Phase Transitions in Neural Networks

Liam Carroll

Supervised by Dr. Daniel Murfet

Thesis for the Masters of Science
October 2021

School of Mathematics and Statistics
The University of Melbourne

Acknowledgements

Completed across the course of the COVID-19 pandemic through 250 days in lockdown in Melbourne, the existence of this thesis owes to the abundant love and support I have received from countless friends, family and mentors. In particular, I would like to share my sincere gratitude towards: my non-mathematics friends Charlie, Oscar, Nicola, Gioele and Rachel for countless phone calls offering moral support; my mathematics friends Luke, Ben, Caleb, Spencer and other members of the Melbourne Deep Learning Group for offering plenty of useful advice and answering umpteen questions along the journey; and to my mentors Prof. Arun Ram and Dr. Thomas Quella for their words of wisdom and encouragement, particularly when the going got tough. I feel very lucky that so many people were willing to read drafts and offer guidance from such different perspectives.

A student is only as good as their teacher helps them to be, and for this I am indebted to my supervisor, Dr. Daniel Murfet, for his enduring passion and care in inspiring me to be the best mathematician I can be. Finally, to my family, Lynette, James and Hannah, thank you for your unconditional love and support throughout this process, despite not quite understanding what I have been working on. I hope I have done you all proud.

Contents

1 Introduction **4**

2 Preliminaries **6**
2.1 Feedforward ReLU Neural Networks . . . 6
2.2 The Objects of Statistical Learning Theory . . . 9
2.2.1 Bayesian statistics . . . 9
2.2.2 The Kullback-Leibler divergence $K(w)$ . . . 11
2.2.3 Empirical estimators of loss and error . . . 13
2.2.4 Tempered posterior . . . 15
2.3 Deep Learning as a Gibbs Ensemble . . . 16

3 Singular Learning Theory **18**
3.1 Singular Models . . . 18
3.2 Free Energy . . . 23
3.3 Asymptotics of the Free Energy . . . 25

4 Symmetries of W_0 **28**
4.1 Topology of Two-Layer Feedforward ReLU Networks . . . 28
4.2 Classification of W_0 for $m = d$ . . . 30
4.2.1 Definitions and hypotheses . . . 30
4.2.2 Lemmas . . . 31
4.2.3 Main theorem for $m = d$ . . . 35
4.3 Classification of W_0 for $m < d$ . . . 36
4.3.1 Hypotheses, definitions and lemmas . . . 36
4.3.2 Main theorem for $m < d$ . . . 39
4.4 Arbitrary Depth . . . 40
4.5 Example - m-symmetric Networks . . . 41

5 Phase Transitions in ReLU Neural Networks **44**
5.1 Phases and Phase Transitions . . . 44
5.2 Experimental Methodology . . . 49
5.2.1 Markov Chain Monte Carlo . . . 49
5.2.2 Experimental setup . . . 50
5.2.3 Machine epsilon and practical limits . . . 51
5.2.4 Visualising the posterior . . . 51
5.3 Phase Transition 1: Deforming to Degeneracy . . . 53
5.3.1 Defining the order parameter . . . 53
5.3.2 Results and discussion . . . 54
5.4 Phase Transition 2: Orientation Reversing Symmetry . . . 57
5.4.1 Defining the order parameter . . . 57
5.4.2 Results and discussion . . . 58
5.4.3 An instructive calculation . . . 58
5.4.4 Phase transition 3: Equal Weight-Annihilation . . . 61

6 Conclusion **63**

A Appendix **65**

Bibliography **69**

Notation

Notation	Meaning	Reference
$\mathbb{N}$	Natural numbers $\{1, 2, \dots\}$.	
$\mathrm{ReLU}(x)$	The ReLU function $\max(0, x)$.	Eq. (2.1)
$w \in W$	A parameter in parameter space W defining a model, where W is a compact subset of $\mathbb{R}^D$.	Definition 2.1
$[d]$	The set of nodes of a network $\{1, \dots, d\}$.	Definition 2.1
$f(x, w)$, $f_w(x)$	A feedforward ReLU neural network.	Definition 2.1
$f_0(x)$	A feedforward ReLU network defining a true distribution.	Hypothesis 4.1
(w_i, b_i), (q_i, c)	The weights and biases of the first and second layer respectively, of a two layer network for a node $i \in [d]$.	Eq. (2.2)
$(p(y\|x, w), q(y\|x), \varphi(w))$	The (model, truth, prior) triple.	Hypothesis 2.1
$q(x)$	The known distribution of inputs.	Hypothesis 2.1
$p^\beta(w\|D_n)$, Z_n^β	The tempered posterior distribution and its partition function at inverse temperature β.	Definition 2.11
$\mathbb{E}_X$, $\mathbb{E}_{\mathcal{W}}^\beta$	Expectation with respect to $q(y, x)$ and the truncated posterior respectively.	Definition 2.3 Definition 3.5
$L_n(w)$, S_n	The negative log likelihood of a model, and the entropy of the true distribution.	Definition 2.9
$K(w)$	The Kullback-Leibler divergence from the model to the truth.	Definition 2.7
W_0	The set of true parameters $\{w \in W \mid K(w) = 0\}$.	Definition 2.8
F_n^β, $F_n^\beta(\mathcal{W})$	The free energy, and the free energy of a compact set $\mathcal{W} \subseteq W$.	Definition 3.6
λ	The RLCT of a (model, truth, prior) triple.	Definition 3.7

Chapter 1

Introduction

In this thesis we provide accessible examples of singular statistical models in the form of simple feedforward ReLU neural networks, in order to illustrate the central message of Sumio Watanabe's *Singular Learning Theory*: singularities lie at the heart of statistical learning [Wat09]. In doing so, we demonstrate why the theory of deep learning should shift from analysing *points* in the space of parameters W to considering *singularities* of the Kullback-Leibler divergence $K(w)$.

Deep learning is a part of Artificial Intelligence (AI) that uses statistical models called *neural networks* to model tasks such as computer vision, voice recognition, machine translation and many more [LBH15]. These models have recently been shown to be highly effective [Bro+20] [Nak+19]. The number of parameters D in modern deep learning models is typically orders of magnitude more than the number of datapoints n that they are trained on [Zha+16]. Because of this, standard results for regular statistical models predict that neural networks should overfit the training data and thus have high generalisation error. Understanding why this is not the case, and understanding other observed phenomena such as scaling laws [Kap+20], are important open problems.

We begin by casting deep learning as a Bayesian statistical learning model in Chapter 2. Here the Kullback-Leibler divergence $K(w)$ from a model to the truth is presented as the fundamental object of study, alongside the set of true parameters $W_0 = \{w \in W \mid K(w) = 0\}$. We then explain how one can draw an analogy between neural networks as Bayesian models and the Gibbs ensemble of statistical physics, hinting at objects and phenomena that arise naturally such as the free energy and phase transitions.

Based on the work of [Wat07] and [Mur+20] we then show that feedforward ReLU neural networks are not regular but rather *singular* models, which is to say, have degenerate Fisher information matrices. Thus points on W_0 are singularities of K in the sense of algebraic geometry. It is then argued that minimising the free energy is the central goal of statistical learning since it measures the posterior density associated to different regions of W, and is related to the generalisation error.

The key result of Singular Learning Theory is then stated, that being the correct asymptotic expansion of the free energy for singular models. By desingularising $K(w)$ using Hironaka's Resolution of Singularities, Watanabe shows that the correct measure of model complexity in singular models is the RLCT $\lambda \leq \frac{D}{2}$ which represents the effective number of parameters associated to a singularity. We interpret this result in the context of Occam's Razor, in line with the approach of [Bal97] which considers only regular models. Thus the model selection process is cast as a trade-off between accuracy and complexity.

Perhaps Watanabe's most profound realisation is that "knowledge to be discovered corresponds to a singularity," [Wat09]. Put differently, "if a statistical model is devised so that it extracts hidden structure from a random phenomenon, then it naturally becomes singular," [Wat13]. This offers a groundbreaking shift in perspective of statistical thought from points to singularities. However,

this message has been underappreciated by practitioners of deep learning, perhaps owing to the heavy algebraic geometry machinery at the heart of the theory, as well as the lack of accessible mental models that exhibit the phenomena it describes. We aim to provide such models.

To this end, in Chapter 4 we set about classifying the symmetries of W_0 for two layer ReLU networks of arbitrary width. We consider the realisable case where the true distribution is defined by such a network, for which the problem then becomes classifying functional equivalence of ReLU networks. This is done in two stages. We first analyse the case where the truth and model have the same number of nodes, $m = d$, where we establish that W_0 exhibits scaling symmetry, permutation symmetry and orientation reversing symmetry. The latter is non-generic, occurring only under the particular condition that weight vectors of the true network sum to zero (weight annihilation). This is then generalised to the case where the model is overparamaterised compared to the truth, $m < d$. Here we prove that each of the $d - m$ excess nodes is either degenerate, meaning it is never meaningfully active, or has the same activation boundary as another model node. It is then shown that suitably adjusted scaling, permutation and orientation reversing symmetries must also hold. A more general result from [PL19] that considers networks of arbitrary depth under particular conditions is then discussed. We conclude the chapter by introducing a class of networks, m-symmetric networks, whose associated W_0 exhibit non-generic symmetries.

Armed with a full classification of all points on W_0 in these models, we then set about analysing these points as singularities of K in Chapter 5. The primary goal here is to show that not all points on W_0 are equally good minimisers of the free energy. In line with the statistical physics analogy, we argue that a phase of a neural network corresponds to a small compact subset of W containing a particular singularity of interest. Phase transitions thus arise naturally from variations in the true distribution due to symmetry breaking of W_0, which causes a meaningful change in the free energy. We demonstrate the existence and differences of such phases in two layer ReLU networks through a posterior estimation procedure using Markov Chain Monte Carlo methods. The key result of the work is showing that a point *not* on W_0 can nonetheless be preferred by the posterior. We first show experimentally that the complexity of a degenerate-node phase is lower than that of a non-degenerate node phase, and demonstrate both first and second order phase transitions associated to this due to changes in the accuracy of both phases. We then show that the complexity of non-weight-annihilation phases is lower than that of weight-annihilation phases, and again demonstrate a second order phase transition accordingly.

Chapter 2

Preliminaries

2.1 Feedforward ReLU Neural Networks

Artificial neural networks, which we will refer to simply as *neural networks* or *networks*, are the fundamental mathematical objects of deep learning. They consist of an *input layer*, a number of *hidden layers*, and an *output layer*. Each layer consists of a finite number of nodes. We call the number of layers the *depth* of the network, and the number of nodes in a given layer the *width* of the layer. In general, the *architecture* of the network is the data consisting of:

- The depth L of the network, implying there are $L-1$ hidden layers. [1]
- The widths of each layer $(d_0, \ldots, d_L) \in \mathbb{N}^{L+1}$.
- The graph describing the connectivity of the layers. Each layer l connects to layer $l+1$.

In the simplest case the collection of neurons form a directed acyclic graph where the subgraph generated by successive layers is fully connected, that is, there exist edges connecting each node from layer l to every node of layer $l+1$. Such a network is called a *feedforward network*. A typical schematic of a feedforward neural network is seen in Fig. 2.1.

Architectures with different graph structures have been recently used with great success, including graphs with the presence of loops (e.g. *recurrent neural networks*), or layers that are not fully connected (e.g. *convolutional neural networks*). We refer the reader to [GBC16] for elaboration on such architectures. In this thesis our study will be restricted to feedforward neural networks.

To each edge between a node $i \in [d_{l-1}] := \{1, \ldots, d_{l-1}\}$ in layer $l-1$ to a node $j \in [d_l]$ in layer l is a weight $w^l_{i,j} \in \mathbb{R}$, and to each node j there is a bias $b^l_j \in \mathbb{R}$. This gives rise to an affine function $A^l : \mathbb{R}^{d_{l-1}} \to \mathbb{R}^{d_l}, A^l(x) = (w^l)^T x + b^l$, where the first term is matrix multiplication. Neurons are then "activated" via composition of the affine function with a vectorised activation function $\sigma(x)$, thus the output a^l from each layer $1 \leq l \in L-1$ can be expressed recursively as

$$a^l = \sigma(w^l a^{l-1} + b^l)\,.$$

Interestingly, there is no widely accepted definition of an activation function, and indeed when one examines the plethora of such functions that are used in practice it is clear that there are no common traits other than the fact that they are non-zero somewhere. In the early literature the activation function was typically the step function

$$\sigma_H(x) = \begin{cases} 1 & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases},$$

[1] In a quirk of terminology, the depth ignores the input layer. For example, a two-layer network has an input layer, one hidden layer and output layer.

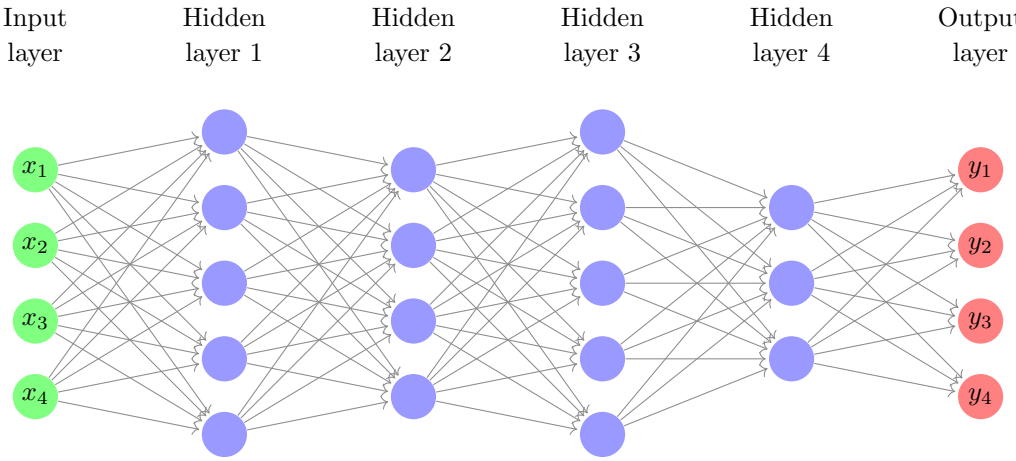


Figure 2.1: Five layer feedforward neural network with widths $(N, d_1, d_2, d_3, d_4, M) = (4, 5, 4, 5, 3, 4)$.

which thus elucidates the use of the term "activation" [Ros62]. Neural networks with $\sigma_H(x)$ as the activation function are called *perceptron* networks. Other common activation functions include the sigmoid function $\sigma_s(x) = \frac{1}{1+e^{-x}}$ and hyperbolic tangent $\sigma_t(x) = \tanh(x)$, but the one we will almost exclusively discuss throughout this thesis is the Rectified Linear Unit (ReLU) defined by

$$\sigma_R(x) = \text{ReLU}(x) = \max\{0, x\} = \begin{cases} x & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases} . \tag{2.1}$$

Note that for $\lambda \in \mathbb{R}$, $\text{ReLU}(\lambda x) = \lambda \,\text{ReLU}(x)$, however $\text{ReLU}(-x) \neq -\,\text{ReLU}(x)$.

We may extend the definition of any of these activation functions to be vectorised by writing $\sigma : \mathbb{R}^n \to \mathbb{R}^n$ for some n such that $\sigma(x_1, \ldots, x_n) = (\sigma(x_1), \ldots, \sigma(x_n))$. From here on we assume that any activation function mentioned has naturally been vectorised.

Remark 2.1. The ReLU function is not analytic at $x = 0$, which we will see is problematic when discussing such networks in the context of Singular Learning Theory in Chapter 3. An analytic alternative to ReLU is the swish function given by

$$\sigma_\gamma(x) = \frac{x}{1 + e^{-\gamma x}}$$

for some $\gamma \in \mathbb{R}$, which satisfies $\lim_{\gamma \to \infty} \sigma_\gamma(x) = \text{ReLU}(x)$ (see [RZL17]).

We now have all of the pieces to define the neural networks we will examine in this thesis.

Definition 2.1. Let $W \subseteq \mathbb{R}^D$ denote the *weight space*, where D is the total number of weights and biases. A *feedforward ReLU neural network* of depth L with widths $(d_0, \ldots, d_L) \in \mathbb{N}^{L+1}$ such that there are $d_0 = N$ inputs and $d_L = M$ outputs, is a feedforward neural network with activation function $\sigma(x) = \text{ReLU}(x)$. That is, it is a function

$$f : \mathbb{R}^N \times W \longrightarrow \mathbb{R}^M$$
$$f(x, w) = (A^L \circ \text{ReLU} \circ A^{L-1} \circ \text{ReLU} \circ \cdots \circ \text{ReLU} \circ A^1)(x)$$

such that for each layer $1 \leq l \leq L$ there is an affine function $A^l : \mathbb{R}^{d_{l-1}} \to \mathbb{R}^{d_l}$, parameterised by weights $w^l \in \mathbb{R}^{d_{l-1} \times d_l}(\mathbb{R})$ and biases $b^l \in \mathbb{R}^{d_l \times 1}(\mathbb{R})$, given by

$$A^l(x) = (w^l)^T x + b^l \,.$$

In the case where w^l is a column vector we may write this as $A^l(x) = \langle w^l, x \rangle + b^l$ where $\langle \, , \rangle$ denotes the dot product. When w is assumed fixed we will often denote $f_w := f(\cdot, w) : \mathbb{R}^N \to \mathbb{R}^M$.

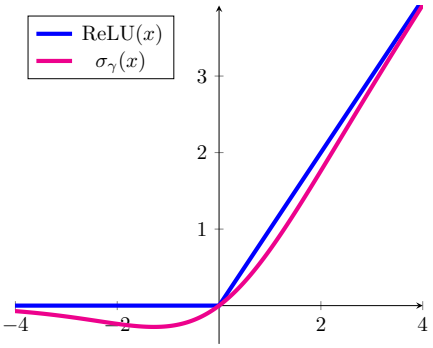


Figure 2.2: ReLU(x) versus $\sigma_\gamma(x)$ for $\gamma = 1$.

Feedforward ReLU networks are, by the definition of ReLU, piecewise affine functions, meaning networks of relatively low depth and widths are quite simple functions. However, this class of networks has been shown to be arbitrarily expressive in the sense that, under suitable conditions, they are universal approximators of arbitrary Lebesgue integrable functions. The following Universal Approximation Theorem of [Lu+17], which we shall not prove here but is included here for completeness, describes this:

Theorem. *Let $g : \mathbb{R}^N \to \mathbb{R}$ be a Lebesgue-integrable function and let $\varepsilon > 0$ be arbitrary. Then there exists a feedforward ReLU neural network $f_w : \mathbb{R}^N \to \mathbb{R}$ of some depth L with bounded widths $d_l \leq N + 4$ such that*

$$\int_{\mathbb{R}^N} |g(x) - f(x, w)| dx < \varepsilon \,.$$

This theorem thus demonstrates the potential of feedforward ReLU neural networks in modern deep learning: they are simple functions to compute, yet they are able to express complicated functions to arbitrary precision. The popularity of such networks lies in this expressivity, together with the fact that in practice good approximations can be found on modern hardware for large datasets via stochastic gradient descent (SGD).

Nearly all networks we will consider in this thesis will have two layers and one output. For readability we make a slight notational adjustment. Let $f_w : \mathbb{R}^N \to \mathbb{R}$ be such a feedforward ReLU neural network with d hidden nodes. For each $i \in [d]$ we let $w_i \in \mathbb{R}^N$ and $b_i \in \mathbb{R}$ denote the weights and biases associated to the first layer, and let $q_i, c \in \mathbb{R}$ be the weights and bias of the second layer. Thus f_w has the form

$$\begin{aligned} f_w(x) &= (A_2 \circ \text{ReLU} \circ A_1)(x) \\ &= \left\langle w^2, \text{ReLU}((w^1)^T x + b^1) \right\rangle + b^2 \\ &= c + \sum_{i=1}^{d} q_i \text{ReLU}(\langle w_i, x \rangle + b_i) \end{aligned} \tag{2.2}$$

where $w = (w_1, \ldots, w_d, b_1, \ldots, b_d, q_1, \ldots, q_d, c) \in W \subseteq \mathbb{R}^{4d+1}$.

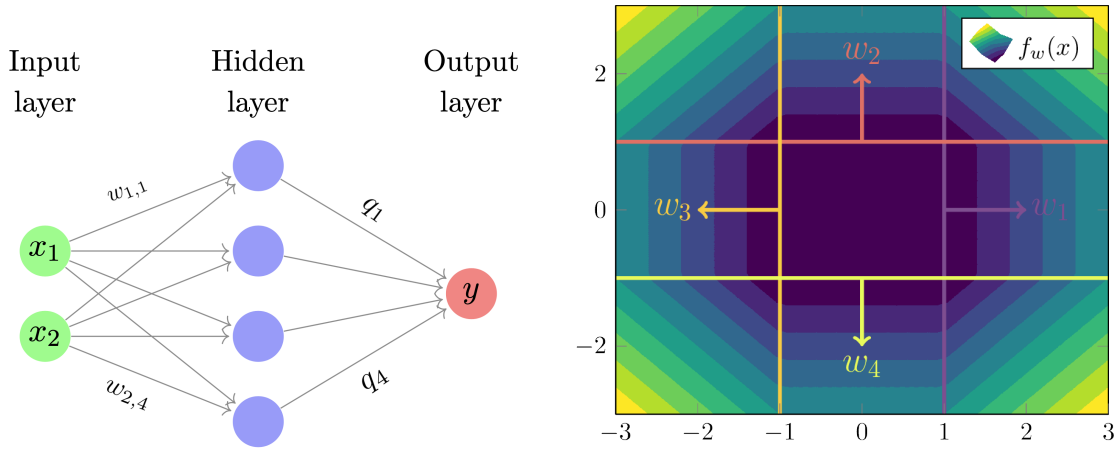
We will return to topological properties of two-layer networks in Section 4.1, but for now we present a single example of such a function for ease of understanding.

Example 2.1. Let $f_w : \mathbb{R}^2 \to \mathbb{R}^1$ be a two-layer neural network with two inputs, one output and $d = 4$ hidden nodes, so

$$\begin{aligned} f_w(x) = q_1 \,\text{ReLU}(w_{1,1}x_1 + w_{2,1}x_2 + b_1) + q_2 \,\text{ReLU}(w_{1,2}x_1 + w_{2,2}x_2 + b_2) \\ + q_3 \,\text{ReLU}(w_{1,3}x_1 + w_{2,3}x_2 + b_3) + q_4 \,\text{ReLU}(w_{1,4}x_1 + w_{2,4}x_2 + b_4) + c \,, \end{aligned}$$

whose network architecture is seen in Fig. 2.3a. Consider $w \in W$ such that

$$f_w(x) = \text{ReLU}(x_1 - 1) + \text{ReLU}(x_2 - 1) + \text{ReLU}(-x_1 - 1) + \text{ReLU}(-x_2 - 1) \,.$$



(a) Neural network graph of $f_w(x)$

(b) Contour plot of $f_w(x)$ and its activation boundaries.

Figure 2.3: The data of $f_w(x)$ in Example 2.1.

A depiction of f_w is seen in Fig. 2.3b, alongside the activation boundaries $\langle w_i, x\rangle + b_i = 0$ that show where each node is activated. Note that the corresponding weight vector w_i, which is normal to its respective activation boundary, points towards the regions of activation when anchored on the activation boundary it defines.

2.2 The Objects of Statistical Learning Theory

Given a dataset D_n of inputs $x \in \mathbb{R}^N$ and outputs $y \in \mathbb{R}^M$ drawn from some true distribution $q(y, x)$, the objective of statistical learning is to train a *model* (or *learning machine*) $p(y|x, w)$ to predict an output for a given input from the true distribution. This amounts to estimating parameters w that minimise a loss metric $K(w)$.

Due to computational benefits such as training parallelisation and scalability, it is standard practice within modern deep learning to view this estimation procedure within a frequentist framework, where w is viewed as being unknown yet fixed [CB02]. Training is then performed using the Stochastic Gradient Descent (SGD) algorithm, which is achieved at scale via the famous back-propagation algorithm [GBC16].

However, as in *Algebraic Geometry and Statistical Learning Theory* [Wat09], our view of the learning procedure will be within the Bayesian framework, whereby the model parameters $w \in W$ are assumed to be drawn from a probability distribution, and the learning goal thus becomes estimating the posterior distribution $p(w|D_n)$.

Remark 2.2. An assumption within the deep learning literature is that training via SGD is approximately equivalent to sampling from a Bayesian posterior, with evidence mounting that this is indeed the case (see [Min+20] and [MHB18]). If true, this justifies our use of Bayesian statistics in drawing conclusions about deep learning, but keep in mind that this connection is not yet rigorous.

The following exposition of Bayesian statistics and related definitions is largely drawn from [Wat18; Wat09] and [CB02].

2.2.1 Bayesian statistics

Let $(\Omega, \mathcal{B}, \mathbb{P})$ be a probability space and $(X, Y) : \Omega \to \mathbb{R}^N \times \mathbb{R}^M$ a jointly random variable subject to the probability density $q(y, x) = q(y|x)q(x)$, where X is the input to the model and Y is the output. Recall that the objective of statistical learning is to estimate the true distribution $q(y, x)$

given a collection of random samples D_n of the form

$$D_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$$

where each (X_i, Y_i) is independent and identically distributed, thus leading to a probability density of the form

$$q((x_1, y_1), \dots, (x_n, y_n)) = q(x_1, y_1) \cdots q(x_n, y_n)\,.$$

We then assume that the data is drawn according to a joint probability distribution $p(y, x|w) = p(y|x, w)q(x)$ which we call the *model*, paramaterised by some parameter $w \in W$. Since the samples are independent and identically distributed, we may define the *likelihood* function as

$$l(w|x, y) := p(y_1, \dots, y_n, x_1, \dots, x_n|w) = \prod_{i=1}^{n} p(y_i, x_i|w) = \prod_{i=1}^{n} p(y_i|x_i, w)q(x_i)\,. \tag{2.3}$$

The statistical learning goal is thus to estimate the posterior density $p(w|x, y)$ subject to a dataset D_n. Let $\varphi(w)$ denote the *prior* probability density of $w \in W$, which is a "subjective distribution based on the experimenter's belief and is formulated before the data is seen" [CB02]. By Bayes' rule, the *posterior* probability density is given by

$$p(w|D_n) := \frac{p(D_n|w)\varphi(w)}{p(D_n)} = \frac{1}{p(D_n)}\varphi(w)\prod_{i=1}^{n} p(y_i, x_i|w) = \frac{1}{p(D_n)}\varphi(w)\prod_{i=1}^{n} p(y_i|x_i, w)q(x_i)\,,$$

where the *evidence* $p(D_n)$ (also called the *marginal likelihood*) is given by

$$p(D_n) = \int_W p(D_n|w)\varphi(w)dw = \int_W \prod_{i=1}^{n} q(x_i)p(y_i|x_i, w)\varphi(w)dw\,,$$

which ensures the posterior is normalised and thus a well defined probability density. But since $\prod_{i=1}^{n} q(x_i)$, which is independent of w, is a factor of both $p(D_n|w)$ and $p(D_n)$, we may simplify this to give a more concise definition:

Definition 2.2. The *posterior* probability density $p(w|D_n)$ is given by

$$p(w|D_n) = \frac{1}{Z_n}\varphi(w)\prod_{i=1}^{n} p(y_i|x_i, w)\,, \quad \text{where} \quad Z_n = \int_W \varphi(w)\prod_{i=1}^{n} p(y_i|x_i, w)dw\,. \tag{2.4}$$

We call Z_n the *partition function*.

Remark 2.3. Clearly the partition function and evidence are related via $p(D_n) = Z_n \prod_{i=1}^{n} q(x_i)$.

Within our setup we are considering the random variable D_n (associated to the random inputs X_i and random outputs Y_i), and the random variable w (which we do not denote with a capital for notational clarity). As such, we define the following expectations:

Definition 2.3. Let $g(X, Y)$ be a function of one sample (X, Y) drawn from the true distribution. Then we write

$$\mathbb{E}_X[g(X, Y)] = \iint_{\mathbb{R}^{N+M}} g(x, y)q(y, x)dxdy\,.$$

In the case where we have a dataset D_n of samples drawn from the true distribution, supposing $g(D_n) = g((X_1, Y_1), \dots, (X_n, Y_n))$, we write

$$\mathbb{E}_{D_n}[g(D_n)] = \iint_{\mathbb{R}^{n(N+M)}} g(D_n)\prod_{i=1}^{n} q(x_i, y_i)dx_i dy_i\,.$$

Let $f(w)$ be a function of the random weights, then the posterior expectation is given by

$$\mathbb{E}_w[f(w)] = \int_W f(w)p(w|D_n)dw\,.$$

Note that due to its dependence on the random variable D_n, $\mathbb{E}_w[f(w)]$ itself is a random variable. For any of these expectations the variance is defined in the usual way,

$$\mathbb{V}(f(x)) = \mathbb{E}[f(x)^2] - \mathbb{E}[f(x)]^2\,.$$

For this thesis we will restrict our attention to the following setup:

Hypothesis 2.1. We consider a (model, truth, prior) triple $(p(y|x,w), q(y|x), \varphi(w))$ associated to the class of feedforward ReLU neural networks $f : \mathbb{R}^N \times W \to \mathbb{R}^M$. We assume that:

- The true conditional distribution $q(y|x)$ is unknown and to be modelled.
- The distribution of inputs, $q(x)$, is known (i.e. not modelled or estimated).
- The prior on parameters, $\varphi(w)$, is a known distribution on a compact space $W \subseteq \mathbb{R}^D$ that contains the origin.
- The model is a standard regression model on f; that is, p is multivariate normally distributed of dimension M with mean $f(x,w)$ and identity covariance matrix, so $p(y|x,w) \sim \mathcal{N}(f(x,w), \mathbb{1}_M)$ with model density given by

$$p(y|x,w) = \frac{1}{(2\pi)^{\frac{M}{2}}} \exp\left(-\frac{1}{2}\|y - f(x,w)\|^2\right)\,,$$

  where $\|.\|^2$ is the standard Euclidean norm on the output space $\mathbb{R}^M$.

Thus we can express the joint densities in terms of conditional densities,

$$q(y,x) = q(y|x)q(x)\,, \quad \text{and} \quad p(y,x|w) = p(y|x,w)q(x)\,.$$

Remark 2.4. The case in which $q(x)$ is to be modelled is of great interest in many real world settings such as natural language processing or image generation. In such situations, *generative models* are used, where the objective is to train a network to generate data similar to its inputs (see [Ope16] for more examples and explanation). Our hypothesis on $q(x)$ is valid for the purposes of this thesis due to the nature of our experiments in Chapter 5.

In Bayesian statistics there are two main ways of estimating a distribution on outputs y given an input x to the learning machine.

Definition 2.4. The *Bayes predictive distribution* is given by

$$p(y|x,D_n) = \mathbb{E}_w[p(y|x,w)] = \int_W p(y|x,w)p(w|D_n)dw\,.$$

Definition 2.5. A *Gibbs estimator* is the model $p(y|x,w)$ evaluated for a single sample drawn from the posterior, $w \sim p(w|D_n)$.

2.2.2 The Kullback-Leibler divergence $K(w)$

The starting point of all supervised[2] statistical learning is to train a given model to minimum loss. In the Bayesian setting it is standard practice to take this loss function to be the Kullback-Leibler divergence $K(w)$. Watanabe shows that the geometry of $K(w)$ strongly affects the learning process, thus it is a central object of our study.

[2] A statistical learning setting is *supervised* if the predictive model is trained via knowledge of the labels (outputs) for each input. This is true of our setting.

Definition 2.6. The *entropy* S of the true conditional distribution $q(y|x)$ is

$$S = \mathbb{E}_X[-\log q(y|x)] = -\iint_{\mathbb{R}^{N+M}} q(y,x)\log q(y|x)dxdy\,.$$

The *negative log loss* (or *negative log likelihood*) $L(w)$ of a model for a given $w \in W$ is

$$L(w) = \mathbb{E}_X[-\log p(y|x,w)] = -\iint_{\mathbb{R}^{N+M}} q(y,x)\log p(y|x,w)dxdy\,.$$

Remark 2.5. We can also define the joint entropy $S_J = -\iint q(y,x)\log q(y,x)dydx$ and the input entropy $S_x = -\iint q(y,x)\log q(x)dxdy$, thus $S_J = S + S_x$. But since $q(x)$ is assumed known and does not depend on w we are really only interested in the quantity S.

Given arbitrary probability distributions $p(z)$ and $q(z)$, one typically defines the Kullback-Leibler divergence, or *relative entropy*, from $p(z)$ to $q(z)$ to be

$$K(q||p) = \int q(z)\log\frac{q(z)}{p(z)}dz\,.$$

Since our $q(y,x)$ and $p(y,x|w)$ both have the known $q(x)$ as factors, we may refine this definition for our purposes as follows:

Definition 2.7. The Kullback-Leibler (KL) divergence of the true distribution $q(y,x)$ from the model $p(y,x|w)$ is a function $K : W \to \mathbb{R}$ defined by

$$K(w) := \mathbb{E}_X\left[\log\frac{q(y|x)}{p(y|x,w)}\right] = \iint_{\mathbb{R}^{N+M}} q(y|x)q(x)\log\frac{q(y|x)}{p(y|x,w)}dxdy\,. \tag{2.5}$$

It satisfies $K(w) = L(w) - S$.

Though it is often thought of as being a distance, $K(w)$ is not a true metric as it is not symmetric in p and q, nor does it satisfy the triangle inequality. It is, however, a loss metric, as the next lemma shows.

Lemma 2.1. *Let $q(y,x)$ and $p(y,x|w) > 0$ be continuous probability density functions. Then:*

- *$K(w) \geq 0$ for all $w \in W$.*
- *$K(w) = 0$ if and only if $p(y|x,w) = q(y|x)$ for almost all $x \in \mathbb{R}^N$, $y \in \mathbb{R}^M$.*

Proof. See Lemma A.1. □

Our statistical learning objective to minimise $K(w)$ thus becomes finding the zero-sets:

Definition 2.8. The set of *true parameters* is defined as

$$W_0 := \{w \in W \,|\, K(w) = 0\} = \{w \in W \,|\, p(y|x,w) = q(y|x)\}\,, \tag{2.6}$$

where the second equality follows from Lemma 2.1. We say that the true distribution $q(y|x)$ is *realisable* by the model $p(y|x,w)$ if W_0 is non-empty. That is, there exists a $w \in W$ such that $q(y|x) = p(y|x,w)$. When $q(y|x)$ is realisable by a true network $f_0(x) = f(x,w^{(0)})$ for some $w^{(0)} \in W$ we will write $W_0(f_0)$.

The definition of realisability should be interpreted as saying that the chosen model is sufficiently expressive to perfectly capture the true distribution in question. This is, of course, unlikely to occur in real world distributions (especially at the scale of datasets at which most deep learning occurs), but the non-realisable case is significantly more technically challenging to deal with, and many of the key results of Singular Learning Theory do not hold under this hypothesis.

Under Hypothesis 2.1 we see that when $q(y|x)$ is realisable, $K(w)$ is just the mean squared error weighted by the prior on inputs $q(x)$:

Lemma 2.2. *Let $q(y|x) = p(y|x, w^{(0)})$ be realisable, defined by a parameter $w^{(0)} \in W$. Then*

$$K(w) = \frac{1}{2} \int_{\mathbb{R}^N} \|f(x, w) - f(x, w^{(0)})\|^2 q(x)\, dx\,. \tag{2.7}$$

Proof. See Lemma A.2. □

In general, for ReLU neural networks, $K(w)$ is *not* analytic. For the minimal counterexample consider $f(x, w) = \text{ReLU}(x - b)$ and truth $f(x, w_0) = \text{ReLU}(x)$ with input distribution $q(x)$ the uniform distribution on $[-a, a]$ for some $a > 0$. Then

$$K(b) = \int_{-a}^{a} \big(\text{ReLU}(x - b) - \text{ReLU}(x)\big)^2\, dx = \begin{cases} -\frac{1}{3a}b^3 + \frac{1}{2}b^2 & b \geq 0 \\ -\frac{1}{6a}b^3 + \frac{1}{2}b^2 & b < 0 \end{cases},$$

which shows that $K(w)$ is C^2 but is not C^3, let alone analytic.

However, we can rectify this by instead considering the swish function σ_γ in Remark 2.1 as an approximation for ReLU, which gives an analytic $K(w)$. When reading the remainder of this thesis, one should keep this correspondence in mind, as it allows us to exploit the properties of either function whenever convenient throughout our analysis.

2.2.3 Empirical estimators of loss and error

In practice, we may only interact with the true distribution $q(y|x)$ by drawing a set of samples $D_n = \{(X_i, Y_i)\}_{i=1}^n$ and calculating an estimator of $K(w)$ based on the observed samples. For notational aesthetics, we let (x_i, y_i) denote the random variables (X_i, Y_i) drawn from $q(y|x)$.

Definition 2.9. Let $D_n = \{(x_i, y_i)\}_{i=1}^n$ be a dataset of inputs and outputs drawn from the true distribution $q(y|x)$ with associated model $p(y|x, w)$. We define the *empirical entropy* S_n of the true distribution to be

$$S_n := -\frac{1}{n} \sum_{i=1}^{n} \log q(y_i|x_i)\,,$$

the *empirical negative log likelihood* $L_n(w)$ (or *empirical negative log loss*) to be

$$L_n(w) := -\frac{1}{n} \sum_{i=1}^{n} \log p(y_i|x_i, w)\,,$$

and the *empirical Kullback-Leibler divergence* to be

$$K_n(w) := \frac{1}{n} \sum_{i=1}^{n} \log \frac{q(y_i|x_i)}{p(y_i|x_i, w)} = L_n(w) - S_n\,. \tag{2.8}$$

The negative log likelihood is so-called due to its relation to the likelihood function (Eq. (2.3))

$$e^{-nL_n(w)} = \prod_{i=1}^{n} p(y_i|x_i, w) = \frac{l(w|x, y)}{\prod_{i=1}^n q(x_i)}\,. \tag{2.9}$$

Using this, we can redefine our posterior distribution to be

$$p(w|D_n) = \frac{1}{Z_n}\varphi(w)e^{-nL_n(w)}\,, \quad \text{where} \quad Z_n = \int_W \varphi(w)e^{-nL_n(w)}\,.$$

This is the form that we will use for the remainder of the thesis, and provides a clear link to the Gibbs distribution of statistical physics in Section 2.3.

Lemma 2.3. *Under Hypothesis 2.1, $L_n(w)$ has the form*

$$L_n(w) = \frac{M}{2} \log 2\pi + \frac{1}{n} \sum_{i=1}^{n} \frac{1}{2} \|y_i - f(x_i, w)\|^2\,,$$

which satisfies $L_n(w) \geq 0$, and is a continuous function of w.

Proof. The first claim follows from a trivial calculation recalling the definition $p(y_i|x_i,w) = \frac{1}{(2\pi)^{\frac{M}{2}}}\exp\left(-\frac{1}{2}\|y_i - f(x_i,w)\|^2\right)$. Both terms are clearly positive by definition of the Euclidean norm. For any ReLU neural network the map $w \mapsto f(x,w)$ is continuous for a fixed x since it is simply a composition of continuous ReLU functions. Thus, since $\|.\|^2$ is continuous, $L_n(w)$ is continuous. □

Lemma 2.3 means that under Hypothesis 2.1 we can interpret the negative log likelihood as simply being the *mean-squared error* plus a constant that only depends on M.

It is important to note that, as stated at the start of the section, we *never have access to* $q(y|x)$. Thus even though $K_n(w)$ is an empirical estimator, we are prohibited from actually estimating it since we can only ever evaluate $p(y|x,w)$. However, the reason we continue to discuss this quantity is because in the limit as $n \to \infty$, S_n can be viewed as a constant that depends on neither the model nor the prior.

Lemma 2.4. *The empirical estimators satisfy*

$$\mathbb{E}_X[K_n(w)] = K(w)\,, \quad \textit{and} \quad \mathbb{E}_X[S_n] = S\,, \quad \textit{and} \quad \mathbb{E}_X[L_n(w)] = L(w)\,.$$

If $K(w), S, L(w) < \infty$ *then as* $n \to \infty$ *we have almost sure convergence*

$$K_n(w) \xrightarrow{a.s.} K(w)\,, \quad \textit{and} \quad S_n \xrightarrow{a.s.} S\,, \quad \textit{and} \quad L_n(w) \xrightarrow{a.s.} L(w)\,.$$

Proof. We will only calculate $K_n(w)$ as the others are identical. Let $w \in W$ be fixed, then

$$\begin{aligned}\mathbb{E}_X[K_n(w)] &= \mathbb{E}_X\left[\frac{1}{n}\sum_{i=1}^n \log\frac{q(y_i|x_i)}{p(y_i|x_i,w)}\right] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}_X\left[\log\frac{q(y_i|x_i)}{p(y_i|x_i,w)}\right] \\ &= \frac{1}{n}\sum_{i=1}^n \int_{\mathbb{R}^{N+M}} q(x,y)\log\frac{q(y|x)}{p(y|x,w)}dxdy = \frac{1}{n}\sum_{i=1}^n K(w) = K(w)\,.\end{aligned}$$

The second statement is a simple corollary of the above calculation by Kolmogorov's Law of Large Numbers [Res99, §7]. □

Remark 2.6. One can define the normalised posterior and normalised partition function to be

$$p(w|D_n) = \frac{1}{Z_n^0}\varphi(w)e^{-nK_n(w)}\,, \quad \text{where} \quad Z_n^0 = \int_W \varphi(w)e^{-nK_n(w)}dw\,,$$

where $Z_n^0 = e^{S_n}Z_n$. Which version you prefer to think of is simply a matter of taste - the inaccessible S_n term is ultimately irrelevant to the learning process in the limit $n \to \infty$.

Given a model defined by some parameters achieves a low loss, it is then natural to ask how well it generalises beyond its training data.

Definition 2.10. Let D_n be a dataset of inputs and outputs drawn from the true distribution $q(y|x)$ with associated model $p(y|x,w)$.

The *Bayesian training loss* is given, in terms of the predictive distribution, by

$$T_n = -\frac{1}{n}\sum_{i=1}^n \log p(y_i|x_i,D_n)\,,$$

and the *Bayesian generalisation loss* is given by

$$G_n = \mathbb{E}_X[-\log p(y|x,D_n)] = -\iint_{\mathbb{R}^{N+M}} q(y,x)\log p(y|x,D_n)dxdy\,,$$

which satisfies $\mathbb{E}_X[T_n] = G_n$. We will typically drop the "Bayesian" prefix.

The *Gibbs training loss* is given by

$$G_t = \mathbb{E}_w[L_n(w)]\,,$$

and the *Gibbs generalisation loss* is given by

$$G_g = \mathbb{E}_w[L(w)]\,,$$

which satisfies

$$G_g = \mathbb{E}_w[\mathbb{E}_X[L_n(w)]] = \mathbb{E}_X[\mathbb{E}_w[L_n(w)]] = \mathbb{E}_x[G_t]$$

by Fubini's theorem.

Using these definitions we can define theoretical error functions

$$\mathrm{TE}_n = T_n - S_n\,, \quad \text{and} \quad \mathrm{GE}_n = G_n - S\,.$$

In particular we have $\mathrm{GE}_n = K(q(y,x)||p(y|x,D_n))$ and a similar result for the training error involving the empirical KL divergence. However, since S_n and S do not depend on the model or prior, minimising loss functions is equivalent to minimising the error. Thus, almost all of our discussions are based on these loss functions and their convergence properties.

2.2.4 Tempered posterior

In accordance with the statistical physics analogy we will present in Section 2.3, we introduce a generalised version of the Bayesian posterior:

Definition 2.11. The *tempered posterior* is defined as

$$p^\beta(w|D_n) := \frac{1}{Z_n^\beta}\varphi(w)e^{-n\beta L_n(w)}\,, \quad \text{where} \quad Z_n^\beta = \int_W \varphi(w)e^{-n\beta L_n(w)}dw\,.$$

We call β the *inverse temperature*. We denote the expectation of some function $f(w)$ with respect to the tempered posterior by

$$\mathbb{E}_w^\beta[f(w)] := \frac{1}{Z_n^\beta}\int_W f(w)\varphi(w)e^{-n\beta L_n(w)}dw\,.$$

Clearly when $\beta = 1$ we have $p^\beta(w|D_n) = p(w|D_n)$, the standard posterior, and $\mathbb{E}_w = \mathbb{E}_w^1$. As such, we will mainly refer to the tempered posterior for the remainder of the thesis, where $\beta = 1$ can be viewed as a special case. Any other variable that we refer to with the β subscript is assumed to be in reference to the tempered posterior, for example, $G_t^\beta = \mathbb{E}_X^\beta[L_n(w)]$.

Remark 2.7. For practical purposes, when p is normally distributed as in Hypothesis 2.1, β manifests itself as the inverse variance of the regression model.

The tempered posterior is well motivated in purely mathematical terms too. In computational Bayesian statistics, the presence of β can be used as a tunable hyperparameter. Furthermore, it arises naturally as the object which minimises information complexity under suitable constraints [Zha06], as the next lemma shows.

Lemma 2.5. *Let $\varphi(w) > 0$ be a prior on W. Suppose $P(w)$ is the unique maximiser of the relative entropy $K(P||\varphi(w))$ subject to the constraint*

$$\mathbb{E}_{w\sim P}[nL_n(w)] = \mu_\beta$$

for some fixed $\mu_\beta \in \mathbb{R}$. Then $P(w) = p^\beta(w|D_n)$ for some $\beta > 0$ that depends on μ_β.

Proof. See Lemma A.3.[3] □

Remark 2.8. When $\beta \to \infty$, the posterior is infinitely concentrated at the maximum likelihood estimators $\hat{w}$. That is, $\lim_{\beta\to\infty} p(w|D_n) = \delta(w - \hat{w})$, where δ is the Dirac delta function [Wat09, §1.3].

[3] My thanks to Matt Farrugia-Roberts for sharing this proof with me.

2.3 Deep Learning as a Gibbs Ensemble

The use of terms from physics such as free energy, entropy and partition function can be justifed by formalising the setting of previous sections into a statistical physics context. The goal of this section is to explain an interpretation of our statistical learning setup as a physical canonical ensemble, shedding light on how it can be viewed as a complex system (see [THK18] for a formal definition of this). This analogy is by no means complete, but we hope it may serve as a useful conceptual framework for physical intuition, and also as a direct mathematical analogy to the widely developed literature of statistical physics.

Thermodynamics is the study of macroscopic observables, such as energy, volume, pressure and mole numbers, associated to equilibrium states. Its central problem is to predict the equilibrium state which eventually results after a change to the total system, for example, after a constraint is removed from a system. We refer the reader to [Cal85] for an extensive overview of this field.

We formulate the learning machine as a *Gibbs ensemble*, in which we imagine the learning process as a physical system in contact with a thermal reservoir. This contact allows an exchange in energy, which corresponds to the ability to cause statistical fluctuations of the system. In physics literature, the Gibbs ensemble, also known as the *canonical ensemble*, is defined as follows:

Definition 2.12. Consider a system with d particles, in a box of volume V, weakly coupled to and in thermal equilibrium with an infinitely large heat reservoir at absolute temperature T. The number of particles in the system is fixed but heat is exchanged with the environment to maintain a temperature T. Let $\Gamma \subseteq \mathbb{R}^D$ denote the configuration space of the particles and their properties, where $\sigma \in \Gamma$ is viewed as a microscopic state (i.e. configuration) of the system of particles. To each microscopic state is associated an energy given by the *Hamiltonian*, denoted $H(\sigma)$. The fundamental postulate of the ensemble is that the probability density of points in phase space $\rho : \Gamma \to \mathbb{R}$ is given by the *Gibbs* (or *Boltzmann*) distribution,

$$\rho(\sigma) = \frac{e^{-\beta H(\sigma)}}{Z}\,, \quad \text{where} \quad Z = \int_\Gamma e^{-\beta H(\sigma)} d\Gamma\,. \tag{2.10}$$

We can immediately make the identification between the phase space Γ of microscopic states and our weight space W, whereby a microstate configuration $\sigma \in \Gamma$ is identified[4] with a weight $w \in W$ (and we assume that the measure on Γ assigns an identical probability to each configuration, thus $d\Gamma = d\sigma = dw$). Comparing the definition of $\rho(\sigma)$ in (2.10) to $p^\beta(w|D_n)$ in Definition 2.11, we note that to identify the Hamiltonian of the system we should rewrite

$$\begin{aligned} p^\beta(w|D_n) &= \frac{1}{Z_n^{0,\beta}} \exp\Big(-\beta(nK_n(w) - \frac{1}{\beta}\log\varphi(w)\Big) \\ \text{with} \quad Z_n^{0,\beta} &= \int_W \exp\Big(-\beta(nK_n(w) - \frac{1}{\beta}\log\varphi(w)\Big) dw\,, \end{aligned}$$

which thus leads us to define:

Definition 2.13. Given a dataset D_n, the (random) *Hamiltonian* of our learning machine is

$$H_n(w) := nK_n(w) - \frac{1}{\beta}\log\varphi(w)\,, \tag{2.11}$$

where $\beta = \frac{1}{T}$ is the inverse temperature of the ensemble.

The Hamiltonian thus measures the violation of two different penalty terms and can be thought of as the cost function that we wish to minimise. The first violation is the loss of the model compared to the true distribution, $nK_n(w)$, and the second violation is imposed by the prior term $-\frac{1}{\beta}\log\varphi(w)$. Notice that if $\varphi(w)$ vanishes (or is negligibly small) then $H_n(w)$ will be very large, meaning that we can interpret $\varphi(w)$ as analogous to walls containing a gas. The posterior will

[4] One may choose to view the d particles of the ensemble as nodes of the neural network with associated configurations given by incoming weights and biases.

thus be concentrated in those regions of W with low Hamiltonian values, which is where $nK_n(w)$ is small and $\varphi(w)$ is large.

The partition function can be viewed as the sum of Boltzmann weights over all possible configurations, meaning one might hope to express an "effective" Hamiltonian F such that

$$e^{-\beta F} \approx \sum_W e^{-\beta H(w)} dw\,.$$

This F is known as the free energy. This will be a crucial part of our analysis, and we will define it formally in Definition 3.4.

Within physics, the free energy as a function of thermodynamic quantities is of fundamental importance, since expectation values of various functions often arise as derivatives of the free energy. For example, consider the canonical ensemble with the number of particles N and the temperature T fixed. Then important physical quantities include: the entropy, $S = -\frac{\partial F}{\partial T}$; the specific heat capacity (at constant volume) $C_V = -T\frac{\partial^2 F}{\partial T^2}$, and the pressure $P = -\frac{\partial F}{\partial V}$.

Using this setup it is thus possible to define macroscopic thermodynamic parameters implicit in statistical learning theory, such as the *average energy* $U = \mathbb{E}_w^\beta[H_n(w)]$ or the *entropy of the Boltzmann distribution* $S = \mathbb{E}_w^\beta[-\log p^\beta(w|D_n)]$. Determining which such quantities are important to statistical learning remains an open problem.

Chapter 3

Singular Learning Theory

In this chapter we outline some of the key concepts and results of Watanabe's *Singular Learning Theory* as described in [Wat07; Wat09; Wat13; Wat18]. The basic observation of the theory is that many statistical models, including neural networks, are strictly singular, which implies that points on the set of true parameters W_0 are degenerate singularities of $K(w)$. This observation draws the link between statistical learning theory and the rich field of singularity theory from algebraic geometry. For such models, regular free energy asymptotic results do not hold. By performing a desingularisation process of $K(w)$ using Hironaka's Resolution of Singularities, Watanabe derives the correct asymptotic forms of the free energy for singular models.

In doing so, Watanabe arrives at a remarkable conjecture: complicated singularities correspond to simpler functions with lower generalisation error. Because of this, he says, singular models are naturally able to infer hidden structure from data. This is a profound statement with far reaching statistical consequences.

The main purpose of the chapter is to provide a short summary of Singular Learning Theory for the uninitiated reader and show that feedforward ReLU neural networks are singular models. The proofs of the free energy asymptotic expansion is beyond the scope of this thesis, but we interpret the result in line with Occam's Razor. We shall provide a mental framework that elucidates the mathematics underpinning the success of deep learning, and further informs our experiments demonstrating phase transitions in Chapter 5.

3.1 Singular Models

Let us begin by outlining what defines a singular model and how the geometry of $K(w)$ associated to such models is fundamentally different to regular models.

Definition 3.1. Let $W \subseteq \mathbb{R}^D$. The elements of the Fisher information matrix $I(w) = \{I_{j,k}(w)\}_{j,k=1}^{D}$ for a given statistical model $p(y|x,w)$ are given by

$$I_{j,k}(w) = \iint_{\mathbb{R}^{N+M}} \left(\frac{\partial}{\partial w_j} \log p(y|x,w)\right) \left(\frac{\partial}{\partial w_k} \log p(y|x,w)\right) p(y|x,w) q(x) dx dy \,,$$

where the derivatives are evaluated at w. We assume $q(x)$ is such that these integrals exist.

Definition 3.2. A statistical model $p(y|x,w)$ is *identifiable* if the map $w \mapsto p(y|x,w)$ is injective for all x, y, and non-identifiable otherwise.

Recall that $I(w)$ is positive definite if for all $x \in \mathbb{R}^D \backslash \{0\}$ and all $w \in W$, I satisfies $x^T I(w) x > 0$. This is equivalent to $I(w)$ having no zero eigenvalues for any w, thus $\det I(w) \neq 0$ for all w. Strictly singular models thus correspond to those models where $\det I(w) = 0$ for some $w \in W$.

Definition 3.3. A statistical model $p(y|x,w)$ is *regular* if it is both identifiable and has positive definite $I(w)$. It is called *strictly singular* if it is not regular.

The distinction between regular and singular models has profound consequences for the geometry of $K(w)$, and therefore the learning process [1]. The main reason for this is that when a model is regular, in a neighbourhood of a true parameter $w^{(0)} \in W_0$, $K(w)$ can be approximated by a quadratic form

$$K(w) \approx \frac{1}{2}(w - w^{(0)})^T I(w^{(0)})(w - w^{(0)}) \,,$$

for which usual convex optimisation applies. However, if $I(w^{(0)})$ is singular then this breaks down. Said differently, regular models obey *asymptotic normality*:

$$p(w|D_n) \xrightarrow{d} N\left(w^{(0)}, \frac{1}{n} I(w^{(0)})^{-1}\right) \,,$$

which is known as the Bernstein-von Moses Theorem [Vaa07, §7]. If $I(w^{(0)})$ is singular then this cannot hold as the inverse of $I(w^{(0)})$ will not exist. Furthermore, the famed *Bayesian information criterion*,

$$\text{BIC} = L_n(w_0) + \frac{D}{2} \log n \,, \qquad L_n(w_0) = \min_{w_0 \in W} L_n(w)$$

used as a tool for comparison between two Bayesian models, is derived by performing a Laplace approximation of $L(w)$ which depends on regularity of $I(w^{(0)})$ for the second order term to exist [KK08]. All of this is to say, regular statistical results of model complexity are inadequate to describe singular models.

The remainder of the section is dedicated to proving the following theorem. For simplicity of the proofs, we restrict our attention to two-layer networks defined in Eq. (2.2), but the proof in full generality can be found in [Wat07] (and see [Mur+20]).

Theorem 3.1. *Let $f : \mathbb{R}^N \times W \to \mathbb{R}^M$ be a ReLU neural network as defined in Definition 2.1, and suppose we have a (model, truth, prior) triple as in Hypothesis 2.1. Then the associated Fisher information matrix $I(w)$ is singular, thus feedforward ReLU neural network models are strictly singular models.*

Lemma 3.2. *Under Hypothesis 2.1, the Fisher information is*

$$I(w)_{j,k} = \int_{\mathbb{R}^N} \left\langle \frac{\partial}{\partial w_j} f(x,w), \frac{\partial}{\partial w_k} f(x,w) \right\rangle_M q(x) dx \,. \tag{3.1}$$

Proof. Let $M = 1$ for simplicity, but note that the proof is easily generalised to higher dimensions using similar Gaussian arguments as in Lemma A.2. We have

$$\begin{aligned} \frac{\partial}{\partial w_j} \log p(y|x,w) &= \frac{\partial}{\partial w_j} \left(-\frac{M}{2} \log 2\pi - \frac{1}{2}(y - f(x,w))^2 \right) \\ &= -\left(\frac{\partial}{\partial w_j} f(x,w) \right) (y - f(x,w)) \end{aligned}$$

which implies

$$I_{j,k}(w) = \int_{\mathbb{R}^N} \left(\frac{\partial}{\partial w_j} f(x,w) \right) \left(\frac{\partial}{\partial w_k} f(x,w) \right) q(x) \left(\int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}} (y - f(x,w))^2 e^{-\frac{1}{2}(y - f(x,w))^2} dy \right) dx \,,$$

and since the second central moment of a Gaussian with $\sigma = 1$ is 1, we get the result. □

The Fisher information matrix is related to the Hessian of $K(w)$, $H_K(w) = \{\frac{\partial^2 K(w)}{\partial w_j \partial w_k}\}_{j,k=1}^D$, in a fundamental way:

[1] Rather loftily, Watanabe states that "almost all statistical models are singular" [Wat07], a statement to be taken with great seriousness, but in mathematical jest.

Lemma 3.3. *Under Hypothesis 2.1, and assuming $q(y|x) = p(y|x, w^{(0)})$ is realisable defined by some fixed $w^{(0)} \in W$, the entries of the Hessian of $K(w)$ satisfy*

$$\frac{\partial^2}{\partial w_j \partial w_k} K(w) = I_{j,k}(w) + \int_{\mathbb{R}^N} \left\langle f(x,w) - f(x,w^{(0)}), \frac{\partial^2}{\partial w_j w_k} f(x,w) \right\rangle_M q(x) dx \,.$$

In particular, $H_K(w^{(0)}) = I(w^{(0)})$.

Proof. The key property is the product rule applied to an inner product. Let $g, h : W \to \mathbb{R}^M$ be two functions, then writing $\partial_j = \frac{\partial}{\partial w_j}$ we have

$$\partial_j \langle g(w), h(w) \rangle = \langle \partial_j g(w), h(w) \rangle + \langle g(w), \partial_j h(w) \rangle \,.$$

This gives

$$\partial_j K = \int_{\mathbb{R}^N} \langle \partial_j f(x,w), f(x,w) - f(x,w^{(0)}) \rangle dx \,.$$

The remaining details are left to the reader. □

To show that $I(w)$ is singular we will show that its rows are linearly dependent.

Lemma 3.4. *Consider a given two-layer ReLU network $f_w : \mathbb{R}^N \to \mathbb{R}$ as defined in Eq. (2.2) with d hidden nodes and a fixed $w \in W$. Given a fixed node $i \in [d]$, f_w satisfies the following differential equation on the open domains for which f_w is differentiable (see Definition 4.2):*

$$\left\{ \sum_{k=1}^{N} w_{ik} \frac{\partial}{\partial w_{ik}} + b_i \frac{\partial}{\partial b_i} - q_i \frac{\partial}{\partial q_i} \right\} f = 0 \,. \tag{3.2}$$

Proof. Let $a_i = \langle w_i, x \rangle + b_i$. Straight calculations show

$$\frac{\partial f}{\partial w_{i,k}} = q_i x_k \mathbb{1}(a_i > 0) \,, \quad \frac{\partial f}{\partial b_i} = q_i \mathbb{1}(a_i > 0) \,, \quad \frac{\partial f}{\partial q_i} = \mathrm{ReLU}(a_i) \,.$$

Recalling that $\mathrm{ReLU}(a_i) = a_i \mathbb{1}(a_i > 0)$, we have

$$\left\{ \sum_{k=1}^{N} w_{ik} \frac{\partial}{\partial w_{ik}} + b_i \frac{\partial}{\partial b_i} - q_i \frac{\partial}{\partial q_i} \right\} f = q_i \left(\sum_{k=1}^{N} w_{i,k} x_k + b_i \right) \mathbb{1}(a_i > 0) - q_i \,\mathrm{ReLU}(a_i) = 0 \,.$$

□

We note that the scaling symmetry that shall be exhibited in Chapter 4 is responsible for this result. Let us now prove the main theorem.

Proof of Theorem 3.1. Observing the form of Eq. (3.1), we first show that $I(w)$ is degenerate if and only if the set

$$\left\{ \frac{\partial}{\partial w_j} f(x,w) \right\}_{j=1}^{D}$$

is linearly dependent. Note that here w_j refers to the jth component of $w \in W \subseteq \mathbb{R}^D$, not to be confused with the specific weight vectors defined in two-layer networks. To see this, first suppose the set is linearly dependent, meaning there is some sequence $r_j \in \mathbb{R}$ such that

$$\sum_{j=1}^{D} r_j \frac{\partial}{\partial w_j} f(x,w) = 0 \,.$$

Then for any fixed $k \in [d]$ we have

$$\begin{aligned}\sum_{j=1}^{D} r_j I_{j,k}(w) &= \sum_{j=1}^{D} r_j \iint_{\mathbb{R}^{N+M}} \left\langle \frac{\partial}{\partial w_j} f(x,w),\, \frac{\partial}{\partial w_k} f(x,w) \right\rangle q(x)dxdy \\ &= \iint_{\mathbb{R}^{N+M}} \left\langle \sum_{j=1}^{D} r_j \frac{\partial}{\partial w_j} f(x,w),\, \frac{\partial}{\partial w_k} f(x,w) \right\rangle q(x)dxdy \\ &= 0\,.\end{aligned}$$

In particular, letting $I^j(w) = [I_{j,k}(w)]_{k=1}^{D}$ denote each row of $I(w)$ we thus have

$$\sum_{j=1}^{D} r_j I^j(w) = \left[\sum_{j=1}^{D} r_j I_{j,k}(w) \right]_{k=1}^{D} = 0\,,$$

thus showing the rows of $I(w)$ are linearly dependent and so $I(w)$ is singular. For the reverse implication, if $I(w)$ is singular then its rows must be dependent by the invertible matrix theorem. We leave the remaining details as an exercise for the reader.

In particular, the differential equation in Eq. (3.2) implies that for *any* $w \in W$ we have a linear dependence relation for each node. The Fisher information restricted to each node is thus singular, and arranging $I(w)$ into block diagonal form where the blocks correspond to each node shows that $I(w)$ itself is singular, proving the claim. □

Connection to algebraic geometry

This series of results provides the key link between statistical learning theory and algebraic geometry, as we now explain.

Given an analytic function $K : W \to \mathbb{R}$, $x \in W$ is a *critical point* of K if $\nabla K(x) = 0$, and if it further satisfies $K(x) = 0$ then it is a *singularity* of K [Har10, §1.5]. In fact, any true parameter $w^{(0)} \in W_0$ is a singularity of K since $K(w^{(0)}) = 0$, and $\nabla K(w^{(0)}) = 0$ because $K(w) \geq 0$. However, what we are interested in are *degenerate* singularities.

A singularity $x \in W \subseteq \mathbb{R}^D$ of K is *non-degenerate* if in a neighbourhood of x one can write

$$K(x) = x_1^2 + \cdots + x_D^2$$

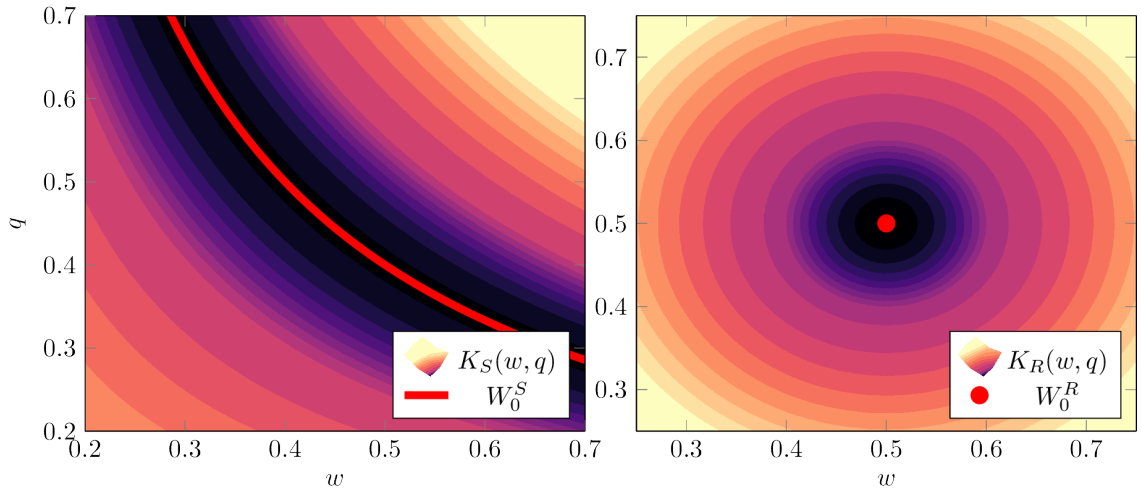
for some set of local coordinates $x_1, \ldots, x_D$. Otherwise, x is *degenerate*. By the Morse lemma [Gil93], if the Hessian of K at a singularity x is non-degenerate, then that singularity is non-degenerate, which corresponds to non-degenerate Fisher information matrix by Lemma 3.3. Then by Theorem 3.1, for feedforward ReLU neural networks, *every point on W_0 is a degenerate singularity of K*.

From the point of view of algebraic geometry, non-degenerate singularities of K are uninteresting. Even if a statistical model is non-identifiable (meaning true parameters are simply isolated minima of K), regular asymptotic results hold in a local sense, as discussed in [Bal97]. The strength of Singular Learning Theory's results are only necessary for dealing with the case where W_0 contains degenerate singularities, where it is shown that the nature of these singularities from an algebraic geometric perspective *strongly affect* the statistical learning process. It is this realisation that makes Watanabe's change in perspective truly groundbreaking for statistical learning.

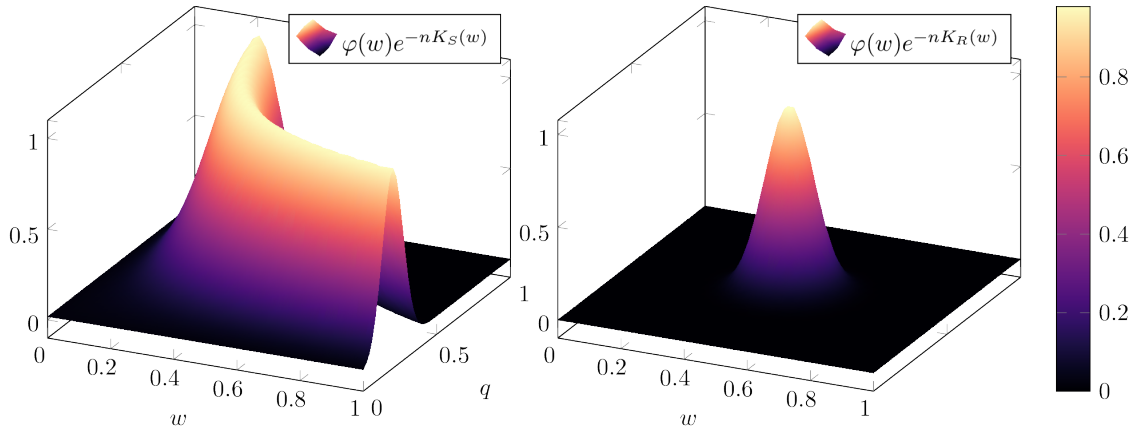
For intuition let us now examine how W_0 typically differs between regular and singular models.

Example 3.1. Let $W \subseteq \mathbb{R}^2$ and denote $w = (w, q) \in W$. Consider a one-node two-layer ReLU network $f : \mathbb{R} \times W \to \mathbb{R}$. Define a model and an underlying truth by

$$f(x,w) = q\,\mathrm{ReLU}(wx)\,, \quad f(x,w^{(0)}) = \theta\,\mathrm{ReLU}(x)\,.$$



(a) Contour plots of $K_S(w,q)$ and $K_R(w,q)$.



(b) Normalised posterior (see Remark 2.6) of singular and regular models, taking $K_n(w) \approx K(w)$.

Figure 3.1: The difference between singular and regular models for $\theta = \frac{1}{5}$ and $\vartheta = \frac{1}{2}$ respectively, where $\varphi(w)$ is uniform on $[0,1]^2$ and $n = 100$.

Assume that $w, q, \theta \geq 0$ where θ is fixed, and assume that $q(x)$ is the uniform distribution on $[-\sqrt{3}, \sqrt{3}]$. Then by Lemma 2.2,

$$K_S(w,q) = (qw - \theta)^2 ,$$

so $W_0^S = \{(w,q) \in W \mid qw = \theta\}$ and thus every point is a degenerate singularity. We observe in Fig. 3.1a that this gives a kind of hyperbolic valley, where W_0 corresponds to the floor of the valley. The degeneracy here comes from the fact that at every point $w^{(0)} \in W_0$, there is a tangent direction in which $K(w)$ is still minimised.

In contrast, consider a constant model and truth $g : \mathbb{R} \times W \to \mathbb{R}^2$ (i.e. a feedforward ReLU network with only the last layer biases) such that

$$g(x,w) = (w,q) , \quad g(x,w^{(0)}) = (\vartheta, \vartheta) ,$$

where we again assume $w, q, \vartheta \geq 0$ and ϑ is fixed. Taking $q(x)$ uniform on $[0,2]$ for ease, this gives

$$K_R(w,q) = (w - \vartheta)^2 + (q - \vartheta)^2 ,$$

and so $W_0^R = (\vartheta, \vartheta)$. Fig. 3.1a shows that this induces a bowl-like paraboloid with a single minimiser at $(w,q) = (\vartheta, \vartheta)$.

For completeness, Fig. 3.1b shows the respective posteriors of the singular and regular model (up to a scale factor).

3.2 Free Energy

As alluded to in Section 2.3, the free energy is a fundamental object of study in both physics and Bayesian statistics. Primarily, one can think of the free energy as being a measure of posterior density associated to a particular region of W. In physical terms, these regions of W are sets of microscopic states associated to particular macroscopic states.

Such a view implies that optimising the free energy is perhaps the *fundamental objective of statistical learning*. Indeed many of the results of Singular Learning Theory suggest that this is the correct approach to understanding how neural networks generalise so effectively, as well as being the key to understanding how neural networks undergo phase transitions.

Definition 3.4. Given a dataset D_n, we define the *total empirical free energy* $F_n^\beta \in \mathbb{R}$ as

$$F_n^\beta = -\log Z_n^\beta = -\log\left(\int_W \varphi(w) e^{-n\beta L_n(w)} dw\right).$$

We will typically just refer to F_n^β as the free energy.

Let us inspect this definition a bit more closely. The free energy F_n^β depends on the choice of model p and prior φ, but more importantly it is inherently a random variable that depends on the random dataset D_n. To investigate the posterior landscape of a given true network in the search for phase transitions we will want to make statements independent of D_n, hence we may instead define the *total free energy* as a function of $\mathbb{E}_X[nL_n(w)] = nL(w)$,

$$\overline{F}_n^\beta = -\log\left(\int_W \varphi(w) e^{-n\beta L(w)} dw\right).$$

Note that $\overline{F}_n^\beta$ still depends on n even though the randomness in D_n has been marginalised out. Indeed, it is stated in [Wat18, §9.4] that F_n^β and $\overline{F}_n^\beta$ are asymptotically equivalent up to constant order, meaning we may interchange statements about either.

Remark 3.1. In physics, $\overline{F}_n^\beta$ is known as the annealed average, whereas $\mathbb{E}[-\log Z_n^\beta]$ is known as the quenched average. This is a subtle but important distinction, which is explored in a statistical mechanics setting in [SST92] by appealing to the replica method.

What makes the free energy more informative than $K(w)$ is that it encodes both regions of maximum likelihood, which correspond to minimisers of $K(w)$, as well as the generalisation of a region of parameters, as the next lemma shows.

Lemma 3.5. *Let F_n denote the free energy when $\beta = 1$. The generalisation loss is the average increase in free energy,*

$$G_n = \mathbb{E}_{X_{n+1}}[F_{n+1}] - F_n\,. \tag{3.3}$$

In particular, the average free energy is the sum of the generalisation loss,

$$\mathbb{E}_{D_n}[F_n] = \sum_{i=1}^{n-1} \mathbb{E}_{D_i}[G_i] + \mathbb{E}_{D_1}[F_1]\,.$$

Proof. The proof hinges on the fact that we may write

$$\frac{Z_{n+1}}{Z_n} = \frac{\int_W p(y_{n+1}|x_{n+1},w)\varphi(w)e^{-n\beta L_n(w)}dw}{\int_W \varphi(w)e^{-n\beta L_n(w)}dw} = \mathbb{E}_w[p(y_{n+1}|x_{n+1},w)] = p(y|x,D_n)$$

which implies

$$F_{n+1} - F_n = -\log p(y|x,D_n)\,.$$

The remaining details can be found in Lemma A.5. □

In the process of model selection, which we will outline formally in Section 5.1, we will be interested in comparing the free energy associated to compact sets.

Definition 3.5. Let $\mathcal{W} \subseteq W$ be a compact subset. We define the *truncated posterior* $p^\beta_{\mathcal{W}}(w|D_n)$ to be the posterior restricted to $\mathcal{W}$, that is,

$$p^\beta_{\mathcal{W}}(w|D_n) = \frac{\varphi(w)e^{-n\beta L_n(w)}\mathbb{1}(w \in \mathcal{W})}{Z^\beta_n(\mathcal{W})}, \quad \text{where} \quad Z^\beta_n(\mathcal{W}) = \int_{\mathcal{W}} \varphi(w)e^{-n\beta L_n(w)}dw\,.$$

We let $\mathbb{E}^\beta_{\mathcal{W}}$ denote expectation with respect to $p^\beta_{\mathcal{W}}(w|D_n)$.

Definition 3.6. For a given β and n, the *empirical free energy* associated to a compact region $\mathcal{W} \subseteq W$ is

$$F^\beta_n(\mathcal{W}) = -\log Z^\beta_n(\mathcal{W}) = -\log\left(\int_{\mathcal{W}} \varphi(w)e^{-n\beta L_n(w)}dw\right).$$

The total free energy $\overline{F}^\beta_n(\mathcal{W})$ is defined in the obvious way.

Remark 3.2. Lemma 3.5 is easily generalised to $F_n(\mathcal{W})$ by considering $\frac{Z_{n+1}(\mathcal{W})}{Z_n(\mathcal{W})}$.

Suppose $\mathcal{W}_1, \mathcal{W}_2 \subseteq W$ are two compact sets, and suppose for simplicity that they have the same measure in W. If the posterior is more densely concentrated in $\mathcal{W}_1$ than $\mathcal{W}_2$, then $\int_{\mathcal{W}_1} p(w|D_n)dw > \int_{\mathcal{W}_1} p(w|D_n)dw$, implying $F^\beta_n(\mathcal{W}_1) < F^\beta_n(\mathcal{W}_2)$. Combining this with its relation to generalisation from Lemma 3.5 shows why minimising the free energy should be our primary statistical goal.

As in statistical physics, derivatives of the free energy with respect to intensive parameters often equate to expectation values and variances of quantities of interest. This is easy to see given the following lemma.

Lemma 3.6. *Let $\mathcal{W} \subseteq W$ be compact. The free energy of $\mathcal{W}$ satisfies*

$$\frac{\partial F^\beta_n(\mathcal{W})}{\partial \beta} = \mathbb{E}^\beta_{\mathcal{W}}[nL_n(w)] = nG^\beta_t(\mathcal{W}),$$

$$\textit{and} \quad \frac{\partial^2 F^\beta_n(\mathcal{W})}{\partial \beta^2} = -\mathbb{E}^\beta_{\mathcal{W}}[(nL_n(w))^2] + \mathbb{E}^\beta_{\mathcal{W}}[nL_n(w)]^2 = -\mathbb{V}^\beta_{\mathcal{W}}[nL_n(w)].$$

Proof. For the first result, by straight calculation we have

$$\begin{aligned}
\frac{\partial F^\beta_n(\mathcal{W})}{\partial \beta} &= -\frac{1}{Z^\beta_n(\mathcal{W})}\frac{\partial Z^\beta_n(\mathcal{W})}{\partial \beta} \\
&= -\frac{1}{Z^\beta_n(\mathcal{W})}\frac{\partial}{\partial \beta}\left(\int_{\mathcal{W}} \varphi(w)e^{-n\beta L_n(w)}dw\right) \\
&= -\frac{1}{Z^\beta_n(\mathcal{W})}\int_{\mathcal{W}} \varphi(w)\frac{\partial}{\partial \beta}e^{-n\beta L_n(w)}dw \\
&= \frac{1}{Z^\beta_n(\mathcal{W})}\int_{\mathcal{W}} nL_n(w)\varphi(w)e^{-n\beta L_n(w)}dw = \mathbb{E}^\beta_{\mathcal{W}}[nL_n(w)],
\end{aligned}$$

where we may take the derivative inside the integral by similar arguments to those in Lemma 3.7. The second equality here follows from the definition G^β_t and $L_n(w) = K_n(w) + S_n$. We leave the second derivative to Lemma A.4. □

Whilst an important quantity, the free energy is both analytically and computationally intractable for most non-trivial models, meaning our model selection process hinges on asymptotic results instead. Deriving these asymptotics for singular models is the main result of Singular Learning Theory, which we will see in Section 3.3. The start of this proof begins with the following simple result.

Lemma 3.7. *Assume that $L_n(w)$ is not a constant in w. Denote $F^\beta_n(\mathcal{W}) = F^{\mathcal{W}}_n(\beta)$ to indicate $\mathcal{W} \subseteq W$ is fixed and β is the function variable. Then*

1. $\mathbb{E}_{\mathcal{W}}^{\beta}[nL_n(w)]$ *is a decreasing function of* β.

2. $F_n^{\mathcal{W}}(\beta)$ *is continuous in* β.

3. *There exists a unique* $\beta^* \in (0,1)$ *satisfying*

$$F_n^{\mathcal{W}}(1) = \mathbb{E}_{\mathcal{W}}^{\beta^*}[nL_n(w)]\,.$$

Proof. From Lemma 3.6, since $\frac{\partial}{\partial\beta}\mathbb{E}_{\mathcal{W}}^{\beta}[nL_n(w)] = -\mathbb{V}_{\mathcal{W}}^{\beta}[nL_n(w)]$, and the variance is always positive by the Cauchy-Schwarz inequality (and $L_n(w)$ is non-constant), we see that $\frac{\partial}{\partial\beta}\mathbb{E}_{\mathcal{W}}^{\beta}[nL_n(w)] < 0$ showing the first claim.

$F_n^{\mathcal{W}}(\beta)$ is continuous via a simple application of the Lebesgue dominated convergence theorem assuming Hypothesis 2.1, for which $\varphi(w)$ and $L_n(w)$ are continuous in w and positive. Merely consider a sequence $\beta_j \to \beta$ and define a sequence of continuous functions $f_j(w) = \varphi(w)e^{-n\beta_j L_n(w)}$. Then $f_j(w) \to f(w) = \varphi(w)e^{-n\beta L_n(w)}$, and $|f(w)|$ is bounded since $\mathcal{W}$ is compact and f is continuous, meaning

$$F_n^{\mathcal{W}}(\beta_j) = \int_{\mathcal{W}} f_j(w)dw \longrightarrow \int_{\mathcal{W}} f(w)dw = F_n^{\mathcal{W}}(\beta)$$

by the dominated convergence theorem [SS05, §2] and so $F_n^{\mathcal{W}}(\beta)$ is continuous.

For the this final claim, first note that by definition $F_n^{\mathcal{W}}(0) = 0$, hence

$$F_n^{\mathcal{W}}(1) = \int_0^1 \frac{\partial F_n^{\mathcal{W}}}{\partial\beta}(\beta)d\beta\,.$$

Since $F_n(\beta)$ is continuous, by the mean value theorem there exists a unique $\beta^* \in (0,1)$ such that

$$\frac{\partial F_n^{\mathcal{W}}}{\partial\beta}(\beta^*) = \mathbb{E}_{\mathcal{W}}^{\beta^*}[nL_n(w)] = \frac{F_n^{\mathcal{W}}(1) - F_n^{\mathcal{W}}(0)}{1-0} = F_n^{\mathcal{W}}(1)\,. \qquad \square$$

Finding the optimal inverse temperature β^* is where the heavy lifting of the resolution of singularities comes into play, which we shall now discuss.

3.3 Asymptotics of the Free Energy

One of the key observations of Watanabe is that the real log canonical threshold λ is the fundamental quantity determining the geometry of $K(w)$, at least as it relates to statistical learning theory. Suppose $\varphi(w) > 0$ on all of W and $K(w)$ is a real analytic function. The zeta function $\zeta(z)$ of a statistical model, where $z \in \mathbb{C}$, is defined as

$$\zeta(z) = \int_W K(w)^z \varphi(w)dw\,.$$

It is standard (see [Wat09, §4]) that $\zeta(z)$ can be analytically continued to a meromorphic function on the whole complex plane with Laurent expansion

$$\zeta(z) = \zeta_0(z) + \sum_{k=1}^{\infty}\sum_{m=1}^{m_k} \frac{c_{km}}{(z+\lambda_k)^m}\,, \tag{3.4}$$

where $\zeta_0(z)$ is holomorphic and $c_{km} \in \mathbb{C}$ are coefficients, each $\lambda_k \in \mathbb{Q}_{>0}$ satisfies $0 < \lambda_1 < \lambda_2 < \ldots$ and m_k is the largest order of the pole λ_k.

Definition 3.7. The *real log canonical threshold* (RLCT) of the (model, truth, prior) triple is $\lambda = \lambda_1$ with *multiplicity* $m = m_1$ of Eq. (3.4).

Remark 3.3. The subsequent results hold on any compact set W, so in particular one may consider local RLCTs associated to some compact subset $\mathcal{W} \subseteq W$ by taking the integral over this domain in the definition of $\zeta(z)$.

Remark 3.4. Recall from Section 2.2.2 that $K(w)$ is not in general analytic for ReLU networks, thus one instead pretends that we are referring to the swish function for the remainder of this discussion.

As shown in [Wat09, Theorem 7.1], the RLCT can be understood as a volume co-dimension, that is, the number of "effective parameters" near the most singular point of $W_0 \subseteq \mathbb{R}^D$. In essence, $\lambda = \frac{D'}{2}$ where $D' \leq D$ is such that for every point $w_0 \in W_0$, $K(w)$ has an expression in local coordinates of the form

$$K(w) = \sum_{i=1}^{D'} c_i w_i^2$$

for some constants $c_i > 0$ that may depend on w_0. With this viewpoint, it becomes clear that the RLCT should be the quantity that describes the geometric behaviour of $K(w)$ near true parameters, and thus the free energy. In fact, in regular models, the RLCT is *precisely equal* to $D/2$.

To analyse the asymptotic expansion of the free energy, Watanabe's idea is to *desingularise* $K(w)$ by employing Hironaka's Resolution of Singularities [Hir64], one of the fundamental results of algebraic geometry. Let us state the main theorem of [Wat13]. The fundamental conditions can be found in [Wat09, §6] and further summarised in [Wat13, §3]. In short, they ensure that W is a compact set whose boundary is defined by several analytic functions, the prior is analytic, the model is sufficiently integrable in the L^s-space for $s \geq 6$, and a local finite-variance property of $K(w)$.

Theorem 3.8. *Consider a (model, truth, prior) triple satisfying the fundamental conditions. Let* $\beta = \frac{\beta_0}{\log n}$ *for some constant* $\beta_0 > 0$*, and let* $L_n(w_0) = \min_{w \in W} L_n(w)$ *(in the realisable case* $S_n = L_n(w_0)$*). Then there exists a random variable* U_n *such that*

$$\mathbb{E}_w^\beta[nL_n(w)] = nL_n(w_0) + \frac{\lambda}{\beta} + U_n\sqrt{\frac{\lambda \log n}{2\beta_0}} + O_p(1)\,, \tag{3.5}$$

where λ *is the RLCT of the triple and* $\{U_n\}$ *is a sequence of random variables which satisfy* $\mathbb{E}[U_n] = 0$ *and converges in distribution to a Gaussian random variable as* $n \to \infty$*.*

Remark 3.5. Recall that $o_p(1)$ denotes a sequence of random vectors that converge to zero in probability, and $O_p(1)$ denotes a sequence that is bounded in probability [Vaa07, §2.2].

Remark 3.6. So long as $\varphi(w) > 0$ on all of W Theorem 3.8 holds independent of the choice of $\varphi(w)$.

Combining this with Lemma 3.7, Watanabe shows:

Theorem 3.9. *The free energy satisifes*

$$F_n = \mathbb{E}_w^{\beta^*}[nL_n(w)]$$

at the optimal inverse temperature

$$\beta^* = \frac{1}{\log n}\left(1 + \frac{U_n}{\sqrt{2\lambda \log n}} + o_p\left(\frac{1}{\sqrt{\log n}}\right)\right).$$

The quantity $\mathbb{E}_w^{\beta^*}[nL_n(w)]$ *is called the* Widely Applicable Bayesian Information Criterion (WBIC)*.*

Remark 3.7. The WBIC result still holds for a truncated posterior on a compact $\mathcal{W} \subseteq W$, that is,

$$F_n(\mathcal{W}) = \mathbb{E}_{\mathcal{W}}^{\beta^*}[nL_n(w)]$$

for the same optimal inverse temperature. This allows us to discuss local free energy associated to different regions of W.

Remark 3.8. These theorems have important consequences for practical computations, too. Firstly, Theorem 3.9 gives us a direct way of estimating the free energy of a given triple. We can estimate the posterior via a sampling procedure at inverse temperature $\beta^* = \frac{1}{\log n}$, and then calculate the average log loss over these samples. Secondly, Theorem 3.8 shows we may estimate the RLCT via a simple linear regression on $\{(x_j, y_j)\}$ for a sequence of points $y_j = \mathbb{E}_w^{\beta_j}[nL_n(w)]$, $x_j = \frac{1}{\beta_j} = \frac{\log n}{\beta_0^j}$, where $\hat{\lambda}$ is the gradient of this line [Wat13, §6.2].

The WBIC is so called because it is the precise generalisation of the BIC for singular models:

Corollary 3.10. *If a model with $W \subseteq \mathbb{R}^D$ is regular, then $\lambda = \frac{D}{2}$ and $U_n = 0$, so*

$$\text{WBIC} = nL_n(w_0) + \frac{D}{2}\log n = \text{BIC}\,. \tag{3.6}$$

It is in this sense that deep learning is "unreasonably effective" as described by Yann Lecun [LeC14]. According to the BIC, if D is very large (for example, $D \sim 10^{11}$ in the state of the art GPT-3 [Bro+20]) one should never use large neural networks as the dimensionality is massively penalised. But as Watanabe shows, it is not D that we care about in model selection, but λ. Let us explore this.

Occam's Razor

We can interpret the asymptotic relationship in Eq. (3.5) as a competition between "energy and entropy", or equivalently, "accuracy and complexity". The above theorems show that in singular models, for any compact $\mathcal{W} \subseteq W$ we may write

$$F_n(\mathcal{W}) \approx nL_n(\omega_0) + \lambda\frac{\log n}{\beta_0}\,, \tag{3.7}$$

where $L_n(\omega_0) = \min_{\omega\in\mathcal{W}} L_n(\omega)$, $\beta_0 = \beta^* \log n$ and λ is the local RLCT associated to the most singular point on $\mathcal{W}$. For the rest of this discussion we assume we are in the realisable case.

In [Bal97] the BIC in Eq. (3.6) is analysed for regular models. The $nL_n(w_0)$ term is called the *accuracy* of the model, which is to say, the smallest loss that one can hope to attain for the model $p(y|x,\omega)$ evaluated at some parameter $\omega \in \mathcal{W}$. The $\frac{D}{2}\log n$ term is a measure of *complexity*, where models with large numbers of parameters are penalised. This, he states, gives a mathematical realisation of Occam's Razor: "plurality should not be posited without necessity," [Bri] or typically known as "the simplest explanation is usually the right one".

This embodiment is even more apparent in the singular setting. Recall that λ can be thought of as measuring the effective number of dimensions associated to a singularity on W_0. The accuracy term in Eq. (3.5) is the same as in the BIC, but the complexity term is now dependent on λ.

Let $\mathcal{W}, \mathcal{W}' \subseteq W$ be two sufficiently small compact sets and $\mathcal{W}_0 = \mathcal{W} \cap W_0$, $\mathcal{W}_0' = \mathcal{W}' \cap W_0$, which we assume are both non-trivial. Let $\omega_0 \in \mathcal{W}_0$ and $\omega_0' \in \mathcal{W}'$. Since they are both true parameters their accuracies are equal, $L_n(\omega_0) = L_n(\omega_0') = S_n$. Suppose, however, that the local RLCTs λ, λ' of $\mathcal{W}, \mathcal{W}'$ respectively satisfy $\lambda < \lambda'$. Then

$$F_n(\mathcal{W}) < F_n(\mathcal{W}')$$

meaning we should prefer models from $\mathcal{W}_0$ over $\mathcal{W}_0'$ since they have lower model complexity as measured by the RLCT. Recalling the relationship between free energy and generalisation from Lemma 3.5, this illustrates Watanabe's equivalence [Wat09, §7.6]:

$$\text{smaller } \lambda \iff \text{more complicated singularity} \iff \text{lower free energy} \iff \text{better generalisation.}$$

This forms the motivation for our experiments in Chapter 5 where we will demonstrate that different singularities can have different free energies with simple examples. First we need to classify the points of W_0 in some simple but nontrivial examples. This is the subject of the next chapter.

Chapter 4

Symmetries of W_0

In order to demonstrate differences in the free energy of regions of W, we must first begin by understanding what points are on W_0. In this chapter we will fully charactersie the symmetries of the set of true parameters W_0 when the true distribution is defined by an activation-distinguished minimal two-layer feedforward ReLU network with two inputs and one output with m hidden nodes, and the model is defined similarly but with d hidden nodes. This classification procedure is equivalent to classifying functional equivalence of these networks, with the key insight being that their activation boundaries must be the same. We begin with the case where $m = d$ and find that W_0 generically exhibits scaling, permutation, and under particular conditions, orientation reversing symmetry. This is then generalised to $m < d$ where it is found that each of the $d - m$ excess nodes is either degenerate or shares an activation boundary with another node. We then state the main theorem of [PL19] which proves similar results for networks of arbitrary depth. The section concludes with an example of networks with non-generic symmetries called m-symmetric networks.

4.1 Topology of Two-Layer Feedforward ReLU Networks

Let $f_w : \mathbb{R}^N \to \mathbb{R}$ be a two-layer feedforward ReLU neural network with one input, two outputs and d hidden nodes for some fixed $w \in W$. For convenience, recall Eq. (2.2):

$$f_w(x) = c + \sum_{i=1}^{d} q_i \mathrm{ReLU}(\langle w_i, x\rangle + b_i) \tag{4.1}$$
$$\text{where} \quad w = (w_1, \ldots, w_d, b_1, \ldots, b_d, q_1, \ldots, q_d, c) \in W \subseteq \mathbb{R}^{4d+1},$$

where for each node $i \in [d] := \{1, \ldots, d\}$ we have $w_i \in \mathbb{R}^2$, and $b_i, q_i, c \in \mathbb{R}$. In this section a network will always refer to a network of this form.

Definition 4.1. We say a node $i \in [d]$ is *degenerate* in a network f_w if either $q_i = 0$ or $w_i = 0$, thus meaning there is no meaningful contribution from i to $f_w(x)$.

Remark 4.1. In the case where $w_i = 0$ but $q_i \neq 0$ and $b_i \neq 0$ we may simply redefine the total bias to be $c' = c + q_i b_i$, thus meaning degeneracy need only be defined in terms of the weights q_i and w_i. Without loss of generality, we exclude this case from all subsequent results.

ReLU neural nets are piecewise affine functions, thus determine a finite collection of regions with constant gradient defined by activation boundaries, which we now formalise.

Definition 4.2. Let $\alpha \in \Lambda$ where Λ is an index set. A *linear domain* $U^\alpha \subseteq \mathbb{R}^N$ of f_w is a connected open set such that:

1. f_w is a simple plane with constant gradient and bias when f_w is restricted to U^α, that is,

$$f_w|_{U^\alpha}(x) = \langle w^\alpha, x\rangle + b^\alpha$$

for some $w^\alpha \in \mathbb{R}^N$ and $b^\alpha \in \mathbb{R}$, and;

2. U^α is the maximal such set for which this plane is defined.

Since any network only contains finitely many nodes, $|\Lambda|$ is necessarily finite.

Remark 4.2. Note that w^α and b^α are the sum of the weights and biases that are active in the region U^α, and also absorb the gradients q_i and bias c. Given some network f_w, the precise size of Λ is non-trivial in general. [1]

Definition 4.3. Let $i \in [d]$ be a non-degenerate node of $f_w(x)$ and let $U^\alpha \in \mathbb{R}^2$ be a linear domain. We say node i is *active* in U^α if $\langle w_i, x\rangle + b_i \geq 0$ for all $x \in U^\alpha$. The *activation boundary* associated to i is the hyperplane

$$H_i = \{x \in \mathbb{R}^N \mid \langle w_i, x\rangle + b_i = 0\}\,. \tag{4.2}$$

When $N = 2$, H_i is merely a line in $\mathbb{R}^2$. We say H_i is degenerate if the corresponding node i is degenerate, meaning H_i is either empty or all of $\mathbb{R}^N$.

Remark 4.3. Since each w_i vector is normal to the activation boundary it defines, recall from Fig. 2.3b that these vectors "point" to the regions in which they are active when anchored on the activation boundary they define.

Following from [PL19], we can make sense of these activation boundaries in the context of foldsets:

Definition 4.4. Let $\mathcal{Z} \subseteq \mathbb{R}^N$ be open and $g : \mathcal{Z} \to \mathbb{R}$ a continuous piecewise linear function. The *foldset* of g is

$$\mathcal{F}(g) = \left\{ x \in \mathcal{Z} \,\middle|\, g \text{ is not differentiable at } x \right\}.$$

Lemma 4.1. *Let $f_w : \mathbb{R}^N \to \mathbb{R}$ be a two-layer feedforward ReLU neural network as in (2.2). Then f_w is continuous, and*

$$\mathcal{F}(f_w) = \bigcup_{i=1}^{d} \{H_i \mid i \text{ is non-degenerate}\}\,. \tag{4.3}$$

Proof. Since $\mathrm{ReLU}(x)$ is continuous and f_w is a sum of ReLU's composed with affine functions, continuity is clear. Note that $\mathrm{ReLU}(x)$ is non-differentiable at $x = 0$ since for $x \neq 0$ we have

$$\frac{d}{dx}\mathrm{ReLU}(x) = \begin{cases} 1 & x > 0 \\ 0 & x < 0 \end{cases}$$

which is clearly discontinuous at $x = 0$. Any degenerate nodes will have at most constant contribution to $f_w(x)$, so these nodes won't contribute to $\mathcal{F}(f_w)$. For any non-degenerate node i, observing the form of $f_w(x)$ in Eq. (2.2) shows that f_w is non-differentiable when $\langle w_i, x\rangle + b_i = 0$, which is the activation boundary H_i, giving the result. □

By definition, the linear domains and foldsets are related by

$$\bigcup_{\alpha\in\Lambda} U^\alpha = \mathbb{R}^2 \setminus (\mathcal{F}(f_w))\,.$$

[1] The reader is referred to [Iva10] for an interesting discussion when these are lines in the plane for $N = 2$.

4.2 Classification of W_0 for $m = d$

Recall from Definition 2.8 that the set of true parameters is defined by

$$W_0 = \{w \in W \mid p(y|x, w) = q(y|x)\}.$$

In classifying W_0 we are obviously only interested in the case that it is non-empty, hence we may assume that $q(y|x) = p(y|x, w^{(0)})$, where we take $w^{(0)} \in W$ to be a fixed parameter defining a two-layer feedforward ReLU neural network $f_0(x) := f(x, w^{(0)})$ with two inputs and one output. Since probability distributions are uniquely defined (up to a set of measure zero), this condition implies that $p(y|x, w)$ and $p(y|x, w^{(0)})$, under Hypothesis 2.1, must have the same mean. Thus

$$W_0 = \{w \in W \mid f(x, w) = f_0(x)\},$$

and so the task of classifying W_0 becomes classifying functional equivalence of this class of networks. We begin in the simplest case where f_0 is activation-distinguished and minimal, and the model and truth have the same number of parameters.

This section is dedicated to proving the following classification theorem. Suppose the model and truth have the same number of hidden nodes, and the true network is minimal and activation-distinguished. Then we will show in Theorem 4.7 that W_0 exhibits three kinds of symmetry:

- *Scaling symmetry* of the incoming and outgoing weights to any node.
- *Permutation symmetry* of the hidden nodes.
- *Orientation reversing symmetry* of the weights, only allowed for collections of weights which sum to zero.

The key observation that guides this proof is that the foldsets of the model and truth must be equal. Once this is accounted for, the rest is merely a matter of good bookkeeping.

4.2.1 Definitions and hypotheses

Inspired by definitions in [Sus92] which deals with tanh neural networks, we begin with some simplifying definitions.

Definition 4.5. Let $\mathcal{Z} \subseteq \mathbb{R}^2$ be an open set. Given a two-layer network $f_w : \mathcal{Z} \to \mathbb{R}$ with d hidden nodes, f_w is *minimal* if: given another two-layer network $f'_{w'} : \mathcal{Z} \to \mathbb{R}$ with $d' \leq d$ hidden nodes such that $f_w(x) = f'_{w'}(x)$ for all $x \in \mathcal{Z}$, then the number of hidden nodes are necessarily equal, $d = d'$. We say f_w is *activation-distinguished* if each non-degenerate H_i is unique, that is, $H_i \neq H_j$ for each $i \neq j \in [d]$.

Remark 4.4. Each node of a minimal network is necessarily non-degenerate.

To see why we impose the activation-distinguished condition, observe that if $f(x, w)$ has d unique activation boundaries then it is necessarily minimal, but the converse is not necessarily true:

Example 4.1. Consider $w = (w_1, w_2, b_1, b_2, q_1, q_2, c) = ((1, 1), (-1, -1), 0, 0, 1, 1, 0)$, so

$$f(x, w) = \mathrm{ReLU}(x_1 + x_2) + \mathrm{ReLU}(-x_1 - x_2).$$

Then H_1 is the line $x_1 + x_2 = 0$ and H_2 is the line $-x_1 - x_2 = 0$, and so clearly $H_1 = H_2$ as subsets of $\mathbb{R}^2$, thus there is only one unique foldset. However $f(x, w)$ is minimal: suppose there was a one-node neural network $f(x, w_r) = c + q\mathrm{ReLU}(\langle x, r\rangle + b)$ which produced the same input output map. Then this network necessarily has a region of inactivation by the definition of ReLU, $\langle x, r\rangle + b < 0$, and $f(x, w_r)$ has zero gradient in this region. But $f(x, w)$ has non-zero gradient in both regions $\{(x_1, x_2) \mid x_2 \geq x_1\}$ and $\{(x_1, x_2) \mid x_2 \leq x_1\}$, thus we could not have $f(x, w_r) = f(x, w)$.

Remark 4.5. It is possible to show that n non-parallel lines divide the plane into $\mathcal{L}_n = \frac{n(n+1)}{2} + 1$ different regions. As such, in principle one can check for the minimality of a given network f by counting $|\Lambda|$ and comparing this to $\mathcal{L}_n$.

For the remainder of this section we will enforce some simplifying hypotheses:

Hypothesis 4.1. Let $q(y|x)$ be a realisable true distribution defined by a two-layer feedforward ReLU neural network $f_0 := f(\cdot, w^{(0)}) : \mathbb{R}^2 \to \mathbb{R}^1$ with m hidden nodes for some fixed parameter $w^{(0)} \in W$. We call f_0 the true network. Let $p(y|x, w)$ be the model as in Hypothesis 2.1 defined by a two-layer feedforward ReLU neural network $f : \mathbb{R}^2 \times W \to \mathbb{R}^1$ with d hidden nodes. We further impose:

- $f_0(x)$ is minimal and distinguished, so all activation boundaries H_i are non-degenerate and unique.
- The width of the true network and the model are equal, so $m = d$.
- $w^{(0)}$ is defined by

$$w^{(0)} = (w_1^{(0)}, \ldots, w_m^{(0)}, b_1^{(0)}, \ldots, b_m^{(0)}, q_1^{(0)}, \ldots, q_m^{(0)}, c^{(0)}),$$

  where $w_i^{(0)} \in \mathbb{R}^2 \backslash \{(0,0)\}$, $b_i^{(0)} \in \mathbb{R}$, $q_i^{(0)} \in \mathbb{R} \backslash \{0\}$ and $c^{(0)} \in \mathbb{R}$ for each $i \in [d]$, since every node is non-degenerate.

To characterise W_0, we thus set $f(x, w) = f_0(x)$ and describe the possible values of w.

Remark 4.6. Clearly the model $f(x, w)$ must also be minimal and activation-distinguished.

Armed with these formulations we are thus ready to begin investigating the symmetries of W_0.

4.2.2 Lemmas

The results of Lemma 4.2, Lemma 4.4, Lemma 4.5, Lemma 4.6 and Theorem 4.7 all assume the conditions of Hypothesis 4.1. The fundamental observation that guides this classification is that the foldsets of a network are the pivotal piece of data. Given that the foldsets of the model and truth must be equal, this restricts the classification process to observing equivalence of lines in the plane, and then ensuring the gradients and biases are equivalent.

Beginning with this observation, the first form of symmetry is a simple permutation of nodes.

Lemma 4.2. *Let H_i denote the activation boundaries associated to $f(x, w)$ as in (4.2) and let $H_i^{(0)}$ denote the activation boundaries associated to $f_0(x)$. Then there exists some permutation $\sigma \in S_m$ (the symmetric group of order m) such that $H_i = H_{\sigma(i)}^{(0)}$ for each $i \in [d]$.*

Proof. Since $f(x, w) = f_0(x)$ for all $x \in \mathbb{R}^2$, and $f(x, w)$ must also be minimal, they must also be non-differentiable at the same points, so

$$\mathcal{F}(f(x, w)) = \bigcup_{i=1}^{d} H_i = \bigcup_{j=1}^{m} H_j^{(0)} = \mathcal{F}(f_0(x)).$$

Each H_i and $H_j^{(0)}$ are distinct by hypothesis, which we will show implies there is a unique $j \in [m]$ such that $H_i = H_j^{(0)}$. First observe that there can only be finitely many intersection points of the lines $\{H_i\}_{i \in [d]}$ (and similarly for $\{H_j^{(0)}\}_{j \in [m]}$), thus let $x \in \mathcal{F}(f(x, w))$ be a non-intersection point uniquely associated to a line H_i for some $i \in [d]$, so $x \in H_i \setminus \left(\bigcup_{j \neq i} H_j\right)$. Then since the foldsets are equal, $x \in \mathcal{F}(f_0(x))$, and since the points of intersection of the sets $\{H_i\}_{i \in [d]}$ coincide with those of $\{H_j^{(0)}\}_{j \in [m]}$, x is a non-intersection point of the true foldset lines and thus is uniquely associated to a line $H_j^{(0)}$.

To see that this implies $H_i = H_j^{(0)}$ for all $x \in H_i$, suppose we pick $d+1$ non-intersection points on H_i. Then since $m = d$, by the pigeonhole principle there must be at least two points associated to some $H_j^{(0)}$. But since lines are uniquely determined by two points, this implies $H_i = H_j^{(0)}$ for all non-intersection points $x \in H_i$. By continuity this also applies to intersection points, thus there is a unique $j \in [m]$ such that $H_i = H_j^{(0)}$.

Finally, since $m = d$, our previous statement simply says that there is a unique bijection $\sigma : [d] \to [m]$ associating lines in $\mathcal{F}(f_0(x))$ to lines in $\mathcal{F}(f(x,w))$, thus giving the desired $\sigma \in S_m$. □

Lemma 4.2 induces a scaling symmetry in the weights and biases. Before demonstrating this, we begin with a more general result.

Lemma 4.3. *Let $w, w' \in \mathbb{R}^2 \setminus \{0\}$ and $b, b' \in \mathbb{R}$ be given and let*

$$H = \{x \in \mathbb{R}^2 \mid \langle w, x\rangle + b = 0\}, \quad \text{and} \quad H' = \{x \in \mathbb{R}^2 \mid \langle w', x\rangle + b' = 0\}.$$

Then $H = H'$ if and only if there exists some scalar $\lambda \in \mathbb{R} \setminus \{0\}$ such that $w = \lambda w'$ and $b = \lambda b'$.

Proof. See Lemma A.6. □

Lemma 4.4. *Given $f(x,w) = f_0(x)$, there exists a unique $\sigma \in S_m$, and for each $i \in [d]$ there exists an $\epsilon_i \in \mathbb{Z}_2$ and $\lambda_i \in \mathbb{R}_{>0}$, such that*

$$w_i = (-1)^{\epsilon_i}\lambda_i w_{\sigma(i)}^{(0)}, \quad \text{and} \quad b_i = (-1)^{\epsilon_i}\lambda_i b_{\sigma(i)}^{(0)}, \quad \text{where} \quad \lambda_i = \frac{q_{\sigma(i)}^{(0)}}{q_i},$$

meaning q_i and $q_{\sigma(i)}^{(0)}$ necessarily have the same sign.

Proof. Lemma 4.2 gives the permutation $\sigma \in S_m$ relating the activation boundaries of $f(x,w)$ and $f_0(x)$. For each $i \in [d]$ Lemma 4.3 gives us a $\mu_i \in \mathbb{R}\setminus\{0\}$, which we can decompose into $\mu = (-1)^{\epsilon_i}\lambda_i$ for some $\epsilon_i \in \mathbb{Z}_2$ and $\lambda_i \in \mathbb{R}_{>0}$ such that we can initially write

$$w_i = (-1)^{\epsilon_i}\lambda_i w_{\sigma(i)}^{(0)}, \quad \text{and} \quad b_i = (-1)^{\epsilon_i}\lambda_i b_{\sigma(i)}^{(0)}. \tag{4.4}$$

Let $i \in [d]$ be fixed and fix a non-intersection point $x \in H_i \setminus \left(\bigcup_{j\neq i} H_j\right)$. Let $U \ni x$ be a sufficiently small open ball around x such that $U \cap \left(\bigcup_{j\neq i} H_j\right) = \varnothing$, which exists since there are only finitely many (and thus isolated) points of intersection. Recall that by hypothesis $H_i \neq H_j$ for any other $j \neq i$, thus excluding the possibility of any other node being activated across the boundary H_i. Then the vector w_i emanating from x points towards a unique linear domain U^α for which x is on the closure of U^α, and node i is active in U^α. Similarly, $-w_i$ points towards a different unique linear domain U^β satisfying the same closure property, but for which node i is inactive in U^β. Thus we can find a unique decomposition $U = U^- \cup U^+$ where

$$U^- = U \cap U^\beta, \quad \text{and} \quad U^+ = U \cap U^\alpha$$

Then we have

$$f|_{U^+}(x,w) = f|_{U^-}(x,w) + q_i(\langle w_i, x\rangle + b_i).$$

Similarly, consider the same set up for the line $H_{\sigma(i)}^{(0)} = H_i$ associated to $f_0(x)$, where $U_0 = U$ is the same sufficiently small neighbourhood and U_0^- and U_0^+ are the regions of inactivation and activation (referring to f_0) respectively. Note that the orientation of w_i will determine whether $U^+ = U_0^+$ or $U^+ = U_0^-$, and similarly for U^-. Explicitly, we then have

$$f_0|_{U^+}(x) = f_0|_{U^-}(x) + q_{\sigma(i)}^{(0)}\left(\langle w_{\sigma(i)}^{(0)}, x\rangle + b_{\sigma(i)}^{(0)}\right). \tag{4.5}$$

First suppose $\epsilon_i = 0$, so w_i and $w^{(0)}_{\sigma(i)}$ are oriented in the same direction, hence $U^+ = U^+_0$ and $U^- = U^-_0$. Since $f(x,w) = f_0(x)$ we have

$$\begin{aligned} q_i(\langle w_i, x\rangle + b_i) &= f|_{U^+}(x,w) - f|_{U^-}(x,w) \\ &= f_0|_{U^+}(x) - f_0|_{U^-}(x) = q^{(0)}_{\sigma(i)}\left(\langle w^{(0)}_{\sigma(i)}, x\rangle + b^{(0)}_{\sigma(i)}\right), \end{aligned}$$

and so by comparing polynomial coefficients we must have

$$w_i = \frac{q^{(0)}_{\sigma(i)}}{q_i} w^{(0)}_{\sigma(i)}, \quad \text{and} \quad b_i = \frac{q^{(0)}_{\sigma(i)}}{q_i} b^{(0)}_{\sigma(i)}. \tag{4.6}$$

If $\epsilon_1 = 1$ then w_i and $w^{(0)}_{\sigma(i)}$ are oriented in different directions, thus $U^+ = U^-_0$ and $U^- = U^+_0$ so we have

$$\begin{aligned} q_i\left(\langle w_i, x\rangle + b_i\right) &= f|_{U^+}(x,w) - f|_{U^-}(x,w) \\ &= f_0|_{U^-}(x) - f_0|_{U^+}(x) = -q^{(0)}_{\sigma(i)}\left(\langle w^{(0)}_{\sigma(i)}, x\rangle + b^{(0)}_{\sigma(i)}\right) \end{aligned}$$

and so again comparing coefficients we have

$$w_i = -\frac{q^{(0)}_{\sigma(i)}}{q_i} w^{(0)}_{\sigma(i)}, \quad \text{and} \quad b_i = -\frac{q^{(0)}_{\sigma(i)}}{q_i} w^{(0)}_{\sigma(i)}. \tag{4.7}$$

Combining (4.6) with (4.7) we can write

$$w_i = (-1)^{\epsilon_i}\frac{q^{(0)}_{\sigma(i)}}{q_i} w^{(0)}_{\sigma(i)}, \quad \text{and} \quad b_i = (-1)^{\epsilon_i}\frac{q^{(0)}_{\sigma(i)}}{q_i} w^{(0)}_{\sigma(i)},$$

thus giving $\lambda_i = \frac{q^{(0)}_{\sigma(i)}}{q_i}$ as advertised. Since λ_i must be positive by (4.4), we see that q_i necessarily has the same sign as $q^{(0)}_{\sigma(i)}$. □

At first glance, one may assume that all orientations must be preserved, that is, all $\epsilon_i = 0$, to yield functional equivalence. But as the next example shows, this is not necessarily the case.

Example 4.2. Consider a simple one dimensional ReLU neural network $f : \mathbb{R} \times W \to \mathbb{R}$ with $d = 2$ hidden nodes. Defining $w^{(0)}, w \in W$ such that

$$\begin{aligned} f_0(x) &= \mathrm{ReLU}(x+1) + \mathrm{ReLU}(-x+1), \\ \text{and} \quad f(x,w) &= 2 + \mathrm{ReLU}(-x-1) + \mathrm{ReLU}(x-1), \end{aligned}$$

we see that $f(x,w) = f_0(x)$ and so $w = (-1,1,-1,-1,1,1,2)$ is also a true parameter for $w^{(0)} = (1,-1,1,1,1,1,0)$. Notice that here we have $w_1 + b_1 = -(w^{(0)}_1 + b^{(0)}_1)$ and similarly $w_2 + b_2 = -(w^{(0)}_2 + b^{(0)}_2)$, meaning $\epsilon_1 = \epsilon_2 = 1$. But this works because the weights of both networks sum to zero.

We use this example as the inspiration for the orientation reversing symmetry. We let $G_\alpha(f(x,w)) \in \mathbb{R}^2$ denote the gradient computed by $f(x,w)$ in the linear domain U^α, and similarly $\mathcal{B}_\alpha(f(x,w)) \in \mathbb{R}$ the bias.

Lemma 4.5. *Let $E = \{i \in [d] \mid \epsilon_i = 1\}$. Given $\sigma \in S_m$, $\lambda_i \in \mathbb{R}\backslash\{0\}$ and $\epsilon_i \in \mathbb{Z}_2$ from Lemma 4.4, the weights of the true network necessarily satisfy $\sum_{i\in E} q^{(0)}_{\sigma(i)} w^{(0)}_{\sigma(i)} = 0$.*

Proof. Let U^α be a fixed domain associated to $f(x,w)$. For notational convenience, define

$$\delta^\alpha_j := \begin{cases} 1 & \text{if} \quad \langle w^{(0)}_j, x\rangle + b^{(0)}_j \geq 0 \\ 0 & \text{otherwise} \end{cases}, \quad \text{and} \quad \overline{\delta_j}^{\alpha} := 1 - \delta^\alpha_j, \tag{4.8}$$

thus δ_j^α indicates whether node $j \in [m]$ of $f_0(x)$ is active in the region U^α, and $\overline{\delta_j^\alpha}$ indicates the converse. Let $i \in [d]$ be a node of $f(x,w)$ and $\sigma(i)$ the corresponding node of $f_0(x)$ such that $H_i = H^{(0)}_{\sigma(i)}$.

Following the result of Lemma 4.4 we can calculate

$$f(x,w) = c + \sum_{i=1}^{d} q_i \mathrm{ReLU}\left(\langle w_i, x\rangle + b_i\right) = c + \sum_{i=1}^{d} q_i \lambda_i \mathrm{ReLU}\left((-1)^{\epsilon_i}\left(\langle w^{(0)}_{\sigma(i)}, x\rangle + b^{(0)}_{\sigma(i)}\right)\right),$$

where $\lambda_i = \frac{q_i^{(0)}}{q_i} > 0$. In particular, the single node map

$$\mathrm{ReLU}(\langle w_i, x\rangle + b_i) \longmapsto \lambda_i \,\mathrm{ReLU}\left((-1)^{\epsilon_i}\left(\langle w^{(0)}_{\sigma(i)}, x\rangle + b^{(0)}_{\sigma(i)}\right)\right)$$

shows that if node $\sigma(i)$ is active in U^α, then we are in one of two situations: either $\epsilon_i = 0$ and $\delta^\alpha_{\sigma(i)} = 1$, or $\epsilon_i = 1$ and $\overline{\delta}^\alpha_{\sigma(i)} = 1$. We can then equate gradients $\mathcal{G}_\alpha(f(x,w)) = \mathcal{G}_\alpha(f_0(x))$, and recall $q^{(0)}_{\sigma(i)} w^{(0)}_{\sigma(i)} = (-1)^{\epsilon_i} q_i w_i$ from Lemma 4.4, to calculate

$$\sum_{i\in[d]} \delta^\alpha_{\sigma(i)} q^{(0)}_{\sigma(i)} w^{(0)}_{\sigma(i)} = \sum_{i\notin E} \delta^\alpha_{\sigma(i)} q_i w_i + \sum_{i\in E} \overline{\delta}^\alpha_{\sigma(i)} q_i w_i = \sum_{i\notin E} \delta^\alpha_{\sigma(i)} q^{(0)}_{\sigma(i)} w^{(0)}_{\sigma(i)} + \sum_{i\in E} \overline{\delta}^\alpha_{\sigma(i)} q^{(0)}_{\sigma(i)} (-w^{(0)}_{\sigma(i)}) \,.$$

But by definition we have

$$\sum_{i\in[d]} \delta^\alpha_{\sigma(i)} q^{(0)}_{\sigma(i)} w^{(0)}_{\sigma(i)} = \sum_{i\notin E} \delta^\alpha_{\sigma(i)} q^{(0)}_{\sigma(i)} w^{(0)}_{\sigma(i)} + \sum_{i\in E} \delta^\alpha_{\sigma(i)} q^{(0)}_{\sigma(i)} w^{(0)}_{\sigma(i)} \,,$$

so subtracting one from the other shows that

$$\sum_{i\in E} \left(\delta^\alpha_{\sigma(i)} + \overline{\delta}^\alpha_{\sigma(i)}\right) q^{(0)}_{\sigma(i)} w^{(0)}_{\sigma(i)} = \sum_{i\in E} q^{(0)}_{\sigma(i)} w^{(0)}_{\sigma(i)} = 0 \,.$$

□

Example 4.2 shows us that we can expect a similar result for the biases.

Lemma 4.6. *With E as in Lemma 4.5, given $\sigma \in S_m$ and $\epsilon_i \in \mathbb{Z}_2$ from Lemma 4.4, the biases of the true network satisfy $c_0 + \sum_{i\in E} q^{(0)}_{\sigma(i)} b^{(0)}_{\sigma(i)} = c$.*

Proof. Let U^α be a fixed linear domain, so the same arguments from Lemma 4.5 regarding active nodes on U^α apply. The bias c is active on every domain, so equating $\mathcal{B}(f(x,w)) = \mathcal{B}(f_0(x))$ gives

$$\begin{aligned} c_0 + \sum_{i\in[d]} \delta^\alpha_{\sigma(i)} q^{(0)}_{\sigma(i)} b^{(0)}_{\sigma(i)} &= c + \sum_{i\notin E} \delta^\alpha_{\sigma(i)} q_i b_i + \sum_{i\in E} \overline{\delta}^\alpha_{\sigma(i)} q_i b_i \\ &= c + \sum_{i\notin E} \delta^\alpha_{\sigma(i)} q^{(0)}_{\sigma(i)} b^{(0)}_{\sigma(i)} + \sum_{i\in E} \overline{\delta}^\alpha_{\sigma(i)} q^{(0)}_{\sigma(i)} (-b^{(0)}_{\sigma(i)}) \,, \end{aligned}$$

but again recalling

$$\sum_{i\in[d]} \delta^\alpha_{\sigma(i)} q^{(0)}_{\sigma(i)} b^{(0)}_{\sigma(i)} = \sum_{i\notin E} \delta^\alpha_{\sigma(i)} q^{(0)}_{\sigma(i)} b^{(0)}_{\sigma(i)} + \sum_{i\in E} \delta^\alpha_{\sigma(i)} q^{(0)}_{\sigma(i)} b^{(0)}_{\sigma(i)} \,,$$

we thus have

$$c_0 + \sum_{i\in E} \delta^\alpha_{\sigma(i)} q^{(0)}_{\sigma(i)} b^{(0)}_{\sigma(i)} = c - \sum_{i\in E} \overline{\delta}^\alpha_{\sigma(i)} q^{(0)}_{\sigma(i)} b^{(0)}_{\sigma(i)} \,,$$

and so $c_0 + \sum_{i\in E} q^{(0)}_{\sigma(i)} b^{(0)}_{\sigma(i)} = c$. □

Remark 4.7. Our convention is to take the empty sum to be zero, so if E is empty then we have $c = c_0$.

4.2.3 Main theorem for $m = d$

We have thus arrived at the main theorem of this section having classified all of the symmetries of W_0 for $m = d$. We first introduce some notation. Let $\sigma \in S_m$ be a fixed permutation and define

$$X_i := \left\{ (\lambda_i, q_i) \in \mathbb{R}_{>0} \times \mathbb{R} \,\middle|\, \lambda_i q_i = q^{(0)}_{\sigma(i)} \right\},$$

$$\text{and} \quad \Upsilon := \left\{ \epsilon : [d] \to \mathbb{Z}_2 \,\middle|\, \sum_{i \in E} q^{(0)}_{\sigma(i)} w^{(0)}_{\sigma(i)} = 0\,, \text{ and } c_0 + \sum_{i \in E} q^{(0)}_{\sigma(i)} b^{(0)}_{\sigma(i)} = c \right\},$$

where $E = \{i \in [d] \mid \epsilon_i = 1\}$.

Theorem 4.7. *Let $f_0 : \mathbb{R}^2 \to \mathbb{R}^1, f_0(x) := f(x, w^{(0)})$ be an activation distinguished minimal feedforward ReLU neural network with two-layers, d hidden nodes defined by some fixed parameter $w^{(0)} \in W$. Then there is a bijection*

$$\Psi : \prod_{i=1}^{m} X_i \times S_m \times \Upsilon \xrightarrow{\cong} W_0$$

$$\left((\lambda_i, q_i)_{i=1}^m, \sigma, \epsilon\right) \longmapsto \left(\left((-1)^{\epsilon_i} \lambda_i w^{(0)}_{\sigma(i)} \right)_{i=1}^d, \left((-1)^{\epsilon_i} \lambda_i b^{(0)}_{\sigma(i)} \right)_{i=1}^d, (q_i)_{i=1}^d, c_0 + \sum_{i \in E} q^{(0)}_{\sigma(i)} b^{(0)}_{\sigma(i)} \right).$$

We refer to $\prod_{i=1}^m X_i$ as *scaling symmetry*, S_m as *permutation symmetry* and Υ as *orientation reversing* symmetry.

Proof. We first verify that Ψ is well defined, that is, $f(x, \Psi(\theta)) = f_0(x)$ as functions. We compute for a fixed $\theta = ((\lambda_i, q_i)_{i=1}^m, \sigma, \epsilon) \in \prod_{i=1}^m X_i \times S_m \times \Upsilon$

$$\begin{aligned} f(x, \Psi(\theta)) &= c_0 + \sum_{i \in E} q^{(0)}_{\sigma(i)} b^{(0)}_{\sigma(i)} + \sum_{i=1}^{d} q_i \mathrm{ReLU}\left(\left\langle (-1)^{\epsilon_i} \lambda_i w^{(0)}_{\sigma(i)}, x \right\rangle + (-1)^{\epsilon_i} \lambda_i b^{(0)}_{\sigma(i)} \right) \\ &= c_0 + \sum_{i \in E} q^{(0)}_{\sigma(i)} b^{(0)}_{\sigma(i)} + \sum_{i=1}^{d} q_i \lambda_i \mathrm{ReLU}\left((-1)^{\epsilon_i} \left(\left\langle w^{(0)}_{\sigma(i)}, x \right\rangle + b^{(0)}_{\sigma(i)} \right) \right) \\ &= c_0 + \sum_{i \in E} q^{(0)}_{\sigma(i)} b^{(0)}_{\sigma(i)} + \sum_{i=1}^{d} q^{(0)}_{\sigma(i)} \mathrm{ReLU}\left((-1)^{\epsilon_i} \left(\left\langle w^{(0)}_{\sigma(i)}, x \right\rangle + b^{(0)}_{\sigma(i)} \right) \right). \end{aligned}$$

Thus we see that $f(x, \Psi(\theta))$ has the same foldsets as $f_0(x)$. It remains to check the gradients and biases in any domain U^α agree. Let $\delta^\alpha_{\sigma(i)}$ be as in Lemma 4.5 so it refers to active nodes of $f_0(x)$. Then for any domain U^α the gradient computed by $f_0(x)$ is

$$\mathcal{G}_\alpha(f_0(x)) = \sum_{i \in [d]} \delta^\alpha_{\sigma(i)} q^{(0)}_{\sigma(i)} w^{(0)}_{\sigma(i)},$$

whereas the gradient computed by $f(x, \Psi(\theta))$ is

$$\begin{aligned} \mathcal{G}_\alpha(f(x, \Psi(\theta))) &= \sum_{i \notin E} \delta^\alpha_{\sigma(i)} q^{(0)}_{\sigma(i)} w^{(0)}_{\sigma(i)} + \sum_{i \in E} \overline{\delta}^\alpha_{\sigma(i)} q^{(0)}_{\sigma(i)} (-w^{(0)}_{\sigma(i)}) \\ &= \sum_{i \notin E} \delta^\alpha_{\sigma(i)} q^{(0)}_{\sigma(i)} w^{(0)}_{\sigma(i)} + \sum_{i \in E} \delta^\alpha_{\sigma(i)} q^{(0)}_{\sigma(i)} w^{(0)}_{\sigma(i)} = \sum_{i} \delta^\alpha_{\sigma(i)} q^{(0)}_{\sigma(i)} w^{(0)}_{\sigma(i)} = \mathcal{G}_\alpha(f_0(x)), \end{aligned}$$

where the second equality followed from our hypothesis on $\epsilon \in \Upsilon$. Similarly, the bias computed by $f_0(x)$ is

$$\mathcal{B}_\alpha(f_0(x)) = c_0 + \sum_{i \in [d]} \delta^\alpha_{\sigma(i)} q^{(0)}_{\sigma(i)} b^{(0)}_{\sigma(i)}$$

whereas for $f(x, \Psi(\theta))$ it is

$$\begin{aligned}\mathcal{B}_\alpha(f(x,\Psi(x))) &= c_0 + \sum_{i\in E} q^{(0)}_{\sigma(i)} b^{(0)}_{\sigma(i)} + \sum_{i\notin E} \delta^\alpha_{\sigma(i)} q^{(0)}_{\sigma(i)} b^{(0)}_{\sigma(i)} + \sum_{i\in E} \overline{\delta}^\alpha_{\sigma(i)} q^{(0)}_{\sigma(i)} (-b^{(0)}_{\sigma(i)}) \\ &= c_0 + \sum_{i\in E} q^{(0)}_{\sigma(i)} b^{(0)}_{\sigma(i)} + \sum_{i\in[d]} \delta^\alpha_{\sigma(i)} q^{(0)}_{\sigma(i)} b^{(0)}_{\sigma(i)} - \sum_{i\in E} q^{(0)}_{\sigma(i)} b^{(0)}_{\sigma(i)} \\ &= c_0 + \sum_{i\in[d]} \delta^\alpha_{\sigma(i)} q^{(0)}_{\sigma(i)} b^{(0)}_{\sigma(i)} = \mathcal{B}_\alpha(f_0(x))\,,\end{aligned}$$

where the second equality follows from $\overline{\delta}^\alpha_{\sigma(i)} = 1 - \delta^\alpha_{\sigma(i)}$. Therefore we have $f(x, \Psi(\theta)) = f_0(x)$ and so Ψ is well defined.

For injectivity, if $\Psi(\theta) = \Psi(\theta')$ for suitable $\theta, \theta' \in \prod_{i=1}^m X_i \times S_m \times \Upsilon$ then we can read off $q_i = q'_i$, hence $\lambda_i = \lambda'_i$ for all i. We may then compare foldsets of $f(x, \Psi(\theta))$ and $f(x, \Psi(\theta'))$ which are identically labelled to recover $\sigma = \sigma'$. Finally we may read off each ϵ_i and ϵ'_i from each weight entry, thus $\epsilon = \epsilon'$ and so $\theta = \theta'$.

For surjectivity, if $f(x, w) = f_0(x)$ then they must have identical foldsets as in Lemma 4.2, where the weights and biases must be equal up to scaling from Lemma 4.3, and ϵ can only be non-zero under the hypothesis on Υ shown in Lemma 4.5. Thus Ψ is a bijection. □

Remark 4.8. We now see that the activation-distinguished condition allows us to uniquely identify the permutation σ relating activation boundaries, and ensures only one node changes across each boundary.

4.3 Classification of W_0 for $m < d$

The key assumption of the previous section was that the number of parameters in the model d and (minimal) truth m were equal. Let us now weaken this condition and examine the case $m < d$. In this section we will show that of the d nodes associated to the model, the symmetries associated to m of them are the same as in Theorem 4.7 - without loss of generality let these nodes be $[m]$. By contrast, each "excess" node $i \in \{m+1, \ldots, d\}$ will either be

- Degenerate, so $q_i = 0$ or $w_i = 0$, or;
- Have the same activation boundary as a node in $[m]$.

Thus in this case W_0 has degenerate-node symmetries, suitably adjusted scaling symmetries, suitably adjusted permutation symmetries and almost identical orientation reversing symmetry.

4.3.1 Hypotheses, definitions and lemmas

Hypothesis 4.2. We assume the same conditions as in Hypothesis 2.1, but with the number of hidden nodes m in the true network $f_0(x)$ strictly less than those d in the model $f(x, w)$, so $m < d$. In particular, we assume that $f_0(x)$ is still minimal and activation-distinguished.

We assume the conditions of Hypothesis 4.2 for Lemma 4.8, Lemma 4.9, Lemma 4.10 and Theorem 4.11.

Lemma 4.8. *Let H_i denote the activation boundaries associated to the model for $i \in [d]$, and $H^{(0)}_j$ those to the truth for $j \in [m]$. Let $K = d - m > 0$ denote the number of excess parameters in the model. Then:*

1. *There exists a $0 \leq k \leq K$ such that k nodes of $f(x, w)$ (and thus their corresponding activation boundaries) are degenerate.*

2. *Consider the remaining $d' = K - k \geq m$ nodes of the model. Without loss of generality we may write this set as $\{1, \dots, d'\}$. Then there is a surjective finite set map*

$$\pi : \{1, \dots, m, m+1, \dots d'\} \longrightarrow \{1, \dots, m\}$$

such that $H_i = H^{(0)}_{\pi(i)}$ for each $i \in [d']$. If $d' = m$ then π is a bijection.

Proof. Once again, the key observation is that since $f(x, w) = f_0(x)$, their foldsets must also be equal, thus

$$\mathcal{F}(f_0(x)) = \bigcup_{j=1}^{m} H_j^{(0)} = \bigcup_{i=1}^{d} \{H_i \mid i \text{ is non-degenerate}\} = \mathcal{F}(f(x, w))\,.$$

Since f_0 is minimal and activation-distinguished, so $\bigcup_{j=1}^{m} H_j^{(0)}$ comprises m distinct lines in the plane, the model requires at least m non-degenerate activation boundaries for these foldsets to be equal. Without loss of generality, suppose the non-degenerate nodes of the model are $\{1, \dots, m\}$. Thus, there exists a permutation $\sigma \in S_m$ such that $H_i = H^{(0)}_{\sigma(i)}$ for $i \in [m]$. For clarity, let us write

$$\mathcal{F}(f(x, w)) = \left(\bigcup_{i=1}^{m} H_i \right) \bigcup \left(\bigcup_{i=m+1}^{d} \{H_i \mid i \text{ is non-degenerate}\} \right)\,.$$

Since $\bigcup_{i=1}^{m} H_i = \bigcup_{j=1}^{m} H_j^{(0)}$, we see that

$$\bigcup_{i=m+1}^{d} \{H_i \mid i \text{ is non-degenerate}\} \subseteq \bigcup_{j=1}^{m} H_j^{(0)}\,.$$

Let H_i be one of the remaining activation boundaries for $m+1 \leq i \leq d$. Then using the same arguments associating unions of lines to one another as in Lemma 4.2, there are two choices:

1. $\{H_i \mid i \text{ is non-degenerate}\}$ is empty, thus H_i is degenerate, or;
2. $H_i = H_j^{(0)}$ for some $j \in [m]$.

Enumerating these choices over each $m+1 \leq i \leq d$, the model thus has $0 \leq k \leq d-m$ degenerate nodes, and $d' = d - k \geq m$ non-degenerate nodes. Let $\upsilon : \{m+1, \dots, d'\} \to \{1, \dots, m\}$ denote the map given as a result of the second choices above. Combining σ and υ gives a map $\pi : \{1, \dots, d'\} \to \{1, \dots, m\}$ which acts as σ on $\{1, \dots, m\}$, and as υ on $\{m+1, \dots, d'\}$. The map π inherits its surjectivity from σ, and if $d' = m$ then $\pi = \sigma$ and thus is a bijection. □

Remark 4.9. If $d' > m$ then π is clearly non-injective, meaning multiple nodes of the model will share the same activation boundaries.

In light of Lemma 4.8, we can introduce a new kind of network that simplifies the degeneracy property of $f(x, w)$.

Definition 4.6. Let $w \in W \subseteq \mathbb{R}^{4d+1}$, and $f : \mathbb{R}^2 \times W \to \mathbb{R}$ be a two-layer feedforward ReLU network with d hidden nodes. Suppose $f(x, w)$ is an activation-distinguished network such that $0 \leq k \leq d$ nodes are degenerate, and let $d' = d - k$. Then by Lemma 4.8 there exists an

- activation-distinguished two-layer ReLU network $g : \mathbb{R}^2 \times W' \to \mathbb{R}$ with d' hidden nodes such that $W' \subseteq \mathbb{R}^{4d'+1}$ and $W' \subseteq W$, and;
- a parameter $w' \in W'$ with d' non-degenerate nodes equal to the non-degenerate nodes of w;

such that $f(x, w) = g(x, w')$ for all $x \in \mathbb{R}^2$. We call $g(x, w')$ the *degenerate reduced* network of $f(x, w)$.

Since the networks in Definition 4.6 are activation distinguished, this definition uses the first form of symmetry in Lemma 4.8 to remove degenerate nodes from both the function and the vector. Assuming $f(x, w)$ to be in its degenerate reduced form with d' hidden nodes, we can now classify the excess nodes $m+1 \leq i \leq d'$ by appealing to the same scaling symmetry as seen in Lemma 4.3.

Lemma 4.9. *Assume we have the data of Lemma 4.8, where $f(x, w)$ is in its degenerate reduced form with d' hidden nodes. Then for each $i \in [d']$, there exists an $\epsilon_i \in \mathbb{Z}_2$ and $\lambda_i \in \mathbb{R}_{>0}$ such that*

$$w_i = (-1)^{\epsilon_i} \lambda_i w^{(0)}_{\pi(i)}\,, \quad b_i = (-1)^{\epsilon_i} \lambda_i b^{(0)}_{\pi(i)}\,. \tag{4.9}$$

Moreover, let $M_j = \{i \in [d'] \,|\, \pi(i) = j\}$ for $j \in [m]$. Then the λ_i are constrained such that

$$\sum_{i \in M_j} q_i \lambda_i = q^{(0)}_j\,.$$

Proof. The proof is nearly identical to Lemma 4.4. Lemma 4.3 gives Eq. (4.9) since $H_i = H^{(0)}_{\pi(i)}$ for each $i \in [d']$. As in Lemma 4.4, we analyse a small neighbourhood U centred at an isolated non-intersection point $x \in H_i \setminus \left(\bigcup_{j \neq i} H_j\right)$, and let U^+ be the region where the truth is active and U^- where it is inactive. Let $\chi^{\pm}_i$ indicate when node i of the model is active in the regions $U^{\pm}$. Then

$$f|_{U^\pm}(x, w) = c + \sum_{i \in [d']} \chi^{\pm}_i q_i \big(\langle w_i, x\rangle + b_i\big)\,.$$

In particular, letting $E_j = \{i \in [d'] \,|\, \pi(i) = j\,,\, \epsilon_i = 1\}$ we have

$$f|_{U^+}(x, w) = c + \sum_{i \notin M_j} \chi^{+}_i q_i \big(\langle w_i, x\rangle + b_i\big) + \sum_{i \notin E_j} q_i \lambda_i \left(\langle w^{(0)}_{\pi(i)}, x\rangle + b^{(0)}_{\pi(i)}\right)\,,$$

$$\text{and} \quad f|_{U^-}(x, w) = c + \sum_{i \notin M_j} \chi^{-}_i q_i \big(\langle w_i, x\rangle + b_i\big) - \sum_{i \in E_j} q_i \lambda_i \left(\langle w^{(0)}_{\pi(i)}, x\rangle + b^{(0)}_{\pi(i)}\right)\,.$$

The key insight is that the only nodes i that flip activation on this boundary are those with $\pi(i) = j$. That is, for $i \notin M_j$, $\chi^+_i = \chi^-_i$. Thus

$$\begin{aligned} f|_{U^+}(x, w) - f|_{U^-}(x, w) &= \sum_{i \notin E_j} q_i \lambda_i \left(\langle w^{(0)}_{\pi(i)}, x\rangle + b^{(0)}_{\pi(i)}\right) + \sum_{i \in E_j} q_i \lambda_i \left(\langle w^{(0)}_{\pi(i)}, x\rangle + b^{(0)}_{\pi(i)}\right) \\ &= \left(\sum_{i \notin E_j} q_i \lambda_i + \sum_{i \in E_j} q_i \lambda_i\right) \left(\langle w^{(0)}_{\pi(i)}, x\rangle + b^{(0)}_{\pi(i)}\right) \\ &= \left(\sum_{i \in M_j} q_i \lambda_i\right) \left(\langle w^{(0)}_{\pi(i)}, x\rangle + b^{(0)}_{\pi(i)}\right)\,. \end{aligned}$$

But since the model and truth are functionally equivalent, recalling Eq. (4.5) we have

$$\begin{aligned} f|_{U^+}(x, w) - f|_{U^-}(x, w) &= \left(\sum_{i \in M_j} q_i \lambda_i\right) \left(\langle w^{(0)}_{\pi(i)}, x\rangle + b^{(0)}_{\pi(i)}\right) \\ &= q^{(0)}_{\pi(i)} \left(\langle w^{(0)}_{\pi(i)}, x\rangle + b^{(0)}_{\pi(i)}\right) = f_0|_{U^+}(x, w) - f_0|_{U^-}(x, w) \end{aligned}$$

and so by comparing coefficients of the polynomials we get the claim. □

Remark 4.10. It is easy to see that Lemma 4.4 is recovered in this more general result. If $m = d$ then π is a bijection so $|M_j| = 1$ for all $j \in [m]$, thus $q_i \lambda_i = q^{(0)}_{\pi(i)}$.

It turns out that the ϵ_i are constrained in effectively the same way as in the minimal activation-distinguished case.

Lemma 4.10. *Given the data of Lemma 4.8, where $E = \{i \in [d'] \mid \epsilon_i = 1\}$, we have*

$$\sum_{i \in E} q_i \lambda_i w^{(0)}_{\pi(i)} = 0\,, \quad \text{and} \quad c_0 + \sum_{i \in E} q_i \lambda_i b^{(0)}_j = c\,.$$

Proof. We perform the same bookkeeping and notation as in Lemma 4.5, where δ^α_j indicates whether node j of the truth is active in the linear domain U^α. We have from Lemma 4.9

$$\sum_{j \in [m]} \delta^\alpha_j q^{(0)}_j w^{(0)}_j = \sum_{j \in [m]} \left(\sum_{i \in M_j} q_i \lambda_i \right) \delta^\alpha_j w^{(0)}_j = \sum_{j \in [m]} \left(\sum_{i \notin E_j} \delta^\alpha_{\pi(i)} q_i \lambda_i + \sum_{i \in E_j} \delta^\alpha_{\pi(i)} q_i \lambda_i \right) w^{(0)}_j\,. \quad (4.10)$$

Making the same observations as in Lemma 4.5, we also have

$$\begin{aligned}
\sum_{j \in [m]} \delta^\alpha_j q^{(0)}_j w^{(0)}_j &= \sum_{i \notin E} \delta^\alpha_{\pi(i)} q_i w_i + \sum_{i \in E} \overline{\delta^\alpha_{\pi(i)}} q_i w_i \\
&= \sum_{i \notin E} \delta^\alpha_{\pi(i)} (-1)^{\epsilon_i} q_i \lambda_i w^{(0)}_{\pi(i)} + \sum_{i \in E} \overline{\delta^\alpha_{\pi(i)}} (-1)^{\epsilon_i} q_i \lambda_i w^{(0)}_{\pi(i)} \\
&= \sum_{i \notin E} \delta^\alpha_{\pi(i)} q_i \lambda_i w^{(0)}_{\pi(i)} - \sum_{i \in E} \overline{\delta^\alpha_{\pi(i)}} q_i \lambda_i w^{(0)}_{\pi(i)} \\
&= \sum_{j \in [m]} \left(\sum_{i \notin E_j} \delta^\alpha_{\pi(i)} q_i \lambda_i - \sum_{i \in E_j} \overline{\delta^\alpha_{\pi(i)}} q_i \lambda_i \right) w^{(0)}_j\,. \qquad (4.11)
\end{aligned}$$

Equating (4.10) and (4.11) thus gives

$$\sum_{j \in [m]} \sum_{i \in E_j} \left(\delta^\alpha_{\pi(i)} + \overline{\delta^\alpha_{\pi(i)}} \right) q_i \lambda_i w^{(0)}_j = \sum_{i \in E} q_i \lambda_i w^{(0)}_{\pi(i)} = 0\,.$$

The proof of the biases is identical. □

Remark 4.11. Using $\sum_{i \in M_j} q_i \lambda_i = (\sum_{i \notin E_j} + \sum_{i \in E_j}) q_i \lambda_i$ we can alternatively express this as

$$\sum_{j \in [m]} q^{(0)}_j w^{(0)}_j = \sum_{i \notin E_j} q_i \lambda_i w^{(0)}_{\pi(i)}\,.$$

Again we note that the $m = d$ case is a corollary of this lemma, which can be seen by substituting in $q_i \lambda_i = q^{(0)}_{\pi(i)}$.

4.3.2 Main theorem for $m < d$

We are once again in a position to fully characterise W_0. Define

$$S^{d'}_m := \left\{ \pi : [d'] \to [m] \;\middle|\; \pi \text{ is surjective} \right\}.$$

Let $\pi \in S^{d'}_m$ be a fixed surjection. For each $i \in [d']$ we write $\pi(i) = j$ and define

$$\begin{aligned}
M_j &:= \{i \in [d'] \mid \pi(i) = j\}\,, \\
E &:= \{i \in [d'] \mid \epsilon_i = 1\}\,,
\end{aligned}$$

Then for each $j \in [m]$ we can define

$$Y_j := \left\{ (\lambda_i, q_i)_{i \in M_j} \in (\mathbb{R}_{>0} \times \mathbb{R})^{|M_j|} \;\middle|\; \sum_{i \in M_j} \lambda_i q_i = q^{(0)}_j \right\},$$

$$\text{and} \quad \Upsilon' := \left\{ \epsilon : [d'] \to \mathbb{Z}_2 \;\middle|\; \sum_{i \in E} q_i \lambda_i w^{(0)}_{\pi(i)} = 0\,, \text{ and } c_0 + \sum_{i \in E} q_i \lambda_i b^{(0)}_j = c \right\}.$$

Theorem 4.11. *Let $f : \mathbb{R}^2 \times W \to \mathbb{R}$ be a two-layer feedforward ReLU neural network function and let $w \in W \subseteq \mathbb{R}^{4d+1}$ define a fixed $f(x, w)$ with d hidden nodes. Let the true network $f_0(x)$ be an activation-distinguished, minimal, two-layer network with $m < d$ hidden nodes defined by a true parameter $w^{(0)} \in W^{(0)} \subseteq \mathbb{R}^{4m+1}$, where $W^{(0)} \subseteq W$. Suppose $f(x, w) = f_0(x)$ for all $x \in \mathbb{R}^2$. Then*

1. *Let $0 \leq k \leq d-m$ be the number of degenerate nodes of w. Then there is a degenerate reduced form $g(x, w')$ of $f(x, w)$ (Definition 4.6), where $g : \mathbb{R}^2 \times W' \to \mathbb{R}$ is a two-layer network with $d' \geq m$ hidden nodes, and $w' \in W' \subseteq \mathbb{R}^{4d'+1}$ is a parameter whose non-degenerate nodes are equal to those of w, where $W' \subseteq W$.*

2. *Let $W_0' = \{w' \in W' \mid g(x, w') = f_0(x)\}$. Then there is a bijection*

$$\Psi : \prod_{j=1}^{m} Y_j \times S_m^{d'} \times \Upsilon' \xrightarrow{\cong} W_0'$$

$$\left(\left((\lambda_i, q_i)_{i \in M_j}\right)_{j=1}^{m}, \pi, \epsilon\right) \longmapsto \left(\left((-1)^{\epsilon_i}\lambda_i w_{\pi(i)}^{(0)}\right)_{i=1}^{d'}, \left((-1)^{\epsilon_i}\lambda_i b_{\pi(i)}^{(0)}\right)_{i=1}^{d'}, (q_i)_{i=1}^{d'}, c_0 + \sum_{i \in E} q_i \lambda_i b_i^{(0)}\right).$$

Proof. One simply needs to perform identical calculations to those in Theorem 4.7. The same justifications about bijectivity apply too, using Lemma 4.8, Lemma 4.9 and Lemma 4.10. □

As we have remarked above, the result in Theorem 4.7 is a corollary of Theorem 4.11 when $m = d$.

Remark 4.12. In both Theorem 4.7 and Theorem 4.11 we have considered symmetries of W_0 when the domain of the model and truth is all of $\mathbb{R}^2$, which allowed us to make conclusions about comparing polynomial coefficients. In the case where $f(x, w)$ and $f_0(x)$ are restricted to some open bounded domain $\mathcal{Z} \subseteq \mathbb{R}^N$, there could in principle be more degeneracies and symmetries of W_0, since the functional equivalence need only be on $\mathcal{Z}$.

To see this more explicitly, consider a true network $f_0(x) = 0$ with single node model $f(x, w) = q\,\mathrm{ReLU}(\langle w, x\rangle + b)$ defined on $\mathcal{Z} = (-a, a)^2$ for some $a > 0$, where H is the single activation boundary. Then so long as $H \cap \mathcal{Z} = \varnothing$ and w points away from the origin, there will be functional equivalence $f(x, w) = f_0(x)$. This observation is important to keep in mind because our experiments in Chapter 5 involve such a situation.

4.4 Arbitrary Depth

In Section 4.2 and Section 4.3 we have considered networks with two layers. In [PL19], Phuong and Lampert were able to show that this can be generalised to networks of arbitrary depth and arbitrary input dimension for networks with non-increasing widths. They show that, other than set of measure zero, there are "no other function preserving parameter transformations besides permutation and scaling". We state the main result with the key assumptions here for completeness, but the proof of this theorem is beyond the scope of this thesis, for which the reader is referred to the paper.

Referring back to Definition 2.1, let $f : \mathbb{R}^N \times W \to \mathbb{R}^1$ be a neural network of depth L with non-increasing widths $N = d_0 \geq d_2 \geq \cdots \geq d_L \geq d_L = 1$. Letting $1 \leq l \leq k \leq L$ be two layers, we introduce notation for truncated networks

$$A^{l:k} := A^k \circ \mathrm{ReLU} \circ A^{k-1} \circ \cdots \circ \mathrm{ReLU} \circ A^l,$$

where $A_i^{l:k}$ denotes the ith component (i.e. the output of node $i \in [d_{k+1}]$).

Definition 4.7. A *general* feedforward ReLU network is one that satisfies the following three conditions:

1. For any node $i \in [d_{l+1}]$ of a layer $1 \leq l \leq L$, the local optima of $A_i^{1:l}$ do *not* have value exactly zero.

2. For all $1 \leq k \leq l \leq L$ and diagonal indicator matrices $(I_k, \ldots, I_l)$ with entries in $\{0, 1\}$,

$$\operatorname{rank}(I_l w^l I_{l-1} \ldots I_k w^k) = \min\{d_{k-1}, \operatorname{rank}(I_k), \ldots, \operatorname{rank}(I_{l-1}), \operatorname{rank}(I_l)\}\,.$$

3. Let $i \in [d_l]$, $j \in [d_k]$ be nodes of layers $1 \leq l, k \leq L$ respectively. Let U^α be a linear domain of $A_i^{1:l}$, and U^β be a linear domain of $A_j^{1:k}$. Then $A_i^{1:l}\big|_{U^\alpha}$ and $A_j^{1:k}\Big|_{U^\beta}$ are not multiples of one another.

Relating to Theorem 4.7 and Theorem 4.11, the conditions of general networks ensure that there is no weight-cancellation (and thus no orientation-reversing symmetries), and that each node is non-degenerate. Thus Theorem 4.13 excludes these symmetries.

The following lemma proven in Theorem 4.13 justifies the study of general networks:

Lemma 4.12. *Almost all feedforward ReLU networks with this architecture are general.*

As Phuong and Lampert put it, "a sufficient condition for a network to be general with probability one is that the weights are sampled from a distribution with a density." We now state the main theorem of [PL19].

Theorem 4.13. *Let $\mathcal{Z} \subseteq \mathbb{R}^N$ be a bounded non-empty connected open set, and let $f : \mathbb{R}^N \times W \to \mathbb{R}^1$ be a neural network of depth L with non-increasing widths $N = d_1 \geq d_2 \geq \cdots \geq d_L \geq d_{L+1} = 1$. Let $f_0(x) = f(x, w_{(0)})$ be the true network defined by a parameter $w_{(0)} \in W$, which we assume is general, and suppose the model $f(x, w)$ is functionally equivalent, so $f(x, w) = f_0(x)$ for all $x \in \mathcal{Z}$. Then there exist permutations $\sigma_1, \ldots, \sigma_{L-1} \in S_m$, and positive diagonal matrices $M_1, \ldots, M_{L-1}$ such that*

$$\begin{aligned} w^1 &= M_1 \sigma_1 w^1_{(0)}\,, & b^1 &= M_1 \sigma_1 b^1_{(0)}\,, \\ w^l &= M_l \sigma_l w^l_{(0)} \sigma_{l-1}^{-1} M_{l-1}^{-1}\,, & b^l &= M_l \sigma_l b^l_{(0)} \sigma_{l-1}^{-1} M_{l-1}^{-1}\,, \\ w^L &= w^L_{(0)} \sigma_{L-1}^{-1} M_{L-1}^{-1}\,, & b^L &= b^L_{(0)}\,, \end{aligned}$$

where $l \in \{2, \ldots, L-1\}$.

In essence, this result shows that the only generic symmetries of W_0 for ReLU networks with non-increasing widths and one output are scaling and permutation symmetries. Though degenerate and orientation-reversing symmetries are non-generic as points on W_0, we will show in Section 5.1 that such points can nonetheless have lower free energies and thus be preferred by the posterior - in other words, be better model choices. In this way, we believe the theory of deep learning should shift its focus from points to singularities.

4.5 Example - m-symmetric Networks

Let us consider a particular class of networks that gives rise to both degenerate-node and orientation reversing symmetries. In particular, the latter implies that Υ (in the notation of Section 4.2.3) is non-trivial.

Definition 4.8. We define an *m-symmetric network* to be $f_m : \mathbb{R}^2 \to \mathbb{R}$ such that

- f_m is a two-layer feedforward ReLU neural network with two inputs and one output with d hidden nodes as defined in (4.1);
- Let m be an integer such that $2 \leq m \leq d$ and let $s_m \in W$ be the parameter defining $f_m(x) = f(x, s_m)$. Let $g \in SO(2)$ denote the rotation matrix by $\frac{2\pi}{m}$ and let $w_0 \neq (0, 0)$ be some fixed initial vector, e.g. $w_0 = \binom{1}{0}$. Then for each $i \in [m]$, the weights are successively rotated by g, hence set $w_i = g^{i-1} w_0$.

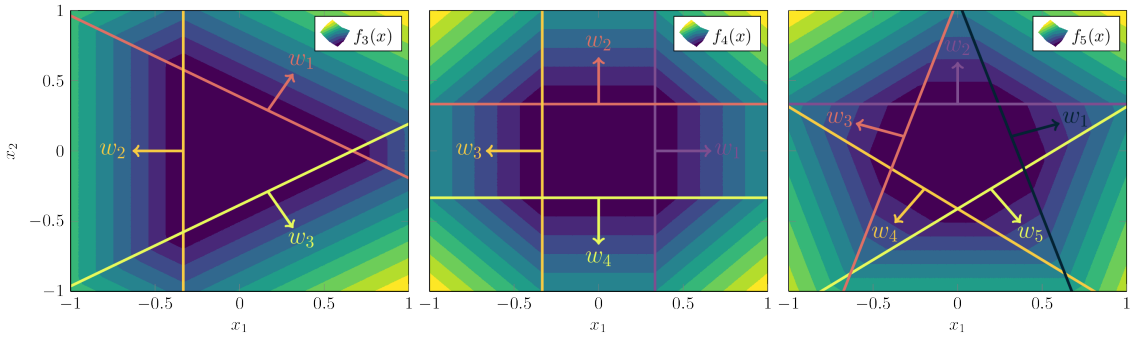


Figure 4.1: $f_m(x)$ for $m = 3, 4, 5$ with suitably defined b and w_0.

- For $i = 1, \ldots, m$ we let $q_i = 1$, $c = 0$, and $b_i = -b$ for some fixed $b > 0$.
- For each $i = m+1, \ldots, d$, node i is degenerate, so $q_i = 0$.

We can observe contour plots of m-symmetric networks (with no degeneracies) in Fig. 4.1. Visually, it is clear that these weight vectors always sum to zero, which will induce orientation reversing symmetry. Let's prove it.

Lemma 4.14. *For any $2 \leq m \leq d$, f_m is a degenerate reducible network. The degenerate reduced form of f_m is minimal distinguished.*

Proof. The first claim follows immediately from the definition of f_m, so assume without loss of generality that f_m is degenerate reduced. For each node $i \in [m]$ the corresponding activation boundary is

$$H_i = \{x \in \mathbb{R}^2 \mid \langle g^{i-1} w_0, x \rangle + b_i = 0\}\,.$$

If $H_k = H_j$ for some $k < j \in [m]$ (without loss of generality), then by Lemma 4.3 we would have for some $\lambda \in \mathbb{R}\backslash\{0\}$

$$g^{j-1} w_0 = \lambda g^{k-1} w_0\,, \quad \text{so} \quad g^{j-k} w_0 = \lambda w_0\,, \quad \text{and} \quad b_j = \lambda b_k\,.$$

This would imply that w_0 was an eigenvector of g^{j-k} with eigenvalue λ. The eigenvalues of g^{j-k} are $\lambda_\pm = e^{\pm i \frac{2\pi(j-k)}{m}}$, which are only real-valued when $j - k = 0$ ($\lambda = 1$) or $j - k = \frac{2}{m}$ ($\lambda = -1$). If $\lambda = -1$ then we would have $b_j = -b_k$. But all biases are equal by definition (and by design), thus we see that $j = k$, showing that each H_j is unique and thus f_m is distinguished. Since it is distinguished, it is necessarily minimal. □

In particular, m-symmetric networks satisfy the weight cancellation that induces orientation reversing symmetry in W_0 (Theorem 4.7).

Lemma 4.15. *Let f_m be an m-symmetric network for some fixed integer m as above. Then the weights associated to f_m satisfy*

$$\sum_{i=1}^{d} q_i w_i = 0\,.$$

Furthermore, suppose m is prime and let $I \subseteq [m]$. Then $\sum_{i \in I} q_i w_i = 0$ if and only if I is empty or $I = [m]$.

Proof. By definition, for weights $m < i \leq d$ we have $w_i = 0$, so we only need to consider the indices $1 \leq i \leq m$. Further, recall that each $q_i = 1$.

Since each w_i is a unit vector and g are rotation matrices, it is easier to reformulate this in terms of complex roots of unity under the isomorphism $U(1) \cong SO(2)$, so let $g = e^{\frac{2\pi}{m} i} \in \mathbb{C}$, which satisfies $g^m - 1 = 0$. The key to the proof rests on the factorisation

$$g^m - 1 = (g-1)(1 + g + g^2 + \cdots + g^{m-1})\,. \tag{4.12}$$

We know $g \neq 1$ since $m \geq 2$, so $g-1$ is invertible, hence

$$\sum_{i=1}^{m} w_i = (1 + g + g^2 + \cdots + g^{m-1})w_0 = (g-1)^{-1}(g^m - 1)w_0 = 0\,,$$

showing the first claim.

For the second claim, if m is prime then it is well known that $1 + X + X^2 + \cdots + X^{m-1}$ is the minimal polynomial of the algebraic number g [Lan02], which is the unique irreducible polynomial of minimal degree such that g is a root. Suppose I is a non-empty subset of $[m]$ such that $\sum_{i\in I} w_i = 0$, and let $J = \{i - 1 \,|\, i \in I\}$. This would imply $\sum_{j\in J} g^j = 0$, that is, g is a root of the polynomial $\sum_{j\in J} X^j = 0$. But since $\sum_{j=0}^{m-1} X^j$ is the minimal polynomial, this implies $J = \{0, \ldots, m-1\}$ or J is empty (since we take the empty sum to be zero), showing the claim. □

Corollary 4.16. *Let f_m be an m-symmetric true network in degenerate reduced form for some prime m, and let $f(x, w)$ be the model with $d > m$ hidden nodes such that $f(x, w) = f_m(x)$. Then $f_m(x)$ is a degenerate reduced form of $f(x, w)$. For ease, now suppose $f(x, w)$ is in its degenerate reduced form with $d = m$ nodes as per Definition 4.6 and define $W_0 = \{w \in W \,|\, f(x, w) = f_m(x)\}$. Then $W_0 \cong \prod_{i=1}^{m} X_i \times S_m \times \Upsilon$ where*

$$\Upsilon = \left\{(0)_{i=1}^{m}, (1)_{i=1}^{m}\right\},$$

meaning

$$W_0(f_m) = \left\{ \left((w_i)_{i=1}^{d}, (b_i)_{i=1}^{d}, (q_i)_{i=1}^{d}, c\right) \;\middle|\; \epsilon \in \{0, -1\}\,,\;\; q_i \in \mathbb{R}_{>0}\,,\;\; \sigma \in S_m \right\},$$

$$\text{where} \quad w_i = \left(\frac{(-1)^{\epsilon}}{q_i} g^{\sigma(i)-1} w_0\right)_{i=1}^{d}, \quad b_i = \left(\frac{(-1)^{\epsilon+1}}{q_i} b\right)_{i=1}^{d}, \quad c = \epsilon m b\,.$$

Proof. By Lemma 4.14 we may reduce f_m to its degenerate reduced network which is then distinguished, meaning we can apply Theorem 4.7. By Lemma 4.15, the only subsets I for which $\sum_{i\in I} q_i w_i = 0$ is the empty set or $[m]$. Thus either no weights are reversed, $\epsilon = (0)_{i=1}^{m}$, or all weights are reversed, $\epsilon = (1)_{i=1}^{m}$. So by Theorem 4.7, Υ has the stated form. □

Remark 4.13. If m is not prime then Υ will have a more complicated form. In particular, not all $\epsilon \in \Upsilon$ will have the same entries for $i \in [m]$. We leave it as an exercise to the reader to think about such cases, for example in the case of $m = 4$.

Chapter 5

Phase Transitions in ReLU Neural Networks

In Chapter 4 we classified the symmetries of the set of true parameters W_0 in order to determine all points that minimise $K(w)$. In this chapter we will show that not all points on W_0 are equally good minimisers of the free energy, and moreover, that parameters not in W_0 can nonetheless be preferred by the posterior due to having lower model complexity. Inspired by statistical physics, we identify phases of F_n as compact sets containing a particular singularity of interest. Phase transitions arise from changes in the accuracy versus complexity trade-off of different phases of W, which we induce by changing the symmetries of the underlying W_0. In particular, we use m-symmetric networks and Markov Chain Monte Carlo methods to analyse the non-generic orientation-reversing and degenerate-node phases found in Chapter 4.

5.1 Phases and Phase Transitions

Physicists typically think of a phase as an aggregate state of a system of many particles with complicated interactions. For example, the solid, liquid or gaseous states of H_2O are all examples of phases. Phases are distinguished by their physical properties, such as the molar volume in the case of the phases of water. A phase transition, then, is a sudden change in such a property as a function of some order parameter θ. Mathematically speaking, this is a non-analyticity of the free energy, which corresponds to a change in the configuration of phases resulting in one being newly preferred over another.

Empirical measures of neural network performance, such as scaling laws [Kap+20] and generalisation error learning curves [Nak+19] offer evidence of phase transitions in neural networks. Furthermore, [SST92] provides a theoretical-physics treatment of phase transitions associated to the generalisation error, where the order parameter is $\frac{n}{D}$. However, as in most statistical literature, incorrect assumptions of regularity of the neural network models are relied upon.

In the context of Bayesian statistics, Watanabe simply defines a phase transition as a "drastic change in the posterior," [Wat18], and in [Wat20] he outlines a useful conceptual framework for phase transitions in neural networks. Let us attempt to add a modicum more rigour to this statement in line with the interpretation of phase transitions in [Cal85] and [Gil93].

For simplicity of interpretation, we suppose the free energy is a function of a level set of W, as projected by a function V. We may think of this V as a macroscopic observable (e.g. volume) which thus extracts information about configurations in W.

Definition 5.1. Let n and β be fixed. Let $V : W \to \mathbb{R}$ be an analytic function, and let $\mathcal{V} = \text{Image}(V)$, which is compact since W is, and define

$$\mathcal{W}_v = \{w \in W \mid V(w) = v\}.$$

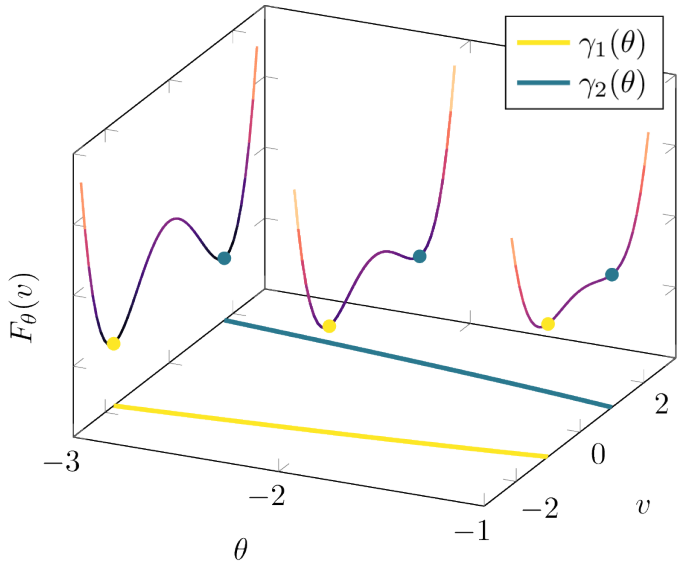


Figure 5.1: A depiction of two phases of $F_\theta(v) = \frac{1}{4}v^4 + \frac{\theta}{2}v^2$ over the interval $I = [-3, -1]$.

Suppose the posterior depends on a non-stochastic order parameter $\theta \in \Theta \subseteq \mathbb{R}$. In particular, we will assume $L_n(w) = L_n(w, \theta)$, but this could also be true of the prior $\varphi(w)$. Let the free energy be defined by

$$F : \mathcal{V} \times \Theta \longrightarrow \mathbb{R}\,, \quad F(v, \theta) = -\log\left(\int_{\mathcal{W}_v} \varphi(w) e^{-n\beta L_n(w,\theta)} dw\right).$$

We define the critical points of F to be

$$\mathrm{Crit}_F = \left\{ (v, \theta) \in \mathcal{V} \times \Theta \;\middle|\; \frac{\partial F}{\partial v}(v, \theta) = 0 \right\},$$

and we let $\mathrm{Crit}_F^{\min} \subseteq \mathrm{Crit}_F$ denote those critical points that are minima of F.

Consider an interval $I = [\theta_1, \theta_2] \subseteq \Theta$. A *phase* over I is a continuous map $\gamma : I \to \Theta$ such that the following diagram commutes

$$\begin{array}{ccc} & & \mathrm{Crit}_F^{\min} \\ & \overset{\gamma}{\nearrow} & \downarrow{\scriptstyle \iota} \\ I & \xhookrightarrow{\iota} & \mathcal{V} \times \Theta \\ & & \downarrow{\scriptstyle \pi_2} \\ & & \Theta \end{array}$$

where ι denotes the respective inclusions and π_2 is the projection of the second component.

In words, a phase is a minimum of the free energy that remains a minimum with small perturbations in θ, and as such can be viewed as a path in $\mathrm{Crit}_F^{\min}$. The precise value of the free energy is irrelevant to the phase structure, but the ordering of the free energy of phases is the substance of phase transitions, whereby the configuration of the phases changes. The following definition is based on [Gil93, §10], where we adopt the Maxwell convention outlined in [Gil93, §8.2].

Definition 5.2. Let $\Gamma = \{\gamma_g\}_{g=1}^G$ be the set of phases of some F on a fixed interval I, and assume $G \geq 2$. Let $I_c = [\theta_0, \theta_1] = [\theta_c - \varepsilon, \theta_c + \varepsilon] \subseteq I$ denote a small critical interval for some critical

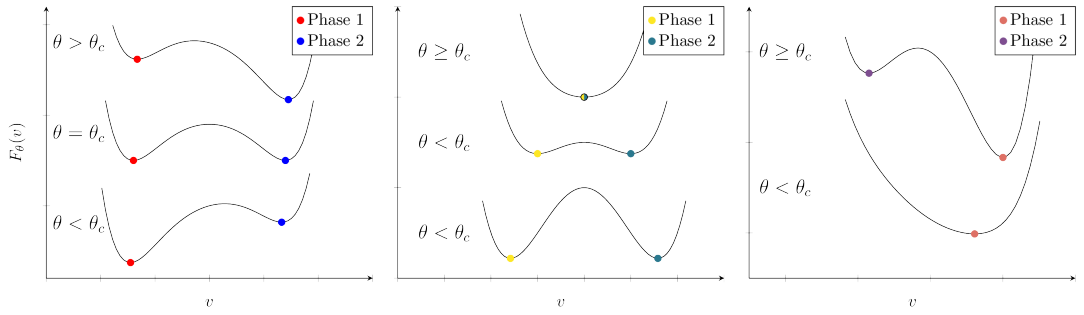


Figure 5.2: A first order phase transition (left), a second order merge phase transition (centre) and a second order creation phase transition (right). All curves are generated by $F_\theta(v) = \frac{1}{6}x^6 + \frac{1}{8}c_1x^4 + \frac{1}{2}c_2x^2 + c_3x + c_4$ for parameters c_i that depend on θ as in [Gil93, §10.8].

parameter $\theta_c \in I$. Without loss of generality, assume that Γ is ordered by the corresponding free energy of each phase at the left endpoint of the interval. That is, for each $1 \leq g < h \leq G$,

$$F(\gamma_g(\theta_0)) \leq F(\gamma_h(\theta_0))\,,$$

where $F(\gamma_1(\theta_0))$ thus denotes the global minimum.

A *first order phase transition* at θ_c is a reordering of Γ that exchanges a local and global minima. That is, for some $g > 1$, varying $\theta \in I_c$ gives

$$\begin{cases} F(\gamma_1(\theta)) < F(\gamma_g(\theta)) & \theta < \theta_c \\ F(\gamma_1(\theta)) = F(\gamma_g(\theta)) & \theta = \theta_c \\ F(\gamma_1(\theta)) > F(\gamma_g(\theta)) & \theta > \theta_c \end{cases}.$$

A *second order phase transition* at θ_c can be one of two things. A *merge* transition occurs when, given two phases γ_1 and γ_g for some $g > 1$, varying $\theta \in I_c$ gives

$$\begin{cases} \gamma_1(\theta) \neq \gamma_g(\theta) & \theta < \theta_c \\ \gamma_1(\theta) = \gamma_g(\theta) & \theta \geq \theta_c \end{cases},$$

such that γ_1, γ_g are global minima of F for $\theta \geq \theta_c$.

A *creation* transition occurs when, given any phase γ_g, varying $\theta \in I_c$ gives

$$\begin{cases} \gamma_g \text{ is } \textit{not} \text{ a phase} & \theta < \theta_c \\ \gamma_g \textit{ is} \text{ a phase} & \theta \geq \theta_c \end{cases}.$$

The direction of θ may be altered in any of these definitions. If the direction of the creation transition is reversed we refer to this as a *destruction* transition. The different types of phase transitions can be seen in Fig. 5.2.

We do not claim to have given a full classification of phase transitions in these definitions. To do so generically requires one to pay careful attention to the differential geometry of F, and the possible types of *catastrophes* that can occur. For further discussion of this, see [Gil93].

Remark 5.1. In principle, one may also care to define phase transitions that exchange or merge local minima which are not global. We have restricted our attention to only consider global minima due to the fact that, in a physical sense, these are the only ones that have a meaningful impact on the state of a system.

Phases as singularities

Let us now explore how phases can be associated to compact sets containing particular singularities of K, thus providing a different kind of candidate solution to the problem of model selection. For this discussion, let $V : W \to \mathbb{R}$ be a fixed analytic function with level sets $\mathcal{W}_v$ as in Definition 5.1.

The reader may wish to review the discussion of singularities in Chapter 3. For any compact set $\mathcal{W}_v \subseteq W$ there is a local RLCT λ_v as per Remark 3.3 which extracts the effective dimensionality of "the most singular" point in $\mathcal{W}_v$. Let $\omega_0^v \in \mathcal{W}_v$ be defined by $L_n(\omega_0^v) = \min_{\omega \in \mathcal{W}_v} L_n(\omega)$, thus has the best accuracy (meaning the lowest loss) of any parameter in $\mathcal{W}_v$. Then according to [Wat09, §7.6] we have that

$$F_n(\mathcal{W}_v) \approx nL_n(\omega_0^v) + \frac{\lambda_v \log n}{\beta_0} .$$

Assume for simplicity that n and β_0 are both fixed. This relation shows that a minimum of F_n will correspond to a compact set $\mathcal{W}_v$ with low $L_n(\omega_0^v)$ and low λ_v compared to nearby v. Thus, a phase of F corresponds to a compact set $\mathcal{W}_v$ that contains a particular singularity of interest. Comparing the free energy of these different $\mathcal{W}_v$ thus becomes the basis of the model selection process, where each phase can be thought of as a different set from which to draw candidate solutions.

Phase transitions, then, correspond to a change in the structure of $K(w)$, which induces a change in either the accuracy or RLCT of the regions $\{\mathcal{W}_v\}_v$ as a function of θ which causes a new phase to be preferred. In simple terms, it is a change in the accuracy versus complexity trade-off between candidate solutions. As the true distribution is varied, so is the geometry of $K(w)$. We thus expect to see phase transitions occur near points where the underlying symmetry of W_0 changes (for example, where a node is newly degenerate). We call such changes *symmetry breaking* of W_0. Naively we may expect the critical point θ_c to occur precisely where the symmetry breaking takes place. But we will show in Section 5.3 that in fact due to the accuracy versus complexity trade-off, these critical values can occur before the point of symmetry breaking.

Remark 5.2. In general, it is not physically or computationally reasonable to measure these level sets $\mathcal{W}_v$. Instead we must rely on *coarse graining*. Let $\mathcal{V} = [v_1, v_H]$, which we can evenly partition into intervals

$$v_1 < v_2 < \cdots < v_{H-1} < v_H$$

for some $H \geq 2$. Then we may define compact sets for each $1 \leq h \leq H-1$

$$\mathcal{W}_h = V^{-1}([v_h, v_{h+1}]) = \{w \in W \mid V(w) = [v_h, v_{h+1}]\} .$$

Since V is an analytic function defined on the compact W, the $\mathcal{W}_h$ are disjoint and compact, meaning we can write

$$Z_n = \sum_{h=1}^{H-1} Z_n(\mathcal{W}_h) .$$

Then the model selection process becomes finding

$$\min_{h=1,\ldots,H-1} F_n(\mathcal{W}_h) .$$

This is the scenario we will consider in our experiments. More precisely, we will consider compact sets that are unions of $\mathcal{W}_h$, for which all of the above setup still applies.

Before we move on to our experimental findings, it is useful to have an intuitive understanding of what we mean by the "nature of a singularity". In algebraic geometry, typical examples of singularities are lines that self-intersect (perhaps more than once), cusps, and tacnodes. As an example, a curve that intersects itself three times at the same singularity will be more complicated than one that intersects itself only once. Let us give an example relevant to our context.

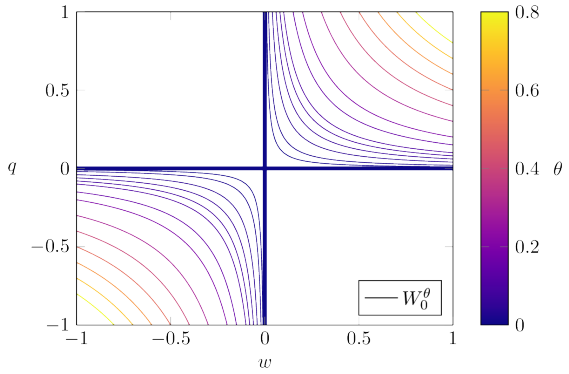


Figure 5.3: W_0^θ of $K_\theta(w,q)$ coloured by different values of θ.

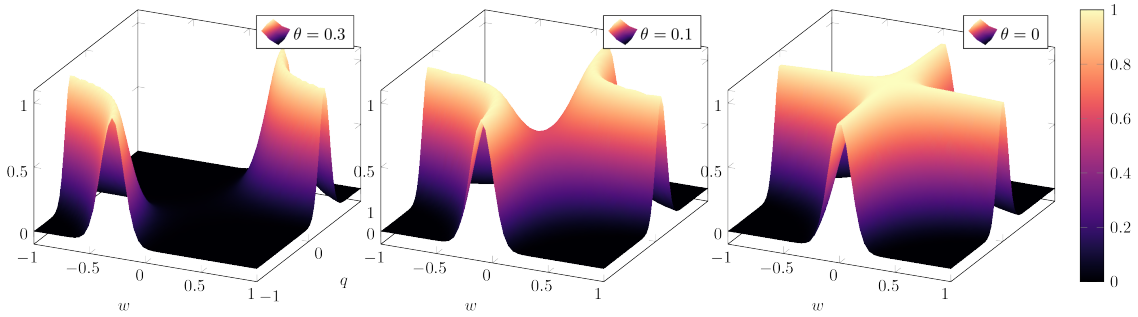


Figure 5.4: $\varphi(w)e^{-nK_\theta(w)}$ for $n = 50$ over different θ values. Notice the posterior concentration of the more singular point, $(w,q) = (0,0)$, when $\theta = 0$.

Example 5.1. Consider $K(w,q)$ from Example 3.1, namely

$$K_\theta(w,q) = (wq - \theta)^2 ,$$

this time for $w, q \in [-1,1]^2$ and some $\theta \geq 0$. Then it is visually clear from Fig. 5.3 that, the geometry of W_0 is different for $\theta = 0$ compared to $\theta > 0$, meaning we should think of $\theta > 0$ as one phase and $\theta = 0$ as another phase. This implies that the RLCT (and/or its multiplicity) is different between these phases. To prove this, however, one must perform a resolution of singularities (an algorithm known as a "blow-up") of the former case in order to calculate $\zeta(z)$.

In Fig. 5.4 we observe the posterior (up to a scale factor) for uniform $\varphi(w)$ as θ varies. Notice that the posterior concentration around the point $(w,q) = (0,0)$ for $\theta = 0$ is different to that around $wq = \theta$ for $\theta > 0$, giving a clear visual picture of why one might expect "more singular points" to have lower free energy.

We now turn our attention to studying phases and phase transitions of ReLU neural networks from an experimental perspective.

5.2 Experimental Methodology

The following section outlines our methodology for estimating the posterior of ReLU neural networks. MCMC methods are introduced for the unfamiliar reader and we outline the details of the experimental procedure. Finally, we explain how to interpret the density plots demonstrating the phase transitions, describing how we quotient out the scaling and permutation symmetries of the networks. The methodology used closely follows the work of [Mur+20].

5.2.1 Markov Chain Monte Carlo

In order to study phases in neural networks, we must first begin with estimating the posterior density $p^\beta(w|D_n)$. To see why such a procedure is non-trivial, recall that whilst the numerator of the posterior $\varphi(w)e^{-n\beta L_n(w)}$ is easily calculated for any given w, the partition function Z_n is intractable even in simple settings.

This problem has given rise to the field of computational Bayesian statistics, whose primary focus is to develop algorithmic methods for estimating probability densities (see [RC04] for an introduction). Mathematically, the goal is to generate a set of samples $\{w^{(k)}\}_{k=1}^K$ such that for any arbitrary function $f(w)$ we have

$$\int_W f(w)p^\beta(w|D_n)dw \approx \frac{1}{K}\sum_{k=1}^K f(w^{(k)})$$

as $K \to \infty$ [Wat18, §7]. In particular we can retrieve the posterior on any open set $A \subseteq W$ by simply taking $f(w) = \mathbb{1}(w \in A)$.

The centrepiece of computational Bayesian statistics is a class of algorithms called Markov Chain Monte Carlo (MCMC). Recall that a Markov chain is a random walk on a space W such that the next step only depends on the present step and is independent of the previous history. The main idea of MCMC is to simulate a Markov chain such that the equilibrium distribution of the chain is equal to the probability density to be estimated. This is achieved in two key steps:

1. Selecting a candidate step from a distribution based on the initial position (Markov step).
2. Choosing to accept the candidate with a probability proportional to $\varphi(w)e^{-n\beta L_n(w)}$ (Monte Carlo step).

The simplest version of this algorithm is known as the *Metropolis algorithm*, but it suffers from some important shortcomings. Namely, it can struggle to sufficiently explore the space of parameters due to potential and entropy barriers in $\varphi(w)e^{-n\beta L_n(w)}$. Further discussions of these problems can be found in [Bro11] and [Wat18, §7].

Instead, let us turn our attention to the MCMC variant used in this thesis, Hamiltonian Monte Carlo, which aims to avoid these barrier problems by replacing the Markov step with a particle simulation through phase space according to Hamilton's equations of motion. Concretely, define the *Hamiltonian* of our model-prior system by

$$H(w) = \beta L_n(w) - \frac{1}{\beta}\log\varphi(w)$$

such that $p^\beta(w|D_n) \propto e^{-H(w)}$. Suppose $v \in \mathbb{R}^d$ is some randomly generated initial velocity, usually from a standard normal for simplicity. Then we may define the *total Hamiltonian* by

$$\mathcal{H}(w,v) = \frac{1}{2}\|v\|^2 + H(w)\,.$$

Note that sampling (w,v) from $e^{-\mathcal{H}(w,v)}$ still ensures that w is still subject to the equilibrium distribution $e^{-H(w)}$. After initialising some random starting point $w^{(1)}$ for $k = 1$, the algorithm thus runs as follows:

1. *Hamiltonian step:* Sample $v_0 \sim \mathcal{N}(0, \mathbb{1}_d)$. Let τ denote the time variable. Simulate a trajectory through phase space W according to the differential equation

$$\begin{aligned} \frac{dw}{d\tau} &= v\,, \qquad \frac{dv}{d\tau} = -\nabla H(w)\,, \\ \text{s.t.} \quad (w, v) &= (w^{(k)}, v_0) \quad \text{at time } \tau = 0\,. \end{aligned}$$

   Run the simulation up to some time τ_{stop} and let $(w', v') = (w(\tau_{\text{stop}}), v(\tau_{\text{stop}}))$ be the candidate parameter.

2. *Monte Carlo step:* Define $\Delta\mathcal{H} = \mathcal{H}(w', v') - \mathcal{H}(w^{(k)}, v_0)$ and $P = \min\{1, \exp(-\Delta\mathcal{H})\}$. Then assign the next step $w^{(k+1)}$ according to the rule:

$$w^{(k+1)} = \begin{cases} w' & \text{with probability } P \\ w^{(k)} & \text{with probability } 1 - P \end{cases}. \tag{5.1}$$

Notice that the candidate is accepted with probability 1 if $\mathcal{H}(w', v') < \mathcal{H}(w^{(k)}, v_0)$, thus ensuring that the sampler moves towards regions of lower energy. If the inequality is reversed but the candidate has relatively similar energy, then the candidate is still accepted with high probability. If the candidate energy is much higher than that of the current position, it is rejected with high probability. The Hamiltonian/Markov-chain step is thus the key to ensuring MCMC is not just a glorified gradient descent algorithm, but rather that it effectively explores the space.

In practice there is an invisible first step called the "burn-in" period, which discards some number of initial samples to allow the sampler to approach the equilibrium distribution before it accepts samples.

Remark 5.3. A useful conceptual framework for Hamiltonian Monte Carlo, as elaborated in great detail in [McE], is to imagine the posterior, or rather the reciprocal of the posterior, as a skate park. Peaks of $p^\beta(w|D_n)$ correspond to the troughs of the bowls in the rink, and zero regions of $p^\beta(w|D_n)$ correspond to infinitely high walls. Then HMC takes a ball (particle) $w^{(k)}$ at some starting point in W, generates a random initial velocity v_0, and kicks the ball with that initial velocity. Hamilton's equations are used to simulate the trajectory across the skate park. After some stopping time τ', we stop the ball, record its position, and restart the process from the chosen position.

The experiments performed in this thesis use a variant on the Hamiltonian Monte Carlo method called the *No U-Turn Sampler* (NUTS), which improves the dynamical simulations by ensuring there are "no U-turns", thus dynamically choosing τ_{stop} to avoid returning to a similar position as was started at. More details can again be found in [Bro11]. For the uninitiated but interested reader, a great introduction to HMC is found in [Bet18].

In this thesis, MCMC methods were implemented using the Python packages PyTorch [Pas+19] and Pyro [Bin+19], with experiments run on Spartan High Performance Computing of the University of Melbourne [Laf17].

5.2.2 Experimental setup

We shall consider the realisable case under Hypothesis 2.1 where $q(y|x) = p(y|x, w^{(0)})$ is defined by a two-layer feedforward ReLU network with two inputs, one output and m hidden nodes as discussed in Eq. (2.2), and the model has the same architecture with $d = m$ hidden nodes. Since the model $p(y|x, w)$ is a normal distribution as in Hypothesis 2.1, the true distribution is also a normal distribution.

In particular, the true network will depend on some order parameter $\theta \in \Theta$ which we denote by $f_0(x, \theta)$. Thus for any dataset of samples $D_n = \{(x_i, y_i)\}_{i=1}^n$ drawn from the true distribution $(x_i, y_i) \sim q(y, x|\theta)$ we write $D_n = D_n(\theta)$, meaning the posterior depends on the order parameter through its variation in the true distribution, and thus the dataset. The exact nature of the order

parameter will differ between experiments. The prior on inputs will be uniform on a square for some $a > 0$,

$$q(x) = \frac{1}{4a^2} \mathbb{1}\left((x_1, x_2) \in [-a, a]^2\right) ,$$

and the prior on parameters will be the standard normal with fixed variance σ_φ^2,

$$\varphi(w) = \frac{1}{(2\pi\sigma_\varphi^2)^{\frac{4d+1}{2}}} e^{-\frac{1}{2\sigma_\varphi^2}\|w\|^2} .$$

An *experiment* will refer to a fixed vector $(\theta, a, \sigma_\varphi)$. In order to account for the randomness in D_n and make statements about the posterior independent of D_n, we will run T repeat trials of the same experiment and average the posterior estimates over the results following a validation procedure (see below). Typically $T = 8$ or $T = 4$.

For a given trial, we begin by generating a dataset D_n for fixed $n = 10000$, and then use HMC (NUTS) to generate a set of samples $\{w^{(k)}\}_{k=1}^K$ from the tempered posterior $w^{(k)} \sim p^{\beta^*}(w|D_n(\theta, a))$ with burn-in period $\frac{K}{20}$. Here we take $\beta^* = \frac{1}{\log n}$ as per Theorem 3.9. Typically $K = 20000$.

MCMC is quite a delicate algorithm and can occasionally produce results contrary to what is anticipated - for example, the sampler can get stuck in particular regions of space of much higher free energy than the true global minimum. Our first form of validation is using standard MCMC chain divergence criterion, where trials with more than $\frac{K}{10}$ chain divergences discarded. We then perform a simple statistical validation process, discarding any outlier trials as measured by the average mean square error across all samples, which in our language is the empirical Gibbs training loss G_t^β for a trial t,

$$G_t^\beta(t) \approx \frac{1}{K} \sum_{k=1}^{K} L_n(w^{(k)}) .$$

Using the central limit theorem, $G_t^\beta \approx \mathcal{N}(\mu_T, s_T^2)$ where μ_T and s_T^2 are the sample mean and sample variance respectively. We discard any trials t such that $\frac{1}{s_T}|G_t^\beta(t) - \mu_T| > \kappa$ for some outlier threshold κ, which we usually set as $\kappa = 1.5$.

5.2.3 Machine epsilon and practical limits

Strictly speaking, our setup violates two key assumptions of Singular Learning Theory, namely that W is compact and that $K(w)$ is analytic. The former is violated by our assumption that $\varphi(w)$ is normally distributed, and thus a density on all of $\mathbb{R}^{4d+1}$. The latter is violated due to the non-analyticity of $f(x, w)$ for ReLU neural networks. In both cases we may appeal to machine epsilon ε_m, which is the finite floating point precision of any given computer.

For the first point, $\varphi(w)$ is computationally zero outside of some (very large) neighbourhood of the origin. Thus we can simply consider W to be this very large compact neighbourhood, for which the experimental results do not meaningfully change.

For the second point, we can use Remark 2.1 to find some γ such that ReLU is approximated by swish, $|\sigma_\gamma(x) - \text{ReLU}(x)| < \varepsilon_m$. Then $K(w)$ will be analytic if we take $\sigma_\gamma(x)$ to be the activation function. But, importantly, for this γ, σ_γ and ReLU are indistinguishable from the point of view of the computer, meaning the experimental results also do not change.

5.2.4 Visualising the posterior

Our main method of studying phases of neural networks will be to examine the posterior of a given neural network, where "drastic changes" in these posteriors as a function of some order parameter will correspond to phase transitions. Since $W \subseteq \mathbb{R}^{4d+1}$ for some $d \geq 2$, we clearly cannot visualise the posterior directly. Nor are we interested in doing so: our experiments only consider variations in the weights, not the biases, so we only need to observe the posterior of the weights.

Figure 5.5: Scatterplot of raw (left) versus effective (right) samples $\{\{(\hat{w}_{i,1}^{(k)}, \hat{w}_{i,2}^{(k)})\}_{i\in[d]}\}_{k=1}^{K}$ for $K = 4000$, $d = 3$, labelled according to their index i, for one trial.

Recall from Theorem 4.7 and Theorem 4.11 that W_0 generically admits scaling and permutation symmetry, which is also a property of more general networks as discussed in Theorem 4.13. With this in mind, these are uninteresting symmetries to analyse as singularities. Instead, we shall focus on the non-generic degenerate-node and orientation-reversing phases and measure the free energy of these as we vary the true distribution.

Definition 5.3. Given a fixed sample $w^{(k)} \sim p^{\beta}(w|D_n)$ we define the *effective parameter* $\hat{w}^{(k)} = (\{\hat{w}_i^{(k)}\}_{i=1}^d, \{\hat{b}_i^{(k)}\}_{i=1}^d, \{1\}_{i=1}^d, c)$, where for each node $i \in [d]$ we define

$$\hat{w}_i^{(k)} = |q_i^{(k)}| w_i^{(k)}, \quad \text{and} \quad \hat{b}_i^{(k)} = |q_i^{(k)}| b_i^k .$$

By Theorem 4.7 we have

$$f(x, w^{(k)}) = f(x, \hat{w}^{(k)}),$$

but now the effective parameter has less degrees of freedom since we have taken the quotient of the scaling symmetry.

Remark 5.4. In Definition 4.1 we defined a degenerate node i to be one such that $q_i = 0$ or $w_i = 0$. In the language of effective parameters, this is equivalent to $\hat{w}_i = 0$.

The posterior is invariant under a permutation of nodes, which implies that for each $i \in [d] = \{1, \ldots, d\}$, w_i is identically distributed (though there is a distributional dependence between each weight). This allows us to project each $\hat{w}_i^{(k)}$ on to the same $(\hat{w}_{i,1}, \hat{w}_{i,2})$ plane. Thus each sample $w^{(k)}$ is represented d times on each posterior plot by the points $\{(\hat{w}_{i,1}^{(k)}, \hat{w}_{i,2}^{(k)})\}_{i\in[d]}$.

We will mostly present density estimates of the posterior, but for clarity, Fig. 5.5 shows a typical example of a scatterplot of points

$$\left\{\{(\hat{w}_{i,1}^{(k)}, \hat{w}_{i,2}^{(k)})\}_{i\in[d]}\right\}_{k=1}^{K}$$

coloured according to the node index i and demonstrating the difference between the effective and non-effective estimates of the weights. On all plots the red dots indicate the true parameters used to generate D_n. All experiment plots have been generated using Seaborn [Was21] and Matplotlib [Hun07].

As we argued in Section 3.2, the free energy of a compact set $\mathcal{W} \subseteq W$ is a measure of posterior density associated to $\mathcal{W}$. It was found that approximate values of the free energy using the WBIC was volatile. Instead, our inference about phases and phase transitions will use the following correspondence:

$$\text{phase} \iff \text{minimum of the free energy} \iff \text{concentrated region of posterior},$$

where we compare minima associated to regions by comparing their respective posterior concentrations.

5.3 Phase Transition 1: Deforming to Degeneracy

Having taken the quotient of scaling and permutation symmetry, in this section we will explore phases associated to degenerate-node and non-degenerate-node phases. Recall that a node $i \in [d]$ is degenerate if either $w_i = 0$ or $q_i = 0$, meaning $\hat{w}_i = 0$.

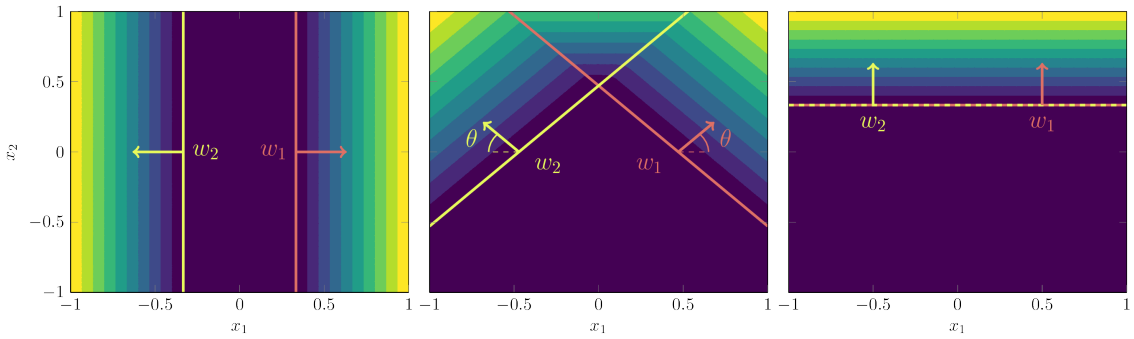


Figure 5.6: $f_2(x, \theta)$ for $\theta = 0$ (left), $\theta = \frac{\pi}{4}$ (middle), $\theta = \frac{\pi}{2}$ (right).

5.3.1 Defining the order parameter

Let $f_2 : \mathbb{R}^2 \times \Theta \to \mathbb{R}$ be an m-symmetric network as in Definition 4.8, with $m = 2$, $w_0 = (1, 0)^T$ and $b = -\frac{1}{3}$, which we take to be the true network. We define an order parameter $\theta \in \Theta = [0, \frac{\pi}{2}]$ that rotates the two weights toward one another, that is, the true weights are

$$w_1^{(0)} = g_\theta w_0 = (\cos\theta, \sin\theta)\,, \qquad w_2^{(0)} = g_{\pi-\theta} w_0 = (-\cos\theta, \sin\theta)\,,$$

where g_θ denotes rotation by θ. Explicitly, the truth is defined by

$$f_2(x, \theta) = \text{ReLU}\left(\cos(\theta)x_1 + \sin(\theta)x_2 - \frac{1}{3}\right) + \text{ReLU}\left(-\cos(\theta)x_1 + \sin(\theta)x_2 - \frac{1}{3}\right)\,.$$

We can see how the foldsets of $f_2(x, \theta)$ change with θ in Fig. 5.6.

Using $f_2(x, \theta)$ we can thus induce symmetry breaking of W_0 which occurs when node-degeneracy becomes one of the symmetries. Recalling that the biases of both nodes are equal, we see that node-degeneracy symmetry of W_0 occurs when $w_1^{(0)} = w_2^{(0)}$, which is only at $\theta = \frac{\pi}{2}$. By contrast, for $\theta \in (0, \frac{\pi}{2})$, W_0 only exhibits the standard scaling and permutation symmetry. At $\theta = \frac{\pi}{2}$ the model nodes $\hat{w}_1, \hat{w}_2 \in W_0(f_2(x, \frac{\pi}{2}))$ can have two possible configurations:

- *Both non-degenerate*: $\hat{w}_1, \hat{w}_2 \neq 0$ such that $\hat{w}_1 + \hat{w}_2 = (0, 2)$,
- *One degenerate, one non-degenerate*: either $\hat{w}_1 = (0, 0)$ and $\hat{w}_2 = (0, 2)$, or $\hat{w}_1 = (0, 2)$ and $\hat{w}_2 = (0, 0)$.

We thus identify neighbourhoods of these singularities as phases to compare. Accordingly, let us define compact subsets of W in order to compare their free energies. Let the analytic projection of Section 5.1 be $V(w_i) = \|w_i\|$, and define an annulus in the plane as

$$\mathcal{A}(r, \varepsilon) = \{\hat{w} \in \mathbb{R}^2 \mid r - \varepsilon \leq \|\hat{w}\| \leq r + \varepsilon\}\,.$$

Then we define the two phases containing the singularities of interest to be

$$\begin{aligned}
\mathcal{A}_{\text{NonDegen}} &= \mathcal{A}(1, \varepsilon) \times \mathcal{A}(1, \varepsilon) \\
&= \left\{(\hat{w}_1, \hat{w}_2) \,|\, \hat{w}_1 \in \mathcal{A}(1, \varepsilon) \text{ and } \hat{w}_2 \in \mathcal{A}(1, \varepsilon)\right\} \\
\mathcal{A}_{\text{Degen}} &= \big(\mathcal{A}(0, \varepsilon) \times \mathcal{A}(2, \varepsilon)\big) \cup \big(\mathcal{A}(2, \varepsilon) \times \mathcal{A}(0, \varepsilon)\big) \\
&= \left\{(\hat{w}_1, \hat{w}_2) \,|\, \hat{w}_1 \in \mathcal{A}(0, \varepsilon) \text{ and } \hat{w}_2 \in \mathcal{A}(2, \varepsilon)\,, \quad \text{or} \quad \hat{w}_1 \in \mathcal{A}(2, \varepsilon) \text{ and } \hat{w}_2 \in \mathcal{A}(0, \varepsilon)\right\}\,.
\end{aligned}$$

For notational ease we then let $\mathcal{A}^c = W \backslash (\mathcal{A}_{\text{NonDegen}} \cup \mathcal{A}_{\text{Degen}})$. We take $\{\theta_j\}_{j=1}^J$ to be a sequence of angles (in radians) such that $\theta_1 = 1.00^c$ and $\theta_J = \frac{\pi}{2}$ and observe changes in the posterior over θ. In the following experiments there were $K = 20,000$ samples taken over $T = 8$ trials with $a = 2$ and $\sigma_\varphi = 1$.

5.3.2 Results and discussion

Fig. 5.7 demonstrates the phase transitions in the posterior as we range over these θ_j values. This is further highlighted by Fig. 5.8a where we take $\varepsilon = 0.3$ as the annuli width.

For $\theta < 1.16^c$ the only phase detected by the posterior is $\mathcal{A}_{\text{NonDegen}}$. At $\theta = 1.16^c$ we see a *second order creation* phase transition, where $\mathcal{A}_{\text{Degen}}$ has suddenly emerged as a region of concentration, and thus a phase - though the $\mathcal{A}_{\text{NonDegen}}$ phase is still the global minima. As $\theta \to 1.26^c$, the phase structure remains the same, though the free energy of $\mathcal{A}_{\text{NonDegen}}$ is increasing, whilst for $\mathcal{A}_{\text{Degen}}$ it is decreasing. Observing Fig. 5.8a, at the critical value $\theta_c = 1.26^c$ we notice a *first order phase transition* where the free energies of both phases are equal and switch roles as local and global minima. Once this transition has occurred and $\mathcal{A}_{\text{Degen}}$ has become the global minima of F, the gap between the respective free energies continues to widen.

Let us inspect the $\theta = \frac{\pi}{2}$ figure in Fig. 5.7 more closely. Although at this θ the true network is defined by two non-degenerate nodes $\hat{w}_1^{(0)} = (0, 1)$ and $\hat{w}^{(0)} = (0, 1)$ in $\mathcal{A}_{\text{NonDegen}}$, the phase $\mathcal{A}_{\text{Degen}}$ is nonetheless preferred. This is our first clear example of the fact that, although all points on W_0 minimise $K(w)$, it is their structure as *singularities* of $K(w)$ that determines which has lower model complexity, and thus lower free energy. Recall from Section 5.1 that for any phase $\mathcal{A} \subseteq W$

$$F_n(\mathcal{A}) \approx nL_n(\omega_0) + \lambda \frac{\log n}{\beta_0}, \tag{5.2}$$

where $L_n(\omega_0) = \min_{\omega \in \mathcal{A}} L_n(\omega)$ is the accuracy and $\lambda \frac{\log n}{\beta_0}$ is the complexity. Since both phases are on W_0, and thus have the same accuracy, these findings suggest that the RLCT λ of $\mathcal{A}_{\text{Degen}}$ is lower than that of $\mathcal{A}_{\text{NonDegen}}$.[1]

The crucial result is this: for $1.26 < \theta < \frac{\pi}{2}$ the degenerate phase $\mathcal{A}_{\text{Degen}}$ does not contain a point on W_0 by Theorem 4.11, thus has worse accuracy, yet it *nonetheless has lower free energy.* We suggest that this is due to the complexity term out-competing the accuracy term in this interval, with the first order phase transition occurring when the accuracy of both phases becomes comparable. Observing Fig. 5.8 we see that for $\theta < 1.3$ the accuracy of $\mathcal{A}_{\text{Degen}}$ is worse than that of $\mathcal{A}_{\text{NonDegen}}$ (meaning $L_n(\omega_0)$ is higher), but then for $\theta \geq 1.3$ the accuracy of the two phases is approximately equal. The preference of $\mathcal{A}_{\text{Degen}}$ over $\mathcal{A}_{\text{NonDegen}}$ in $\theta \in (1.26, \frac{\pi}{2})$ suggests that RLCT of each phase is approximately constant in θ, implying that the first order phase transition is a result of a change in the accuracy of a phase.

Extending this analysis further, we conjecture that as n increases, the critical value $\theta_c = 1.26$ will move closer to $\frac{\pi}{2}$, since the accuracy term is $O(n)$ whereas the complexity is $O(\log n)$. It would be interesting to analyse this in future studies.

As a final remark, as per our discussion in Section 4.4 recall that [PL19] stated that "almost all symmetries of W_0 are scaling and permutation", or alternatively, that degenerate-nodes occur with probability zero in arbitrary-depth ReLU networks. These experiments suggest that while this view is correct, it is incomplete from the perspective of statistical learning. The singularity that determines the phase $\mathcal{A}_{\text{Degen}}$ is non-generic, and yet we have shown that, for these particular networks, it is nonetheless preferred by the posterior even for $\theta < \frac{\pi}{2}$. This implies non-generic points in the space of parameters W can determine the shape of the posterior, and thus influence estimation procedures such as MCMC or Stochastic Gradient Descent. Accordingly, we believe that in order to understand the success of deep learning, the theory should shift perspective from considering points of W to considering singularities of $K(w)$.

[1] In order to add weight to this claim we should attempt to estimate λ, and vary the experiments over n and β_0. We leave this to future work.

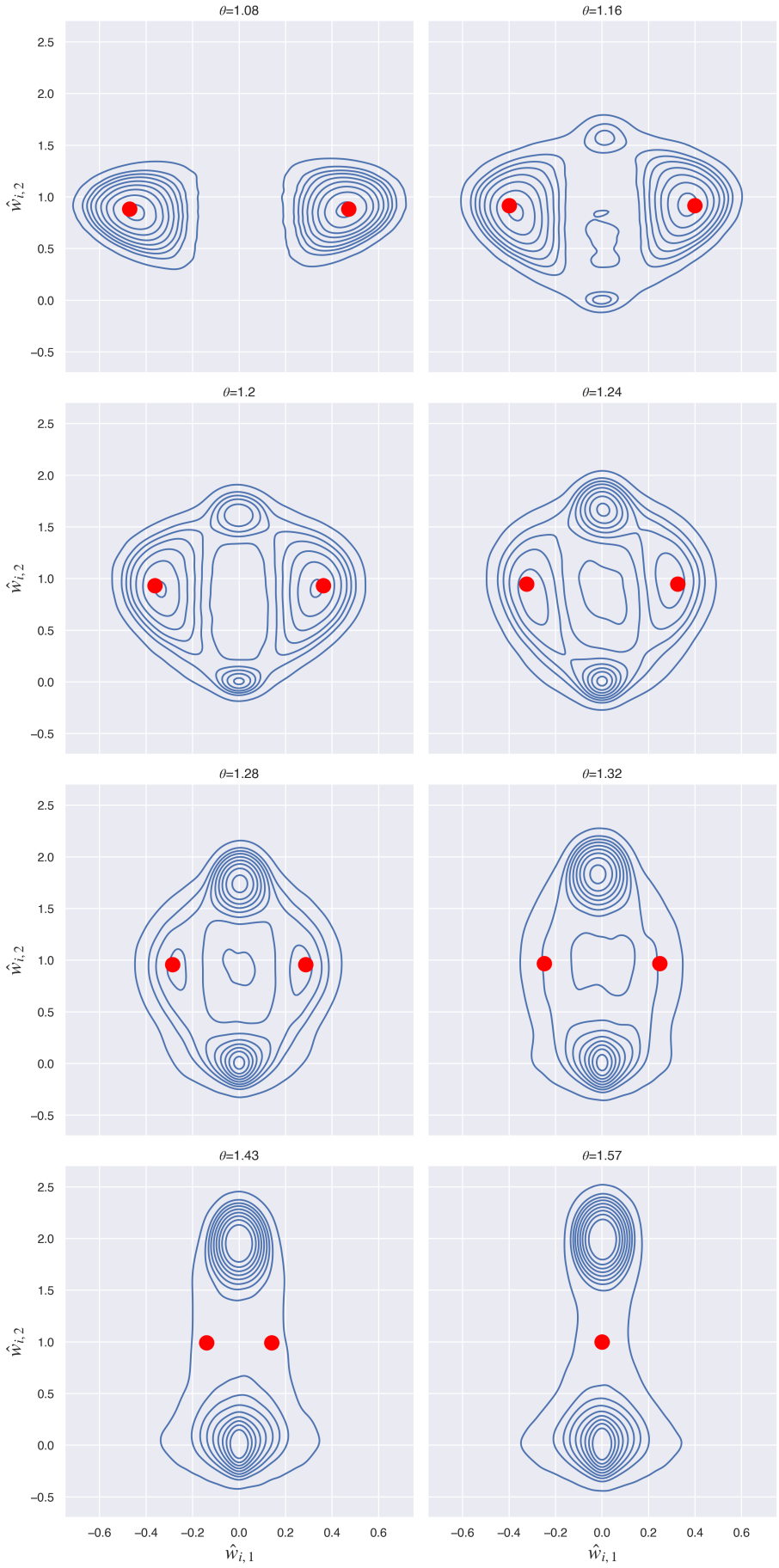
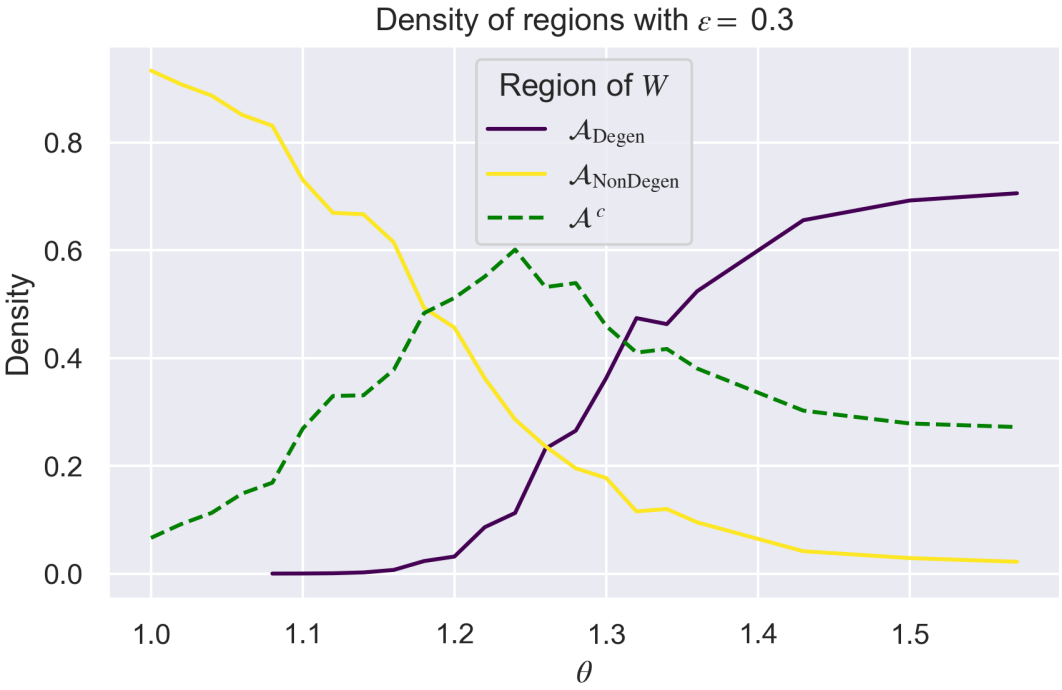
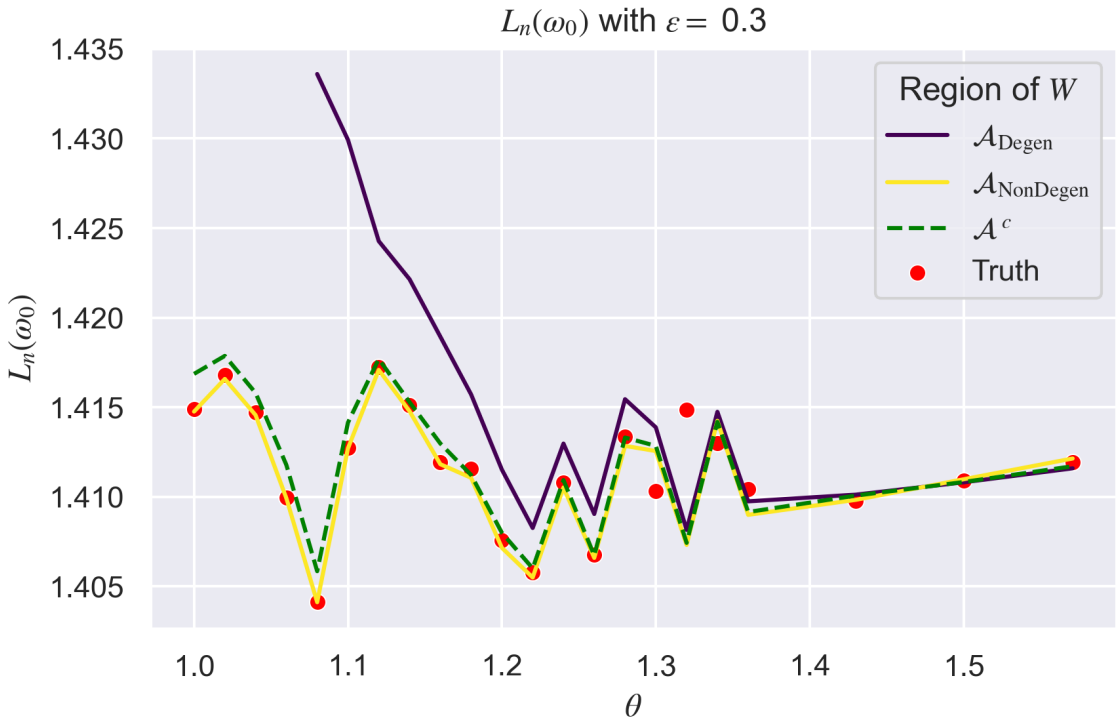


Figure 5.7: Posterior densities of $p^{\beta^*}(w|D_n(\theta))$, where each sample $w^{(k)}$ is represented by two points, $(w_{1,1}^{(k)}, w_{1,2}^{(k)})$ and $(w_{2,1}^{(k)}, w_{2,2}^{(k)})$. The red dots indicate $w_1^{(0)}$ and $w_2^{(0)}$ which are rotated by θ. The $\mathcal{A}_{\text{Degen}}$ phase undergoes a second order creation transition at $\theta \approx 1.16^c$. There is a first order phase transition at $\theta \approx 1.26^c$ as the $\mathcal{A}_{\text{Degen}}$ phase becomes the global minima.



(a) Density (relative frequency) of each phase. Notice how the preferred phase switches at $\theta = 1.26^c$, thus indicating a first order phase transition.



(b) Accuracy $L_n(\omega_0)$ of each phase. Notice that variations in $L_n(\omega_0)$ are closely correlated the accuracy $L_n(w^{(0)}) = S_n$ of the underlying true distribution. These variations in S_n are random (i.e. not constant) since each $y \sim q(y|x)$ which is a normal distribution as outlined in Section 5.2.2.

Figure 5.8: Trajectory of different phases for each θ.

5.4 Phase Transition 2: Orientation Reversing Symmetry

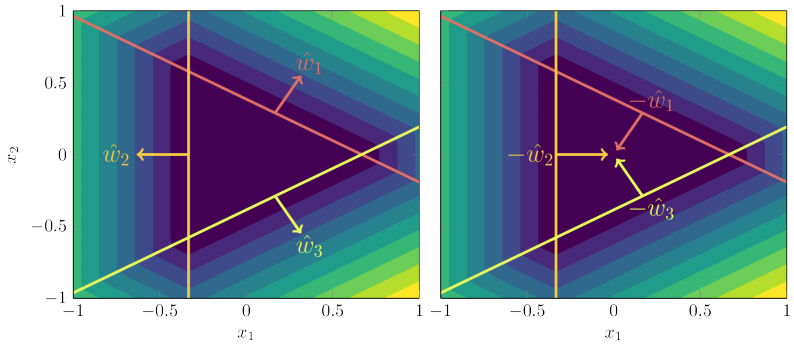


Figure 5.9: Non-weight-annihilation with $\epsilon = (0, 0, 0)$ (left) versus weight-annihilation with $\epsilon = (1, 1, 1)$ (right).

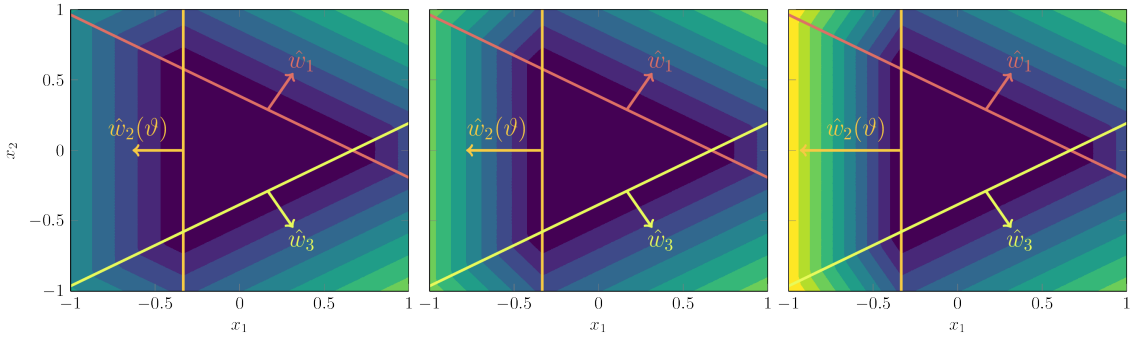


Figure 5.10: $f_3(x, \vartheta)$ for $\vartheta = 1$ (left), $\vartheta = 1.5$ (middle) and $\vartheta = 2$ (right).

5.4.1 Defining the order parameter

Let us now analyse the singularity associated to orientation reversing symmetry of Theorem 4.7, which is present when effective weight vectors in the network sum to zero. Since the true parameter defining W_0 is not unique, we will need to be more careful in distinguishing between the two singularities. To this end, when we refer to *weight annihilation symmetry*, we mean a configuration of weights such that multiple nodes are active in a linear domain, but cancel to give an effective weight of zero. This is shown in Fig. 5.9.

We set the true network to be an m-symmetric network $f_3(x, \vartheta) : \mathbb{R}^2 \times \Theta \to \mathbb{R}$ with $m = 3$ hidden nodes and set $w_0 = g_{\frac{\pi}{3}}(1, 0)^T$ and $b = -\frac{1}{3}$. We define an order parameter $\vartheta \in \Theta = [1, 2.25]$ such that $q_2 = \vartheta$, so

$$f_3(x, \theta) = \text{ReLU}\left(\cos\left(\frac{\pi}{3}\right)x_1 + \sin\left(\frac{\pi}{3}\right)x_2 - \frac{1}{3}\right) + \vartheta\,\text{ReLU}\left(-x_1 - \frac{1}{3}\right) \tag{5.3}$$
$$+ \text{ReLU}\left(\cos\left(\frac{5\pi}{3}\right)x_1 + \sin\left(\frac{5\pi}{3}\right)x_2 - \frac{1}{3}\right).$$

In essence ϑ merely scales the effective weight $\hat{w}_2(\vartheta)$, which can be seen in Fig. 5.10.

Then according to Corollary 4.16 we have the usual scaling and permutation symmetry of W_0, but weight-annihilation is dependent on ϑ:

$$\Upsilon = \begin{cases} \left\{(0)_{i=1}^3\right\}, & \text{if } \vartheta \neq 1 \\ \left\{(0)_{i=1}^3, (1)_{i=1}^3\right\} & \text{if } \vartheta = 1 \end{cases}.$$

Let $R(\theta) = (\cos\theta, \sin\theta)$ and $\mathcal{B}(x, \varepsilon)$ be the closed ball centred at $x \in \mathbb{R}^2$ of radius ε. As regions in the $(w_{i,1}, w_{i,2})$ plane, out setup in Eq. (5.3) dictates that

$$\mathcal{E}_{\text{NonAnn}} = \bigcup_{\sigma \in S_3} \prod_{k=0}^{2} \mathcal{B}\left(R\left(\frac{\pi}{3} + \frac{2\sigma(k)\pi}{3}\right), \varepsilon\right)$$
$$\mathcal{E}_{\text{Ann}} = \bigcup_{\sigma \in S_3} \prod_{k=0}^{2} \mathcal{B}\left(R\left(\frac{2\sigma(k)\pi}{3}\right), \varepsilon\right)$$

are the two phases that we are interested in.

In these experiments there were $K = 10,000$ samples over $T = 4$ trials and $a = 1$, $\sigma_\varphi = 1$.

5.4.2 Results and discussion

The posterior estimates ranging over $\vartheta \in [1, 2.25]$ can be seen in Fig. 5.11. As anticipated we see that both $\mathcal{E}_{\text{NonAnn}}$ and $\mathcal{E}_{\text{Ann}}$ are minima of the free energy. We see that $\mathcal{E}_{\text{NonAnn}}$ is the global minimum, implying the non-annihilation phase has lower model complexity since they have the same accuracy as true parameters. As ϑ increases there is symmetry breaking of W_0 where the singularity defining $\mathcal{E}_{\text{Ann}}$ is no longer on W_0, meaning $L_n(\omega_0)$ of this phase, and therefore the free energy, should increase. Our experiments agree with this conjecture. The free energy of $\mathcal{E}_{\text{NonAnn}}$ increases as ϑ increases until we see a *second order destruction transition* at $\vartheta \approx 2$ where $\mathcal{E}_{\text{NonAnn}}$ ceases to be a minimum of the free energy. This behaviour agrees with our analysis in Section 5.3.2 and reinforces the fact that singularities on W_0 can have different free energies.

5.4.3 An instructive calculation

To illustrate why $\mathcal{E}_{\text{NonAnn}}$ may have a lower free energy, let us consider a perturbation analysis of $K(w)$ centred at singularities corresponding to weight-annihilation and non-weight-annihilation in a simple two-layer ReLU network with one input and one output. We will be interested in measuring the curvature of $K(w)$ at these two points. Since this loosely corresponds to local density of the normalised posterior $\varphi(w)e^{-nK(w)}$, our results in Section 5.4.2 suggest that $\mathcal{E}_{\text{NonAnn}}$ should have the lower curvature of the two phases.

In principle, such a calculation potentially lacks meaning - after all, we have argued that it is not the Hessian at a point that affects the free energy, but the singularity structure. Nevertheless, by removing the singularity in a very simple setting, we can gain some intuition into why it may be that the complexity of $\mathcal{E}_{\text{NonAnn}}$ is less than $\mathcal{E}_{\text{Ann}}$.

Consider a two-layer, one input, one output ReLU network $f(x, w) : \mathbb{R} \times W \to \mathbb{R}$ with $d = 2$ hidden nodes as the model. The true network $f_0(x)$ has the same architecture and is defined by

$$f_0(x) = \text{ReLU}(-x - 1) + \text{ReLU}(x - 1) + 2\,.$$

Let us define two parameters $w_{\text{NA}}, w_{\text{Ann}} \in W$ which depend on a small δ,

$$f_{w_{\text{NA}}}(x, \delta) = \text{ReLU}((-1 + \delta)x - 1) + \text{ReLU}(x - 1) + 2$$
$$f_{w_{\text{Ann}}}(x, \delta) = \text{ReLU}((1 + \delta)x + 1) + \text{ReLU}(-x + 1)\,,$$

where $f_{w_{\text{NA}}}$ is the non-weight-annihilation configuration and $f_{w_{\text{Ann}}}$ is weight-annihilation. Note that both $w_{\text{NA}}, w_{\text{Ann}} \in W_0$ for $\delta = 0$. Let us then define the KL divergence between from each network to $f_0(x)$ as a function of δ, for $q(x)$ uniform on $[-a, a]$ for some $a > 0$,

$$K_{\text{NA}}(\delta) = \int_{-a}^{a} \left(f_{w_{\text{NA}}}(x, \delta) - f_0(x)\right)^2 dx$$
$$K_{\text{Ann}}(\delta) = \int_{-a}^{a} \left(f_{w_{\text{Ann}}}(x, \delta) - f_0(x)\right)^2 dx\,.$$

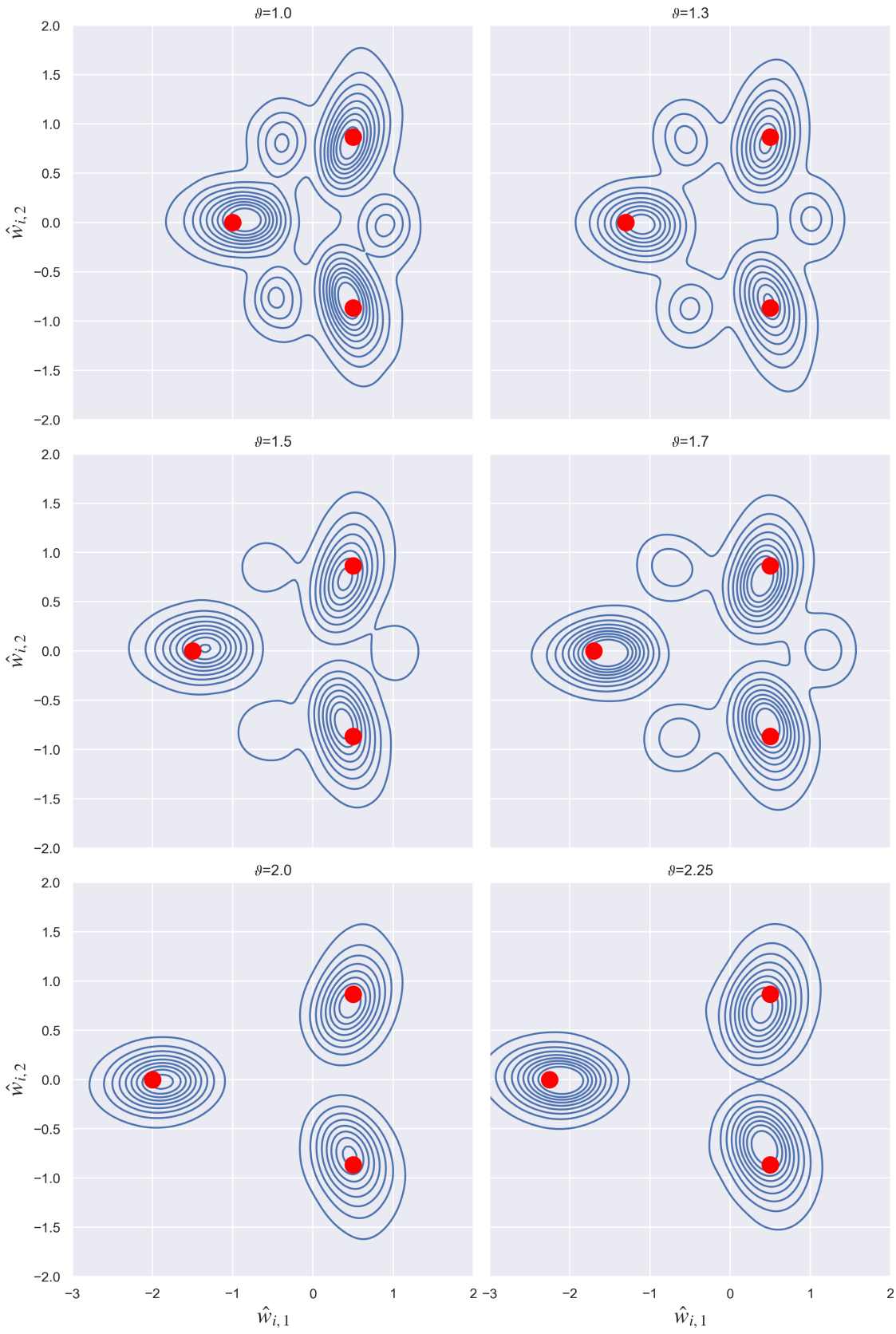


Figure 5.11: Posterior densities of $p^{\beta^*}(w|D_n(\vartheta))$ for Eq. (5.3). The red dots indicate $\hat{w}_1^{(0)}$, $\hat{w}_2^{(0)}$ and $\hat{w}_3^{(0)}$, where $\hat{w}_2^{(0)}$ is scaled by ϑ. The $\mathcal{E}_{\mathrm{Ann}}$ phase always has higher free energy, and undergoes a second order destruction transition somewhere between $\vartheta \in (1.7, 2)$.

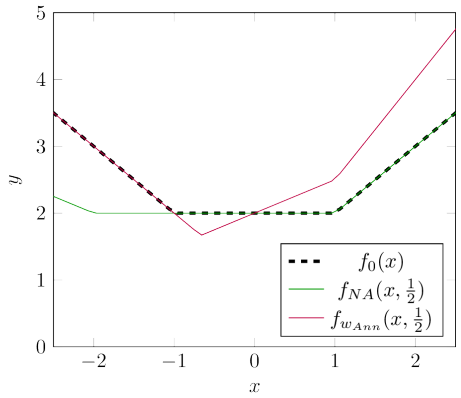


Figure 5.12: The different perturbed networks compared to the truth. Notice the additional area between $f_{w_{\text{Ann}}}$ and f_0 in the region $(-1, 1)$.

Through a careful calculation performing the integrals over domains defined by when each node is active, one finds that

$$K_{\text{NA}}(\delta) = \begin{cases} \dfrac{\delta^2\left(a^3+\delta-1\right)}{12a} & \text{if } \delta < 0 \\ \dfrac{\delta^2\left(a^3(\delta-1)^3-(\delta-1)\right)}{12a(\delta-1)^3} & \text{if } 0 < \delta < 1 \end{cases},$$

$$K_{\text{Ann}}(\delta) = \begin{cases} \dfrac{\delta^2\left(a^3\delta+a^3+1\right)}{12a(\delta+1)} & \text{if } -1 < \delta < 0 \\ \dfrac{\delta^2\left(a^3(\delta+1)^3+\delta+1\right)}{12a(\delta+1)^3} & \text{if } \delta > 0 \end{cases}.$$

In particular we have

$$K''(0) = \frac{a^3-1}{6a} < \frac{a^3+1}{6a} = K''_\epsilon(0)\,,$$

which implies that for any $\delta \in (-1, 1)\backslash\{0\}$,

$$K_{\text{NA}}(\delta) < K_{\text{Ann}}(\delta)\,.$$

Observing Fig. 5.12 gives some insight into the geometry at play here. Small perturbations in $f_{w_{\text{Ann}}}$ lead to a meaningful change in the decision boundaries, which in turn results in additional contributions to the KL divergence from the region $(-1, 1)$. On the other hand, $f_{w_{\text{NA}}}$ still retains a constant region for small perturbations in δ.

Furthermore, notice that

$$\lim_{a\to\infty} |K''_{\text{Ann}}(0) - K''_{\text{NA}}(0)| = \lim_{a\to\infty} \frac{1}{3a} = 0\,,$$

which implies that the free energy of both phases may become equal as $a \to \infty$. This hints at the existence of another first order phase transition. Let us investigate whether $\mathcal{E}_{\text{Ann}}$ becomes a global minmia as we vary a.

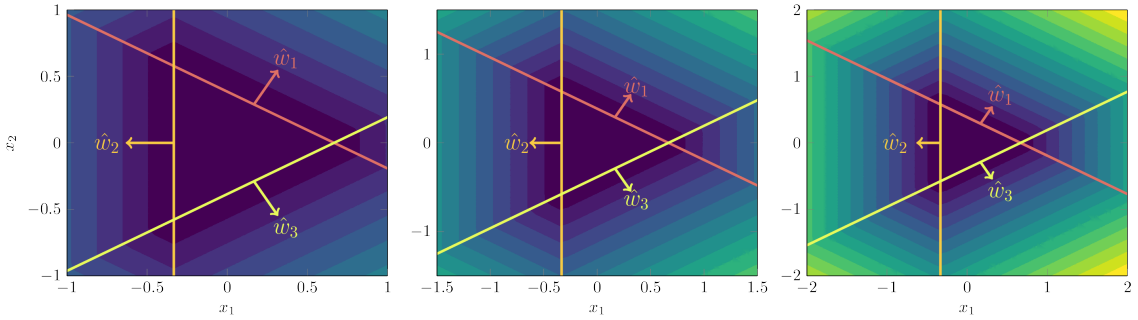


Figure 5.13: The true distribution $q(y, x) = \frac{1}{a^2} p(y|x, w_3)\mathbb{1}((x_1, x_2) \in [-a, a]^2)$ for $a = 1$ (left), $a = 1.5$ (middle) and $a = 2$ (right).

5.4.4 Phase transition 3: Equal Weight-Annihilation

Inspired by the results in Section 5.4.3, let us keep our same network $f_3(x)$ defined in Section 5.4.1, but this time we shall fix $q_2 = \vartheta = 1$ and make the order parameter $a \in \Theta = [1, 2]$ which defines the uniform square of $q(x)$. How a changes the distribution is seen in Fig. 5.13. According to Theorem 4.7, there are no changes to the symmetries of W_0 as we vary a. However, it is worth bearing in mind Remark 4.12, where we observed that our analysis of W_0 in Chapter 4 did not adequately take into account how the form of $q(x)$ affected W_0.

In these experiments there were $K = 1000$ samples over $T = 32$ trials and $\sigma_\varphi = 1$. The reason for more trials is because potential barriers were found, meaning the final distribution of any trial chain was highly dependent on its initial value. Thus these experiments were averaged over random initial values for 32 trials.

The results can be seen in Fig. 5.14. The regions of posterior concentration are not as distinct as in previous experiments, but the results nonetheless agree with what Section 5.4.3 suggested: the free energy of $\mathcal{E}_{\text{Ann}}$ decreases as a increases. This shows how the geometry of $K(w)$ is not just affected by the nature of the ReLU network defining the true distribution, but also the specifications of $q(x)$.

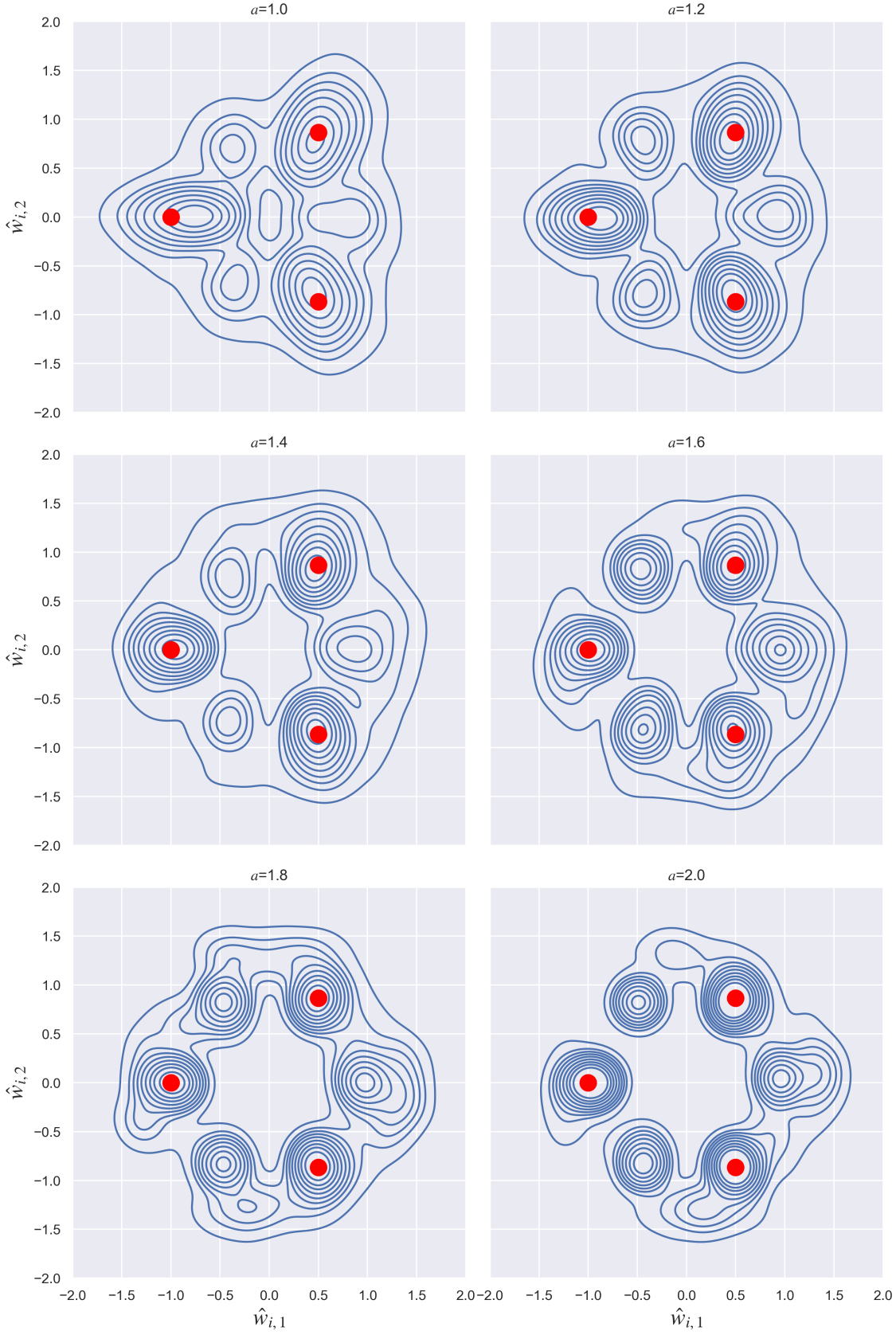


Figure 5.14: Posterior densities of $p^{\beta^*}(w|D_n(a))$ for $f_3(x)$ where $q(x)$ is uniform on $[-a, a]^2$. Notice that the free energies of $\mathcal{E}_0$ and $\mathcal{E}_1$ become comparable as a increases.

Chapter 6

Conclusion

This thesis aimed to provide accessible examples of singular models in the form of small ReLU networks in order to elucidate the key messages of Sumio Watanabe's *Singular Learning Theory* and illuminate the change in statistical perspective from points to singularities.

We started by casting deep learning as a Bayesian statistical learning model in Chapter 2. Here the Kullback-Leibler divergence $K(w)$ between a model and truth was revealed as the fundamental object of study, alongside the set of true parameters $W_0 \subseteq W$. We then explained how one can draw an analogy between neural networks as Bayesian models and the Gibbs ensemble of statistical physics, hinting at objects and phenomena that arise naturally such as the free energy and phase transitions.

An exposition on Singular Learning Theory was then provided in Chapter 3. We began by demonstrating that neural networks have degenerate Fisher information matrices and are therefore singular models, thus showing that singularity theory lies at the heart of statistical learning theory of neural networks. We then explored the free energy $F_n(\mathcal{W})$ associated to compact subsets of W, in particular its relation to the generalisation of a model, and why it is the main quantity of comparison between model. Watanabe's groundbreaking formula for the asymptotics of the free energy in singular models was then explained, where we discussed how the RLCT λ is the correct measure of complexity in singular models and illustrated its interpretation in terms of Occam's Razor.

We then set about establishing the symmetries of W_0 for two layer ReLU networks in the realisable case in Chapter 4, which was equivalent to establishing for which parameters give functional equivalence between a model and a truth network. This was done in two stages. In the first case where the model and truth networks had the same number of nodes, $m = d$, it was found that W_0 exhibited scaling, permutation and orientation reversing symmetry, the latter only occurring when the weight vectors summed to zero. In the second case where the model was had more hidden nodes than the truth, $m < d$, it was proven that the excess nodes were either degenerate or had the same activation boundaries as some other node in the model. We then examined a more general result from the literature for networks of arbitrary depth, before finally providing the example of m-symmetric networks which exhibited interesting symmetries of W_0.

In Chapter 5 we endeavoured to show that not all points on W_0 were equally good minimisers of the free energy due to their difference as singularities. Initially we explained a correspondence between phases and singularities of $K(w)$. This naturally led to the notion of phase transitions, which we argued occurred as a result of a substantial change in the accuracy or RLCT of a compact subset of W, thus being associated to symmetry breaking of W_0. We were able to show that points on W_0 could indeed have different free energies due to having different model complexity. The key finding was that points on $W \backslash W_0$ could still be favoured by the posterior despite not being minimisers of $K(w)$. We first showed that the complexity of degenerate-node singularities was less than that of non-degenerate node singularities. Moreover, we demonstrated a phase transition in these networks corresponding to a change in the accuracy of the degenerate-node phase. Finally we showed that weight-annihilation singularities had greater free energy than non-weight-annihilation

singularities and provided intuition for why this may be the case.

Given the infancy of Singular Learning Theory there remain many key questions that should be examined in future studies. Relating to this thesis, we think it would be interesting to explore:

- Numerical approximations of the RLCT of our established phases.
- Theoretical values of the RLCT for these phases using the swish approximation to ReLU.
- Generalisations of the proofs in Chapter 4 to examine W_0 of networks with arbitrary input dimension, output dimension, depth, and sequences of hidden layer widths.
- Scaling laws near critical values of phase transitions, particularly how these scaling exponents may arise theoretically, and how the RLCT is related.

In summary, not all points on W_0 are equally good. Thus, we believe it is time to evolve statistical analysis of deep learning from considering points to investigating singularities.

Appendix A

Appendix

Lemma A.1. *Let $q(y,x)$ and $p(y,x|w) > 0$ be continuous probability density functions. Then $K(w) \geq 0$ for all $w \in W$, and $K(w) = 0$ if and only if $p(y|x,w) = q(y|x)$ for almost all $x \in \mathbb{R}^N$, $y \in \mathbb{R}^M$.*

Proof. First note that if $q(y,x) = 0$ on some open set $A \subseteq \mathbb{R}^{N+M}$, since $\lim_{x\to 0} x \log x = 0$ we may define in good conscience

$$q(y,x)\log q(y,x) - q(y,x)\log p(y,x|w) := 0.$$

Thus there will be no contribution to $K(w)$ from the region A, so we can assume without loss of generality that $q(y,x) > 0$ on the region of integration.

Consider the real-valued function $S(t) = -\log t + t - 1$ for $t \in (0,\infty)$ which is well defined, continuous and differentiable everywhere on this domain. Then clearly $S(1) = 0$, and indeed we can show that $t = 1$ is the only root. Since $S'(t) = -\frac{1}{t} + 1$, $S(t)$ has a stationary point at $t = 1$, is strictly decreasing on $(0,1)$ and strictly increasing on $(1,\infty)$, thus by continuity we see that $t = 1$ is the only root. Then since $S''(t) = \frac{1}{t^2}$, so $S''(1) = 1 > 0$, we see that S is concave up at $t = 1$, thus showing $S(t) \geq 0$ for all $t \in (0,\infty)$ and $S(t) = 0$ if and only if $t = 1$.

But then since p and q are probability distributions, hence $\iint_{\mathbb{R}^{N+M}} p(y,x|w)dxdy = 1$ and $\iint_{\mathbb{R}^{N+M}} q(y,x)dxdy = 1$, we have

$$\begin{aligned}\iint_{\mathbb{R}^{N+M}} q(y,x)S\left(\frac{p(y,x|w)}{q(y|x)}\right)dxdy &= \iint_{\mathbb{R}^{N+M}} q(y,x)\log\left(\frac{q(y,x)}{p(y,x|w)}\right)dxdy \\ &\quad + \iint_{\mathbb{R}^{N+M}} q(y,x)\frac{p(y,x|w)}{q(y|x)}dxdy - \iint_{\mathbb{R}^{N+M}} q(y,x)dxdy \\ &= K(w)\,.\end{aligned}$$

Since $q(y,x)$, $p(y,x|w) > 0$ we have $0 < \frac{p(y,x|w)}{q(y,x)} < \infty$, hence the integrand in the first integral is non-negative, thus the integral itself is non-negative, so $K(w) \geq 0$.

We have shown that if $p(y,x|w) = q(y,x)$ then $K(w) = 0$, so suppose $K(w) = 0$. Since $S(t) \geq 0$ and $q(y,x) > 0$ are continuous and non-negative on $\mathbb{R}^{N+M}$, by standard real analysis results we must have $S\left(\frac{p(y,x|w)}{q(y,x)}\right) = 0$ for almost all $(x,y) \in \mathbb{R}^{N+M}$, hence $\frac{p(y,x|w)}{q(y,x)} = 1$ as stated. □

Lemma A.2. *Let $q(y|x) = p(y|x,w_0)$ be realisable, defined by a parameter $w_0 \in W$. Then*

$$K(w) = \frac{1}{2}\int_{\mathbb{R}^N} \|f(x,w) - f(x,w_0)\|^2 q(x)\,dx\,. \tag{A.1}$$

Proof. We calculate $K(w)$ to be

$$\iint_{\mathbb{R}^{N+M}} \frac{q(x)}{(2\pi)^{\frac{M}{2}}} \exp\left(-\frac{1}{2}\|y-f(x,w_0)\|^2\right) \log\left(\frac{\frac{1}{(2\pi)^{\frac{M}{2}}}\exp\left(-\frac{1}{2}\|y-f(x,w_0)\|^2\right)}{\frac{1}{(2\pi)^{\frac{M}{2}}}\exp\left(-\frac{1}{2}\|y-f(x,w)\|^2\right)}\right) dxdy\,.$$

$$= \frac{1}{2(2\pi)^{\frac{M}{2}}} \iint_{\mathbb{R}^{N+M}} q(x)\exp\left(-\frac{1}{2}\|y-f(x,w_0)\|^2\right)\left(\|y-f(x,w)\|^2 - \|y-f(x,w_0)\|^2\right) dxdy\,.$$

Let $u = y - f(x,w_0)$, so $du = dy$, and let $a = f(x,w) - f(x,w_0) \in \mathbb{R}^M$ which is fixed, then $y - f(x,w) = u - a$ and so

$$K(w) = \frac{1}{2(2\pi)^{\frac{M}{2}}} \int_{\mathbb{R}^N} q(x)K(w,w_0,x)dx\,, \tag{A.2}$$

where for a fixed $x \in \mathbb{R}^N$ we define

$$K(w,w_0,x) = \int_{\mathbb{R}^M} e^{-\frac{1}{2}\|u\|^2}\left(\|u-a\|^2 - \|u\|^2\right)du = \int_{\mathbb{R}^M} e^{-\frac{1}{2}\|u\|^2}\left(-2a\cdot u + \|a\|^2\right)du\,. \tag{A.3}$$

Recall the standard identity $\int_{\mathbb{R}^M} e^{-\frac{1}{2}\|x\|^2}dx = (2\pi)^{\frac{M}{2}}$. For the dot product term we can show that this contribution is zero by induction on the dimension M. The base case for $M = 1$ is simply $\int_{-\infty}^{\infty} a_1 u_1 e^{-\frac{1}{2}u_1^2}du = 0$ since it is an odd integrand over a symmetric domain. For the inductive step, denote $a = (a_1,\ldots,a_M)$ and $u = (u_1,\ldots,u_M)$ and suppose $\int_{\mathbb{R}^M}(a\cdot u)e^{-\frac{1}{2}\|u\|^2}du = 0$. Then

$$\begin{aligned}
&\int_{\mathbb{R}^M}\int_{-\infty}^{\infty}(a\cdot u + a_{M+1}u_{M+1})e^{-\frac{1}{2}(\|u\|^2+u_{M+1}^2)}du\,du_{M+1}\\
&= \int_{-\infty}^{\infty} e^{-\frac{1}{2}u_{M+1}^2}du_{M+1}\int_{\mathbb{R}^M}(a\cdot u)e^{-\frac{1}{2}\|u\|^2}du\\
&\quad + \int_{-\infty}^{\infty} a_{M+1}u_{M+1}e^{-\frac{1}{2}u_{M+1}^2}du_{M+1}\int_{\mathbb{R}^M} e^{-\frac{1}{2}\|u\|^2}du\\
&= 0\,,
\end{aligned}$$

where the first integral vanishes by the inductive hypothesis and the second due to the odd integral over a symmetric domain. Substituting this into (A.3) gives

$$K(w,w_0,x) = \|a\|^2\int_{\mathbb{R}^M} e^{-\frac{1}{2}\|u\|^2}du = (2\pi)^{\frac{M}{2}}\|a\|^2\,,$$

and so recalling the definition of a and substituting into (A.2) yields the result. □

Lemma A.3. *Let $\varphi(w) > 0$ be a prior on W. Suppose $P(w)$ is the unique maximiser of the relative entropy $K(P||\varphi(w))$ subject to the constraint*

$$\mathbb{E}_{w\sim P}[nL_n(w)] = \mu_\beta$$

for some fixed $\mu_\beta \in \mathbb{R}$. Then $P(w) = p^\beta(w|D_n)$ for some $\beta > 0$ that depends on μ_β

Proof. Given the relative entropy functional

$$K(P||\varphi) = \int_W P(w)\log\frac{P(w)}{\varphi(w)}dw\,,$$

we want to solve for the probability distribution $P(w)$ that maximises $K(P||\varphi)$ subject to the following constraints:

$$-\sum_{i=1}^{n}\int_W P(w)\log p(y_i|x_i,w)dw = \mu_\beta\,,\quad\text{and}\quad \int_W P(w)dw = 1\,,\quad\text{and}\quad \int_W \varphi(w)dw = 1\,,$$

where $\mu_\beta \in \mathbb{R}$ is assumed fixed and given. Let $k(w, P)$ denote the integrand in $K(P||\varphi)$ and let $g_1(w, P)$, $g_2(w, P)$ and $g_3(w, P)$ respectively denote the integrands in the constraints above and let λ_1, λ_2 and λ_3 denote respective Lagrange multipliers. Then we wish to freely optimise the functional

$$\mathcal{F}[\{\lambda_i\}, P(w)] = \int_W \left[k(w, P) - \sum_{j=1}^{3} \lambda_j g_j(w, P) \right] dw + \lambda_1 \mu_\beta + \lambda_2 + \lambda_3 \,.$$

We can appeal to the Euler-Lagrange equation which states that $\mathcal{F}$ is extremised at the function P such that

$$\frac{d}{dw}\left(\frac{\partial k}{\partial P'}\right) - \frac{\partial k}{\partial P} - \sum_{j=1}^{3} \lambda_j \left[\frac{d}{dw}\left(\frac{\partial g_j}{\partial P'}\right) - \frac{\partial g_j}{\partial P} \right] = 0$$

subject to the same constraints as above. This then evaluates to

$$\log \frac{P(w)}{\varphi(w)} + 1 - \lambda_1 \sum_{i=1}^{n} \log p(y_i|x_i, w) + \lambda_2 = 0 \,,$$
$$\text{so} \quad P(w) = e^{-(1+\lambda_2)} \varphi(w) \prod_{i=1}^{n} p^{\lambda_1}(y_i|x_i, w) \,.$$

Then λ_1 and λ_2 can be solved by applying the first two constraints, giving $P(w) = p^\beta(w|D_n)$ for $\beta = \lambda_1$. □

Lemma A.4. *Let $\mathcal{W} \subseteq W$ be compact. The free energy of $\mathcal{W}$ satisfies*

$$\frac{\partial F_n^\beta(\mathcal{W})}{\partial \beta} = \mathbb{E}^\beta_{\mathcal{W}}[nL_n(w)] = nG_t^\beta(\mathcal{W}) \,,$$
$$\textit{and} \quad \frac{\partial^2 F_n^\beta(\mathcal{W})}{\partial \beta^2} = -\mathbb{E}^\beta_{\mathcal{W}}[(nL_n(w))^2] + \mathbb{E}^\beta_{\mathcal{W}}[nL_n(w)]^2 = -\mathbb{V}^\beta_{\mathcal{W}}[nL_n(w)] \,.$$

Proof. The first proof was provided in the main body of the text. For the second derivative we have

$$\begin{aligned}
\frac{\partial^2 F_n^\beta(\mathcal{W})}{\partial \beta^2} &= -\frac{\partial}{\partial \beta}\left(\frac{1}{Z_n^\beta(\mathcal{W})}\frac{\partial Z_n^\beta(\mathcal{W})}{\partial \beta}\right) = -\left(\frac{\partial}{\partial \beta}\frac{1}{Z_n^\beta(\mathcal{W})}\right)\frac{\partial Z_n^\beta(\mathcal{W})}{\partial \beta} - \frac{1}{Z_n^\beta(\mathcal{W})}\frac{\partial^2 Z_n^\beta(\mathcal{W})}{\partial \beta^2} \\
&= \left(\frac{1}{Z_n^\beta(\mathcal{W})}\frac{\partial Z_n^\beta(\mathcal{W})}{\partial \beta}\right)^2 - \frac{1}{Z_n^\beta(\mathcal{W})}\int_{\mathcal{W}} (nL_n(w))^2 \varphi(w) e^{-n\beta L_n(w)} dw \\
&= \mathbb{E}^\beta_{\mathcal{W}}[nL_n(w)]^2 - \mathbb{E}^\beta_{\mathcal{W}}[(nL_n(w))^2] \,.
\end{aligned}$$

□

Lemma A.5. *Let F_n denote the free energy when $\beta = 1$. The generalisation loss is the average increase in free energy,*

$$G_n = \mathbb{E}_{X_{n+1}}[F_{n+1}] - F_n \,. \tag{A.4}$$

In particular, the average free energy is the sum of the generalisation loss,

$$\mathbb{E}_{D_n}[F_n] = \sum_{i=1}^{n-1} \mathbb{E}_{D_i}[G_i] + \mathbb{E}_{D_1}[F_1] \,.$$

Proof. The proof hinges on the fact that we may write

$$\frac{Z_{n+1}}{Z_n} = \frac{\int_W p(y_{n+1}|x_{n+1}, w)\varphi(w)e^{-n\beta L_n(w)}dw}{\int_W \varphi(w)e^{-n\beta L_n(w)}dw} = \mathbb{E}_w[p(y_{n+1}|x_{n+1}, w)] = p(y|x, D_n)$$

which implies

$$F_{n+1} - F_n = -\log p(y|x, D_n)\,.$$

Since F_n does not depend on (X_{n+1}, Y_{n+1}), taking $\mathbb{E}_{X_{n+1}}$ of both sides gives the first result. Taking expectation with respect to D_n of Eq. (A.4) gives

$$\mathbb{E}_{D_n}[G_n] = \mathbb{E}_{D_{n+1}}[F_{n+1}] - \mathbb{E}_{D_n}[F_n]\,.$$

Thus we have

$$\begin{aligned}\sum_{i=1}^{n-1} \mathbb{E}_{D_i}[G_i] + \mathbb{E}_{D_1}[F_1] &= (\mathbb{E}_{D_n}[F_n] - \mathbb{E}_{D_{n-1}}[F_{n-1}]) + (\mathbb{E}_{D_{n-1}}[F_{n-1}] - \mathbb{E}_{D_{n-2}}[F_{n-2}]) \\ &\quad + \cdots + (\mathbb{E}_{D_2}[F_2] - \mathbb{E}_{D_1}[F_1]) + \mathbb{E}_{D_1}[F_1] \\ &= \mathbb{E}_{D_n}[F_n]\,.\end{aligned}$$

□

Lemma A.6. *Let $w, w' \in \mathbb{R}^2 \setminus \{0\}$ and $b, b' \in \mathbb{R}$ be given and let*

$$H = \{x \in \mathbb{R}^2 \,|\, \langle w, x\rangle + b = 0\}, \quad \textit{and} \quad H' = \{x \in \mathbb{R}^2 \,|\, \langle w', x\rangle + b' = 0\}\,.$$

Then $H = H'$ if and only if there exists some scalar $\lambda \in \mathbb{R} \setminus \{0\}$ such that $w = \lambda w'$ and $b = \lambda b'$.

Proof. The first direction is simple: suppose $\lambda \in \mathbb{R} \setminus \{0\}$ is such that $w = \lambda w'$ and $b = \lambda b'$, then if $x \in H'$ we have

$$0 = \langle w', x\rangle + b' = \langle \lambda w, x\rangle + \lambda b = \lambda\left(\langle w, x\rangle + b\right)$$

and so dividing by λ shows that $x \in H$, and by symmetry we clearly have $H' \subseteq H$ too, so $H = H'$.

Now suppose $H = H'$. Let $t \in H$ be a scalar multiple of w, so $t = \mu w$ for some $\mu \in \mathbb{R}$, then

$$0 = \langle w, t\rangle + b = \mu\langle w, w\rangle + b\,, \quad \text{so} \quad \mu = -\frac{b}{\langle w, w\rangle}\,,$$

and so t is the unique point such that $b = -\langle w, t\rangle$. Similarly we have a unique $t' = \mu' w'$, where $\mu' = -\frac{b'}{\langle w', w'\rangle}$, giving $b' = -\langle w', t'\rangle$. Then saying $x \in H$ is now equivalent to $\langle w, x - t\rangle = 0$, but since $x \in H'$ as well we also have $\langle w', x - t'\rangle = 0$. Taking $x = t'$ in the first case and $x = t$ in the second case, noting $\langle w, t' - t\rangle = -\langle w, t - t'\rangle$, we have a system of equations

$$A_w(t - t') := \begin{pmatrix} w_1 & w_2 \\ w_1' & w_2' \end{pmatrix}\begin{pmatrix} t_1 - t_1' \\ t_2 - t_2' \end{pmatrix} = 0\,.$$

Thus either $t = t'$ or $\operatorname{rank}(A_w) = 1$ (w and w' are nonzero by hypothesis, excluding the possibility of $\operatorname{rank}(A_w) = 0$). In the first case we have $t = \mu w = \mu' w' = t'$, thus we can take $\lambda = \frac{\mu}{\mu'} \in \mathbb{R}$ to give $w = \lambda w'$. In the second case, $\operatorname{rank}(A_w) = 1$ implies w and w' are linearly dependent, thus $w = \lambda w'$ for some $\lambda \in \mathbb{R}$. For such a λ we thus have

$$b = -\langle w, t\rangle = -\langle w, t'\rangle = -\lambda\langle w', t'\rangle = \lambda b'\,,$$

where the second equality follows from $\langle w, t - t'\rangle = 0$, thus proving the claim. □

Bibliography

[Bal97] Vijay Balasubramanian. "Statistical Inference, Occam's Razor, and Statistical Mechanics on the Space of Probability Distributions". In: *Neural Computation* 9.2 (Feb. 15, 1997), pp. 349–368. ISSN: 0899-7667. DOI: 10.1162/neco.1997.9.2.349. URL: https://doi.org/10.1162/neco.1997.9.2.349 (visited on 08/06/2021).

[Bet18] Michael Betancourt. "A Conceptual Introduction to Hamiltonian Monte Carlo". In: *arXiv:1701.02434 [stat]* (July 15, 2018). arXiv: 1701.02434. URL: http://arxiv.org/abs/1701.02434 (visited on 09/18/2021).

[Bin+19] Eli Bingham et al. "Pyro: Deep Universal Probabilistic Programming". In: *J. Mach. Learn. Res.* 20 (2019), 28:1–28:6. URL: http://jmlr.org/papers/v20/18-403.html.

[Bri] *Occam's razor*. Encyclopedia Britannica. URL: https://www.britannica.com/topic/Occams-razor (visited on 09/18/2021).

[Bro+20] Tom B. Brown et al. "Language Models are Few-Shot Learners". In: *arXiv:2005.14165 [cs]* (July 22, 2020). arXiv: 2005.14165. URL: http://arxiv.org/abs/2005.14165 (visited on 09/19/2021).

[Bro11] Steve Brooks. *Handbook of Markov Chain Monte Carlo*. OCLC: 751677317. Hoboken: Chapman & Hall/CRC, 2011. ISBN: 9781420079425. URL: http://public.eblib.com/choice/publicfullrecord.aspx?p=762505 (visited on 09/15/2021).

[Cal85] Herbert B. Callen. *Thermodynamics and an introduction to thermostatistics*. 2nd ed. New York: Wiley, 1985. 493 pp. ISBN: 9780471862567.

[CB02] George Casella and Roger L. Berger. *Statistical inference*. 2nd ed. Australia ; Pacific Grove, CA: Thomson Learning, 2002. 660 pp. ISBN: 9780534243128.

[GBC16] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. http://www.deeplearningbook.org. MIT Press, 2016.

[Gil93] Robert Gilmore. *Catastrophe theory for scientists and engineers*. New York: Dover Publications, 1993. 666 pp. ISBN: 9780486675398.

[Har10] Robin Hartshorne. *Algebraic geometry*. Nachdr. Graduate texts in mathematics 52. New York, N.Y: Springer, 2010. 496 pp. ISBN: 9781441928078.

[Hir64] Heisuke Hironaka. "Resolution of Singularities of an Algebraic Variety Over a Field of Characteristic Zero: I". In: *Annals of Mathematics* 79.1 (1964), pp. 109–203. ISSN: 0003-486X. DOI: 10.2307/1970486. URL: https://www.jstor.org/stable/1970486 (visited on 10/06/2021).

[Hun07] J. D. Hunter. "Matplotlib: A 2D graphics environment". In: *Computing in Science & Engineering* 9.3 (2007), pp. 90–95. DOI: 10.1109/MCSE.2007.55.

[Iva10] Ivanov, Oleg A. "On the Number of Regions into Which n Straight Lines Divide the Plane". In: *The American Mathematical Monthly* 117.10 (2010), p. 881. ISSN: 00029890. DOI: 10.4169/000298910x523362. URL: https://www.tandfonline.com/doi/full/10.4169/000298910X523362 (visited on 08/28/2021).

[Kap+20] Jared Kaplan et al. "Scaling Laws for Neural Language Models". In: *arXiv:2001.08361 [cs, stat]* (Jan. 22, 2020). arXiv: 2001.08361. URL: http://arxiv.org/abs/2001.08361 (visited on 09/19/2021).

[KK08] Sadanori Konishi and G. Kitagawa. *Information criteria and statistical modeling*. Springer series in statistics. New York: Springer, 2008. 273 pp. ISBN: 9780387718866 9780387718873.

[Laf17] Lev Lafayette. *Spartan HPC-Cloud Hybrid: Delivering Performance and Flexibility*. figshare. Apr. 10, 2017. DOI: 10.4225/49/58ead90dceaaa. URL: https://melbourne.figshare.com/articles/online_resource/Spartan_HPC-Cloud_Hybrid_Delivering_Performance_and_Flexibility/4768291/1 (visited on 10/05/2021).

[Lan02] Serge Lang. *Algebra*. Rev. 3rd ed. Graduate texts in mathematics 211. New York: Springer, 2002. 914 pp. ISBN: 9780387953854.

[LBH15] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. "Deep learning". In: *Nature* 521.7553 (May 28, 2015), pp. 436–444. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature14539. URL: http://www.nature.com/articles/nature14539 (visited on 09/19/2021).

[LeC14] Yan LeCun. "The Unreasonable Effectiveness of Deep Learning". Johns Hopkins University, Center for Language and Speech Processing, Nov. 18, 2014. URL: https://www.ee.ucl.ac.uk/sahd2014/resources/LeCun.pdf.

[Lu+17] Zhou Lu et al. "The Expressive Power of Neural Networks: A View from the Width". In: *arXiv:1709.02540 [cs]* (Nov. 1, 2017). arXiv: 1709.02540. URL: http://arxiv.org/abs/1709.02540 (visited on 08/12/2021).

[McE] Richard McElreath. *Markov Chains: Why Walk When You Can Flow?* URL: https://elevanth.org/blog/2017/11/28/build-a-better-markov-chain/.

[MHB18] Stephan Mandt, Matthew D. Hoffman, and David M. Blei. "Stochastic Gradient Descent as Approximate Bayesian Inference". In: *arXiv:1704.04289 [cs, stat]* (Jan. 19, 2018). arXiv: 1704.04289. URL: http://arxiv.org/abs/1704.04289 (visited on 08/13/2021).

[Min+20] Chris Mingard et al. "Is SGD a Bayesian sampler? Well, almost". In: *arXiv:2006.15191 [cs, stat]* (Oct. 24, 2020). arXiv: 2006.15191. URL: http://arxiv.org/abs/2006.15191 (visited on 08/12/2021).

[Mur+20] Daniel Murfet et al. "Deep Learning is Singular, and That's Good". In: *arXiv:2010.11560 [cs]* (Oct. 22, 2020). arXiv: 2010.11560. URL: http://arxiv.org/abs/2010.11560 (visited on 09/05/2021).

[Nak+19] Preetum Nakkiran et al. "Deep Double Descent: Where Bigger Models and More Data Hurt". In: *arXiv:1912.02292 [cs, stat]* (Dec. 4, 2019). arXiv: 1912.02292. URL: http://arxiv.org/abs/1912.02292 (visited on 09/19/2021).

[Ope16] *Generative Models*. OpenAI. June 16, 2016. URL: https://openai.com/blog/generative-models/ (visited on 08/12/2021).

[Pas+19] Adam Paszke et al. "PyTorch: An Imperative Style, High-Performance Deep Learning Library". In: *Advances in Neural Information Processing Systems 32*. Ed. by H. Wallach et al. Curran Associates, Inc., 2019, pp. 8024–8035. URL: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

[PL19] Mary Phuong and Christoph H. Lampert. "Functional vs. parametric equivalence of ReLU networks". In: International Conference on Learning Representations. Sept. 25, 2019. URL: https://openreview.net/forum?id=Bylx-TNKvH (visited on 08/28/2021).

[RC04] Christian P. Robert and George Casella. *Monte Carlo statistical methods*. 2nd ed. Springer texts in statistics. New York: Springer, 2004. 645 pp. ISBN: 9780387212395.

[Res99] Sidney I. Resnick. *A probability path*. Boston: Birkhäuser, 1999. 453 pp. ISBN: 9780817640552.

[Ros62] F. Rosenblatt. *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms*. Cornell Aeronautical Laboratory. Report no. VG-1196-G-8. Spartan Books, 1962. URL: https://books.google.com.au/books?id=7FhRAAAAMAAJ.

[RZL17] Prajit Ramachandran, Barret Zoph, and Quoc V. Le. "Searching for Activation Functions". In: *arXiv:1710.05941 [cs]* (Oct. 27, 2017). arXiv: 1710.05941. URL: http://arxiv.org/abs/1710.05941 (visited on 10/06/2021).

[SS05] Elias M. Stein and Rami Shakarchi. *Real analysis: measure theory, integration, and Hilbert spaces*. Princeton lectures in analysis v. 3. Princeton, N.J: Princeton University Press, 2005. 402 pp. ISBN: 9780691113869.

[SST92] H. S. Seung, H. Sompolinsky, and N. Tishby. "Statistical mechanics of learning from examples". In: *Physical Review A* 45.8 (Apr. 1, 1992), pp. 6056–6091. ISSN: 1050-2947, 1094-1622. DOI: 10.1103/PhysRevA.45.6056. URL: https://link.aps.org/doi/10.1103/PhysRevA.45.6056 (visited on 09/29/2021).

[Sus92] Héctor J. Sussmann. "Uniqueness of the weights for minimal feedforward nets with a given input-output map". In: *Neural Networks* 5.4 (July 1, 1992), pp. 589–593. ISSN: 0893-6080. DOI: 10.1016/S0893-6080(05)80037-1. URL: https://www.sciencedirect.com/science/article/pii/S0893608005800371 (visited on 08/28/2021).

[THK18] Stefan Thurner, R. A. Hanel, and Peter Klimek. *Introduction to the theory of complex systems*. OCLC: on1032587876. Oxford : New York: Oxford University Press, 2018. 431 pp. ISBN: 9780198821939.

[Vaa07] Aad W. van der Vaart. *Asymptotic statistics*. 1. paperback ed., 8. printing. Cambridge series in statistical and probabilistic mathematics. Cambridge: Cambridge Univ. Press, 2007. 443 pp. ISBN: 9780521784504 9780521496032.

[Was21] Michael L. Waskom. "seaborn: statistical data visualization". In: *Journal of Open Source Software* 6.60 (2021), p. 3021. DOI: 10.21105/joss.03021. URL: https://doi.org/10.21105/joss.03021.

[Wat07] Sumio Watanabe. "Almost All Learning Machines are Singular". In: *2007 IEEE Symposium on Foundations of Computational Intelligence*. 2007 IEEE Symposium on Foundations of Computational Intelligence. Apr. 2007, pp. 383–388. DOI: 10.1109/FOCI.2007.371500.

[Wat09] Sumio Watanabe. *Algebraic geometry and statistical learning theory*. OCLC: 521946407. Cambridge; New York: Cambridge University Press, 2009. ISBN: 9780511603457 9780511800474 9780511651533. URL: https://doi.org/10.1017/CBO9780511800474 (visited on 08/06/2021).

[Wat13] Sumio Watanabe. "A Widely Applicable Bayesian Information Criterion". In: *Journal of Machine Learning Research* 14 (Mar. 2013), 867897. ISSN: 1533-7928. URL: https://jmlr.csail.mit.edu/papers/v14/watanabe13a.html (visited on 08/06/2021).

[Wat18] Sumio Watanabe. *Mathematical theory of Bayesian statistics*. Boca Raton: CRC Press, Taylor & Francis Group, 2018. ISBN: 9781482238068.

[Wat20] Sumio Watanabe. "Cross Validation, Information Criterion, and Phase Transitions". 2020. URL: http://watanabe-www.math.dis.titech.ac.jp/users/swatanab/slt202113.pdf.

[Zha+16] Chiyuan Zhang et al. "Understanding deep learning requires rethinking generalization". In: (Nov. 4, 2016). URL: https://openreview.net/forum?id=Sy8gdB9xx (visited on 09/19/2021).

[Zha06] Tong Zhang. "From -Entropy to KL-Entropy: Analysis of Minimum Information Complexity Density Estimation". In: *The Annals of Statistics* 34.5 (2006), pp. 2180–2210. ISSN: 0090-5364. URL: https://www.jstor.org/stable/25463505 (visited on 08/12/2021).