Convergence in Linear Neural Networks

The Algebro-Geometric Approach

Ludwig Hedlin

October 13, 2020

Outline

- What is a linear neural network?

Outline

- What is a linear neural network?
- What do we already know about their convergence?

Outline

- What is a linear neural network?
- What do we already know about their convergence?
- What do we want to prove about their convergence and how?

Outline

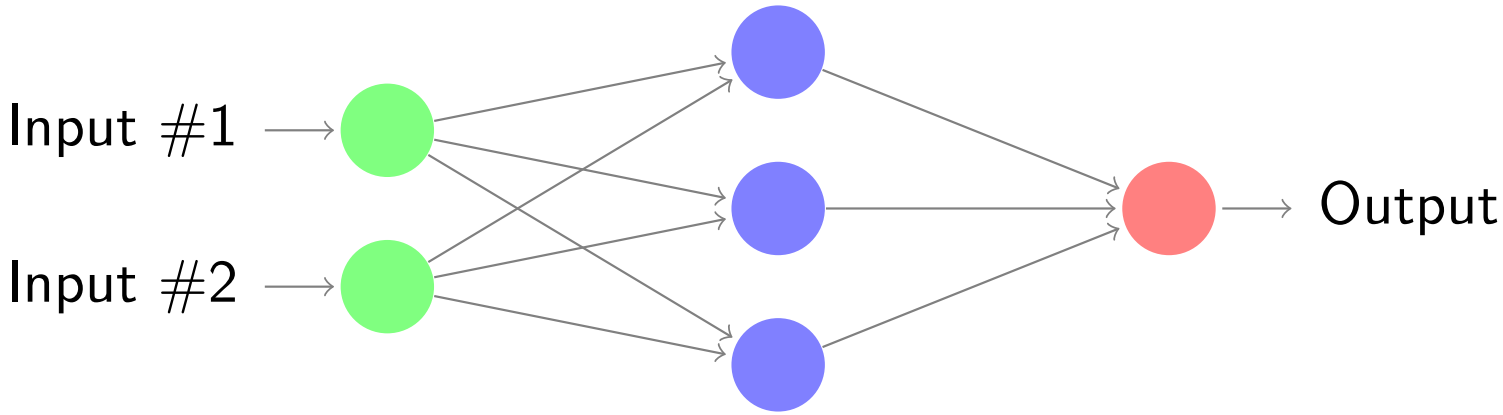
- What is a linear neural network?
- What do we already know about their convergence?
- What do we want to prove about their convergence and how?
- A few successes and the wall.

Linear Neural Networks...

... are simply a glorified form of matrix multiplication.

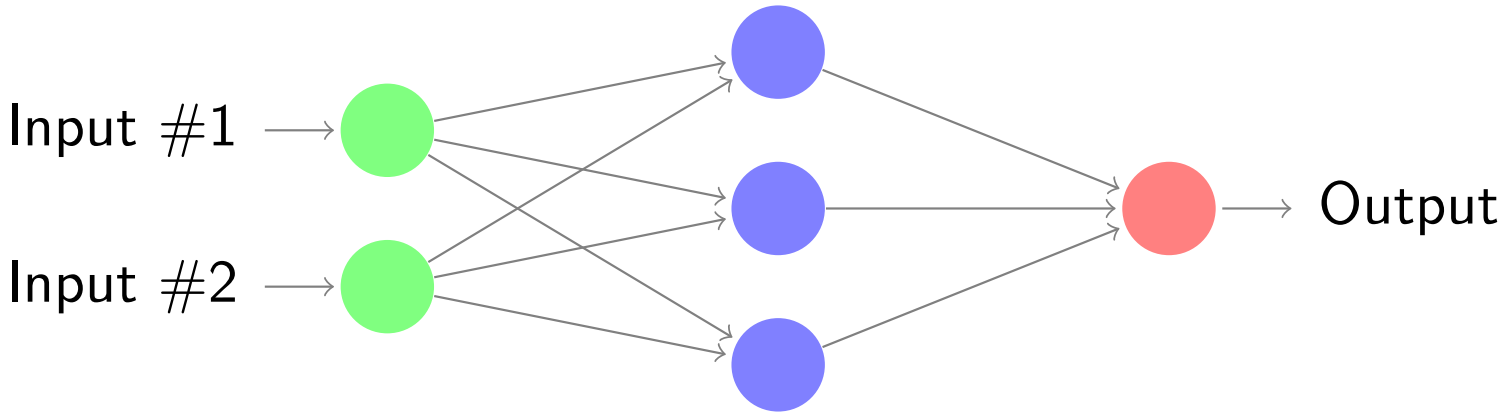
Linear Neural Networks...

... are simply a glorified form of matrix multiplication.



Linear Neural Networks...

... are simply a glorified form of matrix multiplication.



Is simply $W_2 W_1 X$. X being the data matrix, each sample being a column. Each W_i being the matrix where each column is the weights going out of a node.

Notation

We say that a linear network represented by the matrices $(W_N, ..., W_1)$ has depth N.

Notation

We say that a linear network represented by the matrices $(W_N, ..., W_1)$ has depth N.

The tuple of all parameters $(W_N, ..., W_1)$ we call W_a.

Notation

We say that a linear network represented by the matrices $(W_N, ..., W_1)$ has depth N.

The tuple of all parameters $(W_N, ..., W_1)$ we call W_a.

The size of layer i is called d_i, starting from input d_0.

Notation

We say that a linear network represented by the matrices $(W_N, ..., W_1)$ has depth N.

The tuple of all parameters $(W_N, ..., W_1)$ we call W_a.

The size of layer i is called d_i, starting from input d_0.

The full product $W_N ... W_1$ we call W.

Notation

We say that a linear network represented by the matrices $(W_N, ..., W_1)$ has depth N.

The tuple of all parameters $(W_N, ..., W_1)$ we call W_a.

The size of layer i is called d_i, starting from input d_0.

The full product $W_N ... W_1$ we call W.

The data matrices are X, Y with X being input and Y the targets.

Model

We are considering the case of gradient flow with constant data matrices X, Y and the objective function

$$\min_{W_a}(L(W_a, X, Y)) = ||WX - Y||_F^2,$$

i.e. using Euclidean loss.

Model

We are considering the case of gradient flow with constant data matrices X, Y and the objective function

$$\min_{W_a}(L(W_a, X, Y)) = ||WX - Y||_F^2,$$

i.e. using Euclidean loss.

We will also assume that X and XY^T have full rank.

Previous Results

The overarching goal is to show that for almost all initializations, gradient flow will converge to a global minimum. What do we know?

Previous Results

The overarching goal is to show that for almost all initializations, gradient flow will converge to a global minimum. What do we know?

- Trajectories do converge to critical points. ([3], [2])

Previous Results

The overarching goal is to show that for almost all initializations, gradient flow will converge to a global minimum. What do we know?

- Trajectories do converge to critical points. ([3], [2])
- All local minima are global minima. ([3])

Previous Results

The overarching goal is to show that for almost all initializations, gradient flow will converge to a global minimum. What do we know?

- Trajectories do converge to critical points. ([3], [2])
- All local minima are global minima. ([3])
- Every other critical point is a saddle point. ([3])

Previous Results

The overarching goal is to show that for almost all initializations, gradient flow will converge to a global minimum. What do we know?

- Trajectories do converge to critical points. ([3], [2])
- All local minima are global minima. ([3])
- Every other critical point is a saddle point. ([3])
- All critical points that have a negative eigenvalue in their Hessian are avoided almost always. ([1])

What About Vanishing Hessian

A very particular class is not ruled out: critical points with vanishing Hessian. Our aim was to show that these too are avoided almost always.

What About Vanishing Hessian

A very particular class is not ruled out: critical points with vanishing Hessian. Our aim was to show that these too are avoided almost always. These critical points form an algebraic variety A_H, allowing us the use of algebro-geometric techniques.

What About Vanishing Hessian

A very particular class is not ruled out: critical points with vanishing Hessian. Our aim was to show that these too are avoided almost always. These critical points form an algebraic variety A_H, allowing us the use of algebro-geometric techniques.

This implies little on its own, but we also have algebraic invariants from the literature.

Invariants

From ([2]) we get the knowledge that the algebraic map

$$\delta : W_a \mapsto (W_N^T W_N - W_{N-1} W_{N-1}^T, \dots W_2^T W_2 - W_1 W_1^T)$$

is constant under gradient flow.

Invariants

From ([2]) we get the knowledge that the algebraic map

$$\delta : W_a \mapsto (W_N^T W_N - W_{N-1} W_{N-1}^T, \dots W_2^T W_2 - W_1 W_1^T)$$

is constant under gradient flow.
This map is continuous so it tells us that trajectories converging to a critical point must be in the same fiber of the δ-map as that critical point.

We say that the coordinate $W_{i+1}^T W_{i+1} - W_i W_i^T$ is called Δ_i.

The Objective

Conjecture

Main Conjecture: For all N and d_i we have

$$\dim(\overline{im(\delta_{|A_H})}) < \dim(\overline{im(\delta)}).$$

Hessian Variety

Straightforward differentiation of

$$L(\cdot, X, Y) = ||WX - Y||_F^2$$

and some work gets us A_H.

Hessian Variety

Straightforward differentiation of

$$L(\cdot, X, Y) = ||WX - Y||_F^2$$

and some work gets us A_H.

Theorem

Let X and Y be fixed matrices. W_a is a critical point of $L(\cdot, X, Y)$ with vanishing Hessian iff the following conditions hold:
1) There is a least integer n and largest integer $l > n$ such that

$$W_N \dots W_l = 0 \text{ and } W_n \dots W_1 X = 0.$$

2) For $k < j$ we have

$$W_{j-1} \dots W_{k+1} = 0 \text{ or } W_{k-1} \dots W_1 X Y^T W_N \dots W_{j+1} = 0.$$

We define the length of a condition as the number of parameter matrices in it.

N=2,N=3

Turns out very shallow networks are no problem.

N=2,N=3

Turns out very shallow networks are no problem.

- N=2: A_H consists of only the point $W_a = 0$. The main conjecture is trivially true in this case.

N=2,N=3

Turns out very shallow networks are no problem.

- N=2: A_H consists of only the point $W_a = 0$. The main conjecture is trivially true in this case.
- N=3: Among the conditions defining A_H are $W_3 = 0, W_2 = 0$, making this case as trivial as the last.

N=2,N=3

Turns out very shallow networks are no problem.

- N=2: A_H consists of only the point $W_a = 0$. The main conjecture is trivially true in this case.
- N=3: Among the conditions defining A_H are $W_3 = 0, W_2 = 0$, making this case as trivial as the last.
- N=4: Already much more difficult.

N=4

Does the size of the image of δ and A_H force a restriction?

N=4

Does the size of the image of δ and A_H force a restriction?
No. Not in general. We will see that this problem only gets worse.

N=4

Does the size of the image of δ and A_H force a restriction?
No. Not in general. We will see that this problem only gets worse.

But we can prove results specific to this case. It turns out that one of two things are always true in A_H for $N = 4$:

$$W_i = W_j = 0, i \neq j$$

or

$$W_i = W_{i-1}W_{i-2} = 0 \text{ or } W_iW_{i-1} = W_{i-2} = 0.$$

Take care of these and $N = 4$ follows.

Two Zero Matrices

Remember:

$$\delta : W_a \mapsto (W_N^T W_N - W_{N-1} W_{N-1}^T, \ldots W_2^T W_2 - W_1 W_1^T)$$

Two Zero Matrices

Remember:

$$\delta : W_a \mapsto (W_N^T W_N - W_{N-1} W_{N-1}^T, \dots W_2^T W_2 - W_1 W_1^T)$$

Lemma

If A is a variety where $W_i = 0$ and $W_j = 0$ for some $i > j$ then

$$\dim(im(\delta_{|A})) < \dim(im(\delta)).$$

Two Zero Matrices

Remember:

$$\delta : W_a \mapsto (W_N^T W_N - W_{N-1} W_{N-1}^T, \dots W_2^T W_2 - W_1 W_1^T)$$

Lemma

If A is a variety where $W_i = 0$ and $W_j = 0$ for some $i > j$ then

$$\dim(im(\delta_{|A})) < \dim(im(\delta)).$$

The trace vanishes!

Two Zero Matrices

Remember:

$$\delta : W_a \mapsto (W_N^T W_N - W_{N-1} W_{N-1}^T, \dots W_2^T W_2 - W_1 W_1^T)$$

Lemma

If A is a variety where $W_i = 0$ and $W_j = 0$ for some $i > j$ then

$$\dim(im(\delta_{|A})) < \dim(im(\delta)).$$

The trace vanishes!
$\sum_{k=j}^{i-1} \mathrm{tr}(\Delta_k) = \mathrm{tr}(W_i^T W_i) - \mathrm{tr}(W_j W_j^T) = 0$

The Other Case

Lemma

Let A be a variety defined by $W_i = W_{i-1}W_{i-2} = 0$ or $W_iW_{i-1} = W_{i-2} = 0$. Then $\dim(im(\delta_{|A})) < \dim(im(\delta))$.

The Other Case

Lemma

Let A be a variety defined by $W_i = W_{i-1}W_{i-2} = 0$ or $W_iW_{i-1} = W_{i-2} = 0$. Then $\dim(im(\delta_{|A})) < \dim(im(\delta))$.

Turns out that locally on $\Delta_{i-1}, \Delta_{i-2}$ we get our drop in dimension.

N=4

The cases taken care of, we arrive at:

Theorem

The main conjecture holds for $N \leq 4$

N=4

The cases taken care of, we arrive at:

Theorem

The main conjecture holds for $N \leq 4$

If it is already this much more difficult, is there really a proof for arbitrary N? No, as we will see.

Long Subvariety

There is a major problem of conditions of A_H not contributing in a major way. Consider the following lemma:

Lemma

There exists a subvariety A_{long} of A_H defined only by conditions $W_i...W_j = 0$ of length at least $(N-2)/2$.

Lemma

For any $h \geq 0$ there exists an integer M such that for $N \geq M$ there exist $i_j, j = 1, 2, 3, 4$ such that the variety defined by the four conditions $W_i ... W_{i-h} = 0$ is a subvariety of the Hessian variety A_H.

Lemma

For any $h \geq 0$ there exists an integer M such that for $N \geq M$ there exist $i_j, j = 1, 2, 3, 4$ such that the variety defined by the four conditions $W_i ... W_{i-h} = 0$ is a subvariety of the Hessian variety A_H.

For deep networks, a variety defined by a few relatively weak conditions is a subvariety of A_H. If we bound the d_i, then the codimension of A_H is bounded for increasing N.
A_H grows in pace with the parameter space but $\dim(\mathrm{im}(\delta))$ does not in general.

The previous lemma can be used to construct a counter-example to the main conjecture, with a bit of help.

The previous lemma can be used to construct a counter-example to the main conjecture, with a bit of help.

Lemma

If Σ_1, Σ_2 are diagonal 2×2 matrices with distinct eigenvalues, then the map

$$f : O(2) \times O(2) \to Sym_2(\mathrm{R})$$

defined by

$$(U, V) \mapsto U\Sigma_1 U^{-1} - V\Sigma_2 V^{-1}$$

has 2-dimensional image and

$$\overline{im(f)}$$

is a subvariety of $Sym_2(\mathrm{R})$ defined by a constant trace equation.

The previous lemma can be used to construct a counter-example to the main conjecture, with a bit of help.

Lemma

If Σ_1, Σ_2 are diagonal 2×2 matrices with distinct eigenvalues, then the map

$$f : O(2) \times O(2) \to Sym_2(\mathrm{R})$$

defined by

$$(U, V) \mapsto U\Sigma_1 U^{-1} - V\Sigma_2 V^{-1}$$

has 2-dimensional image and

$$\overline{im(f)}$$

is a subvariety of $Sym_2(\mathrm{R})$ defined by a constant trace equation.

Why is this helpful?

Because of the δ-map. If we let $U_i \Sigma_i V_i$ be the SVD of W_i then

$$W_i^T W_i = V_i^{-1} \Sigma_i^T \Sigma_i V_i$$

and

$$W_i W_i^T = U \Sigma_i \Sigma_i^T U^{-1}.$$

So $\Delta_i = V_{i+1}^{-1} \Sigma_{i+1}^T \Sigma_{i+1} V_{i+1} - U_i \Sigma_i \Sigma_i^T U_i^{-1}$.

Because of the δ-map. If we let $U_i \Sigma_i V_i$ be the SVD of W_i then

$$W_i^T W_i = V_i^{-1} \Sigma_i^T \Sigma_i V_i$$

and

$$W_i W_i^T = U \Sigma_i \Sigma_i^T U^{-1}.$$

So $\Delta_i = V_{i+1}^{-1} \Sigma_{i+1}^T \Sigma_{i+1} V_{i+1} - U_i \Sigma_i \Sigma_i^T U_i^{-1}$.
For 2×2 Δ_i we know what the image of δ looks like by the above lemma if we keep all singular values constant and only vary the orthogonal matrices.

Counter-example, N=16

Armed with these insights we can choose $N = 16$, all $d_i = 2$. A subvariety A to A_H is $W_i W_{i-1} W_{i-2} = 0$ for $i = 16, 12, 8, 4$.

Counter-example, N=16

Armed with these insights we can choose $N = 16$, all $d_i = 2$. A subvariety A to A_H is $W_i W_{i-1} W_{i-2} = 0$ for $i = 16, 12, 8, 4$.

Turns out there is a subset $A' \subset A$ where all singular values are constant and $\overline{\mathrm{im}(\delta_{A'})}$ is defined by $\Delta_{i-1}, \Delta_{i-2}$ having constant determinant and trace, all other Δ_j having only constant trace.

Counter-example, N=16

Armed with these insights we can choose $N = 16$, all $d_i = 2$. A subvariety A to A_H is $W_i W_{i-1} W_{i-2} = 0$ for $i = 16, 12, 8, 4$.

Turns out there is a subset $A' \subset A$ where all singular values are constant and $\overline{\mathrm{im}(\delta_{A'})}$ is defined by $\Delta_{i-1}, \Delta_{i-2}$ having constant determinant and trace, all other Δ_j having only constant trace.

Unfreezing the singular values fills up the whole space, so

$$\dim(\overline{\mathrm{im}(\delta_{|A})}) = \dim(\overline{\mathrm{im}(\delta_{|A_H})}) = \dim(\overline{\mathrm{im}(\delta)}).$$

The main conjecture is false.

- Does this mean the overarching goal, to show that for almost all initializations gradient flow will converge to a global minimum, is unreachable? No. The main conjecture was simply too strong.

- Does this mean the overarching goal, to show that for almost all initializations gradient flow will converge to a global minimum, is unreachable? No. The main conjecture was simply too strong.
- The algebraic closure of gradient flow trajectories are in general of high dimension.

- Does this mean the overarching goal, to show that for almost all initializations gradient flow will converge to a global minimum, is unreachable? No. The main conjecture was simply too strong.
- The algebraic closure of gradient flow trajectories are in general of high dimension.
- The δ-map invariants may be the only general algebraic invariants that can be found, effectively closing this line of investigation for deep networks.

References

B. Bah, H. Rauhut, U. Terstiege, and M. Westdickenberg.

Learning deep linear neural networks: Riemannian gradient flows and convergence to global minimizers.

arXiv e-prints, page arXiv:1910.05505, Oct. 2019.

Y. Chitour, Z. Liao, and R. Couillet.

A Geometric Approach of Gradient Descent Algorithms in Neural Networks.

arXiv e-prints, page arXiv:1811.03568, Nov. 2018.

K. Kawaguchi.

Deep Learning without Poor Local Minima.

arXiv e-prints, page arXiv:1605.07110, May 2016.