Landscape Analysis of Overcomplete Tensor and Neural Collapse

Qing Qu

Dept. of EECS, University of Michigan

May 17, 2021
Outline of this Talk

• Introduction

• Overcomplete Tensor Decomposition (Representation Learning)

• Neural Collapse in Deep Network Training
Outline of this Talk

• Introduction

• Overcomplete Tensor Decomposition (Representation Learning)

• Neural Collapse in Deep Network Training
Nonconvex Problems in Representation Learning

\[ \min_x f(x), \text{ s.t. } x \in \mathbb{R}^n \]

Nonconvex landscape

Convex landscape
General Nonconvex Problems

Noncritical Point \((\nabla \varphi \neq 0)\)

Minimizer
\[\nabla^2 \varphi > 0\]

Saddle
\[\lambda_{\text{min}} \nabla^2 \varphi < 0\]
\[\lambda_{\text{max}} \nabla^2 \varphi > 0\]

Maximizer
\[\nabla^2 \varphi < 0\]

Critical Points \((\nabla \varphi = 0)\)
General Nonconvex Problems

\[
\min_x f(x), \quad \text{s.t. } x \in \mathbb{R}^n
\]

“bad” local minimizers  “flat” saddle points

local minima

global minima

“flat” saddle
In the worst case, even finding a local minimizer is NP-hard (Murty et al. 1987)
Optimizing Nonconvex Problems Globally

Benign nonconvex landscapes enable efficient global optimization!
Nonconvex Problems with Benign Landscape

- Generalized Phase Retrieval [Sun’18]
- Low-rank Matrix Recovery [Ma’16, Jin’17, Chi’19]
- (Convolutional) Sparse Dictionary Learning [Sun’16, Qu’20]
- (Orthogonal) Tensor Decomposition [Ge’15]
- Sparse Blind Deconvolution [Zhang’17, Li’18, Kuo’19]
- Deep Linear Network [Kawaguchi’16]
- ...
Outline of this Talk

• Introduction

• Overcomplete Tensor Decomposition (Representation Learning)

• Neural Collapse in Deep Network Training
Landscape Analysis of Overcomplete Learning

Q. Qu, Y. Zhai, X. Li, Y. Zhang, Z. Zhu, Analysis of optimization landscapes for overcomplete learning, *ICLR’20*, (oral, top 1.9%)

- Provide the **global landscape** for overcomplete representation learning problems.

- Explains why they can be **efficiently** optimized to global optimality
Overcomplete Tensor Decomposition

We consider decomposing a 4-th order tensor of rank $m$ in the following form

$$T = \sum_{i=1}^{m} a_i \otimes a_i \otimes a_i \otimes a_i, \quad a_i \in \mathbb{R}^n.$$ 

- Given $T$, our goal is to recover each component $a_i \in \mathbb{R}^n$.
- We are interested in the overcomplete regime that $m > n$.

Core problem for several unsupervised representation learning problems (ICA and mixture of Gaussian [Anandkumar’12], dictionary learning [Barak’14, Qu’20]), and even training neural networks [Ge’17].
Overcomplete Tensor Decomposition

A natural (nonconvex) objective to find one component

$$\min_{q} f(q) = - \sum_{i,j,k,\ell \in [m]^4} T_{i,j,k,\ell} q_i q_j q_k q_\ell = - \sum_{i=1}^{n} \langle a_i, q \rangle^4$$

s.t. $\|q\|_2 = 1$. 
Overcomplete Tensor Decomposition

Let $A = \begin{bmatrix} a_1 & \cdots & a_m \end{bmatrix}$, the problem can be written as

$$\min_{q} - \|A^\top q\|_4^4, \quad \text{s.t.} \quad \|q\|_2 = 1.$$ 

- When $m \leq n$, and $\{a_i\}_{i=1}^m$ are orthogonal, existing result [Ge’15] has shown that the function is a strict saddle function with benign optimization landscape, all global solutions are approximately $\{\pm a_i\}_{i=1}^m$.
- The analysis of orthogonal case cannot be generalized to overcomplete settings.
Overcomplete Tensor Decomposition

Let $A = [a_1 \cdots a_m]$, the problem can be written as

$$\min_q - \|A^\top q\|_4^4, \quad \text{s.t.} \quad \|q\|_2 = 1.$$ 

- For **overcomplete case**, most of existing landscape analysis results [Ge’17] are **local**, or are based on Sum-of-Squares relaxations [Barak’15, Ma’16] which is computationally expensive.
- Empirically, gradient descent or power method find the global solution **efficiently** even when $m \gg n$. 
A Global Result in Overcomplete Settings

\[
\min_q f(q) = -\|A^\top q\|_4^4, \text{ s.t. } \|q\|_2 = 1.
\]

**Theorem (Informal)** Suppose that (i) \(K = m/n\) is a constant, and (ii) \(A\) is near orthogonal with small \(\mu\). Then every critical point of \(f(q)\) is either

- a **strict saddle point** exhibits negative curvature;
- or close to a **target solution**: one column \(a_i\) of \(A\).
Assumptions on $A$ (Near Orthogonal)

- Row orthogonal: unit norm tight frame (UNTF)

$\sqrt{\frac{n}{m}} A A^\top = I, \quad \|a_i\|_2 = 1.$

- Incoherence of the columns (near orthogonal)

$\max_{i \neq j} |\langle a_i, a_j \rangle| \leq \mu.$
Relationship to Dictionary Learning

\[ Y \approx A X \]

Given \( Y = AX \in \mathbb{R}^{n \times p} \), jointly find overcomplete dictionary \( A \in \mathbb{R}^{n \times m} \) and sparse \( X \in \mathbb{R}^{m \times p} \).
Relationship to Dictionary Learning

We can find one column of $A$ via

$$\min_q f_{DL}(q) = -\|Y^\top q\|_4^4, \text{ s.t. } \|q\|_2 = 1.$$ 

The underlying reasoning is that, in expectation

$$\mathbb{E}_X \left[\|Y^\top q\|_4^4\right] = \mathbb{E}_X \left[\|X^\top A^\top q\|_4^4\right] = c_1 \|A^\top q\|_4^4 + c_2$$

for $X$ following some sparse zero-mean distributions (e.g., Bernoulli-Gaussian)
Relationship to Dictionary Learning

\[ \min_{q} f_{DL}(q) = -\left\| Y^\top q \right\|_4^4, \quad \text{s.t.} \quad \|q\|_2 = 1. \]

Theorem (Informal) Suppose that (i) \( K = m/n \) is a constant, (ii) \( A \) is near orthogonal, and (iii) \( p \geq \Omega(\text{poly}(n)) \). Then with high probability every critical point of \( f(q) \) is either

- a **strict saddle point** exhibits negative curvature;
- or close to a **target solution**: one column \( a_i \) of \( A \).
Relationship to Dictionary Learning

\[ \mathbf{A}_0 \in \mathbb{R}^{n \times m} \]

<table>
<thead>
<tr>
<th>Local</th>
<th>Global</th>
<th>Complete $n = m$</th>
<th>Overcomplete $m &gt; n$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>[Qu, Sun, Wright’16]</td>
<td>[Qu, Sun, Wright’16]</td>
</tr>
<tr>
<td>[Zhai et al.’19]</td>
<td>[Sun, Qu, Wright’16]</td>
<td>[Li et al.’18]</td>
<td>[Arora et al.’14&amp;15]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Initialization Required</td>
<td>[Agarwal et al.’16]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>[Chatterji et al.’17]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>[Awasthi et al.’18]</td>
</tr>
</tbody>
</table>
Relationship to Dictionary Learning

practice $m < n^2$
vs. theory $m < Cn$

recover full $A_0$ via repeated independent trials
Outline of this Talk

• Introduction

• Overcomplete Tensor Decomposition (Representation Learning)

• Neural Collapse in Deep Network Training
Understanding Deep Neural Networks


- Analyzes the **global landscape** of the training loss based on the **unconstrained feature model**

- Explains the ubiquity of **Neural Collapse** of the learned representations of the network
Understanding Deep Neural Networks

\[
\psi_\Theta(x) = W_L \sigma(W_{L-1} \cdots \sigma(W_1 x + b_1) + b_{L-1}) + b_L
\]

\[
\Theta := \{W_\ell, b_\ell\}_{\ell=1}^L \quad \sigma(\cdot): \text{nonlinear activations}
\]

weights, bias
Understanding Deep Neural Networks

\[ \min_{\Theta} \sum_{k=1}^{K} \sum_{i=1}^{n_k} \mathcal{L}_{CE}(\psi_{\Theta}(x_{k,i}), y_k) + \|\Theta\|^2_F \]

- \( i \)-th input in the \( k \)-th class
- One-hot vector for the \( k \)-th class
Fundamental Challenges: Optimization

Landscape in Classical Optimization (abundant algorithms & theory)

Landscape of Modern Deep Neural Networks Credited to [Li’17]
Optimization: Existing Results

Existing analysis are based on various simplifications:

• **Go Linear:** deep linear networks [Kawaguchi’16], deep matrix factorizations [Arora’19], etc.

• **Go Shallow:** Two-layer neural networks [Safran’18, Liang’18], etc.

• **Go Wide:** Neural tangent kernels [Jacot’18, Allen-Zhu’18, Du’19], mean-field analysis [Mei’19, Sirignano’19], etc.

Most of results *hardly* provide much insights for practical neural networks.
Features – What NNs (Conceptually) Designed to Learn

Wishful Design: NNs learn rich feature representations across different levels?
Neural Collapse in Classification

Prevalence of neural collapse during the terminal phase of deep learning training

Vardan Panyan, X. Y. Han, and David L. Donoho

See all authors and affiliations

PNAS October 6, 2020 117 (40) 24652-24663; first published September 21, 2020;
https://doi.org/10.1073/pnas.2015509117

Contributed by David L. Donoho, August 18, 2020 (sent for review July 22, 2020; reviewed by Helmut Boelscke and Stéphane Mallat)
Neural Collapse in Classification

\[
\psi_{\Theta}(x) = W_L \sigma(W_{L-1} \cdots \sigma(W_1 x + b_1) + b_{L-1}) + b_L
\]

Last-layer classifier

\[
\phi_{\theta}(x) = :h
\]

Last-layer feature

Data in the Input Space

Neural Collapse in the Feature Space

Simplex Equiangular Tight Frames (Simplex ETF)

\[
W_L = [\mu_1 \cdots \mu_K]^T
\]
Neural Collapse: Symmetry and Structures

Balanced training dataset with \( n = n_1 = n_2 = \cdots = n_K \), and
\[
W := W_L, \quad H := [h_{1,1} \; \cdots \; h_{K,n}].
\]

Neural Collapse (NC) means that

1) **Within-Class Variability Collapse on** \( H \): features of each class collapse to class-mean with zero variability;

2) **Convergence to Simplex ETF on** \( H \): the class means are linearly separable, and maximally distant;

3) **Convergence to Self-Duality** \((W,H)\): the last-layer classifiers are perfected matched with the class-means of features.

4) **Simple Decision Rule** via Nearest Class-Center decision.
Simplification: Unconstrained Features

\[ \psi_\Theta(x) = W_L \sigma(W_{L-1} \cdots \sigma(W_1 x + b_1) + b_{L-1}) + b_L \]

Last-layer classifier

\[ \phi_\theta(x) = h \]

Last-layer feature

Treat \( H = [h_{1,1} \cdots h_{K,n}] \) as a \textbf{free} optimization variable
Simplification: Unconstrained Features

\[ \psi_{\Theta}(x) = W_L \sigma(W_{L-1} \cdots \sigma(W_1 x + b_1) + b_{L-1}) + b_L \]

Treat \( H = [h_{1,1} \cdots h_{K,n}] \) as a free optimization variable

\[
\min_{W,H,b} \frac{1}{K n} \sum_{k=1}^{K} \sum_{i=1}^{n} L_{\text{CE}}(W h_{k,i} + b, y_k) + \frac{\lambda_W}{2} \|W\|_F^2 + \frac{\lambda_H}{2} \|H\|_F^2 + \frac{\lambda_b}{2} \|b\|_2^2
\]
Simplification: Unconstrained Features

\[ \psi_{\Theta}(x) = W_L \sigma(W_{L-1} \cdots \sigma(W_1 x + b_1) + b_{L-1}) + b_L \]

Last-layer classifier \[ \phi_{\Theta}(x) = h \]

Last-layer feature

Treat \( H = [h_{1,1} \cdots h_{K,n}] \) as a free optimization variable

\[
\min_{W,H,b} \frac{1}{Kn} \sum_{k=1}^{K} \sum_{i=1}^{n} \mathcal{L}_{CE}(W h_{k,i} + b, y_k) + \frac{\lambda W}{2} \|W\|_F^2 + \frac{\lambda H}{2} \|H\|_F^2 + \frac{\lambda b}{2} \|b\|_2^2
\]

• **Validity:** Modern network are highly **overparameterized**, that can approximate any point in the feature space [Shaham’18];

• **State-of-the-Art:** also called **Layer-Peeled Model** [Fang’21], existing work [E’20, Lu’20, Mixon’20, Fang’21] only studied global optimality conditions.
Main Theoretical Results

Theorem (Informal) Consider the nonconvex loss with unconstrained feature model with $K < d$ and balanced data

$$\min_{W,H,b} \frac{1}{Kn} \sum_{k=1}^{K} \sum_{i=1}^{n} \mathcal{L}_{CE}(W h_{k,i} + b, y_k) + \frac{\lambda_W}{2} \|W\|_F^2 + \frac{\lambda_H}{2} \|H\|_F^2 + \frac{\lambda_b}{2} \|b\|_2^2$$

- **(Global Optimality)** Any global solution $(W_*, H_*)$ satisfies the NC properties (1-4).
- **(Benign Global Landscape)** The function has no spurious local minimizer and is a strict saddle function, with negative curvature for non-global critical point.
Main Theoretical Results

Theorem (Informal) Consider the nonconvex loss with unconstrained feature model with $K < d$ and balanced data

- **(Global Optimality)** Any global solution $(W_*, H_*)$ satisfies the NC properties (1-4).
- **(Benign Global Landscape)** The function has no spurious local minimizer and is a strict saddle function, with negative curvature for nonglobal critical point.

**Message:** deep networks always learn Neural Collapse features and classifiers, provably
Experiment: NC is Algorithm Independent

CIFAR-10 Dataset, ResNet18, with different training algorithms

Measure of Within-Class Variability  Measure of Between-Class Separation  Measure of Self-Duality Collapse
Experiment: NC Occurs for Random Labels

CIFAR-10 Dataset, ResNet18, random labels with varying network width

Measure of Within-Class Variability  Measure of Between-Class Separation  Measure of Self-Duality Collapse

Validity of Unconstrained Feature Model: Learned last-layer features and classifiers seems to be independent of input!
Implications for Practical Network Training

**Observation:** For NC features, when $K \leq d$ the best classifier is given by the Simplex ETF

\[ W_\star = [\mu_1 \cdots \mu_K]^\top. \]
Implications for Practical Network Training

**Observation:** For NC features, when $K \leq d$ the best classifier is given by the Simplex ETF

$$W_\star = [\mu_1 \cdots \mu_K]^\top.$$  

- **Implication 1:** No need to learn the classifier
  - Just fix them as a Simplex ETF
  - Save 8%, 12%, and 53% parameters for ResNet50, DenseNet169, and ShuffleNet!
Implications for Practical Network Training

Observation: For NC features, when $K \leq d$ the best classifier is given by the Simplex ETF

$$W_* = [\mu_1 \ldots \mu_K]^\top.$$

• **Implication 1:** No need to learn the classifier
  - Just fix them as a Simplex ETF
  - Save 8%, 12%, and 53% parameters for ResNet50, DenseNet169, and ShuffleNet!

• **Implication 2:** No need of large feature dimension $d$
  - Just use feature dim $d = \#\text{class } K$ (e.g., $d=10$ for CIFAR10)
  - Further saves 21% and 4.5% parameters for ResNet18 and ResNet50!
Experiment: Fixed Classifier with $d = K$

ResNet50, CIFAR10, Comparison of Learned vs. Fixed Classifiers of $W$

Measure of Between-Class Separation

Training Accuracy

Testing Accuracy

Training with fixed last-layer classifiers achieves on-par performance with learned classifiers.
Summary and Discussion


- Through landscape analysis under unconstrained feature model, we provide a complete characterization of learned representation of deep networks.

- The understandings of learned representations could shed lights on generalization, robustness, and transferability.
Outline of this Talk

• Introduction

• Overcomplete Tensor Decomposition (Representation Learning)

• Neural Collapse in Deep Network Training
Thank You!