11) is the cumulant-generating function of (ρ times) the random variable $u(w)$ defined by

$$u(w) = \xi \cdot \begin{pmatrix} \mathbf{X}w_1 \\ \cdots \\ \mathbf{X}w_K \end{pmatrix},$$

when w is distributed according to the density proportional to $p_0(w)e^{h(w)}$. Thus, the gradient in ξ is the mean of the vector and the second derivative is the covariance, as are standard properties of derivatives of cumulant-generating functions, wherein the mean and covariance are taken with respect to the probability density tilted by $e^{\rho u(w)}$, which is recognized to be the reverse conditional $p^*(w|\xi)$. ∎

We highlight two important consequences of this result.

Corollary 4.6. *The score* $\nabla \log p^*(\xi)$ *is a linear transformation of the vectors of expected values* $E_{p^*}[w_k|\xi]$ *from the log-concave reverse conditional* $p^*(w|\xi)$.

Proof. This is what is expressed in (4.8) or (4.9). ∎

Remark 4.7. To obtain the score of the marginal density, the $E_{p^*}[w_k|\xi]$ can be estimated using an MCMC method and thus it is readily available for use. Note that to run an MCMC algorithm such as Metropolis adjusted Langevin diffusion to obtain samples from $p^*(\xi)$, its score is needed at each step to update ξ. The conditional means needed for this score at ξ can be computed via an "inner loop" consisting of an MCMC algorithm to obtain samples from $p^*(w|\xi)$ to average. The log-concavity of this reverse conditional provides for the rapid mixing in this inner loop.

Corollary 4.8. *The density* $p^*(\xi)$ *is log-concave if for any unit vector* $u \in \mathbb{R}^{NK}$ *with blocks* $u_k \in \mathbb{R}^N$, *the variance of a particular linear combination of w, namely*

$$v(w) = \sum_{k=1}^{K} u_k^{\mathrm{T}} \mathbf{X} w_k,$$

with respect to the reverse conditional $p^*(w|\xi)$ *is less than* $1/\rho$*:*

$$\mathrm{Var}_{p^*}\big[v(w)|\xi\big] \leq \frac{1}{\rho} \tag{4.12}$$

for ξ *in the convex support set* B.

Proof. This is a simple consequence of (4.10). ∎

Therefore, to show that $p^*(\xi)$ is log-concave, we must provide an upper bound on the covariance of w using the reverse conditional density $p^*(w|\xi)$. Note that such conditional expectation and conditional covariance representations would also hold using $p(\xi)$, which is defined without conditioning on the set B, and thus does not include the $Z(w)$ term in the joint log density. However, the restrictions imposed on maximum inner products by the definition of B will prove useful in upper bounding the reverse conditional covariance.

4.4. Conditional covariance control

The log of $p^*(w|\xi)$ is the log of the prior density plus an additional concave term. Under a log-concave prior, one would expect that adding a concave term to the exponent of an already log-concave density should result in less variance in every direction. Thus one can conjecture the prior covariance would be more than the conditional covariance for any conditioning value,

$$\mathrm{Cov}_{P_0}[w] \succ \mathrm{Cov}_{p^*}[w|\xi] \quad \text{for all } \xi \in B. \tag{4.13}$$

Under a Gaussian prior, such a statement would follow easily from the Brascamp–Lieb inequality [15, Proposition 2.1]. However, for the uniform prior on a convex set, this method does not directly apply.

The covariance matrix of the uniform prior on $(S_1^d)^K$ is diagonal (note the different coordinates are uncorrelated but not independent due to symmetry) with entries

$$\mathrm{Var}_{P_0}(w_{k,j}) = \frac{d}{(d+1)^2(d+2)} \le \frac{1}{d^2},$$

which follows from properties of the Dirichlet distribution. Thus, under the conjecture (4.13), we would expect a bound of the form

$$\begin{aligned}
\rho \mathrm{Var}_{p^*}\big[v(w)|\xi\big] &\le 2a_2 \frac{\beta C_N V}{Kd^2} \sum_{j=1}^{d} \sum_{k=1}^{K} \Big(\sum_{i=1}^{N} u_{i,k} x_{i,j} \Big)^2 \\
&\le 2a_2 \frac{\beta C_N V N}{Kd} \sum_{k=1}^{K} \|u_k\|_1^2 \\
&\le 2a_2 \frac{\beta C_N V N}{Kd} \sum_{k=1}^{K} \|u_k\|_2^2 \\
&= 2a_2 \frac{C_N V \beta N}{Kd}.
\end{aligned}$$

Thus for $Kd > C(\beta N)$ for some value C, we would have log-concavity of the marginal. However, we are unable to prove this conjecture is true. Instead, using a different approach, we will conclude for a specified C that

$$Kd \ge C(\beta N)^2$$

results in log-concavity of the marginal density.

Instead of recreating an inequality like (4.13), we must take a different approach to upper bound the variance in any direction. Denote the function

$$h_\xi(w) = -\beta \ell(w) - \sum_{i=1}^{N} \sum_{k=1}^{K} \frac{\rho}{2} (\xi_{i,k} - w_k \cdot x_i)^2 - Z(w).$$

Denote the function shifted by its prior mean as

$$\widetilde{h}_\xi(w) = h_\xi(w) - E_{P_0}\big[h_\xi(w)\big].$$

Define its cumulant-generating function with respect to the prior as

$$\Gamma_\xi(\tau) = \log E_{P_0}\big[e^{\tau \widetilde{h}_\xi(w)}\big]. \tag{4.14}$$

When bounding the variance using the reverse conditional, the expectation can be thought of as taken with respect to the density proportional to the prior times an $e^{\widetilde{h}_\xi(w)}$ factor. However, because the $h_\xi(w)$ is the sum of N varying terms, it is too large to pull out its supremum even for ξ in B and w in $(S_1^d)^K$. However, with a Hölder's inequality, we can pull out an expectation of $e^{q\widetilde{h}_\xi(w)}$, which is the expectation of the q power of the factor of interest. Then the variance is controlled when that power $q = \ell/(\ell-1)$ is taken to be close enough to 1 and when higher order prior moments of w are controlled. This is captured in the following lemma and in the subsequent analysis of the two factors that arise from this bound.

Lemma 4.9. *For any integer $\ell \geq 1$ and for any vector $u \in \mathbb{R}^{Kd}$, we have the upper bound*

$$\mathrm{Var}_{p^*}(u \cdot w|\xi) \leq \big(E_{P_0}\big[(u \cdot w)^{2\ell}\big]\big)^{1/\ell} e^{((\ell-1)/\ell)\Gamma_\xi(\ell/(\ell-1))-\Gamma_\xi(1)}.$$

Proof. The variance of the inner product $u \cdot w$ is less than its expected square. The reverse conditional density $p^*(w|\xi)$ can be expressed as

$$p^*(w|\xi) = e^{\widetilde{h}_\xi(w)-\Gamma_\xi(1)} p_0(w).$$

We then apply a Hölder's inequality to the integral expression with parameters p and q such that $1/p + 1/q = 1$, to obtain

$$\begin{aligned}\mathrm{Var}_{p^*}(u \cdot w|\xi) &\leq E_{P_0}\big[(u \cdot w)^2 e^{\widetilde{h}_\xi(w)-\Gamma_\xi(1)}\big] \\ &\leq \big(E_{P_0}\big[(u \cdot w)^{2p}\big]\big)^{1/p} \big(E_{P_0}\big[e^{q\widetilde{h}_\xi(w)-q\Gamma_\xi(1)}\big]\big)^{1/q}.\end{aligned}$$

Let $p = \ell$ and $q = \ell/(\ell-1)$. The second factor can be written as

$$e^{((\ell-1)/\ell)\Gamma_\xi(\ell/(\ell-1))-\Gamma_\xi(1)}.$$

■

We then study the moments of the prior density and the behavior of the $\Gamma_\xi(\tau)$ function separately. As we shall see, when the moments are well controlled, the bound is effective, in part, because it only involves (differences) of cumulants at nearby values.

Lemma 4.10. *For any unit vector $u \in \mathbb{R}^{NK}$ with blocks $u_k \in \mathbb{R}^N$, and positive integer ℓ, we have*

$$E_{P_0}\left[\left(\sum_{k=1}^{K} u_k^{\mathrm{T}} \mathbf{X} w_k\right)^{2\ell}\right]^{1/\ell} \le \frac{4\ell N}{\sqrt{e}\, d}.$$

Proof. See Section A.3. ∎

Lemma 4.11. *Assume $\rho = 2\rho_0$. Denote the constants*

$$A_1 = 2a_1 + 8a_2, \tag{4.15}$$

$$A_2 = 2\sqrt{2a_2}. \tag{4.16}$$

Assume positive $\delta \le 0.116$ and $d \ge 2$, $K \ge 2$. For any positive integer $\ell \ge 1$ and any ξ from the constrained set B, we have

$$\frac{\ell - 1}{\ell}\Gamma_\xi\left(\frac{\ell}{\ell - 1}\right) - \Gamma_\xi(1) \le A_1 \frac{C_N V \beta N}{\ell} + A_2 \frac{\sqrt{C_N V \beta N}}{\ell}\sqrt{2\log(2Kd/\delta)}\left(\sqrt{1.4K}\right).$$

Proof. See Section A.3. ∎

We summarize the implications of the conclusions of Lemmas 4.9–4.11 as follows. Ignoring certain constant factors, we have the following upper bound on the variance in (4.12) for any choice of ℓ:

$$\frac{N\ell}{d}\exp\left(\frac{\beta N + \sqrt{\beta N K \log(2Kd/\delta)}}{\ell}\right),$$

which takes the form $(N\ell/d)e^{V_{N,K}/\ell}$, with $V_{N,K} = \beta N + \sqrt{\beta N K \log(2Kd/\delta)}$. Ignoring for now the integer constraint, the optimal continuous choice of ℓ to minimize the expression is seen to be $\ell^* = V_{N,K}$, with which the exponent becomes equal to 1. With this choice, we would have the bound $N V_{N,K} e/d$ on the variance equal to

$$\frac{\beta N^2 e + N^{3/2} e\sqrt{\beta K \log(2Kd/\delta)}}{d}.$$

Multiplying this by $\rho \propto \beta/K$, we would have a bound of order

$$\frac{(\beta N)^2}{Kd}\left(1 + \left[\frac{K\log(2Kd/\delta)}{\beta N}\right]^{1/2}\right).$$

If $K\log(2Kd/\delta) \le \beta N$, then we have an $O((\beta N)^2/Kd)$ bound. With a choice of d and K large enough, we can make this expression be less than 1. We make this statement more precise in the following theorem.

Theorem 4.12. *Assume $\delta \leq 0.116$ and $d \geq 2$, $K \geq 2$. Further assume that*

$$K \log(2Kd/\delta) \leq \beta N,$$

which is essentially a condition that K not be too large, that is, K is less than some multiple of $\beta N/\log(\beta N)$.

With $\rho = 2\rho_0$, define A_1, A_2 as in (4.15) and (4.16), and define

$$A_3 = 8\sqrt{e}a_2 C_N V \left[A_1 C_N V + 2A_2 (1.4 C_N V)^{1/2} + \frac{1}{\beta N} \right].$$

Let d and K satisfy

$$Kd \geq A_3 (\beta N)^2. \tag{4.17}$$

Then $\rho \operatorname{Var}_p^*[v(w)|\xi] \leq 1$ and the marginal density for $p^*(\xi)$ is log-concave with convex support B, using the continuous uniform prior. If equation (4.17) is a strict inequality, then the density is strictly log-concave.

Log concavity of $p^*(\xi)$ with $\rho = 2\rho_0$ is here holding for a larger range of values of δ than the $\delta \leq 0.054$ we require for the log concavity of $p^*(w|\xi)$. So, with this ρ, to get the full log-concave coupling property (with both $p^*(\xi)$ and $p^*(w|\xi)$ log-concave) choose δ not more than 0.054.

The combined conditions of Theorem 4.4 and Theorem 4.12 are collected in the log-concave coupling Theorem 3.1 presented earlier in the document.

Proof. By Corollary 4.8, the Hessian of $\log p^*(\xi)$ is log-concave when, for any unit vector u, we have

$$\rho \operatorname{Var}_{p^*} \left[\sum_{k=1}^{K} u_k^{\mathrm{T}} \mathbf{X} w_k | \xi \right] \leq 1. \tag{4.18}$$

By Lemmas 4.9–4.11, we have an upper bound for this variance for any scalar $\ell > 1$ and $\xi \in B$. Recall A_1, A_2 as defined in expressions (4.15) and (4.16). Set ℓ to be the least integer greater than or equal to

$$\ell^* = A_1 C_N V \beta N + 2A_2 \sqrt{C_N V \beta N \log(2Kd/\delta)} \left(\sqrt{1.4K} \right).$$

This choice makes the cumulant-generating difference bound from Lemma 4.11 not more than 1, simplifying the variance bound from Lemma 4.9. Using Lemma 4.10, it gives an upper bound on the product of ρ and variance that appears in (4.18),

$$\begin{aligned} &2a_2 \frac{\beta C_N V}{K} \frac{4N}{\sqrt{e}d} (\ell^* + 1)e \\ &\quad = 8\sqrt{e} A_1 a_2 \frac{(C_N V \beta N)^2}{Kd} \\ &\qquad + 8\sqrt{e} A_2 a_2 \frac{2(C_N V \beta N)^{3/2} \sqrt{1.4K}}{Kd} \sqrt{\log(2Kd/\delta)} + 8\sqrt{e} a_2 \frac{C_N V \beta N}{Kd} \end{aligned}$$

$$\leq 8\sqrt{e}a_2\frac{(\beta N)^2}{Kd}$$
$$\times C_N V\left[A_1 C_N V) + 2A_2(1.4C_N V)^{1/2}\left(\frac{K\log(2Kd/\delta)}{\beta N}\right)^{1/2} + \frac{1}{\beta N}\right].$$

By assumption,

$$\frac{K\log(2Kd/\delta)}{\beta N} \leq 1,$$

so we have upper bound on (4.18),

$$8\sqrt{e}a_2 C_N V\left[A_1 C_N V + 2A_2(1.4C_N V)^{1/2} + \frac{1}{\beta N}\right]\frac{(\beta N)^2}{Kd}.$$

If Kd satisfies condition (4.17), then $\rho\mathrm{Var}_{p^*}[v(w)|\xi]$ is less than 1 for ξ in B. By Corollary 4.8, this implies log-concavity of the marginal density $p^*(\xi)$. ∎

Remark 4.13. The log-concavity is obtained in the setting of high interior weight dimension d. If the dimension d is not larger than $A_3(\beta N)^2/K$ then the dimension d can be increased to provide satisfaction of this condition. We mention two ways to do this. First, the dimension d can be made artificially larger by repeating the input vectors. Say there are original input vectors $\widetilde{x}_i$ with default dimension $\widetilde{d}$. Define new input vectors by repeating the data L times

$$x_i = (\widetilde{x}_i, \ldots, \widetilde{x}_i) \in \mathbb{R}^{\widetilde{d}L}.$$

We can then consider $\mathbf{X}$ as our data matrix with row dimension $d = L\widetilde{d}$.

The span of the new data matrix with ℓ_1 controlled weights, $\{z = \mathbf{X}w : \|w\|_1 \leq 1\}$, is the same as for the original matrix. So we have the same approximation ability of the network. This can also equivalently be considered as inducing a different log-concave prior on the original weight vectors of dimension $\widetilde{d}$. Nevertheless, it remains convenient to consider the uniform prior in the higher $d = L\widetilde{d}$-dimensional space. It introduces even more multi-modality in the posterior density $p(w)$ as multiple longer weight vectors yield the same output in the neural network. Yet, with sufficient replication size L, by our proceeding theorems the density has a log-concave decomposition.

As indicated, variable replication does not change what is represented by the network model. A second method to increase the dimensionality incorporates additional model flexibility. Akin to polynomial networks as described in [8,22,23], augment the original variables $\widetilde{x}_i$ with a sufficient number of polynomial or spline basis functions of these variables, each bounded by 1 in absolute value, to create the larger vectors x_i, which includes the transformed variables. Computationally excessive dimensionality is avoided by restricting such terms to be additive or first-order interactions. Retain

the activation function with the smoothness conditions we have specified. The network model now represents a larger class of functions of the original variables, yet it retains the desired form as a weighted sum of activation functions applied to linear combinations of (transformed) variables. The log-concave coupling representation holds for any original dimension $\widetilde{d}$, with sufficiently many of these transformed variables.

Remark 4.14. This argument for a log-concave coupling works for any log-concave prior on the ℓ_1 ball whose moments are not more than for the uniform prior. Also, it can be made to work for various likelihoods $e^{-\beta\ell(w)}$ with non-concave $\ell(w)$, including various non-linear regressions, not just those that arise in fitting neural networks.

In our approach, the auxiliary variables are centered at the $w_k \cdot x_i$. Accordingly, this methodology applies with $K = 1$ to models in which the dependence of the log-likelihood $\ell(w)$ on w is expressed via a smooth function $g(w_1 \cdot x)$ of one such linear combination, sometimes called a single-index model, as in generalized linear models [54, 60], wherein the link function may produce a non-concave log-likelihood. Then these generalized linear models can be provided with a log-concave coupling. With additive transformations of original variables included in the x, these represent generalized additive models [38], now with a log-concave coupled representation of the posterior distribution. With $K > 1$, certain multi-index non-linear regression models $g(w_1 \cdot x, \ldots, w_K \cdot x)$ can be addressed. Particular g such as linear combinations of ridge-spline bases create models that correspond to projection pursuit regression [32, 43]. Single-hidden-layer network models $\sum_k c_k \psi_k(w_k \cdot x)$ are of this latter form with a fixed $g(z_1, \ldots, z_K)$ being the additive function $\sum_k c_k \psi_k(z_k)$.

5. Discussion

The main technique used in this paper which facilitates sampling is the use of a measure decomposition, which is taking a target density $p(w)$ and expressing it in a convenient mixture representation with random variable ξ which may be referred to as a hidden, latent, or auxiliary random variable

$$p(w) = \int p(w|\xi)\, p(\xi)\, d\xi.$$

One line of development of this technique has its origins in the statistical physics literature and the study of Ising models or spin glass systems [2, 42]. A tool that emerges in this literature is the Hubbard transform. For any positive definite matrix A, this is simply a rearrangement of a normal density to state

$$e^{(1/2)w^T A w} = \int \frac{e^{-(1/2)\xi^T A \xi}}{(2\pi)^{d/2} \det(A^{-1})^{1/2}} e^{\xi^T A w}\, d\xi.$$

Equivalently, the moment generating function of a mean 0 normal random variable is quadratic. As such, whenever a positive definite term of the form $(1/2)w^T Aw$ appears in a log-likelihood, there is an immediate chance to replace it with the Hubbard transform and introduce and auxiliary random variable ξ.

This approach has been mostly used as an analysis technique for the mixing time of sampling algorithms for Ising models. Modern works of [12,30] showed significant improvements of the log-Sobolev constant for high temperature spin glass systems with this technique.

Moreover, it has been recognized that a measure decomposition also provides a joint distribution which can be used as a sampling algorithm. For instance, Montanari and Wu [58] use a Gibb's chain (alternating between conditional w and ξ draws in the present notation) to sample from a Bayesian regression problem with spike and slab prior. This problem has a natural quadratic structure, and thus can make use of the Hubbard transform.

Measure decomposition exists outside of the use of a normal random variable and the Hubbard technique, as studied in more generality in [29] and [28].

More broadly, there are a number of Markov Chain Monte Carlo algorithms that make use of an auxiliary random variable for sampling, without changing the target marginal measure. These include the Swendsen–Wang algorithm with a spin variable and an auxiliary bond variable [64]; the Hamiltonian Monte Carlo algorithm with a velocity variable and an auxiliary momentum random variable [13]; the Data Augmentation Gibbs Sampling algorithm arising in Hierarchical Bayes models with latent variables [65]; the Stochastic Localization algorithm where the Markov chain moves are motivated by a representation of the target random variable through an auxiliary martingale model [20,27,57]; and the Proximal Sampling algorithm with a proximal random variable serving as an auxiliary [19,41]. The Proximal Sampler perhaps bears the most resemblance to our method here. To sample a density $p(w) \propto e^{-V(w)}$, define a new random variable

$$p(\xi|w) \sim N\left(w, \frac{1}{\lambda} I\right)$$

for a small value of λ. Then sample joint density $p(w)p(\xi|w)$. The method is a Gibb's sampler which alternates between sampling the conditional density $p(\xi|w)$ which is a simple normal with small variance, and then the conditional density $p(w|\xi)$ via a rejection algorithm called the Restricted Gaussian Oracle (RGO) [49]. Any of these methods may be viewed as a sampling algorithm taking advantage of a specific measure decomposition, or specific joint density.

As we show in this work, the use of measure decomposition is not restricted to densities with a predefined latent structure (for Gibbs sampling) or densities with a quadratic term in them (providing a candidate for the Hubbard transform). Functionally, coupling with a normal forward auxiliary random variable introduces a negative

definite corrective term to the Hessian of the log-density, much the same as proximal sampling. Thus we are in essence constructing a measure decomposition "by force", defining the forward coupling we want and then deducing the reverse coupling and induced marginal density from the resulting joint density. The main difference with other techniques is the enforcing of the log-concavity of the reverse coupling and the induced marginal, allowing for provably rapid one pass sampling of each, rather than repeated alternating sampling of the conditionals.

Measure decomposition also has connections with the field of score based diffusion models [62]. Score based diffusions propose starting with a random variable w' from the target density $p(w')$, and then defining the forward SDE

$$dw_t = -w_t\, dt + \sqrt{2}\, dB_t.$$

At every time t, this induces a joint distribution on $p(w', w_t)$ under which the forward conditional distribution $p(w_t|w')$ is a Gaussian distribution with mean being a linear function of w'. Paired with this forward SDE is the definition of a reverse SDE that would transport samples from a standard normal distribution to the target distribution of interest. The drift of the reverse diffusion is defined by the scores of the marginal distribution of the forward process $\nabla \log p_t(w_t)$. If these scores can be computed, the target density can be sampled from.

As is the case in our mixture model, the scores of the marginal are defined by expectations with respect to the reverse conditional $p(w'|w_t)$. For some thresholds τ_1, τ_2, for small times $t \leq \tau_1$ the reverse conditionals $p(w'|w_t)$ are log-concave and easily sampled. For large times $t \geq \tau_2$, the marginal density $p_t(w_t)$ is approaching a standard normal distribution and thus will become log-concave. If $\tau_2 < \tau_1$, these two regions overlap and the original density $p(w')$ can be written as a log-concave mixture of log-concave components

$$p(w') = \int p(w'|w_t) p_t(w_t)\, dw_t.$$

Thus, the entire procedure of reverse diffusion can be avoided and a one shot sample of w_t from its marginal $p_t(w_t)$ and a sample from the reverse conditional $p(w'|w_t)$ can computed. A variation of this idea is the core procedure we use in this paper, simplifying the processes of a reverse diffusion into one specific and useful choice of joint measure with an auxiliary random variable.

We highlight that the score of the marginal density $\nabla \log p^*(\xi)$ is not given as an explicit formula, however it is defined as an expectation with respect to the log-concave reverse conditional for $p^*(w|\xi)$ noted in Corollary 4.6. Thus, the score of the marginal can be computed as needed via its own MCMC sub-routine.

Here we briefly review sampling literature for log-concave densities. Our density $p^*(w|\xi)$ is a weakly log-concave density constrained to a convex set, while $p^*(\xi)$

is a strongly log-concave density also restricted to a convex set. For $p^*(w|\xi)$, the log of the conditional density only depends on the weight vectors w through their interaction with the data matrix $\mathbf{X}w$. The vectors w are d-dimensional with $d > N$, thus for any direction orthogonal to the rows of the data matrix the log of the density is flat and has 0 Hessian, hence this density is weakly log-concave. Nonetheless, [53] shows Ball Walk and Hit and Run algorithms mix in polynomial time for weakly log-concave densities on a bounded convex set. Recent results in [48] improve upon these mixing time bounds. We note that with a strictly log-concave prior, strict log-concavity of $p^*(w|\xi)$ would be forced. We also note, with different construction of the auxiliary random variable ξ, it may be possible to force strict log-concavity in every direction of $p^*(w|\xi)$ using a normal with a different mean and covariance matrix for the forward coupling.

In terms of sampling the marginal $p^*(\xi)$, we have a strictly log-concave distribution restricted to the convex set defined by B. The score $\nabla \log p^*(\xi)$ is expressed as a linear transformation of $E^*[w|\xi]$ and thus can be computed as needed. If the support set was not restricted, we could use Metropolis Adjusted Langevin Diffusion (MALA) and achieve rapid mixing [26]. Instead, to deal with the boundary conditions we must use techniques such as a barrier function [63] or other adaptations of sampling algorithms to restricted support such as Dikin Walks [47] and Hamiltonian Monte Carlo in a constrained space [46].

While in this work we focus on a Bayesian approach and use MCMC for sampling, there have been a number of positive results for training neural networks by optimization in specific instances. For classification problems with well-separated classes and with rather large (potentially overfit) single-hidden-layer networks, [16] shows that gradient descent with large step size converges quickly to an interpolating solution on the training data (i.e. 0 training loss). The article [66] demonstrates this solution still has good generalization risk via a form of "benign overfitting", however this comes at a cost of being susceptible to adversarial perturbations in specific directions that flip model outputs [31].

Another approach to understanding optimization in very large neural networks is to compare them to certain infinite width limits via the Neural Tangent Kernel [44]. With small random initialization and control on the number of gradient steps, gradient methods quickly converge to a near interpolating solution [1,25,68]. Networks trained in this regime approach regression with a fixed kernel determined by the covariance of the gradient of the network at the random initialization. This is tantamount to a linear regression onto a prefixed set of basis functions (the large eigenvalue eigenvectors of the kernel).

A related perspective is in [21] which identifies this regime of an approximate linear model of small weights as "lazy" training, that does not allow the internal weights to have much freedom to adjust the basis. Models trained in this regime can have poor

generalization, compared to models trained in the more difficult non-lazy regime. In contrast, we seek a procedure that performs as well as if the set of directions w_k are adapted to the observed data.

For very wide networks $K > N$, [51] shows neural networks satisfy a Polyak–Łojasiewicz (PL) condition proving convergence of stochastic gradient descent to a global minimizer of the loss function. This is an interesting phenomenon, however without suitable parameter controls (such as ℓ_1 controls), it is not clear if generalization properties will be favorable in this setting for general function learning.

There are also several negative results [24, 33, 35] showing that training a single-hidden-layer network to interpolation (0 training loss) is an NP hard problem. For example, [67] shows that for a network of width K, interior weight dimension d, and using the step activation function, there does not exist a polynomial time algorithm to achieve average squared training error less than $\zeta(Kd)^{-3/2}$ for an absolute constant ζ. It does not rule out in the noise free setting the possibility of computationally feasible algorithms to achieve average squared error less than a constant times $1/K$.

In this paper, the authors have presented posteriors $p(w)$ that sample all K neuron weights $w_1, \ldots, w_K$ jointly. However, the problem can also be constructed as a Greedy Bayes procedure sampling one neuron weight at a time based on the residuals of previous fits. The authors discuss these results in [7, 11, 55, 56].

This paper also contains significant portions of Curtis McDonald's PhD thesis, which can be found at [55]. The thesis document contains more in depth analysis of the points discussed here, and a complete look at the greedy approach to sampling.

6. Conclusion and future work

In this work, we study a mixture form of the posterior density for single-hidden-layer neural nets. For a continuous uniform prior on the ℓ_1 ball, we show the posterior density can be expressed as a mixture with only log-concave components when the total number of parameters Kd larger than $C(\beta N)^2$ for a constant C where β is the inverse temperature and N is the number of data points.

There are a number of future directions for research. While we are able to give computational control for this problem, no statistical guarantees are derived for this choice of inverse temperature $\beta \leq C(\sqrt{Kd}/N)$, and a uniform prior. Other choices of prior, aside from uniform and couplings, different from the normal forward coupling we use here are currently under investigation. These methods may prove better at connecting risk control and a mixture measure form of the target density, but use many of the same ideas outlined in this work.

Further details of the sampling method must also be worked out. The choice of sampling algorithm, hyper-parameter choices such as step size and the number of

MCMC iterations, as well as technical details such as condition number have not been addressed in this work. The choice of ρ we make is in a sense the "smallest" ρ that forces $p(w|\xi)$ to be log-concave by canceling out any positive definite terms in the Hessian arising from non-linearity (that is, terms dependent on the second derivative of the activation function). Larger choices of ρ can result in stronger log-concavity for the reverse conditional distribution $p(w|\xi)$ that can have sampling benefits.

The Hölder inequality approach to upper bound the covariance $\mathrm{Cov}[w|\xi]$ is most likely not a tight bound. It is conjectured, for a constant A, the covariance of the prior could upper bound the conditional covariance $A\,\mathrm{Cov}_{P_0}[w] \succeq \mathrm{Cov}[w|\xi]$. This would require a lesser condition $Kd > C\beta N$ to achieve log-concavity of $p(\xi)$.

A. Additional technical proofs

A.1. Proofs for near constancy of $Z(w)$

In this section, we show the restriction of ξ to the set B is a highly likely event under the base Gaussian distribution, and $Z(w)$ has small magnitude first and second derivatives.

Proof of Lemma 4.2. We have the event

$$B = \left\{ \xi : \max_{j,k} \left| \sum_{i=1}^{N} x_{i,j} \xi_{i,k} \right| \leq N + z_{\delta/2Kd} \right\}.$$

We show that this event is likely for our conditionally independent Gaussian distributions for the $\xi_{i,k}$. This proof follows from standard Gaussian complexity arguments.

The object we must bound is $P(\xi \in B|w)$. If the $\xi_{i,k}$ given w are independent $\mathrm{Normal}(x_i \cdot w_k, 1/\rho)$, we may arrange a representation using independent standard normals Z_k of dimension n:

$$\xi_k = \mathbf{X} w_k + \frac{1}{\sqrt{\rho}} Z_k.$$

Each mean $x_i \cdot w_k$ is in $[-1, 1]$ due to the weight vector having bounded ℓ_1 norm and the data entries having bounded value. Consider the complement of the event we want to study. We wish for this event to have probability less than δ:

$$P\left[\max_{j,k} \left| \sum_{i=1}^{N} x_{i,j} \xi_{i,k} \right| \geq N + z_{\delta/2Kd} \sqrt{\frac{N}{\rho}} \right],$$

where P is the probability using the normal distribution of ξ given w. The max is upper bounded by

$$\max_{j,k} \left| \sum_{i=1}^{N} x_{i,j} \xi_{i,k} \right| \leq N + \max_{j,k} \left| \frac{1}{\sqrt{\rho}} \sum_{i=1}^{N} x_{i,j} Z_{i,k} \right|.$$

Thus we can bound the larger probability event uniformly for $w \in (S_1^d)^K$:

$$P\left[\max_{j,k} \frac{|\sum_{i=1}^{N} x_{i,j} Z_{i,k}|}{\sqrt{N}} \geq z_{\delta/2Kd} \right] \leq \delta,$$

where the conclusion follows from a union bound and a Gaussian tail bound. ∎

Proof of Lemma 4.3. We provide upper bounds on the magnitude of the first and second derivatives of the function $Z(w)$ as defined in equation (4.7). Denote Φ as the standard normal cumulative distribution function (CDF) and φ as the standard normal density function. Throughout the proof, recall that $p(w|\xi)$ treats each $\xi_{i,k}$ as independent normal with $\xi_{i,k} \sim \text{Normal}(x_i \cdot w_k, 1/\rho)$ conditionally independent given w. The absolute value of the gradient of $Z(w)$ inner product with a vector u with blocks u_k is

$$|u \cdot \nabla_w Z(w)| = \left| \rho E\left[\sum_{i=1}^{N} \sum_{k=1}^{K} (u_k \cdot x_i)(\xi_{i,k} - x_i \cdot w_k) \frac{1_B(\xi)}{P(\xi \in B|w)} | w \right] \right|. \quad \text{(A.1)}$$

By Lemma 4.2, the set B has probability at least $1 - \delta$. We note the following upper and lower bounds on the Gaussian CDF, which is provided by the classical results of Gordon [36]:

$$\frac{\varphi(x)}{x + 1/x} \leq 1 - \Phi(x) \leq \frac{\varphi(x)}{x}.$$

Consider the collections of all measurable sets $D \subset \mathbb{R}^{NK}$ such that $P(\xi \in D) \geq 1 - \delta$. This collection contains our original set B as an object in the class. Then, the absolute value of the expected inner product in (A.1) is less than the maximum for any set D in this class,

$$\max_{D: P(\xi \in D|w) \geq 1-\delta} \rho \frac{|E[\sum_{i=1}^{N} \sum_{k=1}^{K} (u_k \cdot x_i)(\xi_{i,k} - x_i \cdot w_k) 1_D(\xi) | w]|}{1 - \delta}. \quad \text{(A.2)}$$

Define the value

$$\widetilde{\sigma} = \sqrt{\frac{\sum_{i=1}^{N} \sum_{k=1}^{K} (u_k \cdot x_i)^2}{\rho}}.$$

Under the normal distribution for ξ, the integrand in question is a scalar mean 0 normal random variable with this variance,

$$\sum_{i=1}^{N} \sum_{k=1}^{K} (u_k \cdot x_i)(\xi_{i,k} - x_i \cdot w_k) \sim \text{Normal}(0, \widetilde{\sigma}^2).$$

Accordingly, since this random variable has mean zero, the expectation in (A.2) restricted to D is the same as (minus) the expectation restricted to its complement D^c, so the quantity of interest is

$$\max_{D:P(\xi\in D^c|w)\le\delta} \rho \frac{|E[\sum_{i=1}^N \sum_{k=1}^K (u_k \cdot x_i)(\xi_{i,k} - x_i \cdot w_k) 1_{D^c}(\xi)|w]|}{1-\delta}.$$

When $0 < \delta \le 1/2$, an extremal set D^c is then seen to correspond to the upper tail of this random variable having the specified probability. That is, the maximal value of this expression is achieved with

$$D^c = \left\{\xi : \frac{\sum_{i=1}^N \sum_{k=1}^K (u_k \cdot x_i)(\xi_{i,k} - x_i \cdot w_k)}{\widetilde{\sigma}} \ge z_\delta\right\},$$

where, as before, z_δ is the upper δ quantile of the standard normal with $1 - \Phi(z_\delta) = \delta$. (Because of the absolute value and the symmetry of the normal, we can also equally well take the left tail, where the standardized normal is less than $-z_\delta$.) From this extremal D^c, we then have the upper bound

$$|u \cdot \nabla_w Z(w)| \le \frac{\rho\widetilde{\sigma}}{1-\delta} \left| \int_{z_\alpha}^\infty z\varphi(z)\, dz \right| = \frac{\rho\widetilde{\sigma}}{1-\delta}\varphi(z_\delta),$$

using $-z\varphi(z) = \varphi'(z)$ and the fundamental theorem of calculus. By Gordon's inequality,

$$\varphi(z_\delta) \le \left(z_\delta + \frac{1}{z_\delta}\right)(1 - \Phi(z_\delta)),$$

which is not more than $\delta(\sqrt{2\log 1/\delta} + 1)$ for $\delta \le 1 - \Phi(1)$. This yields an upper bound on our expression of interest:

$$|u \cdot \nabla_w Z(w)| \le \frac{\rho\widetilde{\sigma}\delta}{1-\delta}\left(\sqrt{2\log(1/\delta)} + 1\right),$$

which notably goes to 0 as $\delta \to 0$.

The Hessian is then a difference in variances,

$$u^{\mathrm{T}}\left[\nabla^2 Z(w)\right]u = -\rho \sum_{i=1}^N \sum_{k=1}^K (x_i \cdot u_k)^2 \tag{A.3}$$

$$+ \rho^2 \operatorname{Var}\left[\sum_{i=1}^N \sum_{k=1}^K (x_i \cdot u_k)\xi_{i,k} | w, B\right]. \tag{A.4}$$

Note that the $\xi_{i,k}$ are independent with variance $1/\rho$ conditional on w, so if we did not constrain to the set B, expressions (A.3) and (A.4) would cancel to 0.

The object whose variance we take in (A.4) is a linear function of ξ, and ξ is a normal random variable given w with diagonal covariance matrix $1/\rho$. By an application of a Brascamp–Lieb inequality (see, for example, [15, Proposition 2.1]), we would have an upper bound on this variance by the norm square of the vector with entries $x_i \cdot u_k$, divided by ρ, which times ρ^2 is exactly expression (A.3). Thus, the term (A.4) is less than or equal to the absolute value of term (A.3). So an upper bound on the quadratic form is 0, that is,

$$u^{\mathrm{T}}[\nabla^2 Z(w)]u \leq 0.$$

We then compute a lower bound on the variance term in (A.4). Note a Cramer–Rao lower bound is not applicable here since restriction to a compact set makes integration by parts inapplicable due to boundary conditions. In particular, the expectation of the score of a constrained distribution is not always 0.

Using a bias-variance decomposition, write the variance as a non-centered expected squared difference minus the square of the expected difference:

$$\begin{aligned}
&\mathrm{Var}\left[\sum_{i=1}^{N}\sum_{k=1}^{K}(x_i \cdot u_k)\xi_{i,k}|w, B\right] \\
&\quad = E\left[\left(\sum_{i=1}^{N}\sum_{k=1}^{K}(x_i \cdot u_k)(\xi_{i,k} - x_i \cdot w_k)\right)^2 |w, \xi \in B\right] \\
&\qquad - \left(\sum_{i=1}^{N}\sum_{k=1}^{K}(u_k \cdot x_i)\big(E[\xi_{i,k}|w, \xi \in B] - x_i \cdot w_k\big)\right)^2 \qquad \text{(A.5)} \\
&\quad \geq E\left[\left(\sum_{i=1}^{N}\sum_{k=1}^{K}(x_i \cdot u_k)(\xi_{i,k} - x_i \cdot w_k)\right)^2 \frac{1_B(\xi)}{P(\xi \in B|w)}|w\right] \qquad \text{(A.6)} \\
&\qquad - \frac{\widetilde{\sigma}^2\delta^2}{(1-\delta)^2}\big(2\log(1/\delta) + 3\big), \qquad \text{(A.7)}
\end{aligned}$$

where we have applied the previously derived bound on expression (A.5) along with

$$\left(z_\delta + \frac{1}{z_\delta}\right)^2 = z_\delta^2 + 2 + \frac{1}{z_\delta},$$

which is not more than $z_\delta^2 + 3$ to deduce expression (A.7).

If we did not impose a condition on the set B, the expression (A.6) would be the variance of a simple normal variable with variance $\widetilde{\sigma}^2$. We will show restricting to B still results in a value very close to $\widetilde{\sigma}^2$.

The set B has probability at least $1-\delta$. Then, the expectation in (A.6) restricted to B is lower bounded by the minimum among sets D with $P(\xi \in D) \geq 1-\delta$:

$$E\left[\left(\sum_{i=1}^{N}\sum_{k=1}^{K}(x_i \cdot u_k)(\xi_{i,k} - x_i \cdot w_k)\right)^2 \frac{1_B(\xi)}{P(\xi \in B|w)}|w\right]$$
$$\geq \min_{D:P(\xi\in D|w)\geq 1-\delta} \frac{E[(\sum_{i=1}^{N}\sum_{k=1}^{K}(x_i \cdot u_k)(\xi_{i,k} - x_i \cdot w_k))^2 1_D(\xi)|w]}{1-\delta}.$$

The integrand in question, as before, is the same normal variable now squared. The minimizing set D^* is then the set placing an upper bound on that expression:

$$D^* = \left\{\xi : -\tau \leq \frac{\sum_{i=1}^{N}\sum_{k=1}^{K}(x_i \cdot u_k)(\xi_{i,k} - x_i \cdot w_k)}{\widetilde{\sigma}} \leq \tau\right\}$$

for some value τ, the proper choice being $\tau = z_{\delta/2}$, which achieves the specified probability $1-\delta$.

As an aside, we note this set D^* can be deduced from the Neyman–Pearson lemma [50, Theorem 3.2.1], comparing the distribution where each $\xi_{i,k}$ has independent normal with mean $x_i \cdot w_k$ and variance $1/\rho$, to the distribution which has this normal density times $(\sum_{i=1}^{N}\sum_{k=1}^{K}(x_i \cdot u_k)(\xi_{i,k} - x_i \cdot w_k))^2$.

We then integrate a squared normal on a truncated range and have a lower bound:

$$\min_{D:P(\xi\in D|w)\geq 1-\delta} \frac{E[(\sum_{i=1}^{N}\sum_{k=1}^{K}(x_i \cdot u_k)(\xi_{i,k} - x_i \cdot w_k))^2 1_D(\xi)|w]}{1-\delta}$$
$$= \frac{\widetilde{\sigma}^2}{1-\delta}\int_{-z_{\delta/2}}^{z_{\delta/2}} z^2\varphi(z)\,dz. \tag{A.8}$$

To evaluate this integral use its complement set and symmetry of the normal pdf:

$$\int_{-z_{\delta/2}}^{z_{\delta/2}} z^2\varphi(z)\,dz = 1 - 2\int_{z_{\delta/2}}^{\infty} z^2\varphi(z)\,dz.$$

Applying integration by parts, using $(z\varphi(z))' = -z^2\varphi(z) + \varphi(z)$, this is

$$1 + 2z\varphi(z)|_{z_{\delta/2}}^{\infty} - 2\int_{z_{\delta/2}}^{\infty}\varphi(z)\,dz = 1 - 2z_{\delta/2}\varphi(z_{\delta/2}) - \delta.$$

Using $z\varphi(z) \leq (z^2+1)(1-\Phi(z))$ (from Gordon's inequality), which at $z_{\delta/2}$ is less than $(2\log(2/\delta)+1)\delta/2$, this gives a lower bound for the expression in (A.8):

$$\frac{\widetilde{\sigma}^2}{1-\delta}\big(1 - 2\delta\big(\log(2/\delta)+1\big)\big), \tag{A.9}$$

which converges to $\widetilde{\sigma}^2$ as $\delta \to 0$. We then combine expressions (A.4), (A.7), and (A.9) to give a lower bound on the Hessian quadratic form:

$$\begin{aligned} u^{\mathrm{T}}\big[\nabla^2 Z(w)\big]u &\geq \rho^2\widetilde{\sigma}^2\bigg[-1 + \frac{1}{1-\delta}\big(1 - 2\delta\big(\log(2/\delta) + 1\big)\big) - \frac{\delta^2}{(1-\delta)^2}\big(2\log(1/\delta) + 3\big)\bigg] \\ &= -\rho^2\widetilde{\sigma}^2\bigg[\frac{\delta}{1-\delta}\big(2\log(2/\delta) + 1\big) + \frac{\delta^2}{(1-\delta)^2}\big(2\log(1/\delta) + 3\big)\bigg], \end{aligned}$$

which converges to 0 as $\delta \to 0$. This bound completes the proof.

For a refinement, one may use the expressions involving z_δ and $z_{\delta/2}$ directly:

$$\begin{aligned} u^{\mathrm{T}}\big[\nabla^2 Z(w)\big]u &\geq \rho^2\widetilde{\sigma}^2\bigg[-1 + \frac{1}{1-\delta}\big(1 - 2z_{\delta/2}\varphi(z_{\delta/2}) - \delta\big) - \frac{1}{(1-\delta)^2}\big(\varphi(z_\delta)\big)^2\bigg] \\ &= -\rho^2\widetilde{\sigma}^2\frac{1}{1-\delta}\bigg[2z_{\delta/2}\varphi(z_{\delta/2}) + \frac{1}{1-\delta}\big(\varphi(z_\delta)\big)^2\bigg]. \end{aligned}$$

■

A.2. Log-concavity of $p^*(w|\xi)$ with conditioning on ξ in the set B

In this section, we show the conditioning of ξ given w to the set B does not affect the log-concavity of the reverse conditional much.

Proof of Theorem 4.4. We prove the reverse conditional is log-concave when restricting ξ to live in the set B. This proof follows much the same way as Theorem 4.1. The log of $p^*(w|\xi)$ is given by

$$\begin{aligned} \log p^*(w|\xi) = &-\beta\ell(w) + B^*(\xi) \\ &- \sum_{i=1}^{N}\sum_{k=1}^{K}\frac{\rho}{2}(\xi_{i,k} - w_k \cdot x_i)^2 \\ &- Z(w) \end{aligned} \tag{A.10}$$

for some function $B^*(\xi)$, which does not depend on w and is only required to make the density integrate to 1. The term (A.10) is a negative quadratic in w which treats each w_k as if it were an independent normal random variable. Thus, the additional Hessian contribution will be a $(Kd) \times (Kd)$ negative definite block diagonal matrix with $d \times d$ blocks of the form

$$\rho\sum_{i=1}^{N} x_i x_i^{\mathrm{T}}.$$

Denote the Hessian as

$$H(w|\xi) \equiv \nabla^2 \log p^*(w|\xi).$$

For any vector $u \in \mathbb{R}^{Kd}$ with blocks $u_k \in \mathbb{R}^d$, the quadratic form $u^{\mathrm{T}} H(w|\xi)u$ can be expressed as

$$-\beta \sum_{i=1}^{N} \Big(\sum_{k=1}^{K} \psi'(w_k \cdot x_i) u_k \cdot x_i \Big)^2$$

$$+ \sum_{k=1}^{K} \sum_{i=1}^{N} (u_k \cdot x_i)^2 \big[\beta \operatorname{res}_i(w) c_k \psi''(w_k \cdot x_i) - \rho \big] \tag{A.11}$$

$$- u^{\mathrm{T}} (\nabla^2 Z(w)) u. \tag{A.12}$$

Let us set $\rho = (1+\varepsilon) a_2 C_N V / K$ with positive ε, so that it is at least a little bigger than the original ρ_0. With $\varepsilon = 1$ this is the $\rho = 2\rho_0$ assumed in the theorem statement, but it may be of interest to explore other choices. By the assumptions on the second derivative of ψ, we have

$$\max_{i,k} \big(\beta \operatorname{res}_i(w) c_k \psi''(w_k \cdot x_i) - \rho \big) \le -\varepsilon a_2 \frac{\beta C_N V}{K},$$

so all the terms in the sum in (A.11) are negative. Recall from the definition of $\widetilde{\sigma}^2$ that

$$\sum_{k=1}^{K} \sum_{i=1}^{N} (u_k \cdot x_i)^2 = \rho \widetilde{\sigma}^2.$$

Therefore, expression (A.11) is less than

$$-(1+\varepsilon)\varepsilon \left(a_2 \frac{\beta C_N V}{K} \right)^2 \widetilde{\sigma}^2 = -\frac{\varepsilon}{1+\varepsilon} \rho^2 \widetilde{\sigma}^2.$$

By Lemma 4.3, for positive $\delta \le 0.158$, a bound on the Hessian term from the correction function Z is

$$-u^{\mathrm{T}} \big(\nabla^2 Z(w) \big) u \le \rho^2 \widetilde{\sigma}^2 \frac{\delta}{1-\delta} \left(2\log(2/\delta) + 1 + \frac{\delta}{1-\delta} \big(2\log(1/\delta) + 3 \big) \right).$$

We recognize the right-hand side as $\rho^2 \widetilde{\sigma}^2 H(\delta)$, where $H(\delta) = H_1(\delta) + H_2(\delta)$ is given by

$$H(\delta) = \frac{\delta}{1-\delta} \big(2\log(2/\delta) + 1 \big) + \frac{\delta^2}{(1-\delta)^2} \big(2\log(1/\delta) + 3 \big),$$

an expression that is monotone in δ for the allowed range and tends to zero as $\delta \to 0$.

Thus, term (A.11) plus (A.12) is less than or equal to $\rho^2 \widetilde{\sigma}^2$ times the following quantity, which we want to be non-positive:

$$-\frac{\varepsilon}{1+\varepsilon} + H(\delta).$$

If $\varepsilon = 1$, then $\varepsilon/(1+\varepsilon)$ is $1/2$. Then by evaluation, $-1/2 + H(\delta)$ is seen to be negative for δ not more than 0.054, and less than $-1/4$ for δ not more than 0.0242. This completes the proof of the log-concavity of $p^*(w|\xi)$ for the specified ρ with $\varepsilon = 1$.

As long as $H(\delta) < 1$, which is so for positive $\delta < 0.115$, log-concavity can be arranged by a suitable choice of ε. Indeed, we see that $\varepsilon = H(\delta)/(1 - H(\delta))$ makes the expression $-\varepsilon/(1+\varepsilon) + H(\delta)$ equal to zero. With this choice, we have

$$1 + \varepsilon = \frac{1}{1 - H(\delta)}.$$

This shows $p^*(w|\xi)$ is log-concave with ρ at least $\rho_\delta = \rho_0/(1 - H(\delta))$, that is,

$$\rho_\delta = \frac{a_2}{1 - H(\delta)} \frac{\beta C_N V}{K}.$$

This allows ρ to be near the original ρ_0 for small δ.

If we use the refined form of the bound on the Hessian of $Z(w)$, the above expression for $H(\delta)$ is replaced by the smaller value

$$H^*(\delta) = \frac{1}{1-\delta}\left[2 z_{\delta/2} \varphi(z_{\delta/2}) + \frac{1}{1-\delta}\big(\varphi(z_\delta)\big)^2\right].$$

Accordingly, we can improve the result to obtain log-concavity for a larger range of δ (as $H^*(\delta)$ remains less than 1 for $\delta < 0.160$ with no need to impose the restriction that z_δ be at least 1), and this log-concavity is obtained for smaller ρ, at least $\rho_0/(1 - H^*(\delta))$. ∎

A.3. Proofs of bounds from the Hölder inequality

In this section, we bound the two factors in the Hölder inequality. First, we need a supporting lemma.

Lemma A.1. *For any vector $x \in [-1, 1]^d$ and any integer $\ell > 0$, the expected inner product with random vector w from the continuous uniform distribution on S_1^d raised to the power 2ℓ is upper bounded by*

$$E_{P_0}\left[\left(\sum_{j=1}^{d} x_j w_j\right)^{2\ell}\right] \le \frac{1}{(d)^\ell} \frac{(2\ell)!}{\ell!}.$$

Proof. The sum $\sum_{j=1}^d x_j w_j$ raised to the power 2ℓ can be expressed as a sum using a multi-index $J = (j_1, \ldots, j_{2\ell})$, where each $j_i \in \{1, \ldots, d\}$ and there are $d^{2\ell}$ terms:

$$E\left[\left(\sum_{j=1}^{d} x_j w_j\right)^{2\ell}\right] = \sum_{j_1,\ldots,j_{2\ell}} \prod_{i=1}^{2\ell} (x_{j_i})\, E\left[\prod_{i=1}^{2\ell} w_{j_i}\right].$$

For a given multi-index vector J, let $r(j, J)$ count the number of occurrences of the value j in the vector

$$r(j, J) = \sum_{i=1}^{2\ell} 1\{j_i = j\}.$$

Then for any multi-index, we would have

$$\prod_{i=1}^{2\ell} w^{j_i} = \prod_{j=1}^{d} w_j^{r(j,J)}.$$

Abbreviate $r_j = r(j, J)$ for a fixed vector J, also note

$$\sum_{j=1}^{d} r_j = 2\ell.$$

Consider the expectation $E[\prod_{i=1}^{d} w_j^{r_j}]$. Due to the symmetry of the prior, if any of the r_j are odd, then the whole expectation is 0. Thus, we only consider vectors

$$\vec{r} = (r_1, \ldots, r_d),$$

where all entries are even. If we fix the signs of the w_j points to live in a given orthant, then the distribution is uniform on the $(d + 1)$-dimensional simplex. Define

$$w_{d+1} = 1 - \sum_{j=1}^{d} |w_j|,$$

then $(|w_1|, \ldots, |w_d|, w_{d+1})$ has a symmetric Dirichlet $(1, \ldots, 1)$ distribution in $d + 1$ dimensions. Note, a general Dirichlet distribution in $d + 1$ dimensions with parameter vector $\vec{\alpha} = (\alpha_1, \ldots, \alpha_{d+1})$ has a properly normalized density as

$$p_{\vec{\alpha}}(w_1, \ldots, w_d) = \frac{\Gamma(\sum_{j=1}^{d} \alpha_j)}{\prod_{j=1}^{d+1} \Gamma(\alpha_j)} \prod_{j=1}^{d} (w_j)^{\alpha_j - 1} \left(1 - \sum_{j=1}^{d} w_j\right)^{\alpha_{d+1} - 1}.$$

Thus, the expectation of $\prod_{j=1}^{d} w_j^{r_j}$ with respect to a symmetric Dirichlet has the form of an unnormalized $\mathrm{Dir}(r_1 + 1, \ldots, r_d + 1, 1)$ distribution. Thus, the expectation is a ratio of their normalizing constants:

$$E\left[\prod_{j=1}^{d} w_j^{r_j}\right] = \frac{\Gamma(d + 1) \prod_{j=1}^{d} \Gamma(r_j + 1)}{\Gamma(d + 1 + \sum_{j=1}^{d} r_j)} = \frac{d! \prod_{j=1}^{d} r_j!}{(d + 2\ell)!}.$$

The number of times a specific vector $\vec{r}$ appears from the multi-index J is

$$\frac{(2\ell)!}{\prod_{j=1}^{d} r_j!},$$

and thus we have

$$\begin{aligned}
E\left[\left(\sum_{j=1}^{d} x_j w_j\right)^{2\ell}\right] &= \sum_{\substack{\vec{r} \text{ even}, \\ \sum_j r_j = 2\ell}} \prod_{j=1}^{d} (x_j)^{r_j} \frac{(2\ell)!}{\prod_{j=1}^{d} r_j!} E\left[\prod_{j=1}^{d} w_j^{r_j}\right] \\
&= \frac{(2\ell!)(d!)}{(d+2\ell)!} \sum_{\substack{\vec{r} \text{ even}, \\ \sum_j r_j = 2\ell}} \prod_{j=1}^{d} (x_j)^{r_j} \\
&= \frac{(2\ell!)(d!)}{(d+2\ell)!} \sum_{\substack{\vec{r} \text{ even}, \\ \sum_j r_j = 2\ell}} \prod_{j=1}^{d} (x_j^2)^{r_j/2} \\
&\leq \frac{(2\ell!)(d!)}{(d+2\ell)!} \frac{(d+\ell-1)!}{\ell!(d-1)!} \qquad \text{(A.13)} \\
&= \frac{(d+\ell-1)\cdots(d)}{(d+2\ell)\cdots(d+1)} \frac{(2\ell)!}{(\ell)!} \leq \frac{1}{d^\ell} \frac{2\ell!}{\ell!},
\end{aligned}$$

where inequality (A.13) follows from each $x_j^2 \leq 1$, thus each term in the sum is less than 1, and there being $\binom{d+\ell-1}{\ell}$ terms in the sum. ∎

Proof of Lemma 4.10. We bound the first term in the Hölder inequality depending on the higher order moments of the prior. We have the unit vector $u \in \mathbb{R}^{nK}$ with n-dimensional blocks u_k. Define vectors in $\mathbb{R}^d$ as $v_k = \mathbf{X}^{\mathrm{T}} u_k$, and the object we study is

$$E\left[\left(\sum_{k=1}^{K} v_k \cdot w_k\right)^{2\ell}\right].$$

Use a multinomial expansion of this power of a sum and we have the expression

$$\begin{aligned}
&E\left[\sum_{\substack{j_1,\ldots,j_K, \\ \sum j_k = 2\ell}} \binom{2\ell}{j_1,\ldots,j_K} \prod_{k=1}^{K} (v_k \cdot w_k)^{j_k}\right] \\
&\quad = \sum_{\substack{j_1,\ldots,j_K \\ \sum j_k = 2\ell}} \binom{2\ell}{j_1,\ldots,j_K} \prod_{k=1}^{K} E\left[(v_k \cdot w_k)^{j_k}\right],
\end{aligned}$$

since the prior treats each neuron weigh vector w_k as independent and uniform on S_1^d.

By the symmetry of the prior, if any j_k are odd, then the whole expression is 0, and thus we only sum using even j_k values:

$$\sum_{\substack{j_1,\ldots,j_K,\\ \sum j_k=\ell}} \binom{2\ell}{2j_1,\ldots,2j_K} \prod_{k=1}^{K} E\big[(v_k \cdot w_k)^{2j_k}\big]. \tag{A.14}$$

Each vector v_k is a linear combination of the rows of the data matrix

$$v_k = \sum_{i=1}^{N} u_{k,i} x_i.$$

Define $s_{k,i} = \text{sign}(u_{k,i})$ and $\alpha_{k,i} = |u_{k,i}|/\|u_k\|_1$. We can then interpret the above inner product as a scaled expectation on the data indexes:

$$v_k \cdot w_k = \big(\|u_k\|_1\big) \sum_{i=1}^{N} \alpha_{k,i} s_{k,i}\, x_i \cdot w_k.$$

The average is then less than the maximum term in index i:

$$\begin{aligned} E\big[(v_k \cdot w_k)^{2j_k}\big] &= \big(\|u_k\|_1\big)^{2j_k} E\left[\left(\sum_{i=1}^{N} \alpha_{k,i} s_{k,i}\, x_i \cdot w_k\right)^{2j_k}\right] \\ &\le \big(\|u_k\|_1\big)^{2j_k} \sum_{i=1}^{N} \alpha_{k,i} E\big[(x_i \cdot w_k)^{2j_k}\big] \\ &\le \big(\|u_k\|_1\big)^{2j_k} \max_i E\big[(x_i \cdot w_k)^{2j_k}\big] \\ &\le \big(\|u_k\|_1\big)^{2j_k} \frac{1}{(d)^{j_k}} \frac{(2j_k)!}{j_k!}, \end{aligned}$$

where we have applied Lemma A.1. We then plug this result into equation (A.14),

$$\frac{1}{d^\ell} \frac{(2\ell!)}{\ell!} \left(\sum_{\substack{j_1,\ldots,j_K,\\ \sum j_k=\ell}} \binom{\ell}{j_1,\ldots,j_K} \prod_{k=1}^{K} \big(\|u_k\|_1\big)^{2j_k} \right) = \frac{1}{d^\ell} \frac{(2\ell)!}{\ell!} \left(\sum_{k=1}^{K} \|u_k\|_1^2 \right)^{\ell}.$$

For each sub-block u_k of dimension n, we have $\|u_k\|_1^2 \le n\|u_k\|_2^2$, and since the full vector u is a unit vector, we have

$$\sum_{k=1}^{K} \|u_k\|^2 = 1,$$

which gives the upper bound

$$\frac{n^\ell (2\ell)!}{d^\ell \ell!}.$$

Via Stirling's bound [61], we obtain

$$\sqrt{2\pi\ell}\Big(\frac{\ell}{e}\Big)^{\ell} e^{1/(12\ell+1)} \le \ell! \le \sqrt{2\pi\ell}\Big(\frac{\ell}{e}\Big)^{\ell} e^{1/12\ell}.$$

Taking the ℓ root, we have

$$\begin{aligned}\left(\frac{n^{\ell}}{d^{\ell}}\frac{(2\ell)!}{\ell!}\right)^{1/\ell} &\le \frac{N}{d}\left(2^{2\ell+1/2}\Big(\frac{\ell}{e}\Big)^{\ell} e^{1/(24\ell)-1/(12\ell+1)}\right)^{1/\ell} \\ &= \frac{2^{2+1/(2\ell)}n\ell}{d} e^{1/(24\ell^2)-1/(12\ell^2+\ell)-1} \\ &\le \frac{4n\ell}{d}\sqrt{2}e^{1/24+1/13-1} \le \frac{4n\ell}{\sqrt{e}d}.\end{aligned}$$

■

Proof of Lemma 4.11. We bound the second term in the Hölder inequality determined by the growth rate of the cumulant-generating function. By the mean value theorem, there exists some value $\widetilde{\tau} \in [1, \ell/(\ell-1)]$ such that

$$\Gamma_{\xi}\Big(\frac{\ell}{\ell-1}\Big) = \Gamma_{\xi}(1) + (\Gamma_{\xi})'(\widetilde{\tau})\Big[\frac{\ell}{\ell-1} - 1\Big].$$

Rearranging, we can express the difference

$$\frac{\ell-1}{\ell}\Gamma_{\xi}\Big(\frac{\ell}{\ell-1}\Big) - \Gamma_{\xi}(1) = (\Gamma_{\xi})'(\widetilde{\tau})\frac{1}{\ell} - \frac{1}{\ell}\Gamma_{\xi}(1).$$

By construction, $\Gamma_{\xi}(\tau)$ is an increasing convex function with $\Gamma_{\xi}(0) = 0$. Therefore, $\Gamma_{\xi}(1) > 0$ and we can study the upper bound

$$\frac{\ell-1}{\ell}\Gamma_{\xi}\Big(\frac{\ell}{\ell-1}\Big) - \Gamma_{\xi}(1) \le (\Gamma_{\xi})'(\widetilde{\tau})\frac{1}{\ell}.$$

Recall $\Gamma_{\xi}(\tau)$, defined in equation (4.14), is a cumulant-generating function of $\widetilde{h}_{\xi}^{N}(w)$. Thus, its derivative at $\widetilde{\tau}$ is the mean of $\widetilde{h}_{\xi}(w)$ under the tilted distribution. The mean is then less than the maximum difference of any two points on the constrained support set:

$$(\Gamma_{\xi})'(\widetilde{\tau}) = E_{\widetilde{\tau}}\big[\widetilde{h}_{\xi}(w)|\xi\big] \le \max_{w,w_0\in(S_1^d)^K}\big(\widetilde{h}_{\xi}(w) - \widetilde{h}_{\xi}(w_0)\big).$$

By the mean value theorem, for any choice of $w, w_0 \in (S_1^d)^K$, there exists a $\widetilde{w} \in (S_1^d)^K$ along the line between w and w_0 such that

$$\widetilde{h}_{\xi}(w) - \widetilde{h}_{\xi}(w_0) = \nabla_w \widetilde{h}_{\xi}(\widetilde{w}) \cdot (w - w_0).$$

For each k, the gradient in w_k, evaluated at $\widetilde{w}$ is

$$\nabla_{w_k} \widetilde{h}_\xi(\widetilde{w}) = \sum_{i=1}^{N} \big(\beta \,\mathrm{res}_i(\widetilde{w}) c_k \psi'(\widetilde{w}_k \cdot x_i) - \rho w_k \cdot x_i\big) x_i \tag{A.15}$$

$$+ \rho\beta \sum_{i=1}^{N} \xi_{i,k} x_i + \nabla_{w_k} Z(w). \tag{A.16}$$

Since $|c_k| \le V/K$ and each $|x_i \cdot \widetilde{w}_k| \le 1$, with the choice $\rho = 2a_2 \beta C_N V/K$, the multipliers in the sum in (A.15), for each i, satisfy

$$\beta \Big| \mathrm{res}_i(\widetilde{w}) c_k \psi'(w_k \cdot x_i) - 2a_2 \frac{C_N V}{K} [w_k \cdot x_i] \Big| \le (a_1 + 2a_2) \frac{\beta C_N V}{K}.$$

Likewise, we have

$$\|x_i \cdot (w_k - w_{0,k})\|_1 \le 2.$$

Accordingly, the inner product of $w_k - w_{0,k}$ with the first term of the gradient is bounded as

$$\Big[\beta \sum_{i=1}^{N} \Big(\mathrm{res}_i(\widetilde{w}) c_k \psi'(\widetilde{w}_k \cdot x_i) - 2a_2 \frac{C_N V}{K}\Big) x_i\Big] \cdot (w_k - w_{0,k})$$

$$\le 2(a_1 + 2a_2) \frac{C_N V \beta N}{K}. \tag{A.17}$$

As for the second term, we have

$$\rho \Big[\sum_{i=1}^{N} \xi_{i,k} x_i\Big] \cdot (w_k - w_{0,k}) \le 2\rho \max_j \Big| \sum_{i=1}^{N} \xi_{i,k} x_{i,j} \Big|.$$

Our original restriction of ξ to the set B is specifically designed to control this term. By definition of the set B, for all k, we have

$$\max_j \Big| \sum_{i=1}^{N} \xi_{i,k} x_{i,j} \Big| \le N + \sqrt{2\log(2Kd/\delta)} \sqrt{\frac{N}{\rho}}.$$

Multiplying by 2ρ, it yields the bound on the second term of

$$2\rho N + 2\sqrt{\rho N} \sqrt{2\log(2Kd/\delta)}$$

$$= \frac{4a_2 C_N V N}{K} + 2\sqrt{\frac{2a_2 C_N V \beta N}{K}} \sqrt{2\log(2Kd/\delta)}. \tag{A.18}$$

For the final term, $Z(w)$ is shown to have small derivative. By Lemma 4.3,

$$\begin{aligned}
&\sum_k \nabla_{w_k} Z(w) \cdot (w_k - w_{0,k}) \\
&\quad \le \sqrt{\rho} \sqrt{\sum_{i=1}^{N} \sum_{k=1}^{K} \big((w_k - w_{0,k}) \cdot x_i\big)^2} \, \frac{\delta}{1-\delta} \big(2\log(1/\delta) + 1\big) \\
&\quad \le 2\sqrt{\rho}\,\sqrt{NK}\, \frac{\delta}{1-\delta} \big(2\log(1/\delta) + 1\big) \\
&\quad = 2\sqrt{2a_2 C_N V \beta N}\, \frac{\delta}{1-\delta} \big(2\log(1/\delta) + 1\big). \qquad \text{(A.19)}
\end{aligned}$$

Summing using index k for the terms (A.17) and (A.18), and combining with the term (A.19), we can upper bound $(\Gamma_\xi)'(\widetilde{\tau})$ as

$$\begin{aligned}
&C_N V \beta N (2a_1 + 8a_2) \\
&\quad + 2\sqrt{2a_2 C_N V \beta N} \left(\sqrt{2\log(2Kd/\delta)}\sqrt{K} + \frac{\delta}{1-\delta} \big(2\log(1/\delta) + 1\big) \right). \qquad \text{(A.20)}
\end{aligned}$$

The last part of order $\delta \log 1/\delta$ is negligible compared to the other contributions. With the assumptions $d \ge 2$ and $K \ge 2$, we find that for all values in the range $0 \le \delta \le 0.116$, we have the inequality

$$\frac{\delta}{1-\delta} \big(2\log(1/\delta) + 1\big) \le \frac{1}{4}\sqrt{2\log(8/\delta)} \le \frac{1}{4}\sqrt{2\log(2Kd/\delta)},$$

where the form of this bound is chosen for convenience in combining with the second term of (A.20). This gives a bound on $(\Gamma_\xi)'(\widetilde{\tau})$ of

$$C_N V \beta N (2a_1 + 8a_2) + \sqrt{2a_2 C_N V \beta N}\, \sqrt{2\log(2Kd/\delta)} \left(\sqrt{K} + \frac{1}{4} \right).$$

Finally, dividing by ℓ gives the upper bound on $((\ell-1)/\ell)\Gamma_\xi(\ell/(\ell-1)) - \Gamma_\xi(1)$ as claimed:

$$\frac{1}{\ell} \left[C_N V \beta N (2a_1 + 8a_2) + \sqrt{2a_2 C_N V \beta N}\, \sqrt{\log(2Kd/\delta)} \left(\sqrt{K} + \frac{1}{4} \right) \right].$$

When convenient, we also use that $\sqrt{K} + 1/4$ is not more than $\sqrt{1.4K}$ for $K \ge 2$. ∎

If in place of ρ we use

$$\rho_\delta = \frac{1}{1 - H(\delta)} a_2 \frac{C_N \beta V}{K},$$

defined in the proof of Theorem 4.4, with the multiplier $1/(1-H(\delta))$ in place of 2, then we have a refined bound valid for all δ with $H(\delta) < 1$ (which includes values of δ less than 0.115). This refined bound on $((\ell-1)/\ell)\Gamma_\xi(\ell/(\ell-1)) - \Gamma_\xi(1)$ is

$$\frac{1}{\ell}\Bigg[C_N V\beta N\left(2a_1 + 4\,\frac{1}{1-H(\delta)}\,a_2\right) + \sqrt{\frac{1}{1-H(\delta)}a_2 C_N V\beta N}\,\sqrt{\log(2Kd/\delta)}\left(\sqrt{K}+\frac{1}{4}\right)\Bigg].$$

It provides smaller bounds when δ is less than 0.054, that is, when $H(\delta) \le 1/2$. Armed with this ρ_δ, in Lemma 4.11 and in Theorem 4.12 we would use

$$A_{1,\delta} = 2a_1 + 4\,\frac{1}{1-H(\delta)}\,a_2, \quad A_{2,\delta} = 2\sqrt{\frac{1}{1-H(\delta)}\,a_2}$$

in place of A_1 and A_2, respectively. In particular, we would plug these in to provide an improved A_3 in Theorem 4.12 .

While primarily studying the properties of $p^*(\xi)$, it could be of interest to know whether $p(\xi)$ is log-concave at least for the relevant part of its domain when ξ is in B (see the implications below). Accordingly, we can also control the conditional variances using $p(w|\xi)$ instead of $p^*(w|\xi)$.

Toward that end, the analysis above also bounds the second factor of the Hölder inequality, for ξ in B, when the cumulant-generating functions uses $p(w|\xi)$. The only difference is the absence of the $Z(w)$ factor, so that there is no $|\nabla Z(\widetilde{w})\cdot(w-w_0)|$ part. Then there is no $(2\delta/(1-\delta))(2\log(1/\delta)+1)$ part in the above bound, and the final displayed constant is improved with $\sqrt{K}$ in place of $\sqrt{1.4K}$, allowing log-concavity for a slightly larger range of K.

For certain choices of the parameters, it could happen that $p^*(w|\xi)$ or $p^*(\xi)$ are not log-concave, while $p(w|\xi)$ and $p(\xi)$ are log-concave for ξ in B. When that happens, it is still possible to establish rapid mixing using certain samplers from $p^*(w|\xi)$ and $p^*(\xi)$ via establishment of log-Sobolev properties, as discussed in the next section.

A.4. Log-Sobolev analysis

Note that the $Z(w)$ is $\log p(\xi \in B|w)$. Since it is log of a probability it is at most 0. Lemma 4.2 gives a lower bound, so $Z(w)$ has the following bounded range:

$$\log(1-\delta) \le Z(w) \le 0.$$

Recall the density $p(w|\xi)$ analyzed in Theorem 4.1 without the $Z(w)$ correction, and using the original $\rho = \rho_0$ (with multiplier 1). This density $p(w|\xi)$ is weakly log-concave and it has support contained within the Euclidean ball $B(0, K)$ of radius K, since each neuron weight vector has independent bounded 1 norm. As such, $p(w|\xi)$ has control on its log-Sobolev constants via standard theory as we discuss here.

Consider all smooth functions f, that is, all bounded functions with bounded derivatives. For a density μ, the log-Sobolev constant $C_{LS}(\mu)$ is the smallest constant C such that for all smooth f, we have

$$\mathrm{Ent}_\mu(f) = E_\mu\big[f^2 \ln(f^2)\big] - E_\mu[f^2] \ln\big(E_\mu[f^2]\big) \le C E_\mu\big[|\nabla f|^2\big].$$

Equivalently, the Kullback divergence from μ of distributions with density w.r.t μ of the form $g = f^2$ is upper bound by a constant times the relative Fisher Information. The seminal conclusion of Bakry–Emery [3, 4] established that strongly log-concave densities have a log-Sobolev constant. Interestingly, when the support is bounded, weakly log-concave densities also have an identified log-Sobolev constant determined by the squared radius of the support.

Lemma A.2 ([14, Corollary 2.3]). *For any log-concave measure with support contained within the Euclidean ball $B(0, r)$ of radius r, the log-Sobolev constant is upper bound by $10r^2$.*

Control of a log-Sobolev constant is a standard condition for rapid mixing of various MCMC algorithms, so we want to understand the log-Sobolev constant of the measure $p^*(w|\xi)$ which includes the $Z(w)$ term. In this section, we allow ρ to be the original ρ_0, so we are not accounting for the Hessian of $Z(w)$ (so we do not necessarily have log-concavity of $p^*(w|\xi)$). Nevertheless, the $Z(w)$ may be thought of as a small magnitude correction to the otherwise concave exponent, as quantified by Lemma 4.2. This concave exponent is not necessarily strictly concave, especially when $d > NK$, leading to choices of direction u where the individual $u \cdot x_i$ in the Hessian quadratic form are zero. Nevertheless, the ℓ_1 constraint sets for w_k, for $k = 1, \ldots, K$, lead to w being contained in a Euclidean ball of radius K, for which we have the quantified log-Sobolev constant. As such, a log-Sobolev constant for the distribution with density $p^*(w|\xi)$ can be given as an application of [14] together with the perturbation lemma of Holley–Stroock ([17, Theorem 1.1], [39]).

Lemma A.3. *Since $p^*(w|\xi)$ is a bounded perturbation to the original $p(w|\xi)$ density, with their ratio between $1 - \delta$ and $1/(1 - \delta)$ and support in $B(0, K)$, it follows that for each ξ, the log-Sobolev constant of this density is upper bounded by*

$$C_{LS}\big(p^*(\cdot|\xi)\big) \le \frac{1}{1-\delta} 10K^2.$$

The same strategy can be applied to the analysis of the marginal density $p^*(\xi)$, which has the bounded support set B. Within B this $p^*(\xi)$ can be seen to be a perturbation of the density $p(\xi)$. Indeed in B, we have

$$\frac{p^*(\xi)}{p(\xi)} = \frac{\int p(w)p(\xi|w)e^{-Z(w)}\,\eta(dw)}{\int p(w)p(\xi|w)\,\eta(dw)},$$

which we recognize as

$$\int p(w|\xi)e^{-Z(w)}\,\eta(dw).$$

Accordingly, this ratio is between 1 and $1/(1-\delta)$, which shows the perturbation property. As for the log-concavity of $p(\xi)$ on the subset B, this holds when

$$\rho \mathrm{Var}_p\big[v(w)|\xi\big] \le 1.$$

As developed above, this can hold for a slightly larger range of values of K than for the log-concavity of $p^*(\xi)$.

It follows that when $p(\xi)$ is log-concave on the subset B, then $p^*(\xi)$ also has a log-Sobelev constant of not more than $10r^2/(1-\delta)$, where r is the radius of the smallest ball containing B. The set B constrains

$$\xi = \big(\xi_{i,k} : i \in [1,N], k \in [1,K]\big)$$

by requiring each column $\xi_k = (\xi_{1,k}, \ldots, \xi_{N,k})'$ to be in the polygonal region determined by d pairs of linear inequality conditions of the form $|\sum_i x_{i,j}\xi_{i,k}|$ being less than a specified value. There is a pair of such conditions for each $j \in [1,d]$. So, with d greater than N, there is the potential for enough linear constraints such that the set B is bounded. However, this depends on the nature of the design

$$X = \big(x_{i,j} : i \in [1,N], j \in [1,d]\big).$$

Rather than tackle that complication for the sake only of obtaining a log-Sobolev property, we have preferred in this paper to be slightly more restrictive regarding the allowed parameters such as d and K and obtain the stronger property of strict log-concavity of $p^*(\xi)$ directly.

A.5. Alternative strategy to show log-concavity of $p^*(\xi)$

There is a slight variant of our approach to establishing the log-concavity of $p^*(\xi)$ that is worthy to mention. One can show that for ξ in B, the variance $\mathrm{Var}_{p^*}[v(w)|\xi]$ is not more than $\mathrm{Var}_p[v(w)|\xi]/(1-\delta)$, using the fact that

$$\frac{p^*(w|\xi)}{p(w|\xi)} \le \frac{1}{1-\delta}.$$

So we get the strong log-concavity we want for $p^*(\xi)$ by a simpler variance calculation showing that

$$\rho \mathrm{Var}_p\big[v(w)|\xi\big] < (1-\delta)$$

for ξ in B. This can be established by the assumption

$$Kd > \frac{A_{3,\delta}(\beta N)^2}{1-\delta}$$

with a smaller $A_{3,\delta}$ that replaces the 1.4 factor with 1. With small δ, Kd satisfies this property more generally than before. This approach has the advantage that the $Z(w)$ term has been removed for the variance calculation using $p(w|\xi)$, so that there is no need to control $\nabla Z(w)$ in the variance bound. Nevertheless, it remains the log-concavity of $p^*(\xi)$ that is being established in this way, and the $Z(w)$ factor is needed in the formulation of the $p^*(w|\xi)$ from which we sample in the log-concave coupling.

References

[1] Z. Allen-Zhu, Y. Li, and Z. Song, A convergence theory for deep learning via over-parameterization. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 242–252, Proceedings of Machine Learning Research 97, PMLR, 2019

[2] G. A. Baker Jr, Certain general order-disorder models in the limit of long-range interactions. *Phys. Rev.* **126** (1962), no. 6, 2071–2078

[3] D. Bakry and M. Émery, Diffusions hypercontractives. In *Séminaire de probabilités, XIX, 1983/84*, pp. 177–206, Lecture Notes in Math. 1123, Springer, Berlin, 1985 Zbl 0561.60080 MR 0889476

[4] D. Bakry, I. Gentil, and M. Ledoux, *Analysis and geometry of Markov diffusion operators*. Grundlehren Math. Wiss. 348, Springer, Cham, 2014 Zbl 1376.60002 MR 3155209

[5] A. R. Barron, Neural net approximation. In *Proceedings of the 7th Yale Workshop on Adaptive and Learning Systems*, pp. 69–72, Center for Systems Science, Yale University, New Haven, CT, 1992

[6] A. R. Barron, Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Trans. Inform. Theory* **39** (1993), no. 3, 930–945 Zbl 0818.68126 MR 1237720

[7] A. R. Barron, Shannon lecture: Information theory and high-dimensional Bayesian computation. IEEE International Symposium on Information Theory, Athens, Greece, 11 July 2024

[8] A. R. Barron and R. L. Barron, Statistical learning networks: A unifying view. In *Computing Science and Statistics: Proceedings of the 20th Symposium on the Interface*, pp. 192–203, American Statistical Association, Alexandria, VA, 1988

[9] A. R. Barron and J. M. Klusowski, Approximation and estimation for high-dimensional deep learning networks. 2018, arXiv:1809.03090v2

[10] A. R. Barron and J. M. Klusowski, Complexity, statistical risk, and metric entropy of deep nets using total path variation. 2019, arXiv:1902.00800v2

[11] A. R. Barron and C. McDonald, Log concave coupling for sampling from neural net posterior distributions. Statistical Machine Learning for High Dimensional Data, Institute for Mathematical Sciences, Singapore, 13–31 May 2024

[12] R. Bauerschmidt and T. Bodineau, A very simple proof of the LSI for high temperature spin systems. *J. Funct. Anal.* **276** (2019), no. 8, 2582–2588 Zbl 1408.81033 MR 3926125

[13] M. Betancourt, A conceptual introduction to Hamiltonian Monte Carlo. [v1] 2017, [v2] 2018, arXiv:1701.02434

[14] S. G. Bobkov, Isoperimetric and analytic inequalities for log-concave probability measures. *Ann. Probab.* **27** (1999), no. 4, 1903–1921 Zbl 0964.60013 MR 1742893

[15] S. G. Bobkov and M. Ledoux, From Brunn–Minkowski to Brascamp–Lieb and to logarithmic Sobolev inequalities. *Geom. Funct. Anal.* **10** (2000), no. 5, 1028–1052 Zbl 0969.26019 MR 1800062

[16] Y. Cai, J. Wu, S. Mei, M. Lindsey, and P. Bartlett, Large stepsize gradient descent for non-homogeneous two-layer networks: Margin improvement and fast optimization. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, pp. 71306–71351, Curran Associates, Red Hook, NY, 2024

[17] P. Cattiaux and A. Guillin, Functional inequalities for perturbed measures with applications to log-concave measures and to some Bayesian problems. *Bernoulli* **28** (2022), no. 4, 2294–2321 Zbl 07594060 MR 4474544

[18] T. Charnock, L. Perreault-Levasseur, and F. Lanusse, Bayesian neural networks. In *Artificial Intelligence for High Energy Physics*, pp. 663–713, World Scientific, 2022

[19] Y. Chen, S. Chewi, A. Salim, and A. Wibisono, Improved analysis for a proximal algorithm for sampling. In *Proceedings of 35th Conference on Learning Theory*, pp. 2984–3014, Proceedings of Machine Learning Research 178, PMLR, 2022

[20] Y. Chen and R. Eldan, Localization schemes: A framework for proving mixing bounds for Markov chains. In *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science*, pp. 110–122, IEEE Computer Society, Los Alamitos, CA, 2022 Zbl 08079907 MR 4537195

[21] L. Chizat, E. Oyallon, and F. Bach, On lazy training in differentiable programming. In *Advances in Neural Information Processing Systems 32*, Curran Associates, Red Hook, NY, 2019

[22] T. M. Cover, Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *EEE Trans. Electron. Comput.* **EC-14** (1965), no. 3, 326–334 Zbl 0152.18206

[23] T. M. Cover, Capacity problems for linear machines. In *Pattern Recognition*, pp. 283–289, Thompson Book Company, Washington DC, 1968

[24] S. S. Dey, G. Wang, and Y. Xie, Approximation algorithms for training one-node ReLU neural networks. *IEEE Trans. Signal Process.* **68** (2020), 6696–6706 Zbl 1543.68324 MR 4188568

[25] S. Du, J. Lee, H. Li, L. Wang, and X. Zhai, Gradient descent finds global minima of deep neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 1675–1685, Proceedings of Machine Learning Research 97, PMLR, 2019

[26] R. Dwivedi, Y. Chen, M. J. Wainwright, and B. Yu, Log-concave sampling: Metropolis–Hastings algorithms are fast. *J. Mach. Learn. Res.* **20** (2019), article no. 183 Zbl 1440.62039 MR 4048994

[27] A. El Alaoui, A. Montanari, and M. Sellke, Sampling from the Sherrington–Kirkpatrick Gibbs measure via algorithmic stochastic localization. In *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science*, pp. 323–334, IEEE Computer Society, Los Alamitos, CA, 2022 Zbl 08079926 MR 4537214

[28] R. Eldan, Taming correlations through entropy-efficient measure decompositions with applications to mean-field approximation. *Probab. Theory Related Fields* **176** (2020), no. 3-4, 737–755 Zbl 1436.82016 MR 4087482

[29] R. Eldan and R. Gross, Decomposition of mean-field Gibbs distributions into product measures. *Electron. J. Probab.* **23** (2018), article no. 35 Zbl 1390.60346 MR 3798245

[30] R. Eldan, F. Koehler, and O. Zeitouni, A spectral condition for spectral gap: Fast mixing in high-temperature Ising models. *Probab. Theory Related Fields* **182** (2022), no. 3-4, 1035–1051 Zbl 1515.82046 MR 4408509

[31] S. Frei, G. Vardi, P. Bartlett, and N. Srebro, The double-edged sword of implicit bias: Generalization vs. robustness in ReLU networks. In *Advances in Neural Information Processing Systems 36*, pp. 8885–8897, Curran Associates, Red Hook, NY, 2023

[32] J. H. Friedman and W. Stuetzle, Projection pursuit regression. *J. Amer. Statist. Assoc.* **76** (1981), no. 376, 817–823 MR 0650892

[33] V. Froese and C. Hertrich, Training neural networks is NP-hard in fixed dimension. In *Advances in Neural Information Processing Systems 36*, pp. 44039–44049, Curran Associates, Red Hook, NY, 2023

[34] V. Gallego and D. Ríos Insua, Current advances in neural networks. *Annu. Rev. Stat. Appl.* **9** (2022), 197–222 MR 4394906

[35] S. Goel, A. Klivans, P. Manurangsi, and D. Reichman, Tight hardness results for training depth-2 ReLU networks. In *12th Innovations in Theoretical Computer Science Conference*, article no. 22, Leibniz Int. Proc. Inform. 185, Schloss Dagstuhl. Leibniz-Zent. Inform., Wadern, 2021 Zbl 08186810 MR 4214266

[36] R. D. Gordon, Values of Mills' ratio of area to bounding ordinate and of the normal probability integral for large values of the argument. *Ann. Math. Statistics* **12** (1941), 364–366 Zbl 67.0467.03 MR 0005558

[37] B. Hanin and A. Zlokapa, Bayesian inference with deep weakly nonlinear networks. 2024, arXiv:2405.16630

[38] T. Hastie and R. Tibshirani, Generalized additive models. *Statist. Sci.* **1** (1986), no. 3, 297–310 Zbl 0645.62068 MR 0858512

[39] R. Holley and D. Stroock, Logarithmic Sobolev inequalities and stochastic Ising models. *J. Statist. Phys.* **46** (1987), no. 5-6, 1159–1194 Zbl 0682.60109 MR 0893137

[40] J. Hron, R. Novak, J. Pennington, and J. Sohl-Dickstein, Wide Bayesian neural networks have a simple weight posterior: Theory and accelerated sampling. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 8926–8945, Proceedings of Machine Learning Research 162, PMLR, 2022

[41] X. Huang, D. Zou, Y.-A. Ma, H. Dong, and T. Zhang, Faster sampling via stochastic gradient proximal sampler. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 20559–20596, Proceedings of Machine Learning Research 235, PMLR, 2024

[42] J. Hubbard, Critical behaviour of the Ising model. *Phys. Lett.* **39A** (1972), 365–367 MR 0329526

[43] P. J. Huber, Projection pursuit. *Ann. Statist.* **13** (1985), no. 2, 435–475 Zbl 0595.62059 MR 0790553

[44] A. Jacot, F. Gabriel, and C. Hongler, Neural tangent kernel: Convergence and generalization in neural networks. *Advances in Neural Information Processing Systems 31*, Curran Associates, Red Hook, NY, 2018

[45] J. M. Klusowski and A. R. Barron, Approximation by combinations of ReLU and squared ReLU ridge functions with ℓ^1 and ℓ^0 controls. *IEEE Trans. Inform. Theory* **64** (2018), no. 12, 7649–7656 Zbl 1432.41003 MR 3883687

[46] Y. Kook, Y.-T. Lee, R. Shen, and S. Vempala, Sampling with Riemannian Hamiltonian Monte Carlo in a constrained space. *Advances in Neural Information Processing Systems 35*, pp. 31684–31696, Curran Associates, Red Hook, NY, 2022

[47] Y. Kook and S. S. Vempala, Gaussian cooling and Dikin walks: The interior-point method for logconcave sampling. In *Proceedings of 37th Conference on Learning Theory*, pp. 3137–3240, Proceedings of Machine Learning Research 247, PMLR, 2024

[48] Y. Kook and S. S. Vempala, Sampling and integration of logconcave functions by algorithmic diffusion. In *Proceedings of the 57th Annual ACM Symposium on Theory of Computing*, pp. 924–932, ACM, New York, NY, 2025 MR 4928485

[49] Y. T. Lee, R. Shen, and K. Tian, Structured logconcave sampling with a restricted Gaussian oracle. In *Proceedings of the 34th Conference on Learning Theory*, pp. 2993–3050, Proceedings of Machine Learning Research 134, PMLR, 2021

[50] E. L. Lehmann, J. P. Romano, and G. Casella, *Testing statistical hypotheses*, Springer Texts Stat., Springer, Cham, 2022

[51] C. Liu, L. Zhu, and M. Belkin, Loss landscapes and optimization in over-parameterized non-linear systems and neural networks. *Appl. Comput. Harmon. Anal.* **59** (2022), 85–116 Zbl 1550.65093 MR 4412181

[52] S. Livingstone, M. Betancourt, S. Byrne, and M. Girolami, On the geometric ergodicity of Hamiltonian Monte Carlo. *Bernoulli* **25** (2019), no. 4A, 3109–3138 Zbl 1428.62375 MR 4003576

[53] L. Lovász and S. Vempala, The geometry of logconcave functions and sampling algorithms. *Random Structures Algorithms* **30** (2007), no. 3, 307–358 Zbl 1122.65012 MR 2309621

[54] P. McCullagh and J. A. Nelder, *Generalized linear models*. Monogr. Statist. Appl. Probab., Chapman and Hall, London, 1989 Zbl 0744.62098 MR 3223057

[55] C. McDonald, Computation and estimation for neural networks via log-concave coupling. PhD thesis, Yale University, 2025

[56] C. McDonald and A. R. Barron, Log-concave coupling for sampling neural net posteriors. In *2024 IEEE International Symposium on Information Theory*, pp. 2251–2256, IEEE, 2024

[57] A. Montanari, Sampling, diffusions, and stochastic localization. [v1] 2023, [v2] 2025, arXiv:2305.10690v2

[58] A. Montanari and Y. Wu, Provably efficient posterior sampling for sparse linear regression via measure decomposition. *J. Amer. Statist. Assoc.* **121** (2026), no. 553, 636–654 Zbl 8188598 MR 5061149

[59] R. M. Neal, *Bayesian learning for neural networks*. Lect. Notes Stat. 118, Springer, New York, NY, 1996 Zbl 0888.62021

[60] J. A. Nelder and R. W. Wedderburn, Generalized linear models. *J. R. Statist. Soc. A* **135** (1972), no. 3, 370–384

[61] H. Robbins, A remark on Stirling's formula. *Amer. Math. Monthly* **62** (1955), 26–29 Zbl 0068.05404 MR 0069328

[62] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, Curran Associates, Red Hook, NY, 2021

[63] V. Srinivasan, A. Wibisono, and A. Wilson, Fast sampling from constrained spaces using the Metropolis-adjusted mirror Langevin algorithm. In *Proceedings of the 37th Conference on Learning Theory*, pp. 4593–4635, Proceedings of Machine Learning Research 247, PMLR, 2024

[64] R. H. Swendsen and J.-S. Wang, Nonuniversal critical dynamics in Monte Carlo simulations. *Phys. Rev. Lett.* **58** (1987), no. 2, 86–88

[65] M. A. Tanner and W. H. Wong, The calculation of posterior distributions by data augmentation. *J. Amer. Statist. Assoc.* **82** (1987), no. 398, 528–540 Zbl 0619.62029 MR 0898357

[66] A. Tsigler and P. L. Bartlett, Benign overfitting in ridge regression. *J. Mach. Learn. Res.* **24** (2023), article no. 123 Zbl 1571.68293 MR 4583284

[67] V. H. Vu, On the infeasibility of training neural networks with small mean-squared error. *IEEE Trans. Inform. Theory* **44** (1998), no. 7, 2892–2900 Zbl 0981.68138 MR 1672043

[68] D. Zou, Y. Cao, D. Zhou, and Q. Gu, Gradient descent optimizes over-parameterized deep ReLU networks. *Mach. Learn.* **109** (2020), no. 3, 467–492 Zbl 1494.68245 MR 4075425

Received 29 April 2025; revised 13 April 2026.

Curtis McDonald
Simons Institute for the Theory of Computing, University of California, Berkeley, 121 Calvin Lab, Berkeley, 94720-2190, USA; c.mcdonald.simons@berkeley.edu
Author IDs: zbMATH mcdonald.curtis MR 1398664 ORCID 0000-0001-6298-2861

Andrew R. Barron
Department of Statistics and Data Science, Yale University, P.O. Box 208290, New Haven, 06520-8290, USA; andrew.barron@yale.edu
Author IDs: zbMATH barron.andrew-r MR 222645