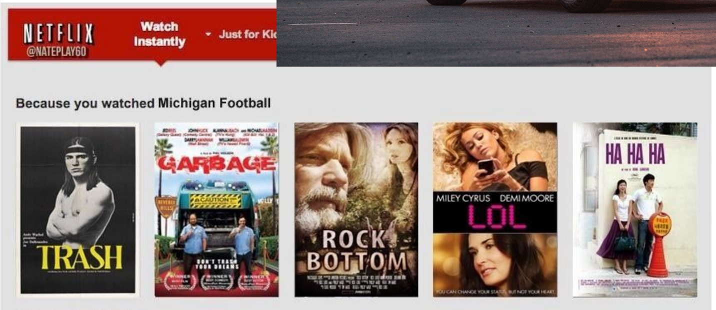
On the Quality of the Initial Basin in Overspecified Neural Networks

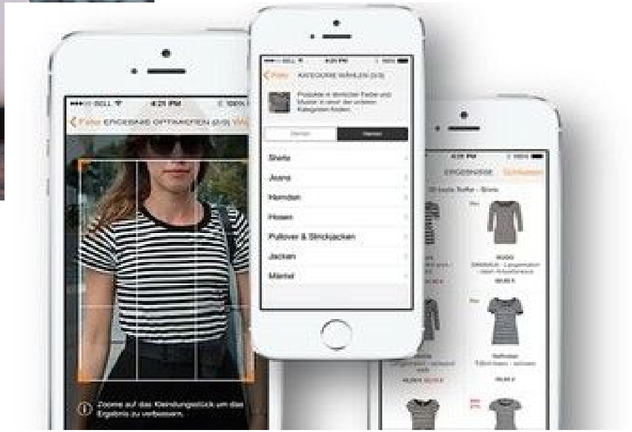
By Itay Safran et al., 2016

Apul Jain, Yunyang Xiong

Introduction

- Deep learning has achieved remarkable success
- Real world applications

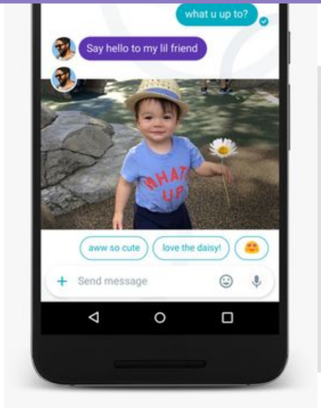


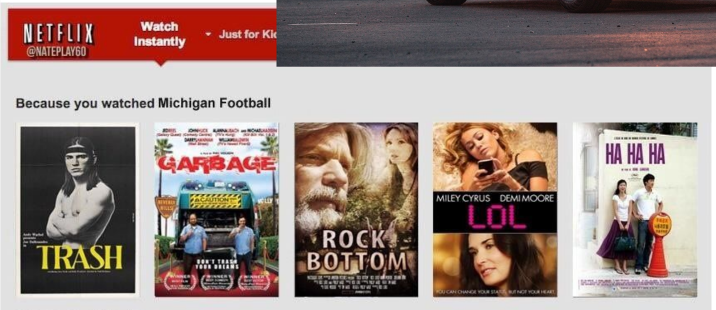


Introduction

- Deep learning has achieved remarkable success
- Real world applications

Little Theoretical explanation





Motivation

- Highly complex non-convex function with neural networks training
- In practice, it converges to a small minimal objective value in most cases

Initialization ← Random, not extreme

Regularization

Big data

This Paper: Main Idea

- Focus on **Random Initialization**
- Identify conditions s.t. with high probability:
  - Initializing at a random point from which there is a monotonically decreasing path to a global minimum
  - Initializing randomly at a **basin** with a small minimal loss value

This Paper: Two parts

<u>Part 1:</u> Focus on Initialization point

- Consider NN of arbitrary depth, weights are initialized at random ➜ random starting point in the parameter space
- Under **mild conditions** on loss function and data set, as size ⬆ we are more likely to begin at a point from which there is a continuous strictly monotonically decreasing path to a global minimum

This Paper: Two parts

Part 2: Focus on 2-layer ReLU with good basin

- 2-layer ReLU Network -- non-convex optimization problem
- Define a partition of the parameter space into convex regions (**basins**)
- Objective function has a relatively simple, ***basin-like structure***:
    - Every **local minima** of the objective function is **global**
    - All sublevel sets are connected, and in particular there is only a single connected set of minima, all global on that basin
- High prob. that a random initialization will land us at a basin with small minimum value*
- *Conditions: Low intrinsic data dimension, or a cluster structure

Notations

- **ReLU Network:** Computes $\mathbb{R}^d \to \mathbb{R}^k$
- Each neuron computes: $\mathbf{x} \mapsto \left[\mathbf{w}^\top \mathbf{x} + b\right]_+$ $\mathbf{w}$ is the weight vector and $\boldsymbol{b}$ is the bias while the ReLU activation function $[z]_+ = \max\{0, z\}$
- For a layer of *n* neurons, let $\mathbf{b} = (b_1, \ldots, b_n)$ and

$$W = \begin{pmatrix} \cdots & \mathbf{w}_1 & \cdots \\ & \vdots & \\ \cdots & \mathbf{w}_n & \cdots \end{pmatrix}$$

- We can define a layer of n neurons as: $\mathbf{x} \mapsto [W\mathbf{x} + \mathbf{b}]_+$

Notations

- Define the output of the network $N : \mathbb{R}^d \to \mathbb{R}^k$ over the set of weights $\mathcal{W}$ and an instance $\mathbf{x} \in \mathbb{R}^d$ by:

$$N(\mathcal{W})(\mathbf{x})$$

- Loss function:

$$L_S(N(\mathcal{W})) = \frac{1}{m} \sum_{t=1}^{m} \ell(N(\mathcal{W})(\mathbf{x}_t), \mathbf{y}_t)$$

Part 1

Focus on Initialization point and Path to Minima

Initialization scheme

Assumption

- The weights of every neuron are initialized independently
- The vector of each neuron’s weights (including bias) is drawn from a spherically symmetric distribution supported on non-zero vectors

Path to Global Minima

Recall Loss Function:

non-convex

convex

$$\underbrace{L(P(\mathcal{W}))}_{\text{non-convex}} = \frac{1}{m}\sum_{t=1}^{m} \underbrace{\ell\left(N\left(\mathcal{W}\right)\left(\mathbf{x}_t\right), \mathbf{y}_t\right)}_{\text{convex}}.$$

Example: L-2 Loss

$$L(P(\mathcal{W})) = \frac{1}{m}\sum_{t=1}^{m}(N(\mathcal{W})(\mathbf{x}_t) - y_t)^2.$$

Path to Global Minima

To Prove:

- If loss is convex in predictions, $\exists$ a **continuous path** in the parameter space $\mathcal{W}$ of multilayer networks (of any depth) which is:
  - Strictly **monotonically decreasing** in the objective value
  - Can reach an **arbitrarily small objective value**, including the global minimum

$$L(P(\mathcal{W})) = \frac{1}{m}\sum_{t=1}^{m}(N(\mathcal{W})(\mathbf{x}_t) - y_t)^2.$$

Path to Global Minima: **Theorem**

If...

- Suppose $L : \mathbb{R}^{m\times k} \to \mathbb{R}$ is convex, initialization point: $\mathcal{W}^{(0)}$, and $\exists$ a continuous path $\mathcal{W}^{(\lambda)}, \lambda \in [0,1]$ in the space of parameter vectors, starting from $\mathcal{W}^{(0)}$, and ending in $\mathcal{W}^{(1)}$ s.t. $(L(P(\mathcal{W}^{(1)})) < L(P(\mathcal{W}^{(0)})))$, and satisfies:
    - For some $\epsilon > 0$, and any $\lambda \in [0,1]$, $\exists\ c_\lambda \geq 0$ s.t.
      $$L(c_\lambda \cdot P(\mathcal{W}^{(\lambda)})) \geq L(P(\mathcal{W}^{(0)})) + \epsilon.$$
    - Initial point satisfies $L(P(\mathcal{W}^{(0)})) > L(\mathbf{0})$

Path to Global Minima: **Theorem**

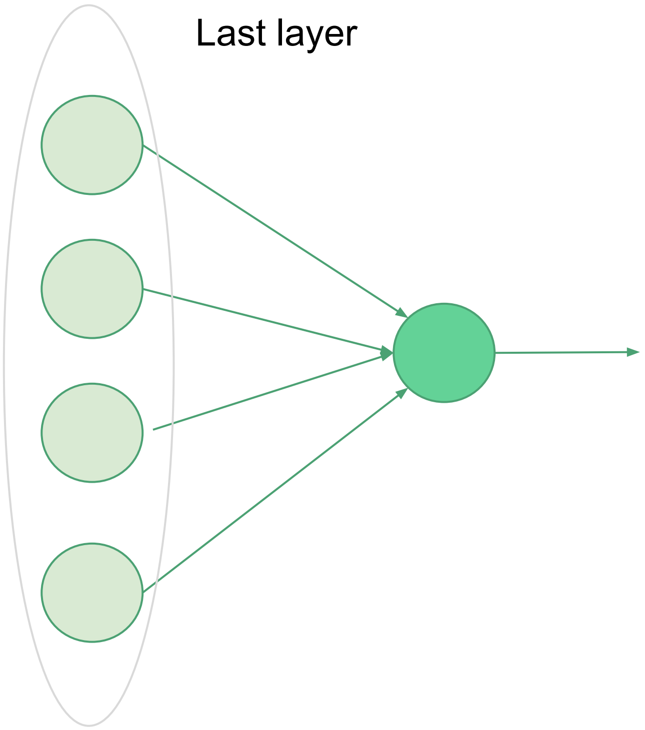
Then...

- $\exists$ a continuous path $\tilde{\mathcal{W}}^{(\lambda)}, \lambda \in [0,1]$ from the initial point $\tilde{\mathcal{W}}^{(0)} = \mathcal{W}^{(0)}$ to some point $\tilde{\mathcal{W}}^{(1)}$ satisfying $L(P(\tilde{\mathcal{W}}^{(1)})) = L(P(\mathcal{W}^{(1)}))$, along which $L(P(\tilde{\mathcal{W}}^{(\lambda)}))$ is strictly monotonically decreasing

Path to Global Minima: **Theorem**

Intuition:

- Linear dependence of output on last layer.
- Given the initial non-monotonic path $w^{(\lambda)}$, we rescale the last layer's parameters at each $w^{(\lambda)}$ by some positive factor c (λ) depending on λ (moving it closer or further from the origin), which changes its output and hence its objective value



Path to Global Minima: **Theorem**

Review: Two conditions:

- For some $\epsilon > 0$, and any $\lambda \in [0,1]$, $\exists\, c_\lambda \geq 0$ s.t.

$$L(c_\lambda \cdot P(\mathcal{W}^{(\lambda)})) \geq L(P(\mathcal{W}^{(0)})) + \epsilon.$$

Satisfied by losses which get very large far away from origin

- Initial point satisfies $L(P(\mathcal{W}^{(0)})) > L(\mathbf{0})$

Can be shown to hold with close to prob 1/2 for losses discussed earlier

$$\mathbb{P}_{\mathcal{W}^{(0)}}\left[L(P(\mathcal{W}^{(0)})) > L(\mathbf{0})\right] \geq \frac{1}{2}\left(1 - 2^{-n_{h-1}}\right)$$

Part 2

Focus on 2-layer ReLU: Initialize at “good” basin

2-layer ReLU Networks

- First layer parameter ***W*** with **n** neurons
- Output neuron parameter **v**
- Network is defined as: $N_n(W, \mathbf{v}) : \mathbb{R}^d \to \mathbb{R}$.
- Then Loss function is:

$$L_S(W, \mathbf{v}) := \frac{1}{m} \sum_{t=1}^{m} \ell\left(N_n(W, \mathbf{v})(\mathbf{x}_t), y_t\right) = \frac{1}{m} \sum_{t=1}^{m} \ell\left(\sum_{i=1}^{n} v_i \cdot [\langle \mathbf{w}_i, \mathbf{x}_t \rangle]_+, y_t\right)$$

Basin

Definition 1. *(Basin) A closed and convex subset B of our parameter space is called a basin if the following conditions hold:*

- *B is connected, and for all $\alpha \in \mathbb{R}$, the set $B_{\leq}\alpha = \{\mathcal{W} \in B : L_S(\mathcal{W}) \leq \alpha\}$ is connected.*
- *If $\mathcal{W} \in B$ is a local minimum of L_S on B, then it is a global minimum of L_S on B.*

We define the basin value $\mathrm{Bas}(B)$ of a basin B as the minimal value[2] attained:

$$\mathrm{Bas}(B) := \min_{\mathcal{W} \in B} L_S(\mathcal{W}).$$

2-layer ReLU Basin Partition

Observation:

1. Partition parameter space s.t. $\operatorname{sign}(\langle \mathbf{w}_i, \mathbf{x}_t \rangle)$ and $\operatorname{sign}(v_i)$ are fixed

2. The objective function becomes: $\frac{1}{m}\sum_{t=1}^{m} \ell\left(\sum_{i \in I_t} v_i \langle \mathbf{w}_i, \mathbf{x}_t \rangle, y_t\right)$ for some index set $I_1, \ldots, I_m \subseteq [n]$.

Defines a Basin!

2-layer ReLU Basin Partition

Formally...

Definition 2. *(Basin Partition) For any* $A \in \{-1,+1\}^{n\times d}$ *and* $\mathbf{b} \in \{-1,+1\}^n$*, define* $B_S^{A,\mathbf{b}}$ *as the topological closure of a set of the form*

$$\{(W,v) : \forall t \in [m]\,, j \in [n]\,, \mathrm{sign}(\langle \mathbf{w}_j, \mathbf{x}_t\rangle) = a_{j,t}, \mathrm{sign}(v_j) = b_j\}\,.$$

2-layer ReLU - Bounding the basin value

Theorem 2. *For any n, let α denote the minimal objective value achievable with a width n two-layer network, with respect to a convex loss ℓ on a training set S where each $\mathbf{x}_t$ is a singleton. Then when initializing $(W, \mathbf{v}) \in \mathbb{R}^{n \times d} \times \mathbb{R}^n$ from a distribution satisfying Assumption 1, we have*

$$\mathbb{P}\left[Bas\left(W, \mathbf{v}\right) \leq \alpha\right] \geq 1 - 2d\left(\frac{3}{4}\right)^n.$$

2-layer ReLU - Power of Overspecification

Overspecified networks are better in terms of basin value

Lemma 2. *Let $N_n(W, \mathbf{v})$ denote a two-layer network of size n, and let*

$$(W, \mathbf{v}) = (\mathbf{w}_1, \ldots, \mathbf{w}_n, v_1, \ldots, v_n) \in \mathbb{R}^{nd+n}$$

be in the interior of some arbitrary basin. Then for any subset $I = (i_1, \ldots, i_k) \subseteq [n]$ we have

$$Bas\,(W, \mathbf{v}) \leq Bas\,(\mathbf{w}_{i_1}, \ldots, \mathbf{w}_{i_k}, v_{i_1}, \ldots, v_{i_k})\,.$$

Where the right hand side is with respect to an architecture of size k.

Can we guarantee more?

Consider special cases:

- Data with Low Intrinsic Dimension
- Clustered or Full-rank Data

Data With Low Intrinsic Dimension

- Large enough amount of overspecification, with high probability, the output will attain global minimum.
    - Intuitively, it is overfitting with overspecification.
    - Exponential number of neurons required.

Data With Low Intrinsic Dimension

Theorem 3. *Assume each training instance* $\mathbf{x}_t$ *satisfies* $\|\mathbf{x}_t\| \leq 1$. *Suppose that the training objective* L_S *refers to the average squared loss, and that* $L_S(W^*, \mathbf{v}^*) = 0$ *for some* $(W^*, \mathbf{v}^*) \in \mathbb{R}^{n \times d} \times \mathbb{R}^n$ *satisfying*

$$|v_i^*| \cdot \|\mathbf{w}_i^*\| \leq B \;\; \forall i \in [n],$$

where B *is some constant. For all* $\epsilon > 0$, *if*

$$p_\epsilon = \frac{1}{2\pi\left(\mathit{rank}(X) - 1\right)} \left(\frac{\sqrt{\epsilon}}{nB}\sqrt{1 - \frac{\epsilon}{4n^2B^2}}\right)^{\mathit{rank}(X)-1} = \Omega\left(\left(\frac{\sqrt{\epsilon}}{nB}\right)^{\mathit{rank}(X)}\right),$$

rank(X) should be modest

and we initialize a two-layer, width $c\lceil \frac{n}{p_\epsilon} \rceil$ *network (for some* $c \geq 2$*), using a distribution satisfying Assumption 1, then*

$$\mathbb{P}\left[\mathit{Bas}(W, \mathbf{v}) \leq \epsilon\right] \geq 1 - e^{-\frac{1}{4}cn}.$$

Full-rank Data

- Training data comprise of k relatively small clusters.
  - Number of training examples m is less than dimension d, overfitting
  - m > d, small clusters have a similar structure with low dimension data

Theorem 4. *Assume rank* $(X) = m$, *and let the target outputs* $y_1, \ldots, y_m$ *be arbitrary. For any* n, *let* α *be the minimal objective value achievable with a width* n *two-layer network. Then if* $(W, \mathbf{v}) \in \mathbb{R}^{n \times d} \times \mathbb{R}^n$ *is initialized according to Assumption 1,*

$$\mathbb{P}\left[Bas\left(W, \mathbf{v}\right) \leq \alpha\right] \geq 1 - m\left(\frac{3}{4}\right)^n.$$

overfitting

Full-rank Data

Theorem 5. *Consider the squared loss, and suppose our data is clustered into $k \leq d$ clusters. Specifically, we assume there are cluster centers $\mathbf{c}_1, \ldots, \mathbf{c}_k \in \mathbb{R}^d$ for which the training data $S = \{\mathbf{x}_t, y_t\}_{t=1}^m$ satisfies the following:*

- $\exists \delta_1, \ldots, \delta_k > 0$ *s.t. for all* $\mathbf{x}_t$*, there is a unique* $j \in [k]$ *such that* $\|\mathbf{c}_j - \mathbf{x}_t\| \leq \delta_j$.
- $\forall j \in [k]\ \ \frac{\delta_j}{\|\mathbf{c}_j\|} \leq 2 \sin\left(\frac{\sqrt{2\pi}}{16d\sqrt{d}}\right)$ *and* $\forall j \in [k]\ \ \|\mathbf{c}_j\| \geq c$ *for some* $c > 0$.
- $\forall t \in [m]\ \ \|\mathbf{x}_t\| \leq B$ *for some* $B \in \mathbb{R}$.
- *For some fixed* γ*, it holds that* $|y_t - y_{t'}| \leq \gamma \|\mathbf{x}_t - \mathbf{x}_{t'}\|_2$ *for any* $t, t' \in [m]$ *such that* $\mathbf{x}_t, \mathbf{x}_{t'}$ *are in the same cluster.*

Let $\delta = \max_j \delta_j$*. Denote as* C *the matrix which rows are* $\mathbf{c}_1, \ldots, \mathbf{c}_k$*, and let* $\sigma_{\max}(C^\top), \sigma_{\min}(C^\top)$ *denote the largest and smallest singular values of* $C^\top$ *respectively. Let* $\mathsf{c}(\mathbf{x}_t) : \mathbb{R}^d \to \mathbb{R}^d$ *denote the mapping of* $\mathbf{x}_t$ *to its nearest cluster center* $\mathbf{c}_j$ *(assumed to be unique), and finally, let* $\hat{\mathbf{y}} = (\hat{y}_1, \ldots, \hat{y}_k) \in \mathbb{R}^k$ *denote the target values of arbitrary instances from each of the* k *clusters. Then if* $(W, \mathbf{v}) \in \mathbb{R}^{n \times d} \times \mathbb{R}^n$ *is initialized from a distribution satisfying Assumption 1,*

$$\mathbb{P}\left[Bas(W, \mathbf{v}) \leq \mathcal{O}\left(\delta^2\right)\right] \geq 1 - d\left(\frac{7}{8}\right)^n$$

clustering

Summary

- Focus on Initial weight vector initialization point
- Analysis for 2-layer networks with ReLU

Limitations:

- Doesn’t consider more general network such as multi-layer networks
- Doesn’t guarantee that Stochastic Gradient Descent will necessarily find the global minimum along the monotonically decreasing path

Q&A

Thanks!

Appendix

Path to Minima: **Theorem Proof**

For any $\lambda \in [-1, 2]$, define

$$v^{(\lambda)} = \begin{cases} L(P(\mathcal{W}^{(0)})) - \frac{\lambda}{2}\epsilon & \lambda \in [-1, 0] \\ \left(1 - \frac{\lambda}{3}\right) \cdot L(P(\mathcal{W}^{(0)})) + \frac{\lambda}{3} \cdot \max\{L(\mathbf{0}), L(P(\mathcal{W}^{(1)}))\} & \lambda \in [0, 2]. \end{cases}$$

Note: It's monotonic in λ

$$L(P(\mathcal{W}^{(0)})) + \epsilon > v^{(-1)} > v^{(0)} = L(P(\mathcal{W}^{(0)})) > v^{(2)} > \max\{L(\mathbf{0}), L(P(\mathcal{W}^{(1)}))\}$$

Path to Minima: **Theorem Proof**

By assumption for any $\lambda \in [0,1]$ $\exists$ $c^{(\lambda)}$ s.t. $L(c^{(\lambda)} \cdot P(\mathcal{W}^{(\lambda)})) \geq L(P(\mathcal{W}^{(0)})) + \epsilon.$

We have: $L(c^{\mathrm{clip}(\lambda)} \cdot P(\mathcal{W}^{\mathrm{clip}(\lambda)})) > v^{(\lambda)}$, $\mathrm{clip}(\lambda) = \min\{1, \max\{0, \lambda\}\}$ (1)

$L(0 \cdot P(\mathcal{W}^{\mathrm{clip}(\lambda)})) = L(\mathbf{0}) < v^{(\lambda)}$ $\lambda \in [-1, 2]$. (2)

Since L is convex and continuous, using (1) and (2) and IVT (Intermediate Val Th)

$$\boxed{\forall \lambda \in [-1, 2], \ \exists \, \tilde{c}^{(\lambda)} \in (0, c^{\mathrm{clip}(\lambda)}) \text{ such that } L(\tilde{c}^{(\lambda)} \cdot P(\mathcal{W}^{\mathrm{clip}(\lambda)})) = v^{(\lambda)}.}$$

$\tilde{c}^{(\lambda)}$ is unique

Path to Minima: **Theorem Proof**

At $\lambda = 0, L(\tilde{c}^{(0)} \cdot P(\mathcal{W}^{(0)})) = v^{(0)} = L(P(\mathcal{W}^{(0)}))$.

Based on the above observations, we have that $\tilde{c}^{(\lambda)}$, as a function of $\lambda \in [0, 1]$, is continuous, begins at $\tilde{c}_0 = 1$, and satisfies $L(\tilde{c}^{(\lambda)} \cdot P(\mathcal{W}^{(\lambda)})) = v^{(\lambda)}$. Moreover, $v^{(\lambda)}$ is strictly decreasing in λ. Therefore, letting

$$\{\tilde{\mathcal{W}}^{(\lambda)}, \lambda \in [0, 1]\} \tag{5}$$

defines the monotonically decreasing path.