THEORETICAL AND APPLIED MECHANICS
Volume 52 (2025) Issue 1, 67–73 DOI: https://doi.org/10.2298/TAM250121008C

ON NON-APPROXIMABILITY OF ZERO LOSS GLOBAL $\mathcal{L}^2$ MINIMIZERS BY GRADIENT DESCENT IN DEEP LEARNING

Thomas Chen and Patricia Muñoz Ewald

Abstract. We analyze geometric aspects of the gradient descent algorithm in Deep Learning (DL), and give a detailed discussion of the circumstance that, in underparametrized DL networks, zero loss minimization cannot generically be attained. As a consequence, we conclude that the distribution of training inputs must necessarily be non-generic in order to produce zero loss minimizers, both for the method constructed in [2, 3], or for gradient descent [1] (which assume clustering of training data).

1. Introduction and Main Results

We analyze some basic geometric aspects of the gradient descent algorithm in Deep Learning (DL) networks. For some thematically related background, see for instance [4,5,7–11] and the references therein. In our previous papers [2,3], we gave an explicit construction of globally minimizing weights and biases for the $\mathcal{L}^2$ cost in underparametrized ReLU DL networks, leading to zero loss (i.e., the value of the cost is zero). In the work at hand, we address the fact that in the underparametrized case, zero loss minimizers do not exist generically. As a consequence, we conclude that the distribution of training inputs must necessarily be non-generic to allow for zero loss minimizers, both for the method constructed in [2, 3] (which assumes clustering of training data), or for gradient descent [1].

We let the input space be given by $\mathbb{R}^M$, with training inputs $x_j^{(0)} \in \mathbb{R}^M$, $j = 1, \dots, N$. We assume that the outputs are given by $y_\ell \in \mathbb{R}^Q$, $\ell = 1, \dots, Q$ where $Q \leq M$. We introduce the map $\omega\colon \{1, \dots, N\} \to \{1, \dots, Q\}$, which assigns the output label $\omega(j)$ to the j-th input label, that is, $x_j^{(0)}$ corresponds to the output $y_{\omega(j)}$. We define $\underline{y}_\omega := (y_{\omega(1)}, \dots, y_{\omega(N)})^T \in \mathbb{R}^{NQ}$, where A^T is the transpose of the matrix A. Let N_i denote the number of training inputs belonging to the output vector y_i, $i = 1, \dots, Q$.

We assume that the DL network contains L hidden layers, with the ℓ-th layer defined on $\mathbb{R}^{M_\ell}$, and recursively determined by

$$x_j^{(\ell)} = \sigma(W_\ell x_j^{(\ell-1)} + b_\ell) \ \in \mathbb{R}^{M_\ell}$$

via the weight matrix $W_\ell \in \mathbb{R}^{M_\ell \times M_{\ell-1}}$, bias vector $b_\ell \in \mathbb{R}^{M_\ell}$, and activation function σ. We assume that σ has a Lipschitz continuous derivative, and that the output layer

$$x_j^{(L+1)} = W_{L+1} x_j^{(L)} + b_{L+1} \ \in \mathbb{R}^Q$$

2020 *Mathematics Subject Classification:* 57R70, 62M45.
Key words and phrases: deep learning, underparametrization, generic training data, zero loss.

contains no activation function.

We let the vector $\underline{\theta} \in \mathbb{R}^K$ enlist all components of all weights W_ℓ and biases b_ℓ, $\ell = 1, \dots, L+1$, including those in the output layer. Accordingly,

$$K = \sum_{\ell=1}^{L+1} (M_\ell M_{\ell-1} + M_\ell)$$

where we define $M_0 \equiv M$ for the input layer.

In the output layer, we denote $x_j^{(L+1)} \in \mathbb{R}^Q$ by $x_j[\underline{\theta}]$ for brevity, and obtain the $\mathcal{L}^2$ cost as

$$\mathcal{C}[\underline{x}[\underline{\theta}]] = \frac{1}{2N} |\underline{x}[\underline{\theta}] - \underline{y}_\omega|^2_{\mathbb{R}^{QN}} = \frac{1}{2N} \sum_j |x_j[\underline{\theta}] - y_{\omega(j)}|^2_{\mathbb{R}^Q},$$

using the notation $\underline{x} := (x_1, \dots, x_N)^T \in \mathbb{R}^{QN}$. Here, $|\bullet|_{\mathbb{R}^n}$ is the Euclidean norm.

1.1. Comparison model. We consider the following toy model for comparison, defined by the gradient flow,

$$\partial_s \underline{x}(s) = -\nabla_{\underline{x}} \mathcal{C}[\underline{x}(s)] \quad \underline{x}(0) = \underline{x}^{(0)} \in \mathbb{R}^{QN} \tag{1.1}$$

parametrized by $s \in \mathbb{R}$, or in components,

$$\partial_s (x_j(s) - y_{\omega(j)}) = -\frac{1}{N} (x_j(s) - y_{\omega(j)})$$

for all $j = 1, \dots, N$. This is trivially solvable,

$$x_j(s) - y_{\omega(j)} = e^{-\frac{s}{N}} (x_j(0) - y_{\omega(j)})$$

with initial data $x_j(0) = x_j^{(0)}$. Because the right hand side converges to zero as $s \to \infty$, we find that $x_j(s) \to y_{\omega(j)}$ as $s \to \infty$ for all j. In particular, this yields a zero loss, global minimum of the cost, since $\mathcal{C}[\underline{x}(s)] \to 0$ as $s \to \infty$.

1.2. Gradient descent flow. The gradient descent algorithm seeks to minimize the cost function by the use of the gradient flow for the vector of weights and biases defined by

$$\partial_s \underline{\theta}(s) = -\nabla_{\underline{\theta}} \mathcal{C}[\underline{x}[\underline{\theta}(s)]] \quad \underline{\theta}(0) = \underline{\theta}_0 \in \mathbb{R}^K, \tag{1.2}$$

where the vector field $\nabla_{\underline{\theta}} \mathcal{C}[\underline{x}[\bullet]] \colon \mathbb{R}^K \to \mathbb{R}^K$ is Lipschitz continuous if the same holds for the derivative of the activation function σ. Accordingly, the existence and uniqueness theorem for ordinary differential equations holds for (1.2). In computational applications, the initial data $\underline{\theta}_0 \in \mathbb{R}^K$ is often chosen at random. Clearly, because of

$$\partial_s \mathcal{C}[\underline{x}[\underline{\theta}(s)]] = -\left|\nabla_{\underline{\theta}} \mathcal{C}[\underline{x}[\underline{\theta}(s)]]\right|^2_{\mathbb{R}^K} \leq 0 \tag{1.3}$$

the cost $\mathcal{C}[\underline{x}[\underline{\theta}(s)]]$ is monotone decreasing in s, and since $\mathcal{C}[\underline{x}[\underline{\theta}(s)]] \geq 0$ is bounded below, the limit $\mathcal{C}_* = \lim_{s\to\infty} \mathcal{C}[\underline{x}[\underline{\theta}(s)]]$ exists for any orbit $\{\underline{\theta}(s) | s \in \mathbb{R}\}$, and depends on the initial data, $\mathcal{C}_0 = \mathcal{C}[\underline{x}[\underline{\theta}(0)]]$.

Convergence of $\mathcal{C}[\underline{x}[\underline{\theta}(s)]]$ implies that $\lim_{s\to\infty} |\partial_s \mathcal{C}[\underline{x}[\underline{\theta}(s)]]| = 0$, and therefore, $\lim_{s\to\infty} |\nabla_{\underline{\theta}} \mathcal{C}[\underline{x}[\underline{\theta}(s)]]|_{\mathbb{R}^K} = 0$ from (1.3). Thus, the basic goal is to find $\mathcal{C}_* = \lim_{s\to\infty} \mathcal{C}[\underline{x}[\underline{\theta}(s)]] = \mathcal{C}[\underline{x}[\underline{\theta}_*]]$ where $\underline{\theta}_*$ is a critical point of the gradient flow (1.2), satisfying $0 = -\nabla_{\underline{\theta}} \mathcal{C}[\underline{x}[\underline{\theta}_*]]$.

REMARK 1.1. Notably, as $s \to \infty$, neither does $\lim_{s\to\infty} \mathcal{C}[\underline{x}[\underline{\theta}(s)]] = \mathcal{C}[\underline{x}[\underline{\theta}_*]]$ imply that $\underline{\theta}(s)$ converges to $\underline{\theta}_*$, nor to any other element of $\{\underline{\theta}_{**} \in \mathbb{R}^K | \mathcal{C}[\underline{x}[\underline{\theta}_{**}]] = \mathcal{C}[\underline{x}[\underline{\theta}_*]]\}$, nor that $\underline{\theta}(s)$ converges at all, without further assumptions on $\mathcal{C}[\underline{x}[\bullet]]$ (for instance, of it being Morse-Bott).

Therefore, while $\mathcal{C}[\underline{x}[\underline{\theta}(s)]]$ always converges to a stationary value of the cost function under the gradient descent flow, $\underline{\theta}(s)$ cannot generally be assumed to

converge to a minimizer $\underline{\theta}_*$. This is a key shortcoming of the gradient descent method, as for the training of a DL network, the main task is to find minimizing weights and biases $\underline{\theta}_*$.

REMARK 1.2. As an elementary 1-dimensional example illustrating the situation addressed in Remark 1.1, we may consider $x[\theta] = \theta\frac{1}{\theta^2+1}$ and $\mathcal{C}[x[\theta]] = \frac{1}{2}(x[\theta])^2 = \frac{1}{2}\theta^2\frac{1}{(\theta^2+1)^2} \geq 0$ for $\theta \in \mathbb{R}$. Here, clearly, $\theta_* = 0$ is a critical value and global minimizer. The gradient descent flow is determined by $\partial_s\theta(s) = -\partial_\theta\mathcal{C}[x[\theta(s)]] = \theta(s)((\theta(s))^2-1)\frac{1}{((\theta(s))^2+1)^3}$, and one easily verifies that given any initial data with $|\theta_0| < 1$, the corresponding orbit converges, $\lim_{s\to\infty}\theta(s) = \theta_* = 0$.

On the other hand, given any initial data with $|\theta_0| > 1$, the corresponding orbit diverges, $\lim_{s\to\infty}|\theta(s)| = \infty$, while nevertheless, $\lim_{s\to\infty} x[\theta(s)] = 0$, and therefore, $\lim_{s\to\infty}\mathcal{C}[x[\theta(s)]] = 0 = \mathcal{C}[x[\theta_*]]$. This is because $|\partial_\theta\mathcal{C}[x[\theta]]| \sim \frac{1}{|\theta|^3}$ for $|\theta| \gg 1$, and one straightforwardly verifies that for $\theta \gg 1$, the solution of $\partial_s\theta(s) \sim \frac{1}{(\theta(s))^3}$ has the asymptotic behavior $\theta(s) \sim s^{\frac{1}{4}} \to \infty$ as $s \to \infty$. The case for $\theta \ll -1$ is similar.

1.3. Dynamics of $\underline{x}(s) := \underline{x}[\underline{\theta}(s)]$. Next, we note that $\mathcal{C}[\underline{x}[\underline{\theta}(s)]]$ depends on $\underline{\theta}(s)$ only through its dependence on $\underline{x}[\underline{\theta}(s)]$. Thus, defining the Jacobi matrix

$$D[\underline{\theta}] := \left[\frac{\partial x_j[\underline{\theta}]}{\partial\theta_\ell}\right]_{j=1,\dots,N\ \ell=1,\dots,K} = \begin{bmatrix} \frac{\partial x_1[\underline{\theta}]}{\partial\theta_1} & \cdots & \frac{\partial x_1[\underline{\theta}]}{\partial\theta_K} \\ \cdots & \cdots & \cdots \\ \frac{\partial x_N[\underline{\theta}]}{\partial\theta_1} & \cdots & \frac{\partial x_N[\underline{\theta}]}{\partial\theta_K} \end{bmatrix} \in \mathbb{R}^{QN\times K}$$

and writing $\underline{x}(s) := \underline{x}[\underline{\theta}(s)]$ for brevity, we find that the gradient descent flow for $\underline{\theta}(s)$ induces the following flow for $\underline{x}(s) \in \mathbb{R}^{QN}$,

$$\begin{aligned} \partial_s\underline{x}(s) = D[\underline{\theta}(s)]\partial_s\underline{\theta} &= -D[\underline{\theta}(s)]\nabla_{\underline{\theta}}\mathcal{C}[\underline{x}[\underline{\theta}(s)]] \\ &= -D[\underline{\theta}(s)]D^T[\underline{\theta}(s)]\nabla_{\underline{x}}\mathcal{C}[\underline{x}[\underline{\theta}(s)]] \end{aligned} \tag{1.4}$$

Passing to the second line, we used (1.2). Here, the matrix $D[\underline{\theta}(s)]D^T[\underline{\theta}(s)] \in \mathbb{R}^{QN\times QN}$ is positive semi-definite; it corresponds to the neural tangent kernel [6]. In the special case when it is strictly positive definite and thus invertible, (1.4) is the gradient flow for $\underline{x}(s)$ in the metric on $\mathbb{R}^{QN}$ defined by the metric tensor $(D[\underline{\theta}(s)]D^T[\underline{\theta}(s)])^{-1}$. Our main results in this paper address the similarity or dissimilarity in the qualitative behavior between solutions to (1.4) and the comparison model (1.1), depending on this invertibility condition.

1.3.1. *The overparametrized case.* In the overparametrized situation where $K \geq QN$, we have the following result.

THEOREM 1.1. *Assume that $\underline{x}[\underline{\theta}_*]$ is a stationary solution,*

$$0 = -D[\underline{\theta}_*]D^T[\underline{\theta}_*]\nabla_{\underline{x}}\mathcal{C}[\underline{x}[\underline{\theta}_*]] \tag{1.5}$$

Then, it corresponds to a global minimum of the $\mathcal{L}^2$ cost,

$$\mathcal{C}[\underline{x}[\underline{\theta}_*]] = 0,$$

if and only if $\nabla_{\underline{x}}\mathcal{C}[\underline{x}[\underline{\theta}_*]] = 0$.

A necessary condition for $\nabla_{\underline{x}}\mathcal{C}[\underline{x}[\underline{\theta}_*]] = 0$ to follow from (1.5) is that

$$\operatorname{rank}(D[\underline{\theta}_*]D^T[\underline{\theta}_*]) = QN$$

has full rank. This, in turn, is only possible if $K \geq QN$ which means that the DL network is overparametrized.

Moreover, if there exist $s_0 \geq 0$ and $\lambda > 0$ such that $D[\underline{\theta}(s)]D^T[\underline{\theta}(s)] > \lambda$ for all $s \geq s_0$ (so that, in particular, $\operatorname{rank}(D[\underline{\theta}(s)]D^T[\underline{\theta}(s)]) = QN$) along the orbit $\underline{\theta}(s)$, the solution of (1.4) converges to the global minimizer for any initial condition $\underline{x}(0) \in \mathbb{R}^{QN}$.

PROOF. In components, $\nabla_{\underline{x}}\mathcal{C}[\underline{x}[\underline{\theta}_*]] = 0$ is explicitly given by

$$\frac{1}{N}(x_j[\underline{\theta}_*] - y_{\omega(j)}) = 0 \quad \forall j \in \{1, \dots, N\}.$$

Therefore, $\nabla_{\underline{x}}\mathcal{C}[\underline{x}[\underline{\theta}_*]] = 0$ is equivalent to $x_j[\underline{\theta}_*] = y_{\omega(j)}$ for all j, and thus holds if and only if $\mathcal{C}[\underline{x}[\underline{\theta}_*]] = 0$.

We recall that $D[\underline{\theta}_*] \in \mathbb{R}^{QN\times K}$ where K is the total number of parameters contained in all weights and biases. Therefore, $\operatorname{rank}(D[\underline{\theta}_*]D^T[\underline{\theta}_*]) \leq \min\{QN, K\}$, and for $D[\underline{\theta}_*]D^T[\underline{\theta}_*]$ to have full rank QN, it is necessary that $QN \leq K$. But, this means that the DL network is overparametrized.

Finally, if there exists $s_0 \geq 0$ such that $D[\underline{\theta}(s)]D^T[\underline{\theta}(s)] > \lambda$ for a positive constant $\lambda > 0$ and all $s \geq s_0$, then

$$\begin{aligned}\partial_s\mathcal{C}[\underline{x}(s)] &= -(\nabla_{\underline{x}}\mathcal{C}[\underline{x}(s)])^T D[\underline{\theta}(s)]D^T[\underline{\theta}(s)]\nabla_{\underline{x}}\mathcal{C}[\underline{x}(s)] \\ &\leq -\lambda|\nabla_{\underline{x}}\mathcal{C}[\underline{x}(s)]|^2_{\mathbb{R}^{QN}} = -2\frac{\lambda}{N}\mathcal{C}[\underline{x}(s)]\end{aligned}$$

for all $s > s_0$. Therefore, $\lim_{s\to\infty}\mathcal{C}[\underline{x}(s)] \leq \lim_{s\to\infty} e^{-2\frac{\lambda}{N}(s-s_0)}\mathcal{C}[\underline{x}(s_0)] = 0$. Since $\mathcal{C}[\underline{x}(s)]$ is a convex function of $\underline{x}(s) - \underline{y}_\omega$, this implies that for any arbitrary initial data $\underline{x}(0) = \underline{x}_0 \in \mathbb{R}^{QN}$, the solution of (1.4) converges to the global minimizer $\underline{x}_* = \lim_{s\to\infty}\underline{x}(s)$ which satisfies $\underline{x}_* - \underline{y}_\omega = 0$. □

REMARK 1.3. We note that while $\underline{x}_* = \lim_{s\to\infty}\underline{x}(s) = \lim_{s\to}\underline{x}[\underline{\theta}(s)]$ converges in the above situation, the vector of weights and biases $\underline{\theta}(s)$ itself nevertheless does not necessarily converge.

1.3.2. *The underparametrized case.* In the underparametrized situation where $K < QN$, we have the following result.

THEOREM 1.2. *Assume that $K < QN$, and that $\underline{\theta}(s)$, $s \in \mathbb{R}$, is an orbit of the gradient descent flow (1.2). Denote by $\mathcal{P}[\underline{\theta}(s)]$ the projector, orthogonal with respect to the Euclidean inner product on $\mathbb{R}^{QN}$, onto the range of $D[\underline{\theta}(s)]D^T[\underline{\theta}(s)]$ where $\operatorname{rank}(D[\underline{\theta}(s)]D^T[\underline{\theta}(s)]) \leq K$ (the latter is not assumed to be constant in s), and let $\mathcal{P}^\perp[\underline{\theta}(s)] := \mathbf{1}_{QN\times QN} - \mathcal{P}[\underline{\theta}(s)]$ denote its complement. Then,*

$$\begin{aligned}\partial_s\underline{x}(s) &= -\mathcal{P}[\underline{\theta}(s)](D[\underline{\theta}(s)]D^T[\underline{\theta}(s)]\nabla_{\underline{x}}\mathcal{C}[\underline{x}[\underline{\theta}(s)]]) \\ \mathcal{P}^\perp[\underline{\theta}(s)]\partial_s\underline{x}(s) &= 0\end{aligned}$$

has the structure of a constrained dynamical system. In particular,

$$\mathcal{P}[\underline{\theta}(s)] = D[\underline{\theta}(s)](D^T[\underline{\theta}(s)]D[\underline{\theta}(s)])^{-1}D^T[\underline{\theta}(s)], \tag{1.6}$$

if $\operatorname{rank}(D[\underline{\theta}(s)]D^T[\underline{\theta}(s)]) = K$ is maximal. Let $\underline{\theta}_*$ be an arbitrary stationary point of the cost function, with $\nabla_{\underline{\theta}}\mathcal{C}[\underline{x}[\theta_*]] = 0$, and $\operatorname{rank}(D[\underline{\theta}_*]D^T[\underline{\theta}_*]) \leq K$. Then,

$$0 = \mathcal{P}[\underline{\theta}_*]\nabla_{\underline{x}}\mathcal{C}[\underline{x}[\underline{\theta}_*]].$$

In particular, the local extremum of the cost function at $\underline{\theta}_*$ is attained at

$$\mathcal{C}[\underline{x}[\underline{\theta}_*]] = \frac{N}{2}\left|\mathcal{P}^\perp[\underline{\theta}_*]\nabla_{\underline{x}}\mathcal{C}[\underline{x}[\underline{\theta}_*]]\right|^2_{\mathbb{R}^{QN}} \tag{1.7}$$

where $\operatorname{rank}(\mathcal{P}^\perp[\underline{\theta}_*]) \geq QN - K$.

PROOF. Due to being a symmetric matrix, $D[\underline{\theta}(s)]D^T[\underline{\theta}(s)] = R^T\Lambda R$ for any given $s \in \mathbb{R}$ (where we notationally suppress the dependence of R and Λ on $\underline{\theta}(s)$ for brevity), where $\Lambda \geq 0$ is diagonal and $R \in SO(QN)$. Then, letting P_Λ denote the projector obtained from replacing all nonzero entries of Λ by 1, we have $\mathcal{P}[\underline{\theta}(s)] = R^T P_\Lambda R$. From $P_\Lambda\Lambda = \Lambda = \Lambda P_\Lambda$ follows that

$$D[\underline{\theta}(s)]D^T[\underline{\theta}(s)] = \mathcal{P}[\underline{\theta}(s)]D[\underline{\theta}(s)]D^T[\underline{\theta}(s)] = D[\underline{\theta}(s)]D^T[\underline{\theta}(s)]\mathcal{P}[\underline{\theta}(s)]. \tag{1.8}$$

In other words, $[D[\underline{\theta}(s)]D^T[\underline{\theta}(s)], \mathcal{P}[\underline{\theta}(s)]] = 0$ commute, and from

$$D[\underline{\theta}(s)]D^T[\underline{\theta}(s)] = \mathcal{P}[\underline{\theta}(s)]D[\underline{\theta}(s)]D^T[\underline{\theta}(s)]\mathcal{P}[\underline{\theta}(s)], \tag{1.9}$$

the ranges and kernels of $D[\underline{\theta}(s)]D^T[\underline{\theta}(s)]$ and $\mathcal{P}[\underline{\theta}(s)]$ coincide, for every $s \in \mathbb{R}$.

If $\mathrm{rank}(D[\underline{\theta}(s)]D^T[\underline{\theta}(s)]) = K$ is maximal, then the matrix $D^T[\underline{\theta}(s)]D[\underline{\theta}(s)] \in \mathbb{R}^{K\times K}$ is invertible, as a consequence of which the expression (1.6) for the ortho-projector $\mathcal{P}[\underline{\theta}(s)]$ is well-defined.

It follows from (1.4) and (1.8) that

$$\partial_s \underline{x}(s) = -\mathcal{P}[\underline{\theta}(s)](D[\underline{\theta}(s)]D^T[\underline{\theta}(s)]\nabla_{\underline{x}}\mathcal{C}[\underline{x}[\underline{\theta}(s)]]),$$

and as a consequence,

$$\mathcal{P}^{\perp}[\underline{\theta}(s)]\partial_s \underline{x}(s) = 0.$$

It follows from (1.4) that $\partial_s\underline{\theta}(s) = 0$ implies $\partial_s\underline{x}(s) = 0$.

Let $\underline{\theta}_*$ denote a stationary point for (1.2), with $\mathrm{rank}(D[\underline{\theta}_*]D^T[\underline{\theta}_*]) \leq K$, so that clearly, $\mathrm{rank}(\mathcal{P}^{\perp}[\underline{\theta}_*]) \geq QN - K$. Then,

$$0 = -D[\underline{\theta}_*]D^T[\underline{\theta}_*]\nabla_{\underline{x}}\mathcal{C}[\underline{x}[\underline{\theta}_*]]$$

from which follows that

$$\mathcal{P}[\underline{\theta}_*]\nabla_{\underline{x}}\mathcal{C}[\underline{x}[\underline{\theta}_*]] = 0,$$

due to (1.9). Then,

$$\begin{aligned}\mathcal{C}[\underline{\theta}_*] = \frac{N}{2}|\nabla_{\underline{x}}\mathcal{C}[\underline{x}[\underline{\theta}_*]]|^2 &= \frac{N}{2}\left(\left|\mathcal{P}[\underline{\theta}_*]\nabla_{\underline{x}}\mathcal{C}[\underline{x}[\underline{\theta}_*]]\right|^2_{\mathbb{R}^{QN}} + \left|\mathcal{P}^{\perp}[\underline{\theta}_*]\nabla_{\underline{x}}\mathcal{C}[\underline{x}[\underline{\theta}_*]]\right|^2_{\mathbb{R}^{QN}}\right)\\ &= \frac{N}{2}\left|\mathcal{P}^{\perp}[\underline{\theta}_*]\nabla_{\underline{x}}\mathcal{C}[\underline{x}[\underline{\theta}_*]]\right|^2_{\mathbb{R}^{QN}}\end{aligned}$$

as claimed. □

1.3.3. *Comparison with constructive minimizers from* [2]. The map $\underline{x}\colon \mathbb{R}^K \to \mathbb{R}^{QN}$, $\underline{\theta} \mapsto \underline{x}[\underline{\theta}]$ from the space of parameters to the output space is determined by the training inputs $\left\{(x^{(0)}_{j,i})_{i=1}^{N_j}\right\}_{j=1}^{Q}$. Accordingly, the map $\underline{x}\colon \mathbb{R}^K \to \mathbb{R}^{QN}$ is generic if the distribution of training inputs $\left\{(x^{(0)}_{j,i})_{i=1}^{N_j}\right\}_{j=1}^{Q}$ is generic (i.e., it is highly random and unstructured).

We, therefore, arrive at the following main result.

THEOREM 1.3. *Zero loss minimizers of underparametrized ReLU DL networks do not exist for generic distributions of training data.*

PROOF. The ReLU activation function σ acts component-wise by the ramp function $(\xi)_+ = \max\{0, \xi\}$ for $\xi \in \mathbb{R}$. Suitably smoothing the latter in an ϵ-neighborhood of the origin for an arbitrary small $\epsilon > 0$, we obtain σ_ϵ, which we assume to be monotone increasing, and to have a Lipschitz continous derivative. Accordingly, the corresponding gradient vector field $\nabla_{\underline{x}}\mathcal{C}[\underline{x}[\underline{\theta}]]$ and the matrix $D[\underline{\theta}]$ are Lipschitz continuous in $\underline{\theta}$. Therefore, Theorem 1.2 can be applied to the flow generated by it. For generic training inputs, the right hand side of (1.7) is strictly positive, due to the nonzero rank of $\mathcal{P}^{\perp}[\underline{\theta}_*]$; accordingly, zero loss does not occur. We conclude that zero loss minimizing weights and biases for the $\mathcal{L}^2$ cost in underparametrized ReLU DL networks do not exist, and cannot be approximated via the gradient descent flow, if the distribution of training inputs is generic. □

The DL network studied in [1–3] is underparametrized, with $M = M_\ell = Q = L$ $\forall \ell$ and $K = (Q+1)^3 + (Q+1)^2 < QN$ (respectively, $\ll QN$), and uses the ReLU activation function. The minimizers obtained in [1–3] are robust under a small deformation of σ to a monotone increasing σ_ϵ (in particular, they involve no derivatives of the activation function). Accordingly, the construction given in [2, 3] with σ replaced by σ_ϵ yields a degenerate zero loss minimum of the cost

function. In [1–3], the training data is clustered, and hence non-generic. Therefore, the existence of zero loss minimizers, constructed explicitly for underparametrized ReLU DL in [2, 3], and via gradient flow in [1], is not in contradiction with the above.

Acknowledgments. We thank the anonymous reviewer for very useful comments. T. C. thanks Cy Mayor for helpful discussions. T. C. gratefully acknowledges the support by the NSF through the grant DMS-2009800, and the RTG Grant DMS-1840314 - *Analysis of PDE*. P. M. E. was supported by the NSF grant DMS-2009800 through T. C.

References

1. T. Chen, *Derivation of effective gradient flow equations and dynamical truncation of training data in deep learning*, (2025), https://arxiv.org/abs/2501.07400.
2. T. Chen, P. Muñoz Ewald, *Geometric structure of Deep Learning networks and construction of global $\mathcal{L}^2$ minimizers*. (2024), https://arxiv.org/abs/2309.10639.
3. T. Chen, P. Muñoz Ewald, *Interpretable global minima of deep ReLU neural networks on sequentially separable data*, (2024), arxiv.org/abs/2405.07098.
4. B. Hanin, D. Rolnick, *How to start training: The effect of initialization and architecture*, In: S. Bengio, H. M. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi (eds), *NIPS'18: Proceedings of the 32^{nd} International Conference on Neural Information Processing Systems*, Montréal, Canada, December 3–8, 2018, Curran Associates Inc., Red Hook, NY, USA, 2018, 569–579.
5. P. Grohs, G. Kutyniok, *Mathematical Aspects of Deep Learning*, Cambridge University Press, Cambridge, 2023.
6. A. Jacot, F. Gabriel, C. Hongler, *Neural tangent kernel: Convergence and generalization in neural networks*, In: S. Bengio, H. M. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi (eds), *NIPS'18: Proceedings of the 32^{nd} International Conference on Neural Information Processing Systems*, Montréal, Canada, December 3–8, 2018, Curran Associates Inc., Red Hook, NY, USA, 2018, 8580–8589.
7. K. Karhadkar, M. Murray, H. Tseran, G. Montufar, *Mildly overparameterized relu networks have a favorable loss landscape*, (2024), arXiv:2305.19510.
8. Y. LeCun, Y. Bengio, G. Hinton, *Deep learning*, Nature **521** (2015), 436–521.
9. S. S. Mannelli, E. Vanden-Eijnden, L. Zdeborová, *Optimization and generalization of shallow neural networks with quadratic activation functions*, In: H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, H. Lin (eds), *NIPS'20: Proceedings of the 34^{th} International Conference on Neural Information Processing Systems*, Vancouver, BC, Canada, December 6–12, 2020, Curran Associates Inc., Red Hook, NY, USA, 2020, 13445–13455.
10. M. Nonnenmacher, D. Reeb, I. Steinwart, *Which minimizer does my neural network converge to?* In: N. Oliver, F. Pérez-Cruz, S. Kramer, J. Read, J. A. Lozano (eds), *Machine Learning and Knowledge Discovery in Databases. Research Track: European Conference, ECML PKDD 2021*, Bilbao, Spain, September 13–17, 2021, Part III, Springer, Cham, 2021, 87–102.
11. V. Papyan, X. Y. Han, D. L. Donoho, *Prevalence of neural collapse during the terminal phase of deep learning training*, Proc. Natl. Acad. Sci. USA **117**(40) (2020), 24652–24663.

О НЕМОГУЋНОСТИ АПРОКСИМАЦИЈЕ ГЛОБАЛНИХ $\mathcal{L}^2$ МИНИМИЗАТОРА БЕЗ ГУБИТАКА ГРАДИЈЕНТНИМ СПУШТАЊЕМ У ДУБОКОМ УЧЕЊУ

Резиме. Анализирамо геометријске аспекте алгоритма градијентног спуштања у дубоком учењу (DL) и дајемо детаљну дискусију о околности да се, у недовољно параметризованим DL мрежама, минимизирање без губитака не може генерички постићи. Као последица тога, закључујемо да дистрибуција улаза за обуку нужно мора бити негенеричка да би се произвели минимизатори без губитака, како за метод конструисан у [2,3], тако и за градијентно спуштање [1] (што претпоставља груписање података за обуку).

Department of Mathematics
University of Texas at Austin
Austin TX
USA
tc@math.utexas.edu
https://orcid.org/0000-0003-2704-1454

Department of Mathematics
University of Texas at Austin
Austin TX
USA
ewald@utexas.edu
https://orcid.org/0000-0001-8400-9114

(Received 21.01.2025)
(Revised 05.05.2025)
(Available online 20.05.2025)