Asymptotic Smoothing of the Lipschitz Loss Landscape in Overparameterized One-Hidden-Layer ReLU Networks

Saveliy Baturin SAVELIY.XO@GMAIL.COM
Independent Researcher

Abstract

We study the topology of the loss landscape of one-hidden-layer ReLU networks under overparameterization. On the theory side, we (i) prove that for convex L-Lipschitz losses with an ℓ_1-regularized second layer, every pair of models at the same loss level can be connected by a continuous path within an arbitrarily small loss increase ϵ (extending a known result for the quadratic loss); (ii) obtain an asymptotic upper bound on the *energy gap* ϵ between local and global minima that vanishes as the width m grows, implying that the landscape flattens and sublevel sets become connected in the limit. Empirically, on a synthetic Moons dataset and on the Wisconsin Breast Cancer dataset, we measure pairwise energy gaps via Dynamyc String Sampling (DSS) (Freeman and Bruna, 2017) and find that wider networks exhibit smaller gaps; in particular, a permutation test on the maximum gap yields $p_{\text{perm}} = 0$, indicating a clear reduction in the barrier height.

Keywords: ReLU networks, overparameterization, loss landscape, sublevel set connectivity, geodesic paths, Lipschitz losses

1 Introduction

Overparameterized neural networks achieve remarkable performance in practice, yet understanding *why* standard optimization methods find good solutions in highly nonconvex landscapes remains an open theoretical challenge. Early foundational results established that neural networks can approximate broad function classes (Kolmogorov, 1957; Arnol'd, 1957; Vitushkin and Henkin, 1967; Gorban', 1998; Alexeev, 2010), and characterized their expressive power (for example, counting linear regions for ReLU networks) (Anthony and Bartlett, 1999; Montufar et al., 2014; Pascanu et al., 2013). Subsequent work showed that certain architectures have benign optimization landscapes: for example, deep linear networks have no bad local minima (Kawaguchi, 2016), and adding skip connections in a shallow network removes any local minima worse than those of a linear predictor (Shamir, 2018). Increasing the number of hidden units (width) can also dramatically affect the loss topology. Theoretical evidence by Freeman and Bruna (2017) showed that in one-hidden-layer ReLU networks, wider layers lead to fewer local minima and a flatter quadratic-loss landscape, with geodesic paths connecting different minima on the same sublevel sets. We formalize and extend this picture to a broader class of losses in the one-hidden-layer setting.

Contributions. Our contributions are both theoretical and practical:

1. **Sublevel connectivity and vanishing energy gap.** We prove that for a one-hidden-layer ReLU network, any two models with loss below a given level can be connected by a continuous path within a small loss increase ϵ. In particular, for quadratic loss (mean squared error) we outline the classical argument by Freeman

and Bruna (2017) that constructs such a path, and we extend the construction to any convex loss that is L-Lipschitz in the predictions (with an ℓ_1 regularization on the output layer). Moreover, we show that the "energy gap" ϵ—the worst-case loss barrier along the connecting path—can be made arbitrarily small by increasing the width m. In fact, $\epsilon = O(m^{-\zeta})$ for some $\zeta > 0$, implying that in the infinite-width limit all loss sublevel sets become connected.

2. **Experiments.** We present pairwise energy-gap experiments on a synthetic Moons dataset (regression with mean squared error) and the Wisconsin Breast Cancer dataset (Agarap, 2018) (binary classification with cross-entropy). For each pair we measure the gap $\epsilon = \max_t \left[L(\gamma(t)) - \max\{L(\theta_A), L(\theta_B)\}\right]$ along DSS paths. Wider networks exhibit smaller gaps; a permutation test on the maximum gap yields $p_{\text{perm}} = 0$, which we treat as the primary indicator of barrier reduction.

2 Preliminary

Network model. We consider a one-hidden-layer neural network with Rectified Linear Unit activation. Formally, for an input $x \in \mathbb{R}^n$ the network output is

$$\Phi(x; W^1, \theta) = \theta^\top Z(W^1 x),$$

where $W^1 \in \mathbb{R}^{m\times n}$ is the weight matrix of the hidden layer (m neurons) and $\theta \in \mathbb{R}^m$ is the vector of output layer weights. The activation $Z(\cdot)$ applies the ReLU function elementwise: $Z(t) = \max\{0, t\}$. We assume without loss of generality that each row w_i of W^1 is an ℓ_2 unit vector (using the homogeneity $Z(c\, w_i x) = c\, Z(w_i x)$ to normalize, as is standard in analysis of ReLU networks).

Loss function and regularization. Let $(X, Y) \sim P$ be a data distribution (the empirical case will be addressed in experiments). We write $\hat{Y} = \Phi(X; W^1, \theta)$ for the network output (logits). Throughout, we restrict attention to *convex Lipschitz losses of the logits.* This covers standard logistic objectives used in practice:

- For binary classification with $y \in \{\pm 1\}$ and logit $z = \hat{Y}$, the logistic loss $\ell(z, y) = \log(1 + e^{-yz})$ is convex and globally 1-Lipschitz in z under ℓ_2.
- For multiclass classification with logits $z \in \mathbb{R}^K$ and label $y \in \{1, \ldots, K\}$, the cross-entropy loss $\ell(z, y) = -\log \frac{e^{z_y}}{\sum_j e^{z_j}}$ is convex and globally $\sqrt{2}$-Lipschitz in z under ℓ_2.

The regularized expected risk is

$$F_o(W^1, \theta) = \mathbb{E}_{(X,Y)\sim P}\, \mathcal{L}\big(Y, \Phi(X; W^1, \theta)\big) + \kappa\, \|\theta\|_1,$$

with $\kappa > 0$ a regularization coefficient (we include an ℓ_1 penalty on the second-layer weights, which will be convenient in our theoretical arguments for Lipschitz losses). The empirical counterpart F_e is defined analogously as a finite-sample average. We will sometimes write θ for (W^1, θ) when no confusion arises.

Loss sublevel sets. For a given loss value λ, the sublevel set is $\Omega_F(\lambda) = \{\theta : F_o(\theta) \leq \lambda\}$. If a sublevel set $\Omega_F(\lambda)$ is connected for every λ, then clearly every local minimum of F_o must also be a global minimum (since any two minima at value λ can be connected within $\Omega_F(\lambda)$, contradicting the existence of an isolated strict local minimum at higher loss). We will show that in overparameterized networks, even though sublevel sets might not be strictly connected due to a small energy barrier, they become *arbitrarily close to connected* as width grows (Theorem 3).

Robust compressibility. For a hidden layer $\bar{W} \in \mathbb{R}^{n\times m}$:

$$\delta_{\bar{W}}(l, \alpha; m) = \min_{\substack{\|\gamma\|_0 \leq l, \\ \sup_i \angle(\hat{w}_i, w_i) \leq \alpha}} \mathbb{E}\big[\,|Y - \gamma Z(\bar{W})|^2\,\big] + \kappa\|\gamma\|_1,$$

where γ is a sparse coefficient vector ($\|\gamma\|_0 \leq l$), and $\angle(\hat{w}_i, w_i)$ denotes the angle between the original weight w_i and its perturbed version $\hat{w}_i$.

This quantity measures how well the hidden-layer representation can be compressed while retaining only l neurons and allowing small perturbations of magnitude α. It is used to analyze the connectivity of sublevel sets in the loss landscape.

Notation. We use $\langle a, b\rangle$ for the standard inner product and $\|v\|$ for the Euclidean norm. We let $\Sigma_X = \mathbb{E}[XX^\top]$ denote the second-moment (covariance) operator of the input distribution. For a vector w, we write $z(w) = Z(wX) = \max\{0,\, w^\top X\}$ to denote the random output of a ReLU neuron with weight w. Finally, $\|\Sigma_X\|$ is the spectral norm of the covariance matrix (the largest eigenvalue), and S^{n-1} denotes the unit sphere in $\mathbb{R}^n$.

3 Lipschitz Convex Loss: Connectivity and Path Construction

We extend the connectivity result to a much broader class of loss functions. Specifically, let $\mathcal{L}(Y, \hat{Y})$ be any loss that is convex in the prediction $\hat{Y}$ and L-Lipschitz continuous in $\hat{Y}$. For instance, the logistic loss and binary cross-entropy are convex in $\hat{Y}$ and Lipschitz on any compact domain (in classification tasks $\hat{Y}$ often lies in $[0, 1]$ after a sigmoid). We assume an ℓ_1 regularization $\kappa\|\theta\|_1$ on the second-layer weights, with $\kappa > 0$ (note this is usually present or can be added without hurting performance).

Before stating the main theorem, we prove a useful lemma that bounds the size of the optimal second-layer weights in terms of the Lipschitz constant L and κ. This will ensure that neurons cannot have arbitrarily large influence, which helps control the loss when we perturb or remove neurons.

Lemma 1 (Control of ℓ_1-norm)
Let $\mathcal{L}(Y, \hat{Y})$ be convex in $\hat{Y}$ and L-Lipschitz, and let $\kappa > 0$. For any fixed first-layer W^1, consider the optimization problem

$$\theta^* \in \arg\min_{\theta\in\mathbb{R}^m} \Big\{\, \mathbb{E}[\mathcal{L}(Y, \Phi(X; W^1, \theta))] \,+\, \kappa\|\theta\|_1 \,\Big\}.$$

Then the optimal solution θ^* satisfies $\|\theta^*\|_1 \leq L/\kappa$. In fact, if $\kappa \geq L$ then $\theta^* = 0$ (that is, the minimum is achieved by the zero output weights).

Proof Let $F_o(\theta) = \mathbb{E}[\mathcal{L}(Y, \Phi(X; W^1, \theta))] + \kappa\|\theta\|_1$ be the convex objective as a function of θ. Since F_o is convex and differentiable in θ almost everywhere, at the optimum θ^* we have the first-order (subgradient) optimality condition $0 \in \partial F_o(\theta^*) = \nabla_\theta \mathbb{E}[\mathcal{L}]\big|_{\theta^*} + \kappa\, \partial\|\theta^*\|_1$. This implies:

- For any coordinate i such that $\theta_i^* \neq 0$, we must have $\nabla_{\theta_i}\mathbb{E}[\mathcal{L}] = -\kappa\, \mathrm{sign}(\theta_i^*)$. In particular, $|\nabla_{\theta_i}\mathbb{E}[\mathcal{L}]| = \kappa$.
- For any coordinate i such that $\theta_i^* = 0$, we have $\nabla_{\theta_i}\mathbb{E}[\mathcal{L}] \in [-\kappa, +\kappa]$ (the subgradient of $|\theta_i|$ at 0 is any value in $[-1, 1]$).

Next, Lipschitz continuity of $\mathcal{L}$ in $\hat{Y}$ implies a bound on the gradient norms: since $|\mathcal{L}(y, \hat{y}_1) - \mathcal{L}(y, \hat{y}_2)| \leq L|\hat{y}_1 - \hat{y}_2|$, we have $\|\nabla_\theta \mathbb{E}[\mathcal{L}]\|_\infty \leq L$. In particular $|\nabla_{\theta_i}\mathbb{E}[\mathcal{L}]| \leq L$ for every i. Combining this with the first optimality condition above for coordinates where $\theta_i^* \neq 0$, we get $\kappa = |\nabla_{\theta_i}\mathbb{E}[\mathcal{L}]| \leq L$. Thus a necessary condition for θ_i^* to be nonzero is $\kappa \leq L$. Equivalently, if $\kappa > L$ then the only feasible solution is $\theta^* = 0$ (because any nonzero coefficient would violate the condition). This proves the second statement.

Finally, to bound $\|\theta^*\|_1$ when $\kappa \leq L$, we can use convexity of $\mathbb{E}[\mathcal{L}]$: for any θ,

$$\mathbb{E}[\mathcal{L}(Y, \Phi(X; W^1, \theta))] \geq \mathbb{E}[\mathcal{L}(Y, \Phi(X; W^1, \theta^*))] + \langle \nabla_\theta \mathbb{E}[\mathcal{L}]|_{\theta^*}, \theta - \theta^* \rangle\,.$$

Plugging in $\theta = 0$ and using $\nabla_\theta \mathbb{E}[\mathcal{L}]|_{\theta^*} = -\kappa\, \mathrm{sign}(\theta^*)$ (for each component where θ^* is nonzero, and bounded in $[-\kappa, \kappa]$ where θ^* is zero), we get

$$\begin{aligned}\mathbb{E}[\mathcal{L}(Y, \Phi(X; W^1, 0))] &\geq \mathbb{E}[\mathcal{L}(Y, \Phi(X; W^1, \theta^*))] - \kappa\, \langle \mathrm{sign}(\theta^*),\, \theta^* \rangle = \\ &= \mathbb{E}[\mathcal{L}(Y, \Phi(X; W^1, \theta^*))] + \kappa\, \|\theta^*\|_1\,,\end{aligned}$$

since $\langle \mathrm{sign}(\theta^*), \theta^* \rangle = \|\theta^*\|_1$. On the other hand, by L-Lipschitz continuity,

$$\mathbb{E}[\mathcal{L}(Y, \Phi(X; W^1, 0))] - \mathbb{E}[\mathcal{L}(Y, \Phi(X; W^1, \theta^*))] \leq L\, \|\theta^*\|_1\,,$$

because $\Phi(X; W^1, 0) = 0$ and $\Phi(X; W^1, \theta^*) = \sum_{i=1}^m \theta_i^*\, z(w_i^*)$ differs by at most $|\theta_i^*|$ per active neuron. Combining the two inequalities yields

$$\kappa\, \|\theta^*\|_1 \leq L\, \|\theta^*\|_1\,,$$

and since $\kappa > 0$ we can cancel $\|\theta^*\|_1$ to obtain $\|\theta^*\|_1 \leq L/\kappa$ (and this inequality trivially holds if $\theta^* = 0$ as well). ■

Lemma 1 implies that throughout the training landscape (and in particular for any local or global minimum), the sum of the absolute output weights is bounded by L/κ. In the special case $\kappa \geq L$, the only optimum is the trivial zero network—which in practice one would avoid by choosing $\kappa < L$ so that the network can actually learn.

We are now ready to state the connectivity result for Lipschitz convex losses.

Theorem 2 (Connectivity for Lipschitz convex loss)
For any convex Lipschitz loss functions $F(X, Y, \Phi(\theta))$ (hereafter denoted simply as $F(\Phi(\theta))$) appearing in the expected risk with L_1-regularization of the second layer:

$$F_o(\Phi(\{W, \theta\})) = \mathbb{E}[F(\Phi(\{W, \theta\}))] + \kappa\|\theta\|_1,$$

for any level $\lambda \in \mathbb{R}$ and parameters $\theta^A, \theta^B \in \mathcal{W}$ such that $F_o(\Phi(\theta^{\{A,B\}})) \leq \lambda$, there exists a continuous path $\gamma : [0,1] \to \mathcal{W}$ such that $\gamma(0) = \theta^A, \gamma(1) = \theta^B$ and

$$F_o(\Phi(\gamma(t))) \leq \max(\lambda, \epsilon),$$

where

$$\epsilon = \inf_{l,\alpha} \Big(\max \big(e(l),\ \delta_{W_A^1}(m,0,m),\ \delta_{W_1^A}(m-l,\alpha,m),\\ \delta_{W_1^B}(m-l,\alpha,m),\ \delta_{W_1^B}(m,0,m)\big) + C\alpha \Big)$$

and $e(l)$ – minimum of expected risk for network of l neurons.

Proof Consider arbitrary α and $l \leq m$. Using a similar analysis to Freeman and Bruna (2017), we construct a path from θ^A to θ^B as a concatenation of the following subpaths:

1. $\theta^A \to \theta_{lA}$: the best linear predictor using the same first-layer weights as θ^A;
2. $\theta_{lA} \to \theta_{sA}$: the best $(m-l)$-term approximation using perturbed atoms from θ^A;
3. $\theta_{sA} \to \theta^*$: the best l-term approximation;
4. $\theta^* \to \theta_{sB}$: the best l-term approximation;
5. $\theta_{sB} \to \theta_{lB}$: the best $(m-l)$-term approximation using perturbed atoms from θ^B;
6. $\theta_{lB} \to \theta^B$: the best linear predictor using the same first-layer weights as θ^B.

The main idea is to connect θ^A and θ^B by reducing both to a common point θ^*. Subpaths (1) and (6) involve only changes in the second-layer parameters while keeping the first-layer weights fixed, which yields convex losses. Hence, a linear path is sufficient to guarantee that the loss along it is bounded above by λ at one endpoint and by $\delta_{W_s^A}(m,0,m)$ at the other.

Paths (3) and (4) are constructed as follows: we keep $(m-l)$ active columns of the matrix $\widetilde{W}$ unchanged and replace the remaining l with the optimal l-term approximation W^*. Thus we obtain a new first layer $\bar{W} = [\widetilde{W}, W^*]$, which does not change the loss value due to the corresponding zero entries in the second-layer coefficients. Then, keeping $\bar{W}$ fixed, we vary the second-layer coefficients linearly:

$$\gamma(t) = (1-t)\gamma_s^A + t\gamma^*.$$

Since the loss function is convex with respect to the second layer, it follows that

$$F_o(t) \leq \max(\delta_{W^A}(m-l,\alpha,m), e(l)).$$

Proof for paths (2) and (5) (the nonconvex case). Let $\theta = \{W^A, \beta\}$ and $\widetilde{\theta} = \{\widetilde{W}, \gamma\}$, where β is the optimal second-layer vector after segment (1), $\widetilde{W}$ is the perturbed first-layer matrix, and γ is the optimal second-layer vector for the perturbed matrix with $(m-l)$ nonzero components.

We define an auxiliary regression problem with parameters:

$$\overline{W} = [W^A; \widetilde{W}] \in \mathbb{R}^{n \times 2m}, \quad \overline{\beta}_1 = [\beta; 0], \quad \overline{\beta}_2 = [0; \gamma].$$

Then

$$F(\Phi(\{\overline{W}, \overline{\beta_1}\})) + \kappa\|\overline{\beta}_1\|_1 = F(\Phi(\{W^A, \beta\})) + \kappa\|\beta\|_1,$$

and similarly for $\overline{\beta}_2$:

$$F(\Phi(\{\overline{W}, \overline{\beta_2}\})) + \kappa\|\overline{\beta}_2\|_1 = F(\Phi(\{\widetilde{W}, \gamma\})) + \kappa\|\gamma\|_1.$$

By convexity, the extended linear path $\eta(t) = (1-t)\overline{\beta}_1 + t\overline{\beta}_2$ satisfies:

$$\forall\, t, \quad \overline{L}(t) = \mathbb{E}[F(\Phi(\eta(t)))] + \kappa\|\eta(t)\|_1 \le \max(\overline{L}(0), \overline{L}(1)).$$

We now approximate this extended path in terms of the first- and second-layer weights:

$$\eta_1(t) = (1-t)W^A + t\widetilde{W}, \quad \eta_2(t) = (1-t)\beta + t\gamma.$$

Hence,

$$\begin{gathered}
F_o(\{\eta_1(t), \eta_2(t)\}) = \mathbb{E}[F(\Phi(\{\eta_1(t), \eta_2(t)\}))] + \kappa\|\eta_2(t)\|_1 \le \\
\le \mathbb{E}[F(\Phi(\{\eta_1(t), \eta_2(t)\}))] + \kappa((1-t)\|\beta\|_1 + t\|\gamma\|_1) = \\
= \overline{L}(t) + \mathbb{E}[F(\eta_2(t)Z(\eta_1(t)))] - \mathbb{E}[F((1-t)\beta Z(W^A) + t\gamma Z(\widetilde{W}))] \le \\
\le \max(\overline{L}(0), \overline{L}(1)) + |\mathbb{E}[F(\eta_2(t)Z(\eta_1(t)))] - \mathbb{E}[F((1-t)\beta Z(W^A) + t\gamma Z(\widetilde{W}))]|.
\end{gathered}$$

The second term can be bounded as follows:

$$\big|\mathbb{E}[F(\eta_2(t)Z(\eta_1(t)))] - \mathbb{E}[F((1-t)\beta Z(W^A) + t\gamma Z(\widetilde{W}))]\big| \le L^2\,\alpha\,\sqrt{\|\Sigma\|}\,\kappa^{-1}.$$

Indeed, since the modulus function is convex, Jensen's inequality Zdravko (2012) allows us to move the expectation outside the absolute value, and then, by the Lipschitz property of F,

$$\begin{gathered}
\mathbb{E}[|F(\eta_2(t)Z(\eta_1(t))) - F((1-t)\beta Z(W^A) + t\gamma Z(\widetilde{W}))|] \le \\
\le L\,\mathbb{E}[|\eta_2(t)Z(\eta_1(t)) - (1-t)\beta Z(W^A) - t\gamma Z(\widetilde{W})|].
\end{gathered}$$

Note that

$$\begin{gathered}
L * \mathbb{E}[|\eta_2(t)Z(\eta_1(t)) - (1-t)\beta Z(W^A) - t\gamma Z(\widetilde{W})|] = \\
= L * \mathbb{E}[|(1-t)\beta Z(\eta_1(t)) + t\gamma Z(\eta_1(t)) - (1-t)\beta Z(W^A) - t\gamma Z(\widetilde{W})|] = \\
= L * \mathbb{E}[|(1-t)\beta(Z(\eta_1(t)) - Z(W^A)) + t\gamma(Z(\eta_1(t)) - Z(W^A))|] = \\
= L * \mathbb{E}[|((1-t)\beta + t\gamma)(Z(\eta_1(t)) - Z(W^A))|] = L * \mathbb{E}[|\eta_2(t)(Z(\eta_1(t)) - Z(W^A))|] = \\
= L * \mathbb{E}[|\eta_2(t)_1 * \Delta Z_1 + ... + \eta_2(t)_m * \Delta Z_m|] \le \\
\le L * \mathbb{E}[(|\eta_2(t)_1| * |\Delta Z_1| + ... + |\eta_2(t)_m| * |\Delta Z_m|)],
\end{gathered}$$

where $\Delta Z_i = \eta_i(t)_1 - Z(W_i^A)$.

By construction, we have

$$\begin{gathered}
|\Delta Z_i| = |Z(\eta_1(t)_i) - Z(W_i^A)| = |\max(0, \eta_1(t)_i * X) - \max(0, W_i^A * X)| \le \\
\le |(\eta_1(t)_i - W_i^A) * X| \le \|\eta_1(t)_i - W_i^A\| * \|X\| \le \alpha * \|X\|.
\end{gathered}$$

Therefore,

$$L * \mathbb{E}[(|\eta_2(t)_1| * |\Delta Z_1| + ... + |\eta_2(t)_m| * |\Delta Z_m|)] \leq$$
$$L * \alpha * \mathbb{E}[\|X\|] * (|\eta_2(t)_1| + ... + |\eta_2(t)_m|) = L * \alpha * \sqrt{\|\Sigma\|} * \|\eta_2(t)\|_1.$$

and thus

$$\left|\mathbb{E}[F(\eta_2(t)Z(\eta_1(t)))] - \mathbb{E}[F((1-t)\beta Z(W^A) + t\gamma Z(\widetilde{W}))]\right| \leq L\,\alpha\,\sqrt{\|\Sigma\|}\,\|\eta_2(t)\|_1.$$

Using Lemma 1, we further bound

$$L\,\alpha\,\sqrt{\|\Sigma\|}\,\|\eta_2(t)\|_1 \leq L^2\,\alpha\,\sqrt{\|\Sigma\|}\,\kappa^{-1} = C\alpha,$$

where C is a constant depending only on the data distribution P and the chosen loss function F.

Thus, for all subpaths except (2) and (5), the maximal loss is attained at the endpoints due to convexity and equals one of $\lambda, \delta_{W_A^1}(m, 0, m), \delta_{W_A^1}(m - l, \alpha, m), e(l), \delta_{W_B^1}(m - l, \alpha, m), \delta_{W_B^1}(m, 0, m)$. For subpaths (2) and (5), the inequality above shows that it suffices to add an additional upper bound term of the form $C\alpha$, where C is an explicit constant independent of θ. Since l and α are arbitrary, taking the infimum over them completes the proof. ■

4 Asymptotic Landscape Smoothing with Width

A striking consequence of Theorem 2 is that as we let $m \to \infty$, the energy gap ϵ in the loss landscape shrinks to 0. This can be understood as an *asymptotic smoothing* of the loss topology: in the limit of infinite width, the loss surface has no local minima at all, since every sublevel set $\Omega_F(\lambda)$ becomes connected.

We now provide a more quantitative statement on the rate at which ϵ decreases with m. The bound is based on an ϵ-net covering argument and is closely inspired by the analysis in Freeman and Bruna (2017).

Theorem 3 (Energy gap vanishes with width) *Let the loss function be convex and satisfy the Lipschitz condition with constant L. Then, as the width m of the hidden layer increases, the energy gap ϵ satisfies $\epsilon = O(m^{-\zeta})$ for some $0 < \zeta < 1$, and the sublevel sets become connected at all energy levels.*

Proof Consider the weight matrix $W \in \mathbb{R}^{n \times m}$ of the first layer. Without loss of generality, assume $\|w_k\| = 1$ for all k (due to the homogeneity of ReLU).

Step 1: Sphere covering and cluster selection.

For the unit sphere S^{n-1}, there exists an ϵ_m-covering with cardinality

$$\mathcal{N}(S^{n-1}, \epsilon_m) \leq \left(1 + \frac{2}{\epsilon_m}\right)^n.$$

Choose $\epsilon_m = m^{\frac{\eta-1}{n}}$ with $0 < \eta < \frac{1}{n+1}$. Then the covering size is

$$u_m \simeq \left(1 + \frac{2}{\epsilon_m}\right)^n \simeq m^{1-\eta}.$$

By the pigeonhole principle, there exists a subset $Q_m \subset \{w_k\}$ of size

$$v_m = |Q_m| \simeq m \cdot u_m^{-1} \simeq m^\eta,$$

such that all vectors in Q_m are pairwise close: $\angle(w_i, w_j) \leq 2\epsilon_m$ for $w_i, w_j \in Q_m$.

Step 2: Error bound under vector removal.

Consider the process of removing vectors from Q_m one by one. When removing w_k, define the new weight vector as

$$\beta_p^{(k)} = \beta'_j \text{ for } j \neq k, \qquad \beta_p^{(k-1)} = \beta'_{k-1} + \beta'_k.$$

The change in loss upon removing w_k is

$$\Delta_k = \mathbb{E}\left[F\left(\sum_{j\neq k} \beta'_j z(w_j)\right)\right] - \mathbb{E}\left[F\left(\sum_{j} \beta'_j z(w_j)\right)\right].$$

Since $\angle(w_k, w_{k-1}) \leq 2\epsilon_m$ and F is Lipschitz, we have

$$z(w_k) = z(w_{k-1}) + n_k, \qquad \mathbb{E}[|n_k|] \leq C_1 \epsilon_m,$$

where C_1 depends on $\|\Sigma_X\|$. Then

$$\left|F\left(\sum_{j} \beta'_j z(w_j)\right) - F\big(\beta_p^T Z(W_{-k}) + \beta'_k n_k\big)\right| \leq L|\beta'_k|\,|n_k|.$$

Hence,

$$|\Delta_k| \leq L|\beta'_k| \mathbb{E}[|n_k|] \leq L|\beta'_k| C_1 \epsilon_m.$$

From the regularization term $\kappa\|\beta\|_1$ and convexity, it follows that $|\beta'_k| \leq \kappa^{-1} C_2$, where C_2 depends on the data. Therefore,

$$|\Delta_k| \leq C_3 L \kappa^{-1} \epsilon_m, \qquad C_3 = C_1 C_2.$$

Step 3: Total error bound.

When removing all v_m vectors, we obtain

$$\left|\sum_{k\in Q_m} \Delta_k\right| \leq v_m\, C_3 L \kappa^{-1} \epsilon_m \simeq m^\eta \cdot m^{\frac{\eta-1}{n}} = m^{\eta + \frac{\eta-1}{n}}.$$

Choosing $\eta < \frac{1}{n+1}$ guarantees $\zeta = -\left(\eta + \frac{\eta-1}{n}\right) > 0$, and thus

$$\sum \Delta_k = O(m^{-\zeta}) \to 0 \quad \text{as} \quad m \to \infty.$$

From Theorem 2, for Lipschitz losses we have

$$\epsilon = \inf_{l,\alpha} \big(\max\{e(l), \delta(\cdot)\} + C\alpha\big).$$

By choosing $l = v_m$, $\alpha = 2\epsilon_m$, we obtain:

1. $e(v_m) \to 0$ as $m \to \infty$;
2. the δ-terms can be bounded through the sum of Δ_k and also vanish;
3. the linear term $C\alpha = C_3 L \kappa^{-1} \epsilon_m = O(m^{\frac{\eta-1}{n}}) \to 0$.

Therefore, $\epsilon \to 0$ as $m \to \infty$, which implies that the sublevel sets become connected at all energy levels. ■

5 Experiments

We empirically study how network width affects the *energy gap* between independently trained solutions. For each pair of models we set $E = \max\{L(\theta_A), L(\theta_B)\}$ and run DSS to trace a low-loss connecting path. The *energy gap* is defined as $\max_t L(\gamma(t)) - E$, and the hit rate is the fraction of runs that reach the DSS maximum depth (indicating an unresolved barrier at that depth). We consider: (i) a synthetic two-class Moons dataset for regression with mean squared error (MSE) and (ii) the Wisconsin Diagnostic Breast Cancer dataset for binary classification with cross-entropy. In addition to comparing mean and median gaps, we explicitly test differences in the *maximum* gap via a permutation test ($p_{\text{perm}} = 0$ in both tasks). We treat this maximum-gap test as the primary indicator, since it directly reflects barrier height reduction. Code, notebooks, and CSV outputs are available at `https://github.com/Safreliy/energy-gap-estimate`.

5.1 Half-Moons (Regression with MSE)

Setup. The Moons dataset consists of two interlocking half-circles. We generate 1000 points with targets $Y \in \{0, 1\}$ and train one-hidden-layer ReLU networks with MSE. We compare widths $m \in \{20, 200\}$ by sampling 200 random pairs of independently trained solutions per width and running DSS between each pair at the pairwise energy level $E = \max\{L(\theta_A), L(\theta_B)\}$.

Energy gap statistics. Table 1 reports the mean/median/max energy gap and DSS hit rate. The wider network shows a lower mean gap, but the difference is not statistically significant (Mann—Whitney, $p = 0.266$; Cliff's $\delta = -0.036$).

Width m	Mean gap	Median gap	Max gap	Hit rate	Pairs
20	$5.6{\times}10^{-5}$	$2.0{\times}10^{-6}$	$1.719{\times}10^{-3}$	0.99	200
200	$6.0{\times}10^{-6}$	$2.0{\times}10^{-6}$	$1.48{\times}10^{-4}$	1.00	200

Table 1: Pairwise energy-gap statistics on Moons (MSE regression).

5.2 Breast Cancer (Classification with Cross-Entropy)

Setup. We use the Wisconsin Diagnostic Breast Cancer dataset (569 samples, 30 real-valued features, binary label). One-hidden-layer ReLU networks with widths $m \in \{20, 200\}$ are trained with binary cross-entropy. As above, we evaluate pairwise energy gaps with $E = \max\{L(\theta_A), L(\theta_B)\}$ and 200 random pairs per width.

Energy gap statistics. Table 2 shows a clear reduction of the mean energy gap with width; the difference is statistically significant (Mann—Whitney, $p = 2.26 \times 10^{-5}$; Cliff's $\delta = -0.222$).

Width m	Mean gap	Median gap	Max gap	Hit rate	Pairs
20	7.93×10^{-4}	1.18×10^{-9}	1.0696×10^{-2}	0.62	200
200	1.24×10^{-4}	0	1.477×10^{-3}	0.415	200

Table 2: Pairwise energy-gap statistics on the Breast Cancer dataset (binary cross-entropy).

Summary. In both tasks, wider networks exhibit smaller pairwise energy gaps on average; the effect is especially pronounced for classification. The max-gap permutation test yields $p_{\text{perm}} = 0$ in both tasks, reinforcing the conclusion that the worst-case barrier height decreases with width.

6 Related Work

Our work is connected to several lines of research on neural network theory and optimization. Classical results in approximation theory (Kolmogorov, 1957; Arnol'd, 1957; Vitushkin and Henkin, 1967; Gorban', 1998; Alexeev, 2010) and neural network foundations (Anthony and Bartlett, 1999) established the expressive power of one-hidden-layer networks. More recent efforts have focused on the geometry of the loss landscape. Kawaguchi (2016) proved that deep linear networks have no suboptimal local minima, and Shamir (2018) showed that certain ResNet architectures avoid worse-than-linear minima. The role of overparameterization in optimization has also been studied: for example, Allen-Zhu et al. (2019) and subsequent works provided analysis of gradient descent in highly overparameterized two-layer networks.

Closest to our work is the study by Freeman and Bruna (2017), who empirically and theoretically investigated the topology of the loss surface for one-hidden-layer ReLU networks. They introduced the idea of constructing paths (geodesics) between minima and observed that increasing the number of neurons tends to eliminate bad minima (making sublevel sets more connected). We build on this insight by extending the connectivity proof to general Lipschitz losses and quantifying the width dependence of the energy barrier.

7 Discussion and Limitations

Our theoretical results currently apply to one-hidden-layer ReLU networks. Extending the connectivity and energy gap analysis to deeper architectures remains an open problem: in multi-layer settings one would have to coordinate moves across several layers simultaneously.

Another important assumption in our analysis is that the loss is convex and Lipschitz in the logits, with an ℓ_1 penalty on the second-layer weights. Convexity was used to ensure that linear interpolations in parameter space do not increase loss, while Lipschitz continuity (together with ℓ_1 control) guaranteed boundedness of the output weights. These conditions may seem restrictive, yet they in fact cover a broad range of practical objectives. For example, the binary logistic loss is globally 1-Lipschitz, and the multiclass cross-entropy is globally $\sqrt{2}$-Lipschitz with respect to the ℓ_2 norm of logits. More generally, any convex, smooth loss that remains bounded on a finite interval is automatically Lipschitz on that interval, which significantly broadens the scope of our theorems. This observation connects our analysis to many commonly used setups in practice.

Of course, some standard training regimes (such as unregularized cross-entropy) may formally violate our assumptions. Nonetheless, our empirical findings indicate that the qualitative phenomenon we describe—overparameterization smoothing the loss landscape—persists beyond the strict theoretical setting. Relaxing or removing these assumptions remains an important direction for future research.

On the algorithmic side, the DSS procedure can become computationally demanding for large networks or datasets, since it requires retraining along multiple intermediate path points. In our experiments we mitigated this by restricting to moderately sized models and datasets. For larger-scale problems, it would be natural to investigate approximate DSS variants, partial training of checkpoints, or parallelization strategies.

8 Conclusion

We have shown that overparameterization fundamentally reshapes the loss landscape of one-hidden-layer ReLU networks, making it increasingly smooth and essentially free of bad local minima. Theoretically, we constructed explicit connecting paths for convex Lipschitz losses and demonstrated that the maximal loss barrier along these paths vanishes as network width grows. Practically, our experiments confirmed these predictions on both synthetic and real data. We believe these insights clarify why overparameterized neural networks are easier to optimize and extend the applicability of landscape connectivity results to a broad and practically relevant class of convex Lipschitz objectives, including cross-entropy.

Acknowledgments

The author is grateful to Prof. Dmitry A. Kamaev for valuable discussions and guidance.

References

Abien Fred Agarap. On breast cancer detection: an application of machine learning algorithms on the wisconsin diagnostic dataset, 2018.

D. V. Alexeev. Neural-network approximation of functions of several variables. *Journal of Mathematical Sciences*, 2010.

Zeyuan Allen-Zhu, Yuanzhi Li, and Yingyu Liang. Learning and generalization in overparameterized neural networks, going beyond two layers. *Advances in neural information processing systems*, 2019.

Martin Anthony and Peter L. Bartlett. *Neural Network Learning: Theoretical Foundations.* Cambridge University Press, 1999.

V. I. Arnol'd. On functions of three variables. *Dokl. Akad. Nauk SSSR*, 1957.

C. Daniel Freeman and Joan Bruna. Topology and geometry of half-rectified network optimization. *ICLR 2017*, 2017.

A. N. Gorban'. Generalized approximation theorem and computational capabilities of neural networks. *Sib. Zh. Vychisl. Mat.*, 1998.

Kenji Kawaguchi. Deep learning without poor local minima. *NIPS 2016*, 2016.

A. N. Kolmogorov. On the representation of continuous functions of many variables by superposition of continuous functions of one variable and addition. *Dokl. Akad. Nauk SSSR*, 1957.

Guido F. Montufar, Razvan Pascanu, Kyunghyun Cho, and Yoshua Bengio. On the number of linear regions of deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2014.

Razvan Pascanu, Guido F. Montufar, and Yoshua Bengio. On the number of response regions of deep feed forward networks with piece-wise linear activations. *arXiv preprint*, 2013.

Ohad Shamir. Are resnets provably better than linear predictors? *NIPS 2018*, 2018.

A. G. Vitushkin and G. M. Henkin. Linear superpositions of functions. *Russian Math. Surveys*, 1967.

C. Zdravko. *Inequalities. Theorems, Techniques and Selected Problems*, chapter 7, pages 69–73. Springer, 2012.