MINIMAL-NORM UNIVARIATE TWO-LAYER RELU CLASSIFICATION: EXACT SOLUTIONS AND GLOBAL OPTIMALITY WITH SKIP CONNECTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

We study minimal-norm interpolation and ℓ_2-regularized logistic-loss minimization for binary classification by univariate two-layer ReLU networks. We give complete geometric characterizations of the optimal classifiers in function space, resolving how the solutions depend on whether hidden-layer biases are included in the parameter norm. When biases are unpenalized, the minimal-norm interpolators are exactly the continuous piecewise-affine functions that hug every label switch and have kinks of the appropriate convexity. When biases are penalized, the minimizer is unique in function space, has exactly one kink in each intermediate same-label segment, and is therefore a sparsest positive-margin classifier. We further show that adding a free affine skip connection leaves these function-space solutions unchanged but fundamentally improves the parameter-space landscape: every KKT point of the constrained problem becomes globally optimal, whereas suboptimal KKT points can occur without the skip connection. We establish analogous global-optimality and geometric results for sufficiently weak ℓ_2-regularization of the logistic loss. In the unpenalized-bias case, we identify an additional sparsity-like restriction, implying that most minimal-norm interpolators cannot arise as small-regularization limits of margin-normalized logistic-loss minimizers. Numerical experiments across varying dataset complexity and network width support the predicted landscape and sparsity phenomena.

1 INTRODUCTION

In practice, a key factor in the success of modern large language models is regularization of the norm of their parameters during training (e.g., Wang & Aitchison, 2025; Bergsma et al., 2025), by means of (decoupled) weight decay such as in the AdamW optimizer (Loshchilov & Hutter, 2019). Meanwhile, in theory, minimization of the $\mathcal{F}_1$ norm (Kurková & Sanguineti, 2001), which is the counterpart in function space of the ℓ_2 norm of two-layer networks in parameter space, has been proposed as of major significance in understanding generalization of modern neural networks (Neyshabur et al., 2015; Bach, 2017). Moreover, many prominent works studying implicit bias (cf. Vardi, 2023) of gradient based algorithms on overparameterized neural networks have established that the algorithms often find solutions that generalize well precisely because they implicitly minimize the ℓ_2 norm (or a closely related measure) of the model parameters, even when regularization is not explicitly enforced (e.g., Lyu & Li, 2020; Ji & Telgarsky, 2020; Chizat & Bach, 2020; Boursier et al., 2022; Cai et al., 2025). For these and further reasons including its foundational role since the 1960s (Tikhonov, 1963), research on optimization with ℓ_2 norm minimization continues to attract considerable attention in the machine learning community.

A substantial part of that research has focused on the setting of univariate data (the basic case of low dimensionality) and two-layer ReLU networks (the canonical universal approximators (e.g., Leshno et al., 1993), with many challenging features of deep neural networks such as nonconvexity). Two questions have been central:

1. Given a function, what is the minimal[1] parameter norm of a network that represents it?

[1] In this work, when we write just "minimal", "minimum", "minimizer", etc., we mean global; and we write "local" explicitly where that meaning is needed.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

2. Given a finite dataset, what functions are represented by its interpolators whose parameter norm is minimal?

Considerable progress has been made in answering both, which has also identified two important distinctions, one optimizational and the other architectural:

- whether biases are penalized (i.e., included in the parameter norm);
- whether the network has a skip connection (i.e., an additional affine term whose parameters are not penalized).

Studying all four combinations is well motivated, e.g.: the seminal implicit margin maximization results of Lyu & Li (2020) and Ji & Telgarsky (2020) involve penalizing the (first layer) biases, but do not apply in the presence of the skip connection (which breaks the homogeneity of the network); whereas, excluding the biases from weight decay has been standard in practice (cf. Goodfellow et al., 2016, Chapter 7); also, modern neural architectures such as the transformer (Vaswani et al., 2017) typically feature skip connections.

In this context, question 1 has been resolved (Savarese et al., 2019; Ongie et al., 2020; Boursier & Flammarion, 2023), and question 2 for regression to a large extent (Parhi & Nowak, 2021; Hanin, 2021; Debarre et al., 2022; Boursier & Flammarion, 2023; Kim et al., 2025). Furthermore, building on these works, Zeno et al. (2023) studied generalization of minimal-norm denoiser networks and Joshi et al. (2024) characterized the nature of overfitting noisy data by minimal-norm interpolators.

For classification (where interpolation means $y\, f(x) \geq 1$ for each input x whose label is $y \in \{\pm 1\}$), convex-duality approaches have shown that in one dimension a globally optimal hinge-loss classifier can be chosen from a finite dictionary of data-aligned ReLU features (Ergen & Pilanci, 2021), while Safran et al. (2022), Boursier & Flammarion (2023), and Kornowski et al. (2023) obtained complementary structural and dynamical results. However, a complete characterization of all finite-sample minimum-norm classifier functions, including their dependence on bias penalization and skip connections, has remained open. In this work, we fill that gap, and do much more by answering also the following questions for classification:

3. How does adding a skip connection change the optimization landscape?
4. Given a finite dataset, what functions are represented by minimizers of the ℓ_2-regularized logistic loss; and how are those functions, with margin normalization and small regularization strengths, related to those of the minimal-norm interpolators?

More broadly, question 3 is pertinent since, in spite of a great deal of research (e.g., Hardt & Ma, 2017; Orhan & Pitkow, 2018; Barzilai et al., 2023; MacDonald et al., 2023; Belfer et al., 2024), theoretical understanding of skip connections is still a long way behind empirical demonstrations such as in the seminal works of He et al. (2016) and Li et al. (2018). Similarly, question 4 addresses the theory-practice gap between minimal-norm interpolation and regularized-loss minimization, including how small regularization strengths λ need to be. Here is a summary of our main contributions.

– **Exact function space characterizations.** We establish complete geometric characterizations of the minimal-norm interpolators and of the minimizers of the regularized loss in function space, show that they are not affected by the absence or presence of the skip connection, and show how they differ depending on whether the biases are penalized.

– **Skip connections guarantee global optimality.** We prove that, with a skip connection, and regardless of whether the biases are penalized, every KKT point of the interpolator norm minimization problem and every positive-margin stationary point of the regularized loss are (global) minimizers, and show that this is not the case without the skip connection.

– **Regularized loss promotes sparsity.** We prove that, when the biases are not penalized, the minimizers of the regularized loss (for any λ below a mild bound) in function space have a sparsity-like geometric property which minimal-norm interpolators do not necessarily have, and that therefore most minimal-norm interpolators cannot be obtained as limits of margin-normalized minimizers of the regularized loss by letting λ tend to 0.

– **Validation by numerical experiments.** For our theoretical results on how the optimization landscapes change with the skip connection, and on sparsity, we performed experiments which validate them on a range of synthetic datasets and with varying degrees of overparameterization.

Implications of our results include the following improvements and extensions of algorithms and results in prior works.

– The $32r + 67$ bound of Safran et al. (2022, Theorem 4.2) on the number of linear regions of KKT points in function space when the biases are penalized reduces to r (where r is the number of label switches in the dataset), provided either a skip connection is added or minimizers are considered instead of KKT points.

– We prove the conjecture of Boursier & Flammarion (cf. 2023, Corollary 1) that minimal-norm interpolators when the biases are penalized and with the skip connection are unique in function space, and show both uniqueness and optimal sparsity also without the skip connection.

– The privacy attack of Smorodinsky et al. (2026) that extracts from any local minimizer, when the biases are penalized and without knowing the training dataset, a set of points at least $1/4$ of which are guaranteed to be in the dataset, can be improved to guarantee that all extracted points are in the dataset and moreover that every same-label segment of the dataset is represented, provided either a skip connection is added (in which case it suffices to consider KKT points) or global minimizers are considered.[2]

We provide proofs of all our results in the appendix, which also contains further results, background material, examples, figures, plots, remarks, and details of our experiments.

2 SETTING

Networks. We consider two-layer ReLU networks with a skip connection and univariate inputs $x \in \mathbb{R}$, whose output is

$$f_{\theta,a_0,b_0}(x) \coloneqq a_0 x + b_0 + \textstyle\sum_{j=1}^{m} a_j\, \sigma(w_j x + b_j), \tag{1}$$

where $\sigma(z) \coloneqq \max\{0, z\}$ is the ReLU nonlinearity, m is the network width, a_0 is the weight and b_0 is the bias for the skip connection, and the remaining parameters $\theta = (a_j, w_j, b_j)_{j=1}^{m} \in \mathbb{R}^{3m}$ consist of output weights a_j, hidden weights w_j, and hidden biases b_j. When the skip connection is absent, instead of $f_{\theta,0,0}$ we may write just f_θ.

Functions. We say that a function $g\colon \mathbb{R} \to \mathbb{R}$ is CPA if and only if it is continuous piecewise affine with a finite number of pieces. We denote the slopes of its pieces by $(s_k)_{k=0}^{q}$. The pieces are separated by its kinks[3] $p_1 < \cdots < p_q$. For a kink p_k, we say that it is convex if its incident slopes satisfy $s_{k-1} < s_k$, and concave if $s_{k-1} > s_k$.

CPA functions are exactly the class representable by univariate two-layer ReLU networks, and this characterization is not affected by whether the skip connection is present. For simplicity of presentation and proximity to practice, in this work we do not consider infinite-width networks.

Datasets. We consider univariate binary classification datasets, i.e., $(x_i, y_i)_{i=1}^{n} \in (\mathbb{R} \times \{\pm 1\})^n$, and without loss of generality we assume that the inputs are sorted, i.e., $x_1 < \cdots < x_n$.

Let $\iota(0) \coloneqq 0$, and for each k with $\iota(k) < n$ let $\iota(k+1) \coloneqq \min\big(\{i > \iota(k) \mid y_i \neq y_{i+1}\} \cup \{n\}\big)$. This defines $\iota(0) < \cdots < \iota(r+1) = n$ such that the dataset has $r+1$ maximal same-label segments and the indices between which the label switches are exactly $\iota(k)$ and $\iota(k)+1$ for all $k \in [r]$.

Assumption 1. *The inputs in the first (resp., last) two segments of the dataset are negative (resp., positive), i.e.,* $x_{\iota(2)} < 0 < x_{\iota(r-1)+1}$.

This assumption avoids some special cases that would clutter the statements of our main results. It combines requiring that the number r of label switches in the dataset is ≥ 3 with a coordinate-dependent sign condition used only in the penalized-bias outer-kink arguments; see appendix D for its precise role. Also, this is the only assumption in this work; in particular, all our results apply regardless of how the dataset is sampled and for any cardinality n.

[2] Also, neither knowledge of the network's parameters nor existence of an omniactive neuron are needed.

[3] The "kinks" are sometimes in the literature called "knots", "breakpoints", or "inflection points".

Table 1: We consider these four variants of the interpolator norm minimization problem for a given dataset $(x_i, y_i)_{i=1}^n \in (\mathbb{R} \times \{\pm 1\})^n$. The skip connection is always "free", i.e., its parameters a_0 and b_0 are never penalized. Note also that the network width m is optimized, being determined by the dimension of the parameters vector $\theta = (a_j, w_j, b_j)_{j=1}^m \in \mathbb{R}^{3m}$.

	biases not penalized		**biases penalized**	
skip connection absent	$\inf_\theta \frac{1}{2}\sum_{j=1}^m (a_j^2 + w_j^2)$ s.t. $\forall i \in [n],\ y_i\, f_\theta(x_i) \geq 1$	(P)	$\inf_\theta \frac{1}{2}\sum_{j=1}^m (a_j^2 + w_j^2 + b_j^2)$ s.t. $\forall i \in [n],\ y_i\, f_\theta(x_i) \geq 1$	(P^{b})
skip connection present	$\inf_{\theta, a_0, b_0} \frac{1}{2}\sum_{j=1}^m (a_j^2 + w_j^2)$ s.t. $\forall i \in [n],\ y_i\, f_{\theta,a_0,b_0}(x_i) \geq 1$	(P_1)	$\inf_{\theta, a_0, b_0} \frac{1}{2}\sum_{j=1}^m (a_j^2 + w_j^2 + b_j^2)$ s.t. $\forall i \in [n],\ y_i\, f_{\theta,a_0,b_0}(x_i) \geq 1$	($\mathrm{P}_1^{\mathrm{b}}$)

3 Main results

The complete geometric characterizations obtained in our first main theorem reveal the exact effect of penalizing the biases: among switch hugging and convexity correct CPAs (of which there are infinitely many in general), it picks the unique one that is single turning, which is also the sparsest possible. In addition, the theorem shows that although the skip connection does not modify the sets of minimizers in function space, it changes the optimization landscape dramatically, making every KKT point a (global) minimizer.

Theorem 1. *Under Assumption 1, for the problems in Table 1, all of the following hold.*

*(i) Problems (*P*) and (*P_1*) have the same minimizers in function space, which are exactly all CPAs $g\colon \mathbb{R} \to \mathbb{R}$ that are:*

- **(switch hugging)** *at every label switch in the dataset, g passes through the two incident points, i.e., for all $l \in [r]$, we have $g(x_{\iota(l)}) = y_{\iota(l)}$ and $g(x_{\iota(l)+1}) = y_{\iota(l)+1}$;*
- **(convexity correct)** *every kink p_k of g is contained in an intermediate same-label segment of the dataset, and its convexity is opposite to the segment's label, i.e., there exists $l \in [r-1]$ with $p_k \in [x_{\iota(l)+1}, x_{\iota(l+1)}]$ and $sgn(s_k - s_{k-1}) = -y_{\iota(l)+1}$.*

(ii) Problems (P^{b}*) and (*$\mathrm{P}_1^{\mathrm{b}}$*) have the same unique minimizer g in function space, which is of the form as in part (i) and moreover:*

- **(single turning)** *every intermediate same-label segment contains exactly one kink of g.*

Thus g has $r-1$ kinks in total, which makes it a sparsest interpolant of the dataset.

(iii) For probs. (P_1*) and (*$\mathrm{P}_1^{\mathrm{b}}$*), every KKT point $(\theta, a_0, b_0) \in \mathbb{R}^{3m+2}$ (with any network width m for which KKT points exist) is a minimizer (globally, including over all m).*

Proof sketch. For parts (i) and (ii), we first prove necessity of the geometric conditions, for every KKT point with the skip connection. Any violation yields an infinitesimal perturbation that decreases the parameter norm without decreasing any tight training margin, contradicting first-order optimality. The skip connection is crucial: since $\sigma(z) - \sigma(-z) = z$, reversing a neuron changes its contribution by an affine function, which the skip connection absorbs while preserving the relevant directional derivatives, allowing us to orient neurons conveniently when constructing descent directions.

When biases are not penalized (still with the skip connection, so prob. (P_1)), these directions exclude kinks of inappropriate convexity, outer kinks, and failures of switch hugging. Conversely, for every switch-hugging and convexity-correct CPA, the prescribed signs of the slope changes make the total variation telescope, so Theorem 4 gives the same function-space norm independently of the admissible kink locations and number. Hence these are exactly the minimizers with a skip connection. When

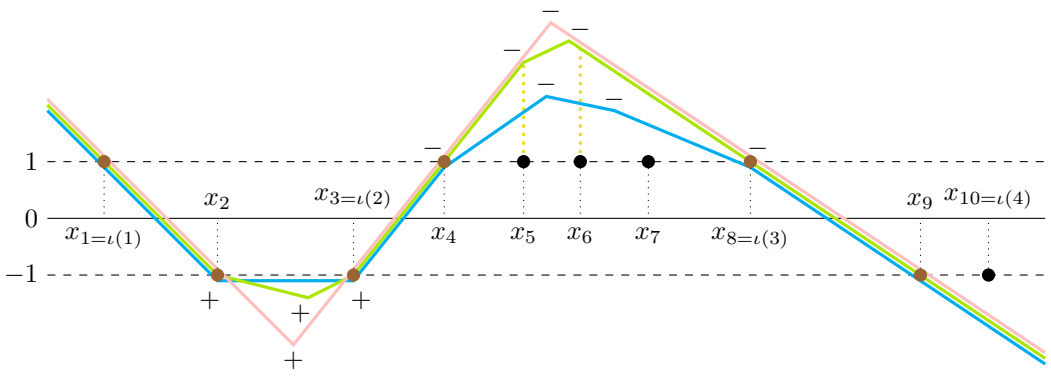


Figure 1: Here we depict a dataset with $n = 10$ points which are indicated by the bold dots, and three interpolating CPAs g_{cyan}, g_{lime}, and g_{pink} which are plotted with those colors respectively. The dataset has $r = 3$ label switches, and its same-label segments end at indices $1 = \iota(1)$, $3 = \iota(2)$, $8 = \iota(3)$, and $10 = \iota(4)$. Each convex kink is indicated by $+$, and each concave kink by $-$. All of g_{cyan}, g_{lime}, and g_{pink} are switch hugging and convexity correct, however only g_{pink} is single turning (cf. Theorem 1). Also, g_{lime} and g_{pink} are interval turning, but g_{cyan} is not (cf. Corollary 1). For a switch hugging and convexity correct interpolating CPA to be interval turning, all its kinks in the same-label segment $[x_4, x_8]$ need to be in the interval $[x_5, x_6]$ between two consecutive inputs that contains the unique single turning kink (which is indicated by the yellow dotted lines).

biases are penalized (prob. $(\mathrm{P}_1^{\mathrm{b}})$), first-order rescaling forces balancedness, and the cost of a kink at p with slope change Δs is $\sqrt{1+p^2}\,|\Delta s|$. The strict triangle inequality then rules out two distinct kinks of the required convexity in one same-label segment. Each intermediate segment therefore has one kink, uniquely determined by the adjacent switch points, giving the unique and sparsest minimizer. For any positive-margin CPA with at least three label switches, a convex combination of two no-skip representations attains the same cost, so removing the skip connection (probs. (P) and $(\mathrm{P}^{\mathrm{b}})$) does not change the minimizing functions.

For part (iii), further descent directions ensure that every KKT parameterization with a skip connection attains the minimal representor norm of its function. Since that function is globally optimal, the parameterization is globally optimal over all widths. The constraint qualification holds throughout the feasible set, so every local minimizer is also covered. □

Although Theorem 1 (i) and (ii) show that adding the free skip connection does not change the optimal functions, its landscape conclusion is specific to the architectures with a skip connection. For the dataset in Figure 2, prob. (P) has plateaus of suboptimal KKT points (the corresponding interpolants have outer kinks and are therefore not convexity correct), which disappear from the plotted profile after adding the skip connection. Theorem 5 in the appendix gives the explicit construction and an analogous penalized-bias counterexample.

Next we state a corollary of two theorems fully given in the appendix. It characterizes regularized-loss minimizers in function space as Theorem 1 does minimal-norm interpolators, but with far less flexibility when biases are unpenalized: all CPA kinks in each dataset segment must lie between two specific consecutive training inputs. This discontinuity between regularized and constrained optimization, more pronounced as the dataset grows denser (more samples per same-label segment), identifies a sparsity-driving mechanism beyond bias penalization. The result also shows that a skip connection makes every positive-margin stationary point a (global) minimizer, and that most minimal-norm interpolators are not limits of margin-normalized regularized-loss minimizers as $\lambda \to 0$.

Corollary 1 (of Theorems 6 and 7)**.** *Under Assumption 1, and for every sufficiently small regularization strength λ (depending only on the dataset), all of the following hold.*

(i) The characterizations in Theorem 1 (i) and (ii) of the minimizers of probs. (P), (P_1), $(\mathrm{P}^{\mathrm{b}})$ and $(\mathrm{P}_1^{\mathrm{b}})$ in function space hold also for the minimizers g of the corresponding ℓ_2-regularized logistic losses (cf. page 41), except that the switch hugging property is replaced by

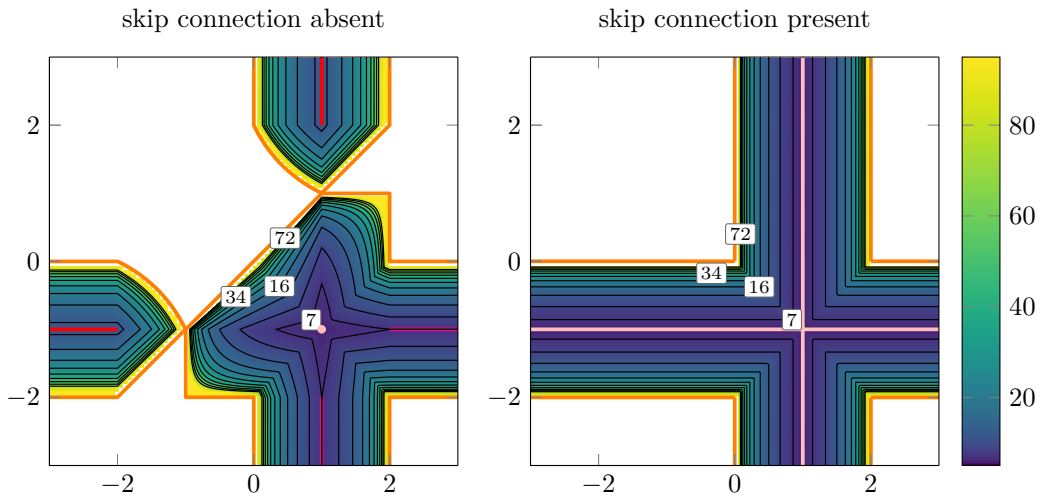


Figure 2: A free skip connection removes suboptimal KKT plateaus from a two-dimensional landscape profile without changing the global optimum. For the dataset $(-2, 1), (-1, -1), (1, 1), (2, -1)$, we plot the norm objective over interpolating symmetric width-four networks with unpenalized biases. The axes give two kink positions $p_\dagger, p_\ddagger$; the corresponding slope changes are optimized subject to interpolation, and orange marks the feasibility boundary. **Left:** Without a skip connection, the profile contains suboptimal plateaus corresponding to KKT points at objective values 11 (colored red) and 7 (colored magenta), whereas the global minimum is 6 (colored pink). **Right:** With a free skip connection, these plateaus disappear from the profile, while the global minimum remains 6. See Theorem 5 for the KKT construction and Remark 1 for additional local-minimality observations.

(switch optimal) *for all $l \in [r]$, the outputs $g(x_{\iota(l)})$ and $g(x_{\iota(l)+1})$ at the inputs incident to the l-th label switch are uniquely determined by the dataset and λ;*

and when the biases are not penalized, the CPA g additionally satisfies

(interval turning) *for all $l \in [r-1]$, letting p and p' be the first and last (respectively) kinks of g in the l-th intermediate same-label segment $[x_{\iota(l)+1}, x_{\iota(l+1)}]$, the interval (p, p') contains no inputs from the dataset.*

(ii) If the skip connection is present, then every Clarke stationary point of the ℓ_2-regularized logistic loss (without or with penalizing the biases) whose margin is positive is a minimizer (globally, including over all m).

(iii) For all four options of whether the skip connection is present and whether the biases are penalized, the limits of margin-normalized Clarke stationary points of the ℓ_2-regularized logistic loss as λ tends to 0 are KKT points of the corresponding interpolator norm minimization problem. If the skip connection is present and the biases are not penalized, then the latter KKT points are necessarily interval turning in function space.

Proof sketch. The proof of Theorem 6 parallels that of Theorem 1, with two changes: every training point contributes to the objective, and the outputs at label switches are selected by a finite-dimensional convex optimization problem rather than fixed at ± 1. For $\tau, \pi \in \mathbb{R}^r_{>0}$, let $g_{\tau,\pi}$ be the canonical CPA whose k-th crossing piece joins $(x_{\iota(k)}, y_{\iota(k)}\tau_k)$ and $(x_{\iota(k)+1}, y_{\iota(k)+1}\pi_k)$. Linear inequalities characterize the compatibility of these pieces across intermediate same-label segments. On this convex feasible set, the empirical logistic loss of $g_{\tau,\pi}$ plus the corresponding minimal representor norm regularizer is strictly convex because of the terms at the switch inputs. Under the stated upper bound on λ (see page 45), each objective has a unique minimizer with strictly positive coordinates, defining the switch-optimal values.

The geometric descent arguments from Theorem 1 exclude failures of convexity correctness and, when biases are penalized, multiple kinks in a same-label segment, giving single turning. Without

bias penalization, if two same-convexity kinks enclose a training input, the graph between them can be moved toward the correct label without increasing the norm and while decreasing the logistic loss. Stationarity therefore requires all kinks in each same-label segment to lie between the same pair of consecutive training inputs, which is exactly interval turning.

Suppose a positive-margin stationary point has the appropriate turning and convexity properties but is not switch optimal. We perturb one or two neurons in each intermediate segment, together with the skip connection, so that the first-order change in the switch margins points from (τ, π) toward the unique convex minimizer $(\tau^\star, \pi^\star)$. The directional derivative of the network regularizer equals that of the corresponding function-space norm, so strict convexity makes the derivative of the full objective negative, contradicting stationarity. This proves the characterizations in Theorem 6. As in Theorem 1, balancing and alignment make the parameter norm equal the function's minimal representor norm; hence every positive-margin stationary point with a skip connection is (globally) minimizing.

For Theorem 7, let $\lambda_k \to 0$ and margin-normalize a sequence of approximate stationary points. Stationarity yields approximate KKT conditions for the corresponding unit-margin interpolation problem, with multipliers $\mu_i^{(k)} = 1/n\lambda_k\big(1 + \exp(y_i\, f_k(x_i))\big)$. The assumed relation between the stationarity error and λ_k makes the KKT equilibrium error vanish and the unnormalized margins diverge. The multipliers at normalized margins strictly larger than 1 therefore vanish, giving asymptotic complementary slackness, so every convergent normalized subsequence limits to a KKT point.

In the unpenalized-bias case with a skip connection, a non-interval-turning limit would yield, uniformly along the sequence, a direction improving a training margin without increasing the quadratic regularizer. The resulting logistic-loss improvement has an exponential scale determined by the margin, and the stronger error condition in Theorem 7 (iii) ensures that it dominates the stationarity error, a contradiction. □

Figure 11 in the appendix shows that the small-λ regularized-loss landscape undergoes the same qualitative transformation as the constrained landscape in Figure 2 when the skip connection is added.

4 Experiments

The main outcomes of our numerical experiments are plotted in Figure 3. In the interval $(-1, 1)$, we sample r_0 class breaks and use them to label a random dataset of $n \coloneqq (r_0/2)^2$ inputs, while satisfying Assumption 1. For an overparameterization hyperparameter ω, we then randomly initialize a network of width $m \coloneqq \omega \cdot r_0$, and train it by Adam (Kingma & Ba, 2015) to get close to a Clarke stationary point of the ℓ_2-regularized logistic loss with small regularization strength λ. Appendix I gives full details and additional sparsity and positive-margin results. Since the global minimum is computed independently through the convex characterization in Theorem 6, the reported excess loss directly tests whether training reaches the theoretically characterized optimum.

5 Discussion

A function-space view of norm-controlled shallow ReLU networks begins with characterizations of the minimal parameter norm needed to represent a prescribed function. For univariate inputs, Savarese et al. (2019) related this cost to the total variation of the derivative, while Ongie et al. (2020) developed a multivariate analogue using Radon-transform methods. Parhi & Nowak (2021) placed such results in a broader Banach-space representor-theorem framework, under which shallow networks arise as ridge-spline solutions of variational problems. These works characterize the regularizer induced in function space, but do not by themselves determine which function is selected by finite classification constraints. In one dimension, Boursier & Flammarion (2023) showed that including hidden biases in the parameter norm spatially weights the variation penalty and promotes sparse, and in their regression setting generically unique, interpolants, whereas omitting the biases permits nonsparse solutions. Our work builds on these representor-norm formulas, but resolves the corresponding classification problems for both choices of bias regularization and both with and without a free affine skip connection.

For finite-sample interpolation, Hanin (2021) showed that minimal-norm univariate ReLU regression exhibits a nearest-neighbor curvature extrapolation rule and established associated generalization

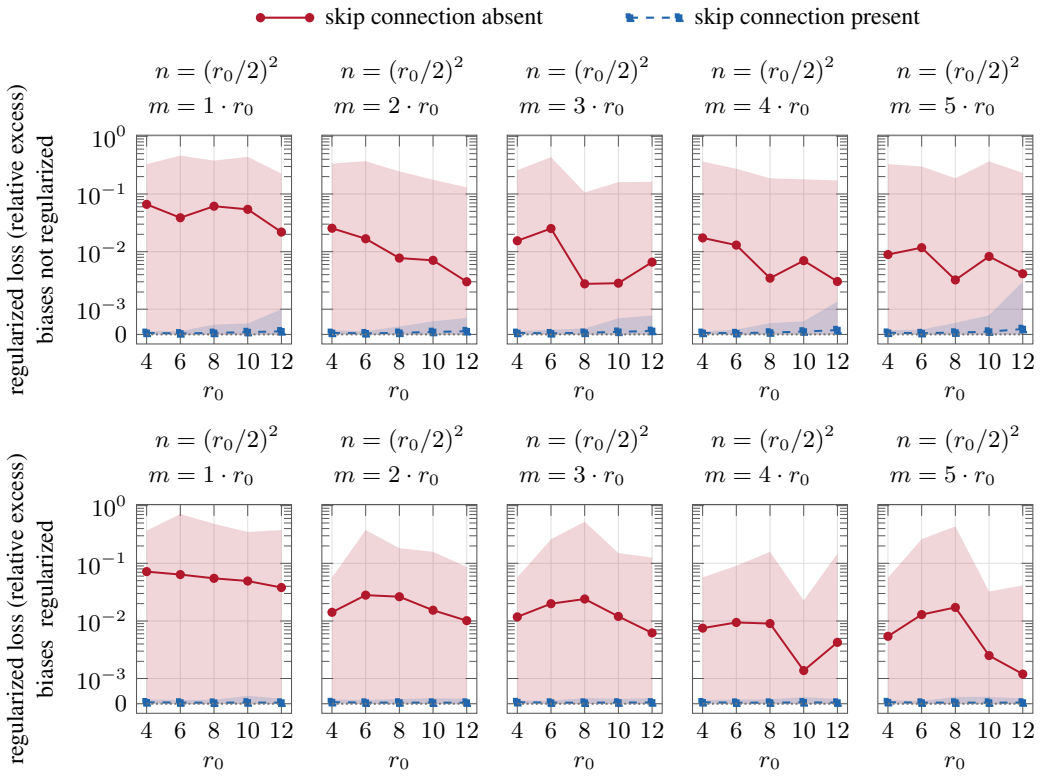


Figure 3: We plot the relative excess of the ℓ_2-regularized logistic loss L at the end of training, i.e., $L/L^\star - 1$, where $L^\star$ is the global minimum for the experiment setting, computed via the convex problem characterized in Theorem 6. Rows indicate whether biases are regularized, columns give the overparameterization factor $\omega = m/r_0$, and the two curves indicate whether the skip connection is absent or present. For each point, we use the runs, among 60 RNG seeds, that reached a positive margin: the curve gives their mean, and the shaded region spans their smallest and largest outcomes (positive-margin failure rates are reported in Figure 15). The shared vertical axis is logarithmic above 10^{-3} and linear below it. **Observations:** With a skip connection, training consistently reaches or closely approaches the global minimum; without one, the best runs reach the minimum but the means and worst outcomes typically remain substantially above it. This is consistent with our theoretical result that a skip connection rules out suboptimal positive-margin stationary points, and with the counterexamples showing that such points can occur without it.

guarantees. Debarre et al. (2022) studied the related, but distinct, objective of obtaining sparsest one-dimensional piecewise-linear regressors, while Kim et al. (2025) connected regularized neural-network landscapes to convex formulations. In complementary multivariate directions, Ardeshir et al. (2023) investigated the intrinsic dimensionality and generalization properties of the R-norm inductive bias, and Park et al. (2023) established an asymptotic connection between regularized empirical-risk minimization and minimal-norm shallow-network interpolation as the sample size and width grow and the regularization vanishes. The latter contrasts with our fixed-dataset characterization: in our classification setting, margin-normalized small-regularization solutions may select only a strict subset of the minimal-norm interpolators.

Ergen & Pilanci (2021) gave a convex-geometric treatment of norm-regularized finite-width two-layer ReLU networks. For univariate binary classification with hinge loss, they showed that an optimum can be chosen from data-aligned extreme-point ReLU features, reducing training to a finite-dimensional minimal-ℓ_1-norm SVM. Our results are complementary: we characterize the entire set of globally optimal classifier functions, rather than the existence of a sparse extreme-point representative, and show that in the unpenalized-bias case optimal functions may have kinks away from training inputs. We also treat penalized hidden biases, a free affine skip connection, the stationary-point landscape of the original nonconvex parameterization, regularized logistic loss, and its small-regularization limit.

Among works focused on gradient dynamics and the complexity of the learned classifier, the closest is Safran et al. (2022): they studied gradient flow for shallow univariate ReLU classifiers and bounded the number of linear regions of the limiting classifier under distributional and initialization assumptions. Kornowski et al. (2023) related the piecewise-linear complexity of learned ReLU functions to tempered and benign overfitting. Other works use minimal-norm structure to study statistical or algorithmic consequences: Zeno et al. (2023) derived closed-form shallow denoisers and associated univariate and multivariate geometric properties; Joshi et al. (2024) showed that minimal-norm univariate ReLU regression can exhibit tempered or catastrophic overfitting depending on the risk criterion; Smorodinsky et al. (2026) constructed provable training-data privacy attacks; and Harzli et al. (2026) gave sufficient conditions for the algorithmic stability of minimal-norm interpolating deep ReLU networks. Complementarily, Nakhleh & Nowak (2025) showed that suitable ℓ_p quasi-norm objectives have globally sparsest interpolating shallow ReLU networks as their global minimizers. Our deterministic results isolate the positive-margin classification geometry underlying such questions: for arbitrary finite datasets satisfying only a mild assumption, we characterize all globally optimal functions, determine exactly when sparsity and uniqueness occur, and, with a skip connection, upgrade first-order stationarity to global optimality. Thus the univariate setting permits a simultaneous and exact treatment of function-space selection, parameter-space optimization, and the regularized-to-constrained limit that is not currently available in broader architectures.

Our landscape results also connect to theoretical explanations of residual architectures. Hardt & Ma (2017) showed that identity parameterizations can give deep linear residual networks favorable optimization properties; Orhan & Pitkow (2018) argued that skip connections remove overlap, elimination, and linear-dependence singularities; and Barzilai et al. (2023) and Belfer et al. (2024) analyzed how residual connections modify convolutional and deep neural tangent kernels. MacDonald et al. (2023) provided a layerwise framework in which skip connections propagate favorable curvature and regularity properties, and identified a mechanism for accelerating gradient optimization. These works explain conditioning, spectral behavior, or trainability in deep models. By contrast, we give an exact finite-network landscape statement: a free affine skip leaves the globally optimal functions unchanged but makes every KKT point of the constrained problem, and every positive-margin stationary point of the regularized problem, globally optimal. This separation of representational and optimizational effects provides a basic setting in which the benefit of a skip connection can be established globally rather than locally or asymptotically.

6 Conclusion

The preceding discussion places our results within broader efforts to understand the function-space bias, statistical consequences, and optimization landscapes induced by norm control. Although we focus on univariate classification by two-layer ReLU networks, this tractable setting allows all three aspects to be resolved together: we completely characterize the optimal classifiers in function space; determine how penalizing biases changes their sparsity and uniqueness; and show that a free affine skip connection eliminates suboptimal KKT points, and hence local minima, without changing the globally optimal functions. These conclusions require only a mild assumption on the dataset.

Natural next steps are to characterize KKT and Clarke stationary points more fully when the skip connection is absent, and to extend the geometric and landscape results to multivariate inputs and deeper networks. The exact univariate theory developed here provides both a benchmark for such extensions and a collection of mechanisms—including label-switch geometry, bias-induced spatial weighting, and skip-enabled descent directions—whose higher-dimensional counterparts remain to be understood.

A complementary empirical direction is to study which of the possible no-skip stationary points are selected by gradient-based dynamics. A natural benchmark is unregularized logistic-loss gradient flow in the homogeneous, bias-penalized architecture, for which implicit-bias results relate long-time normalized trajectories to KKT points of the corresponding margin problem (Lyu & Li, 2020; Ji & Telgarsky, 2020). Comparing these trajectories with our exact global minimizer could determine whether the dynamics select the global solution or one of the suboptimal KKT points that may occur without a skip connection. Such an experiment is challenging, however, because directional convergence is asymptotic and can be extremely slow.

AI USE STATEMENT

In this work, we used generative AI tools to: aid or polish writing in around 20% of the paper; find several related works; create and edit the code for running the experiments and plotting their results, and for depicting the optimization landscapes. We take responsibility for the final content of this work, including the code produced with the aid of generative AI, which we have reviewed and tested.

REPRODUCIBILITY STATEMENT

For the theoretical results in this work, we clarified all nonstandard notation, defined all new notions, stated all required assumptions, and provided all nontrivial proofs (either in the main part of the paper, or in the appendix). For performing the experiments and plotting their results, and for producing the optimization landscape depictions in Figures 2 and 11, we provided complete code and instructions how to run it in the supplementary materials; moreover, doing so does not require specialized software or hardware, and uses a relatively small amount of compute.

REFERENCES

Navid Ardeshir, Daniel J. Hsu, and Clayton Hendrick Sanford. Intrinsic dimensionality and generalization properties of the R-norm inductive bias. In *COLT*, pp. 3264–3303, 2023. 8

Francis R. Bach. Breaking the Curse of Dimensionality with Convex Neural Networks. *J. Mach. Learn. Res.*, 18:19:1–19:53, 2017. 1

Daniel Barzilai, Amnon Geifman, Meirav Galun, and Ronen Basri. A Kernel Perspective of Skip Connections in Convolutional Networks. In *ICLR*, 2023. 2, 9

Yuval Belfer, Amnon Geifman, Meirav Galun, and Ronen Basri. Spectral Analysis of the Neural Tangent Kernel for Deep Residual Networks. *J. Mach. Learn. Res.*, 25:184:1–184:49, 2024. 2, 9

Shane Bergsma, Nolan Dey, Gurpreet Gosal, Gavia Gray, Daria Soboleva, and Joel Hestness. Power Lines: Scaling laws for weight decay and batch size in LLM pre-training. In *NeurIPS*, 2025. 1

Jonathan Borwein and Adrian Lewis. *Convex Analysis and Nonlinear Optimization: Theory and Examples*. Springer, 2nd edition, 2010. 15

Etienne Boursier and Nicolas Flammarion. Penalising the biases in norm regularisation enforces sparsity. In *NeurIPS*, 2023. Also in CoRR abs/2303.01353, which contains a correction to Theorem 4. 2, 3, 7, 17

Etienne Boursier, Loucas Pillaud-Vivien, and Nicolas Flammarion. Gradient flow dynamics of shallow ReLU networks for square loss and orthogonal inputs. In *NeurIPS*, 2022. 1

Yuhang Cai, Kangjie Zhou, Jingfeng Wu, Song Mei, Michael Lindsey, and Peter L. Bartlett. Implicit Bias of Gradient Descent for Non-Homogeneous Deep Networks. In *ICML*, 2025. 1

Lénaïc Chizat and Francis R. Bach. Implicit Bias of Gradient Descent for Wide Two-layer Neural Networks Trained with the Logistic Loss. In *COLT*, 2020. 1

Frank H. Clarke. Generalized gradients and applications. *Trans. Amer. Math. Soc.*, 205:247–262, 1975. 15

Thomas Debarre, Quentin Denoyelle, Michael Unser, and Julien Fageot. Sparsest piecewise-linear regression of one-dimensional data. *J. Comput. Appl. Math.*, 406:114044, 2022. 2, 8

Steven Diamond and Stephen Boyd. CVXPY: A Python-embedded modeling language for convex optimization. *J. Mach. Learn. Res.*, 17(83):1–5, 2016. 59

Joydeep Dutta, Kalyanmoy Deb, Rupesh Tulshyan, and Ramnik Arora. Approximate KKT points and a proximity measure for termination. *J. Glob. Optim.*, 56(4):1463–1499, 2013. 15, 16

Tolga Ergen and Mert Pilanci. Convex Geometry and Duality of Over-parameterized Neural Networks. *J. Mach. Learn. Res.*, 22(212):1–63, 2021. 2, 8

Giorgio Giorgi, Bienvenido Jiménez, and Vicente Novo. *Basic Mathematical Programming Theory*. Springer, 2023. 16

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. 2

Paul J. Goulart and Yuwen Chen. An interior-point solver for conic programs with quadratic objectives. *Math. Prog. Comp.*, 2026. 59

Boris Hanin. Ridgeless Interpolation with Shallow ReLU Networks in $1D$ is Nearest Neighbor Curvature Extrapolation and Provably Generalizes on Lipschitz Functions. *CoRR*, abs/2109.12960, 2021. 2, 7

Moritz Hardt and Tengyu Ma. Identity Matters in Deep Learning. In *ICLR*, 2017. 2, 9

Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, 2020. 59

Ouns El Harzli, Yoonsoo Nam, Ilja Kuzborskij, Bernardo Cuenca Grau, and Ard A. Louis. Sufficient Conditions for Stability of Minimum-Norm Interpolating Deep ReLU Networks. *CoRR*, abs/2602.13910, 2026. 9

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *CVPR*, pp. 770–778, 2016. 2

Ziwei Ji and Matus Telgarsky. Directional convergence and alignment in deep learning. In *NeurIPS*, 2020. 1, 2, 9

Nirmit Joshi, Gal Vardi, and Nathan Srebro. Noisy Interpolation Learning with Shallow Univariate ReLU Networks. In *ICLR*, 2024. 2, 9

Sungyoon Kim, Aaron Mishkin, and Mert Pilanci. Exploring The Loss Landscape Of Regularized Neural Networks Via Convex Duality. In *ICLR*, 2025. 2, 8

Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *ICLR*, 2015. 7, 59

Guy Kornowski, Gilad Yehudai, and Ohad Shamir. From Tempered to Benign Overfitting in ReLU Neural Networks. In *NeurIPS*, 2023. 2, 9

Vera Kurková and Marcello Sanguineti. Bounds on rates of variable-basis and neural-network approximation. *IEEE Trans. Inf. Theory*, 47(6):2659–2665, 2001. 1

Moshe Leshno, Vladimir Ya. Lin, Allan Pinkus, and Shimon Schocken. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural Networks*, 6(6):861–867, 1993. 1

Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the Loss Landscape of Neural Nets. In *NeurIPS*, pp. 6391–6401, 2018. 2

Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *ICLR*, 2019. 1

Kaifeng Lyu and Jian Li. Gradient Descent Maximizes the Margin of Homogeneous Neural Networks. In *ICLR*, 2020. 1, 2, 9, 16

Lachlan E. MacDonald, Jack Valmadre, Hemanth Saratchandran, and Simon Lucey. On skip connections and normalisation layers in deep optimisation. In *NeurIPS*, 2023. 2, 9

Julia B. Nakhleh and Robert D. Nowak. Global Minimizers of ℓ^p-Regularized Objectives Yield the Sparsest ReLU Neural Networks. In *NeurIPS*, 2025. 9

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. In search of the real inductive bias: On the role of implicit regularization in deep learning. In *ICLR (Workshop)*, 2015. 1

Greg Ongie, Rebecca Willett, Daniel Soudry, and Nathan Srebro. A Function Space View of Bounded Norm Infinite Width ReLU Nets: The Multivariate Case. In *ICLR*, 2020. 2, 7, 17

A. Emin Orhan and Xaq Pitkow. Skip Connections Eliminate Singularities. In *ICLR*, 2018. 2, 9

Rahul Parhi and Robert D Nowak. Banach space representer theorems for neural networks and ridge splines. *J. Mach. Learn. Res.*, 22(43):1–40, 2021. 2, 7

Jiyoung Park, Ian Pelakh, and Stephan Wojtowytsch. Minimum norm interpolation by perceptra: Explicit regularization and implicit bias. In *NeurIPS*, 2023. 8

Mert Pilanci and Tolga Ergen. Neural Networks are Convex Regularizers: Exact Polynomial-time Convex Optimization Formulations for Two-layer Networks. In *ICML*, pp. 7695–7705, 2020. 59

Itay Safran, Gal Vardi, and Jason D. Lee. On the Effective Number of Linear Regions in Shallow Univariate ReLU Networks: Convergence Guarantees and Implicit Bias. In *NeurIPS*, 2022. 2, 3, 9, 58

Pedro Savarese, Itay Evron, Daniel Soudry, and Nathan Srebro. How do infinite width bounded norm networks look in function space? In *COLT*, pp. 2667–2690, 2019. 2, 7, 17

Guy Smorodinsky, Gal Vardi, and Itay Safran. Provable Privacy Attacks on Trained Shallow Neural Networks. *Trans. Mach. Learn. Res.*, 2026. 3, 9

A. N. Tikhonov. On the solution of ill-posed problems and the method of regularization. *Dokl. Akad. Nauk SSSR*, 151(3):501–504, 1963. 1

Gal Vardi. On the Implicit Bias in Deep-Learning Algorithms. *Commun. ACM*, 66(6):86–93, 2023. 1

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In *NeurIPS*, pp. 5998–6008, 2017. 2

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. 59

Xi Wang and Laurence Aitchison. How to set AdamW's weight decay as you scale model and dataset size. In *ICML*, 2025. 1

Chen Zeno, Greg Ongie, Yaniv Blumenfeld, Nir Weinberger, and Daniel Soudry. How do Minimum-Norm Shallow Denoisers Look in Function Space? In *NeurIPS*, 2023. 2, 9

Appendix

Table of Contents

A Overview diagram of proof dependencies **14**

B Background on the Clarke subdifferential, KKT points, and MFCQ **15**

C Background on minimal-norm representors **17**

D Role of the dataset assumption **19**

E Proofs for minimal-norm interpolators **20**

F Counterexamples without the skip connection **35**

G Proofs for minimizing ℓ_2-regularized logistic loss **41**

H Proofs for limits of normalized approximate Clarke stationary points **53**

I Further details for and outcomes from the experiments **57**

A Overview diagram of proof dependencies

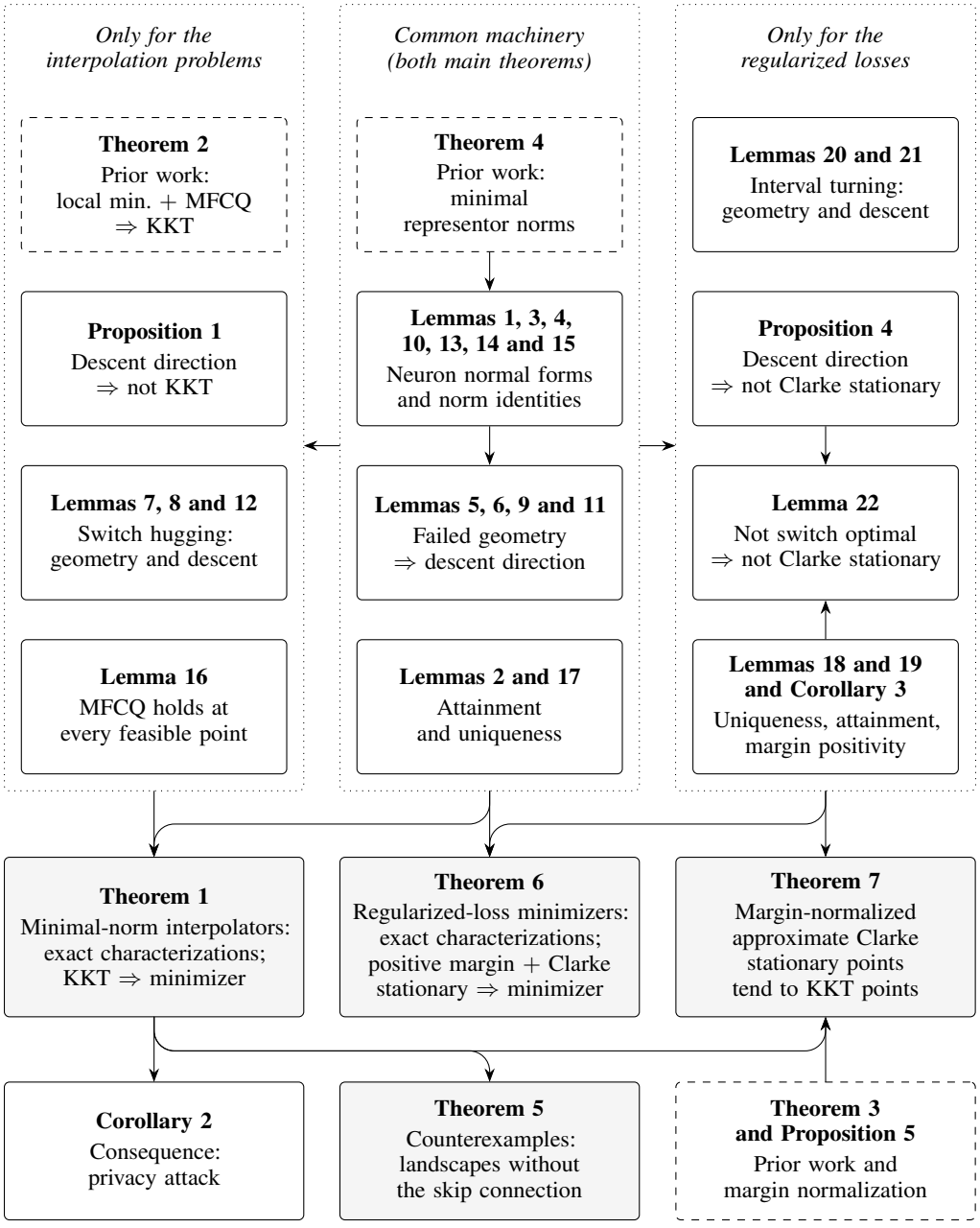


Figure 4: Each box lists in bold the results that play a single role, followed by a description of that role. The three dotted panels separate the machinery used for both Theorem 1 and Theorem 6 from what is needed only for one of them. An arrow means that what is at its head is proved using what is at its tail. Dashed boxes indicate results that are quoted or derived from prior work, and shaded boxes indicate the main theorems.

B Background on the Clarke subdifferential, KKT points, and MFCQ

For the reader's convenience, in this section we review a number of standard definitions and results (and for completeness we prove some of them) from nonsmooth analysis and first-order optimality, that we use in this work. The source of nonsmoothness in our setting is the ReLU nonlinearity, which is not differentiable at 0.

Clarke subdifferential and Clarke directional derivative. Suppose a function $F\colon \mathbb{R}^N \to \mathbb{R}$ is *locally Lipschitz*, i.e., every point $v \in \mathbb{R}^N$ has a neighborhood U such that F is Lipschitz continuous on U. By Rademacher's theorem (see, e.g., Borwein & Lewis, 2010, Theorem 9.1.2), then F is differentiable almost everywhere. The Clarke subdifferential of F at a point v is the convex hull

$$\partial F(v) \coloneqq \operatorname{conv}\Big\{ \lim_{i\to\infty} \nabla F\big(u^{(i)}\big) \;\Big|\; \lim_{i\to\infty} u^{(i)} = v \text{ and } \nabla F\big(u^{(i)}\big) \text{ exists for all } i \Big\}.$$

Its members are the subgradients, it is nonempty and compact for all v, and equals the singleton $\{\nabla F(v)\}$ if F is continuously differentiable at v (Clarke, 1975). The Clarke directional derivative of F at v along u is then the maximum

$$\mathrm{D}^\circ F(v)[u] \coloneqq \max\big\{ \langle \gamma, u\rangle \;\big|\; \gamma \in \partial F(v) \big\},$$

and it equals the ordinary directional derivative $\mathrm{D}F(v)[u]$ if F is continuously differentiable at v.

For example, the Clarke subdifferential of the ReLU function is

$$\partial\sigma(z) = \begin{cases} \{0\} & \text{if } z < 0, \\ [0,1] & \text{if } z = 0, \\ \{1\} & \text{if } z > 0, \end{cases}$$

so its Clarke directional derivatives at 0 along -1 and 1 are

$$\mathrm{D}^\circ\sigma(0)[-1] = 0 \qquad\qquad \mathrm{D}^\circ\sigma(0)[1] = 1.$$

KKT points. Suppose also $G_1, \dots, G_J\colon \mathbb{R}^N \to \mathbb{R}$ are locally Lipschitz. Following Dutta et al. (2013, section 2.2), we have that $v \in \mathbb{R}^N$ is a Karush-Kuhn-Tucker (KKT) point of the problem

$$\inf_v F(v) \quad \text{such that } \forall j \in [J],\ G_j(v) \le 0 \tag{Q}$$

if and only if there exist multipliers $\mu_1, \dots, \mu_J \ge 0$ such that:

(feasibility) $\forall j \in [J],\ G_j(v) \le 0$;

(equilibrium inclusion) $\mathbf{0} \in \partial F(v) + \sum_{j=1}^J \mu_j\, \partial G_j(v)$;

(complementary slackness) $\forall j \in [J],\ \mu_j\, G_j(v) = 0$.

The following basic proposition provides a sufficient condition for impossibility that a feasible point is KKT: existence of a direction along which the derivative of the objective is negative and the derivative of each tight constraint is nonpositive.

Proposition 1. *Suppose $v \in \mathbb{R}^N$ is feasible for prob. (Q). If there exists $u \in \mathbb{R}^N$ such that $\mathrm{D}^\circ F(v)[u] < 0$ and $\mathrm{D}^\circ G_j(v)[u] \le 0$ for all $j \in [J]$ with $G_j(v) = 0$, then v is not a KKT point of prob. (Q).*

Proof. For a contradiction, suppose that v is a KKT point of prob. (Q). Then there exist multipliers $\mu_1, \dots, \mu_J \ge 0$ such that $\mu_j\, G_j(v) = 0$ for all $j \in [J]$, and $\mathbf{0} \in \partial F(v) + \sum_{j=1}^J \mu_j\, \partial G_j(v)$. Hence $\mathbf{0} \in \partial F(v) + \sum_{G_j(v)=0} \mu_j\, \partial G_j(v)$, so there exist a subgradient $\varphi \in \partial F(v)$ and subgradients $\gamma_j \in \partial G_j(v)$ for each $j \in [J]$ with $G_j(v) = 0$, such that $\mathbf{0} = \varphi + \sum_{G_j(v)=0} \mu_j \gamma_j$. But then

$$0 = \left\langle \varphi + \sum_{G_j(v)=0} \mu_j \gamma_j, u \right\rangle$$

$$
\begin{aligned}
&= \langle \varphi, u \rangle + \sum_{G_j(v)=0} \mu_j \langle \gamma_j, u \rangle \\
&\leq \mathrm{D}^\circ F(v)[u] + \sum_{G_j(v)=0} \mu_j \mathrm{D}^\circ G_j(v)[u] \\
&< 0. \qquad \square
\end{aligned}
$$

MFCQ. Moreover, we say that a feasible point $v \in \mathbb{R}^N$ satisfies Mangasarian-Fromovitz constraint qualification (MFCQ) if and only if there exists $u \in \mathbb{R}^N$ such that $\mathrm{D}^\circ G_j(v)[u] < 0$ for all $j \in [J]$ with $G_j(v) = 0$.

Theorem 2 (corollary of, e.g., Giorgi et al. 2023, Theorem 10.61)**.** *If a local minimizer of prob. (*Q*) satisfies MFCQ, then it is a KKT point.*

Approximate KKT points and their limits. Following Lyu & Li (2020, appendix C.1), supposing that $F, G_1, \ldots, G_J \colon \mathbb{R}^N \to \mathbb{R}$ are locally Lipschitz and $\varepsilon, \delta > 0$, we say that $v \in \mathbb{R}^N$ is an ε, δ-KKT point of prob. (Q) if and only if there exist multipliers $\mu_1, \ldots, \mu_J \geq 0$ such that:[4]

$$
\begin{aligned}
&\textbf{(feasibility)} && \forall j \in [J],\ G_j(v) \leq 0; \\
&\textbf{(approximate equilibrium inclusion)} && \varepsilon \mathbb{B}_2^N \cap \left(\partial F(v) + \textstyle\sum_{j=1}^J \mu_j\, \partial G_j(v) \right) \neq \emptyset; \\
&\textbf{(approximate complementary slackness)} && \textstyle\sum_{j=1}^J \mu_j\, G_j(v) \geq -\delta.
\end{aligned}
$$

Theorem 3 (corollary of Dutta et al. 2013, Theorem 3.6)**.** *Suppose that $v_k \in \mathbb{R}^N$ is an ε_k, δ_k-KKT point of prob. (*Q*) for all $k \in \mathbb{N}$, that $v_k \to v_\star$, $\varepsilon_k \to 0$, and $\delta_k \to 0$ as $k \to \infty$, and that $v_\star$ satisfies MFCQ. Then $v_\star$ is a KKT point of prob. (*Q*).*

[4] We write $\mathbb{B}_2^N$ for the closed Euclidean unit ball in $\mathbb{R}^N$.

Table 2: Given a CPA $g\colon \mathbb{R} \to \mathbb{R}$, we consider these four variants of the representor norm minimization problem. The skip connection is always "free", i.e., its parameters a_0 and b_0 are never penalized. Note also that the network width m is optimized, being determined by the dimension of the parameters vector $\theta = (a_j, w_j, b_j)_{j=1}^m \in \mathbb{R}^{3m}$.

	biases not penalized		**biases penalized**	
skip connection absent	$\inf_\theta \frac{1}{2}\sum_{j=1}^m (a_j^2 + w_j^2)$ such that $f_\theta = g$	(R)	$\inf_\theta \frac{1}{2}\sum_{j=1}^m (a_j^2 + w_j^2 + b_j^2)$ such that $f_\theta = g$	(R^b)
skip connection present	$\inf_{\theta, a_0, b_0} \frac{1}{2}\sum_{j=1}^m (a_j^2 + w_j^2)$ such that $f_{\theta,a_0,b_0} = g$	(R_1)	$\inf_{\theta, a_0, b_0} \frac{1}{2}\sum_{j=1}^m (a_j^2 + w_j^2 + b_j^2)$ such that $f_{\theta,a_0,b_0} = g$	(R_1^b)

C Background on minimal-norm representors

After a couple of standard warm-up propositions on the functional expressive power of univariate two-layer ReLU networks, here we collate and state with the same notation the main known results on the representor norm minimization problems in Table 2 (cf. question 1 in section 1), which we build on in this work.

Proposition 2. *The class of functions representable by a network remains the same without a skip connection, i.e., if $a_0 = 0$ and $b_0 = 0$.*

Proof. Observe that $a_0 x + b_0 = \sigma(a_0 x + b_0) - \sigma(-a_0 x - b_0)$. □

Proposition 3. *A function from $\mathbb{R}$ to $\mathbb{R}$ is representable by a two-layer ReLU network if and only if it is CPA.*

Proof. For every two-layer ReLU network, the function it represents is CPA since this class contains constant functions and the identity function, and is closed under linear combinations and under applying the ReLU nonlinearity.

Suppose $g\colon \mathbb{R} \to \mathbb{R}$ is CPA, with kinks $(p_k)_{k=1}^q$ and slopes $(s_k)_{k=0}^q$. Then we have, for all $x \in \mathbb{R}$,

$$g(x) = s_0(x - p_1) + g(p_1) + \sum_{k=1}^{q} (s_k - s_{k-1})\,\sigma(x - p_k). \qquad \square$$

Theorem 4. *[corollary of Savarese et al. 2019, Theorem 3.1, Ongie et al. 2020, Lemma 1, and Boursier & Flammarion 2023, Theorem 4] The minimal norms of representors of a CPA $g\colon \mathbb{R} \to \mathbb{R}$ with kinks $(p_k)_{k=1}^q$ and slopes $(s_k)_{k=0}^q$ are:*

prob. (R)**:** $\max\{\sum_{k=1}^q |s_k - s_{k-1}|, |s_0 + s_q|\}$*;*

prob. (R_1)**:** $\sum_{k=1}^q |s_k - s_{k-1}|$*;*

prob. (R^b)**:** $\sum_{k=1}^q \sqrt{1 + p_k^2}\,|s_k - s_{k-1}|$
$+ \inf_{(\varphi_k)_{k=1}^q \in [-1,1]^q} h\Big(\sum_{k=1}^q \varphi_k(s_k - s_{k-1}) - (s_0 + s_q),$
$\sum_{k=1}^q (|p_k| - \varphi_k p_k)(s_k - s_{k-1}) - 2g(0)\Big)$

where $h(z_1, z_2) \coloneqq \begin{cases} \sqrt{z_1^2 + z_2^2} & \text{if } |z_1| \geq \sqrt{3}\,|z_2|, \\ \frac{\sqrt{3}}{2}|z_1| + \frac{1}{2}|z_2| & \text{if } |z_1| < \sqrt{3}\,|z_2|; \end{cases}$

prob. ($\mathrm{R}_1^{\mathrm{b}}$)**:** $\sum_{k=1}^{q} \sqrt{1+p_k^2}\,|s_k - s_{k-1}|$.

Moreover, for probs. (R_1*), (*R^{b}*) and (*$\mathrm{R}_1^{\mathrm{b}}$*), the minima are attained; and if prob. (*R*) is modified to*

$$\inf_{\theta, b_0} \frac{1}{2} \sum_{j=1}^{m} (a_j^2 + w_j^2) \quad \textit{such that } f_{\theta,0,b_0} = g,$$

i.e., a free output bias is allowed, then the minimum is unchanged and it is attained.

D ROLE OF THE DATASET ASSUMPTION

Assumption 1 combines two logically distinct conditions:

$$r \geq 3 \qquad \text{and} \qquad x_{\iota(2)} < 0 < x_{\iota(r-1)+1}.$$

The first is a lower bound on the number of label switches. The second places the first two same-label segments strictly to the left of the origin and the last two strictly to its right. In this section, we record separately the roles of these two conditions.

At least three label switches. This condition is used in Lemma 1. A positive-margin classifier then has both a negative and a positive turning interval, which allows its free affine skip connection to be eliminated without increasing either representor norm. Consequently, $\mathrm{R}(g) = \mathrm{R}_1(g)$ and $\mathrm{R}^{\mathrm{b}}(g) = \mathrm{R}_1^{\mathrm{b}}(g)$ for every positive-margin CPA g. This is the step that transfers the function-space characterizations obtained with a skip connection to the corresponding problems without one. The weaker condition $r \geq 2$ suffices for the compactness arguments in Lemmas 2 and 19 and for the finite-dimensional problems in Lemma 18.

The sign condition. The inequalities $x_{\iota(2)} < 0 < x_{\iota(r-1)+1}$ are used only when the biases are penalized. Their substantive use is in the outer-kink cases of Lemma 11 and, through the same argument, Lemma 12. For example, in the left outer-kink case one has $p_1 < 0$, and one chooses $\widehat{p} \leq p_1$ so that $-w_j^2(p_1\widehat{p}+1) < 0$. The right outer-kink case is symmetric. No corresponding sign condition is needed when the hidden biases are not penalized, because the perturbations in Lemmas 6 and 8 have regularizer derivative $-a_j^2 < 0$, independently of the kink location. Accordingly, subject only to $r \geq 3$, the conclusions concerning unpenalized biases are unaffected if the sign condition is dropped. These include Theorem 1 (i), the part of Theorem 1 (iii) concerning prob. (P_1), Theorem 6 (i), the part of Theorem 6 (iii) concerning L_1^{λ}, and the corresponding unpenalized-bias conclusions of Corollary 1 and Theorem 7.

By contrast, for penalized biases, the present proofs use the sign condition to exclude certain outer geometric irregularities. Without it, the interior arguments remain valid, including the exclusion of multiple kinks within an intermediate same-label segment, but the present proofs do not establish the full convexity-correct and switch-hugging characterizations. Thus, without additional treatment of the outer cases, the affected statements are Theorem 1 (ii), the part of Theorem 1 (iii) concerning prob. $(\mathrm{P}_1^{\mathrm{b}})$, Theorem 6 (ii), the part of Theorem 6 (iii) concerning $L_1^{\lambda,\mathrm{b}}$, and the penalized-bias conclusions derived from them.

Finally, the sign condition is coordinate dependent when the biases are penalized. Translating x by c moves a kink from p to $p-c$, and changes its contribution to the penalized representor norm from $\sqrt{1+p^2}\,|\Delta s|$ to $\sqrt{1+(p-c)^2}\,|\Delta s|$. Thus, in the penalized-bias setting, an additive translation can enforce the sign condition but is not a without-loss-of-generality reparameterization of the same optimization problem. In the unpenalized-bias setting, the representor cost depends only on the slope changes, and the corresponding conclusions are translation invariant.

E PROOFS FOR MINIMAL-NORM INTERPOLATORS

After restating Theorem 1 from section 3 and showing a privacy attack corollary, we develop and prove a sequence of lemmas culminating in a proof of the theorem.

Theorem 1. *Under Assumption 1, for the problems in Table 1, all of the following hold.*

(i) Problems (P) and (P_1) have the same minimizers in function space, which are exactly all CPAs $g\colon \mathbb{R} \to \mathbb{R}$ that are:

(switch hugging) *at every label switch in the dataset, g passes through the two incident points, i.e., for all $l \in [r]$, we have $g(x_{\iota(l)}) = y_{\iota(l)}$ and $g(x_{\iota(l)+1}) = y_{\iota(l)+1}$;*

(convexity correct) *every kink p_k of g is contained in an intermediate same-label segment of the dataset, and its convexity is opposite to the segment's label, i.e., there exists $l \in [r-1]$ with $p_k \in [x_{\iota(l)+1}, x_{\iota(l+1)}]$ and $sgn(s_k - s_{k-1}) = -y_{\iota(l)+1}$.*

(ii) Problems (P^{b}) and ($\mathrm{P}_1^{\mathrm{b}}$) have the same unique minimizer g in function space, which is of the form as in part (i) and moreover:

(single turning) *every intermediate same-label segment contains exactly one kink of g.*

Thus g has $r-1$ kinks in total, which makes it a sparsest interpolant of the dataset.

(iii) For probs. (P_1) and ($\mathrm{P}_1^{\mathrm{b}}$), every KKT point $(\theta, a_0, b_0) \in \mathbb{R}^{3m+2}$ (with any network width m for which KKT points exist) is a minimizer (globally, including over all m).

As a corollary of the exact geometric characterizations and the skip connection's sanitizing effects in Theorem 1, we obtain a powerful privacy attack which is able to extract reliably all training points that are incident to the label switches, given only the classifier in function space (knowledge of the network's parameters is not needed).

Corollary 2. *Under Assumption 1, suppose θ, a_0, b_0 is either a minimizer of prob. (P) or prob. (P^{b}), or a KKT point of prob. (P_1) or prob. ($\mathrm{P}_1^{\mathrm{b}}$). Let X be the set of all $x \in \mathbb{R}$ such that $f_{\theta,a_0,b_0}(x) = \pm 1$ and f_{θ,a_0,b_0} is not constant on any open interval containing x. Then all elements of X are training inputs. More precisely, X consists exactly of: the last input in the first same-label segment of the training dataset, both first and last inputs in every intermediate segment, and the first input in the last segment. In particular, the number of switches of $f_{\theta,a_0,b_0}(x)$ as x iterates through a sorted enumeration of X equals the number r of label switches in the training dataset.*

Example 1. In Figure 1, the set X of inputs extracted as in Corollary 2 from any of the three CPAs g_{cyan}, g_{lime}, or g_{pink}, is shown by the bold dots colored brown.

Proof. By Theorem 1, f_{θ,a_0,b_0} is switch hugging, and so it equals ± 1 at each $x \in \mathbb{R}$ which is one of: the last input in the first same-label segment of the training dataset, both first and last inputs in every intermediate segment, and the first input in the last segment. For each such x, by the same property of f_{θ,a_0,b_0}, it is not constant on any open interval containing x. Hence it remains to show that the set X contains no other values, so for a contradiction suppose that $x \in X$ is none of: the last input in the first same-label segment of the training dataset, either first or last input in some intermediate segment, or the first input in the last segment. Since by Theorem 1, f_{θ,a_0,b_0} is also convexity correct, necessarily for some intermediate segment with first input x' and last input x'' we have $x' < x < x''$ and $f_{\theta,a_0,b_0}(x') = f_{\theta,a_0,b_0}(x) = f_{\theta,a_0,b_0}(x'')$. But then f_{θ,a_0,b_0} must be constant on the interval $[x', x'']$. □

We now start the sequence of lemmas that leads to the proof of Theorem 1, the first of which shows that, for positive-margin CPAs on datasets with at least three label switches (which is implied by Assumption 1), the skip connection does not lead to lower representational cost. This fact, which is surprising in view of Theorem 4, can be proved either by arguing directly in terms of the minimal costs given by Theorem 4, or as we opt to do here, by showing how the skip connection can be eliminated through constructions at the network level that do not increase the costs. We also remark that the constructions we provide are chosen for their clarity, and can be optimized further to reduce the number of additional neurons to a small constant. Before stating and proving the lemma, we introduce notation for the costs and define the margin precisely.

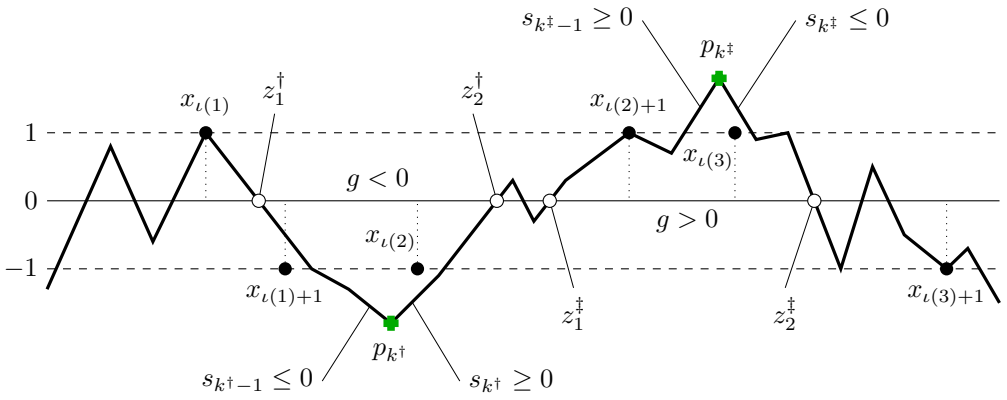


Figure 5: Here we illustrate some of the named quantities in the proof of Lemma 1 and their properties. The bold dots indicate the training points incident to the first three label switches in the dataset, the thick line shows the CPA g, the hollow dots indicate its identified crossings of the x-axis, and the green pluses indicate the two chosen kinks.

For a CPA $g\colon \mathbb{R} \to \mathbb{R}$, we write $\mathrm{R}(g)$, $\mathrm{R}_1(g)$, $\mathrm{R}^{\mathrm{b}}(g)$, and $\mathrm{R}_1^{\mathrm{b}}(g)$ for the minimal norms of its representors that are given in the four cases in Theorem 4, respectively.

Also for a CPA $g\colon \mathbb{R} \to \mathbb{R}$, we write $M(g) \coloneqq \min_{i=1}^n y_i\, g(x_i)$ for its margin over the dataset $(x_i, y_i)_{i=1}^n$. Then for a network θ, a_0, b_0, we may write $M(\theta, a_0, b_0)$ for the margin $M(f_{\theta,a_0,b_0})$ of its function; and when the skip connection is absent, we may write just $M(\theta)$ instead of $M(\theta, 0, 0)$.

Lemma 1. *Suppose the dataset has at least three label switches, i.e., $r \geq 3$, and a CPA $g\colon \mathbb{R} \to \mathbb{R}$ has a positive margin. Then $\mathrm{R}(g) = \mathrm{R}_1(g)$ and $\mathrm{R}^{\mathrm{b}}(g) = \mathrm{R}_1^{\mathrm{b}}(g)$. Moreover, the minimum of prob. (R) is attained.*

Proof. For an illustration of a part of this argument, we refer the reader to Figure 5.

Let $(p_k)_{k=1}^q$ be the kinks of g and $(s_k)_{k=0}^q$ be the slopes of g.

Without loss of generality, $y_1 = 1$. Since $r \geq 3$, we have at least the following three label switches:

$$1 = y_{\iota(1)} \neq y_{\iota(1)+1} = -1 = y_{\iota(2)} \neq y_{\iota(2)+1} = 1 = y_{\iota(3)} \neq y_{\iota(3)+1} = -1.$$

Recalling that $M(g) > 0$, there must exist points $z_1^\dagger \in (x_{\iota(1)}, x_{\iota(1)+1})$ and $z_2^\dagger \in (x_{\iota(1)+1}, x_{\iota(2)+1})$ such that

$$g(z_1^\dagger) = 0 = g(z_2^\dagger) \qquad\qquad g(z) < 0 \text{ for all } z \in (z_1^\dagger, z_2^\dagger),$$

and there must exist points $z_1^\ddagger \in (x_{\iota(2)}, x_{\iota(2)+1})$ and $z_2^\ddagger \in (x_{\iota(2)+1}, x_{\iota(3)+1})$ such that

$$g(z_1^\ddagger) = 0 = g(z_2^\ddagger) \qquad\qquad g(z) > 0 \text{ for all } z \in (z_1^\ddagger, z_2^\ddagger).$$

Then necessarily $z_2^\dagger \leq z_1^\ddagger$. Also it follows that g must have kinks $p_{k^\dagger} \in (z_1^\dagger, z_2^\dagger)$ and $p_{k^\ddagger} \in (z_1^\ddagger, z_2^\ddagger)$ such that their incident slopes satisfy

$$s_{k^\dagger-1} \leq 0 \qquad s_{k^\dagger} \geq 0 \qquad s_{k^\ddagger-1} \geq 0 \qquad s_{k^\ddagger} \leq 0.$$

For each $k \in [q]$, let:

$$\mu_k^\dagger \coloneqq \begin{cases} s_k - s_{k-1} & \text{if } k < k^\dagger, \\ -s_{k^\dagger-1} & \text{if } k = k^\dagger, \\ 0 & \text{if } k > k^\dagger; \end{cases} \qquad \nu_k^\dagger \coloneqq \begin{cases} 0 & \text{if } k < k^\dagger, \\ s_{k^\dagger} & \text{if } k = k^\dagger, \\ s_k - s_{k-1} & \text{if } k > k^\dagger. \end{cases}$$

Then let $\theta^\dagger$ consist of neuron parameters that give the network function

$$f_{\theta^\dagger}(x) = \sum_{k=1}^q \Big(\operatorname{sgn}(\mu_k^\dagger) \sqrt{|\mu_k^\dagger|}\, \sigma\big(-\sqrt{|\mu_k^\dagger|}(x - p_k)\big) + \operatorname{sgn}(\nu_k^\dagger) \sqrt{|\nu_k^\dagger|}\, \sigma\big(\sqrt{|\nu_k^\dagger|}(x - p_k)\big) \Big).$$

It is straightforward to check that $f_{\theta^\dagger}$ has the same kinks and slopes as g, and that $f_{\theta^\dagger}(p_{k^\dagger}) = 0$, which means that $f_{\theta^\dagger}(x) = g(x) - g(p_{k^\dagger})$ for all $x \in \mathbb{R}$.

Analogously, for each $k \in [q]$, let:

$$\mu_k^\ddagger := \begin{cases} s_k - s_{k-1} & \text{if } k < k^\ddagger, \\ -s_{k^\ddagger - 1} & \text{if } k = k^\ddagger, \\ 0 & \text{if } k > k^\ddagger; \end{cases} \qquad \nu_k^\ddagger := \begin{cases} 0 & \text{if } k < k^\ddagger, \\ s_{k^\ddagger} & \text{if } k = k^\ddagger, \\ s_k - s_{k-1} & \text{if } k > k^\ddagger. \end{cases}$$

Then let $\theta^\ddagger$ consist of neuron parameters that give the network function

$$f_{\theta^\ddagger}(x) = \sum_{k=1}^{q} \Big(\operatorname{sgn}(\mu_k^\ddagger) \sqrt{|\mu_k^\ddagger|}\, \sigma\big(-\sqrt{|\mu_k^\ddagger|}(x - p_k)\big) + \operatorname{sgn}(\nu_k^\ddagger) \sqrt{|\nu_k^\ddagger|}\, \sigma\big(\sqrt{|\nu_k^\ddagger|}(x - p_k)\big) \Big).$$

It is straightforward to check that $f_{\theta^\ddagger}$ has the same kinks and slopes as g, and that $f_{\theta^\ddagger}(p_{k^\ddagger}) = 0$, which means that $f_{\theta^\ddagger}(x) = g(x) - g(p_{k^\ddagger})$ for all $x \in \mathbb{R}$.

Now, recalling that $g(p_{k^\dagger}) < 0$ and $g(p_{k^\ddagger}) > 0$, let

$$\xi := \frac{g(p_{k^\ddagger})}{g(p_{k^\ddagger}) - g(p_{k^\dagger})} \in (0, 1).$$

Then we have $g = \xi f_{\theta^\dagger} + (1 - \xi) f_{\theta^\ddagger}$, and hence $g = f_{\theta^\star}$, where $\theta^\star := \big(\sqrt{\xi}\,\theta^\dagger, \sqrt{1-\xi}\,\theta^\ddagger\big)$, i.e., the convex combination of network functions $f_{\theta^\dagger}$ and $f_{\theta^\ddagger}$ is implemented by multiplying the network parameters by $\sqrt{\xi}$ and $\sqrt{1-\xi}$ respectively, and concatenating the results.

Therefore

$$\begin{aligned} \mathrm{R}(g) &\le \xi \sum_{k=1}^{q} \big(|\mu_k^\dagger| + |\nu_k^\dagger|\big) + (1 - \xi) \sum_{k=1}^{q} \big(|\mu_k^\ddagger| + |\nu_k^\ddagger|\big) \\ &= \xi \sum_{k=1}^{q} |s_k - s_{k-1}| + (1 - \xi) \sum_{k=1}^{q} |s_k - s_{k-1}| \\ &= \sum_{k=1}^{q} |s_k - s_{k-1}| \\ &= \mathrm{R}_1(g), \end{aligned}$$

where: the inequality is by the definitions of $\theta^\star$, $\theta^\dagger$, and $\theta^\ddagger$; the first equality is by the definitions of $\mu_k^\dagger$, $\nu_k^\dagger$, $\mu_k^\ddagger$, and $\nu_k^\ddagger$ as well as by recalling that $|s_{k^\dagger} - s_{k^\dagger - 1}| = |s_{k^\dagger - 1}| + |s_{k^\dagger}|$ and $|s_{k^\ddagger} - s_{k^\ddagger - 1}| = |s_{k^\ddagger - 1}| + |s_{k^\ddagger}|$; and the third equality is by Theorem 4.

Hence $\mathrm{R}(g) = \mathrm{R}_1(g)$ also by Theorem 4, so the last inequality above is an equality, and thus the minimum of prob. (R) is attained by the network parameters $\theta^\star$.

Finally, by considering the following alternative expressions

$$\begin{aligned} f_{\theta^\dagger}(x) = \sum_{k=1}^{q} \Bigg(& \operatorname{sgn}(\mu_k^\dagger) \sqrt[4]{1 + p_k^2} \sqrt{|\mu_k^\dagger|}\, \sigma\left(-\frac{\sqrt{|\mu_k^\dagger|}}{\sqrt[4]{1 + p_k^2}} (x - p_k) \right) \\ & + \operatorname{sgn}(\nu_k^\dagger) \sqrt[4]{1 + p_k^2} \sqrt{|\nu_k^\dagger|}\, \sigma\left(\frac{\sqrt{|\nu_k^\dagger|}}{\sqrt[4]{1 + p_k^2}} (x - p_k) \right) \Bigg) \\ f_{\theta^\ddagger}(x) = \sum_{k=1}^{q} \Bigg(& \operatorname{sgn}(\mu_k^\ddagger) \sqrt[4]{1 + p_k^2} \sqrt{|\mu_k^\ddagger|}\, \sigma\left(-\frac{\sqrt{|\mu_k^\ddagger|}}{\sqrt[4]{1 + p_k^2}} (x - p_k) \right) \\ & + \operatorname{sgn}(\nu_k^\ddagger) \sqrt[4]{1 + p_k^2} \sqrt{|\nu_k^\ddagger|}\, \sigma\left(\frac{\sqrt{|\nu_k^\ddagger|}}{\sqrt[4]{1 + p_k^2}} (x - p_k) \right) \Bigg), \end{aligned}$$

we obtain

$$\begin{aligned}\mathrm{R}^{\mathrm{b}}(g) &\le \xi \sum_{k=1}^{q} \sqrt{1+p_k^2}\left(|\mu_k^\dagger| + |\nu_k^\dagger|\right) + (1-\xi)\sum_{k=1}^{q}\sqrt{1+p_k^2}\left(|\mu_k^\ddagger| + |\nu_k^\ddagger|\right) \\ &= \xi \sum_{k=1}^{q} \sqrt{1+p_k^2}\,|s_k - s_{k-1}| + (1-\xi)\sum_{k=1}^{q}\sqrt{1+p_k^2}\,|s_k - s_{k-1}| \\ &= \sum_{k=1}^{q} \sqrt{1+p_k^2}\,|s_k - s_{k-1}| \\ &= \mathrm{R}_1^{\mathrm{b}}(g),\end{aligned}$$

so by Theorem 4 we have $\mathrm{R}^{\mathrm{b}}(g) = \mathrm{R}_1^{\mathrm{b}}(g)$. □

The next lemma shows that the optima of the interpolator norm minimization problems with the skip connection are attained, even for any fixed network width for which interpolation is possible (i.e., $m \ge r-1$).

Lemma 2. *For each of probs. (P_1) and ($\mathrm{P}_1^{\mathrm{b}}$) restricted to any network width $m \in \mathbb{N}$ that is feasible, the minimum is attained.*

Proof. If the dataset has no label switch, then a constant skip connection interpolates it. If it has exactly one label switch, then an affine skip connection interpolates it. In either case, the value zero is attained in both problems. We may therefore assume that $r \ge 2$.

Fix a feasible width m, and consider prob. (P_1). Let $\left(\theta^{(k)}, a_0^{(k)}, b_0^{(k)}\right)_{k\in\mathbb{N}}$ be a minimizing sequence, then the maximum N of its objective values is finite. We first replace each term of the sequence by an equivalent parameterization, without changing its values at the training inputs and without increasing its objective, as follows.

For every neuron with $w_j^{(k)} \ne 0$, write $p_j^{(k)} := -b_j^{(k)}/w_j^{(k)}$ for its kink. If $p_j^{(k)} \notin [x_1, x_n]$, then $x \mapsto a_j^{(k)}\,\sigma\big(w_j^{(k)}x + b_j^{(k)}\big)$ is affine on $[x_1, x_n]$, so its contribution on all training inputs can be absorbed into the free affine skip connection, after which the neuron can be replaced by the zero neuron. Similarly, if $w_j^{(k)} = 0$, then the neuron contributes a constant function, which can be absorbed into the skip bias. Thus, we may assume that every nonzero neuron satisfies $p_j^{(k)} \in [x_1, x_n]$.

By positive homogeneity of the ReLU, each nonzero neuron can also be rescaled, without changing the represented function, so that $|a_j^{(k)}| = |w_j^{(k)}|$. This rescaling minimizes $\frac{1}{2}\Big(\big(a_j^{(k)}\big)^2 + \big(w_j^{(k)}\big)^2\Big)$ over the positive-rescaling orbit of the neuron, and hence does not increase the objective. Consequently, $|a_j^{(k)}|, |w_j^{(k)}| \le \sqrt{2N}$, and also $|b_j^{(k)}| = |w_j^{(k)} p_j^{(k)}| \le \sqrt{2N}\,\max_{i=1}^n |x_i|$.

It follows that the parameter vectors $\theta^{(k)}$ lie in a bounded set. In particular, the pre-skip output vectors $h^{(k)} := \big(f_{\theta^{(k)},0,0}(x_i)\big)_{i=1}^n$ form a bounded sequence in $\mathbb{R}^n$. We next show that the skip parameters are bounded.

Suppose, to the contrary, that after passing to a subsequence, $\big\|a_0^{(k)}, b_0^{(k)}\big\|_2 \to \infty$. Passing to a further subsequence, we may assume that $\big(a_0^{(k)}, b_0^{(k)}\big)/\big\|a_0^{(k)}, b_0^{(k)}\big\|_2 \to (\alpha_0, \beta_0)$ for some unit vector (α_0, β_0). Feasibility gives $y_i\big(a_0^{(k)}x_i + b_0^{(k)} + h_i^{(k)}\big) \ge 1$ for every $i \in [n]$. Dividing by $\big\|a_0^{(k)}, b_0^{(k)}\big\|_2$ and letting $k \to \infty$, using boundedness of $h^{(k)}$, yields $y_i(\alpha_0 x_i + \beta_0) \ge 0$ for every $i \in [n]$. Because $r \ge 2$, there exist three training inputs $x_{i_1} < x_{i_2} < x_{i_3}$ whose labels alternate. Multiplying the affine function $x \mapsto \alpha_0 x + \beta_0$ by the common label of the first and third points if necessary, we obtain

$$\alpha_0 x_{i_1} + \beta_0 \ge 0 \qquad \alpha_0 x_{i_2} + \beta_0 \le 0 \qquad \alpha_0 x_{i_3} + \beta_0 \ge 0.$$

Since x_{i_2} is a strict convex combination of x_{i_1} and x_{i_3}, the value $\alpha_0 x_{i_2} + \beta_0$ is the same strict convex combination of $\alpha_0 x_{i_1} + \beta_0$ and $\alpha_0 x_{i_3} + \beta_0$. It is therefore nonnegative, and hence it must equal zero. Since both outer values are nonnegative, equality of their strict convex combination to zero forces both of them to be zero. Thus the affine function vanishes at two distinct points and is consequently

identically zero. Hence $\alpha_0 = \beta_0 = 0$, contradicting that (α_0, β_0) is a unit vector. It follows that the skip parameters are bounded.

We have therefore obtained a minimizing sequence contained in a compact subset of $\mathbb{R}^{3m+2}$, so we may pass to a convergent subsequence. The feasible set is closed because the network outputs at the training inputs depend continuously on the parameters. Hence the limit is feasible. Also the objective is continuous, so its value at the limit is the infimum. Thus the minimum of prob. (P_1), restricted to width m, is attained.

The proof for prob. ($\mathrm{P}_1^{\mathrm{b}}$) is analogous and slightly simpler. By positive homogeneity, every nonzero neuron may be rescaled so that $|a_j^{(k)}| = \sqrt{\big(w_j^{(k)}\big)^2 + \big(b_j^{(k)}\big)^2}$. This rescaling minimizes $\frac{1}{2}\Big(\big(a_j^{(k)}\big)^2 + \big(w_j^{(k)}\big)^2 + \big(b_j^{(k)}\big)^2\Big)$ over the positive-rescaling orbit of the neuron. The bounded objective of a minimizing sequence therefore directly bounds all coordinates of $\theta^{(k)}$. The pre-skip outputs on the finite training set are consequently bounded. The same normalization argument using three ordered inputs with alternating labels then shows that the skip parameters are bounded. Thus a minimizing sequence can again be chosen in a compact set, and closedness of the feasible set and continuity of the objective imply that the minimum of prob. ($\mathrm{P}_1^{\mathrm{b}}$), restricted to width m, is attained. □

The orientation of a hidden neuron is not intrinsic when a free affine skip connection is present. Reversing its hidden weight and bias changes its ReLU contribution only by an affine function, which can be absorbed exactly by the skip connection. The next lemma records this reparameterization and, crucially, shows that it also transports arbitrary parameter-space directions while preserving the Clarke directional derivatives of the network outputs, their negatives, and both regularizers considered in this work.[5] Consequently, in the subsequent descent constructions we may choose the orientations of the relevant neurons without loss of generality, including when a training input coincides with one of their kinks.

Lemma 3. *Suppose $J \subseteq [m]$, $\widehat{\theta} \coloneqq \big(\widehat{a}_j, \widehat{w}_j, \widehat{b}_j\big)_{j=1}^m$ is obtained from θ by reversing the orientations of the hidden neurons J, i.e.,*

$$\big(\widehat{a}_j, \widehat{w}_j, \widehat{b}_j\big) \coloneqq \begin{cases} (a_j, -w_j, -b_j) & \text{if } j \in J, \\ (a_j, w_j, b_j) & \text{if } j \notin J, \end{cases}$$

and

$$\widehat{a}_0 \coloneqq a_0 + \sum_{j \in J} a_j w_j \qquad\qquad \widehat{b}_0 \coloneqq b_0 + \sum_{j \in J} a_j b_j.$$

(i) We have that $f_{\widehat{\theta}, \widehat{a}_0, \widehat{b}_0} = f_{\theta, a_0, b_0}$.

(ii) For every direction $\widehat{\theta}', \widehat{a}_0', \widehat{b}_0'$, letting $\theta' \coloneqq (a_j', w_j', b_j')_{j=1}^m$ be obtained from $\widehat{\theta}'$ by reversing the orientations at the hidden neurons J, i.e.,

$$(a_j', w_j', b_j') \coloneqq \begin{cases} \big(\widehat{a}_j', -\widehat{w}_j', -\widehat{b}_j'\big) & \text{if } j \in J, \\ \big(\widehat{a}_j', \widehat{w}_j', \widehat{b}_j'\big) & \text{if } j \notin J, \end{cases}$$

and

$$a_0' \coloneqq \widehat{a}_0' + \sum_{j \in J} \big(\widehat{a}_j' \widehat{w}_j + \widehat{a}_j \widehat{w}_j'\big) \qquad\qquad b_0' \coloneqq \widehat{b}_0' + \sum_{j \in J} \big(\widehat{a}_j' \widehat{b}_j + \widehat{a}_j \widehat{b}_j'\big),$$

we have that, for all $x \in \mathbb{R}$,

$$\mathrm{D}^\circ\big(f_{\theta, a_0, b_0}(x)\big)[\theta', a_0', b_0'] = \mathrm{D}^\circ\big(f_{\widehat{\theta}, \widehat{a}_0, \widehat{b}_0}(x)\big)\big[\widehat{\theta}', \widehat{a}_0', \widehat{b}_0'\big]$$
$$\mathrm{D}^\circ\big(-f_{\theta, a_0, b_0}(x)\big)[\theta', a_0', b_0'] = \mathrm{D}^\circ\big(-f_{\widehat{\theta}, \widehat{a}_0, \widehat{b}_0}(x)\big)\big[\widehat{\theta}', \widehat{a}_0', \widehat{b}_0'\big]$$
$$\mathrm{D}\Big(\frac{1}{2}\sum_{j=1}^m (a_j^2 + w_j^2)\Big)[\theta', a_0', b_0'] = \mathrm{D}\Big(\frac{1}{2}\sum_{j=1}^m \big(\widehat{a}_j^2 + \widehat{w}_j^2\big)\Big)\big[\widehat{\theta}', \widehat{a}_0', \widehat{b}_0'\big]$$

[5] The two regularizers are everywhere continuously differentiable, so their Clarke directional derivatives equal their ordinary directional derivatives.

$$\mathrm{D}\Big(\frac{1}{2}\sum_{j=1}^{m}(a_j^2+w_j^2+b_j^2)\Big)[\theta',a_0',b_0'] = \mathrm{D}\Big(\frac{1}{2}\sum_{j=1}^{m}(\widehat{a}_j^2+\widehat{w}_j^2+\widehat{b}_j^2)\Big)[\widehat{\theta}',\widehat{a}_0',\widehat{b}_0'].$$

Proof. Part (i) follows by applying at each $j \in J$ the identity

$$a_j\,\sigma(w_jx+b_j) - a_j\,\sigma(-w_jx-b_j) = a_jw_jx + a_jb_j \quad \text{for all } x\in\mathbb{R}.$$

In part (ii), the equalities of the directional derivatives of the square ℓ_2-norms (without and with the biases) are straightforward and do not involve the skip connection.

The remaining two equations are symmetric, so we show the former one, at an arbitrary $x \in \mathbb{R}$:

$$\begin{aligned}
&\mathrm{D}^\circ\big(f_{\theta,a_0,b_0}(x)\big)[\theta',a_0',b_0'] \\
&= \sum_{j|w_jx+b_j=0} \max\{0, a_j(w_j'x+b_j')\} \\
&\quad + \sum_{j|w_jx+b_j>0} \big(a_j'(w_jx+b_j) + a_j(w_j'x+b_j')\big) \\
&\quad + a_0'x + b_0' \\
&= \sum_{j\in J|\widehat{w}_jx+\widehat{b}_j=0} \Big(\max\Big\{0, \widehat{a}_j\big(-\widehat{w}_j'x-\widehat{b}_j'\big)\Big\} + \widehat{a}_j\big(\widehat{w}_j'x+\widehat{b}_j'\big)\Big) \\
&\quad + \sum_{j\notin J|\widehat{w}_jx+\widehat{b}_j=0} \max\Big\{0, \widehat{a}_j\big(\widehat{w}_j'x+\widehat{b}_j'\big)\Big\} \\
&\quad + \sum_{j\in J|\widehat{w}_jx+\widehat{b}_j<0} \Big(\widehat{a}_j'\big(-\widehat{w}_jx-\widehat{b}_j\big) + \widehat{a}_j\big(-\widehat{w}_j'x-\widehat{b}_j'\big) + \widehat{a}_j'\big(\widehat{w}_jx+\widehat{b}_j\big) + \widehat{a}_j\big(\widehat{w}_j'x+\widehat{b}_j'\big)\Big) \\
&\quad + \sum_{j\notin J|\widehat{w}_jx+\widehat{b}_j>0} \Big(\widehat{a}_j'\big(\widehat{w}_jx+\widehat{b}_j\big) + \widehat{a}_j\big(\widehat{w}_j'x+\widehat{b}_j'\big)\Big) \\
&\quad + \sum_{j\in J|\widehat{w}_jx+\widehat{b}_j>0} \Big(\widehat{a}_j'\big(\widehat{w}_jx+\widehat{b}_j\big) + \widehat{a}_j\big(\widehat{w}_j'x+\widehat{b}_j'\big)\Big) \\
&\quad + \widehat{a}_0'x + \widehat{b}_0' \\
&= \sum_{j|\widehat{w}_jx+\widehat{b}_j=0} \max\Big\{0, \widehat{a}_j\big(\widehat{w}_j'x+\widehat{b}_j'\big)\Big\} \\
&\quad + \sum_{j|\widehat{w}_jx+\widehat{b}_j>0} \Big(\widehat{a}_j'\big(\widehat{w}_jx+\widehat{b}_j\big) + \widehat{a}_j\big(\widehat{w}_j'x+\widehat{b}_j'\big)\Big) \\
&\quad + \widehat{a}_0'x + \widehat{b}_0' \\
&= \mathrm{D}^\circ\big(f_{\widehat{\theta},\widehat{a}_0,\widehat{b}_0}(x)\big)\big[\widehat{\theta}',\widehat{a}_0',\widehat{b}_0'\big].
\end{aligned}$$

□

Several of the perturbations used below move the kink of at most one neuron at any given input. For such directions, the Clarke directional derivative of the network output at an input x_0 can be recovered from the ordinary parameter-directional derivatives evaluated at nearby inputs on the two sides of x_0. The next lemma makes this reduction precise: provided zero hidden neurons are not displaced and no two hidden neurons whose kinks coincide at x_0 are simultaneously moved, the Clarke directional derivative at x_0 is the maximum of the two one-sided limits. This will allow the subsequent proofs to work directly with the affine expressions valid on either side of a kink, rather than repeatedly calculating directly with the Clarke subdifferential at the kink itself.

Lemma 4. *Suppose $x_0 \in \mathbb{R}$ and θ', a_0', b_0' is a direction such that:*

- *for all $j \in [m]$, if $w_j = 0 = b_j$ then $w_j' = 0 = b_j'$;*
- *for all $j, j^\dagger \in [m]$ with $j \neq j^\dagger$, if $w_jx_0 + b_j = 0 = w_{j^\dagger}x_0 + b_{j^\dagger}$ then either $w_j' = 0 = b_j'$ or $w_{j^\dagger}' = 0 = b_{j^\dagger}'$.*

Then we have:

$$\begin{aligned}
&\mathrm{D}^\circ\big(f_{\theta,a_0,b_0}(x_0)\big)[\theta',a_0',b_0'] \\
&= \max\Big\{\lim_{x\uparrow x_0}\mathrm{D}\big(f_{\theta,a_0,b_0}(x)\big)[\theta',a_0',b_0'], \lim_{x\downarrow x_0}\mathrm{D}\big(f_{\theta,a_0,b_0}(x)\big)[\theta',a_0',b_0']\Big\} \\
&\mathrm{D}^\circ\big(-f_{\theta,a_0,b_0}(x_0)\big)[\theta',a_0',b_0'] \\
&= \max\Big\{\lim_{x\uparrow x_0}\mathrm{D}\big(-f_{\theta,a_0,b_0}(x)\big)[\theta',a_0',b_0'], \lim_{x\downarrow x_0}\mathrm{D}\big(-f_{\theta,a_0,b_0}(x)\big)[\theta',a_0',b_0']\Big\}.
\end{aligned}$$

Proof. We show the first equation, the second can be proved analogously. By the second assumption of the lemma, there exists at most one $j \in [m]$ such that $w_j x_0 + b_j = 0$ and either $w_j' \neq 0$ or $b_j' \neq 0$. We consider the case where such a j exists, and we denote it $j^\star$; the other case where such a j does not exist is simpler. Without loss of generality, $w_{j^\star} > 0$. Then we have:

$$\begin{aligned}
&\max\Big\{\lim_{x\uparrow x_0}\mathrm{D}\big(f_{\theta,a_0,b_0}(x)\big)[\theta',a_0',b_0'], \lim_{x\downarrow x_0}\mathrm{D}\big(f_{\theta,a_0,b_0}(x)\big)[\theta',a_0',b_0']\Big\} \\
&= \max\left\{\begin{array}{l}\displaystyle\lim_{x\uparrow x_0}\sum_{j|w_jx+b_j>0}\big(a_j'(w_jx+b_j)+a_j(w_j'x+b_j')\big), \\ \displaystyle\lim_{x\downarrow x_0}\sum_{j|w_jx+b_j>0}\big(a_j'(w_jx+b_j)+a_j(w_j'x+b_j')\big)\end{array}\right\} \\
&\quad + a_0'x_0 + b_0' \\
&= \max\left\{\sum_{j|w_j<0,w_jx_0+b_j=0} a_j(w_j'x_0+b_j'), \sum_{j|w_j>0,w_jx_0+b_j=0} a_j(w_j'x_0+b_j')\right\} \\
&\quad + \sum_{j|w_jx_0+b_j>0}\big(a_j'(w_jx_0+b_j)+a_j(w_j'x_0+b_j')\big) + a_0'x_0 + b_0' \\
&= \max\Big\{0, a_{j^\star}\big(w_{j^\star}'x_0+b_{j^\star}'\big)\Big\} \\
&\quad + \sum_{j|w_jx_0+b_j>0}\big(a_j'(w_jx_0+b_j)+a_j(w_j'x_0+b_j')\big) + a_0'x_0 + b_0' \\
&= \sum_{j|w_jx_0+b_j=0}\max\big\{0, a_j(w_j'x_0+b_j')\big\} \\
&\quad + \sum_{j|w_jx_0+b_j>0}\big(a_j'(w_jx_0+b_j)+a_j(w_j'x_0+b_j')\big) + a_0'x_0 + b_0' \\
&= \mathrm{D}^\circ\big(f_{\theta,a_0,b_0}(x_0)\big)[\theta',a_0',b_0'].
\end{aligned}$$

□

We now come to the central part of the work to prove Theorem 1, which consists of a number of lemmas of three kinds: the first kind show that, if a CPA fails some of the properties that make up the geometric characterizations in Theorem 1, then that is due to at least one of typically several possible local geometric irregularities; the second kind show that, for each such irregularity, there exists a direction for infinitesimally perturbing the network's parameters (so that, informally speaking, the irregularity is lessened) such that the signs of the directional derivatives of the network norm and the interpolation constraints mean that the network cannot be a KKT point; and the third kind is similar to the second, except that instead of handling local geometric irregularities, they apply to networks whose parameters have a local internal irregularity (e.g., a failure of balancedness or alignment). We also remark that the statements of some of these lemmas are tailored not only for contributing to the proof of Theorem 1, but additionally to the proof of Theorem 6 (which we develop in appendix G), where all training points contribute to the objective's logistic loss term (in contrast to only those training points at which the interpolation constraints are tight being significant).

First, we identify the local geometric irregularities that capture the failures of convexity correctness for a positive-margin CPA…

Lemma 5. *If a CPA $g\colon \mathbb{R}\to\mathbb{R}$ whose kinks are $(p_k)_{k=1}^q$ has a positive margin, but it is not convexity correct, then at least one of the following holds.*

(i) Some kinks p_k and p_{k+1} are of opposite convexity and such that either $y_i = -1$ for all $x_i \in [p_k, p_{k+1}]$ or $y_i = 1$ for all $x_i \in [p_k, p_{k+1}]$.

(ii) Either $y_i = -1$ for all $x_i \in (-\infty, p_1]$ or $y_i = 1$ for all $x_i \in (-\infty, p_1]$.

(iii) Either $y_i = -1$ for all $x_i \in [p_q, \infty)$ or $y_i = 1$ for all $x_i \in [p_q, \infty)$.

Proof. Since g is not convexity correct, without loss of generality (i.e., the remaining cases are symmetric):

- either some convex kink p_k and some intermediate segment $\iota(l-1)+1, \ldots, \iota(l)$ with label 1 are such that $p_k \in (x_{\iota(l-1)}, x_{\iota(l)+1})$,
- or some kink is in $(-\infty, x_{\iota(1)+1})$.

In the former case, since we have $g(x_{\iota(l-1)}) < 0$ and $g(x_{\iota(l)+1}) < 0$, but $g(x_i) > 0$ for all $i \in \{\iota(l-1)+1, \ldots, \iota(l)\}$, there must be a concave kink in $(x_{\iota(l-1)}, x_{\iota(l)+1})$, and so there exist kinks $p_{k'}$ and $p_{k'+1}$ of opposite convexity in $(x_{\iota(l-1)}, x_{\iota(l)+1})$. Recalling that $y_i = 1$ for all $i \in \{\iota(l-1)+1, \ldots, \iota(l)\}$, property (i) holds.

In the latter case, we have $p_1 \in (-\infty, x_{\iota(1)+1})$, and recalling that labels y_i are the same for all $i \in \{1, \ldots, \iota(1)\}$, property (ii) holds. □

... and for each of those irregularities, identify a direction along which the bias-free square ℓ_2-norm of the network parameters has a negative derivative while, for each training point, its negative margin has a nonpositive Clarke derivative.

Lemma 6. *If $g = f_{\theta,a_0,b_0}$, whose kinks are $(p_k)_{k=1}^q$, satisfies at least one of properties (i), (ii), or (iii) in Lemma 5, then there exists a direction θ', a_0', b_0' such that $\mathrm{D}\big(\frac{1}{2}\sum_{j=1}^m (a_j^2 + w_j^2)\big)[\theta', a_0', b_0'] < 0$ and $\mathrm{D}^\circ\big(-y_i\, f_{\theta,a_0,b_0}(x_i)\big)[\theta', a_0', b_0'] \le 0$ for all $i \in [n]$.*

Proof. In case property (i) in Lemma 5 holds, without loss of generality, p_k is convex, p_{k+1} is concave, and $y_i = -1$ for all $x_i \in [p_k, p_{k+1}]$. Then θ must contain some neurons (a_j, w_j, b_j) and $(a_{j'}, w_{j'}, b_{j'})$ such that:

- $-b_j/w_j = p_k$ and $a_j > 0$;
- $-b_{j'}/w_{j'} = p_{k+1}$ and $a_{j'} < 0$.

By Lemma 3, after reversing the relevant neurons and absorbing the resulting affine terms into the skip connection, we may assume without loss of generality that $w_j > 0$ and $w_{j'} > 0$. Any direction constructed for the resulting parameterization pulls back, by Lemma 3 (ii), to a direction with the same relevant Clarke and regularizer derivatives at the original parameterization. Moreover, by Lemma 4, the claimed nonpositivity of the Clarke directional derivatives of $-y_i\, f_{\theta,a_0,b_0}(x_i)$ is implied by nonpositivity of the corresponding one-sided limits of the ordinary directional derivatives. Therefore, letting $u \coloneqq (a_{j'} w_{j'}, -a_{j'} w_j, w_j b_{j'} - w_{j'} b_j)$, it suffices to show that:

- $\mathrm{D}_{a_j, a_{j'}, b_{j'}}\big(\frac{1}{2}\sum_{j''=1}^m (a_{j''}^2 + w_{j''}^2)\big)[u] < 0$;
- $\lim_{x \uparrow x_i} \mathrm{D}_{a_j, a_{j'}, b_{j'}}\big(y_i\, f_{\theta,a_0,b_0}(x)\big)[u] \ge 0$ for all $i \in [n]$;
- $\lim_{x \downarrow x_i} \mathrm{D}_{a_j, a_{j'}, b_{j'}}\big(y_i\, f_{\theta,a_0,b_0}(x)\big)[u] \ge 0$ for all $i \in [n]$.

For the first claim, we have

$$\mathrm{D}_{a_j, a_{j'}, b_{j'}}\bigg(\frac{1}{2}\sum_{j''=1}^m (a_{j''}^2 + w_{j''}^2)\bigg)[u] = a_j a_{j'} w_{j'} - a_{j'}^2 w_j < 0$$

because $a_j > 0$, $a_{j'} < 0$, $w_j > 0$, and $w_{j'} > 0$. For the second and third claims, since $y_i = -1$ for all $x_i \in [p_k, p_{k+1}]$, in turn it suffices to observe that:

- for all $x \in (-\infty, p_k)$ we have

$$\begin{aligned}&\mathrm{D}_{a_j,a_{j'},b_{j'}} f_{\theta,a_0,b_0}(x)[u]\\ &= \mathrm{D}_{a_j,a_{j'},b_{j'}}\big(a_j\,\sigma(w_j x + b_j)\big)[u] + \mathrm{D}_{a_j,a_{j'},b_{j'}}\big(a_{j'}\,\sigma(w_{j'} x + b_{j'})\big)[u]\\ &= 0;\end{aligned}$$

- for all $x \in (p_k, p_{k+1})$ we have

$$\begin{aligned}&\mathrm{D}_{a_j,a_{j'},b_{j'}} f_{\theta,a_0,b_0}(x)[u]\\ &= \mathrm{D}_{a_j,a_{j'},b_{j'}}\big(a_j\,\sigma(w_j x + b_j)\big)[u] + \mathrm{D}_{a_j,a_{j'},b_{j'}}\big(a_{j'}\,\sigma(w_{j'} x + b_{j'})\big)[u]\\ &= (w_j x + b_j) a_{j'} w_{j'}\\ &< 0;\end{aligned}$$

- for all $x \in (p_{k+1}, \infty)$ we have

$$\begin{aligned}&\mathrm{D}_{a_j,a_{j'},b_{j'}} f_{\theta,a_0,b_0}(x)[u]\\ &= \mathrm{D}_{a_j,a_{j'},b_{j'}}\big(a_j\,\sigma(w_j x + b_j)\big)[u] + \mathrm{D}_{a_j,a_{j'},b_{j'}}\big(a_{j'}\,\sigma(w_{j'} x + b_{j'})\big)[u]\\ &= (w_j x + b_j) a_{j'} w_{j'} - (w_{j'} x + b_{j'}) a_{j'} w_j + a_{j'}(w_j b_{j'} - w_{j'} b_j)\\ &= 0.\end{aligned}$$

The remaining two cases, when property (ii) or property (iii) in Lemma 5 holds, are symmetric so we show how to handle the former, where without loss of generality, $y_i = -1$ for all $x_i \in (-\infty, p_1]$. Then we have two subcases depending on the convexity of p_1, where we consider the less straightforward one, namely when p_1 is concave. Hence, and by the orientation-reversal reparameterization in Lemma 3, for some neuron (a_j, w_j, b_j) in θ we have

$$-b_j/w_j = p_1 \qquad\qquad a_j < 0 \qquad\qquad w_j < 0.$$

Now pick $\widehat{p} \leq p_1$ such that $\widehat{p} < x_1$, and let $v \coloneqq (-a_j, w_j\widehat{p} + b_j)$. By Lemma 4 again for reducing our proof obligations with Clarke directional derivatives to sign properties of one-sided limits of ordinary directional derivatives, it suffices to show that:

- $\mathrm{D}_{a_j,b_j}\big(\frac{1}{2}\sum_{j'=1}^{m}(a_{j'}^2 + w_{j'}^2)\big)[v] < 0$;
- $\lim_{x\uparrow x_i} \mathrm{D}_{a_j,b_j}\big(y_i\, f_{\theta,a_0,b_0}(x)\big)[v] \geq 0$ for all $i \in [n]$;
- $\lim_{x\downarrow x_i} \mathrm{D}_{a_j,b_j}\big(y_i\, f_{\theta,a_0,b_0}(x)\big)[v] \geq 0$ for all $i \in [n]$.

For the first claim, we have

$$\mathrm{D}_{a_j,b_j}\left(\frac{1}{2}\sum_{j'=1}^{m}(a_{j'}^2 + w_{j'}^2)\right)[v] = -a_j^2 < 0.$$

For the second and third claims, since $\widehat{p} < x_1$ and $y_i = -1$ for all $x_i \in (-\infty, p_1]$, in turn it suffices to observe that:

- for all $x \in (\widehat{p}, p_1)$ we have

$$\begin{aligned}\mathrm{D}_{a_j,b_j} f_{\theta,a_0,b_0}(x)[v] &= \mathrm{D}_{a_j,b_j}\big(a_j\,\sigma(w_j x + b_j)\big)[v]\\ &= -a_j(w_j x + b_j) + a_j(w_j\widehat{p} + b_j)\\ &= -a_j w_j (x - \widehat{p})\\ &< 0;\end{aligned}$$

- for all $x \in (p_1, \infty)$ we have

$$\begin{aligned}\mathrm{D}_{a_j,b_j} f_{\theta,a_0,b_0}(x)[v] &= \mathrm{D}_{a_j,b_j}\big(a_j\,\sigma(w_j x + b_j)\big)[v]\\ &= 0.\end{aligned}$$

□

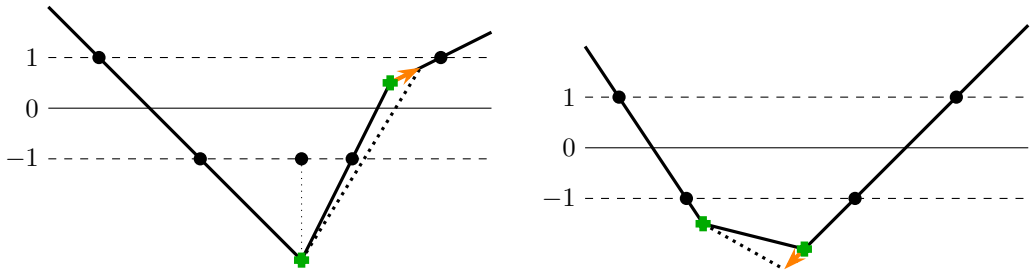


Figure 6: **Left:** Here we illustrate the proof of Lemma 6 for the case of property (i) in Lemma 5; in this example, one of the kinks of the network function happens to coincide with an input in the dataset. **Right:** Here we illustrate the proof of Lemma 11 for the case of property (i) in Lemma 9. **All:** The kinks p_k and p_{k+1} in the proof are indicated by the green pluses. The proof identifies a direction for infinitesimally perturbing the network parameters so that the portion of the network function between p_k and p_{k+1} moves as indicated by the orange arrow and the dotted line segment.

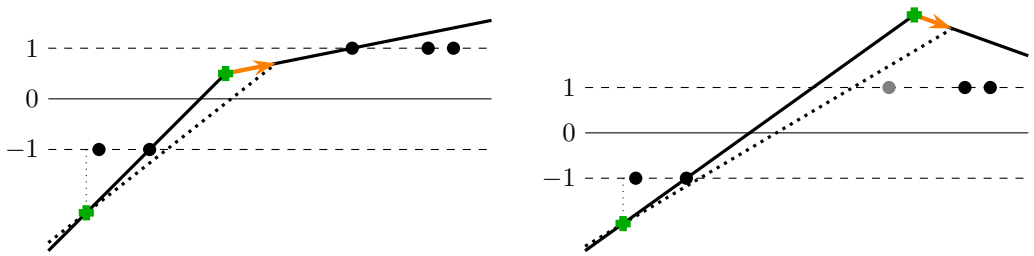


Figure 7: **Left:** Here we illustrate the proofs of Lemmas 6 and 11 for the case of property (ii) in Lemma 5. **Right:** Here we illustrate the proofs of Lemmas 8 and 12 for the case of property (ii) in Lemma 7; this example also shows, by the dataset point colored gray, how the directional derivative of $-y_i\, f_{\theta,a_0,b_0}(x_i)$ may be positive when the interpolation at x_i is not tight (i.e., $f_{\theta,a_0,b_0}(x_i) \neq y_i$). **All:** The first kink p_1 and the center of rotation $\widehat{p}$ in the proofs are indicated by the green pluses. The proof identifies a direction for infinitesimally perturbing the network parameters so that the portion of the network function before p_1 moves as indicated by the orange arrow and the dotted half line.

Next, we identify the local geometric irregularities that capture the failures of switch hugging for a convexity-correct interpolating CPA when there are at least two label switches...

Lemma 7. *Suppose the dataset has at least two label switches, i.e., $r \geq 2$. If a CPA $g\colon \mathbb{R} \to \mathbb{R}$ whose kinks are $(p_k)_{k=1}^q$ interpolates the dataset and is convexity correct, but it is not switch hugging, then at least one of the following holds.*

- *(i) Some kinks p_k and p_{k+1} are of opposite convexity and such that either $g(x_i) \neq 1$ for all $x_i \in [p_k, p_{k+1}]$ or $g(x_i) \neq -1$ for all $x_i \in [p_k, p_{k+1}]$.*
- *(ii) We have $p_1 \in (-\infty, x_{\iota(2)}]$, and either $g(x_i) \neq 1$ for all $x_i \in (-\infty, p_1]$ or $g(x_i) \neq -1$ for all $x_i \in (-\infty, p_1]$.*
- *(iii) We have $p_q \in [x_{\iota(r-1)+1}, \infty)$, and either $g(x_i) \neq 1$ for all $x_i \in [p_q, \infty)$ or $g(x_i) \neq -1$ for all $x_i \in [p_q, \infty)$.*

Proof. Without loss of generality, there exists $l \in [r]$ such that $y_{\iota(l)} = 1$ and $g(x_{\iota(l)}) > 1$.

If $l = 1$, then since $r \geq 2$, and g has a positive margin and is convexity correct, we have that $p_1 \in [x_{\iota(1)+1}, x_{\iota(2)}]$, and so $g(x_i) \neq 1$ for all $x_i \in (-\infty, p_1]$, which means that property (ii) holds.

If $l = r$, then since $r \geq 2$, and g has a positive margin and is convexity correct, we have that $p_q \in [x_{\iota(r-1)+1}, x_{\iota(r)}]$, and so $g(x_i) \neq 1$ for all $x_i \in [p_q, \infty)$, which means that property (iii) holds.

Otherwise $1 < l < r$, so g has a last kink $p_k \leq x_{\iota(l)}$ which is necessarily concave, and the next kink p_{k+1} exists and is necessarily convex. Moreover $g(x_i) \neq 1$ for all $x_i \in [p_k, p_{k+1}]$, so property (i) holds. □

... and for each of those irregularities, identify a direction along which the bias-free square ℓ_2-norm of the network parameters has a negative derivative while, for each training point at which interpolation is tight, its negative margin has a nonpositive Clarke derivative. This lemma and the previous one, in contrast to the preceding two lemmas, are used only in the proof of Theorem 1 (i.e., not also in the proof of Theorem 6).

Lemma 8. *If $g = f_{\theta,a_0,b_0}$, whose kinks are $(p_k)_{k=1}^q$, interpolates the dataset and satisfies at least one of properties (i), (ii), or (iii) in Lemma 7, then there exists a direction θ', a_0', b_0' such that $\mathrm{D}\big(\frac{1}{2}\sum_{j=1}^m (a_j^2 + w_j^2)\big)[\theta', a_0', b_0'] < 0$ and $\mathrm{D}^\circ\big(-y_i\, f_{\theta,a_0,b_0}(x_i)\big)[\theta', a_0', b_0'] \leq 0$ for all $i \in [n]$ with $g(x_i) = y_i$.*

Proof. Properties (i), (ii), and (iii) in Lemma 7 are identical to, respectively, properties (i), (ii), or (iii) in Lemma 5, except that each condition that all training inputs from an interval have a fixed label is weakened to apply only to points at which the corresponding interpolation constraint is tight. The proof of Lemma 8 can therefore be obtained from the proof of Lemma 6, by restricting the reasoning to the latter training points. □

Next, we consider convexity-correct CPAs that fail to be single turning.

Lemma 9. *If a CPA $g\colon \mathbb{R} \to \mathbb{R}$ whose kinks are $(p_k)_{k=1}^q$ is convexity correct, but it is not single turning, then at least one of the following holds.*

(i) Some kinks p_k and p_{k+1} are convex, and $y_i = -1$ for all $x_i \in [p_k, p_{k+1}]$.

(ii) Some kinks p_k and p_{k+1} are concave, and $y_i = 1$ for all $x_i \in [p_k, p_{k+1}]$.

Proof. This is immediate from the definitions of convexity correctness and single turning. □

Before proceeding to handle the two local geometric irregularities identified in Lemma 9, we show that, if the network has any neuron which is imbalanced in the sense that the ℓ_2-norm of its output weight differs from that of its hidden weight and bias, then there exists a direction along which the square ℓ_2-norm of the network parameters has a negative derivative while, for each training point, its negative margin has a nonpositive Clarke derivative.

Lemma 10. *If θ contains a neuron (a_j, w_j, b_j) such that $a_j^2 \neq w_j^2 + b_j^2$, then there exists a direction θ', a_0', b_0' such that $\mathrm{D}\big(\frac{1}{2}\sum_{j=1}^m (a_j^2 + w_j^2 + b_j^2)\big)[\theta', a_0', b_0'] < 0$ and $\mathrm{D}^\circ\big(-y_i\, f_{\theta,a_0,b_0}(x_i)\big)[\theta', a_0', b_0'] \leq 0$ for all $i \in [n]$.*

Proof. We consider the case $a_j^2 > w_j^2 + b_j^2$, and either $w_j \neq 0$ or $b_j \neq 0$; the other cases are analogous or simpler. Letting $v \coloneqq (-a_j, w_j, b_j)$, we are done by observing that

$$\mathrm{D}_{a_j,w_j,b_j}\left(\frac{1}{2}\sum_{j'=1}^m (a_{j'}^2 + w_{j'}^2 + b_{j'}^2)\right)[v] = -a_j^2 + w_j^2 + b_j^2 < 0,$$

that $w_j x + b_j > 0$ implies

$$\begin{aligned}\mathrm{D}_{a_j,w_j,b_j} f_{\theta,a_0,b_0}(x)[v] &= \mathrm{D}_{a_j,w_j,b_j}\big(a_j\,\sigma(w_j x + b_j)\big)[v] \\ &= -a_j(w_j x + b_j) + a_j w_j x + a_j b_j \\ &= 0,\end{aligned}$$

and that $w_j x + b_j < 0$ implies

$$\begin{aligned}\mathrm{D}_{a_j,w_j,b_j} f_{\theta,a_0,b_0}(x)[v] &= \mathrm{D}_{a_j,w_j,b_j}\big(a_j\,\sigma(w_j x + b_j)\big)[v] \\ &= -a_j 0 + w_j 0 + b_j 0 \\ &= 0.\end{aligned}$$

□

Now we handle the two irregularities identified in Lemma 9 which arise from failures of single turning, and also the three irregularities identified in Lemma 5 which capture failures of convexity correctness. In doing so, thanks to the previous lemma, we can assume that all neurons in the network are balanced.

Lemma 11. *Under Assumption 1, if $g = f_{\theta,a_0,b_0}$, whose kinks are $(p_k)_{k=1}^q$, satisfies at least one of properties (i), (ii), or (iii) in Lemma 5, or (i) or (ii) in Lemma 9, then there exists a direction θ', a_0', b_0' such that $\mathrm{D}\big(\frac{1}{2}\sum_{j=1}^m (a_j^2 + w_j^2 + b_j^2)\big)[\theta', a_0', b_0'] < 0$ and $\mathrm{D}^\circ\big(-y_i\, f_{\theta,a_0,b_0}(x_i)\big)[\theta', a_0', b_0'] \le 0$ for all $i \in [n]$.*

Proof. First we observe that the disjunction of properties (i) in Lemma 5, and (i) and (ii) in Lemma 9, is equivalent to the following.

(I) Some kinks p_k and p_{k+1} are such that either at least one of p_k, p_{k+1} is convex and $y_i = -1$ for all $x_i \in [p_k, p_{k+1}]$, or at least one of p_k, p_{k+1} is concave and $y_i = 1$ for all $x_i \in [p_k, p_{k+1}]$.

Now the reasoning follows the same pattern as the proof of Lemma 6, so we focus on the modifications and additions, which are to handle the replacement of property (i) in Lemma 5 by the weaker property (I) above (i.e., now only one convexity combination is excluded depending on the labels) and the enlargement of the cost by including the biases.

In case property (I) above holds, without loss of generality, p_k is convex and $y_i = -1$ for all $x_i \in [p_k, p_{k+1}]$. Then θ must contain some neurons (a_j, w_j, b_j) and $(a_{j'}, w_{j'}, b_{j'})$ such that:

- $-b_j/w_j = p_k$ and $a_j > 0$;
- $-b_{j'}/w_{j'} = p_{k+1}$ and $a_{j'} \neq 0$.

By Lemma 10 we may restrict attention to balanced neurons (in the sense that includes the biases), and by Lemma 3 we may orient the relevant neurons so that their hidden weights are positive, without changing the represented function or any of the directional-derivative inequalities to be established. Thus for the following construction we may assume that:

- $w_j > 0$ and $a_j^2 = w_j^2 + b_j^2$;
- $w_{j'} > 0$ and $a_{j'}^2 = w_{j'}^2 + b_{j'}^2$.

Let $u \coloneqq \big(-w_{j'}, w_j, (w_{j'} b_j - w_j b_{j'})/a_{j'}\big)$. Here vector u has the same direction as in the proof of Lemma 6. Its different scaling helps to handle the unknown sign of $a_{j'}$ in the following derivation of the negativity of the directional derivative of the cost along u:

$$
\begin{aligned}
&\mathrm{D}_{a_j,a_{j'},b_{j'}}\Big(\frac{1}{2}\sum_{j''=1}^m (a_{j''}^2 + w_{j''}^2 + b_{j''}^2)\Big)[u] \\
&= -a_j w_{j'} + a_{j'} w_j + \frac{w_{j'} b_j b_{j'} - w_j b_{j'}^2}{a_{j'}} \\
&= -a_j w_{j'} + \frac{b_j b_{j'} w_{j'} + w_j w_{j'}^2}{a_{j'}} \\
&= -\frac{w_{j'}}{a_{j'}}(a_j a_{j'} - w_j w_{j'} - b_j b_{j'}) \\
&= -\frac{w_{j'}}{|a_{j'}|}\big(a_j |a_{j'}| - \mathrm{sgn}(a_{j'})(w_j w_{j'} + b_j b_{j'})\big) \\
&= -\frac{w_{j'}}{|a_{j'}|}\Big(\sqrt{w_j^2 + b_j^2}\sqrt{w_{j'}^2 + b_{j'}^2} - \mathrm{sgn}(a_{j'})(w_j w_{j'} + b_j b_{j'})\Big) \\
&< 0,
\end{aligned}
$$

where the Cauchy-Bunyakovsky-Schwarz inequality is strict thanks to $p_k \neq p_{k+1}$.

The remaining two cases, when property (ii) or property (iii) in Lemma 5 holds, are symmetric so we consider the former, where without loss of generality, $y_i = -1$ for all $x_i \in (-\infty, p_1]$. Then we have two subcases depending on the convexity of p_1, where again we consider the less straightforward one, namely when p_1 is concave. Hence, and by Lemmas 3 and 10, for some neuron (a_j, w_j, b_j) in θ we have

$$-b_j/w_j = p_1 \qquad a_j < 0 \qquad w_j < 0 \qquad a_j^2 = w_j^2 + b_j^2.$$

As before, pick $\widehat{p} \leq p_1$ with $\widehat{p} < x_1$, and let $v \coloneqq (-a_j, w_j\widehat{p} + b_j)$. It remains to observe that

$$\mathrm{D}_{a_j,b_j}\Big(\frac{1}{2}\sum_{j'=1}^{m}(a_{j'}^2 + w_{j'}^2 + b_{j'}^2)\Big)[v] = -a_j^2 + w_j b_j \widehat{p} + b_j^2 = -w_j^2(p_1\widehat{p} + 1) < 0,$$

where the inequality follows from $\widehat{p} \leq p_1$, since $p_1 < 0$ by Assumption 1. □

Like Lemma 8, the next lemma handles the three irregularities identified in Lemma 7 which arise from failures of switch hugging, the difference being that here the biases are included in the square ℓ_2-norm of the network parameters. Also like Lemma 8, this result is used only in the proof of Theorem 1.

Lemma 12. *Under Assumption 1, if $g = f_{\theta,a_0,b_0}$, whose kinks are $(p_k)_{k=1}^{q}$, interpolates the dataset and satisfies at least one of properties (i), (ii), or (iii) in Lemma 7, then there exists a direction θ', a_0', b_0' such that $\mathrm{D}\big(\frac{1}{2}\sum_{j=1}^{m}(a_j^2 + w_j^2 + b_j^2)\big)[\theta', a_0', b_0'] < 0$ and $\mathrm{D}^\circ\big(-y_i\, f_{\theta,a_0,b_0}(x_i)\big)[\theta', a_0', b_0'] \leq 0$ for all $i \in [n]$ with $g(x_i) = y_i$.*

Proof. As with the proofs of Lemmas 6 and 8, the proof of Lemma 12 proceeds analogously to the proof of Lemma 11, with a further simplification due to property (I) in the latter proof not requiring the two kinks to have opposite convexity whereas property (i) in Lemma 7 requires it. □

The following three lemmas handle further local internal irregularities in the network: first, where some neuron is imbalanced in the sense that the absolute values of its output and hidden weights differ (this mode of balancedness, which is in contrast to that in Lemma 10, is appropriate when the biases are not penalized)…

Lemma 13. *If θ contains a neuron (a_j, w_j, b_j) such that $|a_j| \neq |w_j|$, then there exists a direction θ', a_0', b_0' such that $\mathrm{D}\big(\frac{1}{2}\sum_{j=1}^{m}(a_j^2 + w_j^2)\big)[\theta', a_0', b_0'] < 0$ and $\mathrm{D}^\circ\big(-y_i\, f_{\theta,a_0,b_0}(x_i)\big)[\theta', a_0', b_0'] \leq 0$ for all $i \in [n]$.*

Proof. This is similar to and simpler than the proof of Lemma 10. If $|a_j| > |w_j|$ (resp., $|a_j| < |w_j|$), it suffices to consider directional derivatives at (a_j, w_j, b_j) along $(-a_j, w_j, b_j)$ (resp., $(a_j, -w_j, -b_j)$). □

…second, where some two neurons are aligned in the sense that their kinks are the same, but the signs of their output weights are conflicting…

Lemma 14. *If θ contains neurons (a_j, w_j, b_j) and $(a_{j'}, w_{j'}, b_{j'})$ such that all of a_j, w_j, $a_{j'}$, and $w_{j'}$ are nonzero, $-b_j/w_j = -b_{j'}/w_{j'}$, and $\operatorname{sgn} a_j \neq \operatorname{sgn} a_{j'}$, then there exists a direction θ', a_0', b_0' such that $\mathrm{D}\big(\frac{1}{2}\sum_{j=1}^{m}(a_j^2 + w_j^2)\big)[\theta', a_0', b_0'] < 0$, $\mathrm{D}\big(\frac{1}{2}\sum_{j=1}^{m}(a_j^2 + w_j^2 + b_j^2)\big)[\theta', a_0', b_0'] < 0$, and $\mathrm{D}^\circ\big(-y_i\, f_{\theta,a_0,b_0}(x_i)\big)[\theta', a_0', b_0'] \leq 0$ for all $i \in [n]$.*

Proof. Without loss of generality, we have $a_j > 0$ and $a_{j'} < 0$. Moreover, by the orientation-reversal reparameterization in Lemma 3, we may assume $w_j > 0$ and $w_{j'} > 0$. Letting $v \coloneqq (-w_{j'}, w_j)$ and verifying that Lemma 4 applies because we are not perturbing any hidden-layer parameters, we are done by observing that

$$\mathrm{D}_{a_j,a_{j'}}\Big(\frac{1}{2}\sum_{j''=1}^{m}(a_{j''}^2 + w_{j''}^2 + b_{j''}^2)\Big)[v]$$

$$= \mathrm{D}_{a_j,a_{j'}}\Big(\frac{1}{2}\sum_{j''=1}^{m}(a_{j''}^2 + w_{j''}^2)\Big)[v]$$

$$
\begin{aligned}
&= -a_j w_{j'} + a_{j'} w_j \\
&< 0,
\end{aligned}
$$

that for all $x < -b_j/w_j$ we have

$$
\begin{aligned}
\mathrm{D}_{a_j,a_{j'}} f_{\theta,a_0,b_0}(x)[v] &= \mathrm{D}_{a_j,a_{j'}} \big(a_j\, \sigma(w_j x + b_j) + a_{j'}\, \sigma(w_{j'} x + b_{j'})\big)[v] \\
&= -w_{j'}\, \sigma(w_j x + b_j) + w_j\, \sigma(w_{j'} x + b_{j'}) \\
&= -w_{j'} 0 + w_j 0 \\
&= 0,
\end{aligned}
$$

and that for all $x > -b_j/w_j$ we have

$$
\begin{aligned}
\mathrm{D}_{a_j,a_{j'}} f_{\theta,a_0,b_0}(x)[v] &= \mathrm{D}_{a_j,a_{j'}} \big(a_j\, \sigma(w_j x + b_j) + a_{j'}\, \sigma(w_{j'} x + b_{j'})\big)[v] \\
&= -w_{j'}\, \sigma(w_j x + b_j) + w_j\, \sigma(w_{j'} x + b_{j'}) \\
&= -w_{j'}(w_j x + b_j) + w_j(w_{j'} x + b_{j'}) \\
&= -w_{j'} b_j + w_j b_{j'} \\
&= 0.
\end{aligned}
$$

□

. . . and third, where the hidden weight of some neuron is zero (this is relevant when the biases are penalized, in which case by Lemma 10, we can assume that the neuron's output weight and hidden bias have equal absolute values).

Lemma 15. *If θ contains (a_j, w_j, b_j) such that $|a_j| = |b_j| > 0$ and $w_j = 0$, then there exists θ', a_0', b_0' such that $\mathrm{D}\big(\frac{1}{2}\sum_{j=1}^m (a_j^2 + w_j^2 + b_j^2)\big)[\theta', a_0', b_0'] < 0$ and $\mathrm{D}^\circ\big(-y_i\, f_{\theta,a_0,b_0}(x_i)\big)[\theta', a_0', b_0'] \leq 0$ for all $i \in [n]$.*

Proof. Again this is similar to the proof of Lemma 10. It suffices to consider directional derivatives at (a_j, b_j, b_0) along $(-\operatorname{sgn} a_j, -\operatorname{sgn} b_j, 2\operatorname{sgn}(a_j)\sigma(b_j))$. □

The final lemma in the sequence shows that MFCQ is satisfied everywhere in the feasible sets of the four interpolator norm minimization problems we consider. The proof is based on the familiar phenomenon that increasing the magnitudes of an interpolating network's parameters increases the margin, however some care is required here due to the failure of homogeneity when the skip connection is present.

Lemma 16. *For each of probs. (P), (P_1), (P^b) and (P_1^b), every feasible point satisfies MFCQ.*

Proof. We show the claim for prob. (P_1^b), the cases of probs. (P), (P_1) and (P^b) can be handled analogously.

Suppose θ, a_0, b_0 is feasible. For all $i \in [n]$ (we do not need to consider only the margin) and all $\gamma \in \partial_\theta f_{\theta,a_0,b_0}(x_i)$, we have

$$
\begin{aligned}
&\Big\langle \Big(\frac{1}{2}\theta, a_0, b_0\Big), y_i\big(\gamma, \nabla_{a_0} f_{\theta,a_0,b_0}(x_i), \nabla_{b_0} f_{\theta,a_0,b_0}(x_i)\big) \Big\rangle \\
&= \Big\langle \Big(\frac{1}{2}\theta, a_0, b_0\Big), y_i(\gamma, x_i, 1) \Big\rangle \\
&= y_i\Big(\frac{1}{2}\langle \theta, \gamma\rangle + a_0 x_i + b_0\Big) \\
&= y_i\big(f_\theta(x_i) + a_0 x_i + b_0\big) \\
&= y_i\, f_{\theta,a_0,b_0}(x_i) \\
&\geq 1
\end{aligned}
$$

as required, where the second last equality is by Euler's homogeneous function theorem. □

Lemma 17. *For each $k \in [r]$, let $h_k \colon \mathbb{R} \to \mathbb{R}$ be the affine function whose graph passes through $\big(x_{\iota(k)}, y_{\iota(k)}\big)$ and $\big(x_{\iota(k)+1}, y_{\iota(k)+1}\big)$. There exists a unique CPA $g \colon \mathbb{R} \to \mathbb{R}$ that is switch hugging, convexity correct, and single turning. This g has exactly $r-1$ kinks: for every $k \in [r-1]$, its unique kink in the k-th intermediate same-label segment is the unique intersection of h_k and h_{k+1}.*

Proof. For every $k \in [r-1]$, since labels $y_{\iota(k)}$ and $y_{\iota(k)+1}$ incident to the k-th label switch are opposite, labels $y_{\iota(k)+1}$ and $y_{\iota(k+1)}$ of the k-th intermediate same-label segment are the same, and labels $y_{\iota(k+1)}$ and $y_{\iota(k+1)+1}$ incident to the $(k+1)$-th label switch are opposite, we have that the affine functions h_k and h_{k+1} have a unique intersection; we denote its x-coordinate as p_k.

Letting

$$g(x) \coloneqq \begin{cases} h_1(x) & \text{if } x \le p_1, \\ h_k(x) & \text{if } p_{k-1} \le x \le p_k, \\ h_r(x) & \text{if } p_{r-1} \le x, \end{cases}$$

it is straightforward to verify that g is switch hugging, convexity correct, and single turning.

Conversely, switch hugging and convexity correctness fix the crossing pieces of any such CPA to be $h_1, \dots, h_r$, and continuity forces its unique kink in the k-th intermediate same-label segment to be the intersection p_k. Therefore g is unique. □

Proof. [Proof of Theorem 1] For part (i), suppose $g = f_{\theta,a_0,b_0}$ is a minimizer of prob. (P_1) in function space. By Lemma 16 and Theorem 2, θ, a_0, b_0 is a KKT point, so by Proposition 1 and Lemmas 5, 6, 7 and 8, g is switch hugging and convexity correct.

Conversely, suppose g is switch hugging and convexity correct, but not a minimizer of prob. (P_1) in function space. By Lemma 2, for some $m \in \mathbb{N}$ and some θ', a_0', b_0' which is a minimizer of prob. (P_1) restricted to network width m, we have that $\mathrm{R}_1(f_{\theta',a_0',b_0'}) < \mathrm{R}_1(g)$. Reasoning like above, $f_{\theta',a_0',b_0'}$ is switch hugging and convexity correct. But then by Theorem 4 we have

$$\mathrm{R}_1(f_{\theta',a_0',b_0'}) = \sum_{k=1}^{r-1} \Big(\frac{2}{x_{\iota(k)+1} - x_{\iota(k)}} + \frac{2}{x_{\iota(k+1)+1} - x_{\iota(k+1)}} \Big) = \mathrm{R}_1(g),$$

which is a contradiction.

The remainder of part (i), which is for prob. (P), now follows by Lemma 1.

For part (ii), the reasoning follows the same pattern, with Lemma 9 added for the single turning property and with Lemmas 6 and 8 replaced by Lemmas 11 and 12. Lemma 17 gives a unique switch-hugging, convexity-correct, and single-turning interpolant, with exactly one kink in each intermediate same-label segment. The preceding necessity argument shows that every minimizer has these properties, while the same representor-norm comparison as above shows that this canonical interpolant is a minimizer. It is therefore the unique minimizer in function space and has exactly $r-1$ kinks.

For part (iii), suppose $(\theta, a_0, b_0) \in \mathbb{R}^{3m+2}$ is a KKT point of prob. (P_1). Then by Proposition 1 and Lemmas 5, 6, 7 and 8 we have that f_{θ,a_0,b_0} is switch hugging and convexity correct, and moreover by Theorem 4 and Lemmas 13 and 14 we have that $\frac{1}{2}\sum_{j=1}^{m}(a_j^2 + w_j^2) = \mathrm{R}_1(f_{\theta,a_0,b_0})$, so by part (i) we conclude that θ, a_0, b_0 is a minimizer of prob. (P_1).

Part (iii) for prob. $(\mathrm{P}_1^{\mathrm{b}})$ follows analogously, with Lemma 9 added for the single turning property, with Lemmas 6 and 8 replaced by Lemmas 11 and 12, and with Lemma 13 replaced by Lemmas 10 and 15. □

F COUNTEREXAMPLES WITHOUT THE SKIP CONNECTION

In this section, we show that the skip connection is essential for the benign properties of the optimization landscapes in Theorem 1 (iii) to hold. Namely, in the next theorem, we demonstrate that, if the skip connection is absent, then there exist KKT points that are not (global) minimizers, regardless of whether the biases are penalized.

In addition, for the unpenalized-bias example, Figure 8 displays a two-dimensional landscape profile with multiple valleys and several suboptimal KKT plateaus, together with the corresponding profile after adding a skip connection. When biases are penalized, the explicit KKT examples are isolated rather than forming the same plateaus; Remark 1 records the additional observation that the identified points are strict local minimizers.

Theorem 5. *(i) For the dataset*

$$(-2,1),(-1,-1),(1,1),(2,-1),$$

there exists a two-layer ReLU network θ of width 4 and without a skip connection, which is a KKT point but not a minimizer of prob. (P), and for which the interpolant f_θ is not convexity correct.

(ii) For the dataset

$$(-10,-1),(-6,1),(-5,-1),(-4,1),(4,1),(5,-1),(6,1),(10,-1),$$

there exists a two-layer ReLU network θ of width 6 and without a skip connection, which is a KKT point but not a minimizer of prob. (P^{b}), and for which the interpolant f_θ is not single turning and not sparsest.

Proof. For part (i), let us write $(x_i, y_i)_{i=1}^4$ for the dataset and $\theta = (a_j, w_j, b_j)_{j=1}^4$ for the network. Given any $p_1 < x_1$ and $p_4 > x_4$, we will define θ so that

- f_θ is switch hugging, i.e., $f_\theta(x_i) = y_i$ for all $i \in [4]$,
- the kinks of the four neurons are at

  $$p_1 \qquad p_2 \coloneqq x_2 \qquad p_3 \coloneqq x_3 \qquad p_4$$

  respectively, i.e., $-b_j/w_j = p_j$ for all $j \in [4]$, and
- the first (resp., last) two neurons are active in the positive (resp., negative) direction, i.e., $w_1, w_2 > 0$ and $w_3, w_4 < 0$.

With the latter two properties, the switch hugging constraints are a system of four linear equations in the slopes $t_j \coloneqq a_j w_j$ of the four neurons, which has the following unique solution:

$$t_1 = -\frac{5p_4}{p_4 - p_1} \qquad t_2 = 3 \qquad t_3 = 3 \qquad t_4 = \frac{5p_1}{p_4 - p_1}.$$

To obtain a KKT point, we need to have $a_j^2 = w_j^2$ for all four neurons, so

$$a_1 = -\sqrt{\frac{5p_4}{p_4 - p_1}} \qquad a_2 = \sqrt{3} \qquad a_3 = -\sqrt{3} \qquad a_4 = \sqrt{\frac{-5p_1}{p_4 - p_1}}$$
$$w_1 = \sqrt{\frac{5p_4}{p_4 - p_1}} \qquad w_2 = \sqrt{3} \qquad w_3 = -\sqrt{3} \qquad w_4 = -\sqrt{\frac{-5p_1}{p_4 - p_1}},$$

and the biases are determined by the positions $-b_j/w_j = p_j$ of the kinks, so

$$b_1 = -p_1\sqrt{\frac{5p_4}{p_4 - p_1}} \qquad b_2 = \sqrt{3} \qquad b_3 = \sqrt{3} \qquad b_4 = p_4\sqrt{\frac{-5p_1}{p_4 - p_1}}.$$

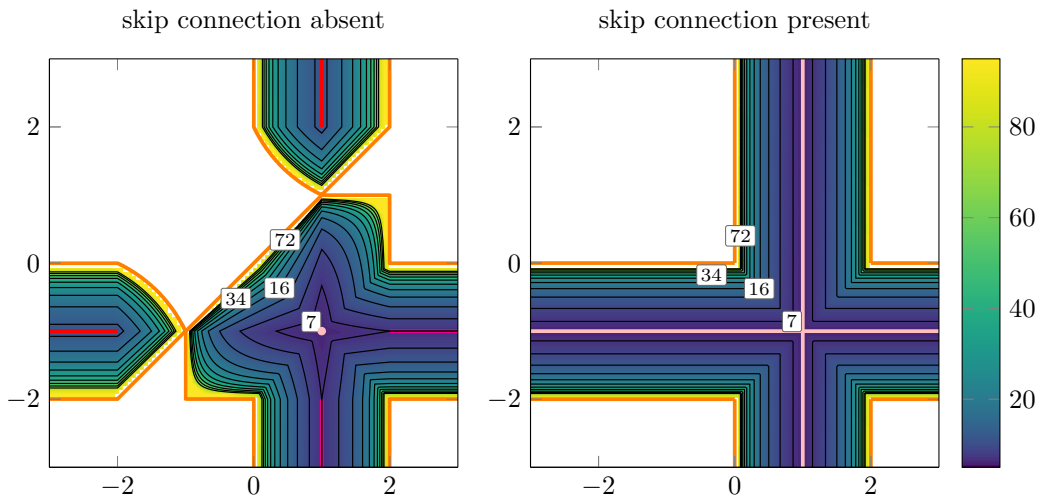


Figure 8: (This is Figure 2, reproduced here for the reader's convenience, and with a more detailed caption.) **Left:** For the dataset in Theorem 5 (i), here we depict the norm landscape of interpolating two-layer ReLU networks of width 4 and without a skip connection, i.e., the objective landscape of prob. (P) with $m = 4$. The symmetry of the dataset allows us to obtain an informative two-dimensional plot as follows. The network consists of: a pair of balanced neurons $(a_\dagger, w_\dagger, b_\dagger), (a'_\dagger, w'_\dagger, b'_\dagger)$ of opposite signs ($a_\dagger < 0$ and $a'_\dagger > 0$), same slopes ($a_\dagger w_\dagger = t_\dagger = a'_\dagger w'_\dagger$), and symmetric kinks ($-b_\dagger/w_\dagger = p_\dagger$ and $-b'_\dagger/w'_\dagger = -p_\dagger$); and another pair of balanced neurons $(a_\ddagger, w_\ddagger, b_\ddagger), (a'_\ddagger, w'_\ddagger, b'_\ddagger)$ of opposite signs ($a_\ddagger > 0$ and $a'_\ddagger < 0$), same slopes ($a_\ddagger w_\ddagger = t_\ddagger = a'_\ddagger w'_\ddagger$), and symmetric kinks ($-b_\ddagger/w_\ddagger = p_\ddagger$ and $-b'_\ddagger/w'_\ddagger = -p_\ddagger$). The horizontal and vertical axes of the plot are parameterized by the kink positions $p_\dagger$ and $p_\ddagger$ respectively, and the slope values $t_\dagger$ and $t_\ddagger$ are optimized. The boundary of the feasible set is depicted in orange. Objective values, which diverge to infinity on approaches to the feasibility boundary, are capped at 95 in this plot. For greater visibility, the contour lines are placed at values $6 + k(k+1)/2$ for $k \in [12]$, and the color map for the objective is also quadratic. Colored red are two plateaus with value 11 (at $p_\dagger \le -2, p_\ddagger = -1$ and at $p_\dagger = 1, p_\ddagger \ge 2$), colored magenta are two plateaus with value 7 (at $p_\dagger = 1, p_\ddagger \le -2$ and at $p_\dagger \ge 2, p_\ddagger = -1$), and colored pink is the minimizer whose value is 6 (at $p_\dagger = 1, p_\ddagger = -1$). The red plateaus depicted here are symmetric one-dimensional slices of two-dimensional plateaus as in Figure 9 top, and the magenta plateaus depicted here are symmetric one-dimensional slices of two-dimensional plateaus as in Figure 9 middle. **Right:** This is the corresponding symmetric profile when a skip connection is added, i.e., here we depict the objective landscape of prob. (P_1) with $m = 4$. The symmetry of the dataset allows us to consider only a skip connection $a_0 x$ with zero bias, and its weight a_0 is optimized. The feasible set becomes connected, and the suboptimal plateaus disappear. Colored pink is the plateau of minimizers whose value is 6 (at $p_\dagger = 1$ and at $p_\ddagger = -1$).

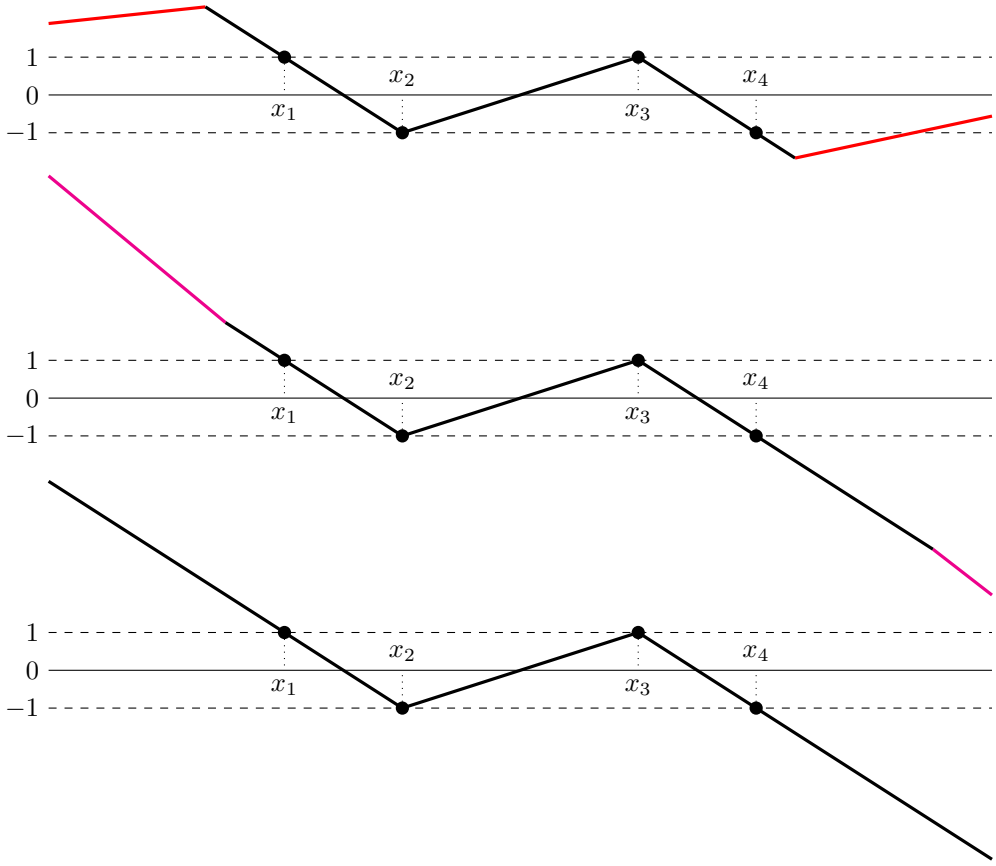


Figure 9: **Top:** Here we plot in function space a KKT point of prob. (P) constructed in the proof of Theorem 5 (i), which is not a minimizer. The red coloring indicates where this interpolant of the depicted dataset fails to be convexity correct, which by Theorem 1 (i) means that this example is suboptimal already in function space. This is one from a two-dimensional plateau of KKT points of cost 11, which can be obtained by placing the first kink arbitrarily in the interval $(-\infty, x_1]$ and placing the last kink arbitrarily in the interval $[x_4, \infty)$ (those choices then determine the initial and final slopes of the KKT point in function space). **Middle:** Here we plot another KKT point of prob. (P) which is not a minimizer, and where the failures of convexity correctness of this interpolant are indicated in magenta. This KKT point is from a different two-dimensional plateau in which the cost is 7. **Bottom:** Here we plot the unique in function space minimizer of prob. (P) for the depicted dataset, whose cost is 6.

We already have that θ is a feasible point of prob. (P), and all four interpolation constraints are tight, so to show that θ is a KKT point, it suffices to exhibit multipliers $\mu_i \geq 0$ such that

$$\partial\Big(\frac{1}{2}\sum_{j=1}^{4}(a_j^2 + w_j^2)\Big) = (a_j, w_j, 0)_{j=1}^4 \in \sum_{i=1}^{4} \mu_i \, \partial(y_i \, f_\theta(x_i)).$$

Thanks to the balancedness $a_j^2 = w_j^2$, in turn it suffices to restrict our attention to the hidden weights and biases, so after a little simplification, the equilibrium inclusion becomes the following system of eight equations in $(\mu_i)_{i=1}^4$ and ρ_2, ρ_3, where the latter variables determine the subgradients in the Clarke subdifferentials of the two ReLUs at x_2, x_3 respectively:

$$\begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \end{bmatrix}^\top \begin{bmatrix} -y_1x_1 & 0 & y_1x_1 & -y_1x_1 \\ -y_2x_2 & \rho_2 y_2 x_2 & y_2x_2 & -y_2x_2 \\ -y_3x_3 & y_3x_3 & \rho_3 y_3 x_3 & -y_3x_3 \\ -y_4x_4 & y_4x_4 & 0 & -y_4x_4 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}^\top$$

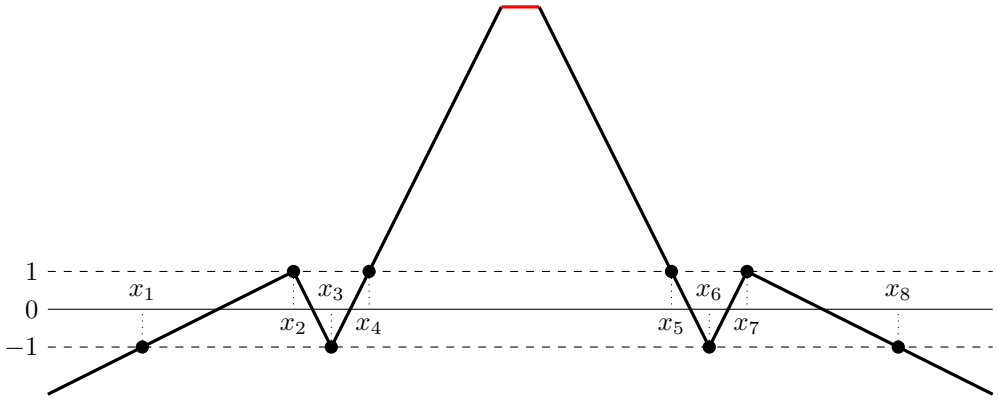


Figure 10: This shows in function space the KKT point of prob. (P^{b}) constructed in the proof of Theorem 5 (ii), which is not a minimizer. The red coloring indicates where this interpolant of the depicted dataset fails to be single turning, which by Theorem 1 (ii) means that this example is suboptimal already in function space (and not sparsest).

$$\begin{bmatrix}\mu_1\\ \mu_2\\ \mu_3\\ \mu_4\end{bmatrix}^\top \begin{bmatrix}-y_1 & 0 & y_1 & -y_1\\ -y_2 & \rho_2 y_2 & y_2 & -y_2\\ -y_3 & y_3 & \rho_3 y_3 & -y_3\\ -y_4 & y_4 & 0 & -y_4\end{bmatrix} = \begin{bmatrix}0\\0\\0\\0\end{bmatrix}^\top .$$

This system has a unique solution

$$\mu_1 = 2 \qquad \mu_2 = \frac{7}{2} \qquad \mu_3 = \frac{7}{2} \qquad \mu_4 = 2$$
$$\rho_2 = \frac{3}{7} \qquad \rho_3 = \frac{3}{7},$$

in which all μ_i are nonnegative and both of ρ_2, ρ_3 are in the closed interval $[0, 1]$, as required.

Using Theorem 4 and Theorem 1 (i), it is straightforward to compute that the minimum of prob. (P) for our dataset is $2(1 + 2) = 6$, however the norm of θ (without the biases) is

$$\sum_{j=1}^{4} |t_j| = \frac{5p_4}{p_4 - p_1} + 3 + 3 - \frac{5p_1}{p_4 - p_1} = 11,$$

so it is not a minimizer. The gap of 5 is exactly the cost of the two kinks of f_θ at p_1 and p_4, and is constant on the two-dimensional plateau obtained by varying p_1 and p_4.

That f_θ is not convexity correct is because of the two kinks at p_1 and p_4, which are not in any intermediate same-label segment of the dataset.

For part (ii), let us write $(x_i, y_i)_{i=1}^8$ for the dataset and $\theta = (a_j, w_j, b_j)_{j=1}^6$ for the network. We will follow the same pattern as for part (i), however the details are more involved.

We will define θ so that

- f_θ is switch hugging, i.e., $f_\theta(x_i) = y_i$ for all $i \in [8]$,
- the kinks of the six neurons are at

  $$p_1 \coloneqq x_2 \qquad p_2 \coloneqq x_3 \qquad p_3 \qquad p_4 \qquad p_5 \coloneqq x_6 \qquad p_6 \coloneqq x_7$$

  respectively, where $p_3, p_4 \in (x_4, x_5)$ are to be determined, and
- the first (resp., last) three neurons are active in the positive (resp., negative) direction, i.e., $w_1, w_2, w_3 > 0$ and $w_4, w_5, w_6 < 0$.

With the latter two properties, the switch hugging constraints are a system of eight equations in the slopes $t_j \coloneqq a_j w_j$ of the six neurons and in the two kinks p_3 and p_4, which has the following unique solution:

$$t_1 = -\frac{5}{2} \qquad t_2 = 4 \qquad t_3 = -2 \qquad t_4 = 2 \qquad t_5 = -4 \qquad t_6 = \frac{5}{2}$$
$$p_3 = -\frac{1}{2} \qquad p_4 = \frac{1}{2}.$$

This satisfies $p_3, p_4 \in (x_4, x_5) = (-4, 4)$, and we remark that the six signs of $a_j = t_j / w_j$ are such that f_θ is convexity correct.

To obtain a KKT point, we need to have $a_j^2 = w_j^2 + b_j^2$ for all $j \in [6]$, so from the values of p_j and t_j we define

$$|a_j| \coloneqq \sqrt[4]{1 + p_j^2}\sqrt{|t_j|} \qquad\qquad |w_j| \coloneqq \frac{\sqrt{|t_j|}}{\sqrt[4]{1 + p_j^2}},$$

then set the sign of w_j as we specified above, set the sign of a_j to satisfy $t_j = a_j w_j$, and define $b_j \coloneqq -w_j p_j$.

We already have that θ is a feasible point of prob. (P^{b}), and all eight interpolation constraints are tight, so to show that θ is a KKT point, it suffices to exhibit multipliers $\mu_i \geq 0$ such that

$$\partial\Big(\frac{1}{2}\sum_{j=1}^{6}(a_j^2 + w_j^2 + b_j^2)\Big) = (a_j, w_j, b_j)_{j=1}^{6} \in \sum_{i=1}^{8} \mu_i \, \partial(y_i \, f_\theta(x_i)).$$

Thanks to the balancedness $a_j^2 = w_j^2 + b_j^2$, in turn it suffices to restrict our attention to the hidden weights and biases, so after a little simplification, the equilibrium inclusion becomes the following system of 12 equations in $(\mu_i)_{i=1}^8$ and $\rho_1, \rho_2, \rho_5, \rho_6$, where the latter variables determine the subgradients in the Clarke subdifferentials of the first two and the last two ReLUs at x_2, x_3, x_6, x_7 respectively:

$$\begin{bmatrix}\mu_1\\ \mu_2\\ \mu_3\\ \mu_4\\ \mu_5\\ \mu_6\\ \mu_7\\ \mu_8\end{bmatrix}^{\top} \begin{bmatrix} 0 & 0 & 0 & y_1x_1 & -y_1x_1 & y_1x_1 \\ -\rho_1 y_2 x_2 & 0 & 0 & y_2x_2 & -y_2x_2 & y_2x_2 \\ -y_3x_3 & \rho_2 y_3 x_3 & 0 & y_3x_3 & -y_3x_3 & y_3x_3 \\ -y_4x_4 & y_4x_4 & 0 & y_4x_4 & -y_4x_4 & y_4x_4 \\ -y_5x_5 & y_5x_5 & -y_5x_5 & 0 & -y_5x_5 & y_5x_5 \\ -y_6x_6 & y_6x_6 & -y_6x_6 & 0 & -\rho_5 y_6 x_6 & y_6x_6 \\ -y_7x_7 & y_7x_7 & -y_7x_7 & 0 & 0 & \rho_6 y_7 x_7 \\ -y_8x_8 & y_8x_8 & -y_8x_8 & 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{1+p_1^2}} \\ \frac{1}{\sqrt{1+p_2^2}} \\ \frac{1}{\sqrt{1+p_3^2}} \\ \frac{1}{\sqrt{1+p_4^2}} \\ \frac{1}{\sqrt{1+p_5^2}} \\ \frac{1}{\sqrt{1+p_6^2}} \end{bmatrix}^{\top}$$

$$\begin{bmatrix}\mu_1\\ \mu_2\\ \mu_3\\ \mu_4\\ \mu_5\\ \mu_6\\ \mu_7\\ \mu_8\end{bmatrix}^{\top} \begin{bmatrix} 0 & 0 & 0 & y_1 & -y_1 & y_1 \\ -\rho_1 y_2 & 0 & 0 & y_2 & -y_2 & y_2 \\ -y_3 & \rho_2 y_3 & 0 & y_3 & -y_3 & y_3 \\ -y_4 & y_4 & 0 & y_4 & -y_4 & y_4 \\ -y_5 & y_5 & -y_5 & 0 & -y_5 & y_5 \\ -y_6 & y_6 & -y_6 & 0 & -\rho_5 y_6 & y_6 \\ -y_7 & y_7 & -y_7 & 0 & 0 & \rho_6 y_7 \\ -y_8 & y_8 & -y_8 & 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} -\frac{p_1}{\sqrt{1+p_1^2}} \\ -\frac{p_2}{\sqrt{1+p_2^2}} \\ -\frac{p_3}{\sqrt{1+p_3^2}} \\ -\frac{p_4}{\sqrt{1+p_4^2}} \\ -\frac{p_5}{\sqrt{1+p_5^2}} \\ -\frac{p_6}{\sqrt{1+p_6^2}} \end{bmatrix}^{\top}.$$

This system has a unique solution

$$\mu_1 = \mu_8 = \frac{5\sqrt{37} - 12\sqrt{5}}{20} \qquad \mu_2 = \mu_7 = \frac{5\sqrt{37} + 4\sqrt{26} - 4\sqrt{5}}{4}$$
$$\mu_3 = \mu_6 = \frac{5\sqrt{37} + 10\sqrt{26} + 6\sqrt{5}}{5} \qquad \mu_4 = \mu_5 = \frac{5\sqrt{26} + 7\sqrt{5}}{5}$$
$$\rho_1 = \rho_6 = \frac{4\big(31\sqrt{37} + 37\sqrt{26}\big)}{37\big(5\sqrt{37} + 4\sqrt{26} - 4\sqrt{5}\big)} \qquad \rho_2 = \rho_5 = \frac{3\big(35\sqrt{26} + 52\sqrt{5}\big)}{26\big(5\sqrt{37} + 10\sqrt{26} + 6\sqrt{5}\big)},$$

in which all μ_i are nonnegative and each of $\rho_1, \rho_2, \rho_5, \rho_6$ is in the closed interval $[0, 1]$, as required.

That θ is not a minimizer of prob. (P^{b}) is due to the suboptimality of the two kinks at p_3 and p_4, whose combined cost of $2 \cdot 2\sqrt{1 + (1/2)^2} = 2\sqrt{5}$ is greater than the cost of 4 for the optimal one kink at 0 with slope change of 4.

Finally, f_θ is not single turning because of the two kinks at p_3 and p_4, and not sparsest because it has 6 kinks rather than the minimal 5. □

Remark 1. Theorem 5 is robust to sufficiently small independent perturbations of the input locations in both parts; the symmetric datasets are used only to simplify the constructions. We also record the following additional properties, whose direct but lengthy verifications we omit because they are not used elsewhere: the KKT plateau constructed in part (i) consists of local minimizers of prob. (P); within that family, only the point whose outer kinks are at x_1 and x_4 is KKT for prob. (P^{b}), and it is suboptimal; and this point and the KKT point constructed in part (ii) are strict local minimizers of prob. (P^{b}).

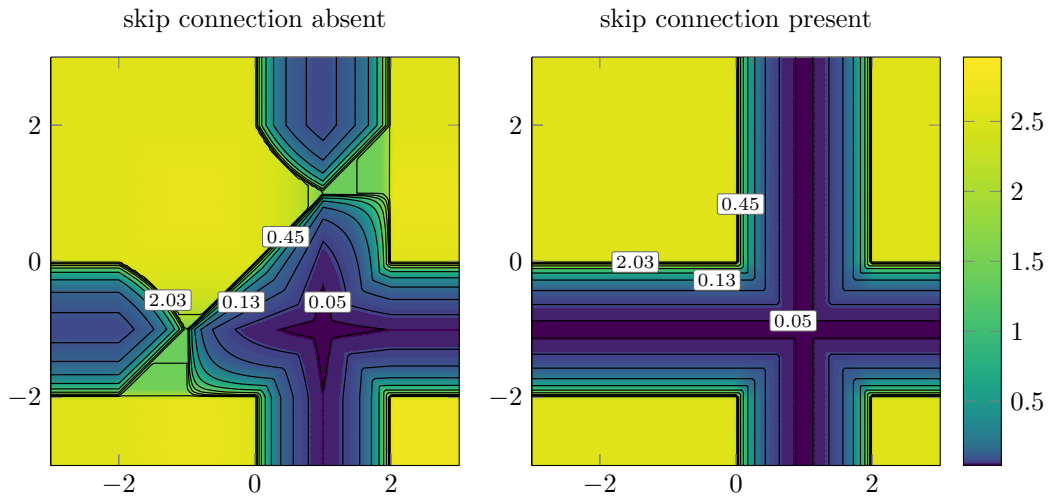


Figure 11: For the dataset in Theorem 5 (i) and two-layer ReLU networks of width 4, here we depict the landscape of the empirical logistic loss with ℓ_2 regularization of strength $\lambda = 0.001$ of network weights (biases are not regularized). The remainder of the setup is analogous to that for Figure 8: the network is symmetric thanks to the symmetry of the dataset, with neuron kinks at $\pm p_\dagger$ and $\pm p_\ddagger$, which parameterize the axes in the plots. The neuron slopes and the skip connection weight (if present) are optimized. For greater visibility, the contour lines and the color map follow superlinear progressions. **Left:** The skip connection is absent, so the landscape is of $L^{0.001}$ with $m = 4$. **Right:** The skip connection is present, so the landscape is of $L_1^{0.001}$ with $m = 4$. Within this two-dimensional profile, adding the skip connection merges the three visible valleys and removes the suboptimal flat regions, paralleling the qualitative change in Figure 8.

G Proofs for minimizing ℓ_2-regularized logistic loss

In this section, we develop several auxiliary results that, together with some of the lemmas from appendix E, contribute to the proof of Theorem 6, which is our next main result and the counterpart of Theorem 1, where here the focus is on minimizing ℓ_2-regularized logistic loss instead of minimizing interpolator norm.

Before that effort, we provide in Figure 11 optimization landscape plots that correspond to those in Figure 8 and therefore provide numerical evidence that an analogous qualitative landscape phenomenon occurs for the small-λ regularized loss.

Losses and regularizations. Let $\ell(z) \coloneqq \log(1 + \mathrm{e}^{-z})$ denote the logistic loss function, and

$$L(\theta, a_0, b_0) \coloneqq \frac{1}{n} \sum_{i=1}^{n} \ell\big(y_i \, f_{\theta, a_0, b_0}(x_i)\big)$$

denote the logistic loss over the dataset $(x_i, y_i)_{i=1}^n$. When the skip connection is absent, instead of $L(\theta, 0, 0)$ we may write just $L(\theta)$.

We consider the following four variants of the empirical logistic loss with ℓ_2 regularization of strength $\lambda > 0$, depending on whether biases are not [eqs. (L^λ) and (L_1^λ)] or are [eqs. ($\mathrm{L}^{\lambda,\mathrm{b}}$) and ($\mathrm{L}_1^{\lambda,\mathrm{b}}$)] regularized, and on whether a skip connection is absent [eqs. (L^λ) and ($\mathrm{L}^{\lambda,\mathrm{b}}$)] or present [eqs. ($\mathrm{L}_1^\lambda$) and ($\mathrm{L}_1^{\lambda,\mathrm{b}}$)]. The skip connection is always "free", i.e., its parameters a_0 and b_0 are never regularized.

$$L^\lambda(\theta) \coloneqq L(\theta) + \frac{\lambda}{2} \sum_{j=1}^{m} (a_j^2 + w_j^2) \tag{L^λ}$$

$$L_1^\lambda(\theta, a_0, b_0) \coloneqq L(\theta, a_0, b_0) + \frac{\lambda}{2} \sum_{j=1}^{m} (a_j^2 + w_j^2) \tag{L_1^λ}$$

$$L^{\lambda,\mathrm{b}}(\theta) \coloneqq L(\theta) + \frac{\lambda}{2}\sum_{j=1}^{m}(a_j^2 + w_j^2 + b_j^2) \tag{$\mathrm{L}^{\lambda,\mathrm{b}}$}$$

$$L_1^{\lambda,\mathrm{b}}(\theta, a_0, b_0) \coloneqq L(\theta, a_0, b_0) + \frac{\lambda}{2}\sum_{j=1}^{m}(a_j^2 + w_j^2 + b_j^2) \tag{$\mathrm{L}_1^{\lambda,\mathrm{b}}$}$$

Our first auxiliary result is a basic proposition that spells out a sufficient condition for impossibility that a network with the skip connection is a Clarke stationary point of the regularized loss (the biases can optionally be penalized). In contrast to showing impossibility of KKT points for interpolator norm minimization, where the directional derivative of the square ℓ_2-norm needed to be negative and Clarke directional derivatives of the negative margins needed to be nonpositive only at training points where interpolation is tight (cf. Proposition 1), here all training points are involved and it suffices that any one of the derivatives is negative (i.e., not necessarily that of the ℓ_2-norm).

Proposition 4. *For any regularization strength $\lambda > 0$, if there exists a direction θ', a_0', b_0' such that $\mathrm{D}\big(\frac{1}{2}\sum_{j=1}^{m}(a_j^2 + w_j^2)\big)[\theta', a_0', b_0'] \leq 0$ (resp., $\mathrm{D}\big(\frac{1}{2}\sum_{j=1}^{m}(a_j^2 + w_j^2 + b_j^2)\big)[\theta', a_0', b_0'] \leq 0$), $\mathrm{D}^\circ\big(-y_i\, f_{\theta,a_0,b_0}(x_i)\big)[\theta', a_0', b_0'] \leq 0$ for all $i \in [n]$, and at least one of those $1 + n$ inequalities is strict, then θ, a_0, b_0 is not a Clarke stationary point of L_1^{λ} (resp., $L_1^{\lambda,\mathrm{b}}$).*

Proof. We consider $L_1^{\lambda,\mathrm{b}}$, the proof for L_1^{λ} is analogous and simpler. For a contradiction, suppose that $\mathbf{0} \in \partial L_1^{\lambda,\mathrm{b}}(\theta, a_0, b_0)$. Then there exist subgradients $\big(\theta^{(i)}, a_0^{(i)}, b_0^{(i)}\big) \in \partial\big(-y_i\, f_{\theta,a_0,b_0}(x_i)\big)$ for each $i \in [n]$ such that

$$\mathbf{0} = \frac{1}{n}\sum_{i=1}^{n}\Big(-\ell'\big(y_i\, f_{\theta,a_0,b_0}(x_i)\big)\Big)\big(\theta^{(i)}, a_0^{(i)}, b_0^{(i)}\big) + \lambda\nabla\Big(\frac{1}{2}\sum_{j=1}^{m}(a_j^2 + w_j^2 + b_j^2)\Big),$$

so recalling that the derivative of the logistic loss function is negative, we have

$$\begin{aligned}
0 &= \Bigg\langle \frac{1}{n}\sum_{i=1}^{n}\Big(-\ell'\big(y_i\, f_{\theta,a_0,b_0}(x_i)\big)\Big)\big(\theta^{(i)}, a_0^{(i)}, b_0^{(i)}\big) + \lambda\nabla\Big(\frac{1}{2}\sum_{j=1}^{m}(a_j^2 + w_j^2 + b_j^2)\Big), \\
&\qquad (\theta', a_0', b_0') \Bigg\rangle \\
&= \frac{1}{n}\sum_{i=1}^{n}\Big(-\ell'\big(y_i\, f_{\theta,a_0,b_0}(x_i)\big)\Big)\Big\langle \big(\theta^{(i)}, a_0^{(i)}, b_0^{(i)}\big), (\theta', a_0', b_0')\Big\rangle \\
&\quad + \lambda\Bigg\langle \nabla\Big(\frac{1}{2}\sum_{j=1}^{m}(a_j^2 + w_j^2 + b_j^2)\Big), (\theta', a_0', b_0')\Bigg\rangle \\
&\leq \frac{1}{n}\sum_{i=1}^{n}\Big(-\ell'\big(y_i\, f_{\theta,a_0,b_0}(x_i)\big)\Big)\mathrm{D}^\circ\big(-y_i\, f_{\theta,a_0,b_0}(x_i)\big)[\theta', a_0', b_0'] \\
&\quad + \lambda\,\mathrm{D}\Big(\frac{1}{2}\sum_{j=1}^{m}(a_j^2 + w_j^2 + b_j^2)\Big)[\theta', a_0', b_0'] \\
&< 0.
\end{aligned}$$

□

Minimizers of the regularized losses. We start with a key lemma that, provided the regularization strength λ is sufficiently small depending on the dataset (and on whether the biases are penalized), shows that the optimization problems that determine the pairs of margins incident to all label switches (cf. the switch optimal property in Theorem 6, defined on page 44) have strictly convex objectives, convex feasibility sets, and unique minimizers. Before stating and proving the lemma, we define the optimization problems.

Given $\tau_k > 0$ and $\pi_k > 0$ for all $k \in [r]$, let $g_{\tau,\pi}\colon \mathbb{R} \to \mathbb{R}$ denote the CPA that has exactly r slopes and, for all $k \in [r]$, passes through points $(x_{\iota(k)}, y_{\iota(k)}\tau_k)$ and $(x_{\iota(k)+1}, y_{\iota(k)+1}\pi_k)$; this is

well defined if and only if vectors τ and π satisfy

$$\begin{aligned}\Phi(\tau,\pi) \coloneqq \forall k \in [r-1],\ -\frac{\tau_k+\pi_k}{x_{\iota(k)+1}-x_{\iota(k)}}(x_{\iota(k+1)}-x_{\iota(k)+1}) \\ \leq \pi_k - \tau_{k+1} \\ \leq \frac{\tau_{k+1}+\pi_{k+1}}{x_{\iota(k+1)+1}-x_{\iota(k+1)}}(x_{\iota(k+1)}-x_{\iota(k)+1}).\end{aligned}$$

Recalling from appendix E the CPA norms R_1 and $\mathrm{R}_1^{\mathrm{b}}$, we consider the following two variants of the problem of minimizing the empirical logistic loss with norm regularization of strength $\lambda > 0$.

$$\inf_{\tau,\pi\in\mathbb{R}^r_{>0}} \left(\frac{1}{n}\sum_{i=1}^n \ell\big(y_i\, g_{\tau,\pi}(x_i)\big) + \lambda\, \mathrm{R}_1(g_{\tau,\pi})\right) \quad \text{such that } \Phi(\tau,\pi) \tag{O_1}$$

$$\inf_{\tau,\pi\in\mathbb{R}^r_{>0}} \left(\frac{1}{n}\sum_{i=1}^n \ell\big(y_i\, g_{\tau,\pi}(x_i)\big) + \lambda\, \mathrm{R}_1^{\mathrm{b}}(g_{\tau,\pi})\right) \quad \text{such that } \Phi(\tau,\pi) \tag{$\mathrm{O}_1^{\mathrm{b}}$}$$

Lemma 18. *Suppose the dataset has at least two label switches, i.e., $r \geq 2$, and let:*

$$d_{\min} \coloneqq \min_{k\in[r]}(x_{\iota(k)+1}-x_{\iota(k)}) \qquad\qquad x_{\max} \coloneqq \max_{i\in[n]}|x_i|.$$

We have that each of probs. (O_1) and ($\mathrm{O}_1^{\mathrm{b}}$) has a strictly convex objective and a convex feasible set, and moreover that:

(i) provided $\lambda \leq d_{\min}/24(r-1)n(\log n)$, prob. ($\mathrm{O}_1$) has a unique minimizer;

(ii) provided $\lambda \leq d_{\min}/24(r-1)n(\log n)\sqrt{1+x_{\max}^2}$, prob. ($\mathrm{O}_1^{\mathrm{b}}$) has a unique minimizer.

Proof. We first consider part (i), where by Theorem 4 we have that

$$\mathrm{R}_1(g_{\tau,\pi}) = \frac{\tau_1+\pi_1}{x_{\iota(1)+1}-x_{\iota(1)}} + 2\sum_{k=2}^{r-1}\frac{\tau_k+\pi_k}{x_{\iota(k)+1}-x_{\iota(k)}} + \frac{\tau_r+\pi_r}{x_{\iota(r)+1}-x_{\iota(r)}},$$

so the regularizing term is linear in vectors τ and π, and thus convex.

For every input x_i that is in an intermediate same-label segment of the dataset, i.e., such that $i \in \{\iota(k)+1,\dots,\iota(k+1)\}$ for some $k \in [r-1]$, we have that

$$\begin{aligned}\ell\big(y_i\, g_{\tau,\pi}(x_i)\big) = \max\Bigg\{&\ell\bigg(\pi_k + \frac{\tau_k+\pi_k}{x_{\iota(k)+1}-x_{\iota(k)}}(x_i - x_{\iota(k)+1})\bigg), \\ &\ell\bigg(\tau_{k+1} + \frac{\tau_{k+1}+\pi_{k+1}}{x_{\iota(k+1)+1}-x_{\iota(k+1)}}(x_{\iota(k+1)} - x_i)\bigg)\Bigg\},\end{aligned}$$

which is the maximum of two compositions of the logistic loss function ℓ with linear functions in τ and π, and thus also convex. For inputs x_i that are in the first or last same-label segments of the dataset, the argument is simpler as no maximum and only one composition is involved.

Hence the objective of prob. (O_1) is convex, and moreover strictly convex since, for every variable τ_k, it contains the term $\frac{1}{n}\ell(\tau_k)$ (at $x_{\iota(k)}$), and for every variable π_k, it contains the term $\frac{1}{n}\ell(\pi_k)$ (at $x_{\iota(k)+1}$).

Since the feasibility condition $\Phi(\tau,\pi)$ is a conjunction of linear inequalities, the feasible set of prob. (O_1) is convex.

To complete the proof for part (i), it suffices to show that the feasible set has a compact convex subset such that the infimum of the objective outside that subset is greater than the value of the objective at a point in the subset. We will obtain such a compact convex subset by bounding all variables τ_k and π_k away from zero and away from infinity; that it is compact and convex will follow from its boundedness together with nonstrictness and affinity of all its defining inequalities.

Let $z \coloneqq \log(2n/\ell(0))$, and let $\tau_k^\dagger \coloneqq z$ and $\pi_k^\dagger \coloneqq z$ for all $k \in [r]$. Then the objective of prob. (O_1) at $\tau^\dagger, \pi^\dagger$ is less than

$$\ell(z) + \frac{4\lambda z(r-1)}{d_{\min}} < \mathrm{e}^{-z} + \frac{z}{6n(\log n)} = \frac{\ell(0)}{2n} + \frac{\log(n) + \log(2/\ell(0))}{6n(\log n)} < \frac{\ell(0)}{2n} + \frac{2\log(n)}{6n(\log n)} = \frac{\ell(0) + 2/3}{2n} < \frac{\ell(0)}{n}.$$

Since the objective of prob. (O_1) at any feasible point τ, π is greater than $\ell\big(\min\{\tau_k, \pi_k \mid k \in [r]\}\big)/n$, it follows that there exists $\varepsilon > 0$ such that the infimum over all feasible points τ, π with $\tau_k < \varepsilon$ or $\pi_k < \varepsilon$ for some $k \in [r]$ is greater than the value at $\tau^\dagger, \pi^\dagger$.

It remains to observe that $R_1(g_{\tau,\pi})$ is linear in every variable τ_k and every variable π_k with a positive coefficient, and so there exists $N > 0$ such that the infimum of the objective of prob. (O_1) over all feasible points τ, π with $\tau_k > N$ or $\pi_k > N$ for some $k \in [r]$ is greater than the value at $\tau^\dagger, \pi^\dagger$.

For part (ii), the reasoning follows the same pattern. From Theorem 4, it is straightforward to derive that

$$\mathrm{R}_1^{\mathrm{b}}(g_{\tau,\pi}) = \sum_{k=1}^{r-1} \sqrt{\left(\frac{\tau_k + \pi_k}{x_{\iota(k)+1} - x_{\iota(k)}} + \frac{\tau_{k+1} + \pi_{k+1}}{x_{\iota(k+1)+1} - x_{\iota(k+1)}}\right)^2 + \left(\frac{x_{\iota(k)+1}\tau_k + x_{\iota(k)}\pi_k}{x_{\iota(k)+1} - x_{\iota(k)}} + \frac{x_{\iota(k+1)+1}\tau_{k+1} + x_{\iota(k+1)}\pi_{k+1}}{x_{\iota(k+1)+1} - x_{\iota(k+1)}}\right)^2}$$

and so the unscaled regularizer is a sum of $r-1$ Euclidean lengths of vectors that are linear in τ and π, thus convex.

The only other significant difference is in bounding the variables away from infinity; for that, it suffices to recall that $\mathrm{R}_1^{\mathrm{b}}(g_{\tau,\pi}) \geq \mathrm{R}_1(g_{\tau,\pi})$. □

Corollary 3. *Suppose $r \geq 2$ and λ satisfies the upper bound in part (i) (resp., (ii)) of Lemma 18. Then the unique minimizer of prob. (O_1) (resp., prob. ($\mathrm{O}_1^{\mathrm{b}}$)) has value less than $\ell(0)/n$ and lies in $\mathbb{R}^{2r}_{>0}$. Also networks whose L_1^λ (resp., $L_1^{\lambda,\mathrm{b}}$) values are less than $\ell(0)/n$ exist, and every such network has a positive margin.*

Proof. Let $z \coloneqq \log(2n/\ell(0))$, and let $\tau_k^\dagger \coloneqq z$ and $\pi_k^\dagger \coloneqq z$ for all $k \in [r]$. Then $\Phi(\tau^\dagger, \pi^\dagger)$ holds, and the calculation in the proof of Lemma 18 gives $\frac{1}{n}\sum_{i=1}^n \ell\big(y_i\, g_{\tau^\dagger,\pi^\dagger}(x_i)\big) + \lambda \mathrm{R}_1(g_{\tau^\dagger,\pi^\dagger}) < \ell(0)/n$. Under the bound in Lemma 18 (ii), the same inequality holds with $\mathrm{R}_1^{\mathrm{b}}$, because $\mathrm{R}_1^{\mathrm{b}}(g_{\tau^\dagger,\pi^\dagger}) \leq \sqrt{1 + x_{\max}^2}\,\mathrm{R}_1(g_{\tau^\dagger,\pi^\dagger})$ as every kink of $g_{\tau^\dagger,\pi^\dagger}$ belongs to $[x_1, x_n]$. By Theorem 4, $g_{\tau^\dagger,\pi^\dagger}$ has a network representor attaining the relevant regularizer.

On the other hand, if a network function f_{θ,a_0,b_0} has a nonpositive margin, then $y_i\, f_{\theta,a_0,b_0}(x_i) \leq 0$ for some $i \in [n]$, and therefore

$$L_1^{\lambda,\mathrm{b}}(\theta, a_0, b_0) \geq L_1^\lambda(\theta, a_0, b_0) \geq \frac{1}{n}\sum_{i'=1}^n \ell\big(y_{i'}\, f_{\theta,a_0,b_0}(x_{i'})\big) \geq \frac{\ell(0)}{n}.$$

Finally, if one allows the closed domain $\tau, \pi \in \mathbb{R}^r_{\geq 0}$ in the finite-dimensional problems, any boundary point has objective at least $\ell(0)/n$, because some switch-incident margin is zero. The strict comparison above therefore places the unique minimizer in $\mathbb{R}^{2r}_{>0}$. □

We recall from appendix E two properties of a CPA $g\colon \mathbb{R} \to \mathbb{R}$ with respect to the dataset: convexity correct and single turning; and we define the following two properties:

(switch optimal) depending on whether biases are not or are regularized, for all $k \in [r]$, we have $g(x_{\iota(k)}) = y_{\iota(k)}\tau_k^\star$ and $g(x_{\iota(k)+1}) = y_{\iota(k)+1}\pi_k^\star$, where $\tau^\star, \pi^\star$ is the unique minimizer of prob. (O_1) or prob. ($\mathrm{O}_1^{\mathrm{b}}$) (respectively);

(interval turning) for all $k \in [r-1]$, letting p and p' be the first and last (respectively) kinks of g in $[x_{\iota(k)+1}, x_{\iota(k+1)}]$, the interval (p, p') contains no inputs from the dataset.

Remark 2. If g is single turning, then it is also interval turning: since within every intermediate segment of the dataset there is exactly one kink, the first kink p and the last kink p' coincide, so the interval (p, p') is empty.

Theorem 6. *Under Assumption 1, and provided $\lambda \le d_{\min}/24(r-1)n(\log n)$ if biases are not regularized (i.e., for losses L^λ and L_1^λ) or $\lambda \le d_{\min}/24(r-1)n(\log n)\sqrt{1+x_{\max}^2}$ if biases are regularized (i.e., for losses $L^{\lambda,\mathrm{b}}$ and $L_1^{\lambda,\mathrm{b}}$), all of the following hold.*

(i) Losses L^λ and L_1^λ have the same minimizers in function space (over all m), which are exactly all CPAs $g\colon \mathbb{R} \to \mathbb{R}$ that are switch optimal, convexity correct, and interval turning.

(ii) Losses $L^{\lambda,\mathrm{b}}$ and $L_1^{\lambda,\mathrm{b}}$ have the same unique minimizer in function space (over all m), which is switch optimal, convexity correct, and single turning; and thus with $r-1$ kinks in total.

(iii) For losses L_1^λ and $L_1^{\lambda,\mathrm{b}}$, we have that every Clarke stationary point $(\theta, a_0, b_0) \in \mathbb{R}^{3m+2}$ whose margin is positive (with any network width m for which such points exist) is a minimizer (globally, including over all m).

Having now stated our main result in this section, we continue the development toward its proof. The following lemma shows that the optima of the regularized loss minimization problems with the skip connection are attained, even for any fixed network width. There need to be at least two label switches in the dataset (which is implied by Assumption 1), however there is no constraint on λ.

Lemma 19. *Suppose the dataset has at least two label switches, i.e., $r \ge 2$. For each of L_1^λ and $L_1^{\lambda,\mathrm{b}}$ restricted to any network width $m \in \mathbb{N}$, and with any fixed regularization strength $\lambda > 0$, the minimum is attained.*

Proof. Fix $m \in \mathbb{N}$ and $\lambda > 0$. We first consider L_1^λ. Let $\big(\theta^{(k)}, a_0^{(k)}, b_0^{(k)}\big)_{k\in\mathbb{N}}$ be a minimizing sequence, then the maximum N of its L_1^λ values is finite. Since both the empirical logistic loss and the regularization term are nonnegative, it follows that

$$\frac{1}{2}\sum_{j=1}^{m}\Big(\big(a_j^{(k)}\big)^2 + \big(w_j^{(k)}\big)^2\Big) \le \frac{N}{\lambda}. \tag{2}$$

As in the proof of Lemma 2, we may alter each parameterization without changing its outputs at the training inputs and without increasing the regularization term. If the kink of a neuron lies outside $[x_1, x_n]$, then the contribution of that neuron is affine on all training inputs. Its contribution there may therefore be absorbed into the free affine skip connection, after which the neuron may be replaced by the zero neuron. A neuron with zero hidden weight contributes a constant and can be treated in the same way. We may thus assume that every nonzero neuron has its kink in $[x_1, x_n]$. By positive homogeneity of the ReLU, each remaining nonzero neuron can be balanced so that $|a_j^{(k)}| = |w_j^{(k)}|$ without changing the represented function or increasing the regularization term. Inequality (2) then bounds $a_j^{(k)}$ and $w_j^{(k)}$. Also $|b_j^{(k)}| \le |w_j^{(k)}| \max_{i=1}^n |x_i|$.

Thus the parameter vectors $\theta^{(k)}$ lie in a bounded set. In particular, the pre-skip output vectors $h^{(k)} := \big(f_{\theta^{(k)},0,0}(x_i)\big)_{i=1}^n$ form a bounded sequence. We now show that the skip parameters are bounded.

Suppose, to the contrary, that after passing to a subsequence, $\big\|a_0^{(k)}, b_0^{(k)}\big\|_2 \to \infty$. Passing to a further subsequence gives $(a_0^{(k)}, b_0^{(k)})/\big\|a_0^{(k)}, b_0^{(k)}\big\|_2 \to (\alpha_0, \beta_0)$ for some unit vector (α_0, β_0). For every $i \in [n]$, nonnegativity of the other summands implies $\ell\Big(y_i\big(a_0^{(k)}x_i + b_0^{(k)} + h_i^{(k)}\big)\Big) \le nN$. Since $\ell(z) = \log(1+\mathrm{e}^{-z})$ is strictly decreasing and tends to infinity as $z \to -\infty$, there exists a finite Q, depending only on n and N, such that $y_i\big(a_0^{(k)}x_i + b_0^{(k)} + h_i^{(k)}\big) \ge -Q$ for all $i \in [n]$ and all $k \in \mathbb{N}$. Dividing by $\big\|a_0^{(k)}, b_0^{(k)}\big\|_2$ and letting $k \to \infty$, using boundedness of $h^{(k)}$, yields $y_i(\alpha_0 x_i + \beta_0) \ge 0$ for every $i \in [n]$. Because the dataset has at least two label switches, there exist three training inputs

$x_{i_1} < x_{i_2} < x_{i_3}$ whose labels alternate. After multiplying $x \mapsto \alpha_0 x + \beta_0$ by the common label of the first and third points if necessary, we obtain

$$\alpha_0 x_{i_1} + \beta_0 \geq 0 \qquad \alpha_0 x_{i_2} + \beta_0 \leq 0 \qquad \alpha_0 x_{i_3} + \beta_0 \geq 0.$$

Since x_{i_2} is a strict convex combination of x_{i_1} and x_{i_3}, the affine value $\alpha_0 x_{i_2} + \beta_0$ is the same strict convex combination of the two outer values. It is therefore nonnegative, and hence it must be zero. The two nonnegative outer values must then both be zero as well. Consequently, the affine function vanishes at two distinct points and is identically zero. This gives $\alpha_0 = \beta_0 = 0$, contradicting that (α_0, β_0) is a unit vector. The skip parameters are therefore bounded.

Hence the minimizing sequence is contained in a fixed compact subset of $\mathbb{R}^{3m+2}$. Passing to a convergent subsequence and using continuity of the ReLU, the logistic loss, and the regularization term, we conclude that its limit attains the infimum of L_1^{λ}.

For $L_1^{\lambda,\mathrm{b}}$, the argument is the same, except that the boundedness of the parameters follows immediately from the bounded regularization term. Indeed, we now have

$$\frac{1}{2}\sum_{j=1}^{m}\Big(\big(a_j^{(k)}\big)^2 + \big(w_j^{(k)}\big)^2 + \big(b_j^{(k)}\big)^2\Big) \leq \frac{N}{\lambda}.$$

Thus all coordinates of $\theta^{(k)}$ are bounded, and hence so are the pre-skip outputs on the finite training set. The preceding normalization argument, which uses the existence of three ordered inputs with alternating labels, then shows that the skip parameters are bounded as well. Compactness and continuity therefore imply that the infimum of $L_1^{\lambda,\mathrm{b}}$ is attained. □

Next, we identify two local geometric irregularities that capture the failures of interval turning for a convexity-correct CPA…

Lemma 20. *If a CPA $g\colon \mathbb{R} \to \mathbb{R}$ whose kinks are $(p_k)_{k=1}^{q}$ is convexity correct, but it is not interval turning, then at least one of the following holds.*

- *(i) Some kinks p_k and $p_{k'}$ with $k < k'$ are convex and such that $y_i = -1$ for all $x_i \in [p_k, p_{k'}]$, and there exists $x_i \in (p_k, p_{k'})$.*
- *(ii) Some kinks p_k and $p_{k'}$ with $k < k'$ are concave and such that $y_i = 1$ for all $x_i \in [p_k, p_{k'}]$, and there exists $x_i \in (p_k, p_{k'})$.*

Proof. This is immediate from the definitions of convexity correctness and interval turning. □

…and for either of those irregularities, identify a direction along which the derivative of the bias-free square ℓ_2-norm of the network parameters is nonpositive, the Clarke derivative of the negative margin of each training point is nonpositive, and at least one of the latter is negative. For an illustration of a part of the proof, we refer the reader to Figure 12 left.

Lemma 21. *If $g = f_{\theta,a_0,b_0}$, whose kinks are $(p_k)_{k=1}^{q}$, satisfies at least one of properties (i) or (ii) in Lemma 20, then there exists a direction θ', a_0', b_0' such that $\mathrm{D}\big(\frac{1}{2}\sum_{j=1}^{m}(a_j^2 + w_j^2)\big)[\theta', a_0', b_0'] \leq 0$, $\mathrm{D}^{\circ}\big(-y_i\, f_{\theta,a_0,b_0}(x_i)\big)[\theta', a_0', b_0'] \leq 0$ for all $i \in [n]$, and at least one of the latter n inequalities is strict.*

Proof. We consider property (i) in Lemma 20; since property (ii) is symmetric, this is without loss of generality. Then θ must contain some neurons (a_j, w_j, b_j) and $(a_{j'}, w_{j'}, b_{j'})$ such that:

- $-b_j/w_j = p_k$ and $a_j > 0$;
- $-b_{j'}/w_{j'} = p_{k'}$ and $a_{j'} > 0$.

By the orientation-reversal reparameterization in Lemma 3, for the following construction we may assume that $w_j > 0$ and $w_{j'} > 0$; and by symmetry, also that $a_j/w_j \geq a_{j'}/w_{j'}$. Moreover, by Lemma 4, the claimed nonpositivity of the Clarke directional derivatives of $-y_i\, f_{\theta,a_0,b_0}(x_i)$ is implied by nonpositivity of the corresponding one-sided limits of the ordinary directional derivatives,

where also the claimed strictness is implied by strictness on both sides. Therefore, letting $u \coloneqq (-a_{j'}w_{j'}, a_{j'}w_j, w_{j'}b_j - w_jb_{j'})$, it suffices to show that:

- $\mathrm{D}_{a_j,a_{j'},b_{j'}}\big(\frac{1}{2}\sum_{j''=1}^{m}(a_{j''}^2 + w_{j''}^2)\big)[u] \leq 0$;
- for all $i \in [n]$ we have

$$\lim_{x \uparrow x_i} \mathrm{D}_{a_j,a_{j'},b_{j'}}\big(y_i\, f_{\theta,a_0,b_0}(x)\big)[u] \geq 0$$
$$\lim_{x \downarrow x_i} \mathrm{D}_{a_j,a_{j'},b_{j'}}\big(y_i\, f_{\theta,a_0,b_0}(x)\big)[u] \geq 0,$$

  and there exists $i \in [n]$ for which both latter inequalities are strict.

For the first claim, we have

$$\mathrm{D}_{a_j,a_{j'},b_{j'}}\bigg(\frac{1}{2}\sum_{j''=1}^{m}(a_{j''}^2 + w_{j''}^2)\bigg)[u] = a_{j'}(-a_jw_{j'} + a_{j'}w_j) \leq 0.$$

For the second claim, since $y_i = -1$ for all $x_i \in [p_k, p_{k'}]$, and there exists $x_i \in (p_k, p_{k'})$, in turn it suffices to observe that:

- for all $x \in (-\infty, p_k)$ we have

$$\begin{aligned}
&\mathrm{D}_{a_j,a_{j'},b_{j'}} f_{\theta,a_0,b_0}(x)[u] \\
&= \mathrm{D}_{a_j,a_{j'},b_{j'}}\big(a_j\,\sigma(w_jx + b_j)\big)[u] + \mathrm{D}_{a_j,a_{j'},b_{j'}}\big(a_{j'}\,\sigma(w_{j'}x + b_{j'})\big)[u] \\
&= 0;
\end{aligned}$$

- for all $x \in (p_k, p_{k'})$ we have

$$\begin{aligned}
&\mathrm{D}_{a_j,a_{j'},b_{j'}} f_{\theta,a_0,b_0}(x)[u] \\
&= \mathrm{D}_{a_j,a_{j'},b_{j'}}\big(a_j\,\sigma(w_jx + b_j)\big)[u] + \mathrm{D}_{a_j,a_{j'},b_{j'}}\big(a_{j'}\,\sigma(w_{j'}x + b_{j'})\big)[u] \\
&= -(w_jx + b_j)a_{j'}w_{j'} \\
&< 0;
\end{aligned}$$

- for all $x \in (p_{k'}, \infty)$ we have

$$\begin{aligned}
&\mathrm{D}_{a_j,a_{j'},b_{j'}} f_{\theta,a_0,b_0}(x)[u] \\
&= \mathrm{D}_{a_j,a_{j'},b_{j'}}\big(a_j\,\sigma(w_jx + b_j)\big)[u] + \mathrm{D}_{a_j,a_{j'},b_{j'}}\big(a_{j'}\,\sigma(w_{j'}x + b_{j'})\big)[u] \\
&= -(w_jx + b_j)a_{j'}w_{j'} + (w_{j'}x + b_{j'})a_{j'}w_j + a_{j'}(w_{j'}b_j - w_jb_{j'}) \\
&= 0.
\end{aligned}$$

□

The final lemma has technically the most involved proof. At its center are constructions, based on the failure of switch optimality, of directions for the network parameters along which the Clarke derivative of the regularized loss is negative. In contrast to the infinitesimal perturbations that we have seen so far in this work, which are local, here we perturb simultaneously one or two neurons per each intermediate same-label segment of the dataset (and also the skip connection). The directions are constructed so that, when they are mapped to the pairs of margins incident to all label switches, they point to the unique minimizer of the corresponding optimization problem considered in Lemma 18. For an illustration of a part of the proof, we refer the reader to Figure 12 right.

Lemma 22. *Suppose the dataset has at least two label switches, i.e., $r \geq 2$, and suppose $\lambda \leq d_{\min}/24(r-1)n(\log n)$ (resp., $\lambda \leq d_{\min}/24(r-1)n(\log n)\sqrt{1 + x_{\max}^2}$). If f_{θ,a_0,b_0} has a positive margin, and it is convexity correct and interval turning (resp., single turning), but it is not switch optimal, then θ, a_0, b_0 is not a Clarke stationary point of L_1^{λ} (resp., $L_1^{\lambda,\mathrm{b}}$).*

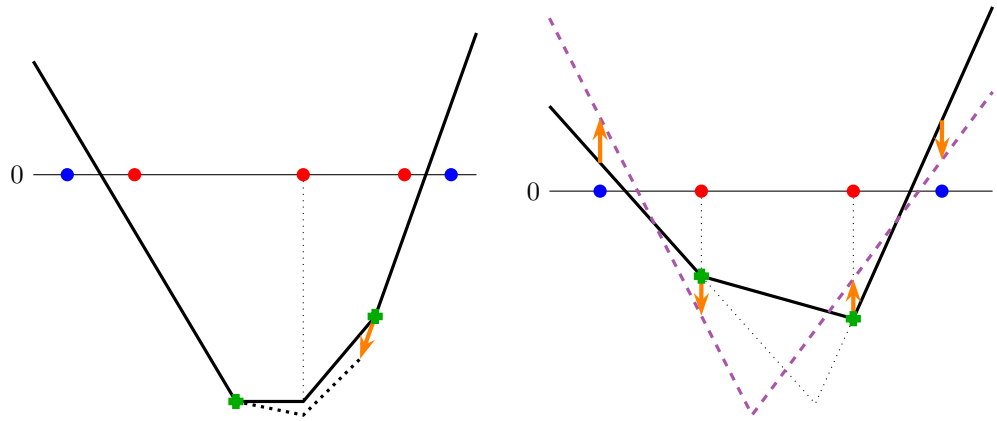


Figure 12: Since the focus in this section is on the regularized losses, we indicate the labels of the inputs in the dataset by the colors blue for 1, and red for -1. **Left:** Here we illustrate the proof of Lemma 21 for the case of property (i) in Lemma 20; in this example, one of the kinks of the network function happens to coincide with an input in the dataset. The kinks p_k and $p_{k'}$ in the proof are indicated by the green pluses. The proof identifies a direction for infinitesimally perturbing the network parameters so that the portion of the network function between p_k and $p_{k'}$ moves as indicated by the orange arrow and the dotted half line. **Right:** Here we illustrate the proof of Lemma 22 for the case when the biases are not regularized, i.e., for the loss L_1^λ; in this example, two of the kinks of the network function happen to coincide with inputs in the dataset. The two kinks are also indicated by the green pluses because they are the first and last in the intermediate dataset segment between the two red inputs, and so two neurons that contribute to those kinks are perturbed infinitesimally in the proof. The perturbations across all the intermediate dataset segments (and of the skip connection) work together to move (as indicated by the orange arrows) all pieces of the network function that cross the $y = 0$ line toward the corresponding ones of the switch optimal function $g_{\tau^\star,\pi^\star}$, which is plotted in violet and dashed.

Proof. For all $k \in [r]$, let

$$\tau_k \coloneqq y_{\iota(k)}\, f_{\theta,a_0,b_0}(x_{\iota(k)}) \qquad\qquad \pi_k \coloneqq y_{\iota(k)+1}\, f_{\theta,a_0,b_0}(x_{\iota(k)+1}).$$

Since f_{θ,a_0,b_0} has a positive margin, and it is convexity correct and interval turning (resp., single turning), we have that $g_{\tau,\pi}$ coincides with f_{θ,a_0,b_0} except possibly on the open intervals between the first and last kinks of f_{θ,a_0,b_0} in intermediate same-label segments of the dataset. Hence τ, π is feasible for prob. (O_1) (resp., prob. ($\mathrm{O}_1^{\mathrm{b}}$)) and

$$f_{\theta,a_0,b_0}(x_i) = g_{\tau,\pi}(x_i) \quad \text{for all } i \in [n].$$

But f_{θ,a_0,b_0} is not switch optimal, so recalling Lemma 18 and letting $\tau^\star, \pi^\star$ denote the unique minimizer of prob. (O_1) (resp., prob. ($\mathrm{O}_1^{\mathrm{b}}$)), we have that $(\tau', \pi') \coloneqq (\tau^\star - \tau, \pi^\star - \pi) \in \mathbb{R}^{2r}$ is a nonzero vector.

Some more notation that is going to be useful in the remainder of the proof is the following. For each $k \in [r]$, let us denote the slope and the bias of the k-th piece of $g_{\tau,\pi}$ as

$$\alpha_{\tau,\pi}^{(k)} \coloneqq -y_{\iota(k)} \frac{\tau_k + \pi_k}{x_{\iota(k)+1} - x_{\iota(k)}} \qquad\qquad \beta_{\tau,\pi}^{(k)} \coloneqq y_{\iota(k)} \frac{x_{\iota(k)+1}\tau_k + x_{\iota(k)}\pi_k}{x_{\iota(k)+1} - x_{\iota(k)}},$$

and for each $k \in [r-1]$, let us denote the k-th kink of $g_{\tau,\pi}$ as

$$\gamma_{\tau,\pi}^{(k)} \coloneqq -\frac{\beta_{\tau,\pi}^{(k+1)} - \beta_{\tau,\pi}^{(k)}}{\alpha_{\tau,\pi}^{(k+1)} - \alpha_{\tau,\pi}^{(k)}}$$

(which is necessarily either in the open interval between the first and last kinks of f_{θ,a_0,b_0} in the k-th intermediate same-label segment, or it is the only kink of f_{θ,a_0,b_0} in that segment). Also, for each $i \in [n]$, let $\widehat{k}(i) \in [r]$ be such that x_i is in the left-closed right-open domain of the $\widehat{k}(i)$-th piece of $g_{\tau,\pi}$.

To proceed, we consider the losses L_1^λ and $L_1^{\lambda,\mathrm{b}}$ as two cases.

CASE L_1^λ. First we define a direction $\theta' = (a'_j, w'_j, b'_j)_{j=1}^m, a'_0, b'_0$ such that perturbing the network along it has the same effect on f_{θ,a_0,b_0} outside the open intervals between the first and last kinks in intermediate same-label segments as perturbing the vector τ, π of margins incident to the label switches in the direction τ', π' has on $g_{\tau,\pi}$. To do this, for each $k \in [r-1]$, we identify as follows either two or one neurons to perturb whose kinks are in the k-th intermediate same-label segment; all other neurons are left fixed.

To begin the definition, we set $\theta'_{(0)} \coloneqq (0,0,0)_{j=1}^m$ and

$$a'_0 \coloneqq \mathrm{D}\big(\alpha^{(1)}_{\tau,\pi}\big)[\tau', \pi'] \qquad\qquad b'_0 \coloneqq \mathrm{D}\big(\beta^{(1)}_{\tau,\pi}\big)[\tau', \pi'].$$

The rest of the definition is inductive, where we consider $k \in [r-1]$ and assume that, for all $i \in [n]$, we have

$$\mathrm{D}^\circ\big(-y_i\, f_{\theta,a_0,b_0}(x_i)\big)[\theta'_{(k-1)}, a'_0, b'_0] = \begin{cases} \mathrm{D}\big(-y_i\, g_{\tau,\pi}(x_i)\big)[\tau', \pi'] & \text{if } \widehat{k}(i) \le k, \\ \mathrm{D}\Big(-y_i\big(\alpha^{(k)}_{\tau,\pi} x_i + \beta^{(k)}_{\tau,\pi}\big)\Big)[\tau', \pi'] & \text{if } \widehat{k}(i) > k. \end{cases} \tag{3}$$

The values of the skip connection parameters a'_0 and b'_0 above ensure that this inductive hypothesis is satisfied initially, i.e., for $k = 1$.

Let $p_k^\dagger$ and $p_k^\ddagger$ denote respectively the first and last kinks of f_{θ,a_0,b_0} in the k-th intermediate same-label segment. We distinguish two subcases depending on whether these coincide.

SUBCASE $p_k^\dagger < p_k^\ddagger$. This subcase is the more complex one, and it involves identifying two neurons to perturb, whose kinks are at $p_k^\dagger$ and $p_k^\ddagger$. By f_{θ,a_0,b_0} having a positive margin and being convexity correct, and by Lemmas 3 and 13 and Proposition 4, there exist $j_k^\dagger, j_k^\ddagger \in [m]$ for which we may assume:

- $a_{j_k^\dagger} > 0$ and $a_{j_k^\ddagger} > 0$ (resp., $a_{j_k^\dagger} < 0$ and $a_{j_k^\ddagger} < 0$) if the k-th intermediate dataset segment has label -1 (resp., 1);
- $w_{j_k^\dagger} = \big|a_{j_k^\dagger}\big|$ and $w_{j_k^\ddagger} = \big|a_{j_k^\ddagger}\big|$;
- $-b_{j_k^\dagger}/w_{j_k^\dagger} = p_k^\dagger$ and $-b_{j_k^\ddagger}/w_{j_k^\ddagger} = p_k^\ddagger$.

Now let

$$\delta_k^\dagger \coloneqq \frac{\mathrm{D}\Big(\big(\alpha^{(k+1)}_{\tau,\pi} - \alpha^{(k)}_{\tau,\pi}\big)p_k^\ddagger + \big(\beta^{(k+1)}_{\tau,\pi} - \beta^{(k)}_{\tau,\pi}\big)\Big)[\tau', \pi']}{p_k^\ddagger - p_k^\dagger}$$

$$\delta_k^\ddagger \coloneqq -\frac{\mathrm{D}\Big(\big(\alpha^{(k+1)}_{\tau,\pi} - \alpha^{(k)}_{\tau,\pi}\big)p_k^\dagger + \big(\beta^{(k+1)}_{\tau,\pi} - \beta^{(k)}_{\tau,\pi}\big)\Big)[\tau', \pi']}{p_k^\ddagger - p_k^\dagger},$$

and let $\theta'_{(k)}$ be obtained from $\theta'_{(k-1)}$ by setting

$$a'_{j_k^\dagger} \coloneqq \frac{\delta_k^\dagger}{2w_{j_k^\dagger}} \qquad w'_{j_k^\dagger} \coloneqq \frac{\delta_k^\dagger}{2a_{j_k^\dagger}} \qquad b'_{j_k^\dagger} \coloneqq -\frac{\delta_k^\dagger p_k^\dagger}{2a_{j_k^\dagger}}$$

$$a'_{j_k^\ddagger} \coloneqq \frac{\delta_k^\ddagger}{2w_{j_k^\ddagger}} \qquad w'_{j_k^\ddagger} \coloneqq \frac{\delta_k^\ddagger}{2a_{j_k^\ddagger}} \qquad b'_{j_k^\ddagger} \coloneqq -\frac{\delta_k^\ddagger p_k^\ddagger}{2a_{j_k^\ddagger}}.$$

These perturbations to the two newly identified neurons $j_k^\dagger$ and $j_k^\ddagger$ have the following contributions to the directional derivative of $f_{\theta,a_0,b_0}(x)$: for $x \in (-\infty, p_k^\dagger)$, zero; for $x \in (p_k^\ddagger, \infty)$, subtracting $\mathrm{D}\big(\alpha^{(k)}_{\tau,\pi} x + \beta^{(k)}_{\tau,\pi}\big)[\tau', \pi']$ and adding $\mathrm{D}\big(\alpha^{(k+1)}_{\tau,\pi} x + \beta^{(k+1)}_{\tau,\pi}\big)[\tau', \pi']$; and for $x \in (p_k^\dagger, p_k^\ddagger)$, the affine interpolation between the former and latter. Moreover, our definitions give

$$w'_{j_k^\dagger} p_k^\dagger + b'_{j_k^\dagger} = 0 \quad \text{and} \quad w'_{j_k^\ddagger} p_k^\ddagger + b'_{j_k^\ddagger} = 0,$$

so the two spatial one-sided limits of the output directional derivative agree at each of the two selected kinks. Hence, since there is no training input in $(p_k^\dagger, p_k^\ddagger)$ because f_{θ,a_0,b_0} is interval turning, and by applying Lemma 4 if either of the two selected kinks is a training input, it follows that the inductive hypothesis eq. (3) is satisfied for $k+1$ and all $i \in [n]$.

SUBCASE $p_k^\dagger = p_k^\ddagger$. Here f_{θ,a_0,b_0} has only this one kink in the k-th intermediate same-label segment, which means that it equals $g_{\tau,\pi}$ everywhere on that segment. By f_{θ,a_0,b_0} having a positive margin and being convexity correct, and by Lemmas 3 and 13 and Proposition 4, there exists $j_k^\dagger \in [m]$ for which we may assume:

- $a_{j_k^\dagger} > 0$ (resp., $a_{j_k^\dagger} < 0$) if the k-th intermediate dataset segment has label -1 (resp., 1);
- $w_{j_k^\dagger} = \left|a_{j_k^\dagger}\right|$;
- $-b_{j_k^\dagger}/w_{j_k^\dagger} = p_k^\dagger$.

Now let $\theta'_{(k)}$ be obtained from $\theta'_{(k-1)}$ by setting

$$a'_{j_k^\dagger} := \frac{\mathrm{D}\big(\alpha_{\tau,\pi}^{(k+1)} - \alpha_{\tau,\pi}^{(k)}\big)[\tau',\pi']}{2w_{j_k^\dagger}}$$

$$w'_{j_k^\dagger} := \frac{\mathrm{D}\big(\alpha_{\tau,\pi}^{(k+1)} - \alpha_{\tau,\pi}^{(k)}\big)[\tau',\pi']}{2a_{j_k^\dagger}}$$

$$b'_{j_k^\dagger} := \frac{\mathrm{D}\big(\alpha_{\tau,\pi}^{(k+1)} - \alpha_{\tau,\pi}^{(k)}\big)[\tau',\pi']\, p_k^\dagger}{2a_{j_k^\dagger}} + \frac{\mathrm{D}\big(\beta_{\tau,\pi}^{(k+1)} - \beta_{\tau,\pi}^{(k)}\big)[\tau',\pi']}{a_{j_k^\dagger}}.$$

This perturbation to the single newly identified neuron $j_k^\dagger$ does not change the directional derivative of $f_{\theta,a_0,b_0}(x)$ for $x < p_k^\dagger$, whereas for $x > p_k^\dagger$ it subtracts $\mathrm{D}\big(\alpha_{\tau,\pi}^{(k)} x + \beta_{\tau,\pi}^{(k)}\big)[\tau',\pi']$ and adds $\mathrm{D}\big(\alpha_{\tau,\pi}^{(k+1)} x + \beta_{\tau,\pi}^{(k+1)}\big)[\tau',\pi']$. Hence, and by Lemma 4, if $p_k^\dagger$ equals a training input x_i, then $\mathrm{D}^\circ\big(-y_i\, f_{\theta,a_0,b_0}(x_i)\big)[\theta'_{(k)}, a'_0, b'_0]$ and $\mathrm{D}\big(-y_i\, g_{\tau,\pi}(x_i)\big)[\tau',\pi']$ both equal

$$\max\Big\{\mathrm{D}\Big(-y_i\big(\alpha_{\tau,\pi}^{(k)} x_i + \beta_{\tau,\pi}^{(k)}\big)\Big)[\tau',\pi'], \mathrm{D}\Big(-y_i\big(\alpha_{\tau,\pi}^{(k+1)} x_i + \beta_{\tau,\pi}^{(k+1)}\big)\Big)[\tau',\pi']\Big\},$$

and so the inductive hypothesis eq. (3) is satisfied for $k+1$ and all $i \in [n]$.

Having completed the two subcases and the induction, letting $\theta' := \theta'_{(r-1)}$ we have that

$$\mathrm{D}^\circ_{\theta,a_0,b_0}\big(-y_i\, f_{\theta,a_0,b_0}(x_i)\big)[\theta', a'_0, b'_0] = \mathrm{D}_{\tau,\pi}\big(-y_i\, g_{\tau,\pi}(x_i)\big)[\tau',\pi'] \quad \text{for all } i \in [n],$$

and hence also

$$\mathrm{D}^\circ_{\theta,a_0,b_0}\ell\big(y_i\, f_{\theta,a_0,b_0}(x_i)\big)[\theta', a'_0, b'_0] = \mathrm{D}_{\tau,\pi}\ell\big(y_i\, g_{\tau,\pi}(x_i)\big)[\tau',\pi'] \quad \text{for all } i \in [n].$$

Also observe that

$$\begin{aligned}
&\mathrm{D}_{\theta,a_0,b_0}\Big(\frac{1}{2}\sum_{j=1}^m (a_j^2 + w_j^2)\Big)[\theta', a'_0, b'_0] \\
&= \sum_{k|p_k^\dagger < p_k^\ddagger} \Big(a_{j_k^\dagger} a'_{j_k^\dagger} + w_{j_k^\dagger} w'_{j_k^\dagger} + a_{j_k^\ddagger} a'_{j_k^\ddagger} + w_{j_k^\ddagger} w'_{j_k^\ddagger}\Big) + \sum_{k|p_k^\dagger = p_k^\ddagger} \Big(a_{j_k^\dagger} a'_{j_k^\dagger} + w_{j_k^\dagger} w'_{j_k^\dagger}\Big) \\
&= \sum_{k=1}^{r-1} \mathrm{sgn}\Big(a_{j_k^\dagger}\Big) \mathrm{D}_{\tau,\pi}\big(\alpha_{\tau,\pi}^{(k+1)} - \alpha_{\tau,\pi}^{(k)}\big)[\tau',\pi'] \\
&= \mathrm{D}_{\tau,\pi} \sum_{k=1}^{r-1} \left(\frac{\tau_k + \pi_k}{x_{\iota(k)+1} - x_{\iota(k)}} + \frac{\tau_{k+1} + \pi_{k+1}}{x_{\iota(k+1)+1} - x_{\iota(k+1)}}\right)[\tau',\pi']
\end{aligned}$$

$$= \mathrm{D}_{\tau,\pi}\mathrm{R}_1(g_{\tau,\pi})[\tau',\pi'].$$

Therefore, we have

$$\mathrm{D}^\circ L_1^\lambda(\theta,a_0,b_0)[\theta',a_0',b_0'] = \mathrm{D}_{\tau,\pi}\Big(\frac{1}{n}\sum_{i=1}^n \ell\big(y_i\, g_{\tau,\pi}(x_i)\big) + \lambda\,\mathrm{R}_1(g_{\tau,\pi})\Big)[\tau',\pi'] < 0,$$

where the latter inequality is by the strict convexity of the objective and the convexity of the feasible set of prob. (O_1) (cf. (i)). Thus θ, a_0, b_0 is not Clarke stationary, as required.

CASE $L_1^{\lambda,\mathrm{b}}$. By f_{θ,a_0,b_0} having a positive margin, and being convexity correct and single turning, and by Lemmas 3 and 10 and Proposition 4, it follows that for all $k \in [r-1]$ there exists $j_k^\dagger \in [m]$ such that, without loss of generality:

- $a_{j_k^\dagger} > 0$ (resp., $a_{j_k^\dagger} < 0$) if the k-th intermediate dataset segment has label -1 (resp., 1);
- $w_{j_k^\dagger} > 0$;
- $a_{j_k^\dagger}^2 = w_{j_k^\dagger}^2 + b_{j_k^\dagger}^2$;
- $-b_{j_k^\dagger}/w_{j_k^\dagger} = \gamma_{\tau,\pi}^{(k)}$.

Next, for each $k \in [r-1]$, we define the perturbation of neuron $j_k^\dagger$ exactly as in subcase $p_k^\dagger = p_k^\ddagger$ above, and define the skip-connection perturbation as before. We then argue like in the former case to conclude that θ, a_0, b_0 is not Clarke stationary, except that we show the equality of the directional derivatives of the unscaled regularizers as follows:

$$\begin{aligned}
&\mathrm{D}_{\theta,a_0,b_0}\Big(\frac{1}{2}\sum_{j=1}^m (a_j^2 + w_j^2 + b_j^2)\Big)[\theta',a_0',b_0'] \\
&= \sum_{k=1}^{r-1}\Big(a_{j_k^\dagger}a'_{j_k^\dagger} + w_{j_k^\dagger}w'_{j_k^\dagger} + b_{j_k^\dagger}b'_{j_k^\dagger}\Big) \\
&= \sum_{k=1}^{r-1}\Bigg[\left(\frac{a_{j_k^\dagger}}{2w_{j_k^\dagger}} + \frac{w_{j_k^\dagger}}{2a_{j_k^\dagger}} - \frac{b_{j_k^\dagger}^2}{2a_{j_k^\dagger}w_{j_k^\dagger}}\right)\mathrm{D}\big(\alpha_{\tau,\pi}^{(k+1)} - \alpha_{\tau,\pi}^{(k)}\big)[\tau',\pi'] + \frac{b_{j_k^\dagger}}{a_{j_k^\dagger}}\,\mathrm{D}\big(\beta_{\tau,\pi}^{(k+1)} - \beta_{\tau,\pi}^{(k)}\big)[\tau',\pi']\Bigg] \\
&= \sum_{k=1}^{r-1}\Bigg[\frac{w_{j_k^\dagger}}{a_{j_k^\dagger}}\,\mathrm{D}\big(\alpha_{\tau,\pi}^{(k+1)} - \alpha_{\tau,\pi}^{(k)}\big)[\tau',\pi'] + \frac{b_{j_k^\dagger}}{a_{j_k^\dagger}}\,\mathrm{D}\big(\beta_{\tau,\pi}^{(k+1)} - \beta_{\tau,\pi}^{(k)}\big)[\tau',\pi']\Bigg] \\
&= \sum_{k=1}^{r-1}\frac{\mathrm{D}\big(\alpha_{\tau,\pi}^{(k+1)} - \alpha_{\tau,\pi}^{(k)}\big)[\tau',\pi'] - \gamma_{\tau,\pi}^{(k)}\,\mathrm{D}\big(\beta_{\tau,\pi}^{(k+1)} - \beta_{\tau,\pi}^{(k)}\big)[\tau',\pi']}{\mathrm{sgn}\Big(a_{j_k^\dagger}\Big)\sqrt{1+\big(\gamma_{\tau,\pi}^{(k)}\big)^2}} \\
&= \sum_{k=1}^{r-1}\frac{\big(\alpha_{\tau,\pi}^{(k+1)} - \alpha_{\tau,\pi}^{(k)}\big)\mathrm{D}\big(\alpha_{\tau,\pi}^{(k+1)} - \alpha_{\tau,\pi}^{(k)}\big)[\tau',\pi'] + \big(\beta_{\tau,\pi}^{(k+1)} - \beta_{\tau,\pi}^{(k)}\big)\mathrm{D}\big(\beta_{\tau,\pi}^{(k+1)} - \beta_{\tau,\pi}^{(k)}\big)[\tau',\pi']}{\sqrt{\big(\alpha_{\tau,\pi}^{(k+1)} - \alpha_{\tau,\pi}^{(k)}\big)^2 + \big(\beta_{\tau,\pi}^{(k+1)} - \beta_{\tau,\pi}^{(k)}\big)^2}} \\
&= \mathrm{D}\left(\sum_{k=1}^{r-1}\sqrt{\big(\alpha_{\tau,\pi}^{(k+1)} - \alpha_{\tau,\pi}^{(k)}\big)^2 + \big(\beta_{\tau,\pi}^{(k+1)} - \beta_{\tau,\pi}^{(k)}\big)^2}\right)[\tau',\pi'] \\
&= \mathrm{D}_{\tau,\pi}\mathrm{R}_1^{\mathrm{b}}(g_{\tau,\pi})[\tau',\pi'].
\end{aligned}$$

□

Proof. [Proof of Theorem 6] For part (i) (resp., part (ii)), suppose f_{θ,a_0,b_0} is a minimizer of L_1^λ (resp., $L_1^{\lambda,\mathrm{b}}$) in function space. Then θ, a_0, b_0 is a Clarke stationary point, and its margin is positive by Corollary 3. By Proposition 4, Lemmas 5 and 6, and Lemmas 20 and 21 (resp., Lemmas 9 and 11), f_{θ,a_0,b_0} is convexity correct and interval turning (resp., single turning), and so by Lemma 22, f_{θ,a_0,b_0} is also switch optimal.

Conversely, suppose h is switch optimal, convexity correct, and interval turning (resp., single turning), but not a minimizer of L_1^λ (resp., $L_1^{\lambda,\mathrm{b}}$) in function space. By Corollary 3 and Lemma 19, for some $m \in \mathbb{N}$ and some θ, a_0, b_0 which is a minimizer of L_1^λ (resp., $L_1^{\lambda,\mathrm{b}}$) restricted to network width m and whose margin is positive, we have that $L_1^\lambda(\theta, a_0, b_0) < \frac{1}{n}\sum_{i=1}^n \ell\big(y_i\, h(x_i)\big) + \lambda\, \mathrm{R}_1(h)$ (resp., $L_1^{\lambda,\mathrm{b}}(\theta, a_0, b_0) < \frac{1}{n}\sum_{i=1}^n \ell\big(y_i\, h(x_i)\big) + \lambda\, \mathrm{R}_1^{\mathrm{b}}(h)$). Reasoning like above, f_{θ,a_0,b_0} is switch optimal, convexity correct, and interval turning (resp., single turning). But then we infer that f_{θ,a_0,b_0} and h have equal values on every input x_i, and have equal R_1 (resp., $\mathrm{R}_1^{\mathrm{b}}$) norms, which is a contradiction.

Since switch optimality and convexity correctness uniquely fix all crossing pieces, a straightforward positive-margin adaptation of the proof of Lemma 17 shows that the single-turning minimizer is unique in function space.

The remainders of parts (i) and (ii), which are for the losses L^λ and $L^{\lambda,\mathrm{b}}$ (respectively), now follow by Lemma 1.

For part (iii), suppose $(\theta, a_0, b_0) \in \mathbb{R}^{3m+2}$ is a Clarke stationary point of L_1^λ (resp., $L_1^{\lambda,\mathrm{b}}$) and its margin is positive. Reasoning like above, f_{θ,a_0,b_0} is switch optimal, convexity correct, and interval turning (resp., single turning). Moreover, by Theorem 4 and Lemmas 13 and 14 (resp., Lemmas 10, 14 and 15), we have that $\frac{1}{2}\sum_{j=1}^m (a_j^2 + w_j^2) = \mathrm{R}_1(f_{\theta,a_0,b_0})$ (resp., $\frac{1}{2}\sum_{j=1}^m (a_j^2 + w_j^2 + b_j^2) = \mathrm{R}_1^{\mathrm{b}}(f_{\theta,a_0,b_0})$), so by part (i) (resp., (ii)), we conclude that θ, a_0, b_0 is a minimizer. □

H PROOFS FOR LIMITS OF NORMALIZED APPROXIMATE CLARKE STATIONARY POINTS

Here we state and prove our final main theoretical result which establishes that, for each of the four variants of the ℓ_2-regularized logistic loss we consider, limits of margin-normalized approximate Clarke stationary points as the regularization strength λ tends to 0 are KKT points of the corresponding interpolator norm minimization problem, provided the approximation tightness ζ tends to 0 sufficiently fast. It also shows that, when the biases are not penalized and the skip connection is present, and again provided ζ tends to 0 sufficiently fast, the limits are necessarily interval turning in function space; therefore, recalling Theorem 1, most KKT points cannot be obtained in this way. Before stating and proving the theorem, we define margin normalization, show a simple proposition about it, and define approximate Clarke stationary points.

Provided the margin involved is positive, let

$$\widetilde{\theta} \coloneqq \frac{\theta}{\sqrt{M(\theta)}} \qquad \widetilde{\theta, a_0, b_0} \coloneqq \frac{\theta}{\sqrt{M(\theta, a_0, b_0)}}, \frac{a_0}{M(\theta, a_0, b_0)}, \frac{b_0}{M(\theta, a_0, b_0)}$$

denote scalings of the parameters that normalize the margin.

Proposition 5. *If $M(\theta, a_0, b_0) > 0$ then:*

$$f_{\widetilde{\theta, a_0, b_0}}(x) = \frac{f_{\theta, a_0, b_0}(x)}{M(\theta, a_0, b_0)} \quad \textit{for all } x \in \mathbb{R}, \qquad \textit{and} \qquad M\big(\widetilde{\theta, a_0, b_0}\big) = 1.$$

Proof. This is straightforward to check using the definition of the network function in eq. (1). □

For locally Lipschitz $F\colon \mathbb{R}^N \to \mathbb{R}$ and $\zeta > 0$, we say that $v \in \mathbb{R}^N$ is a ζ-stationary point of F if and only if $\zeta \mathbb{B}_2^N \cap \partial F(v) \neq \emptyset$.

Theorem 7. *Fix a network width m, and suppose $\lambda_k, \zeta_k > 0$ for all $k \in \mathbb{N}$ and $\lambda_k \to 0$ as $k \to \infty$.*

(i) If $M(\theta^{(k)}) > 0$ and $\theta^{(k)}$ is a ζ_k-stationary point of L^{λ_k} (resp., $L^{\lambda_k, \mathrm{b}}$) for all $k \in \mathbb{N}$ and $\zeta_k / \big(\lambda_k \sqrt{M(\theta^{(k)})}\big) \to 0$ as $k \to \infty$, then every limit point of $\big\{\widetilde{\theta^{(k)}} \;\big|\; k \in \mathbb{N}\big\}$ is a KKT point of prob. (P) (resp., prob. (P^{b})).

(ii) If $M\big(\theta^{(k)}, a_0^{(k)}, b_0^{(k)}\big) \geq 1$ and $\big(\theta^{(k)}, a_0^{(k)}, b_0^{(k)}\big)$ is a ζ_k-stationary point of $L_1^{\lambda_k}$ (resp., $L_1^{\lambda_k, \mathrm{b}}$) for all $k \in \mathbb{N}$ and $\zeta_k / \lambda_k \to 0$ as $k \to \infty$, then every limit point of $\Big\{ \widetilde{\big(\theta^{(k)}, a_0^{(k)}, b_0^{(k)}\big)} \;\Big|\; k \in \mathbb{N} \Big\}$ is a KKT point of prob. (P_1) (resp., prob. ($\mathrm{P}_1^{\mathrm{b}}$)).

(iii) In part (ii) when biases are not penalized, i.e., for loss $L_1^{\lambda_k}$ and prob. (P_1), if moreover $\zeta_k / \Big(\lambda_k^F \sqrt{M\big(\theta^{(k)}, a_0^{(k)}, b_0^{(k)}\big)}\Big) \to 0$ as $k \to \infty$ where $F > 1$ depends only on the dataset,[6] then every limit point of $\Big\{ \widetilde{\big(\theta^{(k)}, a_0^{(k)}, b_0^{(k)}\big)} \;\Big|\; k \in \mathbb{N} \Big\}$ is interval turning in function space.

Proof. For parts (i) and (ii), we show the claim for prob. ($\mathrm{P}_1^{\mathrm{b}}$), the cases of probs. ($\mathrm{P}$), ($\mathrm{P}_1$) and ($\mathrm{P}^{\mathrm{b}}$) can be handled analogously.

By considering a convergent subsequence, we may assume that $\widetilde{\big(\theta^{(k)}, a_0^{(k)}, b_0^{(k)}\big)} \to (\theta^\star, a_0^\star, b_0^\star)$ as $k \to \infty$. Since each $\widetilde{\theta^{(k)}, a_0^{(k)}, b_0^{(k)}}$ is a feasible point of prob. ($\mathrm{P}_1^{\mathrm{b}}$) by Proposition 5 and the feasible set is closed, we have that $\theta^\star, a_0^\star, b_0^\star$ is feasible and thus satisfies MFCQ by Lemma 16.

Since each $\theta^{(k)}, a_0^{(k)}, b_0^{(k)}$ is a ζ_k-stationary point of $L_1^{\lambda_k, \mathrm{b}}$, we have

$$\emptyset \neq \zeta_k \mathbb{B}_2^{3m+2} \cap \partial L_1^{\lambda_k, \mathrm{b}}\big(\theta^{(k)}, a_0^{(k)}, b_0^{(k)}\big)$$

[6] It suffices that, for each $k \in [r-1]$, we have that $F > 1 + y_{\iota(k)}(x_{\iota(k+1)} - x_{\iota(k)+1}) s_{k-1} s_k / (s_{k-1} - s_k)$, where $s_{k-1} \coloneqq 2y_{\iota(k)+1} / (x_{\iota(k)+1} - x_{\iota(k)})$.

$$
\begin{aligned}
&= \zeta_k \mathbb{B}_2^{3m+2} \cap \Big(\partial L\big(\theta^{(k)}, a_0^{(k)}, b_0^{(k)}\big) + \frac{\lambda_k}{2} \partial \|\theta^{(k)}\|_2^2 \Big) \\
&= \zeta_k \mathbb{B}_2^{3m+2} \cap \Bigg(\Big\{ \Big(u, \nabla_{a_0} L\big(\theta^{(k)}, a_0^{(k)}, b_0^{(k)}\big), \nabla_{b_0} L\big(\theta^{(k)}, a_0^{(k)}, b_0^{(k)}\big) \Big) \\
&\qquad\qquad \Big| \; u \in \partial_\theta L\big(\theta^{(k)}, a_0^{(k)}, b_0^{(k)}\big) \Big\} + \frac{\lambda_k}{2} \big\{ \big(\nabla_\theta \|\theta^{(k)}\|_2^2, 0, 0 \big) \big\} \Bigg) \\
&= \zeta_k \mathbb{B}_2^{3m+2} \cap \Big\{ \Big(u + \frac{\lambda_k}{2} \nabla_\theta \|\theta^{(k)}\|_2^2, \nabla_{a_0} L\big(\theta^{(k)}, a_0^{(k)}, b_0^{(k)}\big), \nabla_{b_0} L\big(\theta^{(k)}, a_0^{(k)}, b_0^{(k)}\big) \Big) \\
&\qquad\qquad \Big| \; u \in \partial_\theta L(\theta^{(k)}, a_0^{(k)}, b_0^{(k)}) \Big\},
\end{aligned}
$$

so for some decomposition $\zeta_k^2 = (\zeta_k')^2 + (\zeta_k'')^2$ we have

$$
\begin{aligned}
\zeta_k' \mathbb{B}_2^2 \ni \nabla_{a_0, b_0} L\big(\theta^{(k)}, a_0^{(k)}, b_0^{(k)}\big) &= \frac{1}{n} \sum_{i=1}^n \nabla_{a_0, b_0} \ell\big(y_i \, f_{\theta^{(k)}, a_0^{(k)}, b_0^{(k)}}(x_i)\big) \\
&= -\frac{1}{n} \sum_{i=1}^n \frac{\nabla_{a_0, b_0} \big(y_i \, f_{\theta^{(k)}, a_0^{(k)}, b_0^{(k)}}(x_i)\big)}{1 + \exp\big(y_i \, f_{\theta^{(k)}, a_0^{(k)}, b_0^{(k)}}(x_i)\big)} \\
&= -\frac{1}{n} \sum_{i=1}^n \frac{\nabla_{a_0, b_0} \Big(y_i \, f_{\widetilde{\theta^{(k)}, a_0^{(k)}, b_0^{(k)}}}(x_i)\Big)}{1 + \exp\big(y_i \, f_{\theta^{(k)}, a_0^{(k)}, b_0^{(k)}}(x_i)\big)}
\end{aligned} \tag{4}
$$

and also

$$
\begin{aligned}
\emptyset \neq \zeta_k'' \mathbb{B}_2^{3m} \cap \Big(\partial_\theta L\big(\theta^{(k)}, a_0^{(k)}, b_0^{(k)}\big) + \frac{\lambda_k}{2} \{\nabla_\theta \|\theta^{(k)}\|_2^2\} \Big) &= \zeta_k'' \mathbb{B}_2^{3m} \cap \Big(\frac{1}{n} \sum_{i=1}^n \partial_\theta \ell\big(y_i \, f_{\theta^{(k)}, a_0^{(k)}, b_0^{(k)}}(x_i)\big) + \frac{\lambda_k}{2} \{\nabla_\theta \|\theta^{(k)}\|_2^2\} \Big) \\
&= \zeta_k'' \mathbb{B}_2^{3m} \cap \Big(-\frac{1}{n} \sum_{i=1}^n \frac{\partial_\theta \big(y_i \, f_{\theta^{(k)}, a_0^{(k)}, b_0^{(k)}}(x_i)\big)}{1 + \exp\big(y_i \, f_{\theta^{(k)}, a_0^{(k)}, b_0^{(k)}}(x_i)\big)} + \frac{\lambda_k}{2} \{\nabla_\theta \|\theta^{(k)}\|_2^2\} \Big) \\
&= \zeta_k'' \mathbb{B}_2^{3m} \cap \sqrt{M\big(\theta^{(k)}, a_0^{(k)}, b_0^{(k)}\big)} \Bigg(-\frac{1}{n} \sum_{i=1}^n \frac{\partial_\theta \Big(y_i \, f_{\widetilde{\theta^{(k)}, a_0^{(k)}, b_0^{(k)}}}(x_i)\Big)}{1 + \exp\big(y_i \, f_{\theta^{(k)}, a_0^{(k)}, b_0^{(k)}}(x_i)\big)} \\
&\qquad\qquad + \frac{\lambda_k}{2} \Bigg\{ \nabla_\theta \Bigg\| \frac{\theta^{(k)}}{\sqrt{M\big(\theta^{(k)}, a_0^{(k)}, b_0^{(k)}\big)}} \Bigg\|_2^2 \Bigg\} \Bigg),
\end{aligned}
$$

which once we recall that $M\big(\theta^{(k)}, a_0^{(k)}, b_0^{(k)}\big) \geq 1$ gives us

$$
\emptyset \neq \zeta_k'' \mathbb{B}_2^{3m} \cap \Bigg(-\frac{1}{n} \sum_{i=1}^n \frac{\partial_\theta \Big(y_i \, f_{\widetilde{\theta^{(k)}, a_0^{(k)}, b_0^{(k)}}}(x_i)\Big)}{1 + \exp\big(y_i \, f_{\theta^{(k)}, a_0^{(k)}, b_0^{(k)}}(x_i)\big)} + \frac{\lambda_k}{2} \Bigg\{ \nabla_\theta \Bigg\| \frac{\theta^{(k)}}{\sqrt{M\big(\theta^{(k)}, a_0^{(k)}, b_0^{(k)}\big)}} \Bigg\|_2^2 \Bigg\} \Bigg). \tag{5}
$$

Now putting eqs. (4) and (5) together, and defining

$$
\varepsilon_k \coloneqq \frac{\zeta_k}{\lambda_k} \qquad \mu_i^{(k)} \coloneqq \frac{1}{n \lambda_k \Big(1 + \exp\big(y_i \, f_{\theta^{(k)}, a_0^{(k)}, b_0^{(k)}}(x_i)\big)\Big)} \quad \text{for all } i \in [n],
$$

we obtain

$$
\emptyset \neq \varepsilon_k \mathbb{B}_2^{3m+2} \cap \Bigg(\frac{1}{2} \Bigg\{ \nabla \Bigg\| \frac{\theta^{(k)}}{\sqrt{M\big(\theta^{(k)}, a_0^{(k)}, b_0^{(k)}\big)}} \Bigg\|_2^2 \Bigg\} - \sum_{i=1}^n \mu_i^{(k)} \partial \Big(y_i \, f_{\widetilde{\theta^{(k)}, a_0^{(k)}, b_0^{(k)}}}(x_i)\Big) \Bigg), \tag{6}
$$

so moreover defining

$$\delta_k \coloneqq \sum_{i=1}^{n} \mu_i^{(k)} \Big(y_i \, f_{\widetilde{\theta^{(k)}, a_0^{(k)}, b_0^{(k)}}}(x_i) - 1 \Big),$$

we have that $\widetilde{\theta^{(k)}, a_0^{(k)}, b_0^{(k)}}$ is an ε_k, δ_k-KKT point of prob. $(\mathrm{P}_1^{\mathrm{b}})$.[7]

Recalling that $\varepsilon_k \to 0$ as $k \to \infty$, to conclude by Theorem 3 that $\theta^\star, a_0^\star, b_0^\star$ is a KKT point of prob. $(\mathrm{P}_1^{\mathrm{b}})$, it remains to show that $\delta_k \to 0$ as $k \to \infty$.

For each $k \in \mathbb{N}$, reasoning like in the proof of Lemma 16, from eq. (6) and by Proposition 5 we have

$$\begin{aligned}
\sum_{i=1}^{n} \mu_i^{(k)} &\leq \sum_{i=1}^{n} \mu_i^{(k)} y_i \, f_{\widetilde{\theta^{(k)}, a_0^{(k)}, b_0^{(k)}}}(x_i) \\
&\leq \frac{1}{2} \left\| \frac{\theta^{(k)}}{\sqrt{M\big(\theta^{(k)}, a_0^{(k)}, b_0^{(k)}\big)}} \right\|_2^2 \\
&\quad + \varepsilon_k \left\| \frac{\theta^{(k)}}{2\sqrt{M\big(\theta^{(k)}, a_0^{(k)}, b_0^{(k)}\big)}}, \frac{a_0^{(k)}}{M\big(\theta^{(k)}, a_0^{(k)}, b_0^{(k)}\big)}, \frac{b_0^{(k)}}{M\big(\theta^{(k)}, a_0^{(k)}, b_0^{(k)}\big)} \right\|_2,
\end{aligned}$$

so since $\big(\widetilde{\theta^{(k)}, a_0^{(k)}, b_0^{(k)}}\big) \to (\theta^\star, a_0^\star, b_0^\star)$ as $k \to \infty$, we have that $\sup_{k \in \mathbb{N}} \mu_i^{(k)} < \infty$ for each $i \in [n]$. Hence, since also $\lambda_k \to 0$ as $k \to \infty$, we infer that $M\big(\theta^{(k)}, a_0^{(k)}, b_0^{(k)}\big) \to \infty$ as $k \to \infty$.

Now for all $i, i' \in [n]$ such that $y_i \, f_{\theta^\star, a_0^\star, b_0^\star}(x_i) < y_{i'} \, f_{\theta^\star, a_0^\star, b_0^\star}(x_{i'})$, and all $k \in \mathbb{N}$, we have

$$\begin{aligned}
\frac{\mu_{i'}^{(k)}}{\mu_i^{(k)}} &= \frac{1 + \exp\big(y_i \, f_{\theta^{(k)}, a_0^{(k)}, b_0^{(k)}}(x_i)\big)}{1 + \exp\big(y_{i'} \, f_{\theta^{(k)}, a_0^{(k)}, b_0^{(k)}}(x_{i'})\big)} \\
&< 2 \exp\big(y_i \, f_{\theta^{(k)}, a_0^{(k)}, b_0^{(k)}}(x_i) - y_{i'} \, f_{\theta^{(k)}, a_0^{(k)}, b_0^{(k)}}(x_{i'})\big) \\
&= 2 \exp\Big(\Big(y_i \, f_{\widetilde{\theta^{(k)}, a_0^{(k)}, b_0^{(k)}}}(x_i) - y_{i'} \, f_{\widetilde{\theta^{(k)}, a_0^{(k)}, b_0^{(k)}}}(x_{i'}) \Big) M\big(\theta^{(k)}, a_0^{(k)}, b_0^{(k)}\big) \Big),
\end{aligned}$$

and so $\mu_{i'}^{(k)} \to 0$ as $k \to \infty$.

Therefore, for all $i \in [n]$, we have either $y_i \, f_{\theta^\star, a_0^\star, b_0^\star}(x_i) = 1$ in which case $y_i \, f_{\widetilde{\theta^{(k)}, a_0^{(k)}, b_0^{(k)}}}(x_i) \to 1$ as $k \to \infty$ (and there is necessarily at least one such i), or $y_i \, f_{\theta^\star, a_0^\star, b_0^\star}(x_i) > 1$ in which case $\mu_i^{(k)} \to 0$ as $k \to \infty$. Hence $\delta_k \to 0$ as $k \to \infty$, as required.

For part (iii), suppose for a contradiction that $f_{\theta^\star, a_0^\star, b_0^\star}$ is not interval turning. Since $\theta^\star, a_0^\star, b_0^\star$ is a KKT point of prob. (P_1), we have by Theorem 1 (iii) and (i) that $f_{\theta^\star, a_0^\star, b_0^\star}$ is convexity correct. Hence, by continuity and reasoning as in the proofs of Lemmas 20 and 21, there exists $\eta > 0$ such that, for all sufficiently large k, there exists a direction $(\theta', a_0', b_0') \in \mathbb{B}_2^{3m+2}$ such that:

- $\mathrm{D}\Big(\frac{1}{2} \sum_{j=1}^{m} \big((a_j^{(k)})^2 + (w_j^{(k)})^2 \big) \Big) [\theta', a_0', b_0'] \leq 0$;
- $\mathrm{D}^\circ \big(-y_i \, f_{\theta^{(k)}, a_0^{(k)}, b_0^{(k)}}(x_i) \big) [\theta', a_0', b_0'] \leq 0$ for all $i \in [n]$;
- $\mathrm{D}\big(-y_{i^\dagger} \, f_{\theta^{(k)}, a_0^{(k)}, b_0^{(k)}}(x_{i^\dagger}) \big) [\theta', a_0', b_0'] \leq -\eta \sqrt{M\big(\theta^{(k)}, a_0^{(k)}, b_0^{(k)}\big)}$ for some $i^\dagger \in [n]$.

Then, we have

$$-\frac{\zeta_k}{\lambda_k^F \sqrt{M\big(\theta^{(k)}, a_0^{(k)}, b_0^{(k)}\big)}} \leq \frac{\mathrm{D}^\circ L_1^{\lambda_k}\big(\theta^{(k)}, a_0^{(k)}, b_0^{(k)}\big)[\theta', a_0', b_0']}{\lambda_k^F \sqrt{M\big(\theta^{(k)}, a_0^{(k)}, b_0^{(k)}\big)}}$$

[7] For the definition of approximate KKT points, we refer the reader to appendix B.

$$\leq -\frac{\eta}{n\lambda_k^F\Big(1+\exp\big(y_{i^\dagger}\, f_{\theta^{(k)},a_0^{(k)},b_0^{(k)}}(x_{i^\dagger})\big)\Big)},$$

and so $\lambda_k^F\Big(1+\exp\big(y_{i^\dagger}\, f_{\theta^{(k)},a_0^{(k)},b_0^{(k)}}(x_{i^\dagger})\big)\Big) \to \infty$ as $k \to \infty$.

However, for each $i \in [n]$ with $y_i\, f_{\theta^\star,a_0^\star,b_0^\star}(x_i) = 1$, we have

$$\begin{aligned}
&\big(n\mu_i^{(k)}\big)^F \lambda_k^F\Big(1+\exp\big(y_{i^\dagger}\, f_{\theta^{(k)},a_0^{(k)},b_0^{(k)}}(x_{i^\dagger})\big)\Big) \\
&= \frac{1+\exp\big(y_{i^\dagger}\, f_{\theta^{(k)},a_0^{(k)},b_0^{(k)}}(x_{i^\dagger})\big)}{\Big(1+\exp\big(y_i\, f_{\theta^{(k)},a_0^{(k)},b_0^{(k)}}(x_i)\big)\Big)^F} \\
&< 2\frac{\exp\big(y_{i^\dagger}\, f_{\theta^{(k)},a_0^{(k)},b_0^{(k)}}(x_{i^\dagger})\big)}{\exp\big(y_i\, f_{\theta^{(k)},a_0^{(k)},b_0^{(k)}}(x_i)\big)^F} \\
&= 2\exp\Big(\Big(y_{i^\dagger}\, f_{\widetilde{\theta^{(k)},a_0^{(k)},b_0^{(k)}}}(x_{i^\dagger}) - F\, y_i\, f_{\widetilde{\theta^{(k)},a_0^{(k)},b_0^{(k)}}}(x_i)\Big) M\big(\theta^{(k)},a_0^{(k)},b_0^{(k)}\big)\Big) \\
&\to 0 \quad \text{as } k \to \infty,
\end{aligned}$$

and therefore $\mu_i^{(k)} \to 0$ as $k \to \infty$. But then $\mu_i^{(k)} \to 0$ as $k \to \infty$ for all $i \in [n]$, which is impossible due to the analogue of eq. (6) with biases not penalized. □

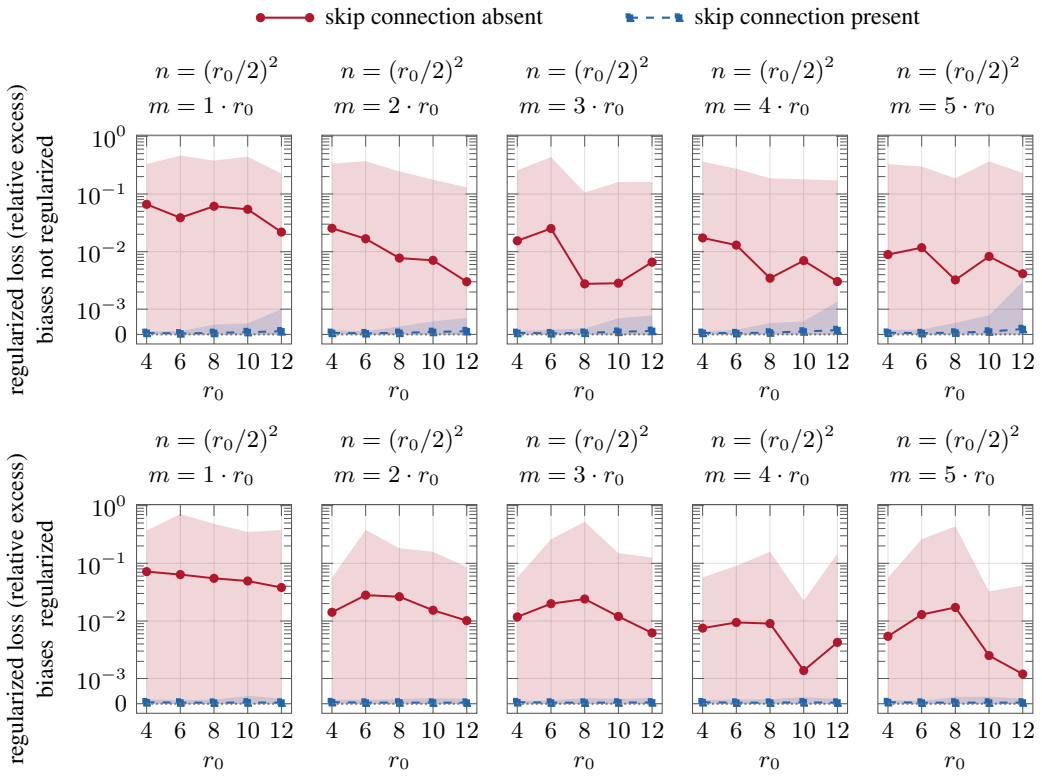


Figure 13: (This is Figure 3, reproduced here for the reader's convenience.) We plot the relative excess of the ℓ_2-regularized logistic loss L at the end of training, i.e., $L/L^\star - 1$, where $L^\star$ is the global minimum for the experiment setting, computed via the convex problem characterized in Theorem 6. Rows indicate whether biases are regularized, columns give the overparameterization factor $\omega = m/r_0$, and the two curves indicate whether the skip connection is absent or present. For each point, we use the runs, among 60 RNG seeds, that reached a positive margin: the curve gives their mean, and the shaded region spans their smallest and largest outcomes (positive-margin failure rates are reported in Figure 15). The shared vertical axis is logarithmic above 10^{-3} and linear below it. **Observations:** With a skip connection, training consistently reaches or closely approaches the global minimum; without one, the best runs reach the minimum but the means and worst outcomes typically remain substantially above it. This is consistent with our theoretical result that a skip connection rules out suboptimal positive-margin stationary points, and with the counterexamples showing that such points can occur without it.

I Further details for and outcomes from the experiments

Datasets. We first sample a data distribution, which has a hyperparameter $r_0 \geq 3$, which is the number of class breaks and which we vary in the experiments. Namely, we sample class breaks $(\beth_k)_{k=1}^{r_0}$ from the uniform distribution on the interval $(-1, 1)$, and index them so that they are increasing, i.e., $\beth_1 < \cdots < \beth_{r_0}$. Also we write $\beth_0 := -1$ and $\beth_{r_0+1} := 1$. We then sample the training inputs $(x_i)_{i=1}^n$ from the uniform distribution on the interval $(-1, 1)$, and assign the labels according to the class breaks, i.e., for each $i \in [n]$, we set $y_i := (-1)^k$, where $x_i \in [\beth_k, \beth_{k+1})$.

Since the training inputs are random, the number r of label switches in the dataset is at most r_0 and possibly smaller. We set the dataset size as $n := (r_0/2)^2$, which ensures that, in expectation, r is only a constant below the hyperparameter r_0.

We use rejection sampling to ensure that all training inputs in the first (resp., last) two same-label segments are negative (resp., positive), i.e., that Assumption 1 is satisfied; also to ensure that the minimal value of $\beth_{k+1} - \beth_k$ is at least $1/(r_0)^2$, so that no class is very unlikely to be populated by a training input; and to ensure that the minimal value of $x_{i+1} - x_i$ when $y_{i+1} \neq y_i$ (i.e., at label

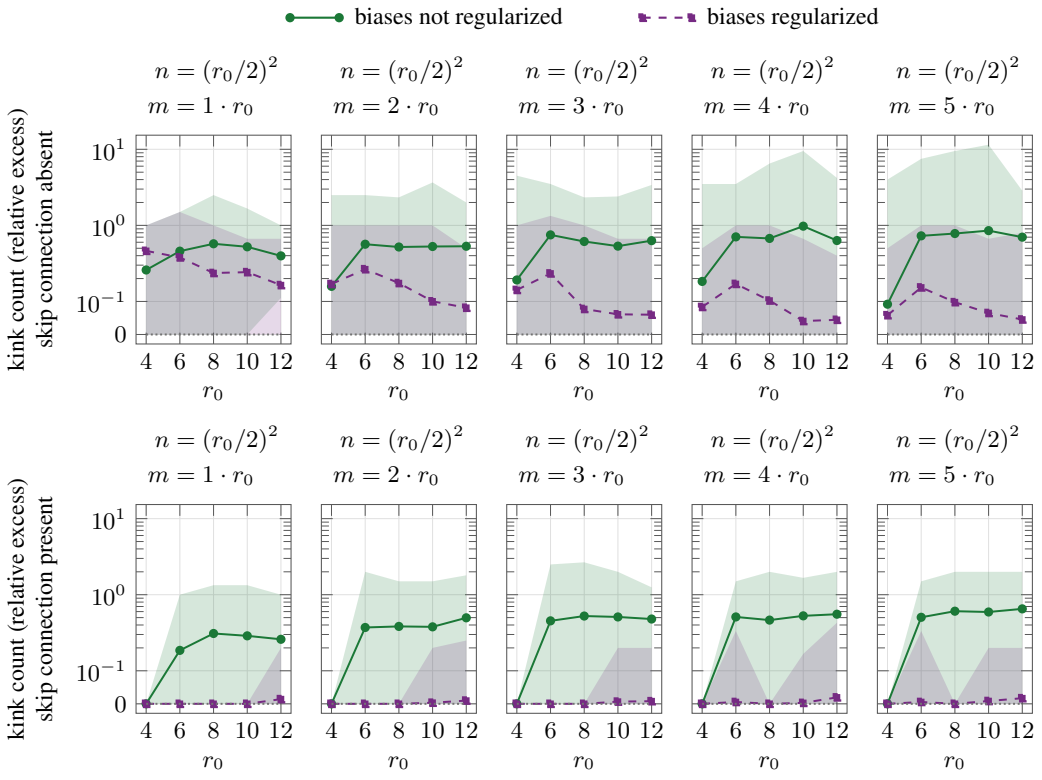


Figure 14: We plot the relative excess number of distinct kinks of the network at the end of training, i.e., $Q/(r-1)-1$, where Q is the number of distinct kinks and r is the number of label switches in the training dataset (and so $r-1$ is the minimal number of distinct kinks of any positive-margin classifier). Rows indicate whether the skip connection is absent or present, columns give the overparameterization factor $\omega = m/r_0$, and the two curves indicate whether biases are regularized. For each point, we use the runs, among 60 RNG seeds, that reached a positive margin: the curve gives their mean, and the shaded region spans their smallest and largest outcomes (positive-margin failure rates are reported in Figure 15). The shared vertical axis is logarithmic above 10^{-1} and linear below it. **Observations:** When biases are regularized, training consistently reaches or closely approaches the minimal kink count, especially with a skip connection; without bias regularization, the final networks typically have more kinks, particularly when the skip connection is absent. This is consistent with our theoretical result that bias regularization makes every (global) minimizer sparsest, and that a skip connection rules out suboptimal positive-margin stationary points, as well as with the counterexample showing that nonsparse suboptimal KKT points can occur without a skip connection.

switches) is at least $2/n\sqrt{r_0}$, so that reaching a positive margin on the dataset will not require very large network parameters.[8]

Networks. We set the network width as $m := \omega \cdot r_0$, where the overparameterization factor ω is a hyperparameter which we vary in the experiments.

Since the focus of this work is norm minimization of interpolating networks, we initialize the network in a way that helps to reach a positive margin early in the training. Namely, following Safran et al. (2022), we initialize the output weights small and the hidden weights large: we set $d_{\min}$ to be the minimal distance between two training inputs with different labels, set ς to be a constant times $1/d_{\min}$, sample the output weights $(a_j)_{j=1}^m$ from the centered normal distribution with variance $1/\varsigma^2$, and sample the hidden weights $(w_j)_{j=1}^m$ from the centered normal distribution with variance ς^2. Then we initialize the hidden biases $(b_j)_{j=1}^m$ so that, for each $j \in [m]$, the kink $-b_j/w_j$ is equally likely to be

[8]These two lower bounds are multiples of the corresponding asymptotic expectations, by small constants.

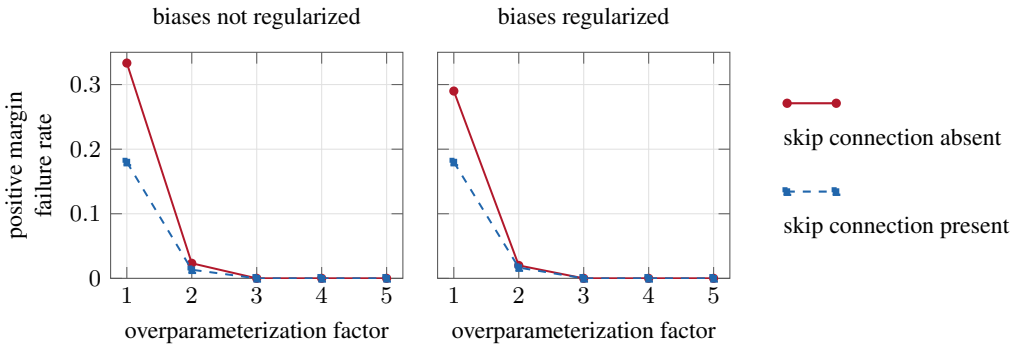


Figure 15: We plot the fraction of runs that failed to reach a positive margin on their training dataset, aggregated over 300 runs for each setting, comprising 60 RNG seeds for each of the five numbers of class breaks $r_0 \in \{4, 6, 8, 10, 12\}$. The plots have the same linear vertical axis. **Observations:** The fractions of failed runs are always (if nonzero) higher when a skip connection is absent, and are significant only for the smallest overparameterization factor of 1, which is when the network width is only an additive constant greater than the minimum required for a positive margin.

in any of the $n-1$ intervals (x_i, x_{i+1}) between two consecutive training inputs, and is independently uniformly distributed on that interval.

If a skip connection is present, we initialize its weight a_0 and its bias b_0 both as zero.

Training. Depending on whether biases are regularized and whether a skip connection is present, we minimize one of the four variants L^λ, L_1^λ, $L^{\lambda,\mathrm{b}}$, and $L_1^{\lambda,\mathrm{b}}$ of the empirical logistic loss with ℓ_2 regularization of strength λ. We use small λ, namely set it to a constant times $1/m\varsigma^2$, so that before a positive margin is reached, the objective is dominated by the logistic loss term; this also ensured that, in all the experiments we ran, λ satisfied the upper bounds in Lemma 18 and Theorem 6.

We use an in-house implementation in NumPy (Harris et al., 2020) of the Adam optimizer (Kingma & Ba, 2015). We leave all of Adam's hyperparameters at their default values, except that before reaching a positive margin the learning rate is 10 times smaller, and after that $1-\beta_2$ is 10 times larger. To get close to a Clarke stationary point, we run $2 \cdot 10^7$ full-batch iterations per experiment.

Final computations. For each experiment, to compare its final value of the regularized loss with the minimum value, we recall that, by Theorem 6, the latter equals the minimum of prob. (O_1) or prob. $(\mathrm{O}_1^\mathrm{b})$ (depending on whether biases are regularized) for the same dataset and the same regularization strength λ. Recalling also Lemma 18, Problems (O_1) and $(\mathrm{O}_1^\mathrm{b})$ have strictly convex objectives on convex feasible sets, and we compute their minima using Clarabel (Goulart & Chen, 2026) (via CVXPY (Diamond & Boyd, 2016) and using the default settings therein). All runs of Clarabel involved in the experiments to produce the plots in Figure 13 returned near optimal solutions. Also we remark that applying a general-purpose convex reformulation (cf. Pilanci & Ergen, 2020) would provide an alternative computation of the same benchmark rather than an additional learning baseline.

For the sparsity plots in Figure 14, we compute the number of distinct kinks as the number of clusters of values $-b_j/w_j$. We use SciPy's (Virtanen et al., 2020) scipy.cluster.hierarchy.fclusterdata function with the "distance" criterion and the "complete" method with a threshold of 0.02 times the minimal distance $d_{\min}$ between two training inputs with different labels. We do not count clusters whose sum of slope changes $a_j|w_j|$ is in the interval $[-0.01, 0.01]$.

Hyperparameters and seeds. We run experiments for each value combination of the following four hyperparameters (the first two numerical and the last two Boolean): number of class breaks $r_0 \in \{4, 6, 8, 10, 12\}$, overparameterization factor $\omega \in \{1, 2, 3, 4, 5\}$, whether biases are regularized, and whether a skip connection is present. For each of those 100 combinations, we repeat everything (class breaks sampling, dataset sampling, network initialization, network training, loss baseline

computation, and sparsity computation) 60 times for different RNG seeds.[9] Only runs that succeed in reaching a positive margin contribute to the plots in Figures 13 and 14; the fractions of failed runs are plotted in Figure 15.

Compute. Single experiments (i.e., with all four hyperparameters fixed and for one RNG seed) took under 20min on a single core of an AMD EPYC 9555P 64-core CPU, and all 6000 experiments that we ran to produce the plots in Figures 13, 14 and 15 took around 12h using two such CPUs.

[9] Class breaks sampling and dataset sampling depend only on r_0 and the seed. Network initialization apart from the skip connection depends additionally on ω.