Advanced Machine Learning
Project Report

Si Kai Lee
sl3950

Introduction

I am interested in the theoretical properties of deep learning methods as even though we get excellent results from them, we do not really understand how they work. For the project, I looked at Kawaguchi's Deep Learning without Poor Local Minima [1], Ge et al.'s Matrix Completion has No Spurious Local Minimum [2] and Hardt and Ma's Identity Maters in Deep Learning [3]. In this project, I prove that the local minima of $\frac{1}{2}||W_3W_2W_1X - Y||_F^2$ are global minima using the frameworks developed in [3].

Related Work

Kawaguchi showed in [1], with a set of assumptions and under certain conditions, that the following are true in both deep linear and non-linear networks:

1. The squared-loss function is non-convex and non-concave
2. Every local minimum is a global minimum
3. Every critical point that is not a global minimum is a saddle point
4. There exist "bad" saddle points (where the Hessian has no negative eigenvalue) for the deeper networks (with more than three layers), whereas there are no "bad" saddle points for the shallow networks (with at most three layers)

Ge et. al [2] proved that the commonly used non-convex objective function for positive semidefinite matrix completion $f(x) = \frac{1}{2}\sum_{(i,j)\in\Omega}[M_{i,j} - (XX^T)_{i,j}]^2$ where $\Omega = \{(i,j) : M_{i,j} \text{ is observed}\}$ has only global minima. They demonstrate an example of a proof technique that aims to be simple enough to be generalisable which might be useful for proving more properties of deep learning techniques.

The authors of [3] demonstrated through a simple proof that arbitrarily deep linear residual networks have no spurious which aids in the argument for the use of identity parameterisation in residual networks. Hardt and Ma also showed that residual networks with ReLu activations can represent any function of a sample of size n as long as the model has more than n parameters. This paper is of particular interest to me as I implemented the All-CNN model last semester for Neural Networks and Deep Learning and their simplified model beats it handily.

A Three Layer Neural Network has only Global Minima

Assumptions and Definitions

Under the framework described by Hardt and Ma [3], we have the following set of assumptions and definitions:

- $y = Rx + \epsilon$ where $R : \mathbb{R}^d \to \mathbb{R}^d$ linear transformation, $\epsilon \in \mathcal{N}(0, I_d)$ gaussian noise
- x is the input distributed according to $\mathcal{D}$, $\Sigma = \mathbb{E}_{x \sim \mathcal{D}}[xx^T]$ the covariance matrix
- Model has 3 layers: $A_3, A_2, A_1 \in \mathbb{R}^{d \times d}$ and $\hat{y} = A_3 A_2 A_1 x$
- Minimise $f(A(x, y)) = \frac{1}{2}\mathbb{E}_{x \sim \mathcal{D}}[||A_3 A_2 A_1 x - y||_F^2] = \frac{1}{2}\mathbb{E}[||A_3 A_2 A_1 x - Rx - \epsilon||_F^2]$

Proof

We start by showing that a global minimum exists in the case where R is a positive semidefinite matrix. If R is that, it can be diagonalised by an orthonormal matrix U where $R = UZU^T$ and $Z = diag(z_1, ..., z_d)$. Then by setting $A_3 = A_2 = A_1 = UZ^{1/3}U^T$, we have $A_3 A_2 A_1 = (UZ^{1/3}U^T)^3 = UZ^{(1/3)3}U^T = UZU^T$. Hence any set of A_3, A_2, A_1 that fulfils $A_3 A_2 A_1 - R = 0$ is the global minimum.

In the case where R is not positive semidefinite, we still would be able to find a global minimum. Start by taking the SVD of R which yields UZV^T. Since the decomposition yields 3 $d \times d$ matrices, we can set $A_3 = U$, $A_2 = Z$ and $A_1 = V^T$ which we can use to reconstruct R. Thus any set of A_3, A_2, A_1 that fulfils $A_3 A_2 A_1 - R = 0$ is still the global minimum for non-positive semidefinite matrices.

At this point, we let $E = A_3A_2A_1 - R$ and work through the proof of Claim 2.3 in [3]

$$\begin{aligned}
f(A) &= \frac{1}{2}\mathbb{E}[||A_3A_2A_1x - Rx - \epsilon||_F^2] \\
&= \frac{1}{2}\mathbb{E}[||Ex - \epsilon||_F^2] \\
&= \frac{1}{2}\mathbb{E}[||Ex||_F^2 - 2\langle Ex, \epsilon\rangle + ||\epsilon||_F^2] \\
&= \frac{1}{2}\mathbb{E}[||Ex||_F^2 + ||\epsilon||_F^2] \text{ since } \langle \mathbb{E}[Ex]^T, \mathbb{E}[\epsilon]\rangle = 0 \text{ as } \mathbb{E}[\epsilon] = 0 \\
&= \frac{1}{2}\mathbb{E}[x^TE^TEx] + \frac{1}{2}\mathbb{E}[||\epsilon||_F^2] \\
&= \frac{1}{2}\mathbb{E}[tr(Exx^TE^T)] + C \\
&= \frac{1}{2}tr(E\Sigma E^T)] + C \\
&= \frac{1}{2}\langle \Sigma^{\frac{1}{2}T}E^TE\Sigma^{\frac{1}{2}}\rangle + C \\
&= \frac{1}{2}||E\Sigma^{\frac{1}{2}}||_F^2 + C
\end{aligned}$$

Hence the non-trivial minimum is C assuming $\Sigma, R \neq 0$.

We move to computing the gradient of $f(A)$ using Taylor expansion:

$$\begin{aligned}
f(A_1, A_2 + \Delta_2, A_3) &= ||(A_3(A_2 + \Delta_2)A_1 - R)\Sigma^{\frac{1}{2}}||_F^2 \\
&= ||(A_3A_2A_1 - R)\Sigma^{\frac{1}{2}} + (A_3\Delta_2A_1)\Sigma^{\frac{1}{2}}||_F^2 \\
&= ||(A_3A_2A_1 - R)\Sigma^{\frac{1}{2}}||_F^2 + 2\langle (A_3A_2A_1 - R)\Sigma^{\frac{1}{2}}, (A_3\Delta_2A_1)\Sigma^{\frac{1}{2}}\rangle + O(||\Delta_2||_F^2) \\
&= ||(A_3A_2A_1 - R)\Sigma^{\frac{1}{2}}||_F^2 + 2\langle (A_3\Delta_2A_1)\Sigma^{\frac{1}{2}}, E\Sigma^{\frac{1}{2}}\rangle + O(||\Delta_2||_F^2) \\
&= ||(A_3A_2A_1 - R)\Sigma^{\frac{1}{2}}||_F^2 + 2(\Sigma^{\frac{1}{2}T}A_1^T\Delta_2^TA_3^TE\Sigma^{\frac{1}{2}}) + O(||\Delta_2||_F^2) \\
&= ||(A_3A_2A_1 - R)\Sigma^{\frac{1}{2}}||_F^2 + 2tr(\Delta_2^TA_3^TE\Sigma A_1^T) + O(||\Delta_2||_F^2) \\
&= ||(A_3A_2A_1 - R)\Sigma^{\frac{1}{2}}||_F^2 + 2\langle \Delta_2, A_3^TE\Sigma A_1^T\rangle + O(||\Delta_2||_F^2) \\
&= ||(A_3A_2A_1 - R)\Sigma^{\frac{1}{2}}||_F^2 + 2\langle A_3^TE\Sigma A_1^T, \Delta_2\rangle + O(||\Delta_2||_F^2)
\end{aligned}$$

Hence the gradient $\frac{\partial f}{\partial A_2} = 2A_3^TE\Sigma A_1^T = 2A_3^T(A_3A_2A_1 - R)\Sigma A_1^T$ and this generalises to both A_3 and A_1.

For the gradient to be 0, E must be 0 as $A_3, A_2, A_1 \neq 0$ as $R \neq 0$ which implies any set of points As that satisfies $A_3A_2A_1 = R$ are both local and global minima.

Discussion

From the above, one can see that the proof techniques developed by Hardt and Ma, without considering the tighter case of critical points being global minima, can provides a simpler proof than that seen in Kawaguchi [1] to show the local minima in 3 $d \times d$ layer linear networks are global minima. Note that R can be perfectly decomposed by SVD and the technique to calculate gradients are still valid if we relax the assumption the 3 layers are of dimensions $d \times d$ and allow layers of different dimensions. However, when dealing with linear networks with arbitrary number of layers, we have to assume that the layers are $d \times d$ dimension. In the case of having an odd number of layers, we could just split them with further SVDs. However, when requiring even numbers of layers, we have do more work. We split the first Z matrix derived from SVD as Z multiplied by an identity matrix and then run SVD as many times on U and V^T as needed to yield an even number of matrices (=layers). Hence the above proof technique generalises to $d \times d$ linear networks of arbitrary length and highlights the potential generalisability of "simple" proofs [2].

References

[1] Kenji Kawaguchi. Deep Learning without Poor Local Minima. *arXiv preprint arXiv:1605.07110*, 2016.

[2] Rong Ge, Jason D Lee, and Tengyu Ma. Matrix Completion has No Spurious Local Minimum. *arXiv preprint arXiv:1605.07272*, 2016.

[3] Moritz Hardt and Tengyu Ma. Identity Matters in Deep Learning. *arXiv preprint arXiv:1611.04231*, 2016.