The Multilinear Structure of ReLU Networks

Anonymous authors
Paper under double-blind review

Abstract

We study the loss surface of neural networks that involve only rectified linear unit (ReLU) nonlinearities from a theoretical point-of-view. Any such network defines a piecewise multilinear form in parameter space. As a consequence, optima of such networks generically occur in non-differentiable regions of parameter space and so any understanding of such networks must carefully take into account their non-smooth nature. We then proceed to leverage this multilinear structure in an analysis of a neural network with one hidden-layer. Under the assumption of linearly separable data, the piecewise bilinear structure of the loss allows us to provide an explicit description of all critical points.

1 Introduction

Empirical practice tends to show that modern neural networks have relatively benign loss surfaces, in the sense that training a deep network proves less challenging than the non-convex and non-smooth nature of the optimization would naïvely suggest. Many theoretical efforts have attempted to explain this phenomenon and, more broadly, the successful optimization of deep networks in general (Gori & Tesi (1992); Choromanska et al. (2015); Kawaguchi (2016)). The properties of the loss surface of neural networks remain poorly understood despite these many efforts. Developing of a coherent mathematical understanding of them is therefore one of the major open problems in deep learning.

We focus on investigating the loss surfaces that arise from feed-forward neural networks where ReLUs $\sigma(x) := \max(x, 0) = (x)_+$ account for all nonlinearities present in the network. We allow the transformations defining the hidden-layers of the network to take the form of fully connected affine transformations, convolutional transformations or some other combination of structured affine maps. For the network criterion we elect to use the hinge loss objective

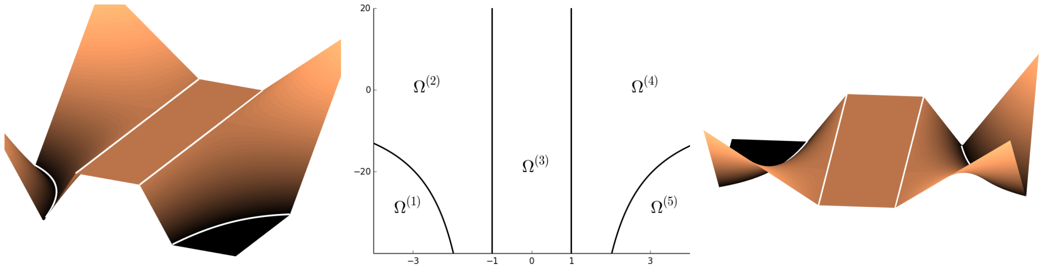
$$\ell(\hat{\mathbf{y}}, r) = \sum_{q=1}^{R} \sigma\big(1 + \hat{y}_q - \hat{y}_r\big)$$

for classification. We use this choice for two reasons. First, the hinge loss is the natural choice if we wish to maintain ReLU nonlinearities throughout the network. As the only nonlinearities from input to loss are ReLUs, each input simply flows through a succession of affine and piecewise linear transformations. This rather homogeneous structure allows us to derive results concerning loss surface of such networks. Second, this choice also allows us to avoid certain pathologies that arise with other objectives; global minimizers generally do not exist, for instance, when using a logistic loss instead of the hinge loss.

To see the type of structure that emerges in these networks, let Ω denote the space of network parameters and let $\mathcal{L}(\boldsymbol{\omega})$ denote the loss. Each nonlinearity involved in the network, including the hinge loss, is either active ($\sigma(x) > 0$) or inactive ($\sigma(x) = 0$) at any point $\boldsymbol{\omega}$ in parameter space. This dichotomy leads to a partition of the parameter space

$$\Omega = \Omega_1 \cup \Omega_2 \cup \ldots \cup \Omega_M \tag{1}$$

into *cells*, where each cell Ω_u corresponds to a given activation pattern of the nonlinearities. Crossing the boundary of a cell Ω_u corresponds to a ReLU switching from active to inactive, or vice-versa. The loss $\mathcal{L}(\boldsymbol{\omega})$ is therefore smooth in the interior of cells and (potentially) non-differentiable on cell boundaries. In this way the decomposition (1) provides a description of the smooth and non-smooth regions of parameter space.



(a) Loss Surface $\mathcal{L}(\boldsymbol{\omega})$ (b) Parameter Space Ω (c) Loss Surface $\mathcal{L}(\boldsymbol{\omega})$

Figure 1: The loss surface corresponding to a piecewise multilinear form. In (a): Local minima are located in the interior of the flat cells $\Omega^{(3)}$ and $\Omega^{(5)}$ (type I), on the boundary between cells $\Omega^{(1)}$ and $\Omega^{(2)}$ (type II) and the boundary between cells $\Omega^{(4)}$ and $\Omega^{(5)}$ (type II). In (b): Parameter space $\Omega = \mathbb{R}^2$ decomposes into a partition of five cells. The loss $\mathcal{L}$ on each cell is a sum of multilinear forms. In (c): A rotation of (a) shows the saddle-like surface of the nontrivial forms on cells $\Omega^{(1)}, \Omega^{(2)}$ and $\Omega^{(4)}$.

We begin by using this decomposition to show that, when restricted to a fixed cell Ω_u, the loss $\mathcal{L}$ is a sum of multilinear forms. Thus the loss $\mathcal{L}(\boldsymbol{\omega})$ **is a piecewise multilinear form**[1], and different multilinear forms characterize the loss on different cells. To see the significance of this structure, recall that a multilinear form is a function $\phi : \mathbb{R}^{d_1} \times \ldots \times \mathbb{R}^{d_n} \to \mathbb{R}$ which is linear with respect to each of its inputs when the other inputs are fixed. That is, each of the n linear relations

$$\phi(\mathbf{v}_1, \ldots, \alpha\mathbf{v}_k + \beta\mathbf{w}_k, \ldots, \mathbf{v}_n) = \alpha\phi(\mathbf{v}_1, \ldots, \mathbf{v}_k, \ldots, \mathbf{v}_n) + \beta\phi(\mathbf{v}_1, \ldots, \mathbf{w}_k, \ldots, \mathbf{v}_n)$$

hold. The functions $\phi(x, y) = xy$ or $\phi(x, y, z) = xyz$ provide canonical examples of multilinear forms, and both of these functions clearly have a saddle like structure. In fact any nontrivial multilinear form has such a saddle-type structure, for the Hessian matrix of a nontrivial multilinear form always has at least one strictly positive and one strictly negative eigenvalue (see appendix for a proof of this statement). Consequently, the graph of a nontrivial multilinear form always has at least one direction of positive curvature and at least one direction of negative curvature. We therefore have the following picture for the loss surface $\mathcal{L}(\boldsymbol{\omega})$ of a piecewise multilinear form: inside each cell Ω_u the loss is either flat, linear or has a saddle-like structure; see figure 1 for a visual example. It is therefore **impossible** for a local minima to occur in the interior of a cell on which the loss has a linear or saddle-like structure, and so neural networks with ReLU nonlinearities have only two types of local minima —

- **Type I (Smooth)**: Those local minima that occur in the interior of a cell with constant loss.
- **Type II (Non-smooth)**: Those local minima that occur on a cell boundary.

We state this result precisely in theorem 1, but figure 1 already shows the presence of these two types of local minima.

This observation has several consequences. A (continuous time) gradient decent algorithm can never reach a type I local minimum. As soon as the algorithm enters a flat cell it must stop since the gradient vanishes at such points. The descent therefore terminates on the boundary of the cell, and so only non-smooth local minima arise when using a local, gradient-based algorithm. Moreover, this structure has potential implications for other various optimization algorithms. An off-the-shelf Newton method, for example, is inappropriate for such networks since the Hessian of $\mathcal{L}$ is never positive definite and typically is indefinite. In other words, any study of algorithms for such a loss must take into account both the nonsmooth structure and the indefinite structure of the loss surface. Local minimizers simply cannot be studied using second-order (i.e. Hessian) information.

We then proceed to leverage this structure in an analysis of a neural network with one hidden-layer. Under the assumption of linearly separable data, the piecewise bilinear structure of the loss allows us to provide an explicit description of critical points. The reasons for this analysis are two-fold. First,

[1] By "piecewise multilinear form" we really mean a *sum of* piecewise multilinear forms.

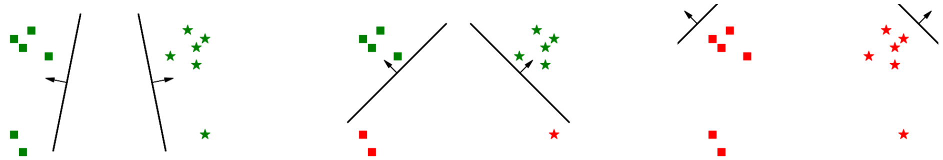

(a) $\mathcal{L}(\boldsymbol{\omega}) = 0$
(b) $\mathcal{L}(\boldsymbol{\omega}) = 3/12$
(c) $\mathcal{L}(\boldsymbol{\omega}) = 12/12$

Figure 2: Three different local minima of the loss $\mathcal{L}(\boldsymbol{\omega})$ for a network with two hidden neurons. Points belonging to class +1 (resp. -1) are denoted by stars (resp. squares). Data points for which the loss is zero (solved points) are colored in green, while data points with non-zero loss (unsolved points) are in red.

it allows us to understand how the addition of depth (i.e. a hidden-layer) affects the loss surface of a simple classification task. With separable data and a purely linear classifier, the corresponding convex optimization problem has only global minimizers with zero loss. Adding a hidden-layer affects this structure and complicates the loss surface, in the sense that non-optimal local minima now occur. Our analysis characterizes this precisely. Secondly, this simple problem serves as a model for the top of a deep network. As we generally expect linearly separable features given enough depth, any non-optimal critical points in a network with one hidden-layer and separable data might manifest in a deep network as well.

To describe the results of this analysis we recall that the hinge loss for binary classification takes the form

$$\ell(\hat{y}, t) = \sigma\big(1 - t\hat{y}\big), \tag{2}$$

where $\hat{y}$ denotes the scalar output of the network and $t \in \{+1, -1\}$ the classification target. For simplicity of exposition, consider the resulting one hidden-layer network without an output bias, i.e.

$$\hat{y} = \sum_{k=1}^{K} v_k \sigma\Big(\langle \mathbf{w}_k, \mathbf{x}\rangle + b_k\Big) \tag{3}$$

with $\langle \cdot, \cdot \rangle$ denoting the Euclidian inner product. The network has K hidden neurons, and each hidden neuron has an associated hyperplane $\langle \mathbf{w}_k, \cdot \rangle + b_k$ as well as a scalar weight v_k used to form the output. Figure 2 shows three different local minima of such a network with two hidden neurons. The first panel, figure 2(a), shows a global minimum where all the data points have zero loss. Figure 2(b) shows a local minimum. All unsolved data points, namely those that contribute a non-zero value to the loss, lie on the “blind side” of the two hyperplanes. For each of these data points the corresponding network output $\hat{y}$ vanishes and so the loss is $\ell(\hat{y}, t) = 1$ for these unsolved points. Small perturbations of the hyperplanes or of the values of the v_k do not change the fact that these data points lie on the blind side of the two hyperplanes. Their loss will not decrease under small perturbations, and so the configuration is, in fact, a local minimum. The same reasoning shows that the configuration in figure 2(c) is also a local minimum. All data points have loss equal to one, and so this local minimum is also a global maximum. Finally, these configurations still define local minimizers if we allow small perturbations to the data points; If the data points represent features of a deep network then such configurations will define local minimizers of a deep network as well.

Despite the presence of sub-optimal local minimizers, the local minima depicted in figure 2 are somehow trivial cases. They simply come from the fact that, due to inactive ReLUs, some data points are completely ignored by the network, and this fact cannot be changed by small perturbations. We show that for binary classification tasks with linearly separable data these are, in fact, the only possible local minima that occur. More precisely, let us say that a hyperplane $\langle \mathbf{w}_k, \cdot \rangle + b_k$ is active if the corresponding v_k is non-zero. Then at any local minima, **a data point with nonzero loss must lie in the blind side of all active hyperplanes**. This result remains true if a bias is added in (3) to the output neuron. For multi-class tasks the answer is more delicate (c.f. section 4), but if we apply an appropriate modifications then the multilinear structure allows us to conclude the analogous result for multi-class partitioning problems as well. In fact, this analysis shows how to reduce the study of *any* multilinear deep network to that of a binary classification problem.

Previous work also address the loss surface of ReLU neural networks, c.f. Safran & Shamir (2016) and Choromanska et al. (2015). The first reference uses ReLU nonlinearities to partition the parameter space into basins that, while similar in spirit, are different from our notion of cells. They estimate the probability of initializing the network in a basin containing a good local minimum under various assumptions on the distribution of data points. However, as noted by the authors, there is no reason to believe that a descent algorithm initialized inside such a basin will actually converge to the local minimum within it. The second reference investigates "randomized" ReLU networks. It provides a description of the quality and asymptotic distribution of the local minima under the assuption that the ReLU activations are independent Bernoulli variables. Similar ideas were pursued in Dauphin et al. (2014) and Kawaguchi (2016). In a different vein, the loss surface of fully connected neural networks with smooth nonlinearities (i.e, sigmoid or tanh) and ℓ^2 loss have also received attention. The dominant strand of this line of work focuses on a search for situations wherein local minima and global minima coincide. For example, if the weight matrices and features at a given layer of the network satisfy certain structural assumptions (e.g. full rank conditions and linear independence) then such a "local equals global" result holds, c.f. Gori & Tesi (1992); Yu & Chen (1995); Frasconi et al. (1997); Nguyen & Hein (2017). Deep linear networks, i.e. a deep network with no nonlinearities, represent the extreme case of this line of work. It is shown in Baldi & Hornik (1989); Baldi & Lu (2012); Kawaguchi (2016) that these networks do not have sub-optimal local minimizers.

2 Piecewise Multilinear Structure

We begin by describing the precise manner in which ReLU networks give rise to piecewise multilinear forms. This will entail both a precise formulation of the decomposition of parameter space into cells as well as an explicit description of the multilinear structure of the loss on each cell. We shall employ the following notation when accomplishing these two tasks. Bold-face Roman and Greek letters such as $\mathbf{x}, \mathbf{y}, \boldsymbol{\lambda}, \boldsymbol{\varepsilon}$ denote vectors in standard Euclidean space, with $\langle \mathbf{x}, \mathbf{y} \rangle := \mathbf{x}^T \mathbf{y}$ the usual inner-product and $\mathbf{x} \otimes \mathbf{y} := \mathbf{x}\mathbf{y}^T$ the standard outer-product of vectors. Their light-face Roman counterparts with sub-scripts $x_i, y_i, \lambda_j, \varepsilon_k$ denote individual entries. Capital Roman letters such as U, V, W will always refer to matrices while the corresponding lower-case letters u_{ij}, v_{ij}, w_{ij} will denote the corresponding matrix entries. We reserve Id for the identity matrix, $\mathbf{0} = (0, \ldots, 0)^t$ for the zero vector and $\mathbf{1} = (1, \ldots, 1)^t$ for the constant vector. We reserve parenthetical super-scripts such as $\mathbf{x}^{(i)}$ or $W^{(\ell)}$ for enumerating a collection of vectors or matrices, respectively. We view a collection of N labelled data points as a set of ordered pairs $(\mathbf{x}^{(i)}, \mathbf{y}^{(i)})$ with the $\mathbf{x}^{(i)}$ representing generic points belonging to $\mathbb{R}^d$ and $\mathbf{y}^{(i)} \in \mathbb{R}^R$ representing one-hot vectors coding for the class of the i^{th} data point. All the proofs are presented in the appendix.

Our analysis considers the following multi-class model with hinge loss. Fix a target $\mathbf{y}$ and let $\hat{\mathbf{y}}$ denote the prediction of the network. Then the expression

$$\ell(\hat{\mathbf{y}}, \mathbf{y}) = -1 + \sum_{q=1}^{R} \sigma\big(1 + \hat{y}_q - \langle \mathbf{y}, \hat{\mathbf{y}} \rangle\big) = -1 + \Big\langle\, \mathbf{1}\,,\ \sigma\Big(\ (\mathrm{Id} - \mathbf{1} \otimes \mathbf{y})\hat{\mathbf{y}} + \mathbf{1}\ \Big)\Big\rangle \tag{4}$$

furnishes the multi-class hinge loss. We consider a neural network with L hidden layers,

$$\begin{aligned} \mathbf{x}^{(i,\ell)} &= \sigma(W^{(\ell)}\mathbf{x}^{(i,\ell-1)} + \mathbf{b}^{(\ell)}) \qquad \text{for} \quad \ell = 1, \ldots, L \\ \hat{\mathbf{y}}^{(i)} &= V\mathbf{x}^{(i,L)} + \mathbf{c} \\ \hat{\mathbf{z}}^{(i)} &= \sigma\Big(\ (\mathrm{Id} - \mathbf{1} \otimes \mathbf{y}^{(i)})\,\hat{\mathbf{y}}^{(i)} \ +\ \mathbf{1}\ \Big), \end{aligned} \tag{5}$$

where $\mathbf{x}^{(i,\ell)}$ denotes the feature vector of the i^{th} data point at the ℓ^{th} layer (with the convention that $\mathbf{x}^{(i,0)} = \mathbf{x}^{(i)}$), $\hat{\mathbf{y}}^{(i)}$ denotes the output of the network for the i^{th} datum and $\hat{\mathbf{z}}^{(i)} \in \mathbb{R}^R$ describes the loss of data point $\mathbf{x}^{(i)}$ associated with each of the R classes. The matrices $W^{(\ell)}$ and vector $\mathbf{b}^{(\ell)}$ define the affine transformation at layer ℓ of the network, and V and $\mathbf{c}$ denote the weights and bias of the output layer. We allow for fully-connected as well as structured models, such as convolutional networks, by imposing the assumption that each $W^{(\ell)}$ is a matrix-valued function that depends *linearly* on some set of parameters $\omega^{(\ell)}$ —

$$W^{(\ell)}\big(\alpha\omega^{(\ell)} + \beta\hat{\omega}^{(\ell)}\big) = \alpha W^{(\ell)}\big(\omega^{(\ell)}\big) + \beta W^{(\ell)}\big(\hat{\omega}^{(\ell)}\big);$$

thus the collection $\boldsymbol{\omega} = (\omega^{(1)}, \ldots, \omega^{(L)}, V, \mathbf{b}^{(1)}, \ldots, \mathbf{b}^{(L)}, \mathbf{c}) \in \Omega$ represent the parameters of the network and Ω denotes parameter space, i.e. a vector space. We let d_ℓ denote the dimension of the features at layer ℓ of the network, with the convention that $d_0 = d$ (dimension of the input data) and $d_{L+1} = R$ (number of classes). We use $D = d_1 + \ldots + d_{L+1}$ for the total number of pointwise nonlinearities and $N_p := \dim(\Omega)$ for the total number of parameters. We then finally arrive at the expression

$$\mathcal{L}(\boldsymbol{\omega}) = \frac{1}{N} \sum_{i=1}^{N} \langle \mathbf{1}, \hat{\mathbf{z}}^{(i)} \rangle - 1$$

for the total loss over all data points and all classes.

2.1 Partitioning Ω into Cells

The lack of differentiability of the nonlinearity $\sigma(x)$ induces a subsequent lack of differentiability of $\mathcal{L}(\boldsymbol{\omega})$, but we may still characterize differentiable regions of the loss precisely. This characterization will prove essential for our analysis. Given a data point $\mathbf{x}^{(i)}$ let us define the collection of functions

$$\begin{aligned} \boldsymbol{\lambda}^{(i,\ell)}(\boldsymbol{\omega}) &:= \sigma'(W^{(\ell)}\mathbf{x}^{(i,\ell-1)} + \mathbf{b}^{\ell}) \qquad \text{for} \quad \ell = 1, \ldots, L \\ \boldsymbol{\varepsilon}^{(i)}(\boldsymbol{\omega}) &:= \sigma'\Big(\big(\mathrm{Id} - \mathbf{1} \otimes \mathbf{y}^{(i)}\big)\, \hat{\mathbf{y}}^{(i)} + \mathbf{1} \Big), \end{aligned} \tag{6}$$

where we make the arbitrary re-definition $\sigma'(0) := 1/2$ to handle those points where $\sigma'(x)$ does not exist. Thus $\boldsymbol{\lambda}^{(i,\ell)} : \Omega \mapsto \{0, 1/2, 1\}^{d_\ell}$ and $\boldsymbol{\varepsilon}^{(i)} : \Omega \mapsto \{0, 1/2, 1\}^{R}$, and by collecting all of these functions into a single *signature function*

$$\mathcal{S}(\boldsymbol{\omega}) = \Big(\boldsymbol{\lambda}^{(1,1)}(\boldsymbol{\omega}), \ldots, \boldsymbol{\lambda}^{(1,L)}(\boldsymbol{\omega}), \boldsymbol{\varepsilon}^{(1)}(\boldsymbol{\omega}); \ldots\ldots; \boldsymbol{\lambda}^{(N,1)}(\boldsymbol{\omega}), \ldots, \boldsymbol{\lambda}^{(N,L)}(\boldsymbol{\omega}), \boldsymbol{\varepsilon}^{(N)}(\boldsymbol{\omega}) \Big)$$

we obtain a function $\mathcal{S} : \Omega \mapsto \{0,\, 1/2,\, 1\}^{ND}$ since there are D total nonlinearities and N total data points. The signature function $\mathcal{S}$ describes how each ReLU in the network activates. These activations take one of three possible states, the fully active state (encoded by a one), the fully inactive state (encoded by a zero), or an in-between state (encoded by a 1/2). If none of the ND entries of $\mathcal{S}(\boldsymbol{\omega})$ equal $1/2$ then all of the ReLUs are differentiable near $\boldsymbol{\omega}$, and so the loss $\mathcal{L}$ is smooth near such points. With this in mind, for a given $u \in \{0,1\}^{ND}$ we define the cell Ω_u as the (possibly empty) set

$$\Omega_u := \mathcal{S}^{-1}(u) := \{\boldsymbol{\omega} \in \Omega : \mathcal{S}(\boldsymbol{\omega}) = u\}$$

of parameter space. By choice $\mathcal{L}$ is smooth on each non-empty cell Ω_u, and so the cells Ω_u provide us with a partition of the parameter space

$$\Omega = \Bigg(\bigcup_{u \in \{0,1\}^{ND}} \Omega_u \Bigg) \bigcup \mathcal{N}.$$

into smooth and potentially non-smooth regions. The set $\mathcal{N}$ contains those $\boldsymbol{\omega}$ for which at least one of the ND entries of $\mathcal{S}(\boldsymbol{\omega})$ takes the value $1/2$, which implies that at least one of the nonlinearities is non-differentiable at such a point. Thus $\mathcal{N}$ consists of points at which the loss is potentially non-differentiable. The following lemma collects the various properties of the cells Ω_u and of $\mathcal{N}$ that we will need in the rest of the paper.

Lemma 1. *Each cell Ω_u for $u \in \{0,1\}^{ND}$ is an open set. If $u \neq u'$ then Ω_u and $\Omega_{u'}$ are disjoint. The set $\mathcal{N}$ is closed and has Lebesgue measure* 0.

2.2 Piecewise Multilinear Structure

Let us briefly assume for the sake of exposition that instead of (5) we have a simplified model without any bias parameters

$$\mathcal{L}(\omega^{(1)}, \ldots, \omega^{(L)}, V) := -1 + \frac{1}{N} \sum_{i} \mathbf{1}^T \sigma(T^{(i)} V \sigma\big(W^{(L)} \cdots \sigma(W^{(2)} \sigma(W^{(1)}\mathbf{x}^{(i)}))) + \mathbf{1}\big)$$

where $T^{(i)} := \mathrm{Id} - \mathbf{1} \otimes \hat{\mathbf{y}}^{(i)}$. By definition, inside a cell Ω_u each nonlinearity $\sigma(x)$ acts as a linear function, i.e. matrix multiplication by a diagonal matrix, or mask, containing zeroes or ones in its diagonal. More precisely, restricted to the cell Ω_u the loss takes the form

$$\mathcal{L}|_{\Omega_u} = -1 + \frac{\mathbf{1}^T \mathcal{E}^{(i,u)} \mathbf{1}}{N} + \frac{1}{N} \sum_i \mathbf{1}^T \mathcal{E}^{(i,u)} T^{(i)} V \Lambda^{(i,L,u)} W^{(L)} \cdots \Lambda^{(i,2,u)} W^{(2)} \Lambda^{(i,1,u)} W^{(1)} \mathbf{x}^{(i)}$$

$$\mathcal{E}^{(i,u)} := \mathrm{diag}(\boldsymbol{\varepsilon}^{(i,u)}), \;\; \Lambda^{(i,\ell,u)} := \mathrm{diag}(\boldsymbol{\lambda}^{(i,\ell,u)});$$

the "u" super-scripts in $\boldsymbol{\varepsilon}^{(i,u)}, \boldsymbol{\lambda}^{(i,\ell,u)}$ indicate the dependence of the masks on the cell. Up to the constant factor $-1 + \langle \mathbf{1}, \mathcal{E}^{(i,u)} \mathbf{1} \rangle / N$ that simply counts the average number of errors on the cell, it is then clear that $\mathcal{L}|_{\Omega_u}$ is a multilinear form of its arguments. As a consequence, the loss is a piecewise multilinear form up to constants and it is therefore smooth in the interior of each cell. As non-zero multilinear forms do not have local minima, it is clear that a local minimum of the loss can only occur (i) in the interior of cells where the loss is constant, or (ii) on the boundary of one or more cells. Going back to our case of interest (5), the presense of bias parameters complicates the picture slightly — the loss on a cell is now a sum of multilinear form rather than a single multilinear form. However, the overall conclusions regarding local minima remain unchanged. The following theorem describes the precise result.

Theorem 1 (Structure of the loss)**.**

(i) *For each cell Ω_u there exist multilinear forms $\phi^u_0, \phi^u_1, \ldots, \phi^u_L$, a linear function ϕ^u_{L+1} and a constant ϕ^u_{L+2} such that*

$$\begin{aligned} \mathcal{L}|_{\Omega_u}(\omega^{(1)}, \ldots, \omega^{(L)}, V, \mathbf{b}^{(1)}, \ldots, \mathbf{b}^{(L)}, \mathbf{c}) = \phi^u_0(\omega^{(1)}, \omega^{(2)}, \omega^{(3)}, \omega^{(4)} \ldots, \omega^{(L)}, V) \\ + \phi^u_1(\mathbf{b}^{(1)}, \omega^{(2)}, \omega^{(3)}, \omega^{(4)} \ldots, \omega^{(L)}, V) \\ + \phi^u_2(\mathbf{b}^{(2)}, \omega^{(3)}, \omega^{(4)} \ldots, \omega^{(L)}, V) \\ + \phi^u_3(\mathbf{b}^{(3)}, \omega^{(4)} \ldots, \omega^{(L)}, V) \\ \vdots \\ + \phi^u_{L-1}(\mathbf{b}^{(L-1)}, \omega^{(L)}, V) \\ + \phi^u_L(\mathbf{b}^{(L)}, V) \\ + \phi^u_{L+1}(\mathbf{c}) \\ + \phi^u_{L+2}. \end{aligned}$$

The constant ϕ^u_{L+2} counts the average number of errors on the cell.

(ii) *The loss $\mathcal{L}$ is smooth on each cell Ω_u. Moreover, if $\boldsymbol{\omega} \in \Omega \setminus \mathcal{N}$ and the Hessian matrix $H\mathcal{L}(\omega)$ does not vanish then it must have at least one strictly positive and one strictly negative eigenvalue.*

(iii) *Local minima and maxima of $\mathcal{L}$ occur only on cell boundaries (i.e. on $\mathcal{N}$) or on those cells Ω_u where the loss is constant. In the latter case, $\mathcal{L}|_{\Omega_u}(\boldsymbol{\omega}) = \phi^u_{L+2}$ for all $\boldsymbol{\omega} \in \Omega_u$.*

3 CRITICAL POINT ANALYSIS

Recall the hinge loss

$$\ell(\hat{y}, y) := \sigma\big(1 - y\hat{y}\big)$$

for binary classification, where $y \in \{+1, -1\}$ denotes the target. For a given set of parameters $\boldsymbol{\omega} = (W, \mathbf{v}, \mathbf{b}, c)$ the expression

$$\mathcal{L}(W, \mathbf{v}, \mathbf{b}, c) = \frac{1}{N} \sum_i \sigma \left[1 - y^{(i)} \left\{ \mathbf{v}^T \sigma(W\mathbf{x}^{(i)} + \mathbf{b}) + c \right\} \right] \qquad (7)$$

then defines the loss associated to a fully connected network with one hidden layer. Let $\{\mathbf{w}_k\}_{1 \le k \le K}$ denote the rows of the linear transformation W defining the hidden layer. A straightforward computation shows that we may specify the multilinear forms in theorem 1 more precisely.

Lemma 2 (Decomposition with $L = 1$)**.** *Let*

$$\mathcal{L}|_{\Omega_u}(W, \mathbf{v}, \mathbf{b}, c) = \phi_0^u(W, \mathbf{v}) + \phi_1^u(\mathbf{b}, \mathbf{v}) + \phi_2^u(c) + \phi_3^u \tag{8}$$

denote the loss on a cell Ω_u. For $1 \leq k \leq K$ define

$$\mathbf{a}_k^{(u)} := \frac{1}{N}\left(\sum_{i:y^{(i)}=1} \varepsilon^{(i,u)}\lambda_k^{(i,u)}\mathbf{x}^{(i)} - \sum_{i:y^{(i)}=-1} \varepsilon^{(i,u)}\lambda_k^{(i,u)}\mathbf{x}^{(i)} \right) \tag{9}$$

$$\alpha_k^{(u)} := \frac{1}{N}\left(\sum_{i:y^{(i)}=1} \varepsilon^{(i,u)}\lambda_k^{(i,u)} - \sum_{i:y^{(i)}=-1} \varepsilon^{(i,u)}\lambda_k^{(i,u)} \right) \tag{10}$$

$$\gamma^{(u)} := \frac{1}{N}\left(\sum_{i:y^{(i)}=1} \varepsilon^{(i,u)} - \sum_{i:y^{(i)}=-1} \varepsilon^{(i,u)} \right) \qquad \delta^u := \frac{1}{N}\sum_i \varepsilon^{(i,u)}. \tag{11}$$

Then $\phi_3^u = \delta^u$ and $\phi_2^u(c) = -\gamma^{(u)}c$, while the relations

$$\phi_0^u(W, \mathbf{v}) = -\sum_k v_k \langle \mathbf{a}_k^{(u)}, \mathbf{w}_k \rangle, \quad \textit{and} \quad \phi_1^u(\mathbf{b}, \mathbf{v}) = -\sum_k v_k \alpha_k^{(u)} b_k, \tag{12}$$

furnish the multilinear forms defining the loss on Ω_u.

With this description in hand, we may now explore the consequences of this decomposition under the assumption of linearly separable data. Since the data are linearly separable there exists a unit vector $\mathbf{q} \in \mathbb{R}^d$, a bias $\beta \in \mathbb{R}$ and a margin $\mu > 0$ such that the family of inequalities

$$\langle \mathbf{q}, \mathbf{x}^{(i)} \rangle + \beta \geq \mu \qquad \text{if} \quad y^{(i)} = +1 \tag{13}$$

$$\langle \mathbf{q}, \mathbf{x}^{(i)} \rangle + \beta \leq -\mu \qquad \text{if} \quad y^{(i)} = -1 \tag{14}$$

hold. By combining (9)–(10) with (13)–(14) we easily obtain the following estimate

$$\langle \mathbf{a}_k^{(u)}, \mathbf{q} \rangle + \alpha_k^{(u)}\beta \geq \mu \left(\frac{1}{N}\sum_i \varepsilon^{(i,u)}\lambda_k^{(i,u)} \right), \tag{15}$$

which we may then use to find a decent direction for the loss whenever the right-hand-side of (15) does not vanish. The idea is simple, i.e. that adding a multiple of $\pm\mathbf{q}$ to $\mathbf{w}_k$ and a multiple of $\pm\beta$ to b_k will usually lead to a decrease of the loss. To see this let $\mathbf{e}_k = (0, \ldots, 1, \ldots, 0)^T$ denote the k^{th} standard basis vector, $\overline{\Omega_u}$ the closure of the cell Ω_u and $\text{sign}(x)$ the signum function that vanishes at zero. The following lemma then makes this idea precise.

Lemma 3. *Let $\boldsymbol{\omega} = (W, \mathbf{v}, \mathbf{b}, c) \in \Omega$ denote any point. Define*

$$\tilde{W} = \text{ sign}(v_k)\, \mathbf{e}_k \otimes \mathbf{q} \qquad \textit{and} \qquad \tilde{\mathbf{b}} = \beta\, \text{sign}(v_k)\, \mathbf{e}_k.$$

For $t \in \mathbb{R}$ let $\boldsymbol{\omega}(t) := (W + t\tilde{W}, \mathbf{v}, \mathbf{b} + t\tilde{\mathbf{b}}, c)$ denote a perturbation of $\boldsymbol{\omega}$. Then

(i) *There exists $t_0 > 0$ and $u \in \{0,1\}^{ND}$ such that $\boldsymbol{\omega}(t) \in \overline{\Omega_u}$ for all $t \in [0, t_0)$.*

(ii) $\mathcal{L}(\boldsymbol{\omega}) \geq \mathcal{L}(\boldsymbol{\omega}(t)) + t|v_k|\frac{\mu}{N}\sum_i \varepsilon^{(i,u)}\lambda_k^{(i,u)}$ *for all $t \in [0, t_0)$.*

We may now state and prove the theorem underyling figure 2 in full generality. If we let $\ell^{(i)}(\boldsymbol{\omega})$ denote the contribution of the i^{th} data point $\mathbf{x}^{(i)}$ to the total loss, so that $\mathcal{L}(\boldsymbol{\omega}) = \frac{1}{N}\sum_i \ell^{(i)}(\boldsymbol{\omega})$, then we may conclude

Theorem 2. *Let $\boldsymbol{\omega} = (W, \mathbf{v}, \mathbf{b}, c)$ be a local minimum of the loss and assume the data $\{\mathbf{x}^{(i)}\}_{i=1}^N$ are linearly separable. Then*

$$\ell^{(i)}(\boldsymbol{\omega}) > 0 \qquad \Longrightarrow \qquad v_k\, \sigma(\langle \mathbf{w}_k, \mathbf{x}^{(i)} \rangle + b_k) = 0 \quad \textit{for all } k \in \{1, \ldots, K\}.$$

Essentially, this theorem says that local minima obey the property sketched in figure 2. If a data point $\mathbf{x}^{(i)}$ has non-zero loss one of v_k or $\sigma(\langle \mathbf{w}_k, \mathbf{x}^{(i)}\rangle + b_k)$ must vanish for *all* hidden neurons. We therefore have a dichotomy. Either $\mathbf{x}^{(i)}$ lies in the blind side of the hyperplane $\langle \mathbf{w}_k, \mathbf{x}^{(i)}\rangle + b_k$ or else $v_k = 0$. In the latter case the k^{th} feature is not used when forming network predictions and so the corresponding hyperplane is inactive. Succinctly, theorem 2 states that **if a data point $\mathbf{x}^{(i)}$ is unsolved it must lie on the blind side of every active hyperplane.** Moreover, this result applies to both critical points as well as to local minimizers. While the proof of theorem 2 only yields the result for minimizers, it has the benefits of both transparency and directness – we invite the reader to read the proof of Lemma 3(ii) and theorem 2 in the appendix, which are particularly simple.

Extending the result of theorem 2 to include critical points is less direct, and it requires an invocation of machinery from non-smooth analysis. To begin, we recall that for a Lipschitz but non-differentiable function $f(\boldsymbol{\omega})$ the *Clarke subdifferential* $\partial_o f(\boldsymbol{\omega})$ of f at a point $\boldsymbol{\omega} \in \Omega$ provides a generalization of both the gradient $\nabla f(\boldsymbol{\omega})$ and the usual subdifferential of a convex function. For a Lipschitz function f we may employ the following definition (c.f. page 133 of Borwein & Lewis (2010)).

Definition 1 (Clarke Subdifferential)**.** *Suppose that a function $f : \Omega \mapsto \mathbb{R}$ is locally Lipschitz around $\boldsymbol{\omega} \in \Omega$, and differentiable on $\Omega \setminus \mathcal{M}$ where $\mathcal{M}$ is a set of Lebesgue measure zero. Then the convex hull*

$$\partial_o f(\boldsymbol{\omega}) := \text{c.h.} \left\{ \lim_k \nabla f(\boldsymbol{\omega}_k) : \boldsymbol{\omega}_k \to \boldsymbol{\omega}, \boldsymbol{\omega}_k \notin \mathcal{M} \right\}$$

is the Clarke subdifferential of f at $\boldsymbol{\omega}$.

With this definition in hand, we can now state the following stronger version of theorem 2:

Theorem 3. *Let $\boldsymbol{\omega} = (W, \mathbf{v}, \mathbf{b}, c)$ and assume that $0 \in \partial_o \mathcal{L}(\boldsymbol{\omega})$. Assume also that the data $\{\mathbf{x}^{(i)}\}_{i=1}^N$ are linearly separable. Then*

$$\ell^{(i)}(\boldsymbol{\omega}) > 0 \qquad \Longrightarrow \qquad v_k\, \sigma(\langle \mathbf{w}_k, \mathbf{x}^{(i)}\rangle + b_k) = 0 \quad \textit{for all } k \in \{1, \ldots, K\}.$$

4 Exact Penalties and Multi-Class Structure

Unfortunately, the result of theorem 2 and 3 does not extend naïvely to the multi-class case. To the contrary, counter-examples show that there exist non-trivial critical points for linearly separable data whenever the number of classes exceeds two. In other words, in the presence of three or more classes a critical point may contain active yet unsolved data points. This begs the question of whether some variant of theorem 3 holds in the multi-class context. A first attempt might simply modify the loss itself. We might hope that substituting the multi-class hinge loss (4) with its one-versus-all variant

$$\bar{\ell}(\hat{\mathbf{y}}, \mathbf{y}) := \sum_{r=1}^{R} \sigma\left(1 + \hat{y}_r^{(i)}(-1)^{y_r^{(i)}}\right) \qquad \bar{\mathcal{L}}(\boldsymbol{\omega}) = \frac{1}{N}\sum_{i=1}^{N} \bar{\ell}(\hat{\mathbf{y}}^{(i)}, \mathbf{y}^{(i)}) \tag{16}$$

would restore the two-class structure of critical points. The idea here is to use the binary hinge loss $\bar{\ell}$ in the hope of decoupling a multi-class problem into R two-class problems. Yet similar counter-examples dash this hope as well, as non-trivial critical points persist for network with modified loss (16) and one hidden layer. The inherent difficulty comes from the fact that all of the parameters $\boldsymbol{\omega}$ in the network still couple through the joint nonlinear minimization of (16), and so simply modifying the hinge loss does not restore the two-class structure.

We may, however, introduce a sufficient amount of decoupling if we modify both the *loss* as well as the *algorithm* used in its optimization. Let us begin this process by recalling that

$$\mathbf{x}^{(i,L)}\big(\omega^{(1)}, \ldots, \omega^{(L)}, \mathbf{b}^{(1)}, \ldots, \mathbf{b}^{(L)}\big) \quad \text{and} \quad \hat{\mathbf{y}}^{(i)} = V\mathbf{x}^{(i,L)} + \mathbf{c}$$

denote the features and predictions of the network with L hidden layers, respectively. The sub-collection of parameters

$$\breve{\boldsymbol{\omega}} := \big(\omega^{(1)}, \ldots, \omega^{(L)}, \mathbf{b}^{(1)}, \ldots, \mathbf{b}^{(L)}\big)$$

therefore determine a common set of features $\mathbf{x}^{(i,L)}$ while the parameters $V, \mathbf{c}$ determine R one-versus-all classifiers utilizing these features. We may write the loss for the r^{th} class as

$$\mathcal{L}^{(r)}(\breve{\boldsymbol{\omega}}, \mathbf{v}_r, c_r) = \frac{1}{N}\sum_{i=1}^{N} \sigma\left(1 + \hat{y}_r^{(i)}(-1)^{y_r^{(i)}}\right)$$

and then sum $\bar{\mathcal{L}}(\boldsymbol{\omega}) := (\mathcal{L}^{(1)} + \cdots + \mathcal{L}^{(R)})(\boldsymbol{\omega})$ to recover the total objective. Thus each classifier in $\bar{\mathcal{L}}$ shares a common set of features, and the joint minimization over features and classifiers couples the R binary problems together.

We may then seek to minimize $\bar{\mathcal{L}}$ by applying a soft-penalty approach. We introduce the R replicates

$$\breve{\boldsymbol{\omega}}^{(r)} = \big(\omega^{(1,r)}, \ldots, \omega^{(L,r)}, \mathbf{b}^{(1,r)}, \ldots, \mathbf{b}^{(L,r)}\big) \quad 1 \le r \le R$$

of the hidden-layer parameters $\breve{\boldsymbol{\omega}}$ and include a soft ℓ^2-penalty

$$\mathcal{R}\big(\breve{\boldsymbol{\omega}}^{(1)}, \ldots, \breve{\boldsymbol{\omega}}^{(R)}\big) := \frac{R}{R-1} \sum_{\ell=1}^{L} \sum_{r=1}^{R} \|\omega^{(\ell,r)} - \bar{\omega}^{(\ell)}\|^2 + \|\mathbf{b}^{(\ell,r)} - \bar{\mathbf{b}}^{(\ell)}\|^2$$

to enforce that the replicated parameters $\omega^{(\ell,r)}, \mathbf{b}^{(\ell,r)}$ remain close to their corresponding means $(\bar{\omega}^{(\ell)}, \bar{\mathbf{b}}^{(\ell)})$ across classes. We then proceed by minimizing the penalized loss

$$\mathcal{E}\big(\boldsymbol{\omega}^{(1)}, \ldots, \boldsymbol{\omega}^{(R)}\big) := \sum_{r=1}^{R} \mathcal{L}^{(r)}\big(\boldsymbol{\omega}^{(r)}\big) + \gamma \mathcal{R}\big(\breve{\boldsymbol{\omega}}^{(1)}, \ldots, \breve{\boldsymbol{\omega}}^{(R)}\big) \tag{17}$$

for $\gamma > 0$ some parameter controlling the strength of the penalty. Remarkably, performing this process yields

Theorem 4 (Exact Penalty and Recovery of Two-Class Structure)**.** *If $\gamma > 0$ then the following hold for* (17) —

(i) *The penalty is exact, that is, at **any** critical point $\big(\boldsymbol{\omega}^{(1)}, \ldots, \boldsymbol{\omega}^{(R)}\big)$ of $\mathcal{E}$ the equalities*

$$\omega^{(\ell,1)} = \cdots = \omega^{(\ell,R)} = \bar{\omega}^{(\ell)} := \frac{1}{R} \sum_{r=1}^{R} \omega^{(\ell,r)}$$

$$\mathbf{b}^{(\ell,1)} = \cdots = \mathbf{b}^{(\ell,R)} = \bar{\mathbf{b}}^{(\ell)} := \frac{1}{R} \sum_{r=1}^{R} \mathbf{b}^{(\ell,r)}$$

hold for all $1 \le \ell \le L$.

(ii) *At **any** critical point of $\mathcal{E}$ the two-class critical point relations*

$$\mathbf{0} \in \partial_0 \mathcal{L}^{(r)}(\breve{\boldsymbol{\omega}}, \mathbf{v}_r, c_r)$$

hold for all $1 \le r \le R$.

In other words, applying a soft-penalty approach to minimizing the coupled problem $\bar{\mathcal{L}}$ actually yields an exact penalty method. By (i), at critical points we obtain a common set of features $\mathbf{x}^{(i,L)}$ for each of the R binary classification problems. Moreover, by (ii) these features simultaneously yield critical points

$$\mathbf{0} \in \partial_0 \mathcal{L}^{(r)}\big(\breve{\boldsymbol{\omega}}, \mathbf{v}_r, c_r\big) \tag{18}$$

for *all* of these binary classification problems. If (18) holds then clearly the weaker critical point relation $\mathbf{0} \in \partial_0 \bar{\mathcal{L}}(\boldsymbol{\omega})$ for the full loss holds as well, and so the penalty approach certainly yields critical points of the original loss. More importantly, the fact that (18) may fail for critical points of $\bar{\mathcal{L}}$ is responsible for the presence of non-trivial critical points in the context of a network with one hidden layer. We may therefore interpret (ii) as saying that the penalty avoids pathological critical points where $\mathbf{0} \in \partial_0 \bar{\mathcal{L}}(\boldsymbol{\omega})$ but (18) does not. To be clear, we may say

Corollary 1. *Assume the hypotheses of theorem 3, and that the $\{\mathbf{x}^{(i)}\}$ are linearly separable. Let $\boldsymbol{\omega}$ denote any critical point of $\mathcal{E}$ and $\ell^{(i,r)}(\boldsymbol{\omega})$ the loss associated to the i^{th} data point and the r^{th} class. Then*

$$\ell^{(i,r)}(\boldsymbol{\omega}) > 0 \quad \Longrightarrow \quad (\mathbf{v}_r)_k\, \sigma(\langle \mathbf{w}_k, \mathbf{x}^{(i)} \rangle + b_k) = 0 \quad \textit{for all } k \in \{1, \ldots, K\}.$$

The corollary follows immediately from (18) and the argument for the two class case. In principle, the penalty approach also provides a path forward for studying multi-class problems. Regardless of the number L of hidden layers, an understanding of the family of critical points (18) reduces to a study of critical points of binary classification problems.

REFERENCES

Pierre Baldi and Kurt Hornik. Neural networks and principal component analysis: Learning from examples without local minima. *Neural networks*, 2(1):53–58, 1989.

Pierre Baldi and Zhiqin Lu. Complex-valued autoencoders. *Neural Networks*, 33:136–147, 2012.

Jonathan Borwein and Adrian S Lewis. *Convex analysis and nonlinear optimization: theory and examples. Second Edition*. Springer Science & Business Media, 2010.

Anna Choromanska, Mikael Henaff, Michael Mathieu, Gérard Ben Arous, and Yann LeCun. The loss surfaces of multilayer networks. In *Artificial Intelligence and Statistics*, pp. 192–204, 2015.

Yann N Dauphin, Razvan Pascanu, Caglar Gulcehre, Kyunghyun Cho, Surya Ganguli, and Yoshua Bengio. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. In *Advances in neural information processing systems*, pp. 2933–2941, 2014.

P Frasconi, M Gori, and A Tesi. Successes and failures of backpropagation: A theoretical. *Progress in Neural Networks: Architecture*, 5:205, 1997.

Marco Gori and Alberto Tesi. On the problem of local minima in backpropagation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(1):76–86, 1992.

Kenji Kawaguchi. Deep learning without poor local minima. In *Advances in Neural Information Processing Systems*, pp. 586–594, 2016.

Quynh Nguyen and Matthias Hein. The loss surface of deep and wide neural networks. In *International Conference on Machine Learning*, pp. 2603–2612, 2017.

Itay Safran and Ohad Shamir. On the quality of the initial basin in overspecified neural networks. In *International Conference on Machine Learning*, pp. 774–782, 2016.

Xiao-Hu Yu and Guo-An Chen. On the local minima free condition of backpropagation learning. *IEEE Transactions on Neural Networks*, 6(5):1300–1303, 1995.

Appendix: Proofs of Lemmas and Theorems

Multilinear forms have a saddle like structure

Lemma. *The Hessian matrix of a non-trivial multilinear form has at least one strictly positive and one strictly negative eigenvalue.*

Proof. A multilinear form $\phi : \mathbb{R}^{d_1} \times \ldots \times \mathbb{R}^{d_n} \to \mathbb{R}$ can always be written as

$$\phi(\mathbf{v}_1, \ldots, \mathbf{v}_n) = \sum_{j_1=1}^{d_1} \ldots \sum_{j_n=1}^{d_n} A_{j_1,\ldots,j_n} v_{1,j_1} \ldots v_{n,j_n} \tag{19}$$

for some tensor $\{A_{j_1,\ldots,j_n} : 1 \leq j_k \leq d_k\}$. Here $v_{k,j}$ denotes the j^{th} component of the vector $\mathbf{v}_k$. From (19) it is clear that $\frac{\partial^2 \phi}{\partial v_{k,j}^2} = 0$ and therefore the trace of the Hessian matrix of ϕ is equal to zero. This implies that the sum of the eigenvalues of the Hessian is equal to zero. So if the Hessian is not the zero matrix, then it has at least one strictly positive and one strictly negative eigenvalue. □

Proof of Lemma 1

The features $\mathbf{x}^{(i,\ell)}$ at each hidden layer depend in a Lipschitz fashion on parameters. Thus each Ω_u defines an open set in parameter space. Moreover, if $u \neq \tilde{u}$ then Ω_u and $\Omega_{\tilde{u}}$ are disjoint by definition. If $\boldsymbol{\omega} \notin \Omega_u$ for all $u \in \{0,1\}^{ND}$ then at least one of the equalities

$$\mathbf{b}_j^{(\ell)} = -\langle \mathbf{w}_j^{(\ell)}, \mathbf{x}^{(i,\ell-1)} \rangle \quad \text{or} \quad \mathbf{c}_s = -\big(1 + \langle \mathbf{v}_s - \mathbf{v}_r, \mathbf{x}^{(i,L)} \rangle\big), \tag{20}$$

must hold. The set of parameters $\mathcal{N} \subset \Omega$ where an equality of the form (20) holds corresponds to a Lipschitz graph in Ω of the bias parameter for that equality. We may therefore conclude that

$$\mathcal{N} := \Omega \setminus \left(\bigcup_{u \in \{0,1\}^{ND}} \Omega_u \right)$$

defines a set contained in a finite union of Lipschitz graphs. Thus $\mathcal{N}$ is $(N_p - 1)$-rectifiable, and in particular, has Lebesgue measure zero. That $\mathcal{N}$ is closed follows from the fact that it is the complement of an open set.

Proof of Theorem 1

Part (i) follows by carefully expanding

$$\mathcal{L} = -1 + \frac{1}{N} \sum_i \mathbf{1}^T \sigma \Big(T^{(i)}(V\sigma(W_L\sigma(\ldots W_2(W_1\mathbf{x}^{(i)} + \mathbf{b}_1) + \mathbf{b}_2 \ldots) + \mathbf{b}_L) + \mathbf{c}) + \mathbf{1} \Big)$$

where $T^{(i)} := \mathrm{Id} - \mathbf{1} \otimes \hat{\mathbf{y}}^{(i)}$. Part (ii) comes from the fact that the trace of the Hessian of a multilinear form is equal to zero.

To prove part (iii), note part (ii) implies that for any $\hat{\boldsymbol{\omega}} \in \Omega_u$ there exists a small neighborhood

$$B_\varepsilon(\hat{\boldsymbol{\omega}}) := \{\boldsymbol{\omega} \in \Omega : \|\boldsymbol{\omega} - \hat{\boldsymbol{\omega}}\| < \varepsilon\} \subset \Omega_u$$

on which $\mathcal{L}|_{\Omega_u}(\boldsymbol{\omega})$ is constant. Thus

$$\mathcal{L}|_{\Omega_u}(\hat{\boldsymbol{\omega}} + \delta\boldsymbol{\omega}) = \mathcal{L}|_{\Omega_u}(\hat{\boldsymbol{\omega}})$$

must hold for all $\boldsymbol{\omega}$ and all δ small enough. Now use part (i) and multilinearity to expand the left-hand-side into powers of δ:

$$\mathcal{L}|_{\Omega_u}(\hat{\boldsymbol{\omega}} + \delta\boldsymbol{\omega}) = \mathcal{L}|_{\Omega_u}(\hat{\boldsymbol{\omega}}) + \sum_{k=1}^{L} \delta^k f_k(\boldsymbol{\omega}) + \delta^{L+1}\big(\phi_0^u + \phi_1^u\big)(\boldsymbol{\omega}). \tag{21}$$

That $\big(\phi_0^u + \phi_1^u\big)(\boldsymbol{\omega})$ is, in fact, the highest-order term is a consequence of the multilinear decomposition from part (i). Since (21) must hold for all δ small enough, all like powers must vanish

$$f_k(\boldsymbol{\omega}) = 0 \qquad \text{and} \qquad \big(\phi_0^u + \phi_1^u\big)(\boldsymbol{\omega}) = 0.$$

Now take any $\boldsymbol{\omega}$ with $\mathbf{b}^{(1)} = 0$ to conclude

$$\phi_0^u(\omega^{(1)}, \omega^{(2)}, \omega^{(3)}, \omega^{(4)} \ldots, \omega^{(L)}, V) = 0$$

for all $\omega^{(\ell)}, V$. But then $\big(\phi_0^u + \phi_1^u\big)(\boldsymbol{\omega}) = 0$ for all $\boldsymbol{\omega}$ implies

$$\phi_1^u(\mathbf{b}^{(1)}, \omega^{(2)}, \omega^{(3)}, \omega^{(4)} \ldots, \omega^{(L)}, V)$$

for all $\omega^{(\ell)}, V, \mathbf{b}^{(1)}$ as well. Thus $\phi_0^u + \phi_1^u$ is the zero function, and so ϕ_2^u is the highest-order multilinear form in the decomposition from part (i). This implies that

$$f_L(\boldsymbol{\omega}) = \phi_2^u(\mathbf{b}^2, \omega^{(3)}, \ldots, \omega^{(L)}, V),$$

but f_L must vanish by (21). Thus ϕ_2^u is the zero function as well. Continuing in this way shows that each ϕ_ℓ^u is the zero function for $0 \leq \ell \leq L+1$, and so in fact

$$\mathcal{L}|_{\Omega_u}(\boldsymbol{\omega}) = \phi_{L+2}^u$$

as claimed.

Proof of Lemma 2

Restricted to a cell Ω_u, the loss can be written

$$\begin{aligned}
\mathcal{L}|_{\Omega_u}(W, \mathbf{v}, \mathbf{b}, c) &= \frac{1}{N}\sum_i \sigma\left[\, -y^{(i)}\left\{\mathbf{v}^T\sigma(W\mathbf{x}^{(i)} + \mathbf{b}) + c\right\} + 1\right] \\
&= \frac{1}{N}\sum_i \varepsilon^{(i,u)}\left[\, -y^{(i)}\left\{\mathbf{v}^T\Lambda^{(i,u)}(W\mathbf{x}^{(i)} + \mathbf{b}) + c\right\} + 1\right]
\end{aligned}$$

Expanding parenthesis after parenthesis the above formula leads to:

$$\begin{aligned}
\mathcal{L}|\Omega_u(W, \mathbf{v}, \mathbf{b}, c) &= \frac{1}{N}\sum_i \varepsilon^{(i,u)}\left[\, -y^{(i)}\left\{\mathbf{v}^T\Lambda^{(i,u)}W\mathbf{x}^{(i)} + \mathbf{v}^T\Lambda^{(i,u)}\mathbf{b} + c\right\} + 1\right] \\
&= \frac{1}{N}\sum_i \varepsilon^{(i,u)}\left[\, -y^{(i)}\mathbf{v}^T\Lambda^{(i,u)}W\mathbf{x}^{(i)} - y^{(i)}\mathbf{v}^T\Lambda^{(i,u)}\mathbf{b} - y^{(i)}c + 1\right] \\
&= \frac{1}{N}\sum_i -\varepsilon^{(i,u)}y^{(i)}\mathbf{v}^T\Lambda^{(i,u)}W\mathbf{x}^{(i)} - \varepsilon^{(i,u)}y^{(i)}\mathbf{v}^T\Lambda^{(i,u)}\mathbf{b} - \varepsilon^{(i,u)}y^{(i)}c + \varepsilon^{(i,u)} \\
&= \phi_0^{(u)}(W, \mathbf{v}) + \phi_1^{(u)}(\mathbf{b}, \mathbf{v}) + \phi_2^{(u)}(c) + \phi_3^{(u)}
\end{aligned}$$

Letting $\mathbf{w}_k$ be the k^{th} row of the matrix W, and noting that $\mathbf{v}^T\Lambda^{(i,u)}W = \sum_k v_k\lambda_k^{(i,u)}\mathbf{w}_k^T$ we find that

$$\begin{aligned}
\phi_0^{(u)}(W, \mathbf{v}) &= -\frac{1}{N}\sum_i \varepsilon^{(i,u)}y^{(i)}\left(\sum_k v_k\lambda_k^{(i,u)}\mathbf{w}_k^T\right)\mathbf{x}^{(i)} \\
&= -\frac{1}{N}\sum_i\sum_k \varepsilon^{(i,u)}y^{(i)}v_k\lambda_k^{(i,u)}\langle\mathbf{w}_k, \mathbf{x}^{(i)}\rangle \\
&= -\sum_k v_k\left\langle \frac{1}{N}\sum_i \varepsilon^{(i,u)}y^{(i)}\lambda_k^{(i,u)}\mathbf{x}^{(i)}, \mathbf{w}_k\right\rangle \\
&= -\sum_k v_k\langle\mathbf{a}_k^{(u)}, \mathbf{w}_k\rangle
\end{aligned}$$

where the vector $\mathbf{a}_k^{(u)}$ is defined by

$$\begin{aligned}
\mathbf{a}_k^{(u)} &= \frac{1}{N}\sum_i y^{(i)}\varepsilon^{(i,u)}\lambda_k^{(i,u)}\mathbf{x}^{(i)} \\
&= \frac{1}{N}\left(\sum_{i:y^{(i)}=1}\varepsilon^{(i,u)}\lambda_k^{(i,u)}\mathbf{x}^{(i)} - \sum_{i:y^{(i)}=-1}\varepsilon^{(i,u)}\lambda_k^{(i,u)}\mathbf{x}^{(i)}\right)
\end{aligned}$$

Similarly we find that

$$\begin{aligned}
\phi_1^{(u)}(\mathbf{b},\mathbf{v}) &= -\frac{1}{N}\sum_i \varepsilon^{(i,u)}y^{(i)}\mathbf{v}^T\Lambda^{(i,u)}\mathbf{b} \\
&= -\frac{1}{N}\sum_i \varepsilon^{(i,u)}y^{(i)}\sum_k\left(v_k\lambda_k^{(i,u)}b_k\right) \\
&= -\sum_k v_k\left(\frac{1}{N}\sum_i \varepsilon^{(i,u)}y^{(i)}\lambda_k^{(i,u)}\right)b_k \\
&= -\sum_k v_k\alpha_k^u b_k
\end{aligned}$$

where

$$\alpha_k^{(u)} = \frac{1}{N}\left(\sum_{i:y^{(i)}=1}\varepsilon^{(i,u)}\lambda_k^{(i,u)} - \sum_{i:y^{(i)}=-1}\varepsilon^{(i,u)}\lambda_k^{(i,u)}\right)$$

and finally we have

$$\phi_1^{(u)}(c) = -\gamma^{(u)}c \qquad \text{where} \qquad \gamma^{(u)} = \frac{1}{N}\left(\sum_{i:y^{(i)}=1}\varepsilon^{(i,u)} - \sum_{i:y^{(i)}=-1}\varepsilon^{(i,u)}\right)$$

Proof of Lemma 3

Proof of (i) ⇒ (ii). We start by using (i) to prove (ii). First note that (8) holds for $\boldsymbol{\omega} \in \overline{\Omega_u}$ due to the continuity of the loss. By part (i), $\boldsymbol{\omega}(t)$ remains in some fixed $\overline{\Omega_u}$ for all t small enough. Thus (9-11) apply. The bilinearity of ϕ_0^u and ϕ_1^u then yield

$$\mathcal{L}(\boldsymbol{\omega}(t)) - \mathcal{L}(\boldsymbol{\omega}) = t\phi_0^u(\tilde{W},\mathbf{v}) + t\phi_1^u(\tilde{\mathbf{b}},\mathbf{v}) = -t|v_k|\big(\langle\mathbf{a}_k^{(u)},\mathbf{q}\rangle + \alpha_k^{(u)}\beta\big),$$

which combined with (15) proves (ii).

Proof of (i). We now prove part (i). While straightforward, the proof is a little longer. Let us denote by $\boldsymbol{\omega}(t) = (W + t\tilde{W}, \mathbf{v}, \mathbf{b} + t\tilde{\mathbf{b}}, c) = (W(t), \mathbf{v}, \mathbf{b}(t), c)$ the perturbation considered in the lemma. Without loss of generality, let us choose $k = 1$. Then the first row of $W(t)$ and the first entry of $\mathbf{b}(t)$ are given by

$$\mathbf{w}_1(t) = \mathbf{w}_1 + t\text{sign}(v_1)\mathbf{q}, \qquad b_1(t) = b_1 + t\text{sign}(v_1)\beta$$

whereas the other rows and entries remains unchanged,

$$\mathbf{w}_k(t) = \mathbf{w}_k, \qquad \text{and} \qquad b_k(t) = b_k \qquad \text{for } k \geq 2.$$

Define the activations,

$$\begin{aligned}
&\gamma_1^{(i)}(t) = \langle\mathbf{w}_1(t),\mathbf{x}^{(i)}\rangle + b_1(t) = \langle\mathbf{w}_1,\mathbf{x}^{(i)}\rangle + b_1 + t\text{sign}(v_1)\left(\langle\mathbf{q},\mathbf{x}^{(i)}\rangle + \beta\right) \\
&\gamma_k^{(i)} = \langle\mathbf{w}_k,\mathbf{x}^{(i)}\rangle + b_k \qquad \text{for } k \geq 2.
\end{aligned}$$

so that the functions involved in the signature $\mathcal{S}(\boldsymbol{\omega}(t))$ can be written as:

$$\lambda_1^{(i)}(\boldsymbol{\omega}(t)) = \sigma'(\gamma_1^{(i)}(t))$$

$$\lambda_k^{(i)}(\boldsymbol{\omega}(t)) = \sigma'(\gamma_k^{(i)}) \qquad \text{for } k \geq 2 \tag{22}$$

$$\varepsilon^{(i)}(\boldsymbol{\omega}(t)) = \sigma' \left[1 - y^{(i)} \left\{ c + v_1 \sigma(\gamma_1^{(i)}(t)) + \sum_{k=2}^{K} v_k \sigma(\gamma_k^{(i)}) \right\} \right] \tag{23}$$

Recall that that signature function, for the network considered here, is simply given by the collection of all the functions $\boldsymbol{\lambda}^{(i)} = (\lambda_1^{(i)}, \ldots, \lambda_k^{(i)})^T$ and $\varepsilon^{(i)}$:

$$\mathcal{S}(\boldsymbol{\omega}(t)) = \Big(\boldsymbol{\lambda}^{(1)}(\boldsymbol{\omega}(t)), \varepsilon^{(1)}(\boldsymbol{\omega}(t)); \ldots; \boldsymbol{\lambda}^{(N)}(\boldsymbol{\omega}(t)), \varepsilon^{(N)}(\boldsymbol{\omega}(t)) \Big)$$

For the network considered here, since there are K hidden neurons, one output neuron, and N data points, we have that $\mathcal{S} : \Omega \mapsto \{0, 1/2, 1\}^{N(K+1)}$. We now make the following claim, that will be proven at the end of this section:

Claim. *There exists* $t_0 > 0$ *such that the function* $t \mapsto \mathcal{S}(\boldsymbol{\omega}(t))$ *is constant on* $(0, t_0)$.

Note that the above claim implies that for $t \in (0, t_0)$, $\boldsymbol{\omega}(t)$ either remains in a fixed cell Ω_u (if none of the entries of $\mathcal{S}(\boldsymbol{\omega}(t))$ are equal to 1/2) or on the boundary of a fixed cell Ω_u (if some of the entries of $\mathcal{S}(\boldsymbol{\omega}(t))$ are equal to 1/2). In both cases we have that $\boldsymbol{\omega}(t) \in \overline{\Omega_u}$ for all $t \in (0, t_0)$. Since $\boldsymbol{\omega}(t)$ is continuous and since $\overline{\Omega_u}$ is closed, we then clearly have that $\boldsymbol{\omega}(t) \in \overline{\Omega_u}$ for all $t \in [0, t_0)$, which conclude the proof of Lemma 3(i). We now prove the claim:

Proof of the claim. Let us fix $i \in \{1, \ldots, N\}$. Note that the function $t \mapsto \gamma_1^{(i)}(t)$ is monotone and continuous (it is simply an affine function of t). As a consequence, the quantity appearing inside $\sigma'[\cdot]$ in equation (23), that is,

$$g(t) = 1 - y^{(i)} \left\{ c + v_1 \sigma(\gamma_1^{(i)}(t)) + \sum_{k=2}^{K} v_k \sigma(\gamma_k^{(i)}) \right\}$$

is also continuous and monotone. Without loss of generality, let assume that g is non-decreasing. Continuity and monotonicity then implies that there are 3 intervals $(-\infty, a)$, $[a, b]$ and (b, ∞) on which g is strictly negative, equal to zero, then strictly positive (with the understanding that $a \leq b$, and both a and b can take infinite values, in which case we would have less than three intervals). As a consequence $\varepsilon^{(i)}(\boldsymbol{\omega}(t))$ is equal to 0, then 1/2, then 1 on each of these three intervals. Note that one can always choose $\tau^{(i)} > 0$ small enough so that $(0, \tau^{(i)})$ is fully contained in one of these three intervals, and this implies that $\varepsilon^{(i)}(\boldsymbol{\omega}(t))$ is constant on the interval $(0, \tau^{(i)})$. A similar line of reasoning shows that there exists $\hat{\tau}^{(i)} > 0$ such that $\lambda_1^{(i)}(\boldsymbol{\omega}(t))$ is constant on $(0, \hat{\tau}^{(i)})$. The interval $(0, t_0)$ is then obtained by taking the intersection of all these intervals:

$$(0, t_0) = \bigcap_{i=1}^{N} (0, \tau^{(i)}) \cap (0, \hat{\tau}^{(i)})$$

□

Proof of Theorem 2

The proof is by contradiction. Suppose $\ell^{(i)}(\boldsymbol{\omega}) > 0$ and for some k both $v_k \neq 0$ and $\sigma(\langle \mathbf{w}_k, \mathbf{x}^{(i)} \rangle + b_k) \neq 0$ hold. Consider the perturbation $\boldsymbol{\omega}(t)$ of lemma 3. Then there exists $u \in \{0, 1\}^{ND}$ and $t_0 > 0$ such that $\boldsymbol{\omega}(t) \in \overline{\Omega}_u$ for $t \in [0, t_0)$. By continuity of $\boldsymbol{\omega}(t)$ there exists $\hat{\boldsymbol{\omega}} = (\hat{W}, \hat{\mathbf{v}}, \hat{\mathbf{b}}, \hat{c}) \in \Omega_u$ such that $\ell^{(i)}(\hat{\boldsymbol{\omega}}) > 0$ and $\sigma(\langle \hat{\mathbf{w}}_k, \mathbf{x}^{(i)} \rangle + \hat{b}_k) \neq 0$. Thus $\varepsilon^{(i,u)} = 1$ and $\lambda_k^{(i,u)} = 1$ in Ω_u. As $|v_k| > 0$ lemma 3(ii) implies that the perturbation leads to a strict decrease of the loss, which contradicts the assumption that $\boldsymbol{\omega}$ is a local minimizer.

Proof of Theorem 3

Definition 1 and theorem 1 allow us to compute the Clarke subdifferential of $\mathcal{L}$ at $\boldsymbol{\omega}$ relatively easily. First recall that the open cells Ω_u fill parameter space up to a set $\mathcal{N}$ of measure zero. If $\boldsymbol{\omega} \in \mathcal{N}$ then $\boldsymbol{\omega}$ must lie on the boundary $\partial\Omega_u$ of some cell. Define the *incidence set*

$$\mathcal{I}(\boldsymbol{\omega}) := \big\{u \in \{0,1\}^{ND} : \boldsymbol{\omega} \in \partial\Omega_u\big\}$$

of such a point $\boldsymbol{\omega} \in \mathcal{N}$ as the collection of all such possible cells. Thus $\mathcal{I}(\boldsymbol{\omega})$ is both non-empty and finite. If $\boldsymbol{\omega}_k \to \boldsymbol{\omega}$ and $\boldsymbol{\omega}_k \notin \mathcal{N}$ we may, by passing to a subsequence if necessary, assume that $\boldsymbol{\omega}_k \in \Omega_u$ for some $u \in \mathcal{I}(\boldsymbol{\omega})$ and all k sufficiently large. But then $\nabla\mathcal{L}(\boldsymbol{\omega}_k) = \nabla\mathcal{L}|_{\Omega_u}(\boldsymbol{\omega}_k)$, and since $\nabla\mathcal{L}|_{\Omega_u}$ is a continuous function (i.e. a sum of multilinear gradients), it can be extended by continuity to the point $\boldsymbol{\omega} \in \partial\Omega_u$ and the limit $\nabla\mathcal{L}(\boldsymbol{\omega}_k) \to \nabla\mathcal{L}|_{\Omega_u}(\boldsymbol{\omega})$ follows. We therefore conclude from definition 1 that the Clark subdifferential at $\boldsymbol{\omega} \in \mathcal{N}$ is given by the convex hull

$$\partial_0\mathcal{L}(\boldsymbol{\omega}) = \left\{ \sum_{u\in\mathcal{I}(\boldsymbol{\omega})} \theta^{(u)}\nabla\mathcal{L}|_{\Omega_u}(\boldsymbol{\omega}) : \theta^{(u)} \geq 0,\ \sum_u \theta^{(u)} = 1 \right\} \tag{24}$$

where it has to be understood that $\nabla\mathcal{L}|_{\Omega_u}(\boldsymbol{\omega})$ denotes the extension by continuity of $\nabla\mathcal{L}|_{\Omega_u}$ to the point $\boldsymbol{\omega}$. If $\boldsymbol{\omega} \in \Omega_u$ for some u we let $\mathcal{I}(\boldsymbol{\omega}) = \{u\}$ denote its incidence set. As the gradient $\nabla\mathcal{L}$ depends continuously on $\boldsymbol{\omega}$ in cells we therefore have $\nabla\mathcal{L}(\boldsymbol{\omega}_k) \to \nabla\mathcal{L}(\boldsymbol{\omega})$ and so (24) also gives the Clark subdifferential in this case with $\theta^{(u)} = 1$.

Suppose now that $\mathbf{0} \in \partial_o\mathcal{L}(\boldsymbol{\omega})$, then we must have that

$$\mathbf{0} = \sum_{u\in\mathcal{I}(\boldsymbol{\omega})} \theta^{(u)}\nabla\mathcal{L}|_{\Omega_u}(\boldsymbol{\omega}) \tag{25}$$

for some collection of positive coefficients $\theta^{(u)}$ due to the characterization (24) of the subdifferential. Using the explicit formula from lemma 2 we can compute the gradients $\nabla\mathcal{L}|_{\Omega_u}(\boldsymbol{\omega})$. In particular, from equations (12) we find that

$$\frac{\partial\mathcal{L}|_{\Omega_u}}{\partial\mathbf{w}_k}(\boldsymbol{\omega}) = -v_k\mathbf{a}_k^{(u)} \qquad \text{and} \qquad \frac{\partial\mathcal{L}|_{\Omega_u}}{\partial b_k}(\boldsymbol{\omega}) = -v_k\alpha_k^{(u)}$$

Equation (25) then obviously implies:

$$\sum_{u\in\mathcal{I}(\boldsymbol{\omega})} \theta^{(u)}v_k\mathbf{a}_k^{(u)} = 0 \qquad \text{and} \qquad \sum_{u\in\mathcal{I}(\boldsymbol{\omega})} \theta^{(u)}v_k\alpha_k^{(u)} = 0$$

for all k. The precise formula for $\mathbf{a}_k^{(u)}$ and b_k provided in lemma 2 then give the equalities

$$\mathbf{0} = v_k\left(\sum_u \sum_{i:y^{(i)}=1} \theta^{(u)}\varepsilon^{(i,u)}\lambda_k^{(i,u)}\mathbf{x}^{(i)} - \sum_u \sum_{i:y^{(i)}=-1} \theta^{(u)}\varepsilon^{(i,u)}\lambda_k^{(i,u)}\mathbf{x}^{(i)}\right)$$

$$0 = v_k\left(\sum_u \sum_{i:y^{(i)}=1} \theta^{(u)}\varepsilon^{(i,u)}\lambda_k^{(i,u)} - \sum_u \sum_{i:y^{(i)}=-1} \theta^{(u)}\varepsilon^{(i,u)}\lambda_k^{(i,u)}\right)$$

If $v_k \neq 0$ then we may interchange summations to find

$$\sum_{i:y^{(i)}=1} \varrho_k^{(i)}\mathbf{x}^{(i)} = \sum_{i:y^{(i)}=-1} \varrho_k^{(i)}\mathbf{x}^{(i)} \tag{26}$$

$$\sum_{i:y^{(i)}=1} \varrho_k^{(i)} = \sum_{i:y^{(i)}=-1} \varrho_k^{(i)} \qquad \text{where} \qquad \varrho_k^{(i)} := \sum_u \theta^{(u)}\varepsilon^{(i,u)}\lambda_k^{(i,u)} \tag{27}$$

We now claim that equalities (26)–(27) cannot happen unless all the $\varrho_k^{(i)}$ vanish. To see this, note that if the $\varrho_k^{(i)}$ do not vanish, we can set $Q = \sum_{i:y^{(i)}=1}\varrho_k^{(i)} = \sum_{i:y^{(i)}=-1}\varrho_k^{(i)}$, and dividing both side of equality (26) by Q to obtain

$$\sum_{i:y^{(i)}=1} \frac{\varrho_k^{(i)}}{Q}\ \mathbf{x}^{(i)} = \sum_{i:y^{(i)}=-1} \frac{\varrho_k^{(i)}}{Q}\ \mathbf{x}^{(i)}$$

The above equation shows that a convex combination of data points of class +1 is equal to a convex combination of data points of class -1, which is not possible since the data points are linearly separable.

Assume now that for some data point $\mathbf{x}^{(i)}$ we have $\ell^{(i)}(\boldsymbol{\omega}) > 0$. Then by continuity of the loss we also have $\ell^{(i)} > 0$ on each neighboring cell Ω_u, $u \in \mathcal{I}(\boldsymbol{\omega})$. Using the definition (6) of $\varepsilon^{(i,u)}$ we then see that $\varepsilon^{(i,u)} = 1$ for all $u \in \mathcal{I}(\boldsymbol{\omega})$. If $v_k \neq 0$ for some k, then the corresponding $\varrho_k^{(i)}$ must be equal to zero, which necessarily implies that $\lambda_k^{(i,u)} = 0$ some u since the $\varepsilon^{(i,u)}$ are all equal to one and at least one of the $\theta^{(u)}$ is nonzero. This in turn implies $\sigma(\langle \mathbf{w}_k, \mathbf{x}^{(i)}\rangle + b_k) = 0$ due to the definition of the $\lambda_k^{(i)}$.

Proof of Theorem 4

The notion of a cell Ω_u for the model (17) consists of sets (Cartesian products) of the form

$$\Omega_u = \Omega_{u^{(1)}} \times \Omega_{u^{(2)}} \times \cdots \times \Omega_{u^{(R)}},$$

where each $u^{(r)} \in \{0,1\}^{ND}$ denotes a signature collection for the individual two-class losses $\mathcal{L}^{(r)}$ and thus

$$u = \big(u^{(1)}, \ldots, u^{(R)}\big) \in \{0,1\}^{NDR}$$

defines a signature collection for the full model. That sets of this form cover the product space $\Omega \times \cdots \times \Omega$ (R-copies) up to a set of measure zero follows easily from the fact that if $\big(\boldsymbol{\omega}^{(1)}, \ldots, \boldsymbol{\omega}^{(R)}\big) \notin \Omega_u$ for all u then at least one of the $u^{(r)}$ (say $u^{(1)}$ WLOG) lies in the set

$$\mathcal{N} := \Omega \setminus \left(\bigcup_{u^{(1)} \in \{0,1\}^{ND}} \Omega_{u^{(1)}} \right)$$

which has measure zero in Ω. Thus u must lie in the set

$$\mathcal{N} \times \Omega \times \cdots \times \Omega$$

which has measure zero in the product space $\Omega \times \cdots \times \Omega$, and so the union of the R measure zero sets of the form

$$\Omega \times \cdots \times \mathcal{N} \times \cdots \times \Omega$$

contains all parameters $\big(\boldsymbol{\omega}^{(1)}, \ldots, \boldsymbol{\omega}^{(R)}\big)$ that do not lie in a cell.

Now let $\big(\boldsymbol{\omega}^{(1)}, \ldots, \boldsymbol{\omega}^{(R)}\big)$ denote any critical point. For each (ℓ, r) we have

$$\nabla_{\omega^{(\ell,r)}} \mathcal{R} = \gamma\big(\omega^{(\ell,r)} - \tilde{\omega}^{(\ell,r)}\big) \qquad \tilde{\omega}^{(\ell,r)} := \frac{1}{R-1} \sum_{s \neq r} \omega^{(\ell,s)}$$

$$\nabla_{\mathbf{b}^{(\ell,r)}} \mathcal{R} = \gamma\big(\mathbf{b}^{(\ell,r)} - \tilde{\mathbf{b}}^{(\ell,r)}\big) \qquad \tilde{\mathbf{b}}^{(\ell,r)} := \frac{1}{R-1} \sum_{s \neq r} \mathbf{b}^{(\ell,s)}$$

by straightforward calculation. By definition of a critical point, for each cell Ω_u adjacent to the critical point there exist corresponding constants $\theta^{(u)} \geq 0$ with $\sum_u \theta^{(u)} = 1$ so that the equalities

$$\begin{aligned}
\mathbf{0} &= \sum_u \theta^{(u)} \nabla_{\mathbf{v}_r} \bar{\mathcal{L}}|_{\Omega_u} \\
\mathbf{0} &= \sum_u \theta^{(u)} \left(\nabla_{\omega^{(\ell,r)}} \bar{\mathcal{L}}|_{\Omega_u} + \nabla_{\omega^{(\ell,r)}} \mathcal{R} \right) = \gamma\big(\omega^{(\ell,r)} - \tilde{\omega}^{(\ell,r)}\big) + \sum_u \theta^{(u)} \nabla_{\omega^{(\ell,r)}} \bar{\mathcal{L}}|_{\Omega_u} \\
\mathbf{0} &= \sum_u \theta^{(u)} \left(\nabla_{\mathbf{b}^{(\ell,r)}} \bar{\mathcal{L}}|_{\Omega_u} + \nabla_{\mathbf{b}^{(\ell,r)}} \mathcal{R} \right) = \gamma\big(\mathbf{b}^{(\ell,r)} - \tilde{\mathbf{b}}^{(\ell,r)}\big) + \sum_u \theta^{(u)} \nabla_{\mathbf{b}^{(\ell,r)}} \bar{\mathcal{L}}|_{\Omega_u} \qquad (28)
\end{aligned}$$

hold for all $1 \leq \ell \leq L$ and $1 \leq r \leq R$, where the final equalities in the second and third line follow from the fact that $\mathcal{R}$ is smooth and so its gradients do not depend upon the cell. Now on any cell we may decompose each $\mathcal{L}^{(r)}$ into a sum of multilinear forms

$$\begin{aligned}
\mathcal{L}^{(r)}|_{\Omega_u} = \phi_0^{(u,r)}\big(\omega^{(1,r)}, \ldots, \omega^{(L,r)}, \mathbf{v}_r\big) + \sum_{\ell=1}^{L-1} \phi_\ell^{(u,r)}\big(\mathbf{b}^{(\ell,r)}, \omega^{(\ell+1,r)}, \ldots, \omega^{(L,r)}, \mathbf{v}_r\big) \\
+ \phi_L^{(u,r)}\big(\mathbf{b}^{(L,r)}, \mathbf{v}_r\big) + \phi_{L+1}^{(u,r)}(c_r) + \phi_{L+2}^{(u,r)}
\end{aligned}$$

by theorem 1. For any multilinear form $\phi(\mathbf{v}_1, \ldots, \mathbf{v}_n)$ we have

$$\phi(\mathbf{v}_1, \ldots, \mathbf{v}_n) = \langle \mathbf{v}_k, \nabla_{\mathbf{v}_k} \phi(\mathbf{v}_1, \ldots, \mathbf{v}_n) \rangle$$

for all $1 \leq k \leq n$ by Euler's theorem for homogeneous functions. Taking the inner-product of (28) with $\mathbf{v}_r, \omega^{(L,r)}$ and $\mathbf{b}^{(L,r)}$ then shows

$$\begin{aligned} 0 &= \sum_u \theta^{(u)} \big(\phi_0^{(u,r)} + \cdots + \phi_L^{(u,r)} \big) \\ 0 &= \sum_u \theta^{(u)} \big(\phi_0^{(u,r)} + \cdots + \phi_{L-1}^{(u,r)} \big) + \gamma \big(\|\omega^{(L,r)}\|^2 - \langle \omega^{(L,r)}, \tilde{\omega}^{(L,r)} \rangle \big) \\ 0 &= \sum_u \theta^{(u)} \big(\phi_L^{(u,r)} \big) + \gamma \big(\|\mathbf{b}^{(L,r)}\|^2 - \langle \mathbf{b}^{(L,r)}, \tilde{\mathbf{b}}^{(L,r)} \rangle \big) \end{aligned} \tag{29}$$

which upon adding the second and third equalities yields

$$\|\omega^{(L,r)}\|^2 + \|\mathbf{b}^{(L,r)}\|^2 = \langle \omega^{(L,r)}, \tilde{\omega}^{(L,r)} \rangle + \langle \mathbf{b}^{(L,r)}, \tilde{\mathbf{b}}^{(L,r)} \rangle$$

for all $1 \leq r \leq R$. By the definitions of $\tilde{\omega}^{(L,r)}$ and $\tilde{\mathbf{b}}^{(L,r)}$ (c.f. lemma 4), this can happen if and only if

$$\omega^{(L,1)} = \cdots = \omega^{(L,R)} \qquad \text{and} \qquad \mathbf{b}^{(L,1)} = \cdots = \mathbf{b}^{(L,R)}.$$

Using this in the second and third equations in (29) then shows that

$$0 = \sum_u \theta^{(u)} \big(\phi_0^{(u,r)} + \cdots + \phi_{L-1}^{(u,r)} \big) = \sum_u \theta^{(u)} \big(\phi_L^{(u,r)} \big) \tag{30}$$

for all $1 \leq r \leq R$ as well. Now take the inner-product of (28) with $\omega^{(L-1,r)}$ and $\mathbf{b}^{(L-1,r)}$ to find

$$\begin{aligned} 0 &= \sum_u \theta^{(u)} \big(\phi_0^{(u,r)} + \cdots + \phi_{L-2}^{(u,r)} \big) + \gamma \big(\|\omega^{(L-1,r)}\|^2 - \langle \omega^{(L-1,r)}, \tilde{\omega}^{(L-1,r)} \rangle \big) \\ 0 &= \sum_u \theta^{(u)} \big(\phi_{L-1}^{(u,r)} \big) + \gamma \big(\|\mathbf{b}^{(L-1,r)}\|^2 - \langle \mathbf{b}^{(L-1,r)}, \tilde{\mathbf{b}}^{(L-1,r)} \rangle \big) \end{aligned}$$

Adding these equations and using (30) then reveals

$$\omega^{(L-1,1)} = \cdots = \omega^{(L-1,R)} \qquad \text{and} \qquad \mathbf{b}^{(L-1,1)} = \cdots = \mathbf{b}^{(L-1,R)}$$

must hold as well, and so also

$$0 = \sum_u \theta^{(u)} \big(\phi_0^{(u,r)} + \cdots + \phi_{L-2}^{(u,r)} \big) = \sum_u \theta^{(u)} \big(\phi_{L-1}^{(u,r)} \big)$$

must hold. Continuing from $\ell = L - 2$ to $\ell = 1$ by induction reveals

$$\omega^{(\ell,1)} = \cdots = \omega^{(\ell,R)} \qquad \text{and} \qquad \mathbf{b}^{(\ell,1)} = \cdots = \mathbf{b}^{(\ell,R)}.$$

for all $1 \leq \ell \leq L$ and so the penalty is exact as claimed. Part (ii) then follows from part (i) since the equalities

$$\tilde{\omega}^{(\ell,r)} = \bar{\omega}^{(\ell)} = \omega^{(\ell,r)} \qquad \tilde{\mathbf{b}}^{(\ell,r)} = \mathbf{b}^{(\ell)} = \mathbf{b}^{(\ell,r)}$$

for all (ℓ, r) at any critical point. Thus (28) yields

$$\begin{aligned} \mathbf{0} &= \sum_u \theta^{(u)} \nabla_{\mathbf{v}_r} \mathcal{L}^{(r)}|_{\Omega_{u^{(r)}}} \\ \mathbf{0} &= \sum_u \theta^{(u)} \nabla_{\omega^{(\ell,r)}} \mathcal{L}^{(r)}|_{\Omega_{u^{(r)}}} \\ \mathbf{0} &= \sum_u \theta^{(u)} \nabla_{\mathbf{b}^{(\ell,r)}} \mathcal{L}^{(r)}|_{\Omega_{u^{(r)}}} \end{aligned} \tag{31}$$

for all $1 \leq \ell \leq L, 1 \leq r \leq R$. Now consider (31) for $r = 1$. Any cells appearing in the sum (31) satisfy either $(\breve{\boldsymbol{\omega}}, \mathbf{v}_1, c_1) \in \Omega_{u^{(1)}}$ or $(\breve{\boldsymbol{\omega}}, \mathbf{v}_1, c_1) \in \partial\Omega_{u^{(1)}}$. If $(\breve{\boldsymbol{\omega}}, \mathbf{v}_1, c_1) \in \Omega_{u^{(1)}}$ for some $u^{(1)}$ then (31) must consist only of gradients on the single cell $\Omega_{u^{(1)}}$ and so $(\breve{\boldsymbol{\omega}}, \mathbf{v}_1, c_1) \in \Omega_{u^{(1)}}$ is a critical point of $\mathcal{L}^{(1)}$ in the classical sense. If $(\breve{\boldsymbol{\omega}}, \mathbf{v}_1, c_1) \in \partial\Omega_{u^{(1)}}$ for some $u^{(1)}$ in the sum then $(\breve{\boldsymbol{\omega}}, \mathbf{v}_1, c_1) \in \partial\Omega_{u^{(1)}}$ for all cells u the sum. Thus (31) consists of a positive combination of gradients of $\mathcal{L}^{(1)}$ on cells adjacent to $(\breve{\boldsymbol{\omega}}, \mathbf{v}_1, c_1)$, and so $(\breve{\boldsymbol{\omega}}, \mathbf{v}_1, c_1)$ defines a critical point of $\mathcal{L}^{(1)}$ in the extended Clarke sense. Applying this reasoning for $r = 2, \ldots, R$ then yields part (ii) and proves the theorem.

Lemma 4. *For any R vectors $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(R)} \in \mathbb{R}^d$, if*

$$\|\mathbf{x}^{(r)}\|^2 = \frac{1}{R-1} \sum_{s \neq r} \langle \mathbf{x}^{(s)}, \mathbf{x}^{(r)} \rangle \qquad \textit{for all} \qquad r \in \{1, \dots, R\}$$

then $\mathbf{x}_1 = \cdots = \mathbf{x}_R$.

Proof. By relabelling if necessary, assume $\mathbf{x}^{(1)}$ has largest norm. Thus $\|\mathbf{x}^{(1)}\| \geq \|\mathbf{x}^{(r)}\|$ for all $1 \leq r \leq R$. If $\|\mathbf{x}^{(1)}\| = 0$ then there is nothing to prove. Otherwise apply Cauchy-Schwarz and the hypothesis of the lemma to find

$$\|\mathbf{x}^{(1)}\|^2 \leq \frac{1}{R-1} \sum_{s \neq 1} \|\mathbf{x}^{(s)}\| \| \|\mathbf{x}^{(1)}\|$$

$$\|\mathbf{x}^{(1)}\| \leq \frac{1}{R-1} \sum_{s \neq 1} \|\mathbf{x}^{(s)}\|.$$

The latter inequality implies $\|\mathbf{x}^{(1)}\| = \cdots = \|\mathbf{x}^{(R)}\|$ since $\mathbf{x}^{(1)}$ has largest norm. Thus

$$\|\mathbf{x}^{(1)}\|^2 = \frac{1}{R-1} \sum_{s \neq 1} \cos\theta_r \|\mathbf{x}^{(1)}\|^2$$

$$1 = \frac{1}{R-1} \sum_{s \neq 1} \cos\theta_r$$

by the hypothesis of the lemma. The latter equality implies $\cos\theta_r = 1$ for all r, and so the lemma is proved.

□