Appendix: Proofs of Lemmas and Theorems

All numbering in the appendix (corollaries, lemmas, theorems and equations) begin at 100 to distinguish them from their counterparts in the main text; any equation number or theorem number below 100 refers to a theorem or equation in the main text.

Proofs of lemmas and theorems from section 2: Global Structure of the Loss

Lemma (Lemma 1 from the paper)**.** *For each $u \in \{-1,1\}^{ND}$ the cell Ω_u 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.*

Proof. 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. That $\mathcal{N}$ is closed follows from the fact that it is the complement of an open set. If $\boldsymbol{\omega} \notin \Omega_u$ for all $u \in \{-1,1\}^{ND}$ then at least one of the equalities

$$\mathbf{b}^{(\ell)}_j = -\langle \mathbf{w}^{(\ell)}_j, \mathbf{x}^{(i,\ell-1)}\rangle \quad \text{or} \quad \mathbf{c}_s = -\big(1 + \langle \mathbf{v}_s - \mathbf{v}_r, \mathbf{x}^{(i,L)}\rangle\big), \tag{100}$$

must hold. In the above equation $\mathbf{b}^{(\ell)}_j$ and $\mathbf{c}_j$ stands for the j^{th} entry of the bias vectors $\mathbf{b}^{(\ell)}$ and $\mathbf{c}$ appearing in equation (4), whereas $\mathbf{w}^{(\ell)}_j$ and $\mathbf{v}_j$ stands for the j^{th} row of the weight matrices $W^{(\ell)}$ and V. The set of parameters $\mathcal{N} \subset \Omega$ where an equality of the form (100) holds corresponds to a Lipschitz graph in Ω of the bias parameter for that equality. It therefore follows that

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

defines a set contained in a finite union of Lipschitz graphs. Thus $\mathcal{N}$ is $(n_p - 1)$-rectifiable, where $n_p := \dim(\Omega)$ denotes the total number of parameters. This implies that $\mathcal{N}$ has Lebesgue measure zero. □

Theorem (Theorem 1 from the paper)**.** *For each cell Ω_u there exist multilinear forms $\phi^u_0, \ldots, \phi^u_{L+1}$ and a constant ϕ^u_{L+2} such that*

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

Proof. Let us define the collection of functions

$$\begin{aligned}
\boldsymbol{\lambda}^{(i,\ell)}(\boldsymbol{\omega}) &:= \sigma'_\alpha(W^{(\ell)}\mathbf{x}^{(i,\ell-1)} + \mathbf{b}^\ell) \qquad \text{for} \quad \ell \in [L] \\
\boldsymbol{\varepsilon}^{(i)}(\boldsymbol{\omega}) &:= \sigma'\Big(\, (\mathrm{Id} - \mathbf{1} \otimes \mathbf{y}^{(i)}) \, \hat{\mathbf{y}}^{(i)} + \mathbf{1} \, \Big),
\end{aligned} \tag{101}$$

Comparing these equations with the definition (6) of the signature functions $\mathbf{s}^{(i,\ell)}(\boldsymbol{\omega})$ it is obvious that $\boldsymbol{\lambda}^{(i,\ell)}(\boldsymbol{\omega})$ and $\boldsymbol{\varepsilon}^{(i)}(\boldsymbol{\omega})$ remain constant on each cell Ω_u. We may therefore refer unambiguously to these functions by referencing a given cell Ω_u instead of a point $\boldsymbol{\omega}$ in parameter space. We shall therefore interchangably use the more convenient notation

$$\begin{aligned}\boldsymbol{\lambda}^{(i,\ell,u)} &:= \boldsymbol{\lambda}^{(i,\ell)}(\boldsymbol{\omega}) \qquad \text{for all } \boldsymbol{\omega} \in \Omega_u \\ \boldsymbol{\varepsilon}^{(i,u)} &:= \boldsymbol{\varepsilon}^{(i)}(\boldsymbol{\omega}) \qquad \text{for all } \boldsymbol{\omega} \in \Omega_u \end{aligned} \tag{102}$$

when referring to these constants.

For simplicity of the exposition let us temporarily assume that that the network has no bias, and let us ignore the vector $\mathbf{1}$ appearing in equation (5). Also let us define the matrix $T^{(i)} = \mathrm{Id} - \mathbf{1} \otimes \mathbf{y}^{(i)}$. The loss (5) then becomes:

$$\mathcal{L}(W^{(1)}, \ldots, W^{(L)}, V) = \sum_i \mathbf{1}^T \sigma(\, T^{(i)} V \sigma_\alpha(W^{(L)} \ldots \sigma_\alpha(W^{(2)} \sigma_\alpha(W^{(1)} \mathbf{x}^{(i)}))))$$

Since inside a cell Ω_u the activation pattern of the ReLU's or leaky ReLUs does not change, each σ and σ_α in the above equation can be replaced by a diagonal matrix with 0, α or ones in its diagonal. To be more precise, restricted to the cell Ω_u, the loss can be written:

$$\mathcal{L}|_{\Omega_u}(W^{(1)}, \ldots, W^{(L)}, V) = \sum_i \mathbf{1}^T \mathcal{E}^{(i,u)} T^{(i)} V \Lambda^{(i,L,u)} W^{(L)} \;\ldots\; \Lambda^{(i,2,u)} W^{(2)} \Lambda^{(i,1,u)} W^{(1)} \mathbf{x}^{(i)}$$

where $\mathcal{E}^{(i,u)} = \mathrm{diag}(\boldsymbol{\varepsilon}^{(i,u)})$ and $\Lambda^{(i,\ell,u)} = \mathrm{diag}(\boldsymbol{\lambda}^{(i,\ell,u)})$. From the above equation it is clear that $\mathcal{L}|_{\Omega_u}$ is a multilinear form of its arguments.

Going back to our case of interest, where we do have biases and where we do not ignore the vector $\mathbf{1}$, the picture becomes slightly more complex: the loss restricted to a cell is now a sum of multilinear form rather than a single multilinear form. The exact formula follows by carefully expanding

$$\begin{aligned}\mathcal{L}|_{\Omega_u} &= -1 + \frac{1}{N} \sum_i \mathbf{1}^T \sigma\left(T^{(i)}(V \sigma_\alpha(W^{(L)} \sigma_\alpha(\ldots W^{(2)} \sigma_\alpha(W^{(1)} \mathbf{x}^{(i)} + \mathbf{b}_1) + \mathbf{b}_2 \ldots) + \mathbf{b}_L) + \mathbf{c}) + \mathbf{1}\right) \\ &= -1 + \frac{1}{N} \sum_i \mathbf{1}^T \mathcal{E}^{(i,u)} \left(T^{(i)}(V \Lambda^{(i,L,u)}(W^{(L)} \Lambda^{(i,L-1,u)}(\ldots W^{(2)} \Lambda^{(i,1,u)}(W^{(1)} \mathbf{x}^{(i)} + \mathbf{b}_1) + \mathbf{b}_2 \ldots) + \mathbf{b}_L) + \mathbf{c}) + \mathbf{1}\right)\end{aligned}$$

□

Corollary (Corollary 2 from the paper)**.** *If $\boldsymbol{\omega} \in \Omega \setminus \mathcal{N}$ and the Hessian matrix $D^2\mathcal{L}(\boldsymbol{\omega})$ does not vanish, then it must have at least one strictly positive and one strictly negative eigenvalue.*

Proof. it suffices to note that a multilinear form $\phi : \mathbb{R}^{d_1} \times \ldots \times \mathbb{R}^{d_n} \to \mathbb{R}$ can always be written as

$$\phi(\mathbf{v}_1, \ldots, \mathbf{v}_n) = \sum_{j_1=1}^{d_1} \ldots \sum_{j_n=1}^{d_n} A_{j_1,\ldots,j_n} v_{1,j_1} \ldots v_{n,j_n} \tag{103}$$

for some tensor $\{A_{j_1,\ldots,j_n} : 1 \leq j_k \leq d_k\}$, with $v_{k,j}$ denoting the j^{th} component of the vector $\mathbf{v}_k$. From (103) it is clear that

$$\frac{\partial^2 \phi}{\partial v_{k,j}^2} = 0$$

and therefore the trace of the Hessian matrix of ϕ vanishes. Thus the (real) eigenvalues of the (real, orthogonally diagonalizable) Hessian sum to zero, and so if the Hessian is not the zero matrix then it has at least one strictly positive and one strictly negative eigenvalue. □

Corollary (Corollary 1 from the paper)**.** *Local minima and maxima of the loss occur only on the boundary set $\mathcal{N}$ or on those cells Ω_u where the loss is constant. In the latter case, $\mathcal{L}|_{\Omega_u}(\boldsymbol{\omega}) = \phi^u_{L+2}$.*

Proof. The proof relies on the so called *maximum principle* for harmonic functions. We first recall that a real valued function $f : \mathbb{R}^d \to \mathbb{R}$ is said to be harmonic at a point x if it twice differentiable at x and if its Laplacian vanishes at x, that is:

$$\Delta f(x) := \textstyle\sum_{i=1}^d \frac{\partial^2 f}{\partial x_i^2}(x) = 0 \tag{104}$$

Note that (104) is equivalent to saying that the Hessian matrix $D^2 f(x)$ has zero trace. A function is said to be harmonic on an open set $\mathcal{O}$ if it harmonic at every points $x \in \mathcal{O}$. In the proof of the previous corollary we have shown that the trace of $D^2\mathcal{L}(\boldsymbol{\omega})$ is equal to zero if $\boldsymbol{\omega} \notin \mathcal{N}$, which exactly means that the loss is harmonic on the open set $\Omega \backslash \mathcal{N}$.

The strong maximum principle states that a non-constant harmonic function cannot attain a local minimum or maximum at an interior point of an open connected set. Therefore if $\hat{\boldsymbol{\omega}}$ is a local minimum of $\mathcal{L}$ that occurs in a cell Ω_u, then 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

$$\mathcal{L}|_{\Omega_u}(\hat{\boldsymbol{\omega}} + \delta\boldsymbol{\omega}) = \mathcal{L}|_{\Omega_u}(\hat{\boldsymbol{\omega}}) \tag{105}$$

must hold for all $\boldsymbol{\omega}$ and all δ small enough. Now use the multilinearity of $\mathcal{L}|_{\Omega_u}$ to expand the left-hand-side into powers of δ:

$$\mathcal{L}|_{\Omega_u}(\hat{\boldsymbol{\omega}} + \delta\boldsymbol{\omega}) = \mathcal{L}|_{\Omega_u}(\hat{\boldsymbol{\omega}}) + \sum_{k=1}^{L} \delta^k f_k(\hat{\boldsymbol{\omega}}; \boldsymbol{\omega}) + \delta^{L+1}\big(\phi_0^u + \phi_1^u\big)(\boldsymbol{\omega}). \tag{106}$$

That $\big(\phi_0^u + \phi_1^u\big)(\boldsymbol{\omega})$ is, in fact, the highest-order term come from the fact that, among all the multilinear forms appearing in the statement of theorem 1, ϕ_0^u and ϕ_1^u are the only one having $L+1$ inputs. The terms of order $k \le L$ appearing in the expansion (106) depends both on the minimizer $\hat{\boldsymbol{\omega}}$ and the perturbation $\boldsymbol{\omega}$, and we denote them by $f_k(\hat{\boldsymbol{\omega}}; \boldsymbol{\omega})$.

Combining (105) and (106) we have that

$$\sum_{k=1}^{L} \delta^k f_k(\hat{\boldsymbol{\omega}}; \boldsymbol{\omega}) + \delta^{L+1}\big(\phi_0^u + \phi_1^u\big)(\boldsymbol{\omega}) = 0. \tag{107}$$

Since (107) must hold for all δ small enough, all like powers must vanish

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

The second equation can be written as

$$\phi_0^u(\omega^{(1)}, \omega^{(2)}, \omega^{(3)}, \omega^{(4)} \ldots, \omega^{(L)}, V) = -\phi_1^u(\mathbf{b}^{(1)}, \omega^{(2)}, \omega^{(3)}, \omega^{(4)} \ldots, \omega^{(L)}, V)$$

Since ϕ_0^u depends linearly on $\omega^{(1)}$ whereas ϕ_0^u does not depend on $\omega^{(1)}$, and since ϕ_1^u depends linearly on $\mathbf{b}^{(1)}$ whereas ϕ_0^u does not depend on $\mathbf{b}^{(1)}$, the only way for the above equality to hold for all perturbation $\boldsymbol{\omega}$ is that both functions are the zero function. 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 theorem 1. This implies that

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

actually just depends on $\boldsymbol{\omega}$, and that f_L must vanish by (107). 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 $0 \le \ell \le L+1$, and so in fact

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

as claimed. □

Theorem (Theorem 2 from the paper)**.** *If $\boldsymbol{\omega}$ is a type II local minimum then $\mathcal{L}(\boldsymbol{\omega}) > 0$.*

Proof. We will show the contrapositive: if $\mathcal{L}(\boldsymbol{\omega}) = 0$, then $\boldsymbol{\omega}$ is a type I local minimum. Given some $\lambda > 1$ and $\varepsilon^{(\ell)} > 0$ put

$$\tilde{\boldsymbol{\omega}} := (\boldsymbol{\omega}^1, \ldots, \boldsymbol{\omega}^{(L)}, \lambda V, \mathbf{b}^{(1)} + \varepsilon^{(1)}, \ldots, \mathbf{b}^{(L)} + \varepsilon^{(L)}, \lambda \mathbf{c})$$

and let $\tilde{\mathbf{x}}^{(i,\ell)}$ denote the corresponding features at layer ℓ using these parameters. Then there exists some constant $C \geq 1$ so that

$$\left|\Delta\mathbf{x}^{(i,\ell)}\right| := \left|\tilde{\mathbf{x}}^{(i,\ell)} - \mathbf{x}^{(i,\ell)}\right| \leq C \max\left\{\varepsilon^{(1)}, \ldots, \varepsilon^{(\ell)}\right\}.$$

For each hidden layer $\ell \in [L]$ and each corresponding feature $k \in [d_\ell]$ define the activation levels

$$\begin{aligned}
\eta_k^{(i,\ell)} &:= \langle W_k^{(\ell)}(\omega^{(\ell)}), \mathbf{x}^{(i,\ell-1)}\rangle + \mathbf{b}_k^{(\ell)} \\
\tilde{\eta}_k^{(i,\ell)} &:= \langle W_k^{(\ell)}(\omega^{(\ell)}), \tilde{\mathbf{x}}^{(i,\ell-1)}\rangle + \mathbf{b}_k^{(\ell)} + \varepsilon^{(\ell)} = \eta_k^{(i,\ell)} + \langle W_k(\omega^{(\ell)}), \Delta\mathbf{x}^{(i,\ell-1)}\rangle + \varepsilon^{(\ell)}
\end{aligned}$$

for each set of parameters. Define

$$\eta^{(-,\ell)} := \max_{i\in[N], k\in[d_\ell]} \{\eta_k^{(i,\ell)} : \eta_k^{(i,\ell)} < 0\}$$

with the convention that $\eta^{(-,\ell)} = -\infty$ if $\eta_k^{(i,\ell)} \geq 0$ for all $k \in [d_\ell], i \in [N]$. Set

$$\eta^{(-)} := \max_{\ell \in [L]} \eta^{(-,\ell)}$$

as the largest negative activation level of the network. Put

$$W_* := \max_{\ell \in [L}\} \, \|W^{(\ell)}\|_2$$

as the largest norm of the matrices $W^{(\ell)}$ across the network. Take any $0 < \varepsilon^{(L)}$ so that

$$\varepsilon^{(L)} < \frac{|\eta^{(-)}|}{2C\max\{1, W_*\}},$$

then for $\ell = L-1, \ldots, 1$ inductively choose

$$0 < \varepsilon^{(\ell)} < \min\left\{\frac{|\eta^{(-)}|}{2C\max\{1, W_*\}}, \frac{\varepsilon^{(\ell+1)}}{C\max\{1, W_*\}}\right\}.$$

In particular, $\varepsilon^{(L)} > \varepsilon^{(L-1)} > \cdots > \varepsilon^{(1)} > 0$. For any such choice

$$\begin{aligned}
\tilde{\eta}_k^{(i,\ell)} &\geq \varepsilon^{(\ell)} - |\langle W_k^{(\ell)}, \Delta\mathbf{x}^{(i,\ell-1)}\rangle| \geq \varepsilon^{(\ell)} - C\max\{1, W_*\}\varepsilon^{(\ell-1)} > 0 \qquad \text{if} \quad \eta_k^{(i,\ell)} \geq 0 \\
\tilde{\eta}_k^{(i,\ell)} &\leq -\left|\eta^{(-)}\right| + C\max\{1, W_*\}\varepsilon^{(\ell-1)} + \varepsilon^{(\ell)} \leq -\frac{1}{2}\left|\eta^{(-)}\right| + \varepsilon^{(\ell)} < 0 \quad \text{if} \quad \eta_k^{(i,\ell)} < 0.
\end{aligned}$$

Now let $\mathrm{cl}(\mathbf{x}^{(i)}) \in \{1, \ldots, R\}$ denote the class of a data point. Put

$$\begin{aligned}
\mathbf{z}^{(i,r)}(\boldsymbol{\omega}) &:= \langle \mathbf{v}_r - \mathbf{v}_{\mathrm{cl}(\mathbf{x}^{(i)})}, \mathbf{x}^{(i,L)}(\boldsymbol{\omega})\rangle + (c_r - c_{\mathrm{cl}(\mathbf{x}^{(i)})}) \\
\tilde{\mathbf{z}}^{(i,r)}(\tilde{\boldsymbol{\omega}}) &:= \lambda\langle \mathbf{v}_r - \mathbf{v}_{\mathrm{cl}(\mathbf{x}^{(i)})}, \mathbf{x}^{(i,L)}(\tilde{\boldsymbol{\omega}})\rangle + \lambda(c_r - c_{\mathrm{cl}(\mathbf{x}^{(i)})}) = \lambda\mathbf{z}^{(i,r)} + \lambda\langle \mathbf{v}_r - \mathbf{v}_{\mathrm{cl}(\mathbf{x}^{(i)})}, \Delta\mathbf{x}^{(i,L)}\rangle
\end{aligned}$$

as the outputs of the network. That $\mathcal{L}(\boldsymbol{\omega}) = 0$ implies

$$\sigma\big(1 + \mathbf{z}^{(i,r)}\big) = 0 \quad \text{for all} \quad r \neq \mathrm{cl}(\mathbf{x}_i)$$

and so $\mathbf{z}^{(i,r)} \leq -1$ for all $i, r \neq \mathrm{cl}(\mathbf{x}_i)$. Thus

$$\tilde{\mathbf{z}}^{(i,r)} \leq -\lambda + 2C\|V\|_2\lambda\varepsilon^{(L)},$$

and so if

$$0 < \varepsilon^{(L)} < \frac{\lambda - 1}{2C\|V\|_2\lambda}$$

then $\tilde{\mathbf{z}}^{(i,r)} < -1$ for all $i, r \neq \mathrm{cl}(\mathbf{x}_i)$. For any $\lambda > 1$ and corresponding choices of $\varepsilon^{(\ell)}$ it follows that $\mathcal{L}(\tilde{\boldsymbol{\omega}}) = 0$. Moreover, by construction the signature function $\mathcal{S}(\tilde{\boldsymbol{\omega}})$ is constant as $\lambda, \varepsilon^{(\ell)}$ vary and so $\tilde{\boldsymbol{\omega}}$ lies in some fixed cell Ω_u for all $\lambda > 1$. But $\tilde{\boldsymbol{\omega}} \to \boldsymbol{\omega}$ as $\lambda \to 1$, so $\boldsymbol{\omega} \in \overline{\Omega}_u$. Finally, for this cell Ω_u it follows that $\mathcal{L}|_{\Omega_u} = 0$ by definition of the signature function. □

Theorem (Theorem 3 from the paper)**.** *Consider the loss (5) for a fully connected network. Assume that $\alpha > 0$ and that the data points $(\mathbf{x}^{(i)}, \mathbf{y}^{(i)})$ are generic. Then $\mathcal{L}(\boldsymbol{\omega}) = 0$ at any type I local minimum.*

In order to prove this theorem we will need two lemmas.

Lemma 100. *Let $\mathbf{x}_i, i \in [N]$ denote arbitrary points in $\mathbb{R}^d$ and $D_i, i \in [N]$ an arbitrary family of $m \times m$ diagonal matrices. Then*

$$\sum_{i=1}^{N} D_i A x_i = 0 \quad \textit{for all} \quad A \in \mathbb{M}_{m\times d} \quad \textit{if and only if}$$
$$\sum_{i=1}^{N} D_i(j,j) x_i = 0 \quad \textit{for all} \quad j \in [m]$$

Proof. Letting $A = [\mathbf{a}_1, \ldots, \mathbf{a}_d]$ and writing out the expression

$$\sum_{i=1}^{N} D_i A x_i = 0$$

column-wise shows that the first statement is equivalent to

$$\sum_{k=1}^{d} \left(\sum_{i=1}^{N} D_i x_i(k) \right) \mathbf{a}_k = 0 \quad \text{for all} \quad (\mathbf{a}_1, \ldots, \mathbf{a}_d) \in \mathbb{R}^d,$$

which in turn is equivalent to

$$\left(\sum_{i=1}^{N} D_i x_i(k) \right) = 0 \quad \text{for all} \quad k \in [d] \tag{108}$$

For each k the left hand side determines a diagonal matrix E_k, which vanishes if and only if all of its diagonal entries vanish. Thus (108) is equivalent to

$$\sum_{i=1}^{N} D_i(j,j) x_i(k) = 0 \quad \text{for all} \quad k \in [d], j \in [m],$$

which is exactly the conclusion of the lemma written component-by-component. □

Lemma 101. *Let Ω_u denote a cell on which the loss $\mathcal{L}(\boldsymbol{\omega})$ of a fully-connected network is constant. Let $\varepsilon_r^{(i,u)} \in \{0,1\}$ denote the error indicator for $\mathbf{x}^{(i)}$ and class $r \in [R]$ on the cell. Define*

$$\varepsilon^{(i,u)} := \sum_{r:\mathbf{y}_r^{(i)}=0} \varepsilon_r^{(i,u)}$$

as the total number of errors for $\mathbf{x}^{(i)}$ on the cell. Then there exist scalars $\lambda^{(i)} \in \{1, \alpha, \ldots, \alpha^L\}$ so that the equalities

$$\begin{aligned} \sum_{\mathbf{y}_r^{(i)}=1} \varepsilon^{(i,u)} \lambda^{(i)} \mu^{(i)} \mathbf{x}^{(i)} &= \sum_{\mathbf{y}_r^{(i)}=0} \varepsilon_r^{(i,u)} \lambda^{(i)} \mu^{(i)} \mathbf{x}^{(i)} \\ \sum_{\mathbf{y}_r^{(i)}=1} \varepsilon^{(i,u)} \lambda^{(i)} \mu^{(i)} &= \sum_{\mathbf{y}_r^{(i)}=0} \varepsilon_r^{(i,u)} \lambda^{(i)} \mu^{(i)} \\ \sum_{\mathbf{y}_r^{(i)}=1} \varepsilon^{(i,u)} \mu^{(i)} &= \sum_{\mathbf{y}_r^{(i)}=0} \varepsilon_r^{(i,u)} \mu^{(i)} \end{aligned} \tag{109}$$

hold for all $r \in [R]$.

Proof. The $\mathcal{L}|_{\Omega_u}$ is constant if and only if all of the multilinear forms in the decomposition of $\mathcal{L}|_{\Omega_u}$ vanish, that is, each of the multilinear forms is the zero function (c.f. the proof of theorem 1). Let $\mathrm{cl}(\mathbf{x}^{(i)})$ denote the class of a data point. Put

$$\boldsymbol{\mu}^{(i)} := (\mu^{(i)}, \ldots, \mu^{(i)}) \in \mathbb{R}^R \qquad\qquad T^{(i)} := \mathrm{Id} - \mathbf{1} \otimes \mathbf{y}^{(i)}$$

and for such a cell define the diagonal matrices

$$\begin{aligned} \varepsilon_r^{(i,u)} &:= \sigma'\Big(1 + \big(\langle \mathbf{v}_r - \mathbf{v}_{\mathrm{cl}(\mathbf{x}^{(i)})}, \mathbf{x}^{(i,L)} \rangle + c_r - c_{\mathrm{cl}(\mathbf{x}^{(i)})}\big)\Big) & \lambda_k^{(i,\ell,u)} &:= \sigma'_\alpha\Big(\langle \mathbf{a}_k^{(\ell)}, \mathbf{x}^{(i,\ell-1)} \rangle + \mathbf{b}_k^{(\ell)}\Big) \\ \mathcal{E}^{(i,u)} &:= \mathrm{diag}\big(\varepsilon_1^{(i,u)}, \ldots, \varepsilon_R^{(i,u)}\big) & \Lambda^{(i,\ell,u)} &:= \mathrm{diag}\big(\lambda_1^{(i,\ell,u)}, \ldots, \lambda_{d_\ell}^{(i,\ell,u)}\big). \end{aligned}$$

These diagonal matrices remain constant for all $\boldsymbol{\omega} \in \Omega_u$ by the definition of a cell. Set

$$\boldsymbol{\nu}^{(i,u)} := (T^{(i)})^T \mathcal{E}^{(i,u)} \boldsymbol{\mu}^{(i)}.$$

Then the multilinear forms in theorem 1 have the following form

$$\begin{aligned} \phi_0^u\big(A^{(1)}, \ldots, A^{(L)}, V\big) &= \sum_{i=1}^N \langle V \Lambda^{(i,L,u)} A^{(L)} \cdots A^{(2)} \Lambda^{(i,1,u)} A^{(1)} \mathbf{x}^{(i)}, \boldsymbol{\nu}^{(i,u)} \rangle \\ \phi_\ell^u\big(\mathbf{b}^{(\ell)}, A^{(\ell+1)}, \ldots, A^{(L)}, V\big) &= \sum_{i=1}^N \langle V \Lambda^{(i,L,u)} A^{(L)} \cdots A^{(\ell+1)} \Lambda^{(i,\ell,u)} \mathbf{b}^{(\ell)}, \boldsymbol{\nu}^{(i,u)} \rangle \\ \phi_L^u\big(\mathbf{b}^{(L)}, V\big) &= \sum_{i=1}^N \langle V \Lambda^{(i,L,u)} \mathbf{b}^{(L)}, \boldsymbol{\nu}^{(i,u)} \rangle \\ \phi_{L+1}^u(\mathbf{c}) &= \sum_{i=1}^N \langle \mathbf{c}, \boldsymbol{\nu}^{(i,u)} \rangle. \end{aligned}$$

Now ϕ_0^u vanishes for all $V, A^{(\ell)}$ if and only if

$$\sum_{i=1}^N \big(\Lambda^{(i,L,u)} A^{(L)} \cdots A^{(2)} \Lambda^{(i,1,u)} A^{(1)} \mathbf{x}^{(i)}\big) \otimes \boldsymbol{\nu}^{(i,u)} = 0$$

for all $A^{(\ell)}$. In other words, each of the R columns

$$\sum_{i=1}^N \big(\Lambda^{(i,L,u)} A^{(L)} \cdots A^{(2)} \Lambda^{(i,1,u)} A^{(1)} \mathbf{x}^{(i)}\big) \boldsymbol{\nu}_r^{(i,u)} = 0 \quad \text{for all} \quad r \in [R]$$

must vanish. Now $\Lambda^{(i,L,u)}$ is diagonal, and so lemma 100 shows this can happen if and only if

$$\sum_{i=1}^N \lambda_{k_L}^{(i,L,u)} \big(\Lambda^{(i,L-1,u)} A^{(L-1)} \cdots A^{(2)} \Lambda^{(i,1,u)} A^{(1)} \mathbf{x}^{(i)}\big) \boldsymbol{\nu}_r^{(i,u)} = 0 \quad \text{for all} \quad (r, k_L) \in [R] \times [d_L].$$

Now $\lambda_{k_L}^{(i,L,u)} \Lambda^{(i,L-1,u)}$ is once again diagonal, and so this can happen if and only if

$$\sum_{i=1}^N \lambda_{k_L}^{(i,L,u)} \lambda_{k_{L-1}}^{(i,L-1,u)} \big(\Lambda^{(i,L-2,u)} A^{(L-2)} \cdots A^{(2)} \Lambda^{(i,1,u)} A^{(1)} \mathbf{x}^{(i)}\big) \boldsymbol{\nu}_r^{(i,u)} = 0 \quad \text{for all} \quad (r, k_L, k_{L-1}) \in [R] \times [d_L] \times [d_{L-1}].$$

Continuing in this way shows ϕ_0^u vanishes if and only if

$$\sum_{i=1}^N \boldsymbol{\nu}_r^{(i,u)} \big(\lambda_{k_L}^{(i,L,u)} \lambda_{k_{L-1}}^{(i,L-1,u)} \cdots \lambda_{k_1}^{(i,1,u)}\big) \mathbf{x}^{(i)} = 0 \quad \text{for all} \quad (r, k_L, \ldots, k_1) \in [R] \times [d_L] \times \cdots \times [d_1].$$

A similar argument shows $\phi^u_\ell, \ell \in [L]$ vanish if and only if

$$\sum_{i=1}^{N} \boldsymbol{\nu}^{(i,u)}_r \lambda^{(i,L,u)}_{k_L} \cdots \lambda^{(i,\ell,u)}_{k_\ell} = 0 \quad \text{for all} \quad (r, k_L, \ldots, k_\ell) \in [R] \times [d_L] \times \cdots \times [d_\ell],$$

while ϕ^u_{L+1} vanishes if and only if

$$\sum_{i=1}^{N} \boldsymbol{\nu}^{(i,u)}_r = 0 \quad \text{for all} \quad r \in [R].$$

Pick some $(k_L, \ldots, k_1) \in [d_L] \times \cdots \times [d_1]$ arbitrary and set

$$\lambda^{(i)} := \lambda^{(i,L,u)}_{k_L} \cdots \lambda^{(i,1,u)}_{k_1} \in \{1, \alpha, \ldots, \alpha^L\}.$$

For any such a choice the fact that ϕ^u_0 vanishes implies the first equation in (109) due to the definition of $\boldsymbol{\nu}^{(i,u)}_r$, while the fact that ϕ^u_1 vanishes implies the second equation. Finally, that the third equation in (109) holds comes from the fact that ϕ^u_{L+1} vanishes. □

Proof of the theorem. The claims are immediate from the preceeding two lemmas. For generic data

$$\begin{aligned} \sum_{\mathbf{y}^{(i)}_r = 1} \varepsilon^{(i,u)} \lambda^{(i)} \mu^{(i)} \mathbf{x}^{(i)} &= \sum_{\mathbf{y}^{(i)}_r = 0} \varepsilon^{(i,u)}_r \lambda^{(i)} \mu^{(i)} \mathbf{x}^{(i)} \\ \sum_{\mathbf{y}^{(i)}_r = 1} \varepsilon^{(i,u)} \lambda^{(i)} \mu^{(i)} &= \sum_{\mathbf{y}^{(i)}_r = 0} \varepsilon^{(i,u)}_r \lambda^{(i)} \mu^{(i)} \\ \sum_{\mathbf{y}^{(i)}_r = 1} \varepsilon^{(i,u)} \mu^{(i)} &= \sum_{\mathbf{y}^{(i)}_r = 0} \varepsilon^{(i,u)}_r \mu^{(i)} \end{aligned}$$

can hold only if all of the coefficients vanish, i.e. $\varepsilon^{(i,u)}_r \lambda^{(i)} = 0$ for all $i \in [N], r \in [R]$. But $\lambda^{(i)} > 0$ for all $i \in [N]$ since $\alpha > 0$, and so $\varepsilon^{(i,u)}_r = 0$ for all $i \in [N], r \in [R]$. That is, $\mathcal{L}(\boldsymbol{\omega}) = 0$ on any flat cell Ω_u. □

Proofs of theorems from section 3: Critical Point Analysis

We begin by showing that no sub-optimal minimizers exist in the simplest case $\alpha = 1$, i.e. for deep linear networks with binary hinge loss. The loss here is

$$\sum \mu^{(i)} \sigma\left(1 - y^{(i)}\big(\langle \mathbf{v}, W^{(L)} \cdots W^{(1)} \mathbf{x}^{(i)} \rangle + c\big)\right). \tag{110}$$

Note that if we define $\bar{\mathbf{v}} := (W^{(L)} \cdots W^{(1)})^T \mathbf{v}$ then (110) corresponds to a convex loss $E(\bar{\mathbf{v}}, c)$ whose first argument is parametrized as a multilinear product.

Theorem 100 (Deep Linear Networks)**.** *Consider the loss* (110) *with arbitrary data and assume that $\boldsymbol{\omega}$ is any critical point in the Clarke sense. Then the following hold —*

(i) If $\bar{\mathbf{v}} \neq \mathbf{0}$ then $\boldsymbol{\omega}$ is a global minimum.

(ii) If the $\mu^{(i)}$ weight both classes equally weighted and $\boldsymbol{\omega}$ is a local minimum with $\bar{\mathbf{v}} = \mathbf{0}$ then it is a global minimum.

Proof. Recall that the loss takes the form

$$\mathcal{L}(\boldsymbol{\omega}) = \sum_{i=1}^{N} \mu^{(i)} \sigma\big(1 - y^{(i)}(\langle \mathbf{v}, \mathbf{x}^{(i,L)} \rangle + c)\big) \qquad y^{(i)} \in \{-1, 1\}$$

for a deep linear network, where

$$\mathbf{x}^{(i,L)} = W^{(L)} \cdots W^{(1)} \mathbf{x}^{(i)}$$

denote the features from the linear network at the L^{th} hidden layer. Let $W := W^{(L)} \cdots W^{(1)}$ and $\bar{\mathbf{v}} := W^T\mathbf{v}$, and recall that on a cell Ω_u the expression

$$\mathcal{L}|_{\overline{\Omega}_u}(\boldsymbol{\omega}) = \sum_{i=1}^{N} \mu^{(i)}\varepsilon^{(i,u)} - \sum_{i=1}^{N} \mu^{(i)}\varepsilon^{(i,u)} y^{(i)}\big(\langle \mathbf{v}, \mathbf{x}^{(i,L)}\rangle + c\big)$$

defines the loss. The expressions

$$\begin{aligned} -\nabla_{W^{(1)}}\mathcal{L}|_{\overline{\Omega}_u} &= \sum_{i=1}^{N} (W^{(L)}\cdots W^{(2)})^T\mathbf{v} \otimes \mathbf{x}^{(i)}\mu^{(i)}\varepsilon^{(i,u)}y^{(i)} \\ -\nabla_c\mathcal{L}|_{\overline{\Omega}_u} &= \sum_{i=1}^{N} \mu^{(i)}\varepsilon^{(i,u)}y^{(i)} \end{aligned}$$

then furnish the gradient of $\mathcal{L}$ with respect to the parameters $(W^{(1)}, c)$ on the cell. By definition, it therefore follows that

$$\begin{aligned} &\mathbf{0} = (W^{(L)}\cdots W^{(2)})^T\mathbf{v} \otimes \mathbf{z} \qquad \mathbf{z} := \sum_{i=1}^{N} \mu^{(i)}\lambda^{(i)}y^{(i)}\mathbf{x}^{(i)}, \quad \lambda^{(i)} := \sum_{u\in\mathcal{I}(\boldsymbol{\omega})} \theta^{(u)}\varepsilon^{(i,u)} \\ &0 = \sum_{i=1}^{N} \mu^{(i)}\lambda^{(i)}y^{(i)} \end{aligned}$$

a critical point, where the non-negative constants $\theta^{(u)} \geq 0$ sum to one. In particular, the coefficients $\lambda^{(i)} \in [0,1]$ lie in the unit interval. Multiplying by transpose of $W^{(1)}$ shows

$$\mathbf{0} = \bar{\mathbf{v}} \otimes \mathbf{z}$$

at a critical point. This gives a dichotomy — either $\mathbf{z}$ vanishes or $\bar{\mathbf{v}}$ vanishes.

Consider first the case where $\mathbf{z}$ vanishes. Then

$$\mathbf{0} = \sum_{i=1}^{N} \mu^{(i)}\lambda^{(i)}y^{(i)}\mathbf{x}^{(i)} \qquad \text{and} \qquad 0 = \sum_{i=1}^{N} \mu^{(i)}\lambda^{(i)}y^{(i)} \tag{111}$$

both hold. Moreover, for each $i \in [N]$ the coefficients $\lambda^{(i)}$ obey

$$\lambda^{(i)} \in \begin{cases} \{1\} & \text{if} \quad 1 - \big(\langle \bar{\mathbf{v}}, \mathbf{x}^{(i)}\rangle + c\big)y^{(i)} > 0 \\ [0,1] & \text{if} \quad 1 - \big(\langle \bar{\mathbf{v}}, \mathbf{x}^{(i)}\rangle + c\big)y^{(i)} = 0 \\ \{0\} & \text{if} \quad 1 - \big(\langle \bar{\mathbf{v}}, \mathbf{x}^{(i)}\rangle + c\big)y^{(i)} < 0 \end{cases} \tag{112}$$

as well. This follows from the observation that $\varepsilon^{(i,u)} = 1$ for all cells $u \in \mathcal{I}(\boldsymbol{\omega})$ in the first case, while $\varepsilon^{(i,u)} = 0$ for all cells $u \in \mathcal{I}(\boldsymbol{\omega})$ in the last case. Now for each $i \in [N]$ define the functions

$$f^{(i)}(\mathbf{w}, d) := \mu^{(i)}\sigma\big(1 - \big(\langle \mathbf{w}, \mathbf{x}^{(i)}\rangle + d\big)y^{(i)}\big),$$

which are clearly convex. The subdifferential $\partial f^{(i)}(\mathbf{w}, d)$ of $f^{(i)}(\mathbf{w}, d)$ at a point $(\mathbf{w}, d)$ is easily computed as

$$\partial f^{(i)}(\mathbf{w}, d) = \begin{cases} \{-\mu^{(i)}y^{(i)}(\mathbf{x}^{(i)}, 1)^T\} & \text{if} \quad 1 - \big(\langle \mathbf{w}, \mathbf{x}^{(i)}\rangle + d\big)y^{(i)} > 0 \\ -[0,1]\mu^{(i)}y^{(i)}(\mathbf{x}^{(i)}, 1)^T & \text{if} \quad 1 - \big(\langle \mathbf{w}, \mathbf{x}^{(i)}\rangle + d\big)y^{(i)} = 0 \\ \{\mathbf{0}\} & \text{if} \quad 1 - \big(\langle \mathbf{w}, \mathbf{x}^{(i)}\rangle + d\big)y^{(i)} < 0, \end{cases}$$

and so (111,112) imply that the inclusion

$$\mathbf{0} \in \sum_{i=1}^{N} \partial f^{(i)}(\bar{\mathbf{v}}, c)$$

holds. As each $f^{(i)}(\mathbf{w}, d)$ is Lipschitz, the composite function

$$E(\mathbf{w}, d) := \sum_{i=1}^{N} f^{(i)}(\mathbf{w}, d)$$

obeys the calculus rule

$$\partial E(\mathbf{w}, d) = \sum_{i=1}^{N} \partial f^{(i)}(\mathbf{w}, d)$$

(c.f. (Borwein & Lewis, 2010) theorem 3.3.5). Thus $\mathbf{0} \in \partial E(\bar{\mathbf{v}}, c)$ and so $(\bar{\mathbf{v}}, c)$ is a global minimizer.

It remains to address the case where $\bar{\mathbf{v}}$ vanishes. Then as a function of c the loss remains constant for all c in the unit interval,

$$E(\mathbf{0}, c) = \sum_{i=1}^{N} \mu^{(i)} \sigma\big(1 - y^{(i)} c\big) = \frac{\sigma(1-c) + \sigma(1+c)}{2} = 1,$$

and moreover if $c \notin [-1, 1]$ then the convex function $E(\mathbf{0}, c)$ attains its minimum in the unit interval. This follows from the equal mass hypothesis

$$\sum_{y^{(i)}=1} \mu^{(i)} = \sum_{y^{(i)}=-1} \mu^{(i)} = \frac{1}{2}$$

on the weights. At a local minimum $\boldsymbol{\omega}$ where $\bar{\mathbf{v}}$ vanishes the parameter c must therefore lie in the unit interval. It therefore suffices to assume that $c \in (-1, 1)$ without loss of generality. But then the loss $\mathcal{L}$ is differentiable (in fact smooth) near $\boldsymbol{\omega}$, and so theorem 1 from (Laurent & von Brecht, 2018) yields the result in this case. □

Theorem (Theorem 4 from the paper)**.** *Consider the loss (9) with $\alpha > 0$ and data $\mathbf{x}^{(i)}, i \in [N]$ that are linearly separable. Assume that $\boldsymbol{\omega} = (W, \mathbf{v}, \mathbf{b}, c)$ is any critical point of the loss in the Clarke sense. Then either $\mathbf{v} = \mathbf{0}$ or $\boldsymbol{\omega}$ is a global minimum.*

The proof of theorem 4 relies on the following pair of auxiliary lemmas. The former gives an explicit description of the multilinear decomposition for the loss in a network with one hidden layer; the latter computes the Clarke subdifferential in terms of the decomposition of parameter space into cells.

Lemma 102 (Decomposition with $L = 1$)**.** *Let*

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

denote the loss on a cell Ω_u. For $k \in [K]$ define

$$\mathbf{a}_k^{(u)} := \sum_i \mu^{(i)} y^{(i)} \varepsilon^{(i,u)} \lambda_k^{(i,u)} \mathbf{x}^{(i)} \qquad \alpha_k^{(u)} := \sum_i \mu^{(i)} y^{(i)} \varepsilon^{(i,u)} \lambda_k^{(i,u)}$$
$$\gamma^{(u)} := \sum_i \mu^{(i)} y^{(i)} \varepsilon^{(i,u)} \qquad \delta^{(u)} := \sum_i \mu^{(i)} \varepsilon^{(i,u)}.$$

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

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

furnish the multilinear forms defining the loss on Ω_u.

Proof. Restricted to a cell Ω_u the loss can be written as

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

Expanding this expression leads to

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

Now let $\mathbf{w}_k$ denote the k^{th} row of the matrix W and note that $\mathbf{v}^T\Lambda^{(i,u)}W = \sum_k v_k\lambda_k^{(i,u)}\mathbf{w}_k^T$ to find

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

A similar argument reveals

$$\begin{aligned}\phi_1^{(u)}(\mathbf{b},\mathbf{v}) &= -\sum_i \mu^{(i)}\varepsilon^{(i,u)}y^{(i)}\mathbf{v}^T\Lambda^{(i,u)}\mathbf{b}\\ &= -\sum_i \mu^{(i)}\varepsilon^{(i,u)}y^{(i)}\sum_k\Big(v_k\lambda_k^{(i,u)}b_k\Big)\\ &= -\sum_k v_k\left(\sum_i \mu^{(i)}\varepsilon^{(i,u)}y^{(i)}\lambda_k^{(i,u)}\right)b_k\\ &= -\sum_k v_k\alpha_k^{(u)}b_k\\ \phi_2^{(u)}(c) &= -\gamma^{(u)}c \qquad \gamma^{(u)} := -\sum_i \mu^{(i)}\varepsilon^{(i,u)}y^{(i)}\end{aligned}$$

as claimed. □

Lemma 103 (Subdifferential Calculation)**.** *Fix a point $\boldsymbol{\omega}\in\Omega$ and let $\mathcal{I}(\boldsymbol{\omega})$ denote its incidence set. Then the convex hull*

$$\partial_0\mathcal{L}(\boldsymbol{\omega}) = \left\{\sum_{u\in\mathcal{I}(\boldsymbol{\omega})}\theta^{(u)}\nabla\mathcal{L}|_{\overline{\Omega}_u}(\boldsymbol{\omega}) : \theta^{(u)}\geq 0,\ \sum_u\theta^{(u)} = 1\right\} \tag{113}$$

is the Clarke subdifferential of the loss $\mathcal{L}$ at $\boldsymbol{\omega}$. In particular, if $\mathcal{I}(\boldsymbol{\omega}) = \{u\}$ is a singleton then $\partial_0\mathcal{L}(\boldsymbol{\omega})$ is single-valued: $\partial_0\mathcal{L}(\boldsymbol{\omega}) = \{\nabla\mathcal{L}|_{\Omega_u}(\boldsymbol{\omega})\}$.

Proof. For a given point $\boldsymbol{\omega}\in\Omega$ recall that $\mathcal{I}(\boldsymbol{\omega})$ denotes the set of indices of the cells that are adjacent to the point $\boldsymbol{\omega}$,

$$\mathcal{I}(\boldsymbol{\omega}) := \{u\in\{-1,1\}^{ND} : \boldsymbol{\omega}\in\overline{\Omega}_u\}$$

where $\overline{\Omega}_u$ stands for the closure of the cell Ω_u. Assume $\boldsymbol{\omega}_k\to\boldsymbol{\omega}$ and $\boldsymbol{\omega}_k\notin\mathcal{N}$. As $\mathcal{I}(\boldsymbol{\omega})$ is clearly finite it suffices to assume, by passing to a subsequence if necessary, 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) the limit $\nabla\mathcal{L}(\boldsymbol{\omega}_k)\to\nabla\mathcal{L}|_{\overline{\Omega}_u}(\boldsymbol{\omega})$ follows. As $\mathcal{N}$ has measure zero definition 2 reveals that the convex hull

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

is the Clark subdifferential at $\boldsymbol{\omega}$. In particular, if $\boldsymbol{\omega} \in \Omega_u$ for some u then $\mathcal{I}(\boldsymbol{\omega}) = \{u\}$ and (113) simply becomes $\partial_0 \mathcal{L}(\boldsymbol{\omega}) = \{\nabla \mathcal{L}|_{\Omega_u}(\boldsymbol{\omega})\}$ as expected. $\square$

Proof of Theorem 6. By assumption $\mathbf{0} \in \partial_0 \mathcal{L}(\boldsymbol{\omega})$ and so

$$\mathbf{0} = \sum_{u \in \mathcal{I}(\boldsymbol{\omega})} \theta^{(u)} \nabla \mathcal{L}|_{\overline{\Omega}_u}(\boldsymbol{\omega}) \tag{114}$$

for some collection of positive coefficients $\theta^{(u)}$ due to the characterization (113) of the subdifferential. The explicit formulas from lemma 102 show

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

which by (114) then obviously implies that both

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

must hold for all k. Substituting the expressions $\mathbf{a}_k^{(u)}$ and b_k provided in lemma 2 then gives the equalities

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

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

If there exists a k for which $v_k \neq 0$ then an interchange of summations reveals

$$\sum_{i: y^{(i)}=1} \varrho_k^{(i)} \mathbf{x}^{(i)} = \sum_{i: y^{(i)}=-1} \varrho_k^{(i)} \mathbf{x}^{(i)} \tag{115}$$

$$\sum_{i: y^{(i)}=1} \varrho_k^{(i)} = \sum_{i: y^{(i)}=-1} \varrho_k^{(i)} \qquad \text{where} \qquad \varrho_k^{(i)} := \sum_u \theta^{(u)} \mu^{(i)} \varepsilon^{(i,u)} \lambda_k^{(i,u)} \tag{116}$$

The claim then follows since (115,116) cannot hold unless all the $\varrho_k^{(i)}$ vanish. To see this, note that if the $\varrho_k^{(i)}$ do not vanish then

$$0 < Q := \sum_{i: y^{(i)}=1} \varrho_k^{(i)} = \sum_{i: y^{(i)}=-1} \varrho_k^{(i)},$$

and upon dividing by Q the equality (115) implies

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

In other words a convex combination of data points of class 1 equals a convex combination of data points of class -1, a contradiction since the data points are linearly separable.

The theorem then easily follows. If $v_k \neq 0$ for some k then all $\varrho_k^{(i)}$ must vanish. But $\mu^{(i)} > 0$ (by definition) and $\lambda_k^{(i,u)} > 0$ (since $\alpha > 0$). The $\theta^{(u)}$'s are nonnegative and sum to 1 and so at least one of them is strictly positive, say $\theta^{(u_0)} > 0$. By (116) it is then clear that $\varrho_k^{(i)} = 0$ for all $i \in [N]$ if and only if $\varepsilon^{(i,u_0)} = 0$ for all $i \in [N]$. Recall by definition that

$$\varepsilon^{(i,u_0)} = \sigma' \left(1 - y^{(i)} \hat{y}^{(i)} \right)$$

inside the cell Ω_{u_0}, and so $\varepsilon^{(i,u_0)} = 0$ for all $i \in [N]$ implies that

$$0 = \sigma \left(1 - y^{(i)} \hat{y}^{(i)} \right)$$

for all $i \in [N]$ as well. Thus $\mathcal{L}|_{\Omega_{u_0}} = 0$, and since $\boldsymbol{\omega} \in \overline{\Omega}_{u_0}$ the continuity of the loss implies $\mathcal{L}(\boldsymbol{\omega}) = 0$ as well. $\square$

Theorem (Theorem 5 from the paper)**.** *Consider the loss (9) with $\alpha > 0$ and data $\mathbf{x}^{(i)}, i \in [N]$ that are linearly separable. Assume that the $\mu^{(i)}$ weight both classes equally. Then every local minimum of $\mathcal{L}(\boldsymbol{\omega})$ is a global minimum.*

The proof of theorem 5 relies on the following four auxiliary lemmas.

Lemma 104. *Let $\mathbb{R} = I_1 \cup \ldots \cup I_N$ be a partition of the real line into a finite number of non-empty intervals. Let $f(t)$ be a function defined by*

$$f(t) = P_j(t) \quad \text{if } t \in I_j,$$

where the $P_1(t), \ldots, P_N(t)$ are polynomials. Then there exists $t_0 > 0$ such that the function $t \mapsto \operatorname{sign}(f(t))$ is constant on $(0, t_0)$.

Proof. First note that there exists a t^* such that the interval $(0, t^*)$ is contained in one of the intervals I_j. On this interval $(0, t^*)$ the function $f(t)$ is simply the polynomial $P_j(t)$. If $P_j(t)$ is the zero polynomial, then $\operatorname{sign}(f(t)) = 0$ for all $t \in (0, t^*)$ and choosing $\tau = t^*$ leads to the claimed result. If $P_j(t)$ is a non-trivial polynomial, it has either no roots on $(0, t^*)$ or a finite number of roots on $(0, t^*)$. In the first case $\operatorname{sign}(f(t))$ is clearly constant on $(0, t^*)$ and so choosing $\tau = t^*$ gives the claim. In the second case simply choose τ to be the first root of $P_j(t)$ that is larger than 0. □

Before turning to the remaining three auxiliary lemmas it is beneficial to recall the decomposition

$$\mathcal{L}|_{\Omega_u}(W, \mathbf{v}, \mathbf{b}, c) = \phi_0^u(W, \mathbf{v}) + \phi_1^u(\mathbf{b}, \mathbf{v}) + \phi_2^u(c) + \phi_3^u \tag{117}$$

$$\phi_0^u(W, \mathbf{v}) = -\sum_k v_k \langle \mathbf{a}_k^{(u)}, \mathbf{w}_k \rangle, \quad \phi_1^u(\mathbf{b}, \mathbf{v}) = -\sum_k v_k \alpha_k^{(u)} b_k, \quad \phi_2^u(c) = -\gamma^{(u)} c, \tag{118}$$

for the loss on a cell, as well as the constants

$$\mathbf{a}_k^{(u)} := \sum_i \mu^{(i)} \varepsilon^{(i,u)} \lambda_k^{(i,u)} y^{(i)} \mathbf{x}^{(i)} \qquad \alpha_k^{(u)} := \sum_i \mu^{(i)} \varepsilon^{(i,u)} \lambda_k^{(i,u)} y^{(i)} \tag{119}$$

$$\gamma^{(u)} := \sum_i \mu^{(i)} y^{(i)} \varepsilon^{(i,u)} \qquad \phi_3^u := \sum_i \mu^{(i)} \varepsilon^{(i,u)}. \tag{120}$$

used to define the decomposition. By assumption the data $\mathbf{x}^{(i)}, i \in [N]$ are linearly separable and so there exists a unit vector $\mathbf{q} \in \mathbb{R}^d$, a bias $\beta \in \mathbb{R}$ and a margin $m > 0$ such that the family of inequalities

$$\langle\ \mathbf{q}\ ,\ y^{(i)} \mathbf{x}^{(i)}\ \rangle + \beta y^{(i)} \geq m \tag{121}$$

hold. Combining (119) with (121) gives the estimate

$$\langle \mathbf{q}, \mathbf{a}_k^{(u)} \rangle + \beta \alpha_k^{(u)} \geq m \sum_i \mu^{(i)} \varepsilon^{(i,u)} \lambda_k^{(i,u)}, \tag{122}$$

that will be used repeatedly when proving the remaining auxiliary lemmas.

Lemma 105 (First perturbation)**.** *Let $\boldsymbol{\omega} = (W, \mathbf{v}, \mathbf{b}, c) \in \Omega$ denote any point in the parameter space. Define*

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

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

(i) There exists $t_0 > 0$ and $u \in \mathcal{I}(\boldsymbol{\omega})$ such that $\boldsymbol{\omega}(t) \in \overline{\Omega}_u$ for all $t \in [0, t_0)$.

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

Proof. To prove (i) let $\boldsymbol{\omega}(t) = (W + t\tilde{W}, \mathbf{v}, \mathbf{b} + t\tilde{\mathbf{b}}, c) = (W(t), \mathbf{v}, \mathbf{b}(t), c)$ denote the perturbation considered in the lemma. Without loss of generality, it suffices to consider the case $k = 1$. 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 constants $A_k^{(i)}$ and $B_k^{(i)}$ as

$$\begin{aligned}
&\langle \mathbf{w}_1(t), \mathbf{x}^{(i)}\rangle + b_1(t) = \langle \mathbf{w}_1, \mathbf{x}^{(i)}\rangle + b_1 + t\mathrm{sign}(v_1)\left(\langle \mathbf{q}, \mathbf{x}^{(i)}\rangle + \beta\right) = A_1^{(i)} + B_1^{(i)}t \\
&\langle \mathbf{w}_k, \mathbf{x}^{(i)}\rangle + b_k = A_k^{(i)} \qquad \text{for } k \geq 2,
\end{aligned}$$

so that the signature functions can be written as:

$$\begin{aligned}
&\mathbf{s}_1^{(i,1)}(\boldsymbol{\omega}(t)) = \mathrm{sign}(A_1^{(i)} + B_1^{(i)}t) \\
&\mathbf{s}_k^{(i,1)}(\boldsymbol{\omega}(t)) = \mathrm{sign}(A_k^{(i)}) \qquad \text{for } k \geq 2 \\
&\mathbf{s}^{(i,2)}(\boldsymbol{\omega}(t)) = \mathrm{sign}\left[1 - y^{(i)}\left\{c + v_1\sigma_\alpha(A_1^{(i)} + B_1^{(i)}t)) + \sum_{k=2}^{K} v_k\sigma_\alpha(A_k^{(i)})\right\}\right].
\end{aligned}$$

The functions appearing inside the sign functions are clearly piecewise defined polynomials, and therefore lemma 104 implies that there exists $t_0 > 0$ such that $t \mapsto \mathcal{S}(\boldsymbol{\omega}(t))$ is constant on $(0, t_0)$. This 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 0) or on the boundary of a fixed cell Ω_u (if some of the entries of $\mathcal{S}(\boldsymbol{\omega}(t))$ are equal to 0). 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, $\boldsymbol{\omega}(t) \in \overline{\Omega_u}$ for all $t \in [0, t_0)$ and so (i) holds. To prove (ii), first note that due to the continuity of the loss, equality (117) holds not only for $\boldsymbol{\omega} \in \Omega_u$, but also for any $\boldsymbol{\omega} \in \overline{\Omega}_u$. By part (i), $\boldsymbol{\omega}(t)$ remains in some fixed $\overline{\Omega}_u$ for all t small enough. Thus (117-118) 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 (122) proves (ii). □

Lemma 106 (Second perturbation)**.** *Let $\boldsymbol{\omega} = (W, \mathbf{v}, \mathbf{b}, c) \in \Omega$ denote a point in the parameter space. Assume that $\mathbf{v} = \mathbf{0}$ and $c \notin \{-1, +1\}$. Define*

$$\tilde{\mathbf{v}} = \mathbf{e}_k.$$

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

(i) There exists $t_0 > 0$ such that $\boldsymbol{\omega}(t) \in \overline{\Omega}_u$ for all $u \in \mathcal{I}(\boldsymbol{\omega})$ and all $t \in (-t_0, t_0)$.

(ii) $\mathcal{L}(\boldsymbol{\omega}(t)) - \mathcal{L}(\boldsymbol{\omega}) = -t\left(\langle \mathbf{a}_k^{(u)}, \mathbf{w}_k\rangle - \alpha_k^u b_k\right)$ for all $t \in (-t_0, t_0)$ and all $u \in \mathcal{I}(\boldsymbol{\omega})$.

Proof. To prove (i), note that $\mathbf{v} = \mathbf{0}$ implies the equalities

$$\begin{aligned}
&\mathbf{s}^{(i,1)}(\boldsymbol{\omega}(t)) = \mathrm{sign}(W\mathbf{x}^{(i)} + \mathbf{b}) \\
&\mathbf{s}^{(i,2)}(\boldsymbol{\omega}(t)) = \mathrm{sign}\left[1 - cy^{(i)} - ty^{(i)}\tilde{\mathbf{v}}^T\sigma_\alpha(W\mathbf{x}^{(i)} + \mathbf{b})\right]
\end{aligned}$$

for the signature function. But then $1 - cy^{(i)} \neq 0$ since $c \notin \{+1, -1\}$, and so there exists an interval $(-t_0, t_0)$ on which all the functions $\mathbf{s}^{(i,2)}(\boldsymbol{\omega}(t))$ do not change. Obviously the functions $\mathbf{s}^{(i,1)}(\boldsymbol{\omega}_2(t))$ do not change as well since W and $\mathbf{b}$ are not perturbed. So the signature $\mathcal{S}(\boldsymbol{\omega}(t))$ does not change on $(-t_0, t_0)$, which yields (i). To prove (ii), choose an arbitrary $u \in \mathcal{I}(\boldsymbol{\omega})$. Since $\boldsymbol{\omega}(t)$ remains in $\overline{\Omega}_u$ for all $t \in (-t_0, t_0)$ the relations (117)-(118) imply

$$\mathcal{L}(\boldsymbol{\omega}(t)) - \mathcal{L}(\boldsymbol{\omega}^*) = t\left(\phi_0^u(W, \tilde{\mathbf{v}}) + \phi_1^u(\mathbf{b}, \tilde{\mathbf{v}})\right) = -t\left(\langle \mathbf{a}_k^{(u)}, \mathbf{w}_k\rangle - \alpha_k^u b_k\right)$$

for all $t \in (-t_0, t_0)$, which is the desired result. □

Lemma 107 (Third perturbation)**.** *Let $\boldsymbol{\omega} = (W, \mathbf{v}, \mathbf{b}, c) \in \Omega$ denote a point in the parameter space. Assume that $\mathbf{v} = \mathbf{0}$. Define*

$$\tilde{W} = \mathbf{e}_k \otimes \mathbf{q}, \qquad \tilde{\mathbf{v}} = \mathbf{e}_k \qquad \text{and} \qquad \tilde{\mathbf{b}} = \beta\mathbf{e}_k.$$

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

(i) There exists $t_0 > 0$ and $u \in \mathcal{I}(\boldsymbol{\omega})$ such that $\boldsymbol{\omega}(t) \in \overline{\Omega}_u$ for all $t \in [0, t_0)$.

(ii) $\mathcal{L}(\boldsymbol{\omega}) \geq \mathcal{L}(\boldsymbol{\omega}(t)) + t\left(\langle \mathbf{a}_k^{(u)}, \mathbf{w}_k\rangle + \alpha_k^{(u)} b_k\right) + t^2 m \sum_i \mu^{(i)} \varepsilon^{(i,u)} \lambda_k^{(i,u)}$ *for all* $t \in [0, t_0)$.

Proof. To prove (i) let $\boldsymbol{\omega}(t) = (W + t\tilde{W}, \mathbf{v} + t\tilde{\mathbf{v}}, \mathbf{b} + t\tilde{\mathbf{b}}, c) = (W(t), \mathbf{v}(t), \mathbf{b}(t), c)$ denote the perturbation considered in the lemma. Without loss of generality, if suffices to consider the case $k = 1$. Define the constants $A_k^{(i)}$ and $B_k^{(i)}$ as

$$\begin{aligned}\langle \mathbf{w}_1(t), \mathbf{x}^{(i)}\rangle + b_1(t) &= \langle \mathbf{w}_1, \mathbf{x}^{(i)}\rangle + b_1 + t\left(\langle \mathbf{q}, \mathbf{x}^{(i)}\rangle + \beta\right) = A_1^{(i)} + B_1^{(i)} t \\ \langle \mathbf{w}_k(t), \mathbf{x}^{(i)}\rangle + b_k(t) &= \langle \mathbf{w}_k, \mathbf{x}^{(i)}\rangle + b_k = A_k^{(i)} \qquad \text{for } k \geq 2.\end{aligned}$$

The fact that $\mathbf{v}(t) = t\mathbf{e}_1$ then gives

$$\begin{aligned}\mathbf{s}_1^{(i,1)}(\boldsymbol{\omega}(t)) &= \operatorname{sign}(A_1^{(i)} + B_1^{(i)} t) \\ \mathbf{s}_k^{(i,1)}(\boldsymbol{\omega}(t)) &= \operatorname{sign}(A_k^{(i)}) \qquad \text{for } k \geq 2 \\ \mathbf{s}^{(i,2)}(\boldsymbol{\omega}(t)) &= \operatorname{sign}\Big[\, 1 - y^{(i)}\Big\{c + t\sigma_\alpha(A_1^{(i)} + B_1^{(i)} t)\Big\}\Big]\end{aligned}$$

for the signature functions. As in the proof of part (i) of lemma 105, the arguments of the sign functions are piecewise defined polynomials and so lemma 104 gives the claim. To prove (ii), note that since $\boldsymbol{\omega}(t)$ remains in a fixed cell $\overline{\Omega}_u$ for all $t \in (0, t_0)$ the formulas (117)-(118) apply. Expanding the bilinear forms gives

$$\mathcal{L}(\boldsymbol{\omega}(t)) \;-\; \mathcal{L}(\boldsymbol{\omega}) \;=\; t\Big(\phi_0^u(\tilde{W}, \mathbf{v}) \;+\; \phi_0^u(W, \tilde{\mathbf{v}}) \;+\; \phi_1^u(\tilde{\mathbf{b}}, \mathbf{v}) \;+\; \phi_1^u(\mathbf{b}, \tilde{\mathbf{v}})\Big) \;+\; t^2\Big(\phi_0^u(\tilde{W}, \tilde{\mathbf{v}}) \;+\; \phi_1^u(\tilde{\mathbf{b}}, \tilde{\mathbf{v}})\Big),$$

and $\mathbf{v} = \mathbf{0}$ the first order terms are

$$\phi_0^u(W, \tilde{\mathbf{v}}) + \phi_1^u(\mathbf{b}, \tilde{\mathbf{v}}) = -\langle \mathbf{a}_k^{(u)}, \mathbf{w}_k\rangle - \alpha_k^u b_k.$$

Applying (122) then yields

$$\phi_0^u(\tilde{W}, \tilde{\mathbf{v}}) + \phi_1^u(\tilde{\mathbf{b}}, \tilde{\mathbf{v}}) = -\langle \mathbf{a}_k^{(u)}, \mathbf{q}\rangle - \alpha_k^{(u)} \beta \leq -m \sum_i \mu^{(i)} \varepsilon^{(i,u)} \lambda_k^{(i,u)}$$

for second order terms, giving the claim. □

Proof of Theorem 5. The proof is in two steps. The first step shows that a sub-optimal local minimizer must necessarily be of the form $\boldsymbol{\omega} = (W, \mathbf{0}, \mathbf{b}, \pm 1)$, while the second step shows that such a sub-optimal minimizer cannot exist if the two classes are equally weighted.

STEP 1: Assume that $\boldsymbol{\omega} = (W, \mathbf{v}, \mathbf{b}, c) \in \Omega$ is a sub-optimal local minimum. Then $\ell^{(i)}(\boldsymbol{\omega}) > 0$ for some data point $\mathbf{x}^{(i)}$. Take an arbitrary $u \in \mathcal{I}(\boldsymbol{\omega})$. By continuity of the loss, there exists $\hat{\boldsymbol{\omega}} \in \Omega_u$ such that $\ell^{(i)}(\hat{\boldsymbol{\omega}}) > 0$, and, as a consequence, $\varepsilon^{(i,u)} = 1$. Thus

$$\varepsilon^{(i,u)} = 1 \qquad \text{for all} \quad u \in \mathcal{I}(\boldsymbol{\omega}). \tag{123}$$

since u was arbitrary. Now choose an arbitrary $k \in [K]$ and consider the perturbation $\boldsymbol{\omega}(t) := (W + t\tilde{W}, \mathbf{v}, \mathbf{b} + t\tilde{\mathbf{b}}, c)$ described in lemma 105. As $\alpha > 0$ the $\lambda_k^{(i,u)}$ are all strictly positive. By (123), the term $\sum_i \mu^{(i)} \varepsilon^{(i,u)} \lambda_k^{(i,u)}$ appearing in statement (ii) of lemma 105 is strictly positive as well. Since $\boldsymbol{\omega}$ is a local minimum, v_k must necessary be equal to zero, otherwise the considered perturbation would lead to a strict decrease of the loss. Thus $\mathbf{v} = \mathbf{0}$ since k was arbitrary. Assume that $c \notin \{-1, +1\}$ for the sake of contradiction. The perturbation described in lemma 106 gives

$$\langle \mathbf{a}_k^{(u)}, \mathbf{w}_k\rangle + \alpha_k^{(u)} b_k = 0 \qquad \text{for all} \quad u \in \mathcal{I}(\boldsymbol{\omega}), \tag{124}$$

which combines with (123), (124) and the perturbation described in lemma (107) to give a strict decrease in the loss. This contradicts the fact that $\boldsymbol{\omega}$ is a local minimum, and so in fact $c \in \{-1, 1\}$.

STEP 2. By step 1 a sub-optimal local minimizer must be of the form $\boldsymbol{\omega} = (W, \mathbf{0}, \mathbf{b}, \pm 1)$. Assume $c = 1$, as the argument for the case $c = -1$ is similar. Thus $\boldsymbol{\omega} = (W, \mathbf{0}, \mathbf{b}, 1)$. Consider the perturbation $\boldsymbol{\omega}(t) = (W, \mathbf{0}, \mathbf{b}, 1 - t)$. For $t \in [0, 2]$ it then follows that

$$
\begin{aligned}
\mathcal{L}(\boldsymbol{\omega}(t)) &= \sum_i \mu^{(i)}\sigma\left(1 - y^{(i)} + y^{(i)}t\right) \\
&= \sum_{i:y^{(i)}=1} \mu^{(i)}\sigma(t) + \sum_{i:y^{(i)}=-1} \mu^{(i)}\sigma(2-t) \\
&= \sum_{i:y^{(i)}=1} \mu^{(i)}t + \sum_{i:y^{(i)}=-1} \mu^{(i)}(2-t) \\
&= 2\sum_{i:y^{(i)}=-1} \mu^{(i)} = 1
\end{aligned}
$$

where the equal mass hypothesis

$$
\sum_{i:y^{(i)}=1} \mu^{(i)} = \sum_{i:y^{(i)}=-1} \mu^{(i)} = \frac{1}{2}
$$

justifies the last two equalities. Therefore $\mathcal{L}(\boldsymbol{\omega}) = \mathcal{L}(\boldsymbol{\omega}(t)) = 1$ for t small enough. But if $t \neq 0$ then $\boldsymbol{\omega}(t)$ cannot be a local minimizer by stem 1. Thus the point $\boldsymbol{\omega} = (W, \mathbf{0}, \mathbf{b}, 1) \in \Omega$ has arbitrarily close neighbors $\boldsymbol{\omega}(t) \in \Omega$ that have same loss and that are not local minima. This implies that $\boldsymbol{\omega}$ can not be a local minimum. □

Theorem (Theorem 6 from the paper)**.** *Consider the loss (9) with $\alpha = 0$ and data $\mathbf{x}^{(i)}, i \in [N]$ that are linearly separable. Assume that $\boldsymbol{\omega} = (W, \mathbf{v}, \mathbf{b}, c)$ is a critical point in the Clarke sense, and that $\mathbf{x}^{(i)}$ is any data point that contributes a nonzero value to the loss. Then for each hidden neuron $k \in [K]$ either*

$$
\textit{(i)}\ \langle \mathbf{w}_k, \mathbf{x}^{(i)}\rangle + b_k \leq 0, \quad \textit{or} \quad \textit{(ii)}\ v_k = 0.
$$

Proof. The proof of theorem 6 shows

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

whenever $\boldsymbol{\omega}$ is a critical point. If $\ell^{(i)}(\boldsymbol{\omega}) > 0$ for some data point $\mathbf{x}^{(i)}$ then $\ell^{(i)} > 0$ on each neighboring cell Ω_u, $u \in \mathcal{I}(\boldsymbol{\omega})$ by continuity of the loss. This implies that $\varepsilon^{(i,u)} = 1$ for all $u \in \mathcal{I}(\boldsymbol{\omega})$. If $v_k \neq 0$ for some k then

$$
\begin{aligned}
\sum_{i:y^{(i)}=1} \varrho_k^{(i)}\mathbf{x}^{(i)} &= \sum_{i:y^{(i)}=-1} \varrho_k^{(i)}\mathbf{x}^{(i)} \\
\sum_{i:y^{(i)}=1} \varrho_k^{(i)} &= \sum_{i:y^{(i)}=-1} \varrho_k^{(i)} \qquad \text{where} \qquad \varrho_k^{(i)} := \sum_u \theta^{(u)}\mu^{(i)}\varepsilon^{(i,u)}\lambda_k^{(i,u)}
\end{aligned}
$$

and the corresponding $\varrho_k^{(i)}$ must all vanish since the data $\mathbf{x}^{(i)}, i \in [N]$ are separable. If $\ell^{(i)} > 0$ this necessarily implies that $\lambda_k^{(i,u)} = 0$ for some u since the $\varepsilon^{(i,u)} = 1$ and at least one $\theta^{(u)}$ does not vanish. This in turn implies $\sigma(\langle \mathbf{w}_k, \mathbf{x}^{(i)}\rangle + b_k) = 0$ due to the definition of the $\lambda_k^{(i)}$. □

Proofs of theorems from section 4: Exact Penalties and Multi-Class Structure

The proof of theorem 7 requires modifying the notion of a cell. This modification is straightforward; it simply accounts for the fact that the penalized loss

$$
\mathcal{E}_\gamma\big(\boldsymbol{\omega}^{(1)}, \ldots, \boldsymbol{\omega}^{(R)}\big) := \sum_{r=1}^{R} \mathcal{L}^{(r)}\big(\boldsymbol{\omega}^{(r)}\big) + \gamma\mathcal{R}\big(\breve{\boldsymbol{\omega}}^{(1)}, \ldots, \breve{\boldsymbol{\omega}}^{(R)}\big) \tag{125}
$$

has the R-fold Cartesian product $\Omega \times \cdots \times \Omega$ as its parameter domain. The notion of a cell Ω_u for the model (125) consists of sets (Cartesian products) of the form

$$\Omega_u = \Omega_{u^{(1)}} \times \Omega_{u^{(2)}} \times \cdots \times \Omega_{u^{(R)}}, \tag{126}$$

where each $u^{(r)} \in \{-1, 1\}^{ND}$ denotes a signature for the individual two-class losses. Thus a binary vector of the form

$$u = \big(u^{(1)}, \ldots, u^{(R)}\big) \in \{-1, 1\}^{NDR} \qquad u^{(r)} \in \{-1, 1\}^{ND}$$

defines a signature for the full model. That sets of the form (126) cover the product space $\Omega \times \cdots \times \Omega$ 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. The proof also relies following auxiliary lemma.

Lemma 108. *For any R vectors $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(R)} \in \mathbb{R}^d$, if*

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

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

Proof. By relabelling if necessary, it suffices to assume $\mathbf{x}^{(1)}$ has largest norm amongst the $\mathbf{x}^{(r)}$. 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. □

Theorem (Theorem 7 from the paper)**.** *If $\gamma > 0$ then the following hold for* (125) —

(i) The penalty is exact, that is, at ***any*** *critical point $\big(\boldsymbol{\omega}^{(1)}, \ldots, \boldsymbol{\omega}^{(R)}\big)$ of $\mathcal{E}_\gamma$ 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 $\ell \in [L]$.

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

$$\mathbf{0} \in \partial_0 \mathcal{L}^{(r)}(\breve{\boldsymbol{\omega}}, \mathbf{v}_r, c_r) \tag{127}$$

hold for all $r \in [R]$.

Proof. Let $\big(\boldsymbol{\omega}^{(1)}, \ldots, \boldsymbol{\omega}^{(R)}\big)$ denote any critical point. For each (ℓ, r) the equalities

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

follow from a 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

$$\begin{aligned}
\mathbf{0} &= \sum_u \theta^{(u)} \nabla_{\mathbf{v}_r} \bar{\mathcal{L}}|_{\Omega_u} \\
\mathbf{0} &= \sum_u \theta^{(u)} \left(\nabla_{\omega^{(\ell,r)}} \bar{\mathcal{L}}|_{\Omega_u} + \nabla_{\omega^{(\ell,r)}} \mathcal{R}\right) = \gamma\big(\omega^{(\ell,r)} - \tilde{\omega}^{(\ell,r)}\big) + \sum_u \theta^{(u)} \nabla_{\omega^{(\ell,r)}} \bar{\mathcal{L}}|_{\Omega_u} \\
\mathbf{0} &= \sum_u \theta^{(u)} \left(\nabla_{\mathbf{b}^{(\ell,r)}} \bar{\mathcal{L}}|_{\Omega_u} + \nabla_{\mathbf{b}^{(\ell,r)}} \mathcal{R}\right) = \gamma\big(\mathbf{b}^{(\ell,r)} - \tilde{\mathbf{b}}^{(\ell,r)}\big) + \sum_u \theta^{(u)} \nabla_{\mathbf{b}^{(\ell,r)}} \bar{\mathcal{L}}|_{\Omega_u}
\end{aligned} \tag{128}$$

hold for all $\ell \in [L]$ and $r \in [R]$, 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 Ω_u the loss $\mathcal{L}^{(r)}$ decomposes 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 $\phi(\mathbf{v}_1, \ldots, \mathbf{v}_n)$ the equality

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

holds for all $k \in [n]$ by Euler's theorem for homogeneous functions. Taking the inner-product of (128) with $\mathbf{v}_r, \omega^{(L,r)}$ and $\mathbf{b}^{(L,r)}$ then shows

$$\begin{aligned}
0 &= \sum_u \theta^{(u)}\big(\phi_0^{(u,r)} + \cdots + \phi_L^{(u,r)}\big) \\
0 &= \sum_u \theta^{(u)}\big(\phi_0^{(u,r)} + \cdots + \phi_{L-1}^{(u,r)}\big) + \gamma\big(\|\omega^{(L,r)}\|^2 - \langle \omega^{(L,r)}, \tilde{\omega}^{(L,r)} \rangle\big) \\
0 &= \sum_u \theta^{(u)}\big(\phi_L^{(u,r)}\big) + \gamma\big(\|\mathbf{b}^{(L,r)}\|^2 - \langle \mathbf{b}^{(L,r)}, \tilde{\mathbf{b}}^{(L,r)} \rangle\big)
\end{aligned} \tag{129}$$

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 $r \in [R]$. By the definitions of $\tilde{\omega}^{(L,r)}$ and $\tilde{\mathbf{b}}^{(L,r)}$ (c.f. lemma 108), 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 (129) then shows that

$$0 = \sum_u \theta^{(u)}\big(\phi_0^{(u,r)} + \cdots + \phi_{L-1}^{(u,r)}\big) = \sum_u \theta^{(u)}\big(\phi_L^{(u,r)}\big) \tag{130}$$

for all $r \in [R]$ as well. Now take the inner-product of (128) 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 (130) 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 $\ell \in [L]$, and so part (i) is proved. Part (ii) then follows from part (i) since the equalities

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

hold for all (ℓ, r) at any critical point. Thus (128) yields

$$
\begin{aligned}
\mathbf{0} &= \sum_u \theta^{(u)} \nabla_{\mathbf{v}_r} \mathcal{L}^{(r)}|_{\Omega_{u^{(r)}}} \\
\mathbf{0} &= \sum_u \theta^{(u)} \nabla_{\omega^{(\ell,r)}} \mathcal{L}^{(r)}|_{\Omega_{u^{(r)}}} \\
\mathbf{0} &= \sum_u \theta^{(u)} \nabla_{\mathbf{b}^{(\ell,r)}} \mathcal{L}^{(r)}|_{\Omega_{u^{(r)}}} \qquad (131)
\end{aligned}
$$

for all $\ell \in [L], r \in [R]$. Now consider (131) for $r = 1$. Any cells appearing in the sum (131) 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 (131) 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 (131) 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. □

The following preliminary lemma will aid the proofs of the stated corollaries to this theorem.

Lemma 109. *Consider a piecewise multilinear loss*

$$
\bar{\mathcal{L}}(\boldsymbol{\omega}) = \sum_{r=1}^{R} \bar{\mathcal{L}}^{(r)}(\breve{\boldsymbol{\omega}}, \mathbf{v}_r, c_r)
$$

that satisfies theorem 7, and for $\gamma > 0$ let

$$
\mathcal{E}_\gamma(\boldsymbol{\omega}^{(1)}, \ldots, \boldsymbol{\omega}^{(R)}) := \sum_{r=1}^{R} \bar{\mathcal{L}}^{(r)}\big(\boldsymbol{\omega}^{(r)}\big) + \gamma\mathcal{R}\big(\breve{\boldsymbol{\omega}}^{(1)}, \ldots, \breve{\boldsymbol{\omega}}^{(R)}\big) \qquad \boldsymbol{\omega}^{(r)} = (\breve{\boldsymbol{\omega}}^{(r)}, \mathbf{v}_r, c_r)
$$

denote its corresponding exact penalization. If $\big(\boldsymbol{\omega}^{(1)}, \ldots, \boldsymbol{\omega}^{(R)}\big)$ is a local minimum of $\mathcal{E}_\gamma$ and $\breve{\boldsymbol{\omega}} := (\breve{\boldsymbol{\omega}}^{(1)} + \cdots + \breve{\boldsymbol{\omega}}^{(R)})/R$ then $\boldsymbol{\omega} := (\breve{\boldsymbol{\omega}}, \mathbf{v}_1, c_1, \ldots, \mathbf{v}_R, c_R)$ is a local minimum of $\bar{\mathcal{L}}$.

Proof. As $\big(\boldsymbol{\omega}^{(1)}, \ldots, \boldsymbol{\omega}^{(R)}\big)$ is a local minimum of $\mathcal{E}_\gamma$ it must be a critical point. Thus each $\boldsymbol{\omega}^{(r)}$ takes the form

$$\boldsymbol{\omega}^{(r)} = (\breve{\boldsymbol{\omega}}, \mathbf{v}_r, c_r)$$

by theorem 7. If $\breve{\boldsymbol{\omega}}$ is not a local minimizer of $\bar{\mathcal{L}}$ then there exists a sequence $\boldsymbol{\omega}_k \to \boldsymbol{\omega}$ for which $\bar{\mathcal{L}}(\boldsymbol{\omega}_k) < \bar{\mathcal{L}}(\boldsymbol{\omega})$ holds. Define the identically replicated points

$$\boldsymbol{\omega}^{(r,k)} := (\breve{\boldsymbol{\omega}}_k, \mathbf{v}_r^{(k)}, c_r^k), \quad \text{so that} \quad \big(\boldsymbol{\omega}^{(1,k)}, \ldots, \boldsymbol{\omega}^{(R,k)}\big) \to \big(\boldsymbol{\omega}^{(1)}, \ldots, \boldsymbol{\omega}^{(R)}\big)$$

and moreover

$$\mathcal{E}_\gamma\big(\boldsymbol{\omega}^{(1,k)}, \ldots, \boldsymbol{\omega}^{(R,k)}\big) = \bar{\mathcal{L}}(\boldsymbol{\omega}_k) < \bar{\mathcal{L}}(\boldsymbol{\omega}) = \mathcal{E}_\gamma\big(\boldsymbol{\omega}^{(1)}, \ldots, \boldsymbol{\omega}^{(R)}\big)$$

which contradicts the fact that $\big(\boldsymbol{\omega}^{(1)}, \ldots, \boldsymbol{\omega}^{(R)}\big)$ is a local minimizer of $\mathcal{E}_\gamma$. □

For our multiclass analysis we begin at $\alpha = 1$ and study the deep linear problem

$$\begin{aligned} \mathcal{L}(\boldsymbol{\omega}) &= \sum_{r=1}^{R} \mathcal{L}^{(r)}(\boldsymbol{\omega}) \quad \text{for} \\ \mathcal{L}^{(r)}(\boldsymbol{\omega}) &:= \sum \mu^{(i,r)} \sigma\big(1 - y^{(i,r)}(\langle \mathbf{v}_r, \mathbf{x}^{(i,L)} \rangle + c_r)\big) \end{aligned} \tag{132}$$

using the soft penalty approach. The features $\mathbf{x}^{(i,L)} := W^{(L)} \cdots W^{(1)} \mathbf{x}^{(i)}$ result from a deep linear network, and so if we define $\bar{\mathbf{v}}^{(r)} := (W^{(L)} \cdots W^{(1)})^T \mathbf{v}_r$ then we may once again view (132) as a convex loss

$$E^{(1)}(\bar{\mathbf{v}}_1, c_1) + \cdots + E^{(R)}(\bar{\mathbf{v}}_R, c_R)$$

with over-parametrized arguments. If the positive weights $\mu^{(i,r)} > 0$ satisfy $\sum_{y^{(i,r)}=1} \mu^{(i,r)} = \sum_{y^{(i,r)}=-1} \mu^{(i,r)} = \frac{1}{2}$ then we say that the $\mu^{(i,r)}$ give equal weight to all classes. Directly appealing to the critical point relations (127) gives our first simple corollary using this approach.

Corollary 100 (Multiclass Deep Linear Networks, I)**.** *Consider the loss* (132) *and its corresponding penalty* (125) *with* $\gamma > 0$ *and arbitrary data. Assume that* $\boldsymbol{\omega} = (\boldsymbol{\omega}^{(1)}, \ldots, \boldsymbol{\omega}^{(R)})$ *is any critical point of* $\mathcal{E}_\gamma$ *in the Clarke sense. If* $\bar{\mathbf{v}}^{(r)} \neq \mathbf{0}$ *for all* $r \in [R]$ *then* $\boldsymbol{\omega}$ *is a global minimum of* $\mathcal{L}$ *and of* $\mathcal{E}_\gamma$*.*

Proof. By theorem 7 any critical point of $\mathcal{E}_\gamma$ yields a common set of weights $W^{(\ell)}, \ell \in [L]$ for which the two-class critical point relations

$$\mathbf{0} \in \partial_0 \mathcal{L}^{(r)}\big(W^{(1)}, \ldots, W^{(L)}, \mathbf{v}_r, c_r\big)$$

hold for all $r \in [R]$. By theorem 100, if $\bar{\mathbf{v}}_r \neq \mathbf{0}$ then $(\bar{\mathbf{v}}_r, c_r)$ is a global minimum of the convex function

$$E^{(r)}(\mathbf{w}_r, c_r) := \sum_{i=1}^{N} \mu^{(i,r)} \sigma\big(1 - y^{(i,r)}(\langle \mathbf{w}_r, \mathbf{x}^{(i)} \rangle + c_r)\big)$$

and so $\mathbf{0} \in \partial E^{(r)}(\bar{\mathbf{v}}_r, c_r)$ by definition of the subgradient for convex functions. If $\bar{\mathbf{v}}_r \neq \mathbf{0}$ for all $r \in [R]$ it therefore follows that

$$\mathbf{0} \in \partial E^{(r)}(\bar{\mathbf{v}}_r, c_r)$$

for all $r \in [R]$. Finally, define the convex function $E(\mathbf{w}_1, c_1, \ldots, \mathbf{w}_R, c_R) := E^{(1)}(\mathbf{w}_1, c_1) + \cdots + E^{(R)}(\mathbf{w}_R, c_R)$ and note that the sum rule

$$\partial E(\bar{\mathbf{v}}_1, c_1, \ldots, \bar{\mathbf{v}}_R, c_R) = \sum_{r=1}^{R} \partial E^{(r)}(\bar{\mathbf{v}}_r, c_r)$$

holds since each $E^{(r)}$ is Lipschitz. Thus $\mathbf{0} \in \partial E(\bar{\mathbf{v}}_1, c_1, \ldots, \bar{\mathbf{v}}_R, c_R)$, and as $\gamma > 0$ it follows that $\boldsymbol{\omega}$ is a global minimizer. □

Corollary 101 (Multiclass Deep Linear Networks, II)**.** *Consider the loss* (132) *and its corresponding penalty* (125) *with* $\gamma > 0$ *and arbitrary data. Assume that* $\boldsymbol{\omega} = (\boldsymbol{\omega}^{(1)}, \ldots, \boldsymbol{\omega}^{(R)})$ *is a local minimum of* $\mathcal{E}_\gamma$ *with* $\bar{\mathbf{v}}^{(r)} = \mathbf{0}$ *for some* $r \in [R]$*. If the* $\mu^{(i,r)}$ *give equal weight to all classes then* $\boldsymbol{\omega}$ *is a global minimum of* $\mathcal{L}$ *and of* $\mathcal{E}_\gamma$*.*

Proof. Any local minimum $(\boldsymbol{\omega}^{(1)},\ldots,\boldsymbol{\omega}^{(R)})$ of $\mathcal{E}_\gamma$ is necessarily a critical point, and so each $\boldsymbol{\omega}^{(r)}$ takes the form

$$\boldsymbol{\omega}^{(r)} = (W^{(1)},\ldots,W^{(L)},\mathbf{v}_r,c_r)$$

for some common weight matrices $W^{(\ell)}, \ell \in [L]$. Moreover, $\mathbf{0} \in \partial_0 \mathcal{L}^{(r)}(W^{(1)},\ldots,W^{(L)},\mathbf{v}_r,c_r)$ for all $r \in [R]$ as well.

Consider any $r \in [R]$ for which $\bar{\mathbf{v}}^{(r)} = \mathbf{0}$. Then as a function of c_r the convex function $E^{(r)}$ obeys

$$E^{(r)}(\mathbf{0},c_r) = \sum_{i=1}^{N} \mu^{(i,r)} \sigma\big(1 - y^{(i,r)}c_r\big) = \frac{\sigma(1-c_r)+\sigma(1+c_r)}{2}$$

due to the equal weight hypothesis. Thus $\mathcal{E}_\gamma$ can only attain a local minimum if c_r lies in the unit interval $-1 \leq c_r \leq 1$, and moreover $E(\mathbf{0},c_r) \equiv 1$ is constant on the unit interval. It therefore suffices to assume that $-1 < c_r < 1$ without loss of generality, and so in particular, that the function $E^{(r)}$ is differentiable (in fact smooth) near $\boldsymbol{\omega}$. Define the perturbation $\tilde{\boldsymbol{\omega}}$ of $\boldsymbol{\omega} = (\boldsymbol{\omega}^{(1)},\ldots,\boldsymbol{\omega}^{(R)})$ as

$$\begin{aligned}\tilde{W}^{(\ell,r)} &:= W^{(\ell)} + \delta^{(\ell)} X^{(\ell)} \qquad \tilde{W}^{(\ell,s)} := W^{(\ell)} \quad (s \neq r),\\ \tilde{\mathbf{v}}_r &:= \mathbf{v}_r + \delta^{(0)}\mathbf{w}_r,\end{aligned}$$

where the $\delta^{(\ell)}, \ell \in [L]$ and $\delta^{(0)}$ are small scalars, the $X^{(\ell)}$ are arbitrary matrices and $\mathbf{w}_r$ is an arbitrary vector. Then the energy $\mathcal{E}_\gamma$ becomes

$$\mathcal{E}_\gamma(\tilde{\boldsymbol{\omega}}) - \mathcal{E}_\gamma(\boldsymbol{\omega}) = E^{(r)}\big((\tilde{W}^{(L,r)}\cdots\tilde{W}^{(1,r)})^T\tilde{\mathbf{v}}_r,c_r\big) - E^{(r)}(\mathbf{0},c_r) + \gamma\sum_{\ell=1}^{L}\big(\delta^{(\ell)}\big)^2\|X^{(\ell)}\|^2.$$

Define the vector

$$\mathbf{z}_r := \sum_{i=1}^{N} \mu^{(i,r)} y^{(i,r)} \mathbf{x}^{(i)},$$

and note that

$$E^{(r)}\big((\tilde{W}^{(L,r)}\cdots\tilde{W}^{(1,r)})^T\tilde{\mathbf{v}}_r,c_r\big) - E^{(r)}(\mathbf{0},c_r) = -\langle(\tilde{W}^{(L,r)}\cdots\tilde{W}^{(1,r)})\mathbf{z}_r,\tilde{\mathbf{v}}_r\rangle$$

for all $\delta^{(0)},\delta^{(\ell)}$ sufficiently small. As $\boldsymbol{\omega}$ is a local minimizer, the inequality

$$\gamma\sum_{\ell=1}^{L}\big(\delta^{(\ell)}\big)^2\|X^{(\ell)}\|^2 \geq \langle(\tilde{W}^{(L,r)}\cdots\tilde{W}^{(1,r)})\mathbf{z}_r,\tilde{\mathbf{v}}_r\rangle \tag{133}$$

must therefore hold for all $X^{(\ell)},\mathbf{w}_r$ and corresponding $\delta^{(\ell)},\delta^{(0)}$ sufficiently small.

First, apply (133) with $\delta^{(\ell)} = 0$ for all $\ell \in [L]$ to find that

$$0 \geq \langle W^{(L)}\cdots W^{(1)}\mathbf{z}_r, \mathbf{v}_r + \delta^{(0)}\mathbf{w}_r\rangle$$

for any $\mathbf{w}_r$ arbitrary and corresponding $\delta^{(0)}$ sufficiently small. But $\bar{\mathbf{v}}^{(r)} = \mathbf{0}$ and so $0 = \langle W^{(L)}\cdots W^{(1)}\mathbf{z}_r,\mathbf{w}_r\rangle$ for all $\mathbf{w}_r$, whence

$$W^{(L)}\cdots W^{(1)}\mathbf{z}_r = \mathbf{0}.$$

Now, if $\mathbf{z}_r = \mathbf{0}$ then

$$\mathbf{0} = \nabla E^{(r)}(\mathbf{0},c_r),$$

and so if $\mathbf{z}_r$ for all $r \in [R]$ for which $\bar{\mathbf{v}}^{(r)} = \mathbf{0}$ then $\mathbf{0} = \nabla E^{(r)}(\mathbf{0},c_r)$ for all such $r \in [R]$. If $\bar{\mathbf{v}}_r \neq \mathbf{0}$ the relation $\mathbf{0} \in \partial E^{(r)}(\bar{\mathbf{v}}_r,c_r)$ holds as well (as in the previous corollary), and so $\boldsymbol{\omega}$ is a global minimum.

It therefore remains to consider the case where there exists $r \in [R]$ for which $\bar{\mathbf{v}}^{(r)} = \mathbf{0}$ but $\mathbf{z}_r \neq \mathbf{0}$. As $W^{(L)}\cdots W^{(1)}\mathbf{z}_r = \mathbf{0}$ there exists an index $0 \leq k \leq L-1$ for which

$$W^{(k)}\cdots W^{(1)}\mathbf{z}_r \neq \mathbf{0} \qquad W^{(k+1)}\cdots W^{(1)}\mathbf{z}_r = \mathbf{0},$$

where clearly $k = 0$ means $\mathbf{z}_r \neq \mathbf{0}$ but $W^{(1)}\mathbf{z}_r = \mathbf{0}$. Apply (133) with $\delta^{(\ell)} = 0$ for all $\ell \neq k+1$ to find that

$$\gamma\big(\delta^{(k+1)}\big)^2 \|X^{(k+1)}\|^2 \geq \langle W^{(L)} \cdots W^{(k+2)}(W^{(k+1)} + \delta^{(k+1)}X^{(k+1)})W^{(k)} \cdots W^{(1)}\mathbf{z}_r, \mathbf{v}_r + \delta^{(0)}\mathbf{w}_r\rangle$$

for all $X^{(k+1)}, \mathbf{w}_r$ arbitrary and corresponding $\delta^{(k+1)}, \delta^{(0)}$ sufficiently small. But $W^{(k+1)} \cdots W^{(1)}\mathbf{z}_r = \mathbf{0}$ and so

$$\gamma\big(\delta^{(k+1)}\big)^2 \|X^{(k+1)}\|^2 \geq \delta^{(k+1)}\langle W^{(L)} \cdots W^{(k+2)}X^{(k+1)}W^{(k)} \cdots W^{(1)}\mathbf{z}_r, \mathbf{v}_r + \delta^{(0)}\mathbf{w}_r\rangle \tag{134}$$

must hold for all $X^{(k+1)}, \mathbf{w}_r$ arbitrary and corresponding $\delta^{(k+1)}, \delta^{(0)}$ sufficiently small as well. Now apply (134) with $\delta^{(0)} = 0$ to see

$$\gamma\big(\delta^{(k+1)}\big)^2 \|X^{(k+1)}\|^2 \geq \delta^{(k+1)}\langle W^{(L)} \cdots W^{(k+2)}X^{(k+1)}W^{(k)} \cdots W^{(1)}\mathbf{z}_r, \mathbf{v}_r\rangle,$$

but as $W^{(k)} \cdots W^{(1)}\mathbf{z}_r \neq \mathbf{0}$ and $X^{(k+1)}$ is arbitrary this can happen if and only if $(W^{(L)} \cdots W^{(k+2)})^T\mathbf{v}_r = \mathbf{0}$. But then (134) shows

$$\gamma\big(\delta^{(k+1)}\big)^2 \|X^{(k+1)}\|^2 \geq \delta^{(k+1)}\delta^{(0)}\langle W^{(L)} \cdots W^{(k+2)}X^{(k+1)}W^{(k)} \cdots W^{(1)}\mathbf{z}_r, \mathbf{w}_r\rangle$$

must hold for all $X^{(k+1)}, \mathbf{w}_r$ and corresponding $\delta^{(k+1)}, \delta^{(0)}$ sufficiently small. Take $\delta^{(k+1)} = \big(\delta^{(0)}\big)^2$ for $\delta^{(0)}$ small to find

$$\gamma\big(\delta^{(0)}\big)^4 \|X^{(k+1)}\|^2 \geq \big(\delta^{(0)}\big)^3\langle W^{(L)} \cdots W^{(k+2)}X^{(k+1)}W^{(k)} \cdots W^{(1)}\mathbf{z}_r, \mathbf{w}_r\rangle$$

and so in fact $\langle W^{(L)} \cdots W^{(k+2)}X^{(k+1)}W^{(k)} \cdots W^{(1)}\mathbf{z}_r, \mathbf{w}_r\rangle = 0$ for $X^{(k+1)}, \mathbf{w}_r$ arbitrary. This cannot happen unless $k \leq L-2$ and $\mathbf{0} = W^{(L)} \cdots W^{(k+2)}$, in which case in fact $\bar{\mathbf{v}}^{(r)} = \mathbf{0}$ for **all** $r \in [R]$. But then $E^{(r)}$ is differentiable for **all** $r \in [R]$ near $(\mathbf{0}, c_r)$, and so $(W^{(1)}, \ldots, W^{(L)}, \mathbf{v}_1, \ldots, \mathbf{v}_R, c_1, \ldots, c_R)$ is a differentiable local minimum of $\bar{\mathcal{L}}$ by lemma 109. As $\bar{\mathcal{L}}$ is piecewise multilinear and the minimum is differentiable, $\bar{\mathcal{L}}$ must be constant near the minimum. Take $X^{(\ell)}, \mathbf{w}_r$ arbitrary and define

$$\tilde{W}^{(\ell)} = W^{(\ell)} + \delta X^{(\ell)} \qquad \tilde{\mathbf{v}}_r = \mathbf{v}_r + \delta\mathbf{w}_r$$

for all δ sufficiently small. Then

$$\bar{\mathcal{L}}\big(\tilde{W}^{(1)}, \ldots, \tilde{W}^{(L)}, \tilde{V}\big) - \bar{\mathcal{L}}\big(W^{(1)}, \ldots, W^{(R)}, V\big) = 0$$

for all δ small enough, and since

$$\bar{\mathcal{L}}\big(\tilde{W}^{(1)}, \ldots, \tilde{W}^{(L)}, \tilde{V}\big) - \bar{\mathcal{L}}\big(W^{(1)}, \ldots, W^{(R)}, V\big) = \langle\big(\tilde{W}^{(L)} \cdots \tilde{W}^{(1)}\big)^T\tilde{V}, Z\rangle \qquad Z = [\mathbf{z}_1, \ldots, \mathbf{z}_R]$$

it follows that $\langle\big(\tilde{W}^{(L)} \cdots \tilde{W}^{(1)}\big)^T\tilde{V}, Z\rangle = 0$ for all δ small enough. Expanding in powers of δ yields

$$0 = f_0 + \delta f_1\big(X^{(1)}, \ldots, X^{(L)}, W\big) + \cdots + \delta^{(L+1)}f_{L+1}\big(X^{(1)}, \ldots, X^{(L)}, W\big).$$

for some constant f_0 and functions $f_\ell, \ell \in [L+1]$ and all δ small enough. But then

$$f_{L+1}\big(X^{(1)}, \ldots, X^{(L)}, W\big) = 0$$

for all $X^{(\ell)}, W$ arbitrary. As $f_{L+1}\big(X^{(1)}, \ldots, X^{(L)}, W\big) = \langle\big(X^{(L)} \cdots \tilde{X}^{(1)}\big)^T W, Z\rangle$ this can happen if and only if $Z = 0$. Thus

$$\mathbf{z}_r = \nabla E^{(r)}(\mathbf{0}, c_r) = \mathbf{0}$$

for all $r \in [R]$ and so $\boldsymbol{\omega}$ is a global minimum. □

To finish the analysis, we first recall the loss

$$\mathcal{L}(\boldsymbol{\omega}) = \sum_{r=1}^{R} \mathcal{L}^{(r)}(\boldsymbol{\omega}) \quad \text{for} \tag{135}$$

$$\mathcal{L}^{(r)}(\boldsymbol{\omega}) := \sum \mu^{(i,r)}\sigma\big(1 - y^{(i,r)}(\langle\mathbf{v}_r, \mathbf{x}^{(i,1)}\rangle + c_r)\big)$$

for a leaky network with one hidden layer.

Corollary (Corollary 3 from the paper)**.** *Consider the loss* (135) *and its corresponding penalty* (125) *with* $\gamma > 0, 0 < \alpha < 1$ *and data* $\mathbf{x}^{(i)}, i \in [N]$ *that are linearly separable.*

(i) Assume that $\boldsymbol{\omega} = (\boldsymbol{\omega}^{(1)}, \ldots, \boldsymbol{\omega}^{(R)})$ *is a critical point of* $\mathcal{E}_\gamma$ *in the Clarke sense. If* $\mathbf{v}^{(r)} \neq \mathbf{0}$ *for all* $r \in [R]$ *then* $\boldsymbol{\omega}$ *is a global minimum of* $\mathcal{L}$ *and of* $\mathcal{E}_\gamma$*.*

(ii) Assume that the $\mu^{(i,r)}$ *give equal weight to all classes. If* $\boldsymbol{\omega} = (\boldsymbol{\omega}^{(1)}, \ldots, \boldsymbol{\omega}^{(R)})$ *is a local minimum of* $\mathcal{E}_\gamma$ *and* $\mathbf{v}_r = \mathbf{0}$ *for some* $r \in [R]$ *then* $\boldsymbol{\omega}$ *is a global minimum of* $\mathcal{L}$ *and of* $\mathcal{E}_\gamma$*.*

Proof. For part (i), note once again that theorem 7 any critical point of $\mathcal{E}_\gamma$ yields a common set of weights $(W, \mathbf{b})$ for which the two-class critical point relations

$$\mathbf{0} \in \partial_0 \mathcal{L}^{(r)}\big(W, \mathbf{b}, \mathbf{v}_r, c_r\big)$$

hold for all $r \in [R]$. By theorem 4, if $\mathbf{v}_r \neq \mathbf{0}$ for all $r \in [R]$ then $(W, \mathbf{b}, \mathbf{v}_r, c_r)$ is a global minimum of $\mathcal{L}^{(r)}$ for all $r \in [R]$. But the $\mathbf{x}^{(i)}$ are separable, and so

$$0 = \mathcal{L}^{(r)}(W, \mathbf{b}, \mathbf{v}_r, c_r)$$

for all $r \in [R]$ and therefore $\boldsymbol{\omega}$ is a global minimum.

For part (ii), define the sets

$$[R]_0 := \{r \in [R] : \mathbf{v}_r = \mathbf{0}\} \qquad \text{and} \qquad [R]_+ := \{r \in [R] : \mathbf{v}_r \neq \mathbf{0}\}$$

as those classes where $\mathbf{v}_r$ does and does not vanish, respectively. If $r \in [R]_0$ then $\mathbf{v}_r = \mathbf{0}$ and so as a function of c_r the corresponding loss $\mathcal{L}^{(r)}$ takes the form

$$\mathcal{L}^{(r)}(W, \mathbf{b}, \mathbf{v}_r, c_r) = \frac{\sigma(1 - c_r) + \sigma(1 + c_r)}{2}$$

due to the equal weight hypothesis. A local minimum $\boldsymbol{\omega}$ must therefore have $c_r \in [-1, 1]$, and as $\mathcal{L}^{(r)}$ is constant in for $c_r \in [-1, 1]$ it suffices to assume that $c_r \in (-1, 1)$ for all $r \in [R]_0$ without loss of generality. If $r \in [R]_+$ then $(W, \mathbf{b}, \mathbf{v}_r, c_r)$ is a global minimum of $\mathcal{L}^{(r)}$ by part (i), and since the $\mathbf{x}^{(i)}$ are separable a global minimum of $\mathcal{L}^{(r)}$ must have zero loss. Thus each of the N equalities

$$y^{(i,r)}\big(\langle \mathbf{v}_r, \mathbf{x}^{(i,1)} \rangle + c_r\big) \geq 1$$

must hold. By replacing $(\mathbf{v}_r, c_r)$ with $(\lambda \mathbf{v}_r, \lambda c_r)$ for any $\lambda > 1$ with λ arbitrarily close to 1 it therefore suffices to assume that

$$y^{(i,r)}\big(\langle \mathbf{v}_r, \mathbf{x}^{(i,1)} \rangle + c_r\big) > 1$$

for all $i \in [N], r \in [R]_+$ without loss of generality. In other words, by combining the case $r \in [R]_0$ and the case $r \in [R]_+$ it suffices to assume that the local minimizer $\boldsymbol{\omega}$ obeys

$$\operatorname{sign}\big(1 - y^{(i,r)}(\langle \mathbf{v}_r, \mathbf{x}^{(i,1)} \rangle + c_r)\big) \in \{-1, 1\}$$

for all $i \in [N], r \in [R]$ without loss of generality. For such a local minimizer, let $\mathbf{w}_h, h \in [H]$ denote the rows of W. By relabelling the data points if necessary assume that

$$\langle \mathbf{w}_h, \mathbf{x}^{(1)} \rangle \leq \langle \mathbf{w}_h, \mathbf{x}^{(2)} \rangle \leq \cdots \leq \langle \mathbf{w}_h, \mathbf{x}^{(N-1)} \rangle \leq \langle \mathbf{w}_h, \mathbf{x}^{(N)} \rangle,$$

and let i_1 respectively denote the greatest index for which the equality

$$\langle \mathbf{w}_h, \mathbf{x}^{(i)} \rangle + \mathbf{b}_h = 0$$

holds. Thus $0 = \langle \mathbf{w}_h, \mathbf{x}^{(i_1)} \rangle + \mathbf{b}_h < \langle \mathbf{w}_h, \mathbf{x}^{(i_1+1)} \rangle + \mathbf{b}_h$, and so decreasing $\mathbf{b}_h$ by any amount smaller than

$$\langle \mathbf{w}_h, \mathbf{x}^{(i_1)} - \mathbf{x}^{(i_1+1)} \rangle$$

gives a row/bias pair $(\mathbf{w}_h, \mathbf{b}_h)$ for which

$$\operatorname{sign}\big(\langle \mathbf{w}_h, \mathbf{x}^{(i)} \rangle + \mathbf{b}_h\big) \in \{-1, 1\}$$

for all $i \in [N]$. Applying such a decrease to all $\mathbf{b}_h, h \in [H]$ if necessary gives

$$\operatorname{sign}\big(W\mathbf{x}^{(i)} + \mathbf{b}\big) \in \{-1, 1\}^H$$

for all $i \in [N]$. Taking the size of these decreases sufficiently small thus yields a local minimizer $\boldsymbol{\omega}$ of $\mathcal{E}_\gamma$ for which the signature functions obey

$$\mathbf{s}^{(i,1)}(W, \mathbf{b}, \mathbf{v}_1, c_1, \ldots, \mathbf{v}_R, c_r) \in \{-1, 1\}^{NH} \quad \text{and} \quad \mathbf{s}^{(i,2)}(W, \mathbf{b}, \mathbf{v}_1, c_1, \ldots, \mathbf{v}_R, c_r) \in \{-1, 1\}^{NR}$$

for all $i \in [N]$, that is the point $\breve{\boldsymbol{\omega}} := (W, \mathbf{b}, \mathbf{v}_1, c_1, \ldots, \mathbf{v}_R, c_R)$ lies in the interior of a cell on which the loss $\mathcal{L}(\breve{\boldsymbol{\omega}})$ is smooth. Moreover, the corresponding replicated point $\boldsymbol{\omega} = (\boldsymbol{\omega}^{(1)}, \ldots, \boldsymbol{\omega}^{(R)})$ is a local minimum of $\mathcal{E}_\gamma$ and, by lemma 109, of $\mathcal{L}$ as well. Thus $\mathcal{L}$ attains a local minimum on the interior of a cell. But as $\alpha > 0$ and the $\mathbf{x}^{(i)}$ are separable, the decomposition of lemma 102 shows that this can happen only if

$$\mathcal{L}^{(r)}(W, \mathbf{b}, \mathbf{v}_r, c_r) = 0$$

for all $r \in [R]$, and so $\boldsymbol{\omega}$ is a global minimizer as claimed. □

References

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

Laurent, T. and von Brecht, J. H. Deep linear neural networks with arbitrary loss: All local minima are global. *ICML*, 2018.