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), and to statistical risk guarantees. 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 multimodal 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.

For a discrete uniform prior restricted to a grid, we study the statistical risk (generalization error) of procedures based on the posterior. Using an inverse temperature that is a fractional power of $1/N$ namely $\beta = C[(\log d)/N]^{1/4}$, we demonstrate that notions of squared error are on the 4th root order $O([(\log d)/N]^{1/4})$. If one further assumes independent Gaussian data with a variance σ^2 that matches the inverse temperature, $\beta = 1/\sigma^2$, we show that the Kullback divergence decays as an improved cube root power $O([(\log d)/N]^{1/3})$.

We extend these risk results to the continuous uniform prior as well. With polynomial time sampling algorithms for log-concave target densities, this represents a polynomial time training method for neural networks with statistical risk control.

Contents

2020 Mathematics Subject Classification. Primary 62F15; Secondary 62M45, 65C05.
Keywords. Neural Networks, Bayesian Methods, Sampling, Statistical Learning.

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 $c_k \in \{-\frac{V}{K}, \frac{V}{K}\}$. Thus, $f_w(x)$ is a convex combination of K signed neurons scaled 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.

We are interested in potentially wide networks where K may be large. The study of deep nets (i.e. multi-layered) nets is a separate topic not addressed in this work, as we focus on the single-hidden-layer class.

The approximation ability of these networks has been studied for many years, which we briefly review here. Restrict input vectors $x \in \mathbb{R}^d$ as having bounded entries, $x \in [-1, 1]^d$. The early work of [4] 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 $O(\frac{1}{K})$.

These original results put no restrictions on how large the components of the internal weight vectors w_k can be. To facilitate computation, we wish to work only with weight vectors w_k with bounded ℓ_1 norm, $\|w_k\|_1 \leq 1$. Denote the set of signed neurons with ℓ_1 controlled interior weights as the collection of functions $h : [-1, 1]^d \to \mathbb{R}$

$$\Psi = \{h : h(x) = \pm\psi(w \cdot x), \|w\|_1 \leq 1\}. \tag{1.2}$$

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 [3],[4],[25], networks of the form (1.1) provide accurate approximation for functions f with $\frac{f}{V}$ 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. In [25] a variant of the condition on the Fourier components of f is also given that would allow f to have finite variation V_f and hence to be accurately approximated using convex combinations of elements of Ψ, with bounded ℓ_1 norm on the weights. For target functions f of this form and any probability distribution P_X on $[-1, 1]^d$, using a squared ReLU activation function, there exists a network f_{w^*} of the form (1.1) with added constant and linear terms, with K neurons with ℓ_1 controlled internal weights such that [25]

$$\|f_{w^*} - f\|^2 \leq \frac{V_f^2}{K}, \tag{1.3}$$

where $\|\cdot\|^2$ is the $L_2(P_X)$ norm.

The approximation with bound (1.3) is an existence result, a useful ingredient in neural net analysis. Yet, by itself, it does not imply anything about the estimation ability of training algorithms based on a finite set of N data points $(x_i, y_i)_{i=1}^N$ independently and identically distributed (iid) from a data distribution $P_{X,Y}$. Currently, the best known results show that for a bounded target function $|f| \leq b$, finding the set of neuron parameters that minimize the empirical squared error,

$$\hat{w} = \text{argmin}_{\|w_k\|_1 \leq 1, k \in \{1,\cdots K\}} \sum_{i=1}^{N} (y_i - f_w(x_i))^2, \tag{1.4}$$

with a network width $K = O([N/\log(d)]^{1/2})$ yields a statistical risk control of the order [7]

$$E[\|f_{\hat{w}} - f\|^2] = O(\Big(\frac{\log(d)}{N}\Big)^{\frac{1}{2}}), \tag{1.5}$$

provided there is sub-Gaussian control of the distribution of the response Y. The expectation here is with respect to the training data, with the norm square provides the expectation for the loss at an independent new input and response pair. Analogous deep net conclusions are also in [6], [7].

There has been much research to understand theoretically the optimization of neural networks via gradient based methods [10, 16, 24, 29, 38]. These approaches work by comparing the network to a certain infinite width limit under initialization and scaling assumptions (called the neural tangent kernel, NTK) where the network becomes linear around its initialization point. They also utilize a scaling of $1/\sqrt{K}$ on their outer

weights rather than the $1/K$ scaling we use. A consequence of this NTK regime is the internal weights w_k become random objects not dependent on the data, and then training the outer weights (which we call c_k) is a linear regression problem. We instead fix the outer weights c_k and explicitly train the interior weights w_k to fit the data.

When choosing network size for favorable statistical risk, we prefer to work with $K < N$. Indeed, our later results will show $K = O[(N/\log(d))^{1/4}]$ is an appropriate size for statistical risk control. Then, 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, 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. Nevertheless, for statistical risk analysis of the Bayes estimator, we retain the commonly adopted frequentist statistical learning framework.

Bayesian neural networks have been studied for many years [11, 18, 34], 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 [21, 22]. These approaches require $K/N \to \infty$ to achieve that simplification of the posterior density. The bounded K/N setting is shown in [21, 22] to be distinct with potentially more flexible non-Gaussian process behavior. Indeed, such flexibility arises in our model where the internal weights adapted by the posterior.

We quantify when sampling can be accomplished in polynomial time using MCMC, as well as statistical risk guarantees for the resulting posterior distribution. We will not achieve the optimal square root rate with our Bayesian methods, in fact we will get a fourth root power in the most general case, but we give up some of the accuracy for the sake of computational ability. That is, we adopt a sampling problem we can solve instead of an optimization problem we cannot.

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 a sequence of posterior densities trained on subsets of the data $x^n, y^n \equiv (x_i, y_i)_{i=1}^n$ for every $n \le N$ by

$$p_n(w|x^n, y^n) \propto p_0(w) e^{-\frac{\beta}{2}\sum_{i=1}^n (y_i - f(w,x_i))^2}, \tag{1.6}$$

and the associated posterior mean at a given x value

$$\mu_n(x) = E_{P_n}[f(x, w)|x^n, y^n]. \tag{1.7}$$

Define the posterior predictive density as

$$p_n(y|x, x^n, y^n) = E_{P_n}[\frac{e^{-\frac{\beta}{2}(y-f(x,w))^2}}{\sqrt{2\pi\frac{1}{\beta}}}|x^n, y^n]. \tag{1.8}$$

Define the Cesàro average posterior as the average of the different posteriors,

$$q^{\text{avg}}(w|x^N, y^N) = \frac{1}{N+1}\sum_{n=0}^{N} p_n(w|x^n, y^n). \tag{1.9}$$

Also define the Cesàro mean estimator and Cesàro average predictive density

$$\hat{g}(x) = \frac{1}{N+1}\sum_{n=0}^{N} \mu_n(x), \tag{1.10}$$

and

$$q^{\text{avg}}(y|x, x^N, y^N) = \frac{1}{N+1}\sum_{n=0}^{N} p_n(y|x, x^n, y^n). \tag{1.11}$$

The estimation ability of these posterior densities is measured by their performance according to a choice of risk control. We consider two classes of risk control: arbitrary sequence regret and predictive risk control for iid data.

For arbitrary sequence regret, let $(x_i, y_i)_{i=1}^N$ be an arbitrary sequence of inputs and response values with no assumption on the underlying data relationship between x_i and y_i. Consider g as an arbitrary competitor function we wish to measure our Bayesian posteriors against. The average squared error regret is defined as

$$R_N^{\text{square}} = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{2}\left[(y_n - \mu_{n-1}(x_n))^2 - (y_n - g(x_n))^2\right]. \tag{1.12}$$

We demonstrate bounds on this regret of the order $O([(\log d)/N]^{1/4})$ for a discrete uniform prior. This bound requires control on the maximum magnitude of the observations in the sequence $\max_{n\leq N}|y_n|$, as well as a bounded function g. This is not as good as the square root bound of optimization, but the posterior means can be computed by sampling, as we demonstrate. As such, for any competitor g in the class of neural networks, including the optimal fit for the data sequence, the sequence of posterior densities achieves average performance arbitrarily close to the competitor if the data size N is sufficiently large relative to $\log d$.

Another form of risk control relies on further assumptions about the incoming data. Assume $(x_i, y_i)_{i=1}^N$ come iid from a data distribution $P_{X,Y}$ with the conditional mean of Y being a function $E[Y|X] = g(X)$ and having conditional variance bound σ^2. Let $X_{N+1} = X$ be a new data point independently drawn from the same input data

distribution. We define our statistical loss function (squared generalization error) as half the squared error averaged with respect to the distribution of the new X,

$$\|g - \hat{g}\|^2 = \int \frac{1}{2}(g(x) - \hat{g}(x))^2 P_X(dx). \tag{1.13}$$

The corresponding notion of statistical risk is half the expected squared error with the expectation taken with respect to the training data and the new observation,

$$E[\|g - \hat{g}\|^2]. \tag{1.14}$$

We demonstrate mean squared risk control of the order $O([(\log d)/N]^{1/4})$ for a discrete uniform prior. This bound requires no moment control on Y higher than the variance. These regret and risk bounds require a gain β of the order $O([\log(d)/N]^{1/4})$, which is atypical to most Bayesian posterior problems where the β would not be a value decaying in N but rather a fixed constant. However, in this formulation we do not have to match the β to the inverse variance $1/\sigma^2$ and we still have the fourth root risk bound.

If we further assume the data is iid Gaussian with $Y|X = x$ having the Normal$(g(x), \sigma^2)$ distribution, and the reciprocal variance of the data matches our gain $\beta = 1/\sigma^2$, we can give bounds on Kullback divergence. For the Kullback divergence between the Cesàro predictive density and the data generating density, we demonstrate a bound of $O([(\log d)/N]^{1/3})$ for the discrete uniform prior.

Our statistical risk analysis is first presented for a discrete uniform prior on an ℓ_1 controlled grid. This allows explicit control on the number of points in the support of the distribution, and control of the minimum probability of a single point. Using a coupled Dirichlet-Multinomial distribution to link the continuous and discrete prior cases, we are able to extend some of the discrete prior risk results to the desired continuous prior setting. When the target function has variation not more than the specified V with respect to the dictionary of neurons (i.e. f/V lives in the closure of the convex hull), the continuous prior inherits risk control of the same order as the discrete prior for iid data.

The barrier to implementing the Bayesian approaches defined above is being able to sample from the densities $p_n(w|x^n, y^n)$ and thus compute posterior averages $\mu_n(x)$ as well as predictive probabilities $p_n(y|x, x^n, y^n)$ which are defined by expectations with respect to $p_n(w|x^n, y^n)$. The densities $p_n(w|x^n, y^n)$ will be high-dimensional and multi-modal densities with no immediately inherent 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. However, an MCMC method is only useful if 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 [1, 2, 15, 31]. 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 densities $p_n(w|x^n, y^n)$ can be re-written as a mixture density (also called a measure decomposition in the language of [33])

$$p_n(w|x^n, y^n) = \int p_n(w|\xi, x^n, y^n) p_n(\xi|x^n, y^n) d\xi, \tag{1.15}$$

using a reverse conditional density $p_n(w|\xi) = p_n(w|\xi, x^n, y^n)$ and an induced marginal density $p_n(\xi) = p_n(\xi|x^n, y^n)$. 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 $p_n(w, \xi)$ that preserves the target marginal $p_n(w)$ and has a log-concave marginal distribution $p_n(\xi)$ and a log-concave conditional distribution $p_n(w|\xi)$ 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 terms when the total number of parameters Kd is large enough such that

$$Kd \geq C(\beta N)^2, \tag{1.16}$$

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 [15, 27, 30, 31]. We leave the specific of this algorithm in our setting (e.g. Metropolis Adjusted Langevin Diffusion (MALA), Hamiltonian Monte Carlo, tuning of parameters, etc.) 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. With access to polynomial time sampling algorithms for log-concave densities, using the continuous uniform prior on the ℓ_1 ball, and appropriately scaled choices of β, K, d and N, this represents a polynomial time training algorithm for single-hidden-layer neural networks with statistical risk control of the order $O([\log(d)/N]^{1/4})$.

The remainder of the paper is organized as follows. In Section 2 and 3 we lay out the specifics of our Bayesian model. In Section 4 we summarize the main conclusions of the paper. Details of the log-concave coupling are given in Section 5. Statistical risk control is discussed in Sections 6, 7. Further discussion with existing literature is given in Section 8, and conclusions in Section 9.

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.
- $a^{\mathrm{T}}, \mathbf{X}^{\mathrm{T}}$ refers to the transpose of a vector or matrix, so quadratic forms of a vector a with the matrix $\mathbf{X}$ will be written as $a^{\mathrm{T}}\mathbf{X}a$.
- $\|w\|_p$ refers to the ℓ_p norm, $\|w\|_p = (\sum_j (w_j)^p)^{\frac{1}{p}}$.
- 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, superscripts indicate the set of variables $X^n = (X_i)_{i=1}^n$.
- For a data sequence $(x_i, y_i)_{i=1}^N$, given a function f associate it with the vector with coordinates equal to the function outputs $f_i = f(x_i)$. For any two vectors of length N define the empirical squared norm and inner product

$$\|h_1 - h_2\|_N^2 = \sum_{i=1}^N (h_{1,i} - h_{2,i})^2 \qquad \langle h_1, h_2 \rangle_N = \sum_{i=1}^N h_{1,i} h_{2,i}$$

- Logarithms in the paper are natural logarithms.

3. Bayesian Model

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 equation (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 $c_k \in \{-\frac{V}{K}, \frac{V}{K}\}$. 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 $\frac{V}{K}$. For non-symmetric activation functions, we can use twice the variation $\tilde{V} = 2V$ and use twice the number of neurons $\tilde{K} = 2K$. For the first K neurons set $c_k = \frac{V}{K}$ and for the second set of K neurons set $c_k = -\frac{V}{K}$. 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 and have weight vector 0. In either case, we consider the outer weights c_k as being 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). We will discuss a couple choices of prior shortly. For each index $i \in \{1, \ldots, N\}$ define the residual of a neural network as

$$\text{res}_i(w) = y_i - \sum_{k=1}^{K} c_k \psi(w_k \cdot x_i). \tag{3.1}$$

For any subset of the data $n \leq N$, define the n-fold loss function as half the sum of squares of the first n residuals

$$\ell_n(w) = \frac{1}{2} \sum_{i=1}^{n} (\text{res}_i(w))^2. \tag{3.2}$$

For any gain parameter $\beta > 0$, we define the n-th posterior density (with respect to η) and the associated posterior mean

$$p_n(w) = \frac{p_0(w) e^{-\beta \ell_n(w)}}{\int e^{-\beta \ell_n(w)} p_0(w) \eta(dw)} \tag{3.3}$$

$$\mu_n(x) = E_{P_n}[f(w, x)], \tag{3.4}$$

where $E_{P_n}[\cdot]$ denotes expectation with respect to the indicated distribution. Note our posterior densities $p_n(w)$ are defined by the data points x^n, y^n we condition on, so we

will also denote them as $p_n(w|x^n, y^n)$. For a given weight vector w, define the predictive density $p(y|x, w)$ to be Normal$(f(x, w), \frac{1}{\beta})$. Define the n-th posterior predictive density as

$$p_n(y|x) = E_{P_n}[p(y|x, w)]. \tag{3.5}$$

Note that these predictive densities are also conditioned on the x^n, y^n data that define the posterior. Define also the Cesàro average posterior, mean, and predictive density as in equations (1.9), (1.10), and (1.11).

3.1. Choice of Prior

We consider two priors in the paper. The first prior we will consider is 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

$$p_0(w) = \prod_{k=1}^{K} \Big(1\{\|w_k\|_1 \le 1\} \frac{1}{\mathrm{Vol}(S_1^d)} \Big). \tag{3.6}$$

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

We will also 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 $\frac{1}{M}$. Define this set as $S_{1,M}^d$,

$$S_{1,M}^d = \{w : Mw \in \{-M, M\}^d, \ \|w\|_1 \le 1\}, \tag{3.7}$$

That is, each coordinate $w_{k,j}$ can only be integer multiples of the grid size $\frac{1}{M}$ 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 $S_{1,M}^d$. This has probability mass function

$$p_0(w) = \prod_{k=1}^{K} \Big(1\{w_k \in S_{1,M}^d\} \frac{1}{|S_{1,M}^d|} \Big). \tag{3.8}$$

with respect to counting measure in $(S_{1,M}^d)^K$. When d is large one may choose a smaller order M to arrange sparsity in the weight vector, as at most M of the d coordinates can be non-zero. Furthermore, we have a bound on the cardinality of the support set $|S_M^d| < (2d + 1)^M$ which will prove useful in future statistical risk analysis. Most notably, $\log |S_M^d|$ only grows logarithmically in the dimension d of the weight vectors.

We can also consider both of these priors as specific marginals of a joint coupled Dirichlet and Multinomial distribution. We arrange a continuous vector $w^{\mathrm{cont}} \in (S_1^d)^K$

and a discrete vector $w^{\text{disc}} \in (S^d_{1,M})^K$. Say the signs of each coordinate $w^{\text{cont}}_{k,j}$ are distributed independent Rademacher. Then, for each index k, the absolute values vectors $(|w^{\text{cont}}_k|, |w^{\text{disc}}_k|)$ are independent and distributed as follows. $|w^{\text{cont}}_k|$ is uniform on the $d+1$ dimensional simplex, which is symmetric Dirichlet using the all 1's parameter vector. Then $|w^{\text{disc}}_k|$ conditioned on $|w^{\text{cont}}_k|$ is distributed as $1/M$ times a Multinomial$(M, |w^{\text{cont}}|)$ distribution. This results in w^{cont} and w^{disc} being marginally uniform on $(S^d_1)^K$ and $(S^d_{1,M})^K$ respectively, but being coupled via this joint distribution.

The continuous prior will be used to prove the log-concave coupling form of our target density, but the finite size of the support of the discrete prior will prove useful for statistical risk control. In the paper, we first prove statistical risk control of the discrete prior, and then extend this to the continuous prior using this joint Dirichlet and Multinomial construction.

4. Summary of Main Results

Our results are two-fold; demonstration of a log-concave mixture form using a continuous uniform prior, and risk control for the discrete uniform prior. We then extend the risk control to the continuous prior as well.

4.1. Log-Concave Coupling

The log-concave coupling result is as follows:

Theorem 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*

$$C_N = \max_{n \in \{1,N\}} |y_n| + a_0 V \tag{4.1}$$

$$A_1 = 2a_1 + 4\sqrt{\frac{3}{2}} a_2 \tag{4.2}$$

$$A_2 = \left(1 + \frac{1}{\sqrt{\pi}}\right)\sqrt{2a_2\sqrt{\frac{3}{2}}} \tag{4.3}$$

$$A_3 = 4\sqrt{\frac{3}{2e}} a_2 (C_N V)^{\frac{3}{2}} [A_1 + A_2 (C_N V)^{\frac{1}{2}}]. \tag{4.4}$$

Define a value

$$\delta = \min\left(\frac{1}{300}, \sqrt{\frac{2\pi}{11}} \frac{K}{a_2 \beta C_N V}\right). \tag{4.5}$$

Let d and K satisfy

$$K[\log(2Kd/\delta)] \leq \beta N \tag{4.6}$$

and

$$Kd \geq A_3(\beta N)^2. \tag{4.7}$$

Using a continuous uniform prior on $(S_1^d)^K$, for each $n \leq N$ the posterior distribution $p_n(w)$ can be written as a mixture distribution with an auxiliary random variable ξ,

$$p_n(w) = \int p_n(w|\xi)p_n(\xi)d\xi, \tag{4.8}$$

where $p_n(w|\xi)$ is a log-concave density for each ξ, and $p_n(\xi)$ is a log-concave density. If equation (4.7) is a strict inequality, $p(\xi)$ is strictly log-concave.

Further details on the proof of this theorem and choice of the auxiliary random variable ξ are presented in Section 5.

4.2. Statistical Risk Control for the Discrete Prior

Consider $(x_i, y_i)_{i=1}^N$ as an arbitrary sequence of inputs and response values. Let g be a competitor function to which we want to compare our performance. The individual squared error regret is defined as

$$r_n^{\text{square}} = \frac{1}{2}\big[(y_n - \mu_{n-1}(x_n))^2 - (y_n - g(x_n))^2\big], \tag{4.9}$$

and average squared regret is defined as

$$R_N^{\text{square}} = \frac{1}{N}\sum_{n=1}^{N} r_n^{\text{square}}. \tag{4.10}$$

Theorem 2. *Let g be a target function with absolute value bounded by b and let $\tilde{g}$ be its projection into the closure of the convex hull of signed neurons scaled by V. Let P_0 be the uniform prior on the discrete set $(S_{1,M}^d)^K$. Assume the neuron activation function is odd symmetric and set all outer weights as $c_k = \frac{V}{K}$. For any data sequence $(x_i, y_i)_{i=1}^N$ with all $x_i \in [-1, 1]^d$, define*

$$\epsilon_n = y_n - g(x_n) \qquad \tilde{\epsilon}_n = y_n - \tilde{g}(x_n). \tag{4.11}$$

Then average squared regret is upper bound by

$$R_N^{square} \leq \frac{MK\log(2d+1)}{\beta N} + \frac{a_0^2V^2}{2K} + \frac{(VC_Na_2 + V^2a_1^2)}{2M} \tag{4.12}$$

$$+ 2\beta\Big(\frac{a_0V+b}{2}C_N + \Big(\frac{a_0V+b}{2}\Big)^2\Big)^2 + \frac{1}{2}\frac{1}{N}\sum_{n=1}^{N}(\tilde{\epsilon}_n^2 - \epsilon_n^2). \tag{4.13}$$

In particular, if g lives in the closure of the convex hull of signed neurons scaled by V we may set $g = \tilde{g}$ *and* $\tilde{\epsilon} = \epsilon$*. With specific choice of* β, M, K*, we can achieve an upper bound*

$$R_N^{square} \leq 4\Big(a_0V(\frac{a_0V+b}{2})\Big)^{\frac{1}{2}}\Big((C_N + \frac{a_0V+b}{2})^2(\frac{a_2VC_N + a_1^2V^2}{2})\Big)^{\frac{1}{4}}\Big(\frac{\log(2d+1)}{N}\Big)^{\frac{1}{4}}. \tag{4.14}$$

Further details on the proof of this result can be found in Section 6.3. This theorem places no further assumptions on the data sequence. For Theorem 2 all that is needed of the g and the network functions are the vectors in $\mathbb{R}^N$ of the function evaluated at the specified inputs values $x_1, ..., x_N$. Then the convex hull is a subset of $\mathbb{R}^N$ and its closure and the projection $\tilde{g}$ is taken in the Euclidean sense. Note any network could be used in place of $\tilde{g}$ here, but the projection is by definition the minimizer of the euclidean distance. In contrast, for control of risk of more general points in $[-1,1]^d$, we treat the networks and comparator g as functions and use the $L_2(P_X)$ projection for $\tilde{g}$.

With more specific assumptions, we can have bounds on the risk of generalization. Suppose (X_i, Y_i) are iid from a distribution with Y having conditional mean $g(X)$ and conditional variance bounded by σ^2. Then we can recover the arbitrary sequence bounds for mean squared risk.

Theorem 3. *Let g be a target function with absolute value bounded by b and let* $\tilde{g}$ *be its* $L_2(P_X)$ *projection into the closure of the convex hull of signed neurons scaled by V. Let* P_0 *be the uniform prior on the discrete set* $(S_{1,M}^d)^K$*. Assume the neuron activation function is odd symmetric and set all outer weights as* $c_k = \frac{V}{K}$*. Let* $(X_i, Y_i)_{i=1}^N$ *be training data iid with conditional mean* $g(X_i)$ *and conditional variance* $\sigma_{X_i}^2$*, with variance bounded by* $\max_{x\in[-1,1]^d}\sigma_x^2 \leq \sigma^2$*. Assume the support of the data distribution* P_X *is* $[-1,1]^d$*. Then the statistical risk of the Cesàro mean* $\hat{g}$ *as an estimator of g is upper bounded as,*

$$E[\|g - \hat{g}\|^2] \leq \frac{MK\log(2d+1)}{\beta(N+1)} + \frac{a_0^2V^2}{2K} + \frac{(V(a_0V+b)a_2 + V^2a_1^2)}{2M} \tag{4.15}$$

$$+ 2\beta(\frac{a_0V+b}{2})^2(\sigma + \frac{a_0V+b}{2})^2 + \|g - \tilde{g}\|^2. \tag{4.16}$$

With specific choice of β, M, K we can have a bound on the mean squared risk of the form

$$4\Big(a_0V(\frac{a_0V+b}{2})(\sigma+\frac{a_0V+b}{2})\Big)^{\frac{1}{2}}\Big(\frac{V(a_0V+b)a_2+V^2a_1^2}{2}\Big)^{\frac{1}{4}}\Big(\frac{\log(2d+1)}{N}\Big)^{\frac{1}{4}} \tag{4.17}$$

$$+\|g-\tilde{g}\|^2. \tag{4.18}$$

Further assume that the data is normally distributed with constant variance, that is $p(y_i|x_i)$ is the Normal$(g(x_i),\sigma^2)$ density (i.e. the typical independent Gaussian errors model), and that the gain of the Bayesian model β matches the inverse variance, $\beta = 1/\sigma^2$. Then we can have a bound on Kullback divergence.

Theorem 4. *Let g be a target function with absolute value bounded by b and let $\tilde{g}$ be its projection into the closure of the convex hull of signed neurons scaled by V. Let P_0 be the uniform prior on the discrete set $(S^d_{1,M})^K$. Assume the neuron activation function is odd symmetric and set all outer weights as $c_k = \frac{V}{K}$. Assuming the data distribution has $Y|X \sim$ Normal$(g(X),\frac{1}{\beta})$ we bound the Kullback divergence as,*

$$E[D(P_{Y|X}\|Q^{avg}_{Y|X,X^N,Y^N})] \leq \frac{MK\log(2d+1)}{N+1}+\beta\frac{a_0^2V^2}{2K}+\beta\frac{V(a_0V+b)a_2+V^2a_1^2}{2M} \tag{4.19}$$

$$+\beta\|g-\tilde{g}\|^2. \tag{4.20}$$

With specific choice of M and K, we can achieve a bound of the form,

$$3(\frac{\beta}{2})^{\frac{2}{3}}(a_0V)^{\frac{2}{3}}(V(a_0V+b)a_2+V^2a_1^2)^{\frac{1}{3}}\Big(\frac{\log(2d+1)}{N+1}\Big)^{\frac{1}{3}}+\beta\|g-\tilde{g}\|^2. \tag{4.21}$$

Further details such as specific constants and proofs for the risk results can be found in Section 6.4. Note $\beta\|g-\tilde{g}\|^2$ is the expected Kullback divergence between normals with mean function g and $\tilde{g}$ and variance $1/\beta$. So the theorem bounds the additional Kullback risk beyond this value.

4.3. Statistical Risk Control for the Continuous Prior

Finally, we can extend the risk control of the discrete prior to the continuous prior, but pay a price of twice the risk. If the term $\|g-\tilde{g}\|^2$ is not too large, then we do not mind paying twice its cost in our final bound. For target functions g with well controlled Fourier components, the previously discussed representation result shows that (adjusted by a constant and linear term), it is in the convex hull of $V\Psi$ for suitable variation V, and hence $\|g-\tilde{g}\|^2$ is zero.

Theorem 5. *Let g be a target function with absolute value bounded by b and let $\tilde{g}$ be its projection into the closure of the convex hull of signed neurons scaled by V. Let P_0 be the uniform prior on the continuous set $(S_1^d)^K$. Assume the neuron activation function is odd symmetric and set all outer weights as $c_k = \frac{V}{K}$. Let $(x_i, y_i)_{i=1}^N$ be training data iid with conditional mean $g(x_i)$ and conditional variance $\sigma_{x_i}^2$ with variance bound $\sigma_{x_i}^2 \leq \sigma^2$. Assume the data distribution P_X has support $[-1,1]^d$. Then statistical risk is upper bound by*

$$E[\|g - \hat{g}\|^2] \leq 2\frac{MK\log(2d+1)}{\beta(N+1)} + \frac{a_0^2V^2}{K} + \frac{3a_2V(a_0V+b) + 2V^2a_1^2}{M} \tag{4.22}$$

$$+ 4\beta(\frac{a_0V+b}{2})^2(\sigma + \frac{a_0V+b}{2})^2 + 2\|g-\tilde{g}\|^2 \tag{4.23}$$

$$+ O(\frac{1}{MK}) \tag{4.24}$$

Note with proper choice of parameters M, K, β this can be shown to be of the order

$$E[\|g - \hat{g}\|^2] \leq 2\|g-\tilde{g}\|^2 + O(\Big(\frac{\log(d)}{N}\Big)^{\frac{1}{4}})$$

Further details on this result are found in Section 7, making use of the joint Dirichlet and Multinomial form of the prior.

Remark 1. *Note these results are stated for odd-symmetric neurons (e.g. sigmoids), similar results for non-odd symmetric neurons (e.g. squared ReLU) can be derived with factors of 2 in some of the constants, but the order of dependence in d and N is the same. The signs of the outer weights c_k must also be handled more specifically. Further discussion on symmetric vs non-symmetric neurons is found in Section 6.*

5. Posterior Densities and Log-Concave Coupling

5.1. Posterior Density

Consider the log-likelihood of the posterior densities $p_n(w)$ as defined in equation (3.3), with the continuous uniform prior on $(S_1^d)^K$. The log likelihood and score of the posterior within the constrained set are

$$\log p_n(w) \propto -\beta \ell_n(w) \tag{5.1}$$

$$\nabla_{w_k} \log p_n(w) = \beta \sum_{i=1}^{n} \mathrm{res}_i(w)(c_k\psi'(w_k \cdot x_i)x_i). \tag{5.2}$$

Denote the Hessian as $H_n(w) \equiv \nabla^2 \log p_n(w)$. The density $p_n(w)$ is log-concave if $H_n(w)$ is negative definite for all choices of w. For any vector $a \in \mathbb{R}^{Kd}$, with blocks

$a_k \in \mathbb{R}^d$, the quadratic form $a^{\mathrm{T}} H_n(w) a$ can be expressed as

$$-\beta \sum_{i=1}^{n} \Big(\sum_{k=1}^{K} c_k \psi'(w_k \cdot x_i) a_k \cdot x_i \Big)^2 \tag{5.3}$$

$$+\beta \sum_{i=1}^{n} \mathrm{res}_i(w) \sum_{k=1}^{K} c_k \psi''(w_k \cdot x_i)(a_k \cdot x_i)^2. \tag{5.4}$$

It is clear that for any vector a the first line (5.3) is a negative term, but term (5.4) 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 $p_n(w)$ may not be a log-concave density.

The term (5.4) 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 (5.4) to be negative. Define constants

$$C_n = \max_{i \le n} |y_i| + a_0 V \tag{5.5}$$

$$\rho_n = a_2 \frac{\beta C_n V}{K}. \tag{5.6}$$

We will consider our posterior densities one fixed value of n at a time, so we will work with $\rho = \rho_n$ when it is clear we are talking about a fixed n value.

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 $\frac{1}{\rho}$,

$$\xi_{i,k} \sim \mathrm{Normal}(w_k \cdot x_i, \frac{1}{\rho}). \tag{5.7}$$

This then defines a forward conditional density (or coupling)

$$p_n(\xi | w) \propto e^{-\frac{\rho}{2} \sum_{i,k} (\xi_{i,k} - x_i \cdot w_k)^2}, \tag{5.8}$$

and a joint density for w, ξ,

$$p_n(w, \xi) = p_n(w) p_n(\xi|w). \tag{5.9}$$

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

$$p_n(w, \xi) = p_n(\xi) p_n(w|\xi). \tag{5.10}$$

As we will show, this choice of forward coupling provides a negative definite correction to the log likelihood of $p_n(w|\xi)$ relative to $p_n(w)$, resulting in a negative definite reverse conditional density.

5.2. Reverse Conditional Density $p_n(w|\xi)$

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

Theorem 6. *Under the continuous uniform prior and $\xi_{i,k} \sim Normal(x_i \cdot w_k, 1/\rho)$ for the given choice of ρ, the reverse conditional density $p_n(w|\xi)$ is log-concave for the given ξ coupling.*

Proof. The log likelihood for $p_n(w|\xi)$ is given by

$$\log p_n(w|\xi) = -\beta \ell_n(w) + B_n(\xi) \tag{5.11}$$

$$- \sum_{i=1}^{n} \sum_{k=1}^{K} \frac{\rho}{2} (\xi_{i,k} - w_k \cdot x_i)^2, \tag{5.12}$$

for some function $B_n(\xi)$ which does not depend on w and is only required to make the density integrate to 1. The term (5.12) 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 $H_n(w|\xi) \equiv \nabla^2 \log p_n(w|\xi)$. For any vector $a \in \mathbb{R}^{Kd}$, with blocks $a_k \in \mathbb{R}^d$, the quadratic form $a^{\mathrm{T}} H_n(w|\xi) a$ can be expressed as

$$-\beta \sum_{i=1}^{n} \Big(\sum_{k=1}^{K} \psi'(w_k \cdot x_i) a_k \cdot x_i \Big)^2 \tag{5.13}$$

$$+ \sum_{k=1}^{K} \sum_{i=1}^{n} (a_k \cdot x_i)^2 [\beta \mathrm{res}_i(w) c_k \psi''(w_k \cdot x_i) - \rho]. \tag{5.14}$$

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

$$\max_{i,k} (\beta \mathrm{res}_i(w) c_k \psi''(w_k \cdot x_i) - \rho) \leq 0. \tag{5.15}$$

So all the terms in the sum in (5.14) are negative. Thus, the Hessian of the log likelihood of $p_n(w|\xi)$ is negative definite and $p_n(w|\xi)$ is a log-concave density. ∎

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

$$\sum_{k=1}^{K}\sum_{j=1}^{d} w_{k,j}\sum_{i=1}^{n}\xi_{i,k}x_{i,j}. \tag{5.16}$$

We want some control on how large this term can become, so we restrict the allowed support of ξ. We define a slightly larger ρ value than before,

$$\rho_n = \sqrt{\frac{3}{2}}a_2\frac{\beta C_n V}{K}. \tag{5.17}$$

For some positive $\delta \le 1/16$, we also define a constrained set,

$$B = \left\{\xi_{i,k} : \max_{j,k}|\sum_{i=1}^{n} x_{i,j}\xi_{i,k}| \le n + \sqrt{2\log(\frac{2Kd}{\delta})}\sqrt{\frac{n}{\rho}}\right\}. \tag{5.18}$$

We then define our forward conditional distribution for $p_n^*(\xi|w) = p_n(\xi|w, B)$ as the normal distribution restricted to the set B,

$$p_n^*(\xi|w) = p_n(\xi|w, B) = 1_B(\xi)\frac{\prod_{i=1}^{n}\prod_{k=1}^{K} e^{-\frac{\rho}{2}(\xi_{i,k}-x_i\cdot w_k)^2}}{\int_B \prod_{i=1}^{n}\prod_{k=1}^{K} e^{-\frac{\rho}{2}(\xi_{i,k}-x_i\cdot w_k)^2}d\xi}. \tag{5.19}$$

Under this constrained density, the term (5.16) will be bounded for any choice of $\xi \in B$ and $w_k \in [-1, 1]^d$, 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 restricting set B. Denote the log normalizing constant as

$$Z(w) = \log\int_B \prod_{i=1}^{n}\prod_{k=1}^{K} e^{-\frac{\rho}{2}(\xi_{i,k}-x_i\cdot w_k)^2}d\xi. \tag{5.20}$$

An equivalent expression for the forward coupling is then

$$p_n^*(\xi|w) = 1_B(\xi)e^{-\frac{\rho}{2}\sum_{i,k}(\xi_{i,k}-x_i\cdot w_k)^2 - Z(w)}. \tag{5.21}$$

This construction also defines reverse conditional density $p_n^*(w|\xi)$ with respect to reference measure η and induced marginal density $p_n^*(\xi)$ with respect to Lebesgue measure,

$$p_n^*(w|\xi) = \frac{p_n(w)e^{-\frac{\rho}{2}\sum_{i,k}(\xi_{i,k}-x_i\cdot w_k)^2 - Z(w)}}{\int p_n(w)e^{-\frac{\rho}{2}\sum_{i,k}(\xi_{i,k}-x_i\cdot w_k)^2 - Z(w)}\eta(dw)} \tag{5.22}$$

$$p_n^*(\xi) = \frac{1_B(\xi)\int p_n(w)e^{-\frac{\rho}{2}\sum_{i,k}(\xi_{i,k}-x_i\cdot w_k)^2 - Z(w)}\eta(dw)}{\int_B \int p_n(w)e^{-\frac{\rho}{2}\sum_{i,k}(\xi_{i,k}-x_i\cdot w_k)^2 - Z(w)}\eta(dw)d\xi}. \tag{5.23}$$

Note these densities differ from the $p_n(w|\xi)$ and $p_n(\xi)$ defined before without restricting to the set B due to the presence of the $Z(w)$ function. We then show that $p_n^*(w|\xi)$ is a very similar density to $p(w|\xi)$ and also log-concave.

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 1. *For any weight vector w with $\|w_k\|_1 \leq 1$ the set B in (5.18) has probability under $p(w|\xi)$ at least*

$$P(w \in B|\xi) \leq 1 - \frac{\delta}{\sqrt{2\log(2Kd/\delta)}}$$

Proof. See Appendix Section A. ∎

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 2. *For any vector $a \in R^{Kd}$, define the value*

$$\tilde{\sigma}^2 = \frac{\sum_{i=1}^n \sum_{k=1}^K (a_k \cdot x_i)^2}{\rho}. \tag{5.24}$$

For positive values $\delta \leq 1/16$ with $Kd \geq 4$, we then have upper bounds,

$$|a \cdot \nabla Z(w)| \leq \frac{\rho\tilde{\sigma}}{1-\delta}\frac{\delta}{\sqrt{2\pi}} \tag{5.25}$$

$$|a^T(\nabla^2 Z(w))a| \leq \frac{\rho^2\tilde{\sigma}^2}{\sqrt{2\pi}}\frac{\delta}{1-\delta}\left(2\sqrt{2\log(1/\delta)} + \frac{\rho^2\tilde{\sigma}^2}{\sqrt{2\pi}}\frac{\delta}{1-\delta}\right) \tag{5.26}$$

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

Proof. See Appendix Section A. ∎

Thus, with restriction to the set B whose radius is determined by δ, and a slightly larger ρ, we can give a similar result to Theorem 6. Note this result is for $p_n^*(w|\xi)$ which is distinct from $p_n(w|\xi)$ due to the presence of the $Z(w)$ function in the log likelihood.

Theorem 7. *Define the notation*

$$H_1(\delta) = \frac{2}{\sqrt{2\pi}} \frac{\delta}{1-\delta} \sqrt{2 \log \frac{2}{\delta}} \tag{5.27}$$

$$H_2(\delta) = \left(a_2 \frac{\beta C_n V}{K}\right)^2 \frac{1}{2\pi} \frac{\delta^2}{(1-\delta)^2}. \tag{5.28}$$

Assume a small $\delta \leq \frac{1}{16}$ *that satisfies*

$$H_1(\delta) \leq \frac{1}{100} \tag{5.29}$$

$$H_2(\delta) \leq \frac{1}{10}. \tag{5.30}$$

Under the continuous uniform prior, with ξ *restricted to the set B defined by* δ *and* ρ *scaled as in equation* (5.17)*, the reverse conditional density* $p_n^*(w|\xi)$ *is a log-concave density in w for any* ξ *in B.*

Proof. See Appendix Section B. ∎

Corollary 1. *A positive* δ *which satisfies,*

$$\delta \leq \min\left(\frac{1}{300}, \sqrt{\frac{2\pi}{11}} \frac{K}{a_2 \beta C_N V}\right), \tag{5.31}$$

will satisfy conditions (5.29)*,* (5.30)*.*

The pairing of a normal forward coupling to $p_n^*(\xi|w)$ with a target density $p_n(w)$ to produce a reverse conditional $p_n^*(w|\xi)$ which is log-concave is not a new idea. As we will later discuss, the same concept is used in proximal sampling methods and diffusion models. However, in this work we go further in stating that the induced marginal on $p_n^*(\xi)$ is itself log-concave, which we call a log-concave coupling.

5.3. Marginal Density $p_n^*(\xi)$

Lemma 3. *The score and Hessian of the induced marginal density for* $p_n^*(\xi)$ *for* $\xi \in B$ *are expressed as*

$$\partial_{\xi_{i,k}} \log p_n^*(\xi) = -\rho\, \xi_{i,k} + \rho\, x_i \cdot E_{P_n^*}[w_k|\xi] \tag{5.32}$$

$$\partial_{\xi_{i_1,k_1},\xi_{i_2,k_2}} \log p_n^*(\xi) = -\rho 1\{(i_1,k_1) = (i_2,k_2)\} \tag{5.33}$$

$$+ \rho^2 Cov_{P_n^*}[w_{k_1} \cdot x_{i_1}, w_{k_2} \cdot x_{i_2}|\xi]. \tag{5.34}$$

Equivalently in vector form using the n by d data matrix $\boldsymbol{X}$*,*

$$\nabla \log p_n^*(\xi) = \rho\Big(-\xi + E_{P_n^*}\begin{bmatrix} \boldsymbol{X}w_1 \\ \cdots \\ \boldsymbol{X}w_K \end{bmatrix} | \xi \Big) \tag{5.35}$$

$$\nabla^2 \log p_n^*(\xi) = \rho\Big(-I + \rho\, Cov_{P_n^*}\begin{bmatrix} \boldsymbol{X}w_1 \\ \cdots \\ \boldsymbol{X}w_K \end{bmatrix} | \xi \Big). \tag{5.36}$$

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

The log likelihood of the induced marginal for $p_n^*(\xi)$ is equal to the log of the joint density with w integrated out,

$$\log p_n^*(\xi) = \log\Big(\int p_n(w) p_n^*(\xi|w) \eta(dw) \Big). \tag{5.37}$$

Rearranging the log likelihood of 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 (5.20). Denote the function

$$h(w) = -\beta \ell_n(w) - \frac{\rho}{2} \sum_{i=1}^{n} \sum_{k=1}^{K} (w_k \cdot x_i)^2 - Z(w), \tag{5.38}$$

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

$$\log p_n^*(\xi) = -\frac{\rho}{2} \|\xi\|_2^2 \tag{5.39}$$

$$+ \log\Big(\int p_0(w) e^{h(w)} e^{\rho \sum_{i=1}^{n} \sum_{k=1}^{K} \xi_{i,k} w_k \cdot x_i} \eta(dw) \Big) + C, \tag{5.40}$$

for some constant C which makes the density integrate to 1. Note that ξ is restricted to have support only on the set B, so there is an indicator of the set B we do not write in the expression for simplicity.

It is clear the term (5.40) is the cumulant generating function of the random variable $u(w)$ defined by

$$u(w) = \xi \cdot \begin{pmatrix} \mathbf{X}w_1 \\ \cdots \\ \mathbf{X}w_K \end{pmatrix}, \tag{5.41}$$

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. The density being integrated is a tilting of the log likelihood defined by $h(w)$, and this tilted density is the reverse conditional $p_n^*(w|\xi)$. ∎

We highlight two important consequences of this result.

Corollary 2. *The score* $\nabla \log p_n^*(\xi)$ *is expressed implicitly as a linear transformation of the expected value of the log-concave reverse conditional* $p_n^*(w|\xi)$.

Proof. This is a simple consequence of (5.32) or (5.35). ∎

Remark 2. *Therefore, while we do not have an explicit closed form expression for the score of the marginal density, it can be estimated using an MCMC method and thus is readily available for use. In particular, to run an MCMC algorithm such as MALA on the marginal density* $p_n^*(\xi)$*, the score is needed. Any time the score needs to be evaluated, it can be computed via its own MCMC algorithm for* $p_n^*(w|\xi)$ *as needed and then utilized in the sampling algorithm for* ξ *itself.*

Corollary 3. *The density* $p_n^*(\xi)$ *is log-concave if for any unit vector* $a \in \mathbb{R}^{nK}$*, with blocks* $a_k \in \mathbb{R}^n$*, the variance of a particular linear combination of w, namely*

$$v(w) = \sum_{k=1}^{K} a_k^T \boldsymbol{X} w_k, \tag{5.42}$$

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

$$Var_{P_n^*}[v(w)|\xi] \leq 1/\rho. \tag{5.43}$$

Proof. This is a simple consequence of (5.36). ∎

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

5.4. Conditional Covariance Control

The log-likelihood for $p_n^*(w|\xi)$ is the log likelihood 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,

$$\mathrm{Cov}_{P_0}[w] \succ \mathrm{Cov}_{P_n^*}[w|\xi] \quad \forall \xi \in B. \tag{5.44}$$

Under a Gaussian prior, such a statement would follow easily from the Brascamp-Lieb inequality [9, Proposition 2.1]. However, for priors which are log-concave but

not strongly log-concave, such a uniform prior on a convex set, such an approach will not work.

The covariance matrix of the uniform prior on $(S_1^d)^K$ is diagonal with entries $\mathrm{Var}_{P_0}(w_{k,j}) = \frac{d}{(d+1)^2(d+2)} \leq \frac{1}{d^2}$ which follows from properties of the Dirichlet distribution. Thus, under conjecture (5.44) we would expect a bound of the form

$$\rho \mathrm{Var}_{P_n^*}[v(w)|\xi] \leq \sqrt{\frac{3}{2}} a_2 \frac{\beta C_n V}{Kd^2} \sum_{j=1}^{d} \sum_{k=1}^{K} \sum_{i=1}^{n} (a_{i,k} x_{i,j})^2 \tag{5.45}$$

$$\leq \sqrt{\frac{3}{2}} a_2 \frac{\beta C_n V}{Kd} \sum_{k=1}^{K} \|a_k\|_1^2 \tag{5.46}$$

$$\leq \sqrt{\frac{3}{2}} a_2 \frac{\beta n C_n V}{Kd} \sum_{k=1}^{K} \|a_k\|_2^2 \tag{5.47}$$

$$= \sqrt{\frac{3}{2}} a_2 C_n V \frac{\beta n}{Kd} \tag{5.48}$$

$$\leq \sqrt{\frac{3}{2}} a_2 \frac{C_N V \beta N}{Kd}. \tag{5.49}$$

Thus for $Kd > C(\beta N)$ for some constant 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 constant C,

$$Kd \geq C(\beta N)^2 \tag{5.50}$$

results in log-concavity of the marginal density.

Instead of recreating an inequality like (5.44), we must take a different approach to upper bound the variance in any direction. Denote the function,

$$h_\xi^n(w) = -\beta l^n(w) - \sum_{i=1}^{n} \sum_{k=1}^{K} \frac{\rho}{2} (\xi_{i,k} - w_k \cdot x_i)^2 - Z(w). \tag{5.51}$$

Denote the function shifted by its mean under the prior as

$$\tilde{h}_\xi^n(w) = h_\xi^n(w) - E_{P_0}[h_\xi^n(w)]. \tag{5.52}$$

Define its cumulant generating function with respect to the prior as

$$\Gamma_\xi^n(\tau) = \log E_{P_0}[e^{\tau \tilde{h}_\xi^n(w)}]. \tag{5.53}$$

Lemma 4. *For any integer $\ell \geq 1$ and for any vector $a \in \mathbb{R}^{Kd}$ we have the upper bound*

$$Var_{P_n^*}(a \cdot w|\xi) \leq \left(E_{P_0}[(a \cdot w)^{2\ell}]\right)^{\frac{1}{\ell}} e^{\frac{\ell-1}{\ell} \Gamma_\xi^n\left(\frac{\ell}{\ell-1}\right) - \Gamma_\xi^n(1)}. \tag{5.54}$$

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

$$p_n^*(w|\xi) = e^{\tilde{h}_\xi^n(w) - \Gamma_\xi^n(1)} p_0(w). \tag{5.55}$$

We then apply a Hölder's inequality to the integral expression with parameters p and q such that $\frac{1}{p} + \frac{1}{q} = 1$

$$\mathrm{Var}_{P_n^*}(a \cdot w|\xi) \le E_{P_0}[(a \cdot w)^2 e^{\tilde{h}_\xi^n(w) - \Gamma_\xi^n(1)}] \tag{5.56}$$

$$\le \Big(E_{P_0}[(a \cdot w)^{2p}]\Big)^{\frac{1}{p}} \Big(E_{P_0}[e^{q\tilde{h}_\xi^n(w) - q\Gamma_\xi^n(1)}]\Big)^{\frac{1}{q}}. \tag{5.57}$$

Let $p = \ell$ and $q = \frac{\ell}{\ell-1}$. The second term can be written as

$$\Big(\int e^{\frac{\ell}{\ell-1}\tilde{h}_\xi^n(w) - \frac{\ell}{\ell-1}\Gamma_\xi^n(1)} p_0(w)\eta(dw)\Big)^{\frac{\ell-1}{\ell}} = \Big(\int e^{\frac{\ell}{\ell-1}\tilde{h}_\xi^n(w)} p_0(w)\eta(dw)\Big)^{\frac{\ell-1}{\ell}} e^{-\Gamma_\xi^n(1)} \tag{5.58}$$

$$= e^{\frac{\ell-1}{\ell}\Gamma_\xi^n(\frac{\ell}{\ell-1}) - \Gamma_\xi^n(1)}. \tag{5.59}$$

■

We then study the moments of the prior density and the behavior of the $\Gamma_\xi^n(\tau)$ function separately.

Lemma 5. *For any unit vector $a \in \mathbb{R}^{nK}$, with blocks $a_k \in \mathbb{R}^n$,*

$$E_{P_0}[(\sum_{k=1}^K a_k^T \boldsymbol{X} w_k)^{2\ell}]^{\frac{1}{\ell}} \le \frac{4\ell n}{\sqrt{e}\, d}. \tag{5.60}$$

Proof. See Appendix Section C. ■

Lemma 6. *Denote the constants*

$$A_1 = 2a_1 + 4\sqrt{\frac{3}{2}}a_2 \tag{5.61}$$

$$A_2 = \Big(1 + \frac{1}{\sqrt{\pi}}\Big)\sqrt{2a_2\sqrt{\frac{3}{2}}}. \tag{5.62}$$

Assume positive $\delta \le \frac{1}{16}, d \ge 2, K \ge 2$. For any positive integer $\ell \ge 1$ and any ξ from the constrained set B, we have

$$\frac{\ell-1}{\ell}\Gamma_\xi^n(\frac{\ell}{\ell-1}) - \Gamma_\xi^n(1) \le A_1\frac{C_n V\beta n}{\ell} + A_2\frac{\sqrt{C_n V\beta n}}{\ell}\Big(\sqrt{\log(\frac{2Kd}{\delta})}\sqrt{K}\Big). \tag{5.63}$$

Proof. See Appendix Section C. ∎

We summarize the conclusions of Lemmas 4,5,6 as follows. Ignoring certain constant factors, we have an upper bound on the variance in (5.43) for any choice of ℓ,

$$\frac{n\ell}{d}\exp\Big(\frac{\beta n+\sqrt{\beta nK\log(\frac{2Kd}{\delta})}}{\ell}\Big). \tag{5.64}$$

Ignoring for now the integer constraint, the optimal continuous choice of ℓ to minimize the expression is the numerator in the exponent. With this choice of ℓ, we would have bound

$$\frac{\beta n^2+n^{\frac{3}{2}}\sqrt{\beta K\log(\frac{2Kd}{\delta})}}{d}. \tag{5.65}$$

Multiplying this by $\rho\propto\frac{\beta}{K}$ and upper bounding with $n\le N$, we would have the bound

$$\frac{(\beta N)^2}{Kd}\Big(1+\Big[\frac{K\log\left(\frac{2Kd}{\delta}\right)}{\beta N}\Big]^{\frac{1}{2}}\Big). \tag{5.66}$$

If $K\log(2Kd/\delta)\le\beta N$, then we have a $O(\frac{(\beta N)^2}{Kd})$ 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 8. *Assume $\delta\le\frac{1}{16}, d\ge 2, K\ge 2, \beta N\ge 2$. Further assume that*

$$K\log\left(\frac{2Kd}{\delta}\right)\le\beta N, \tag{5.67}$$

which is essentially a condition than K not be too large (that is, K is less than some multiple of βN).

Define A_1, A_2 as in (5.61), (5.62) and define the constant

$$A_3=4\sqrt{\frac{3}{2e}}a_2(C_NV)^{\frac{3}{2}}[A_1+A_2(C_NV)^{\frac{1}{2}}]. \tag{5.68}$$

Let d and K satisfy

$$Kd\ge A_3(\beta N)^2. \tag{5.69}$$

Then for all $n\le N$, the marginal density for $p_n^*(\xi)$ is log-concave under the continuous uniform prior. If equation (5.69) is a strict inequality, the density is strictly log-concave.

A relevant δ may be $1/Kd$ or a power thereof, though a small constant value such as say $1/300$ is also acceptable (to satisfy Corollary 1 for example).

Proof. Fix any $n \le N$. By Corollary 3, the Hessian of $\log p_n^*(\xi)$ is log-concave when for any unit vector a, we have

$$\rho \mathrm{Var}_{P_n^*}[\sum_{k=1}^{K} a_k^{\mathrm{T}} \mathbf{X} w_k | \xi] \le 1. \tag{5.70}$$

By Lemma 4, 5, 6 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 (5.61), (5.62). Fix the choice,

$$\ell^* = A_1 C_n V \beta n + A_2 \sqrt{C_n V K \beta n \log(\frac{2Kd}{\delta})}. \tag{5.71}$$

This gives upper bound on ρ times the variance,

$$\sqrt{\frac{3}{2}} a_2 \frac{\beta C_n V}{K} \frac{4n}{\sqrt{e}d} \ell^* \tag{5.72}$$

$$= 4\sqrt{\frac{3}{2e}} A_1 a_2 \frac{(C_n V \beta n)^2}{Kd} \tag{5.73}$$

$$+ 4\sqrt{\frac{3}{2e}} A_2 a_2 \frac{(C_n V \beta n)^{\frac{3}{2}} \sqrt{K}}{Kd} \sqrt{\log(\frac{2Kd}{\delta})} \tag{5.74}$$

$$\le 4\sqrt{\frac{3}{2e}} a_2 \frac{(\beta N)^2}{Kd} \Big[A_2 (C_N V)^2 + A_1 (C_N V)^{\frac{3}{2}} \Big(\frac{K(\log(\frac{2Kd}{\delta})}{\beta N} \Big)^{\frac{1}{2}} \Big]. \tag{5.75}$$

By assumption,

$$\frac{K \log(\frac{2Kd}{\delta})}{\beta N} \le 1, \tag{5.76}$$

so we have upper bound on (5.70),

$$4\sqrt{\frac{3}{2e}} a_2 (C_N V)^{\frac{3}{2}} [A_1 + A_2 (C_N V)^{\frac{1}{2}}] \frac{(\beta N)^2}{Kd}. \tag{5.77}$$

If Kd satisfies condition (5.69), then ρ times the variance is less than 1 in expression (5.70). By Corollary 3, this implies log-concavity of the induced marginal density on ξ. ∎

Remark 3. *Note that ℓ as used in the proof via the Hölder Inequality must be an integer. Thus the ℓ^* in equation (5.71) is the optimal continuous value. We would have to round up or down to the nearest integer. This would result in $\ell^* \pm \epsilon$ for a number $|\epsilon| < 1$ in equation (5.72) instead of ℓ^*. This would give an additional term $\beta N/(Kd)$ in the expression (5.77), yet this is a lower order dependence that $(\beta N)^2/(Kd)$, so it would still be controlled.*

Remark 4. *Note the interior weight dimension d can be made artificially larger by repeating the input vectors. Say the original input vectors x_i have a default dimension of $\tilde{d}$. Define new input vectors by repeating the data L times*

$$\tilde{x}_i = (x_i, \cdots, x_i) \in \mathbb{R}^{\tilde{d}L}. \tag{5.78}$$

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

The span of the new data matrix under ℓ_1 controlled input vectors, $\{z = \tilde{\boldsymbol{X}}w, \|w\|_1 \leq 1\}$, is the same as the original matrix. So we have the same approximation ability of the network. This can also equivalently be considered as inducing some different prior on the original w_k weight vectors of dimension $\tilde{d}$ that is more concentrated than uniform. However, it is more convenient to consider a uniform prior in a higher $d = L\tilde{d}$ dimensional space. This is introducing even more multi-modality into the original density $p_n(w)$ as multiple longer weight vectors yield the same output in the neural network. Yet by our proceeding theorems we have shown the density can be decomposed into a log-concave mixture.

6. Risk Bounds

6.1. Introductory Concepts in Risk Control

For risk control, we want to compare the performance of our Bayesian posterior to the best possible approximation in the model class. Note our previous sampling results are for the continuous uniform prior on $(S_1^d)^K$. When bounding posterior risk, we will provide bounds for the discrete uniform prior on $(S_{1,M}^d)^K$. To recall, the discrete prior forces coordinate values to be whole number multiples of $\frac{1}{M}$ for an integer M. The finite size of the support of the discrete prior makes it easier to analyze under our approach, which relies on the prior probability of any single point being not too small. In Section 7, we will extend these risk results to the continuous uniform prior as well.

Consider $(x_i, y_i)_{i=1}^N$ as an arbitrary sequence of inputs and response values. Let $p_n(w|x^n, y^n)$ be the posterior density trained on data up to index n with gain β. Recall the definitions of posterior mean and predictive density

$$\mu_n(x) = E_{P_n}[f(x, w)|x^n, y^n] \tag{6.1}$$

$$p_n(y|x, x^n, y^n) = E_{P_n}[\frac{\sqrt{\beta}}{\sqrt{2\pi}} e^{-\frac{\beta}{2}(y - f(x,w))^2}|x^n, y^n]. \tag{6.2}$$

Let g be a competitor function we want to compare our performance to. Define its predictive density $q(y|x)$ as Normal$(g(x), \frac{1}{\beta})$. The individual squared error regret is

defined as

$$r_n^{\text{square}} = \frac{1}{2}\Big[(y_n - \mu_{n-1}(x_n))^2 - (y_n - g(x_n))^2\Big]. \tag{6.3}$$

We also define the randomized regret and log regret as

$$r_n^{\text{rand}} = \frac{1}{2}\Big[E_{P_{n-1}}[(y_n - f(x_n, w))^2] - (y_n - g(x_n))^2\Big] \tag{6.4}$$

$$r_n^{\log} = \frac{1}{\beta}\Big[\log \frac{1}{p_{n-1}(y_n|x_n, x^{n-1}, y^{n-1})} - \log \frac{1}{q(y_n|x_n)}\Big]. \tag{6.5}$$

We then have the following ordering of the regrets [32].

Lemma 7. *Assume f_w, g are bounded in absolute value by b_f, b_g. Define*

$$\epsilon_n = y_n - g(x_n) \quad b = \frac{b_f + b_g}{2} \quad \lambda_n = b|\epsilon_n| + b^2. \tag{6.6}$$

Then we have

$$r_n^{log} \le r_n^{rand} \tag{6.7}$$

$$r_n^{square} \le r_n^{rand} \le r_n^{log} + 2\beta\lambda_n^2. \tag{6.8}$$

Proof. $r_n^{\text{square}} \le r_n^{\text{rand}}$ and $r_n^{\log} \le r_n^{\text{rand}}$ by Jensen's inequality. Consider

$$\frac{1}{2}[(y_n - f(x_n, w))^2 - (y_n - g(x_n))^2], \tag{6.9}$$

as a random variable in w. Then r_n^{rand} is its expected value and $r_n^{\log}$ is $\frac{1}{\beta}$ times its cumulant generating function at β. Note that by a difference in squares identity,

$$\frac{1}{2}[(y_n - f(x_n, w))^2 - (y_n - g(x_n))^2] = (g(x_n) - f(x_n, w))(\epsilon_n + \frac{g(x_n) - f(x_n, w)}{2}) \tag{6.10}$$

$$\le 2b(|\epsilon_n| + b) \tag{6.11}$$

$$= 2\lambda_n. \tag{6.12}$$

By second order Taylor expansion, the cumulant generating function of a bounded random matches the mean to within half the range squared. Thus, we have

$$r_n^{\text{rand}} \le r_n^{\log} + 2\beta\lambda_n^2. \tag{6.13}$$

■

Define the averaged quantities as

$$R_N^{\text{square}} = \frac{1}{N}\sum_{n=1}^{N} r_n^{\text{square}} \qquad R_N^{\text{rand}} = \frac{1}{N}\sum_{n=1}^{N} r_n^{\text{rand}} \tag{6.14}$$

$$R_N^{\log} = \frac{1}{N}\sum_{n=1}^{N} r_n^{\log} \qquad \Lambda_N^2 = \frac{1}{N}\sum_{n=1}^{N} \lambda_n^2. \tag{6.15}$$

The average regrets follow the same ordering as the pointwise components,

$$R_N^{\text{square}} \le R_N^{\text{rand}} \le R_N^{\log} + 2\beta\Lambda_N^2. \tag{6.16}$$

The easiest of the regrets to bound is the log regret as it has a telescoping cancellation of log terms.

Lemma 8. *The average log regret is upper bound as*

$$R_N^{log} \le -\frac{1}{\beta N}\log E_{P_0}[e^{-\frac{\beta}{2}\sum_{n=1}^{N}(y_n - f(x_n,w))^2}] - \frac{1}{2}\frac{1}{N}\sum_{n=1}^{N}(y_n - g(x_n))^2. \tag{6.17}$$

Proof. Denote the Bayes factor as

$$Z_n = E_{P_0}[\frac{e^{-\frac{\beta}{2}\sum_{i=1}^{n}(y_i - f(x_i,w))^2}}{(2\pi/\beta)^{\frac{n}{2}}}]. \tag{6.18}$$

The predictive density for p_{n-1} is then the ratio of Z_n to Z_{n-1},

$$p_{n-1}(y_n|x_n, x^{n-1}, y^{n-1}) = \frac{Z_n}{Z_{n-1}}. \tag{6.19}$$

Note this result requires reciprocal variance in our predictive density to match the β gain used in the definition of our Bayesian model. The sum of logs then becomes a telescoping product of canceling terms.

$$-\frac{1}{N}\sum_{n=1}^{N}\log p_{n-1}(y_n|x_n, x^{n-1}, y^{n-1}) \tag{6.20}$$

$$= -\frac{1}{N}\log\prod_{n=1}^{N}\frac{Z_n}{Z_{n-1}} \tag{6.21}$$

$$= -\frac{1}{N}\log\frac{Z_N}{Z_0} \tag{6.22}$$

$$= -\frac{1}{2}\log\left(\frac{\beta}{2\pi}\right) - \frac{1}{N}\log E_{P_0}[e^{-\frac{\beta}{2}\sum_{n=1}^{N}(y_n - f(x_n,w))^2}]. \tag{6.23}$$

The $\beta/2\pi$ terms appear in both p and q, and cancel. ∎

The key term for bounding risk performance will ultimately depend on a cumulant generating function of loss using the prior,

$$-\frac{1}{\beta N}\log E_{P_0}[e^{-\frac{\beta}{2}\sum_{n=1}^{N}(y_n-f(x_n,w))^2}]. \tag{6.24}$$

Providing upper bounds on this term is the main driving force of risk control. With this key expression controlled by a choice of prior, various notions of risk such as Kullback divergence, mean squared risk, and arbitrary sequence regret can be deduced.

One way to upper bound this cumulant generating function is through the index of resolvability [5] approach, which relies on the prior probability of a set of good approximators.

Lemma 9 (Index of Resolvability). *Let the prior distribution P_0 have support S and let A be any measurable subset of S. Then we have upper bound*

$$-\frac{1}{\beta N}\log E_{P_0}[e^{-\frac{\beta}{2}\sum_{n=1}^{N}(y_n-f(x_n,w))^2}] \le \frac{-\log P_0(A)}{\beta N} + \max_{w\in A}\frac{1}{N}\sum_{n=1}^{N}\frac{1}{2}(y_n-f(x_n,w))^2. \tag{6.25}$$

Proof. The proof of this approach is quite simple. The integral on the full space is more than the integral on a subset, thus restricting to a set A upper bounds the negative log integral,

$$-\frac{1}{\beta N}\log E_{P_0}[e^{-\frac{\beta}{2}\sum_{n=1}^{N}(y_n-f(x_n,w))^2}] \le -\frac{1}{\beta N}\log\int_A e^{-\frac{\beta}{2}\sum_{n=1}^{N}(y_n-f(x_n,w))^2}P_0(dw). \tag{6.26}$$

Multiply and divide by the prior probability of the set $P_0(A)$.

$$\frac{-\log P_0(A)}{\beta N} - \frac{1}{\beta N}\log E_{P_0}[e^{-\frac{\beta}{2}\sum_{n=1}^{N}(y_n-f(x_n,w))^2}|w\in A]. \tag{6.27}$$

Then upper bound the conditional mean by the largest value of the object in the exponent for w in A. ∎

This philosophy makes risk control quite clear. First, there must exist at least one point in the support of the prior which produces a good fit for the data. Second, the prior must place enough probability around this point (or rather, at this point in the case of discrete priors) so that the prior probability of the set A is not exponentially small in N. Then both terms of the index of resolvability are controlled.

Note that our finite width neural networks can approximate functions well when the target function lives in V times the convex hull of signed neurons. For a given input data $x^N = (x_i)_{i=1}^N$ and a weight vector w, consider the vector in $\mathbb{R}^N$ of a single neuron

evaluated at the $w \cdot x_i$ points. Let the subset of $\mathbb{R}^N$ denoted $\text{Hull}_N(V\Psi)$ be the closure of the set of convex combinations of V times signed neurons in Ψ evaluated at x^N. This is (the closure of) the set of single hidden layer neural networks with variation at most V, evaluated at the given data. For vector of target function values $(g(x_i))_{i=1}^N$, we denote its projection as

$$\tilde{g} = \text{argmin}_{f \in \text{Hull}_N(V\Psi)} \|g - f\|_N. \tag{6.28}$$

Note $\tilde{g}$ is the vector of numerical values $\tilde{g} = (\tilde{g}_1, \cdots, \tilde{g}_N) \in \mathbb{R}^N$, which the vector of outputs of some network evaluated at the x_i points, not the network itself that would give rise to these outputs.

We will also have consideration of $\text{Hull}(V\Psi)$ defined as the $L_2(P_X)$ closure of the set of convex combinations of V times signed neurons in Ψ as functions on $[-1, 1]^d$. The $L_2(P_X)$ projection of a function g defined as $\tilde{g}$, the corresponding minimizer of $\|g - f\|^2$ within $\text{Hull}(V\Psi)$, is then a function itself not a vector of specific output values.

For the arbitrary sequence regret bounds the best comparator $\tilde{g}$ is the Euclidean projection into $\text{Hull}_N(V\Psi)$, and for the statistical mean square risk bounds it is the $L_2(P_X)$ projection into $\text{Hull}(V\Psi)$.

We now review results for functions g in V times the convex hull of Ψ, how well a finite width network can approximate them.

6.2. Approximation Ability of Single-Hidden-Layer Neural Networks

First, we recall some known results about the approximation ability of neural networks. We have the following established approximation result from previous work [25].

Lemma 10. *Assume g is a target function with variation V, that is $\frac{g}{V}$ lives in the closure of the convex hull of neurons with ℓ_1 controlled weight vectors. Then there exists a finite width network with K neurons and some choice of continuous neurons weights $w_1^*, \cdots, w_K^* \in (S_1^d)^K$ and outer weights $c_1, \cdots c_K \in \{-\frac{V}{K}, \frac{V}{K}\}^K$ such that*

$$\sum_{i=1}^{N} (f(x_i, w^*) - g(x_i))^2 \leq N \frac{V^2}{K}. \tag{6.29}$$

We can slightly modify this result to focus on discrete neuron weight vectors in $S_{1,M}^d$ as opposed to the full continuous space.

Lemma 11. *Assume g lives in the closure of the convex hull of signed neurons scaled by V. Then for any sequence $(x_i, y_i)_{i=1}^N$ with all $x_i \in [-1, 1]^d$, there exists a choice of K discrete-valued interior weights $(w_1^*, \cdots, w_K^*) \in (S_{1,M}^d)^K$ and signed outer weights*

$c_k \in \{-\frac{V}{K}, \frac{V}{K}\}$ *such that the regret compared to* g *is bound by*

$$\sum_{i=1}^{N} \left(y_i - \sum_{k=1}^{K} c_k \psi(x_i \cdot w_k^*)\right)^2 - (y_i - g(x_i))^2 \leq N \frac{a_0^2 V^2}{K} + N \frac{(V C_N a_2 + V^2 a_1^2)}{M}, \tag{6.30}$$

where a_0, a_1, a_2 *are the bounds on* ψ *and its derivatives, and* $C_N = \max_{n \leq N} |y_n| + a_0 V$.

Proof. Since g lives in the closure of the convex hull of signed neurons scaled by V, for every $\epsilon > 0$ there exists some finite width neural network with continuous-valued weight vectors $w_\ell \in S_1^d$ and outer weights c_ℓ with $\sum_\ell |c_\ell| = 1$ such that

$$\tilde{g}(x) = V \sum_{\ell} c_\ell \psi(x \cdot w_\ell), \quad \sum_{i=1}^{N} (g(x_i) - \tilde{g}(x_i))^2 \leq \epsilon. \tag{6.31}$$

Let L be a random draw of neuron index where $L = \ell$ with probability $|c_\ell|$. Define $w^{\text{cont}} = w_L$ as the continuous neuron vector at the selected random index L, and $s^{\text{cont}} = \text{sign}(c_L)$ as the sign of the outer weight.

Given a continuous vector w^{cont} of dimension d, we then make a random discrete vector as follows. Define a $d+1$ coordinate, $w_{d+1}^{\text{cont}} = 1 - \|w_{1:d}^{\text{cont}}\|_1$, to have a $d+1$ length vector which sums to 1. Consider a random index $J \in \{1, \cdots d+1\}$ where $J = j$ with probability $|w_j^{\text{cont}}|$. Given w^{cont}, this defines a distribution on $\{1, \cdots, d+1\}$. Draw M iid random indices $J_1, \cdots, J_M$ from this distribution and define the counts of each index

$$m_j = \sum_{i=1}^{M} 1\{J_i = j\}. \tag{6.32}$$

We then define the discrete vector $w^{\text{disc}} \in S_{1,M}^d$ with coordinate values

$$w_j^{\text{disc}} = \text{sign}(w_j^{\text{cont}}) \frac{m_j}{M}. \tag{6.33}$$

Consider then K iid draws of random indexes $L_1, \cdots L_K$, as well as corresponding signs $s_k = \text{sign}(c_{L_k})$. For each L_k consider M iid drawn indexes $J_1^k, \cdots, J_M^k$. This also defines continuous vectors w_k^{cont} and discrete vectors w_k^{disc}. Denote the neural network using a random set of weights and signs,

$$f(x, w, s) = \sum_{k=1}^{K} \frac{V}{K} s_k \psi(x \cdot w_k). \tag{6.34}$$

Recall the empirical norm and inner product definitions $\|\cdot\|_N^2, \langle\cdot,\cdot\rangle_N$ from the notation section. Consider the expected regret using random discrete neuron weights.

$$E\Big[\|y - f(\cdot, w^{\text{disc}}, s)\|_N^2 - \|y - g\|_N^2\Big]. \tag{6.35}$$

Note this expectation is with respect to the previously defined distribution for w^{disc}, w^{cont}, and s. The data $(x_i, y_i)_{i=1}^N$ are fixed.

Add and subtract the norm using continuous weight vectors, noting that the discrete and continuous vectors of the same index are dependent via the construction,

$$E\Big[\|y - f(\cdot, w^{\text{cont}}, s)\|_N^2 - \|y - g\|_N^2\Big] \tag{6.36}$$

$$+E\Big[\|y - f(\cdot, w^{\text{disc}}, s)\|_N^2 - \|y - f(\cdot, w^{\text{cont}}, s)\|_N^2\Big]. \tag{6.37}$$

Note that using continuous weight vectors the expected value of the random neural network is exactly $\tilde{g}$,

$$E[\frac{V}{K}\sum_{k=1}^{K} s_k \psi(x_i \cdot w_k^{\text{cont}})] = \sum_{i=1}^{N} \tilde{g}(x_i). \tag{6.38}$$

Thus using a bias variance decomposition we have the bound on expression (6.36),

$$E\Big[\|y - f(\cdot, w^{\text{cont}}, s)\|_N^2 - \|y - g\|_N^2\Big] \tag{6.39}$$

$$=\sum_{n=1}^{N} \text{Var}(f(x_i, w^{\text{cont}}, s))^2 + \|y - \tilde{g}\|_N^2 - \|y - g\|_N^2 \tag{6.40}$$

$$\leq N\frac{a_0^2 V^2}{K} + 2\|y - g\|_N \|g - \tilde{g}\|_N + \|\tilde{g} - g\|_N^2 \tag{6.41}$$

$$=N\frac{a_0^2 V^2}{K} + 2\sqrt{N} C_N \sqrt{\epsilon} + \epsilon. \tag{6.42}$$

Where we have used that $f(x, w^{\text{cont}}, s)$ is an average of K iid terms each bounded by $a_o V$, so its variance is less than $a_0^2 V^2/K$.

For expression (6.37), perform a second order Taylor expansion of $\|y - f(\cdot, w^{\text{disc}}, s)\|_N^2$ as a function of w^{disc} centered at w^{cont}. For any other vector $\tilde{w}$, denote the expressions

$$\text{res}_i(w, s) = y_i - \sum_{k=1}^{K} s_k \frac{V}{K}\psi(x_i \cdot w_k) \tag{6.43}$$

$$a_{i,k} = -s_k \frac{2V}{K}\text{res}_i(w^{\text{cont}}, s)\psi'(x_i \cdot w_k^{\text{cont}}) \tag{6.44}$$

$$b_{i,k,k'}(\tilde{w}, s) = -s_k \frac{2V}{K}\text{res}_i(\tilde{w}, s)\psi''(x_i \cdot \tilde{w}_k)\delta_{k=k'} + 2s_k s_{k'} \frac{V^2}{K^2}\psi'(x_i \cdot \tilde{w}_k)\psi'(x_i \cdot \tilde{w}_{k'}). \tag{6.45}$$

Then for any continuous valued vector w^{cont} and discrete valued vector w^{disc}, there exists some vector $\tilde{w}$ (in fact along the line between w^{disc} and w^{cont}) such that the second order expansion is exact using that $\tilde{w}$ in the second derivative terms,

$$\|y - f(\cdot, w^{\text{disc}}, s)\|_N^2 \tag{6.46}$$

$$= \|y - f(\cdot, w^{\text{cont}}, s)\|_N^2 + \sum_{i=1}^{N}\sum_{k=1}^{K} a_{i,k}(x_i \cdot (w_k^{\text{disc}} - w_k^{\text{cont}})) \tag{6.47}$$

$$+ \frac{1}{2}\sum_{i=1}^{n}\sum_{k,k'=1}^{K} b_{i,k,k'}(\tilde{w}, s)(x_i \cdot (w_k^{\text{disc}} - w_k^{\text{cont}}))(x_i \cdot (w_{k'}^{\text{disc}} - w_{k'}^{\text{cont}})). \tag{6.48}$$

Expanding the terms we have the expression,

$$E\left[\|y - f(\cdot, w^{\text{disc}}, s)\|_N^2 - \|y - f(\cdot, w^{\text{cont}}, s)\|_N^2\right] \tag{6.49}$$

$$= \sum_{i=1}^{N}\sum_{k=1}^{K} a_{i,k} E[x_i \cdot (w_k^{\text{disc}} - w_k^{\text{cont}})] \tag{6.50}$$

$$- \frac{V}{K}\sum_{i=1}^{N}\sum_{k=1}^{K} E\left[\text{res}_i(\tilde{w}, s)\psi''(x_i \cdot \tilde{w}_k)(x_i \cdot (w_k^{\text{disc}} - w_k^{\text{cont}}))^2\right] \tag{6.51}$$

$$+ \sum_{i=1}^{N} E\left[\left(\sum_{k=1}^{K} s_k \frac{V}{K}\psi'(\tilde{w}_k)(x_i \cdot (w_k^{\text{disc}} - w_k^{\text{cont}}))\right)^2\right]. \tag{6.52}$$

By construction of the distribution, $E[w_k^{\text{disc}} | w_k^{\text{cont}}] = w_k^{\text{cont}}$ so the first order term (6.50) is mean 0. Note that for each i, $|\text{res}_i(\tilde{w}, s)| \le C_N, \psi'(\cdot) \le a_1, \psi''(\cdot) \le a_2$ so we have upper bound

$$= (VC_N a_2 + V^2 a_1^2)\sum_{i=1}^{N}\sum_{k=1}^{K} \frac{1}{K} E[(x_i \cdot (w_k^{\text{disc}} - w_k^{\text{cont}}))^2] \tag{6.53}$$

$$= (VC_N a_2 + V^2 a_1^2)\sum_{i=1}^{N} E[\text{Var}[x_i \cdot w_1^{\text{disc}} | w_1^{\text{cont}}]], \tag{6.54}$$

since the distribution of $(w_k^{\text{disc}}, w_k^{\text{cont}})$ is the same for $k = 1, \cdots, K$.

For a fixed choice of continuous w_1^{cont}, let $x_{i,d+1} = 0$ and consider x_i as a $d+1$ dimension vector. Then $x_i \cdot w_1^{\text{disc}}$ is the inner product of x_i with a vector defined by counts of the independent random indexes $J_1^1, \cdots, J_M^1$. Therefore, the inner product

can equivalently be written as an average of M iid random variables using these indexes,

$$\mathrm{Var}[x_i \cdot w_1^{\mathrm{disc}} | w_1^{\mathrm{cont}}] = \mathrm{Var}[\frac{1}{M} \sum_{t=1}^{M} x_{i,J_t^1} | w_1^{\mathrm{cont}}] \tag{6.55}$$

$$= \frac{1}{M} \mathrm{Var}[x_{i,J_1^1} | w_1^{\mathrm{cont}}] \tag{6.56}$$

$$\leq \frac{1}{M}, \tag{6.57}$$

since the $|x_{i,j}|$ are all bounded by 1.

The support of the product measure on discrete weights and outer signs is $(S_{1,M}^d)^K \times \{-1, 1\}^K$. There must be at least one element of the support that has a regret equal to or lower than the average regret. Then taking $\epsilon \to 0$ and complete the proof. ∎

Remark 5. *We make a note here about odd symmetric activation functions, such as as the tanh function, and non-odd symmetric functions, such as the ReLU squared. In the general convex hull approximator $\tilde{g}$, the signs of the outer weights c_r are not known to us in defining our model. Yet in our Bayesian model we fix the signs of our outer neuron scalings c_k as specific signed values, and they are not modeled as flexible in the posterior distribution.*

For odd symmetric activation functions, we can consider all signed outer weights to be positive, and any negative outer scalings could be equivalently generated by using negative inner weight vectors. Thus, we can consider all $c_k = \frac{V}{k}$ in our model and the signed discussion in the previous proof becomes irrelevant.

For non-odd symmetric activation functions, if we use double the variation $\tilde{V} = 2V$ and double the number of neurons $\tilde{K} = 2K$, fix the first K outer weights to be positive and the second K to be negative. Then by setting half of inner the weights to be the zero vector, any selection of K inner weights and K signed outer weights can be generated by the model twice as wide. In essence, a non-odd symmetric activation function uses twice the variation and twice the number of neurons to ensure any signed network of size K and variation V can be generated by a certain choice of interior weights alone and fixed outer weights.

6.3. Arbitrary Sequence Risk Control

We now apply these results to a specific choice of prior. The discrete uniform prior on $(S_{1,M}^d)^K$ is a uniform distribution with less than $(2d+1)^{MK}$ possible values. As such, the negative prior log probability of a single point only grows logarithmically in the dimension. By Lemma 11, for any target function of the given variation, the set $(S_{1,M}^d)^K$ contains at least one choice of parameters that is a good approximation to the function. This yields the following result.

Theorem 9 (Odd-Symmetric Neurons). *Let g be a target function and let h be any element of the closure of the convex hull of signed neurons scaled by V. Let P_0 be the uniform prior on $(S^d_{1,M})^K$. Assume the neuron activation function is odd symmetric and set all outer weights as $c_k = \frac{V}{K}$. For any data sequence $(x_i, y_i)_{i=1}^N$ with all $x_i \in [-1,1]^d$, define the terms*

$$\epsilon_n = y_n - g(x_n) \qquad \tilde{\epsilon}_n = y_n - h(x_n). \tag{6.58}$$

Then the average log regret of the sequence of posterior predictive distributions is upper bounded by

$$R_N^{\log} \le \frac{MK\log(2d+1)}{\beta N} + \frac{a_0^2 V^2}{2K} + \frac{(VC_N a_2 + V^2 a_1^2)}{2M} + \frac{1}{2}\frac{1}{N}\sum_{n=1}^{N}(\tilde{\epsilon}_n^2 - \epsilon_n^2). \tag{6.59}$$

In particular, h can be considered as the $Hull_N(V\Psi)$ projection of g denoted $\tilde{g}$.

Proof. Recall the definition

$$\|h_1 - h_2\|_N^2 = \sum_{n=1}^{N}(h_1(x_i) - h_2(x_i))^2 \quad \langle h_1, h_2\rangle_N = \sum_{i=1}^{N} h_1(x_i)h_2(x_i), \tag{6.60}$$

for functions of the x_i sequence. By Lemmas 8 and 9, for any set A of discrete neuron values, we can upper bound the average log regret as

$$-\frac{\log P_0(A)}{\beta N} + \frac{1}{2N}\max_{w\in A}((\|y - f_w\|_N^2 - \|y - g\|_N^2) \tag{6.61}$$

$$= -\frac{\log P_0(A)}{\beta N} + \frac{1}{2N}\max_{w\in A}((\|y - f_w\|^2 - \|y - h\|_N^2) + \frac{1}{2N}(\|y - h\|_N^2 - \|y - g\|_N^2). \tag{6.62}$$

By Lemma 11, there exists a single discrete point with bounded regret from h. Select A as the singleton set at this point. We then consider the number of points in the support of the prior.

Let w be a vector of length d with ℓ_1 norm less than or equal to 1. To make a vector with only positive entries, use double the coordinates and set $\tilde{w}_j = w_j$ if $w_j > 0$ and $\tilde{w}_{d+j} = -w_j$ else. Then add one more coordinate to count how far the ℓ_1 norm is from 1, $\tilde{w}_{2d+1} = 1 - \|w\|_1$. Thus, each w vector can be uniquely expressed as a $2d+1$ size vector of positive entries that sums to exactly 1.

Consider the entries of $\tilde{w}$ as having to be multiples of $\frac{1}{M}$. Each $\tilde{w}$ vector is then a histogram on $2d+1$ locations where the heights at each location can be $\{0, 1, \cdots, M\}/M$. An over-counting of the number of possible histograms is then $(2d+1)^M$. The product

prior on K independent weight vectors gives an additional K power. Since the discrete uniform prior support set has less than $(2d+1)^{MK}$ points,

$$-\log P_0(A) \leq (MK)\log(2d+1). \tag{6.63}$$

Combined with the bound from Lemma 11 this completes the proof.

■

In general, for a non-odd symmetric activation function (e.g. squared ReLU) we use twice the number of neurons with fixed outer weights to ensure any choice of signed neurons of half the width can be generated. Thus, we can prove the same order bounds but with slightly different constants. Here, we give the explicit changes, but all future theorems will be given for the odd-symmetric case and the non-odd symmetric version can be similarly derived.

Corollary 4 (Non-Odd Symmetric Neurons). *For a neural network with non-odd symmetric neurons, use twice the number of neurons $\tilde{K} = 2K$ neurons and twice the variation $\tilde{V} = 2V$. Set the first K outer weights as positive $c_k = \frac{V}{K}$ and the second K outer weights as negative $c_k = -\frac{V}{K}$. Then we have the bound of*

$$R_N^{\log} \leq \frac{M\tilde{K}\log(2d+1)}{\beta N} + \frac{a_0^2\tilde{V}^2}{\tilde{K}} + \frac{(\tilde{V}C_N a_2 + \tilde{V}^2 a_1^2)}{2M} + \frac{1}{2}\frac{1}{N}\sum_{n=1}^{N}(\tilde{\epsilon}_n^2 - \epsilon_n^2). \tag{6.64}$$

Proof. By Lemma 11, there exists some signed neural network of width K that achieves the given regret bound with target function g. Our chosen network of width $\tilde{K}$ of fixed signed neurons has the flexibility to generate arbitrary signed (i.e. any number proportion of positive or negative signs) networks of width $K = \frac{\tilde{K}}{2}$. The proof then follows. ■

Theorem 10. *Let $g(x)$ be a target function bounded by a value b and let $\tilde{g}$ be its projection into $Hull_N(V\Psi)$. Let P_0 be the uniform prior on $(S_{1,M}^d)^K$. Assume the neuron activation function is odd symmetric and set all outer weights as $c_k = \frac{V}{K}$. For any data sequence $(x_i, y_i)_{i=1}^N$ with all $x_i \in [-1,1]^d$, the average squared regret of the posterior mean predictions is upper bounded by*

$$R_N^{square} \leq \frac{MK\log(2d+1)}{\beta N} + \frac{a_0^2V^2}{2K} + \frac{(VC_N a_2 + V^2 a_1^2)}{2M} \tag{6.65}$$

$$+ 2\beta\frac{1}{N}\sum_{n=1}^{N}\left(\frac{a_0V+b}{2}|\tilde{\epsilon}_n| + \left(\frac{a_0V+b}{2}\right)^2\right)^2 + \frac{1}{2}\frac{1}{N}\sum_{n=1}^{N}(\tilde{\epsilon}_n^2 - \epsilon_n^2). \tag{6.66}$$

Proof. Apply Lemma 7 and Theorem 9 to upper bound squared regret by log regret and an additional β term. Note that f_w is bounded by a_0V and g is bounded by b. ■

Note that the residuals $\tilde{\epsilon}$ are not known to us., but we do have upper bound $|\tilde{\epsilon}_n| \leq C_N$. Thus we can replace the $\tilde{\epsilon}_n$ with C_N and derive the optimal choices ofβ, M, K which can make the expression small.

Corollary 5. *Replace the residuals $\tilde{\epsilon}_n$ with C_N in expression (6.66). Denote the term*

$$B_1 = (C_N + \frac{a_0 V + b}{2})^2 \tag{6.67}$$

Let

$$\beta^* = \gamma_1 \Big(\frac{\log(2d+1)}{N}\Big)^{\frac{1}{4}} \tag{6.68}$$

$$K^* = \gamma_2 \Big(\frac{N}{\log(2d+1)}\Big)^{\frac{1}{4}} \tag{6.69}$$

$$M^* = \gamma_3 \Big(\frac{N}{\log(2d+1)}\Big)^{\frac{1}{4}}, \tag{6.70}$$

where

$$\gamma_1 = \frac{(a_0 V)^{\frac{1}{2}} (\frac{a_2 V C_N + a_1^2 V^2}{2})^{\frac{1}{4}}}{2(\frac{a_0 V + b}{2})^{\frac{3}{2}} (B_1)^{\frac{3}{4}}} \tag{6.71}$$

$$\gamma_2 = \frac{(a_0 V)^{\frac{3}{2}}}{2(\frac{a_0 V + b}{2})^{\frac{1}{2}} (B_1)^{\frac{1}{4}} (\frac{a_2 V C_N + a_1^2 V^2}{2})^{\frac{1}{4}}} \tag{6.72}$$

$$\gamma_3 = \frac{(\frac{a_2 V C_N + a_1^2 V^2}{2})^{\frac{3}{4}}}{(a_0 V)^{\frac{1}{2}} (\frac{a_0 V + b}{2})^{\frac{1}{2}} (B_1)^{\frac{1}{4}}}. \tag{6.73}$$

Then we have a bound on the squared regret of the form

$$4\Big(a_0 V (\frac{a_0 V + b}{2})\Big)^{\frac{1}{2}} \Big(B_1 (\frac{a_2 V C_N + a_1^2 V^2}{2})\Big)^{\frac{1}{4}} \Big(\frac{\log(2d+1)}{N}\Big)^{\frac{1}{4}} + \frac{1}{2}\frac{1}{N}\sum_{n=1}^{N} (\tilde{\epsilon}_n^2 - \epsilon_n^2). \tag{6.74}$$

In particular, if the function g lives in the convex hull scaled by V so $\tilde{g} = g$ then $\epsilon_n = \tilde{\epsilon}_n$ and we have an upper bound of

$$R_N^{square} = O(\|y\|_\infty^{\frac{3}{4}} \Big(\frac{\log(2d+1)}{N}\Big)^{\frac{1}{4}}). \tag{6.75}$$

In the algorithm M, K must be integers. The closest integer values to the stated continuous values achieve a similar bound.

Remark 6. *Replacing $\tilde{\epsilon}_n$ with C_N in equation (6.66), equations (6.68), (6.69), (6.70) represent the choice of modeling parameters that optimize our derived bound independent of $\tilde{\epsilon}$. However, we do not advocate plugging in these specific parameter choices directly into the model and training only one model based on these values. If for example, the dependence on say K in equation (6.65) was an improved $\frac{1}{K^2}$ rather than $\frac{1}{K}$, the given bounds would not provide the optimal choice We instead advocate adaptive modeling by putting a prior on a number of possible M, K, β values, say 100-1000 possible values each.*

Corollary 5 shows one choice of β^*, K^*, M^* that can achieve bounded regret. If we include these values in our prior set, by a further index of resolvability argument we can show using a uniform prior on a finite number of M, K, β possible values, we would pay a log number of possible values divided by βN price in the bound, which can be easily controlled. We note that computationally, all different M, K, β combinations result in different models that can be sampled in parallel and independently on different cores at the same time and the results combined at the end. Thus, such an approach is amenable to GPU usage and distributed computing from a practical perspective.

6.4. IID Sequence Predictive Risk Control

In the previous section, we studied risk control for arbitrary data sequences with no assumptions on the data. We compared performance in terms of regret to a competitor fit. Here, we assume training data iid from a data distribution and prove bounds on predictive risk for future data pairs.

Suppose $(x_i, y_i)_{i=1}^N$ are independent with y having conditional mean $E[Y|X = x] = g(x)$ and conditional variance $\text{Var}[Y|X = x] = \sigma_x^2$, with bound on the variance $\max_x \sigma_x^2 \leq \sigma^2$. Recall that our neural network is trained with a gain β. In a typical setting with assumed independent Gaussian errors, $\sigma_x^2 = \sigma^2$ for each x value and β would be set as a constant matching the inverse variance $\beta = \frac{1}{\sigma^2}$. However, we would also like to consider gains decaying in N, such as $\beta = [\log(d)/N]^{\frac{1}{4}}$. Using such a β, we can reproduce the arbitrary regret results above and show for the Cesàro mean estimator $\hat{g}$,

$$E[\|g - \hat{g}\|^2] = O\left(\left(\frac{\log(d)}{N+1}\right)^{\frac{1}{4}}\right). \tag{6.76}$$

Note that this statistical risk bound makes no assumptions about the distribution of Y given X aside from its mean and variance. In particular, the distribution of the data need not be Gaussian even though we use quadratic loss to define our posterior densities. Additionally, our sampling gain β does not have to match any data specific value exactly (that is β does not depend on σ^2 which may not be known).

If we further assume the conditional distribution is independent normal with constant variance, $Y|X \sim \text{Normal}(g(X), \sigma^2)$, and the gain β accurately represents the

inverse variance $\beta = \frac{1}{\sigma^2}$. We can give a similar bound for Kullback divergence which has an improved $1/3$ power

$$E[D(P_{Y|X}\|Q^{\text{avg}}_{Y|X,X^N,Y^N})] = O(\Big(\frac{\log(d)}{N+1}\Big)^{\frac{1}{3}}). \tag{6.77}$$

We first bound the mean squared risk without any assumptions on β and no normality assumptions.

Theorem 11. *Let $g(x)$ be a target function with absolute value bounded by b and let $\tilde{g}$ be its $L_2(P_X)$ projection into the closure of the convex hull of signed neurons scaled by V. Let P_0 be the uniform prior on $(S^d_{1,M})^K$. Assume the neuron activation function is odd symmetric and set all outer weights as $c_k = \frac{V}{K}$. Let $(x_i, y_i)_{i=1}^N$ be training data iid with conditional mean $g(x_i)$ and conditional variance $\sigma^2_{x_i}$ with variance bound $\sigma^2_x \le \sigma^2$. Assume the data distribution P_X has support $[-1,1]^d$. Then the mean squared statistical risk of the averaged posterior mean estimator $\hat{g}$ is upper bounded by*

$$E[\|g-\hat{g}\|^2] \le \frac{MK\log(2d+1)}{\beta(N+1)} + \frac{a_0^2V^2}{2K} + \frac{(V(a_0V+b)a_2 + V^2a_1^2)}{2M} \tag{6.78}$$

$$+ 2\beta(\frac{a_0V+b}{2})^2(\sigma + \frac{a_0V+b}{2})^2 + \|g-\tilde{g}\|^2. \tag{6.79}$$

Let

$$\beta^* = \gamma_1\Big(\frac{\log(2d+1)}{N+1}\Big)^{\frac{1}{4}} \tag{6.80}$$

$$K^* = \gamma_2\Big(\frac{N+1}{\log(2d+1)}\Big)^{\frac{1}{4}} \tag{6.81}$$

$$M^* = \gamma_3\Big(\frac{N+1}{\log(2d+1)}\Big)^{\frac{1}{4}}, \tag{6.82}$$

where

$$\gamma_1 = \frac{(a_0V)^{\frac{1}{2}}(\frac{V(a_0V+b)a_2+V^2a_1^2}{2})^{\frac{1}{4}}}{2(\frac{a_0V+b}{2})^{\frac{3}{2}}(\sigma+\frac{a_0V+b}{2})^{\frac{3}{2}}} \tag{6.83}$$

$$\gamma_2 = \frac{(a_0V)^{\frac{3}{2}}}{2(\frac{a_0V+b}{2})^{\frac{1}{2}}(\sigma+\frac{a_0V+b}{2})^{\frac{1}{2}}(\frac{V(a_0V+b)a_2+V^2a_1^2}{2})^{\frac{1}{4}}} \tag{6.84}$$

$$\gamma_3 = \frac{(\frac{V(a_0V+b)a_2+V^2a_1^2}{2})^{\frac{3}{4}}}{(a_0V)^{\frac{1}{2}}(\frac{a_0V+b}{2})^{\frac{1}{2}}(\sigma+\frac{a_0V+b}{2})^{\frac{1}{2}}}. \tag{6.85}$$

Then we have a bound on the mean squared risk of the form

$$4\Big(a_0V(\frac{a_0V+b}{2})(\sigma+\frac{a_0V+b}{2})\Big)^{\frac{1}{2}}\Big(\frac{V(a_0V+b)a_2+V^2a_1^2}{2}\Big)^{\frac{1}{4}}\Big(\frac{\log(2d+1)}{N}\Big)^{\frac{1}{4}} \tag{6.86}$$

$$+\|g-\tilde{g}\|^2. \tag{6.87}$$

Proof. Note the following expectations are with respect to training data $(X_i,Y_i)_{i=1}^N$ and a new input and response pair $(X,Y)=(X_{N+1},Y_{N+1})$ all iid from the data distribution $P_{X,Y}$. Note that since there are many expectations with respect to different random variables in the proof, we will make explicit use of subscripts to indicate which random variable each expectation is with respect to. The initial expectation is for the data distribution $P_{X,Y}$ for the training data as well as the new X point which we are evaluating at. Bring the average of the Cesàro mean outside the square to upper bound

$$\frac{1}{2}E_{P_{X^{N+1},Y^{N+1}}}[(g(X)-\hat{g}(X))^2]\leq\frac{1}{2}\sum_{n=0}^{N}\frac{1}{N+1}E_{P_{X^{N+1},Y^{N+1}}}[(g(X)-\mu_n(X))^2] \tag{6.88}$$

$$=\frac{1}{2}E_{P_{X^{N+1},Y^{N+1}}}\Big[\sum_{n=0}^{N}\frac{(g(X_{n+1})-\mu_n(X_{n+1}))^2}{N+1}\Big] \tag{6.89}$$

$$=\frac{1}{2}E_{P_{X^{N+1},Y^{N+1}}}\Big[\sum_{n=0}^{N}\frac{(Y_{n+1}-\mu_n(X_{n+1}))^2-(Y_{n+1}-g(X_{n+1}))^2}{N+1}\Big], \tag{6.90}$$

where we have added the Y in using the fact that $Y_{n+1}-g(X_{n+1})$ is mean 0 under $P_{X,Y}$. This is then exactly the expectation of a squared regret. Define notation $R_{N+1}^{\log}(X^{N+1},Y^{N+1})$, $R_{N+1}^{\text{square}}(X^{N+1},Y^{N+1})$ as the log and squared regret at the random $(X_i,Y_i)_{i=1}^{N+1}$ values. Then by Lemma 7 we have,

$$E_{P_{X^{N+1},Y^{N+1}}}\Big[R_{N+1}^{\text{square}}(X^{N+1},Y^{N+1})\Big]\leq E_{P_{X^{N+1},Y^{N+1}}}\Big[R_{N+1}^{\log}(X^{N+1},Y^{N+1})\Big] \tag{6.91}$$

$$+2E_{P_{X^{N+1},Y^{N+1}}}\Big[\beta\frac{1}{N+1}\sum_{n=0}^{N}(\frac{a_0V+b}{2}|Y_{n+1}-g(X_{n+1})|+(\frac{a_0V}{2})^2)^2\Big] \tag{6.92}$$

$$\leq E_{P_{X^{N+1},Y^{N+1}}}\Big[R_{N+1}^{\log}(X^{N+1},Y^{N+1})\Big]+2\beta(\frac{a_0V+b}{2})^2(\sigma+\frac{a_0V+b}{2})^2. \tag{6.93}$$

Then by Lemma 8,

$$E_{P_{X^{N+1},Y^{N+1}}}[R_{N+1}^{\log}(X^{N+1},Y^{N+1})]\leq-\frac{1}{2}\frac{1}{N+1}\sum_{n=0}^{N}E_{P_{X^{N+1},Y^{N+1}}}[(Y_{n+1}-g(X_{n+1}))^2] \tag{6.94}$$

$$+\frac{1}{\beta(N+1)}E_{P_{X^{N+1},Y^{N+1}}}[-\log\int e^{-\frac{\beta}{2}\sum_{n=0}^{N}(Y_{n+1}-f(X_{n+1},w))^2}P_0(dw)]. \tag{6.95}$$

Use the $\|\cdot\|_{N+1}^2$ and $\langle\cdot,\cdot\rangle_{N+1}$ notation defined earlier. Note the outer expectation in (6.95) is with respect to X^{N+1}, Y^{N+1} from the data distribution and the inner integral is for w using the prior, as a consequence of our index of resolvability bound. Recall that our prior P_0 is absolutely continuous with respect to a reference η with density $p_0(w)$. In this proof, η can be considered as counting measure on $(S_{1,M}^d)^K$ for the discrete uniform prior, but in other instances it could be considered as Lebesgue measure.

Add and subtract $g(X_{n+1})$ inside each of the terms in the exponent of (6.95), expand the terms and note the cancellation of the first quadratic term,

$$-\frac{1}{2}\frac{1}{N+1}E_{P_{X^{N+1},Y^{N+1}}}[\|Y-g\|_{N+1}^2] \tag{6.96}$$

$$+\frac{1}{\beta(N+1)}E_{P_{X^{N+1},Y^{N+1}}}[-\log\int p_0(w)e^{-\frac{\beta}{2}\|Y-g+g-f_w\|_{N+1}^2}\eta(dw)] \tag{6.97}$$

$$=\frac{1}{\beta(N+1)}E_{P_{X^{N+1},Y^{N+1}}}[-\log\int p_0(w)e^{-\frac{\beta}{2}\|g-f_w\|_{N+1}^2-\beta\langle Y-g,g-f_w\rangle_{N+1}}\eta(dw)]. \tag{6.98}$$

Inside the log, multiply and divide by $\int p_0(w)e^{-\frac{\beta}{2}\|g-f_w\|_{N+1}^2}\eta(dw)$, which acts as the normalizing constant of a density with respect to η,

$$\frac{E_{P_{X^{N+1},Y^{N+1}}}[-\log\int\left(\frac{p_0(w)e^{-\frac{\beta}{2}\|g-f_w\|_{N+1}^2}}{\int p_0(w)e^{-\frac{\beta}{2}\|g-f_w\|_{N+1}^2}\eta(dw)}\right)e^{-\beta\langle Y-g,g-f_w\rangle_{N+1}}\eta(dw)]}{\beta(N+1)} \tag{6.99}$$

$$+\frac{E_{P_{X^{N+1},Y^{N+1}}}[-\log\int p_0(w)e^{-\frac{\beta}{2}\|g-f_w\|_{N+1}^2}\eta(dw)]}{\beta(N+1)}. \tag{6.100}$$

Interestingly, the density in equation (6.99) can be viewed as a pseudo posterior $p_n(w|g)$ using the $g(x_i)$ data points in place of the y_i to define the likelihood. This cannot be used for actual training since the function g is not known to us, but is a tool for risk analysis.

We can then bring the $-\log$, which is a convex function, inside the integral to get an upper bound in (6.99). This brings the inner product in the exponent down. Then switch the order of the inner w integral and outer $Y^{N+1}|X^{N+1}$ expectation. Note in this analysis, the distribution of w is the prior distribution P_0 and is independent of the X^{N+1}, Y^{N+1} values. Under the data distribution, Y^{N+1} conditioned on X^{N+1} is independent of w and mean $g(X^{N+1})$, thus the expected value of the inner product is

0 for any choice of w. Thus expression (6.99) is less than 0.

$$\frac{E_{P_{X^{N+1},Y^{N+1}}}[-\log\int\left(\frac{p_0(w)e^{-\frac{\beta}{2}\|g-f_w\|^2_{N+1}}}{\int p_0(w)e^{-\frac{\beta}{2}\|g-f_w\|^2_{N+1}\eta(dw)}}\right)e^{-\beta\langle Y-g,g-f_w\rangle_{N+1}}\eta(dw)]}{\beta(N+1)} \tag{6.101}$$

$$\leq\frac{E_{P_{X^{N+1}}}[\int\left(\frac{p_0(w)e^{-\frac{\beta}{2}\|g-f_w\|^2_{N+1}}}{\int p_0(w)e^{-\frac{\beta}{2}\|g-f_w\|^2_{N+1}\eta(dw)}}\right)E_{P_{Y^{N+1}|X^{N+1}}}[\langle Y-g,g-f_w\rangle_{N+1}|X^{N+1}]\eta(dw)]}{N+1} \tag{6.102}$$

$$=0. \tag{6.103}$$

Then consider expression (6.100). This term can be bounded by the logic in Lemma 11. Add and subtract $\|\tilde{g}-g\|^2_{N+1}$ in the exponent and we have the expression

$$\frac{E_{P_{X^{N+1}}}[-\log\int e^{-\frac{\beta}{2}(\|g-f_w\|^2_{N+1}-\|g-\tilde{g}\|^2_{N+1})}P_0(dw)]}{\beta(N+1)}+\frac{1}{2}\frac{E_{P_{X^{N+1}}}[\|\tilde{g}-g\|^2_{N+1}]}{N+1} \tag{6.104}$$

$$=\frac{E_{P_{X^{N+1}}}[-\log\int e^{-\frac{\beta}{2}(\|g-f_w\|^2_{N+1}-\|g-\tilde{g}\|^2_{N+1})}P_0(dw)]}{\beta(N+1)}+\frac{E_{P_X}[(g(X)-\tilde{g}(X))^2]}{2}. \tag{6.105}$$

To bound this further, think of $g(x_i)$ as the "y_i" observations in Lemma 11, and $\tilde{g}(x_i)$ as our competitor function in the regret. The result of Lemma 11 would then apply. However, our $g(x_i)$ are now bounded which offers an improvement. Each instance of $C_{N+1}=\max_{1\leq n\leq N+1}|y_n|+a_0V$ in the result of Lemma 11 can be replaced with

$$\max_{1\leq n\leq N+1}|g(x_n)|+a_0V\leq a_0V+b, \tag{6.106}$$

which is not y dependent. Thus, the random variable y can have unbounded range, yet its mean function is bounded and the range of the mean function is the relevant term for the bound. An expression like Theorem 9 then follows replacing C_N with a_0V+b. Returning to expression (6.93) and applying this bound, we have our final expression,

$$\frac{MK\log(2d+1)}{\beta(N+1)}+\frac{a_0^2V^2}{2K}+\frac{(V(a_0V+b)a_2+V^2a_1^2)}{2M} \tag{6.107}$$

$$+2\beta(\frac{a_0V+b}{2})^2(\sigma+\frac{a_0V+b}{2})^2+\frac{1}{2}E[(g(X)-\tilde{g}(X))^2]. \tag{6.108}$$

Plugging in the stated β^*, M^*, K^* gives the more specific bound. ∎

A corollary of this result is not only is the risk of our estimator $\hat{g}$ close to the risk of projection $\tilde{g}$ (which is the minimum risk attainable by any network), but also by a Pythagorean inequality $\hat{g}$ is close to $\tilde{g}$ itself in squared distance.

Corollary 6. *Let g be the target function and $\tilde{g}$ its $L_2(P_X)$ projection into the closure of the convex hull of signed neurons scaled by V. Assume the risk of the Cesàro mean estimator is bounded by*

$$E[\|g - \hat{g}\|^2] \leq \|g - \tilde{g}\|^2 + O(\Big(\frac{\log(d)}{N}\Big)^{\frac{1}{4}}) \tag{6.109}$$

Then the distance from $\hat{g}$ to the projection $\tilde{g}$ is bounded by this error term decaying N,

$$E[\|\tilde{g} - \hat{g}\|^2] = O(\Big(\frac{\log(d)}{N}\Big)^{\frac{1}{4}}) \tag{6.110}$$

Proof. The closure of the convex hull of signed neurons is a convex set, thus $\tilde{g}$ being the projection of g into the set provides a separating hyper-plane. This means for all points inside the closure of the convex hull, of which $\hat{g}$ is a member, we have a Pythagorean inequality,

$$\|g - \tilde{g}\|^2 + \|\tilde{g} - \hat{g}\|^2 \leq \|g - \hat{g}\|^2, \tag{6.111}$$

and thus

$$\|\tilde{g} - \hat{g}\|^2 \leq \|g - \hat{g}\|^2 - \|g - \tilde{g}\|^2. \tag{6.112}$$

The conclusion follows by taking the expectation. ∎

For a target function g, consider the distribution for $Y|X$ as Normal$(g(X), \frac{1}{\beta})$. Consider X^N, Y^N as training data used to train our Bayesian model independent according to $P_{X,Y}$ and a pair X_{N+1}, Y_{N+1} as a new data input and response pair from the same distribution not in our training set. We then bound the Kullback divergence between $P_{Y_{N+1}|X_{N+1}}$ and $Q^{\text{avg}}_{Y_{N+1}|X_{N+1},X^N,Y^N}$.

Theorem 12. *Assuming the data distribution is $Y|X \sim Normal(g(X), \frac{1}{\beta})$ we bound the we bound the Kullback risk of the posterior predictive distribution as*

$$E[D(P_{Y_{N+1}|X_{N+1}} \| Q^{avg}_{Y_{N+1}|X_{N+1},X^N,Y^N})] \leq E\Big[\frac{-\log E_{P_0}[e^{-\frac{\beta}{2}\sum_{i=1}^{N+1}(f(X_i,w)-g(X_i))^2}]}{N+1}\Big]. \tag{6.113}$$

Proof. The proof of this theorem follows much the same as the arbitrary log regret proof, with a few changes using the iid nature of the data.

The Cesàro average predictive density is a mixture of $N+1$ predictive densities $p_n(y_{n+1}|x_{n+1}, x^n, y^n)$. Since Kullback divergence is a convex function, this is less than the average of individual divergences

$$\frac{1}{N+1}\sum_{n=0}^{N} E[D(P_{Y_{N+1}|X_{N+1}} \| P_{Y_{N+1}|X_{N+1},X^n,Y^n})]. \tag{6.114}$$

We assume the training data and new data come iid from the same distribution. Therefore, the predictive distribution for any $P_{Y_{i^*}|X_{i^*},X^n,Y^n}$ is the same distribution for all $i^* > n$. That is, if a Bayesian model is only trained on data up to index n, all data of higher index is predicted the same. Thus, we have

$$\frac{1}{N}\sum_{n=0}^{N} E[D(P_{Y_{N+1}|X_{N+1}} \| P_{Y_{N+1}|X_{N+1},X^n,Y^n})] \tag{6.115}$$

$$= \frac{1}{N}\sum_{n=0}^{N} E[D(P_{Y_{n+1}|X_{n+1}} \| P_{Y_{n+1}|X_{n+1},X^n,Y^n})]. \tag{6.116}$$

Consider each individual term in (6.116), we will see a similar telescoping cancellation as in the log regret proof. Denote the Bayes factor,

$$Z_n = E_{P_0}[\frac{e^{-\frac{\beta}{2}\sum_{i=1}^{n}(y_i - f(x_i,w))^2}}{(2\pi/\beta)^{\frac{n}{2}}}]. \tag{6.117}$$

Then the predictive density $p_n(y_{n+1}|x_{n+1}, x^n, y^n)$ is the ratio of Z_{n+1} to Z_n,

$$p_n(y_{n+1}|x_{n+1}, x^n, y^n) = \frac{Z_{n+1}}{Z_n}. \tag{6.118}$$

For each individual Kullback term we have

$$E[D(P_{Y_{n+1}|X_{n+1}} \| P_{Y_{n+1}|X_{n+1},X^n,Y^n})] = E[-\frac{\beta}{2}(Y_{n+1} - g(X_{n+1}))^2 - \log\frac{Z_{n+1}}{Z_n}] \tag{6.119}$$

$$- \frac{1}{2}\log(\frac{2\pi}{\beta}). \tag{6.120}$$

Use notation $\|\cdot\|_{N+1}, \langle\cdot,\cdot\rangle_{N+1}$ as before. The sum of Kullback risks divided by $N+1$ is

$$-\frac{\beta}{2}E[\frac{\|Y-g\|_{N+1}^2}{N+1}] - \frac{1}{2}\log(\frac{2\pi}{\beta}) - \frac{1}{N+1}E[\log\prod_{n=0}^{N}\frac{Z_{n+1}}{Z_n}] \tag{6.121}$$

$$= -\frac{\beta}{2}E[\frac{\|Y-g\|_{N+1}^2}{N+1}] - \frac{1}{2}\log(\frac{2\pi}{\beta}) - \frac{1}{N+1}E[\log\frac{Z_{N+1}}{Z_0}]. \tag{6.122}$$

We now proceed with an argument similar to bounding equation (6.95). Consider the negative log of Z_{N+1}. Recall the prior is absolutely continuous with respect to reference

measure η. Add and subtract g inside the exponent and simplify

$$E[-\log Z_{n+1}] = E[-\log E_{P_0}[e^{-\frac{\beta}{2}\|Y-f_w\|^2_{N+1}}] + \frac{N+1}{2}\log(\frac{2\pi}{\beta}) \tag{6.123}$$

$$= E[-\log E_{P_0}[e^{-\frac{\beta}{2}\|g-f_w\|^2_{N+1}}] + \frac{\beta}{2}\|Y-g\|^2_{N+1}] + \frac{N+1}{2}\log(\frac{2\pi}{\beta}) \tag{6.124}$$

$$+ E[-\log \int \frac{p_0(w)e^{-\frac{\beta}{2}\|g-f_w\|^2_{N+1}}}{E_{P_0}[e^{-\frac{\beta}{2}\|g-f_w\|^2_{N+1}}]} e^{-\beta\langle Y-g, g-f_w\rangle_{N+1}}\eta(dw)]. \tag{6.125}$$

The second and third terms in (6.124) will cancel with the first and second terms in the Kullback risk (6.122). Term (6.125) is the same expression as (6.99), and was shown to be less than 0. ∎

Theorem 13. *Let $g(x)$ be a target function with absolute value bounded by b and let $\tilde{g}$ be its $L_2(P_X)$ projection into the closure of the convex hull of signed neurons scaled by V. Let P_0 be the uniform prior on $(S^d_{1,M})^K$. Assume the neuron activation function is odd symmetric and set all outer weights as $c_k = \frac{V}{K}$. Assuming the data distribution has $Y|X \sim Normal(g(X), \frac{1}{\beta})$, with P_X having support $[-1,1]^d$. We bound the Kullback divergence as*

$$E[D(P_{Y|X}\|Q^{avg}_{Y|X,X^N,Y^N})] \le \frac{MK\log(2d+1)}{N+1} + \beta\frac{a_0^2V^2}{2K} + \beta\frac{V(a_0V+b)a_2 + V^2a_1^2}{2M} \tag{6.126}$$

$$+ \beta\|g-\tilde{g}\|^2. \tag{6.127}$$

In particular, with the choice

$$K^* = \frac{(\frac{\beta}{2}V^4)^{\frac{1}{3}}(a_0^2)^{\frac{2}{3}}}{(V(a_0V+b)a_2 + V^2a_1^2)^{\frac{1}{3}}}\Big(\frac{(N+1)}{\log(2(d+1))}\Big)^{\frac{1}{3}} \tag{6.128}$$

$$M^* = \frac{(((a_0V+b)a_2 + V^2a_1^2)^{\frac{2}{3}}(\frac{\beta}{2})^{\frac{1}{3}}}{(a_0V)^{\frac{2}{3}}}\Big(\frac{(N+1)}{\log(2(d+1))}\Big)^{\frac{1}{3}}, \tag{6.129}$$

we would have a bound of

$$3(\frac{\beta}{2})^{\frac{2}{3}}(a_0V)^{\frac{2}{3}}(V(a_0V+b)a_2 + V^2a_1^2)^{\frac{1}{3}}\Big(\frac{\log(2d+1)}{N+1}\Big)^{\frac{1}{3}} + \beta\|g-\tilde{g}\|^2. \tag{6.130}$$

Proof. Add and subtract $\|g-\tilde{g}\|^2_{N+1}$ in the exponent of equation (6.113) to get the expression

$$\frac{E[-\log E_{P_0}[e^{-\frac{\beta}{2}(\|g-f_w\|^2_{N+1} - \|g-\tilde{g}\|^2_{N+1})}]]}{(N+1)} + \beta\frac{1}{2}\frac{E[\|\tilde{g}-g\|^2_{N+1}]}{N+1}. \tag{6.131}$$

This is the same expression as (6.104), scaled by a β. Doing the same analysis gives the bound

$$\frac{MK\log(2d+1)}{(N+1)} + \beta\frac{a_0^2V^2}{2K} + \beta\frac{(V(a_0V+b)a_2 + V^2a_1^2)}{2M} + \beta\|g-\tilde{g}\|^2. \qquad (6.132)$$

Note now that β, being the inverse variance of the data distribution, is not a design parameter we can choose. However, M and K are modeling choices. Setting M^* and K^* as given yields the final expression. ∎

7. Risk Control for the Continuous Prior

Our approach to risk control relies on the log regret being an upper bound on other forms of regret, and statistical risk being interpreted as an expected regret. The log regret can be upper bound via the index of resolvability, which utilizes the prior probability of a good set of approximators. Our approximation results show for any element of the closure of the convex hull of signed neurons, there is one set of neuron weights from the discrete lattice $(S_{1,M}^d)^K$ that is a good approximator. Also, there are not too many points in the lattice so under the uniform discrete prior, the probability of any one point is more than $(2d+1)^{-MK}$ (importantly, not exponentially small in N).

However, our sampling results are for the continuous uniform prior on $(S_1^d)^K$, under which any single point has 0 probability. Additionally, small balls around any point will have probability exponentially small in the dimension d, which we cannot afford. Thus, we would like to utilize finite state space of the discrete prior in an index of resolvability bound on the log regret, but apply these results to the continuous prior.

The key to connecting these two results is recognizing a joint distribution on discrete and continuous weight vectors which couples the vectors to be close together, but the marginal prior for each variable is uniform on $(S^d)^K$ and $(S_{1,M}^d)^K$ respectively. Then the continuous and discrete prior can be considered as different marginals of this joint distribution.

Consider P_0 as a joint prior distribution on $(S_1^d)^K \times (S_{1,M}^d)^K$, with the continuous random vector $w^{\text{cont}} \in (S_1^d)^K$ and the discrete random vector $w^{\text{disc}} \in (S_{1,M}^d)^K$. Consider the marginal distribution on w^{cont} as treating each w_k^{cont} vector as independent uniform on S_1^d. Consider an additional coordinate for each w_k^{cont} vector to track it's ℓ_1 distance from 1, $w_{k,d+1}^{\text{cont}} = 1 - \sum_{j=1}^d |w_{k,j}^{\text{cont}}|$.

Then define the conditional distribution on $w_k^{\text{disc}}|w_k^{\text{cont}}$ as follows. Force the signs of the coordinates to stay the same, $\text{sign}(w_{k,j}^{\text{disc}}) = \text{sign}(w_{k,j}^{\text{cont}})$, and have the absolute values be distributed as $1/M$ times a Multinomial$(M, |w_{k,1}^{\text{cont}}|, \cdots, |w_{k,d+1}^{\text{cont}}|)$ distribution. That

is, the conditional pmf of the absolute values of the discrete vector can be written as

$$p_0(|w_k^{\text{disc}}| \,\big|\, |w_k^{\text{cont}}|) = \frac{M!}{\prod_{j=1}^{d+1}(M|w_{k,j}^{\text{disc}}|)!} \prod_{j=1}^{d+1} |w_{k,j}^{\text{cont}}|^{M|w_{k,j}^{\text{disc}}|} \tag{7.1}$$

Note the discrete vector's coordinates themselves are whole number multiples of $1/M$, thus M times the discrete vector coordinates are whole numbers between 0 and M. There is also a $w_{k,d+1}^{\text{disc}}$ coordinate in this construction which is 1 minus the sum of the other coordinates. Then the overall prior is of the form

$$p_0(w^{\text{cont}}, w^{\text{disc}}) = \prod_{k=1}^{K} p_0(w_k^{\text{cont}}) p_0(w_k^{\text{disc}}|w_k^{\text{cont}}) \tag{7.2}$$

$$= \prod_{k=1}^{K} \text{Uniform}_{S_1^d}(w_k^{\text{cont}}) \text{Multinomial}_{M,|w_k^{\text{cont}}|}(M|w_k^{\text{disc}}|) \prod_{j=1}^{d+1} 1\{\text{sign}(w_{k,j}^{\text{cont}}) = \text{sign}(w_{k,j}^{\text{disc}})\}. \tag{7.3}$$

This results in the marginal distribution for w^{disc} to treat each w_k^{disc} as uniform on $S_{1,M}^d$. This is a special case of the Dirichlet-Multinomial distribution using the all 1's vector in the parameter vector of the Dirichlet distribution [35, Chapter 6].

Lemma 12. *Consider the joint distribution outlined in expression* (7.3)*. The marginal distribution on w^{disc} treats each w_k^{disc} as uniform on $S_{1,M}^d$.*

Proof. The signs of the continuous vector coordinates are independent and equally likely to be ± 1, which is inherited by the discrete vector. The different k indexes are also clearly independent due to the product structure.

Focus then on the vectors of absolute values. Note the form of the Dirichlet distribution. For a vector of positive values $v_1, \cdots, v_{d+1}$ which sum to 1, the Dirichlet distribution with parameter vector α is written as

$$q_\alpha(v) = \frac{\Gamma(\sum_{j=1}^{d+1} \alpha_j)}{\prod_{j=1}^{d+1} \Gamma(\alpha_j)} \prod_{j=1}^{d+1} (v_j)^{\alpha_j - 1}. \tag{7.4}$$

Note the Gamma function is equal to factorial at integer values, $\Gamma(z) = (z-1)!, z \in \mathbb{N}$.

The absolute values of the continuous vector $|w_k^{\text{cont}}|$ are uniform on the simplex, which is also the symmetric Dirichlet distribution in $d+1$ dimensions with all ones parameter vector. Then, the marginal probability of the absolute values of the discrete vector is found by integrating out this Dirichlet distribution times the Multinomial distribution, which turns out to exactly cancel and give a constant value. This is a

special case of the Dirichlet-Multinomial distribution.

$$p_0(|w_k^{\text{disc}}|) = \int p_0(|w_k^{\text{cont}}|)p_0(|w_k^{\text{disc}}|\big||w_k^{\text{cont}}|)d|w_k^{\text{cont}}| \tag{7.5}$$

$$= \int \frac{\Gamma(d+1)}{\prod_{j=1}^{d+1}\Gamma(1)} \prod_{j=1}^{d+1}(|w_{k,j}^{\text{cont}}|)^{1-1} \frac{\Gamma(M+1)}{\prod_{j=1}^{d+1}\Gamma(M|w_{k,j}^{\text{disc}}|+1)} \prod_{j=1}^{d+1}|w_{k,j}^{\text{cont}}|^{(M|w_{k,j}^{\text{disc}}+1)-1} d|w_k^{\text{cont}}| \tag{7.6}$$

$$= \frac{(d!)(M!)}{(d+M)!} \int \frac{\Gamma(\sum_{j=1}^{d+1}(M|w_{k,j}^{\text{disc}}|+1))}{\prod_{j=1}^{d+1}\Gamma(M|w_{k,j}^{\text{disc}}|+1)} \prod_{j=1}^{d+1}|w_{k,j}^{\text{cont}}|^{(M|w_{k,j}^{\text{disc}}|+1)-1} d|w^{\text{cont}}| \tag{7.7}$$

$$= \frac{1}{\binom{M+d}{M}}. \tag{7.8}$$

The integral is equal to 1 as it represents the integral of a properly normalized Dirichlet distribution in $d+1$ dimensions using parameters $M|w_{k,j}^{\text{disc}}|+1$. ∎

Then we can relate expectations using either the continuous marginal or the discrete marginal as integrals with respect to the same joint distribution with one variable potentially marginalized out. The object which is used in our regret bound is the cumulant generating function of the loss function using the discrete vector,

$$-\log E_{P_0}[e^{-\frac{\beta}{2}\|Y-f_{w^{\text{disc}}}\|_{N+1}^2}]. \tag{7.9}$$

This object has been controlled in our previous proofs. The object we must understand is this same expression with the continuous vector in place of the discrete,

$$-\log E_{P_0}[e^{-\frac{\beta}{2}\|Y-f_{w^{\text{cont}}}\|_{N+1}^2}]. \tag{7.10}$$

If we can upper bound the continuous version by an expression using the discrete version and additional terms, we can upper bound continuous risk by an expression using discrete risk. We have the following upper bound.

Lemma 13. *Using the joint distribution defined above, the cumulant generating function using the continuous vector is less than twice the cumulant generating function using the discrete vector plus an additional term,*

$$-\log E_{P_0}[e^{-\frac{\beta}{2}\|Y-f_{w^{cont}}\|_{N+1}^2}] \tag{7.11}$$

$$\leq 2\left(-\log E_{P_0}[e^{-\frac{\beta}{2}\|Y-f_{w^{disc}}\|_{N+1}^2}]\right) + \log E_{P_0}[e^{-\frac{\beta}{2}\|Y-f_{w^{disc}}\|_{N+1}^2+\frac{\beta}{2}\|Y-f_{w^{cont}}\|_{N+1}^2}]. \tag{7.12}$$

Proof. We show that (7.11) minus (7.12) is less than 0. Collecting all log terms under one expression, (7.11) minus (7.12) is written as

$$-\log\left(\frac{E_{P_0}[e^{-\frac{\beta}{2}\|Y-f_{w^{\text{cont}}}\|_{N+1}}]E_{P_0}[e^{-\frac{\beta}{2}\|Y-f_{w^{\text{disc}}}\|^2_{N+1}+\frac{\beta}{2}\|Y-f_{w^{\text{cont}}}\|^2_{N+1}}]}{\left(E_{P_0}[e^{-\frac{\beta}{2}\|Y-f_{w^{\text{disc}}}\|^2_{N+1}}]\right)^2}\right) \tag{7.13}$$

Note the square in the denominator is due to the factor of 2 in (7.12). Distribute one of these factors in the denominator to each expectation in the numerator and separate into two log expressions,

$$-\log E_{P_0}[\frac{e^{-\frac{\beta}{2}\|Y-f_{w^{\text{cont}}}\|^2_{N+1}}}{E_{P_0}[e^{-\frac{\beta}{2}\|Y-f_{w^{\text{disc}}}\|^2_{N+1}}]}]-\log E_{P_0}[\frac{e^{-\frac{\beta}{2}\|Y-f_{w^{\text{disc}}}\|^2_{N+1}+\frac{\beta}{2}\|Y-f_{w^{\text{cont}}}\|^2_{N+1}}}{E_{P_0}[e^{-\frac{\beta}{2}\|Y-f_{w^{\text{disc}}}\|^2_{N+1}}]}] \tag{7.14}$$

We wish to consider the expectation in the denominators as the normalizing constant of a density. In the first expression, add and subtract $\frac{\beta}{2}\|Y-f_{w^{\text{disc}}}\|^2_{N+1}$ in the exponent. Then treat each term as an expectation using a properly normalized density,

$$-\log\int\frac{e^{-\frac{\beta}{2}\|Y-f_{w^{\text{disc}}}\|^2_{N+1}}}{E_{P_0}[e^{-\frac{\beta}{2}\|Y-f_{w^{\text{disc}}}\|^2_{N+1}}]}E_{P_0}[e^{-\frac{\beta}{2}\|Y-f_{w^{\text{cont}}}\|^2_{N+1}+\frac{\beta}{2}\|Y-f_{w^{\text{disc}}}\|^2_{N+1}}|w^{\text{disc}}]P_0(dw^{\text{disc}}) \tag{7.15}$$

$$-\log\int\frac{e^{-\frac{\beta}{2}\|Y-f_{w^{\text{disc}}}\|^2_{N+1}}}{E_{P_0}[e^{-\frac{\beta}{2}\|Y-f_{w^{\text{disc}}}\|^2_{N+1}}]}E_{P_0}[e^{\frac{\beta}{2}\|Y-f_{w^{\text{cont}}}\|^2_{N+1}}|w^{\text{disc}}]P_0(dw^{\text{disc}}). \tag{7.16}$$

Apply Jensen's inequality on each term twice to bring the negative log into the inner most expectation. This will bring the terms in the exponent down with a negative sign, so we have upper bound

$$\int\frac{e^{-\frac{\beta}{2}\|Y-f_{w^{\text{disc}}}\|^2_{N+1}}}{E_{P_0}[e^{-\frac{\beta}{2}\|Y-f_{w^{\text{disc}}}\|^2_{N+1}}]}E_{P_0}[\frac{\beta}{2}\|Y-f_{w^{\text{cont}}}\|^2_{N+1}-\frac{\beta}{2}\|Y-f_{w^{\text{disc}}}\|^2_{N+1}|w^{\text{disc}}]P_0(dw^{\text{disc}}) \tag{7.17}$$

$$+\int\frac{e^{-\frac{\beta}{2}\|Y-f_{w^{\text{disc}}}\|^2_{N+1}}}{E_{P_0}[e^{-\frac{\beta}{2}\|Y-f_{w^{\text{disc}}}\|^2_{N+1}}]}E_{P_0}[-\frac{\beta}{2}\|Y-f_{w^{\text{cont}}}\|^2_{N+1}|w^{\text{disc}}]P_0(dw^{\text{disc}}). \tag{7.18}$$

These expectations are then with respect to the same distribution, so we can collect into a common integral. The norms with $f_{w^{\text{cont}}}$ are of opposite sign and cancel, while the norm with $f_{w^{\text{disc}}}$ remains with a negative sign. Thus we have,

$$-\frac{\beta}{2}\int\frac{e^{-\frac{\beta}{2}\|Y-f_{w^{\text{disc}}}\|^2_{N+1}}}{E_{P_0}[e^{-\frac{\beta}{2}\|Y-f_{w^{\text{disc}}}\|^2_{N+1}}]}\|Y-f_{w^{\text{disc}}}\|^2_{N+1}P_0(dw^{\text{disc}})\leq 0. \tag{7.19}$$

Since the loss function is always non-negative, this expectation is always positive, and the negative in front makes it less than or equal to 0. ∎

For iid data, we can use this result to relate the square risk using the continuous prior to the risk using the discrete prior which we have already controlled. An unfortunate consequence of this method is it does not take into account the distance of the target g to its projection into the closure of the convex hull $\tilde{g}$. Thus our risk for the continuous prior will pay a price of twice the risk of the discrete prior, plus an additional $1/M$ term. Thus the minimum risk of the continuous prior can only be shown to be $2\|g - \tilde{g}\| + O([\log(d)/N)^{1/4}])$. For large data sets, this is twice the risk of the best estimator available, $\tilde{g}$, whereas for the discrete prior we have shown our Bayesian estimator is arbitrarily close to the best estimator. If g lives in the closure of the convex hull, this term is 0 and we perform arbitrarily close to the optimal estimator. Future work hopes to fix this gap.

First, we show the additional term in Lemma 13 has an expected value of $O(1/M)$ when the expectation is for iid data.

Lemma 14. *Let g be the target function bounded by b and assume the data distribution P_X is iid with data with support $[-1,1]^d$. Then we have the upper bound*

$$E_{P_{X^{N+1}}}[\log E_{P_0}[e^{-\frac{\beta}{2}\left(\|g-f_{w^{disc}}\|^2_{N+1}-\|g-f_{w^{cont}}\|^2_{N+1}\right)}]] \tag{7.20}$$

$$\leq 2a_2V(b+a_0V)\frac{\beta(N+1)}{M} \tag{7.21}$$

$$+\frac{(N+1)K}{2M}\left(e^{4a_1\beta\frac{V(b+a_0V)}{K}}-1-4a_1\beta\frac{V(b+a_0V)}{K}\right) \tag{7.22}$$

$$+\frac{(N+1)K}{M^2}\left(e^{4a_2\beta\frac{V(b+a_0V)}{K}}-1-4a_2\beta\frac{V(b+a_0V)}{K}\right) \tag{7.23}$$

Note for K of smaller order than N the first term (7.21) is dominant and this is approximately equal to $2a_2V(b+a_0V)\beta(N+1)/M$.

Proof. See Appendix Section 10.4 for full proof. The proof follows from a Taylor expansion to focus on quadratic terms in the exponent, independence in random variables to simplify the sum in the exponent into a product, and then a Bernstein inequality to control the individual moment generating functions at each index i and k. ∎

Combining these results, we can upper bound the continuous squared risk by twice the discrete squared risk plus an additional $O(\frac{1}{M})$ term.

Theorem 14. *Let $g(x)$ be a target function with absolute value bounded by b and let $\tilde{g}$ be its $L_2(P_X)$ projection into the closure of the convex hull of signed neurons scaled by V. Let P_0 be the uniform prior on the continuous set $(S_1^d)^K$. Assume the neuron*

activation function is odd symmetric and set all outer weights as $c_k = \frac{V}{K}$. *Let* $(x_i, y_i)_{i=1}^N$ *be training data iid with conditional mean* $g(x_i)$ *and conditional variance* $\sigma_{x_i}^2$ *with variance bound* $\sigma_{x_i}^2 \leq \sigma^2$. *Assume the data distribution* P_X *has support* $[-1,1]^d$. *Then statistical risk is upper bound by*

$$E[\|g - \hat{g}\|^2] \leq 2\frac{MK\log(2d+1)}{\beta(N+1)} + \frac{a_0^2V^2}{K} + \frac{3a_2V(a_0V+b) + 2V^2a_1^2}{M} \tag{7.24}$$

$$+ 4\beta(\frac{a_0V+b}{2})^2(\sigma + \frac{a_0V+b}{2})^2 + 2\|g - \tilde{g}\|^2 \tag{7.25}$$

$$+ O(\frac{1}{MK}) \tag{7.26}$$

Note with proper choice of parameters M, K, β *this can be shown to be of the order*

$$E[\|g - \hat{g}\|^2] \leq 2\|g - \tilde{g}\|^2 + O(\Big(\frac{\log(d)}{N}\Big)^{\frac{1}{4}})$$

Proof. This proof will follow much the same as the proof of Theorem 11. Note we are considering our posterior means as utilizing the continuous uniform prior in their definition, which is one of the marginals of the joint distribution we have defined for continuous and discrete values. Thus, we will write w^{cont} and w^{disc} inside the integrals to indicate which variable is arising in the expectation, even though all expectations with respect to P_0 are really joint integrals for both variables at the same time, with one potentially marginalized out.

The initial stages of the proof of Theorem 11 makes no explicit reference to the prior, so we can follow the same steps up to equation (6.93). We then apply Lemma 8 but explicitly note we are upper bounding using the continuous marginal of the prior, as this is the prior used to define the posterior means,

$$E_{P_{X^{N+1},Y^{N+1}}}[R_{N+1}^{\log}(X^{N+1}, Y^{N+1})] \leq -\frac{1}{2}\frac{1}{N+1}\sum_{n=0}^{N} E_{P_{X^{N+1},Y^{N+1}}}[(Y_{n+1} - g(X_{n+1}))^2] \tag{7.27}$$

$$+\frac{1}{\beta(N+1)}E_{P_{X^{N+1},Y^{N+1}}}[-\log E_{P_0}[e^{-\frac{\beta}{2}\sum_{n=0}^{N}(Y_{n+1} - f(X_{n+1}, w^{\text{cont}}))^2}]]. \tag{7.28}$$

We can then again upper bound by placing g in the exponent instead of Y, dropping the resulting linear term via a Jensen's inequality, which makes no explicit use of the form of the prior. This gives us the upper bound

$$\frac{E_{P_{X^{N+1}}}[-\log E_{P_0}[e^{-\frac{\beta}{2}(\|g - f_{w^{\text{cont}}}\|_{N+1}^2)}]]}{\beta(N+1)}. \tag{7.29}$$

By Lemma 13, we can upper bound this expression by twice its discrete counterpart plus an additional term,

$$2\Big(\frac{E_{P_{X^{N+1}}}[-\log E_{P_0}[e^{-\frac{\beta}{2}\|g-f_{w^{\text{disc}}}\|^2_{N+1}}]]}{\beta(N+1)}\Big) \tag{7.30}$$

$$+\frac{E_{P_{X^{N+1}}}[\log E_{P_0}[e^{-\frac{\beta}{2}\|g-f_{w^{\text{disc}}}\|^2_{N+1}+\frac{\beta}{2}\|g-f_{w^{\text{cont}}}\|^2_{N+1}}]]}{\beta(N+1)}. \tag{7.31}$$

Equation (7.30) is then twice the object we study in the remainder of Theorem 11 and thus inherits twice its final bound. This has the unfortunate effect of depending on twice the minimum achievable error $\|g-\tilde{g}\|^2$, meaning for large N we are not arbitrarily close to the projection. However, if the target function g does live in the closure of the Hull of signed neurons, this term is 0.

For the term (7.31), apply Lemma 14 to get a bound of the form

$$2\frac{a_2V(a_0V+b)}{M}+O(\frac{1}{MK}). \tag{7.32}$$

We have incorporated this term into the similar error term appearing in the analysis of the discrete object. This gives the factor of 3 in (7.24). ∎

8. Discussion

The use of an auxiliary random variable to create log-concavity is not a new idea, and has connections to existing methods. The critical structure of our sampling problem is that our target distribution of interest can be expressed as a mixture distribution with easy to sample components,

$$p_n(w)=\int p_n(w|\xi)p_n(\xi)d\xi. \tag{8.1}$$

The structure of a mixture distribution has been recognized in a number of recent papers. For spin glass systems (Sherrington–Kirkpatrick models) of high temperature, [8] expanded the range of known temperatures under which a Log Sobolev constant can be established by using such a mixture structure. For a Bayesian regression problem with a spike and slab (i.e. multi-modal) prior, [33] used the mixture structure to perform easy MCMC sampling. Thus, it is clear this approach of a mixture distribution can be applied to a number of sampling problems of interest. However, the posterior densities in these problems where much simpler than ours, making explicit use of the quadratic terms of their log-likelihoods which simplifies the analysis. Our view of a log-concave coupling as a mixture distribution applicable to any target distribution via a forward coupling is much more general.

Our method of creating the mixture is via forward coupling with a Gaussian auxiliary random variable ξ whose mean is determined by the target variable w. This has connections to proximal sampling algorithms and score based diffusion models. A proximal sampling algorithm would sample from the same joint distribution for $p(w, \xi)$ as we define here. However, the sampling method would be the Gibb's sampler alternating between sampling $p(w|\xi)$ and $p(\xi|w)$ which are both log-concave distributions [12], [23]. The mixing time of this sampling procedure must then be determined. If the original density of interest satisfies conditions such being Lipschitz and having a specified Log Sobolev constant, mixing time bounds can be established for the Gibb's sampler. It remains unclear what the mixing times bounds would be for a more difficult target density such as the one we study here. We instead explicitly examine the log-concavity of the induced marginal density $p_n(\xi)$ and propose to sample ξ from its marginal, followed by a sample of $w|\xi$ from its conditional.

We highlight that the score of the marginal density $\nabla \log p_n(\xi)$ 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 2. Thus, the score of the marginal can be computed as needed via its own MCMC sub-routine.

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 $\nabla \log p(w_t)$. 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 $p(w_t)$ 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 $p(w') = \int p(w'|w_t) p(w_t) dw_t$. Thus, the entire procedure of reverse diffusion can be avoided and a one shot sample of w_t from its marginal $p(w_t)$ 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.

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 likelihood 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 density is flat and has 0 Hessian, hence weakly log-concave. Nonetheless, [31] shows Ball Walk and Hit and Run algorithms mix in polynomial time for weakly log-concave densities on a convex set. Additionally, sampling $p(w|\xi)$ to compute expectations $E[w|\xi]$ may not always be required. Approximate methods such as accept-reject sampling [12, Section 4.2], importance sampling, or variational methods may be effective and faster than direct MCMC sampling to compute expectations $E[w|\xi]$ as needed. 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 [15]. Instead, to deal with the boundary conditions we must use techniques such as a barrier function [37] or other adaptations of sampling algorithms to restricted support such as Dikin Walks [27] and Hamiltonian Monte Carlo in a constrained space [26].

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, [10] shows that gradient descent with large step size converges quickly to an interpolating solution on the training data (i.e. 0 training loss). [38] 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 [16].

Another approach to understanding optimization in very large neural networks is to compare them to certain infinite width limits via the Neural Tangent Kernel [24]. With restrictions on the initialization distribution, at an infinite width limit the network is approximately linear around its initialization point with a fixed Gaussian kernel representation. Gradient methods quickly converge to a near interpolating solution. These methods generalize well for functions approximated by the large weight eigenfunctions of the kernel. The authors in [13] call this linearization of the network the "lazy regime" of training, and demonstrate models trained in this regime can have poor generalization, compared to models trained in the more difficult non-lazy regime. Our network

with the chosen scale of parameters adapts the directions of the internal weights to provide a more flexible span.

For very wide networks $K > N$, [29] 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 $K > N$ setting for general function learning.

There are also several negative results [14, 17, 19] showing that training a single-hidden-layer network to interpolation (0 training loss) is an NP hard problem. For example, [39] 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 $\zeta(Kd)^{-\frac{3}{2}}$ for an absolute constant ζ.

9. Conclusion and Future Work

In this work, we study a mixture form of the posterior density and statistical risk guarantees 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 is large enough that $Kd \geq C(\beta N)^2$ for a constant C where β is the inverse temperature and N is the number of data points. For a discrete uniform prior on the ℓ_1 ball (that is, restricted to a grid), we show notions of risk are on the order of $O([(\log d)/N]^{1/4})$. We extend these statistical risk control to the continuous prior as well, with a factor of 2. When the target function is itself in the closure of the convex hull of signed neurons, the continuous risk control is also of the same order as the discrete.

There are a number of future directions for research. The further details of sampling must 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 $p(\xi)$.

Finally, in the risk results we prove, we have assumed the V we use in defining our neural network matches the variation V of our target function. However, we would have no way of knowing what this value would be. In practice, we would place a prior on V from a finite set of possible values $(V_1, \cdots, V_{M_1})$. The log prior probability in the index of resolvability will have an additional M_1 in the numerator, for small M_1 relative to N this will not change the risk bounds much. We would also place a prior on a finite number of β and K values to consider multiple different models.

For each choice of hyperparameter V, β, K, we can demonstrate the mixture decomposition of $p(w)$ as studied in this paper. Therefore, we would run the sampling separately for all choices of V, β, K and get a different posterior mean for each choice (note this can easily be done on different machines simultaneously or on a GPU in parallel as there is no interaction between the different samplings at different hyperparameter choices). Our estimate would then be a weighted average of these different means. The weight of each choice of V, β, K would be the associated posterior probability, proportional to the normalizing constant (partition function) of $p(w)$ for that choice of V, β, K. These would have to be computed, which amounts to computing the partition function for a density we can sample from.

10. Appendix

10.1. 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 1:

Proof. We show that the set B is likely for conditionally independent Gaussian distributions for each variable. 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 Normal$(x_i \cdot w_k, 1/\rho)$ we may arrange a representation using independent standard normals Z_k of dimension n,

$$\xi_k = \mathbf{X} w_k + \frac{1}{\sqrt{\rho}} Z_k. \tag{10.1}$$

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.

$$P(\max_{j,k} |\sum_{i=1}^{n} x_{i,j}\xi_{i,k}| \geq n + \sqrt{2\log\frac{2Kd}{\delta}}\sqrt{\frac{n}{\rho}}). \tag{10.2}$$

The max is upper bound by

$$\max_{j,k} |\sum_{i=1}^{n} x_{i,j}\xi_{i,k}| \leq n + \max_{j,k} |\frac{1}{\sqrt{\rho}}\sum_{i=1}^{n} x_{i,j}Z_{i,k}|. \tag{10.3}$$

Thus we can bound the larger probability event,

$$P(\max_{j,k} \frac{|\sum_{i=1}^{n} x_{i,j}Z_{i,k}|}{\sqrt{n}} \geq \sqrt{2\log\frac{2Kd}{\delta}}) \leq \frac{\delta}{\sqrt{2\log(2Kd/\delta)}}. \tag{10.4}$$

Where the conclusion follows from a union bound and Gaussian tail bound.

■

Proof of Lemma 2:

Proof. We provide upper bounds on the magnitude of the first and second derivatives of the function $Z(w)$ as defined in equation (5.20). Denote Φ as the normal CDF and φ as the normal pdf. Throughout the proof recall that $p(w|\xi)$ treats each $\xi_{i,k}$ as independent normal with $\xi_{i,k} \sim \text{Normal}(x_i \cdot w_k, \frac{1}{\rho})$ conditionally independent given w. The gradient of $Z(w)$ inner product with a vector a with blocks a_k is

$$\left|a \cdot \nabla_w Z(w)\right| = \left|\rho E[\sum_{i=1}^{n}\sum_{k=1}^{K}(a_k \cdot x_i)(\xi_{i,k} - x_i \cdot w_k)\frac{1_B(\xi)}{P(\xi \in B|w)}|w]\right|. \tag{10.5}$$

By Lemma 1, the set B has probability at least $1 - \delta/\sqrt{2\log(2Kd/\delta)}$ when the $\xi_{i,k}$ are distributed according to their normal distribution at a fixed choice of w. We note the following upper and lower bounds on the Gaussian CDF provided by the classical results of Gordon [20], we have bounds on the Gaussian CDF

$$\frac{\varphi(x)}{x + \frac{1}{x}} \leq 1 - \Phi(x) \leq \frac{\varphi(x)}{x}. \tag{10.6}$$

Consider then the value

$$\delta^* = \Phi(-\sqrt{2\log(1/\delta)}). \tag{10.7}$$

For our problem, $Kd \geq 2$ by construction. Then for all positive $\delta \leq 1/e$, it can be shown that δ^* is larger than the term which defines the probability of our set B,

$$\frac{\delta}{\sqrt{2\log(2Kd/\delta)}} \leq \delta^*. \tag{10.8}$$

Then consider the collections of all measurable sets $\{D : D \subset \mathbb{R}^{NK}, P(\xi \in D) \geq 1 - \delta^*\}$. This contains our original set B as an object in the class. Then, the absolute value of the expected inner product in (10.5) is less than the maximum for any set D in this class,

$$\max_{\substack{D \\ P(\xi \in D|w) \geq 1-\delta}} \rho \frac{|E[\sum_{i=1}^{n} \sum_{k=1}^{K} (a_k \cdot x_i)(\xi_{i,k} - x_i \cdot w_k) 1_D(\xi)|w]|}{1 - \delta}. \tag{10.9}$$

Define the value

$$\tilde{\sigma} = \sqrt{\frac{\sum_{i=1}^{n} \sum_{k=1}^{K} (a_k \cdot x_i)^2}{\rho}}. \tag{10.10}$$

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

$$\sum_{i=1}^{n} \sum_{k=1}^{K} (a_k \cdot x_i)(\xi_{i,k} - x_i \cdot w_k) \sim \text{Normal}(0, \tilde{\sigma}^2). \tag{10.11}$$

The set D which maximizes expression (10.9) is then the set which controls the size of this integrand,

$$D^* = \{\xi : \frac{\sum_{i=1}^{n} \sum_{k=1}^{K} (a_k \cdot x_i)(\xi_{i,k} - x_i \cdot w_k)}{\tilde{\sigma}} \leq \tau\}, \tag{10.12}$$

for some choice of τ. We can also equally consider the set D^* where the object in the expression being more than some negative τ, due to symmetry. The proper choice of τ is $\sqrt{2\log(1/\delta)}$.We then have upper bound

$$|a \cdot \nabla_w Z(w)| \leq \frac{\rho\tilde{\sigma}}{1 - \delta} \left| \int_{-\infty}^{\sqrt{2\log(1/\delta)}} z\varphi(z)dz \right| = \frac{\rho\tilde{\sigma}\delta}{\sqrt{2\pi}1 - \delta}, \tag{10.13}$$

using the fact that $-z\varphi(z) = \varphi'(z)$ and fundamental theorem of calculus. This yields an upper bound on our expression of interest,

$$|a \cdot \nabla_w Z(w)| \leq \frac{\rho\tilde{\sigma}}{1 - \delta} \frac{\delta}{\sqrt{2\pi}}. \tag{10.14}$$

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

The Hessian is then a difference in variances,

$$a^{\mathrm{T}}[\nabla^2 Z(w)]a = -\rho \sum_{i=1}^{n} \sum_{k=1}^{K} (x_i \cdot a_k)^2 \tag{10.15}$$

$$+ \rho^2 \text{Var}[\sum_{i=1}^{n} \sum_{k=1}^{K} (x_i \cdot a_k)\xi_{i,k}|w, B]. \tag{10.16}$$

Note that $\xi_{i,k}$ is independent normal with variance $1/\rho$, so if we did not constrain the set B, expressions (10.15) and (10.16) would cancel to 0. That is, (10.15) is the variance of the linear function of ξ if we did not constrain to the set B, and (10.16) is the variance constrained to the set B.

Note that the object whose variance we are taking in (10.16) is a linear function of ξ, and ξ is a normal random variable with diagonal covariance matrix $\frac{1}{\rho}$. By an application of a Brascamp-Lieb inequality, see for example [9, Proposition 2.1], we would have an upper bound on this variance by the norm of this linear vector divided by ρ, which times ρ^2 is exactly expression (10.15). Thus, term (10.16) is less than or equal to the absolute value of term (10.15) so an upper bound on the quadratic form is 0, that is $a^{\mathrm{T}}[\nabla^2 Z(w)]a \le 0$.

We then compute a lower bound on the variance term in (10.16). Note a standard 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, we can write the variance as a non-centered expected squared difference minus a bias correction,

$$\mathrm{Var}[\sum_{i=1}^{n}\sum_{k=1}^{K}(x_i \cdot a_k)\xi_{i,k}|w, B] \tag{10.17}$$

$$= E[\Big(\sum_{i=1}^{n}\sum_{k=1}^{K}(x_i \cdot a_k)(\xi_{i,k} - x_i \cdot w_k)\Big)^2|w, \xi \in B] \tag{10.18}$$

$$- \Big(\sum_{i=1}^{n}\sum_{k=1}^{K}(a_k \cdot x_i)\Big(E[\xi_{i,k}|w, \xi \in B] - x_i \cdot w_k\Big)\Big)^2 \tag{10.19}$$

$$\ge E[\Big(\sum_{i=1}^{n}\sum_{k=1}^{K}(x_i \cdot a_k)(\xi_{i,k} - x_i \cdot w_k)\Big)^2 \frac{1_B(\xi)}{P(\xi \in B|w)}|w] \tag{10.20}$$

$$- \frac{\rho^2\tilde{\sigma}^2}{(1-\delta)^2}\frac{\delta^2}{2\pi}, \tag{10.21}$$

where we have applied the previously derived bound on the score to expression (10.19) to deduce expression (10.21), which is the square of the previous bond.

If we did not constrain to the set B, the expression (10.20) would be the variance of a simple normal variable with variance $\tilde{\sigma}^2$. We will show restricting to B still results in a value very close to $\tilde{\sigma}^2$.

The set B has probability at least $1 - \delta/\sqrt{2\log(2Kd/\delta)}$. Define the value

$$\delta^{**} = 2\Phi(-\sqrt{2\log(1/\delta)}). \tag{10.22}$$

If $Kd \geq 4$, for all positive $\delta \leq 1/16$ we have that δ^{**} is larger than the term which defines the set B probability,

$$\frac{\delta}{\sqrt{2\log(2Kd/\delta)}} \leq \delta^{**}.$$

Then, the expected value of the variable in question restricted to B is lower bound by the minimum for any set D with $P(\xi \in D) \geq 1 - \delta^{**}$,

$$E[\Big(\sum_{i=1}^{n}\sum_{k=1}^{K}(x_i \cdot a_k)(\xi_{i,k} - x_i \cdot w_k)\Big)^2 \frac{1_B(\xi)}{P(\xi \in B|w)}|w] \tag{10.23}$$

$$\geq \min_{\substack{D \\ P(\xi \in D|w) \geq 1-\delta^{**}}} \frac{E[\Big(\sum_{i=1}^{n}\sum_{k=1}^{K}(x_i \cdot a_k)(\xi_{i,k} - x_i \cdot w_k)\Big)^2 1_D(\xi)|w]}{1-\delta}. \tag{10.24}$$

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 the expression in question,

$$D^* = \{\xi : -\tau \leq \frac{\sum_{i=1}^{n}\sum_{k=1}^{K}(x_i \cdot a_k)(\xi_{i,k} - x_i \cdot w_k)}{\tilde{\sigma}} \leq \tau\}, \tag{10.25}$$

for some value τ, the proper choice being $\tau = \sqrt{2\log(1/\delta)}$.

Note this set D^* can be deduced from the Neyman-Pearson Lemma [28, Theorem 3.2.1], comparing the distribution where each $\xi_{i,k}$ is independent normal with mean $x_i \cdot w_k$ and variance $\frac{1}{\rho}$, to the distribution which has this normal density times $(\sum_{i=1}^{n}\sum_{k=1}^{K}(x_i \cdot a_k)(\xi_{i,k} - x_i \cdot w_k))^2$.

We are then integrating a squared normal on a truncated range and have lower bound,

$$\min_{\substack{D \\ P(\xi \in D|w) \geq 1-\delta}} \frac{E[\Big(\sum_{i=1}^{n}\sum_{k=1}^{K}(x_i \cdot a_k)(\xi_{i,k} - x_i \cdot w_k)\Big)^2 1_D(\xi)|w]}{1-\delta} \tag{10.26}$$

$$= \frac{\tilde{\sigma}^2}{1-\delta}\int_{-\sqrt{2\log(1/\delta)}}^{\sqrt{2\log(1/\delta)}} z^2\varphi(z)dz. \tag{10.27}$$

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

$$\int_{-\sqrt{2\log(1/\delta)}}^{\sqrt{2\log(1/\delta)}} z^2\varphi(z)dz = 1 - 2\int_{-\infty}^{-\sqrt{2\log(1/\delta)}} z^2\varphi(z)dz. \tag{10.28}$$

Then apply integration by parts,

$$-\int_{-\infty}^{-\sqrt{2\log(1/\delta)}} z^2\varphi(z)dz = z\varphi(z)|_{-\infty}^{-\sqrt{2\log(1/\delta)}} - \Phi(-\sqrt{2\log(1/\delta)}). \tag{10.29}$$

This gives a lower bound for expression for (10.27)

$$\frac{\tilde{\sigma}^2}{1-\delta}\Big(1-\frac{2\delta}{\sqrt{2\pi}}(\sqrt{2\log(1/\delta)}+\frac{1}{\sqrt{2\log(1/\delta)}})\Big), \tag{10.30}$$

which converges to $\tilde{\sigma}^2$ as $\delta \to 0$. We then combine expression (10.16), (10.21), and (10.30) to give a lower bound on Hessian quadratic form,

$$a^{\mathrm{T}}[\nabla^2 Z(w)]a \geq -\rho^2\tilde{\sigma}^2+\rho^2\tilde{\sigma}^2(\frac{1}{1-\delta}-\frac{2\delta}{(1-\delta)\sqrt{2\pi}}(\sqrt{2\log(1/\delta)}+\frac{1}{\sqrt{2\log(1/\delta)}})) \tag{10.31}$$

$$-\frac{\rho^4\tilde{\sigma}^2}{(1-\delta)^2}\frac{\delta^2}{2\pi} \tag{10.32}$$

$$=-\frac{\rho^2\tilde{\sigma}^2}{\sqrt{2\pi}}\frac{\delta}{1-\delta}\Big(-\sqrt{2\pi}+2\sqrt{2\log(1/\delta)}(1+\frac{1}{2\log(1/\delta)})+\frac{\rho^2\tilde{\sigma}^2}{\sqrt{2\pi}}\frac{\delta}{1-\delta}\Big) \tag{10.33}$$

$$\geq -\frac{\rho^2\tilde{\sigma}^2}{\sqrt{2\pi}}\frac{\delta}{1-\delta}\Big(2\sqrt{2\log(1/\delta)}+\frac{\rho^2\tilde{\sigma}^2}{\sqrt{2\pi}}\frac{\delta}{1-\delta}\Big) \tag{10.34}$$

which converges to 0 as $\delta \to 0$. ∎

10.2. Log-Concavity of $p(w|\xi)$ with Set B Restriction

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

Proof of Theorem 7

Proof. 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 6. The log likelihood for $p_n^*(w|\xi)$ is given by

$$\log p_n^*(w|\xi) = -\beta\ell_n(w)+H(\xi) \tag{10.35}$$

$$-\sum_{i=1}^{n}\sum_{k=1}^{K}\frac{\rho}{2}(\xi_{i,k}-w_k\cdot x_i)^2 \tag{10.36}$$

$$-Z(w), \tag{10.37}$$

for some function $H(\xi)$ which does not depend on w and is only required to make the density integrate to 1. The term (10.36) 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 $\rho\sum_{i=1}^{n}x_ix_i^{\mathrm{T}}$. Denote the Hessian as $H_n(w|\xi)\equiv\nabla^2\log p_n^*(w|\xi)$.

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

$$-\beta \sum_{i=1}^{n} \Big(\sum_{k=1}^{K} \psi'(w_k \cdot x_i) a_k \cdot x_i \Big)^2 \tag{10.38}$$

$$+ \sum_{k=1}^{K} \sum_{i=1}^{n} (a_k \cdot x_i)^2 \Big[\beta \mathrm{res}_i(w) c_k \psi''(w_k \cdot x_i) - \rho) \Big] \tag{10.39}$$

$$+ a^{\mathrm{T}} (\nabla^2 Z(w)) a. \tag{10.40}$$

By the assumptions on the second derivative of ψ and the definition of ρ in equation (5.17) we have

$$\max_{i,k} (\beta \mathrm{res}_i(w) c_k \psi''(w_k \cdot x_i) - \rho) \leq -(\sqrt{\frac{3}{2}} - 1) a_2 \frac{\beta C_n V}{K}, \tag{10.41}$$

so all the terms in the sum in (10.39) are negative. Recall the definition of $\tilde{\sigma}^2$,

$$\tilde{\sigma}^2 = \frac{\sum_{k=1}^{K} \sum_{i=1}^{n} (a_k \cdot x_i)^2}{\rho}. \tag{10.42}$$

Therefore, expression (10.39) is less than

$$-(\sqrt{\frac{3}{2}} - 1) \sqrt{\frac{3}{2}} \Big(a_2 \frac{\beta C_n V}{K} \Big)^2 \tilde{\sigma}^2. \tag{10.43}$$

By Lemma 2, the largest the Hessian term from the correction function Z can be is

$$a^{\mathrm{T}} (\nabla^2 Z(w)) a \leq \frac{\rho^2 \tilde{\sigma}^2}{\sqrt{2\pi}} \frac{\delta}{1-\delta} \Big(2\sqrt{2 \log(1/\delta)} + \frac{\rho^2 \tilde{\sigma}^2}{\sqrt{2\pi}} \frac{\delta}{1-\delta} \Big). \tag{10.44}$$

Thus term (10.39) plus (10.40) is less than

$$-\tilde{\sigma}^2 \Big(a_2 \frac{\beta C_n V}{K} \Big)^2 \Big(\sqrt{\frac{3}{2}} - 1 \Big) \Big(\sqrt{\frac{3}{2}} \Big) \tag{10.45}$$

$$+\tilde{\sigma}^2 \Big(a_2 \frac{\beta C_n V}{K} \Big)^2 \Big(\sqrt{\frac{3}{2}} \Big)^2 \frac{2}{\sqrt{2\pi}} \frac{\delta}{1-\delta} \sqrt{2 \log \frac{\delta}{2}} \tag{10.46}$$

$$+\tilde{\sigma}^2 \Big(a_2 \frac{\beta C_n V}{K} \Big)^4 \Big(\sqrt{\frac{3}{2}} \Big)^4 \frac{1}{2\pi} \frac{\delta^2}{(1-\delta)^2}. \tag{10.47}$$

Recall the definitions of H_1 and H_2 in the theorem statement,

$$H_1(\delta) = \frac{2}{\sqrt{2\pi}} \frac{\delta}{1-\delta} \sqrt{2 \log \frac{\delta}{2}} \tag{10.48}$$

$$H_2(\delta) = \Big(a_2 \frac{\beta C_n V}{K} \Big)^2 \frac{1}{2\pi} \frac{\delta^2}{(1-\delta)^2}. \tag{10.49}$$

Simplifying expressions (10.45) to (10.47) by dividing out common terms, to have a negative expression for the Hessian we require,

$$\sqrt{\frac{3}{2}}(-1 + H_1(\delta)) + \left(\sqrt{\frac{3}{2}}\right)^3 H_2(\delta) \leq -1. \tag{10.50}$$

By assumption $H_1(\delta) \leq \frac{1}{100}, H_2(\delta) \leq \frac{1}{10}$ satisfies the inequality,

$$\sqrt{\frac{3}{2}}(-1 + H_1(\delta)) + \left(\sqrt{\frac{3}{2}}\right)^3 H_2(\delta) \leq \sqrt{\frac{3}{2}}(-\frac{99}{100}) + \left(\sqrt{\frac{3}{2}}\right)^3 \frac{1}{10} \tag{10.51}$$

$$= -\frac{21}{25}\sqrt{\frac{3}{2}} < -1. \tag{10.52}$$

■

10.3. Hölder Inequality Proofs

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

Lemma 15. *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 bound by,*

$$E_{P_0}[(\sum_{j=1}^{d} x_j w_j)^{2\ell}] \leq \frac{1}{(d)^\ell} \frac{(2\ell)!}{\ell!}. \tag{10.53}$$

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

$$E[\Big(\sum_{j=1}^{d} x_j w_j\Big)^{2\ell}] = \sum_{j_1, \cdots, j_{2\ell}} \prod_{i=1}^{2\ell} (x_{j_i}) E[\prod_{i=1}^{2\ell} w_{j_i}]. \tag{10.54}$$

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,

$$\prod_{i=1}^{2\ell} w^{j_i} = \prod_{j=1}^{d} w_j^{r(j,J)}. \tag{10.55}$$

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 $\vec{r} = (r_1, \cdots, r_d)$ 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 $(|w_1|, \cdots, |w_d|, w_{d+1})$ has a symmetric Dirichlet $(1, \cdots, 1)$ distribution in $d+1$ dimensions. Note a general Dirichlet distribution in $d+1$ dimensions with parameter vector $\vec{\alpha} = (\alpha_1, \cdots, \alpha_{d+1})$ has a properly normalized density as

$$p_{\vec{\alpha}}(w_1, \cdots, w_d) = \frac{\Gamma(\sum_{j=1}^{d} \alpha_j)}{\prod_{j=1}^{d+1} \Gamma(\alpha_j)} \prod_{j=1}^{d} (w_j)^{\alpha_j - 1} (1 - \sum_{j=1}^{d} w_j)^{\alpha_{d+1} - 1}. \tag{10.56}$$

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

$$E[\prod_{j=1}^{d} w_j^{r_j}] = \frac{\Gamma(d+1) \prod_{j=1}^{d} \Gamma(r_j + 1)}{\Gamma(d + 1 + \sum_{j=1}^{d} r_j)} \tag{10.57}$$

$$= \frac{d! \prod_{j=1}^{d} r_j!}{(d + 2\ell)!}. \tag{10.58}$$

The number of times a specific vector $\vec{r}$ appears from the multi-index J is $\frac{(2\ell)!}{\prod_{j=1}^{d} r_j!}$ thus we have,

$$E[(\sum_{j=1}^{d} x_j w_j)^{2\ell}] = \sum_{\substack{\vec{r} \text{ even} \\ \sum_j r_j = 2\ell}} \prod_{j=1}^{d} (x_j)^{r_j} \frac{(2\ell)!}{\prod_{j=1}^{d} r_j!} E[\prod_{j=1}^{d} w_j^{r_j}] \tag{10.59}$$

$$= \frac{(2\ell!)(d!)}{(d + 2\ell)!} \sum_{\substack{\vec{r} \text{ even} \\ \sum_j r_j = 2\ell}} \prod_{j=1}^{d} (x_j)^{r_j} \tag{10.60}$$

$$= \frac{(2\ell!)(d!)}{(d + 2\ell)!} \sum_{\substack{\vec{r} \text{ even} \\ \sum_j r_j = 2\ell}} \prod_{j=1}^{d} (x_j^2)^{\frac{r_j}{2}} \tag{10.61}$$

$$\leq \frac{(2\ell!)(d!)}{(d + 2\ell)!} \frac{(d + \ell - 1)!}{\ell!(d-1)!} \tag{10.62}$$

$$= \frac{(d + \ell - 1) \cdots (d)}{(d + 2\ell) \cdots (d+1)} \frac{(2\ell)!}{(\ell)!} \tag{10.63}$$

$$\leq \frac{1}{d^\ell} \frac{2\ell!}{\ell!}, \tag{10.64}$$

where inequality (10.62) 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 5:

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

$$E[(\sum_{k=1}^{K} v_k \cdot w_k)^{2\ell}]. \tag{10.65}$$

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

$$E[\sum_{\substack{j_1,\cdots,j_K \\ \sum j_k = 2\ell}} \binom{2\ell}{j_1 \cdots j_K} \prod_{k=1}^{K} (v_k \cdot w_k)^{j_k}] = \sum_{\substack{j_1,\cdots,j_K \\ \sum j_k = 2\ell}} \binom{2\ell}{j_1 \cdots j_K} \prod_{k=1}^{K} E[(v_k \cdot w_k)^{j_k}], \tag{10.66}$$

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 the whole expression is 0 thus we only sum using even j_k values,

$$\sum_{\substack{j_1,\cdots,j_K \\ \sum j_k = \ell}} \binom{2\ell}{2j_1 \cdots 2j_K} \prod_{k=1}^{K} E[(v_k \cdot w_k)^{2j_k}]. \tag{10.67}$$

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

$$v_k = \sum_{i=1}^{n} a_{k,i} x_i. \tag{10.68}$$

Define $s_{k,i} = \mathrm{sign}(a_{k,i})$ and $\alpha_{k,i} = \frac{|a_{k,i}|}{\|a_k\|_1}$. We can then interpret the above inner product as a scaled expectation on the data indexes,

$$v_k \cdot w_k = (\|a_k\|_1) \sum_{i=1}^{n} \alpha_{k,i} s_{k,i}\, x_i \cdot w_k. \tag{10.69}$$

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

$$E[(v_k \cdot w_k)^{2j_k}] = (\|a_k\|_1)^{2j_k} E[\Big(\sum_{i=1}^{n} \alpha_{k,i} s_{k,i}\, x_i \cdot w_k\Big)^{2j_k}] \tag{10.70}$$

$$\leq (\|a_k\|_1)^{2j_k} \sum_{i=1}^{n} \alpha_{k,i} E[\Big(x_i \cdot w_k\Big)^{2j_k}] \tag{10.71}$$

$$\leq (\|a_k\|_1)^{2j_k} \max_i E[(x_i \cdot w_k)^{2j_k}] \tag{10.72}$$

$$\leq (\|a_k\|_1)^{2j_k} \frac{1}{(d)^{j_k}} \frac{(2j_k)!}{j_k!}, \tag{10.73}$$

where we have applied Lemma 15. We then plug this result into equation (10.67),

$$\frac{1}{d^\ell}\frac{(2\ell!)}{\ell!}\Big(\sum_{\substack{j_1,\cdots,j_K\\ \sum j_k=\ell}}\binom{\ell}{j_1\cdots j_K}\prod_{k=1}^{K}(\|a_k\|_1)^{2j_k}\Big)=\frac{1}{d^\ell}\frac{(2\ell)!}{\ell!}\Big(\sum_{k=1}^{K}\|a_k\|_1^2\Big)^\ell. \tag{10.74}$$

For each sub block a_k of dimension n we have $\|a_k\|_1^2 \le n\|a_k\|_2^2$ and $\|a\|^2=\sum_{k=1}^K\|a_k\|^2=1$ is a unit vector which gives upper bound

$$\frac{n^\ell(2\ell)!}{d^\ell\ell!}. \tag{10.75}$$

Via Stirling's bound [36],

$$\sqrt{2\pi\ell}(\frac{\ell}{e})^\ell e^{\frac{1}{12\ell+1}}\le\ell!\le\sqrt{2\pi\ell}(\frac{\ell}{e})^\ell e^{\frac{1}{12\ell}}. \tag{10.76}$$

Taking the ℓ root we have

$$\Big(\frac{n^\ell}{d^\ell}\frac{(2\ell)!}{\ell!}\Big)^{\frac{1}{\ell}}\le\frac{n}{d}\Big(2^{2\ell+\frac{1}{2}}(\frac{\ell}{e})^\ell e^{\frac{1}{24\ell}-\frac{1}{12\ell+1}}\Big)^{\frac{1}{\ell}} \tag{10.77}$$

$$=\frac{2^{2+\frac{1}{2\ell}}n\ell}{d}e^{\frac{1}{24\ell^2}-\frac{1}{12\ell^2+\ell}-1} \tag{10.78}$$

$$\le\frac{4n\ell}{d}\sqrt{2}e^{\frac{1}{24}+\frac{1}{13}-1} \tag{10.79}$$

$$\le\frac{4n\ell}{\sqrt{e}d}. \tag{10.80}$$

■

Proof of Lemma 6:

Proof. 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 $\tilde{\tau}\in[1,\frac{\ell}{\ell-1}]$ such that

$$\Gamma_\xi^n(\frac{\ell}{\ell-1})=\Gamma_\xi^n(1)+(\Gamma_\xi^n)'(\tilde{\tau})[\frac{\ell}{\ell-1}-1]. \tag{10.81}$$

Rearranging, we can express the difference

$$\frac{\ell-1}{\ell}\Gamma_\xi^n(\frac{\ell}{\ell-1})-\Gamma_\xi^n(1)=(\Gamma_\xi^n)'(\tilde{\tau})\frac{1}{\ell}-\frac{1}{\ell}\Gamma_\xi^n(1). \tag{10.82}$$

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

$$\frac{\ell-1}{\ell}\Gamma_\xi^n(\frac{\ell}{\ell-1})-\Gamma_\xi^n(1)\le(\Gamma_\xi^n)'(\tilde{\tau})\frac{1}{\ell}. \tag{10.83}$$

Recall $\Gamma^n_\xi(\tau)$ defined in equation (5.53) is a cumulant generating function of $\tilde{h}^n_\xi(w)$. Thus, its derivative at $\tilde{\tau}$ is the mean of $\tilde{h}^n_\xi(w)$ under the tilted distribution. The mean is then less than the maximum difference of any two points on the constrained support set,

$$(\Gamma^n_\xi)'(\tilde{\tau}) = E_{\tilde{\tau}}[\tilde{h}^n_\xi(w)|\xi] \leq \max_{w,w_0 \in (S_1^d)^K} (\tilde{h}^n_\xi(w) - \tilde{h}^n_\xi(w_0)). \tag{10.84}$$

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

$$\tilde{h}^n_\xi(w) - \tilde{h}^n_\xi(w_0) = \nabla_w \tilde{h}^n_\xi(\tilde{w}) \cdot (w - w_0). \tag{10.85}$$

For each k, the gradient in w_k is

$$\nabla_{w_k} \tilde{h}^n_\xi(\tilde{w}) = \beta \sum_{i=1}^n (\text{res}_i(\tilde{w}) c_k \psi'(w_k \cdot x_i) + a_2 \sqrt{\frac{3}{2}} \frac{C_n V}{K}) x_i \tag{10.86}$$

$$+ a_2 \sqrt{\frac{3}{2}} \frac{\beta C_n V}{K} \sum_{i=1}^n \xi_{i,k} x_i + \nabla_{w_k} Z(w). \tag{10.87}$$

The scalar terms in the sum in (10.86) satisfy

$$|\text{res}_i(\tilde{w}) c_k \psi'(w_k \cdot x_i) + a_2 \sqrt{\frac{3}{2}} \frac{C_n V}{K}| \leq (a_1 + a_2 \sqrt{\frac{3}{2}}) \frac{C_n V}{K}, \tag{10.88}$$

for each i. The vector $w_k - w_{0,k}$ satisfies $\|w_k - w_{0,k}\|_1 \leq 2$. Since each x_i vector has bounded entries between -1 and 1, the inner product with the first term is bounded as

$$\Big[\beta \sum_{i=1}^n (\text{res}_i(\tilde{w}) c_k \psi'(w_k \cdot x_i) - \frac{C_n V}{K}) x_i\Big] \cdot (w_k - w_{0,k}) \leq 2\Big(a_1 + a_2 \sqrt{\frac{3}{2}}\Big) \frac{C_n V \beta n}{K}. \tag{10.89}$$

As for the second term,

$$\Big[\sum_{i=1}^n \xi_{i,k} x_i\Big] \cdot (w_k - w_{0,k}) \leq 2 \max_j |\sum_{i=1}^n \xi_{i,k} x_{i,j}|. \tag{10.90}$$

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,

$$\max_j |\sum_{i=1}^n \xi_{i,k} x_{i,j}| \leq n + \sqrt{2 \log(\frac{2Kd}{\delta})} \sqrt{\frac{n}{\rho}} \tag{10.91}$$

$$= n + \sqrt{2 \log \frac{2Kd}{\delta}} \sqrt{\sqrt{\frac{2}{3}} \frac{nK}{a_2 \beta C_n V}}. \tag{10.92}$$

For the final term, $Z(w)$ is shown to have small derivative. By Lemma 2,

$$\sum_k \nabla_{w_k} Z(w) \cdot (w_k - w_{0,k}) \leq \sqrt{\rho} \sqrt{\sum_{i=1}^{n} \sum_{k=1}^{K} ((w_k - w_{0,k}) \cdot x_i)^2} \frac{1}{(1-\delta)} \frac{\delta}{\sqrt{2\pi}} \tag{10.93}$$

$$\leq \sqrt{4a_2 \sqrt{\frac{3}{2}} C_n V \beta n} \frac{\delta}{\sqrt{2\pi}} \frac{1}{1-\delta}. \tag{10.94}$$

Summing using index k for terms (10.89), (10.92) and combining with term (10.94), we can upper bound the difference in the CGF as,

$$2\Big(a_1 + a_2\sqrt{\frac{3}{2}}\Big)\frac{C_n V \beta n}{\ell} + a_2\sqrt{\frac{3}{2}}\frac{\beta C_n V}{\ell}\Big(n + \sqrt{2\log\frac{2Kd}{\delta}}\sqrt{\sqrt{\frac{2}{3}}\frac{nK}{a_2 \beta C_n V}}\Big) \tag{10.95}$$

$$+\sqrt{4a_2\sqrt{\frac{3}{2}}C_n V\beta n}\frac{\delta}{\sqrt{2\pi}}\frac{1}{1-\delta} \tag{10.96}$$

$$= \frac{C_n V \beta n}{\ell}(2a_1 + 4a_2\sqrt{\frac{3}{2}}) + \frac{\sqrt{C_n V \beta n}}{\ell}\sqrt{2a_2\sqrt{\frac{3}{2}}}\Big(\sqrt{\log\frac{2Kd}{\delta}}\sqrt{K} + \frac{\delta}{\sqrt{\pi}(1-\delta)}\Big). \tag{10.97}$$

By assumption $d \geq 2, K \geq 2, \delta \leq \frac{1}{16}$. For all values $0 < z \leq \frac{1}{2}$ we have the inequality

$$\frac{z}{\sqrt{\pi}(1-z)} \leq \frac{1}{\sqrt{\pi}}\sqrt{\log\frac{2}{z}} \leq \frac{1}{\sqrt{\pi}}\sqrt{\log\frac{2Kd}{z}}\sqrt{K}. \tag{10.98}$$

This gives the final upper bound

$$\frac{C_n V \beta n}{\ell}(2a_1 + 4a_2\sqrt{\frac{3}{2}}) + \frac{\sqrt{C_n V \beta n}}{\ell}(1 + \frac{1}{\sqrt{\pi}})\sqrt{2a_2\sqrt{\frac{3}{2}}}\Big(\sqrt{\log\frac{2Kd}{\delta}}\sqrt{K}\Big). \tag{10.99}$$

■

10.4. Bounding Additional Continuous Risk Term Lemma 14

Proof. Bring the log outside the outer expectation to provide an upper bound, since log is a concave function,

$$E_{P_{X^{N+1}}}[\log E_{P_0}[e^{-\frac{\beta}{2}\left(\|g - f_{w^{\text{disc}}}\|^2_{N+1} - \|g - f_{w^{\text{cont}}}\|^2_{N+1}\right)}]] \tag{10.100}$$

$$\leq \log E_{P_{X^{N+1}}}[E_{P_0}[e^{-\frac{\beta}{2}\left(\|g - f_{w^{\text{disc}}}\|^2_{N+1} - \|g - f_{w^{\text{cont}}}\|^2_{N+1}\right)}]] \tag{10.101}$$

Then recall the Taylor expansion for $\|g - f_{w^{\text{disc}}}\|^2_{N+1}$ centered at w^{cont}, using some vector $\tilde{w}$ in the second derivative terms. Note the outer weights of the network c_k are $\pm\frac{V}{K}$, which will will write as $c_k = s_k\frac{V}{K}$ using a sign $s_k \in \{-1, 1\}$. Define

$$\text{res}_i(w) = g(x_i) - \sum_{k=1}^{K} s_k \frac{V}{K}\psi(x_i \cdot w_k) \tag{10.102}$$

$$a_{i,k} = -s_k\frac{2V}{K}\text{res}_i(w^{\text{cont}}, s)\psi'(x_i \cdot w_k^{\text{cont}}) \tag{10.103}$$

$$\begin{aligned} b_{i,k,k'}(\tilde{w}) = &-s_k\frac{2V}{K}\text{res}_i(\tilde{w})\psi''(x_i \cdot \tilde{w}_k)\delta_{k=k'} \\ &+ 2s_k s_{k'}\frac{V^2}{K^2}\psi'(x_i \cdot \tilde{w}_k)\psi'(x_i \cdot \tilde{w}_{k'}). \end{aligned} \tag{10.104}$$

Then for any continuous valued vector w^{cont} and discrete valued vector w^{disc}, there exists some vector $\tilde{w}$ (in fact along the line between w^{disc} and w^{cont}) such that the second order expansion is exact using that $\tilde{w}$ in the second derivative terms,

$$-\frac{\beta}{2}\Big(\|g - f(\cdot, w^{\text{disc}}, s)\|^2_{N+1} - \|g - f(\cdot, w^{\text{cont}}, s)\|^2_{N+1}\Big) \tag{10.105}$$

$$= -\frac{\beta}{2}\sum_{i=1}^{N+1}\sum_{k=1}^{K} a_{i,k}x_i \cdot (w_k^{\text{disc}} - w_k^{\text{cont}}) \tag{10.106}$$

$$+\frac{\beta}{2}\frac{V}{K}\sum_{i=1}^{N+1}\sum_{k=1}^{K} s_k\text{res}_i(\tilde{w}, s)\psi''(x_i \cdot \tilde{w}_k)(x_i \cdot (w_k^{\text{disc}} - w_k^{\text{cont}}))^2 \tag{10.107}$$

$$-\frac{\beta}{2}\sum_{i=1}^{N+1}\Big(\sum_{k=1}^{K} s_k\frac{V}{K}\psi'(\tilde{w}_k)(x_i \cdot (w_k^{\text{disc}} - w_k^{\text{cont}}))\Big)^2. \tag{10.108}$$

This last term (10.108) is always negative, so we can upper bound by ignoring this term. Then each $|\text{res}_i(\tilde{w})\psi''(x_i \cdot \tilde{w}_k)| \le (b + a_0V)a_2$ by assumptions on bounded g and the activation function. This gives upper bound on the difference in loss functions,

$$-\frac{\beta}{2}\Big(\|g - f(\cdot, w^{\text{disc}}, s)\|^2_{N+1} - \|g - f(\cdot, w^{\text{cont}}, s)\|^2_{N+1}\Big) \tag{10.109}$$

$$\le -\frac{\beta}{2}\sum_{i=1}^{N+1}\sum_{k=1}^{K} a_{i,k}x_i \cdot (w_k^{\text{disc}} - w_k^{\text{cont}}) \tag{10.110}$$

$$+\frac{\beta}{2}\frac{V}{K}(b + a_0V)a_2\sum_{i=1}^{N+1}\sum_{k=1}^{K}(x_i \cdot (w_k^{\text{disc}} - w_k^{\text{cont}}))^2 \tag{10.111}$$

Note the $a_{i,k}$ are functions of the continuous vectors, not the discrete. Switch the order of the expectations to have the outer expectation be with respect to w^{cont}, and the inner

expectation with respect to w^{disc} and X^{N+1},

$$\log E_{P_0}[E_{P_{X^{N+1}}}[E_{P_0}[e^{\sum_{i,k} -\frac{\beta}{2}a_{i,k}x_i\cdot(w_k^{\text{disc}}-w_k^{\text{cont}})+\frac{\beta}{2}\frac{a_2V(b+a_0V)}{K}(x_i\cdot(w_k^{\text{disc}}-w_k^{\text{cont}}))^2}|w^{\text{cont}}]]]. \tag{10.112}$$

Under the prior, the w_k^{disc} are conditionally independent given w^{cont}, and by assumption of iid data the data distribution P_X treats each index i as independent. The inner expectation can be written as a product of individual moment generating functions for each i and k, conditioned on w^{cont}.

$$\log E_{P_0}[\prod_{i,k} E_{P_X}[E_{P_0}[e^{-\frac{\beta}{2}a_{i,k}x_i\cdot(w_k^{\text{disc}}-w_k^{\text{cont}})+\frac{\beta}{2}\frac{a_2V(b+a_0V)}{K}(x_i\cdot(w_k^{\text{disc}}-w_k^{\text{cont}}))^2}|w^{\text{cont}}]]]. \tag{10.113}$$

For simplicity and to avoid odd power terms, use a Cauchy Schwartz inequality to upper bound the cumulant generating function of the linear and quadratic terms with two separate expectations with a factor of 2

$$E_{P_0}[e^{-\frac{\beta}{2}a_{i,k}x_i\cdot(w_k^{\text{disc}}-w_k^{\text{cont}})+\frac{\beta}{2}\frac{a_2V(b+a_0V)}{K}(x_i\cdot(w_k^{\text{disc}}-w_k^{\text{cont}}))^2}|w_k^{\text{cont}}] \tag{10.114}$$

$$\leq\left(E_{P_0}[e^{-\beta a_{i,k}x_i\cdot(w_k^{\text{disc}}-w_k^{\text{cont}})}|w_k^{\text{cont}}]E_{P_0}[e^{\beta\frac{a_2V(b+a_0V)}{K}(x_i\cdot(w_k^{\text{disc}}-w_k^{\text{cont}}))^2}|w_k^{\text{cont}}]\right)^{\frac{1}{2}} \tag{10.115}$$

We then can upper bound these moment generating functions with a Bernstein inequality using the first, second, and fourth conditional moments of $x_i\cdot(w_k^{\text{disc}}-w_k^{\text{cont}})$ conditioned on w_k^{cont}. This bound is useful for small inputs to the moment generating function, and can be considered a concise statement of a sub-exponential random variable. This utilizes the Multinomial conditional distribution which defines our prior. Note that this random variable is bounded by 2, and is mean 0. Thus with m_2 being its second moment and m_4 its fourth moment we have upper bound via Bernstein inequality [40, Lemma 7.26] for any scaling t,

$$E_{P_0}[e^{tx_i\cdot(w_k^{\text{disc}}-w_k^{\text{cont}})}|w_k^{\text{cont}}]\leq\exp\left(t^2m_2\frac{e^{2t}-1-2t}{4t^2}\right) \tag{10.116}$$

$$E_{P_0}[e^{t(x_i\cdot(w_k^{\text{disc}}-w_k^{\text{cont}}))^2}|w_k^{\text{cont}}]\leq\exp\left(t^2(m_4-m_2^2)\frac{e^{4t}-1-4t}{16t^2}\right)e^{tm_2}. \tag{10.117}$$

Note that the Multinomial distribution inner product with the vector x_i can be considered as a normalized sum of M iid bounded random variables. Consider a random index $J\in\{1,\cdots d+1\}$ where $J=j$ with probability $|w_{k,j}^{\text{cont}}|$. Given w_k^{cont}, this defines a distribution on $\{1,\cdots,d+1\}$. Draw M iid random indices $J_1,\cdots,J_M$ from this distribution. We can construct the Multinomial distribution with these random index

selections and write

$$x_i \cdot (w_k^{\text{disc}} - w_k^{\text{cont}}) = \frac{1}{M} \sum_{t=1}^{M} (x_{i,J_t} - E[x_{i,J_t}]) \tag{10.118}$$

This is then an average of M mean 0 iid random variables each bounded by 2. One can bound the moments by expanding the expectations and some algebra,

$$m_1 = E_{P_0}[x_i \cdot (w_k^{\text{disc}} - w_k^{\text{cont}}) | w_k^{\text{cont}}] = 0 \tag{10.119}$$

$$m_2 = E_{P_0}[(x_i \cdot (w_k^{\text{disc}} - w_k^{\text{cont}}))^2 | w_k^{\text{cont}}] \leq \frac{4}{M} \tag{10.120}$$

$$m_4 - m_2^2 = E_{P_0}[((x_i \cdot (w_k^{\text{disc}} - w_k^{\text{cont}})^2 - m_2)^2 | w_k^{\text{cont}}] \leq 32 \frac{1}{M^2}. \tag{10.121}$$

Looking back at expression (10.115), we have bound $|a_{i,k}| \leq 2\frac{a_1 V(b+a_0 V)}{K}$. Thus for any $x_i \in [-1,1]^d$ and any $w_k^{\text{cont}} \in S_1^d$ we have the upper bound,

$$t_1 = 2a_1 \frac{V(b + a_0 V)\beta}{K} \quad t_2 = a_2 \frac{V(b + a_0 V)\beta}{K} \tag{10.122}$$

$$E_{P_0}[e^{-\frac{\beta}{2} a_{i,k} x_i \cdot (w_k^{\text{disc}} - w_k^{\text{cont}}) + \frac{\beta}{2} \frac{a_2 V(b+a_0 V)}{K} (x_i \cdot (w_k^{\text{disc}} - w_k^{\text{cont}}))^2} | w_k^{\text{cont}}] \tag{10.123}$$

$$\leq \exp\Big(\frac{1}{2} m_2 \frac{e^{2t_1} - 1 - 2t_1}{4}\Big) \exp\Big(\frac{1}{2}(m_4 - m_2^2) \frac{e^{4t_2} - 1 - 4t_2}{16}\Big) e^{\frac{1}{2} t_2 m_2} \tag{10.124}$$

$$\leq \exp\Big(\frac{1}{2M}(e^{2t_1} - 1 - 2t_1) + \frac{1}{M^2}(e^{4t_2} - 1 - 4t_2) + \frac{2t_2}{M}\Big) \tag{10.125}$$

This bound holds for all x_i and w_k^{cont} in their relevant support, thus the product in (10.113) can be upper bound by this object to the power $(N+1)K$. The outer expectation is then irrelevant and we have the final upper bound,

$$(N+1)K\Big(\frac{1}{2M}(e^{2t_1} - 1 - 2t_1) + \frac{1}{M^2}(e^{4t_2} - 1 - 4t_2) + \frac{2t_2}{M}\Big). \tag{10.126}$$

■

References

[1] D. Bakry and M. Emery, Diffusions hypercontractives. *Seminaire de probabilites de Strasbourg* **19** (1985), 177–206

[2] D. Bakry, I. Gentil, and M. Ledoux, *Analysis and geometry of Markov diffusion operators*. 103, Springer, 2014

[3] A. R. Barron, Neural net approximation. In *Proc. 7th Yale Workshop on Adaptive and Learning Systems*, pp. 69–72, 1, 1992

[4] A. R. Barron, Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information theory* **39** (1993), no. 3, 930–945

[5] A. R. Barron, Information-theoretic characterization of Bayes performance and the choice of priors in parametric and nonparametric problems. *In Proc. Valencia Conference, Bayesian Statistics* **6** (1998), 22–52

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

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

[8] R. Bauerschmidt and T. Bodineau, A very simple proof of the LSI for high temperature spin systems. *Journal of Functional Analysis* **276** (2019), no. 8, 2582–2588

[9] S. G. Bobkov and M. Ledoux, From Brunn-Minkowski to Brascamp-Lieb and to logarithmic Sobolev inequalities. *Geometric and Functional Analysis* **10** (2000), 1028–1052

[10] 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 *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024

[11] T. Charnock, L. Perreault-Levasseur, and F. Lanusse, Bayesian neural networks. In *Artificial Intelligence for High Energy Physics*, pp. 663–713, WORLD SCIENTIFIC, 2020

[12] Y. Chen, S. Chewi, A. Salim, and A. Wibisono, Improved analysis for a proximal algorithm for sampling. In *Conference on Learning Theory*, pp. 2984–3014, PMLR, 2022

[13] L. Chizat, E. Oyallon, and F. Bach, On lazy training in differentiable programming. *Advances in Neural Information Processing Systems* **32** (2019)

[14] S. S. Dey, G. Wang, and Y. Xie, Approximation algorithms for training one-node relu neural networks. *IEEE Transactions on Signal Processing* **68** (2020), 6696–6706

[15] R. Dwivedi, Y. Chen, M. J. Wainwright, and B. Yu, Log-concave sampling: Metropolis-Hastings algorithms are fast. *Journal of Machine Learning Research* **20** (2019), no. 183, 1–42

[16] S. Frei, G. Vardi, P. Bartlett, and N. Srebro, The double-edged sword of implicit bias: Generalization vs. robustness in relu networks. In *The Thirty-seventh Annual Conference on Neural Information Processing Systems*, 2023

[17] V. Froese and C. Hertrich, Training neural networks is np-hard in fixed dimension. *Advances in Neural Information Processing Systems* **36** (2024)

[18] V. Gallego and D. Ríos Insua, Current advances in neural networks. *Annual Review of Statistics and Its Application* **9** (2022), no. 1, 197–222

[19] 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 (ITCS 2021)*, pp. 22:1–22:14, 185, Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2021

[20] R. D. Gordon, Values of Mills' ratio of area to bounding ordinate and of the normal probability integral for large values of the argument. *The Annals of Mathematical Statistics* **12** (1941), no. 3, 364–366

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

[22] J. Hron, R. Novak, J. Pennington, and J. Sohl-Dickstein, Wide Bayesian neural networks have a simple weight posterior: theory and accelerated sampling. In *International Conference on Machine Learning*, pp. 8926–8945, PMLR, 2022

[23] X. Huang, D. Zou, Y.-A. Ma, H. Dong, and T. Zhang, Faster sampling via stochastic gradient proximal sampler. *arXiv:2405.16734* (2024)

[24] A. Jacot, F. Gabriel, and C. Hongler, Neural tangent kernel: Convergence and generalization in neural networks. *Advances in Neural Information Processing Systems* **31** (2018)

[25] 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 Transactions on Information Theory* **64** (2018), no. 12, 7649–7656

[26] 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** (2022), 31684–31696

[27] Y. Kook and S. S. Vempala, Gaussian cooling and Dikin walks: the interior-point method for log-concave sampling. *arXiv:2307.12943* (2023)

[28] E. L. Lehmann, J. P. Romano, and G. Casella, *Testing Statistical Hypotheses*. 3, Springer, 1986

[29] C. Liu, L. Zhu, and M. Belkin, Loss landscapes and optimization in over-parameterized non-linear systems and neural networks. *Applied and Computational Harmonic Analysis* **59** (2022), 85–116

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

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

[32] C. McDonald and A. R. Barron, Log concave coupling for sampling from neural net posterior distributions. In *Proc. IMS-NUS Singapore Workshop on Statistical Machine Learning for High Dimensional Data*, 2024

[33] A. Montanari and Y. Wu, Provably efficient posterior sampling for sparse linear regression via measure decomposition. *arXiv:2406.19550* (2024)

[34] R. M. Neal, *Bayesian Learning for Neural Networks*. Lecture Notes in Statistics 118, Springer, New York, NY, 1996

[35] K. W. Ng, G.-L. Tian, and M.-L. Tang, *Dirichlet and Related Distributions: Theory, Methods and Applications*. John Wiley & Sons, 2011

[36] H. Robbins, A remark on Stirling's formula. *The American mathematical monthly* **62** (1955), no. 1, 26–29

[37] V. Srinivasan, A. Wibisono, and A. Wilson, Fast sampling from constrained spaces using the Metropolis-adjusted mirror Langevin algorithm. *arXiv:2312.08823* (2023)

[38] A. Tsigler and P. L. Bartlett, Benign overfitting in ridge regression. *Journal of Machine Learning Research* **24** (2023), no. 123, 1–76

[39] V. H. Vu, On the infeasibility of training neural networks with small mean-squared error. *IEEE Transactions on Information Theory* **44** (1998), no. 7, 2892–2900

[40] L. Wasserman, J. Lafferty, and H. Liu, Concentration of measure. Retrieved 2025, URL https://www.stat.cmu.edu/~larry/=sml/Concentration.pdf, Chapter 7 of Statistical Machine Learning course notes

Curtis McDonald
Department of Statistics and Data Science, Yale University, P.O. Box 208290, 06520-8290 New Haven, United States of America; Curtis.Mcdonald@yale.edu

Andrew R Barron
Department of Statistics and Data Science, Yale University, P.O. Box 208290, 06520-8290 New Haven, United States of America; Andrew.Barron@yale.edu