Identity Matters in Deep Learning

Gene Li[1]

[1]Department of Electrical Engineering
Princeton University

April 10, 2018

Introduction

- Initialization of weights "near-zero" can be problematic
- Linear residual networks: layers represented by $x + h(x)$.
- Zero parameterization represents the identity transformation!

Deep Linear Residual Networks - Setup

- $R : \mathbb{R}^d \to \mathbb{R}^d$ from measurements $y = Rx + \zeta$, where $x, y \in \mathbb{R}^d$ and $\zeta \sim \mathcal{N}(0, I_d)$.
- Define $\mathcal{D}$ the distribution of x, with covariance $\Sigma = \mathbb{E}_{x\sim\mathcal{D}}[xx^T]$.
- Max of spectral norms:

$$|||A||| := \max_{i \in [l]} ||A_i||. \tag{1}$$

- Population Risk:

$$f(A) := \mathbb{E}||y - (I + A_l)...(I + A_1)x||^2. \tag{2}$$

How do we solve this problem?

- Solve the Least Squares:

$$\min_{A} \mathbb{E}||y - Ax||^2 \tag{3}$$

- Overparameterize:

$$\min_{A1,\ldots,A_l} \mathbb{E}||y - (I + A_l)\ldots(I + A_1)x||^2, \tag{4}$$

Results

Theorem

Let l be the number of layers. Under some conditions, there exists $A^* = A_1, \dots A_l$ such A^* minimizes population risk $f(\cdot)$ and

$$|||A^*||| \leq O(\frac{1}{l})$$

Theorem

The overparameterized optimization problem for any $\tau < 1$:

$$\min_{A1,\dots,A_l} \mathbb{E}||y - (I + A_l)\dots(I + A_1)x||^2$$
$$\text{s.t. } |||A||| \leq \tau.$$

has the property that all critical points of the objective function $f(\cdot)$ are global minimum.

Results

- Provided that the spectral norms of your A_i are bounded...
- An algorithm like gradient descent is able to reach the global minimum, if your iterates fall within this region.
- Even though the function is nonconvex, it still has a nice landscape!

Deep Nonlinear Residual Networks - What can we say?

- dataset of n training examples: $\{(x^{(i)}, y^{(i)})\}_{i\in[n]}$. Our data is $x^{(i)} \in \mathbb{R}^d$, labels $y^{(i)} \in \mathbb{R}^r$ encoded as one-hot basis vectors $e_1, ..., e_r$.
- Building block: $\mathcal{T}_{U,V,s}(\cdot) : \mathbb{R}^k \to \mathbb{R}^k$, where $\mathcal{T}_{U,V,s}(h) = V\text{ReLU}(Uh + s)$.

Result

- We would like to construct a deep residual network with blocks of the form $x + \mathcal{T}_{U,V,s}(x)$.

Theorem

Assume that for all $i, j \in [n], i \neq j$, we have $||x^{(i)} - x^{(j)}||^2 \geq \rho$ for some small constant $\rho \geq 0$. Then, there exists a residual network N with $O(n \log n + r^2)$ parameters that expresses the training data: N maps each $x^{(i)}$ to $y^{(i)}$.

Construction

- In the first layer, we map our input x to $h_0 = A_0 x$, where $A_0 \in \mathbb{R}^{kxd}$. This A_0 is taken to be a random matrix.
- The middle "residual" layers all take the form of:

$$h_j = h_{j-1} + \mathcal{T}_{A_j,B_j,b_j}(h_{j-1}), \forall j \in [l].$$

- The last layer maps h_l to the predicted y value: i.e.:

$$\hat{y} = \mathcal{T}_{A_{l+1},B_{l+1},b_{l+1}}(h_l), \forall j \in [l].$$

  Here we have to map $h_l \to \mathbb{R}^r$, so we need $A \in \mathbb{R}^{rxk}$, $b \in \mathbb{R}^r$, $B \in \mathbb{R}^{kxk}$.
- Take $k \sim O(\log n)$, $l = \lceil n/k \rceil$.

Summary

- The optimization landscape for deep linear residual networks, while nonconvex, is still "nice".
- Deep nonlinear residual networks are able to express the training data perfectly.
- Outlook
  - Extending optimization landscape result to nonlinear networks.
  - Can we use simple architectures to achieve better results?

Bibliography

Moritz Hardt and Tengyu Ma.
Identity matters in deep learning.
CoRR, abs/1611.04231, 2016.