The Success of Deep Learning

computer vision
(Credit: Appen. (2019))

natural language processing
(Credit: Andrey Suslov (2023))

gameplay
(Credit: AlphaGo)

autonomous driving
(Credit: Phil Brown (2019))
The Trend of Large Models...

![Graph showing the trend of large models](image)

**Figure:** Accuracy vs. model size for image classification on ImageNet dataset

In principle, deep network can fit *any* training labels! (i.e., not only clean, but also corrupted labels)
The Challenges & Opportunities in Large Models...

- Tremendous cost of computation
- Difficult to interpret
- Vulnerable to data corruptions

Figure: Accuracy vs. model size for image classification on ImageNet dataset
The Challenges & Opportunities in Large Models...

- Tremendous cost of computation
- Difficult to interpret
- Vulnerable to data corruptions

Figure: Accuracy vs. model size for image classification on ImageNet dataset
The Challenges & Opportunities in Large Models...

- Tremendous cost of computation
- Difficult to interpret
- Vulnerable to data corruptions

Theory and principles behind its success?

Figure: Accuracy vs. model size for image classification on ImageNet dataset
Low-Dimensional Structures Are Largely Ignored...

Low-Complexity Structures

\[ Y = A_0 \cdot X_0 \]

Generative Models

\[ \min_{W=[A,X]} \varphi(Y,W) \]

Optimization
Low-Dimensional Structures Are Largely Ignored...

- **Sparse Recovery**  
  [Donoho’06, Candes’08]

- **Low-rank Matrix Recovery**  
  [Candes’08, Recht’11, Candes’11]

- **(Sparse) Phase Retrieval**  
  [Candes’13, Shechtman’15]

- **Super-resolution**  
  [Candes’14, Fernandez-Granda’16]

- **(Sparse) Blind Deconvolution**  
  [Ahmed’14, Zhang’17, Kuo’20]

- **(Convolutional) Dictionary Learning**  
  [Aharon’06, Sun’16, Bristow’13, Papyan’17]
The Emergence of Low-Dim Models in Deep Learning

Network Architectures

[Gregor’10, Liu’18, Sulam’18, Papyan’18, Monga’19]

- image credited to Monga et al., Yu et al. & Azizan et al.

Representations

[Pennington’17, Bansal’18, Xiao’18, Wang’20, Ye’20, Qi’20, Han’20, Zhu’21, Fang’21]

Regularizations & Generalization

[Neyshabur’17, Mianjy’18, Ulyanov’18, Gidel’19, Arora’19, Belkin’19, Nakkiran’19, Yang’20]
Outline of Today’s Course

**Lec.1** Low-dimensional Models & Noconvex Optimization (1hrs)

**Lec.2** Low-dimensional Representations in Deep Learning I: Neural Collapse (1hrs)

**Lec.3** Low-dimensional Representations in Deep Learning II: Law-of-Parsimony in GD (1.5hrs)

**Lec.4** Low-dimensional Models for Robust Learning (0.5hrs)
Lec.1 Low-dimensional Models & Noconvex Optimization (1hrs)
Outline of Today’s Course

Credit: Han et al.


Lec.2 Low-dimensional Representations in Deep Learning I: Neural Collapse (1hrs)
Lec.3 Low-dimensional Representations in Deep Learning II: Law-of-Parsimony in GD (1.5hrs)
Outline of Today’s Course

\[ y_i = f(x_i; \Theta^*) + s_i \]

noisy label
input params.
sparse label noise

Exact Separation of Sparse Corruption with Incoherence between Data and Noise

Lec.4 Low-dimensional Models for Robust Learning (0.5hrs)
Outline

1 Introduction

2 Symmetry & Geometry for Nonconvex Problems in Practice
   Problems with Rotational Symmetry
   Problems with Discrete Symmetry

3 Efficient Nonconvex Optimization

4 Conclusion
Most of the Machine Learning Problems are Nonconvex...

**Figure**: Convex vs. Nonconvex Optimization Problems.
Basic Calculus

Critical points or stationary points: gradient vanishes

- **convex function**: critical point = minimizer
- **nonconvex function**: not all critical points are minimizers
## Basic Calculus

Critical points with non-singular hessian

- **local minimizer:** hessian is positive definite
- **saddle points:** hessian has both positive and negative eigenvalues
- **local maximizer:** hessian is negative definite

![Noncritical Point ($\nabla \varphi \neq 0$) vs. Critical Points ($\nabla \varphi = 0$)]

- **Minimizer**
  - $\nabla^2 \varphi > 0$
- **Saddle**
  - $\lambda_{\text{min}} \nabla^2 \varphi < 0$
  - $\lambda_{\text{max}} \nabla^2 \varphi > 0$
- **Maximizer**
  - $\nabla^2 \varphi < 0$
Challenges of Nonconvex Optimization – Pessimistic Views

Consider the problem of minimizing a general nonlinear function:

\[
\min_z \varphi(z), \quad z \in C. \quad (1)
\]

In the worst case, even finding a local minimizer can be NP-hard\(^1\).

---

\(^1\)Some NP-complete problems in quadratic and nonlinear programming, K.G Murty and S. N. Kabadi, 1987
Challenges of Nonconvex Optimization – Pessimistic Views

Consider the problem of minimizing a general nonlinear function:

$$\min_{z} \varphi(z), \quad z \in C. \quad (1)$$

In the worst case, even finding a local minimizer can be NP-hard\(^1\).

Hence, typically people seek to work with mild guarantees for nonconvex problems:

1. convergence to some critical point $\bar{z}$ such that $\nabla \varphi(\bar{z}) = 0$;
2. or convergence to some local minimizer $\nabla^2 \varphi(\bar{z}) \succeq 0$.

\(^1\)Some NP-complete problems in quadratic and nonlinear programming, K.G Murty and S. N. Kabadi, 1987
Benign Nonconvex Optimization Landscape

General Case

Structured Case
Benign Nonconvex Optimization Landscape

General nonconvex problems

Our training problem

General Case

Structured Case
Example I: Low-rank Matrix Completion

We observe:

\[ Y = \mathcal{P}_\Omega X \]

where:

- \( Y \) are the observed ratings
- \( \mathcal{P}_\Omega \) is the projection onto the observed ratings
- \( X \) are the complete ratings

\[ Y \] is a matrix with unknown values represented by `?`.
Example I: Low-rank Matrix Completion

We observe:

\[
\mathbf{Y} = \mathcal{P}_\Omega \mathbf{X}.
\]

Matrix completion via nonconvex Burer-Monteiro factorization

\[
\min_{\mathbf{U}, \mathbf{V}} f(\mathbf{U}, \mathbf{V}) = \sum_{(i,j) \in \Omega} [(\mathbf{U} \mathbf{V}^*)_{i,j} - Y_{i,j}]^2 + \frac{\lambda}{2} \|\mathbf{U}\|_F^2 + \frac{\lambda}{2} \|\mathbf{V}\|_F^2 + \underbrace{\text{reg}(\mathbf{U}, \mathbf{V})}_{\text{reg}}.
\]
Example II: Dictionary for Image Representation

Image processing (e.g. denoising or super-resolution) against a known sparsifying dictionary:

\[ I_{\text{noisy}} = \underbrace{A_{\text{dictionary}} \times x_{\text{sparse}}} + z_{\text{noise}}. \] (2)

**Dictionary learning**: the motifs or atoms of the dictionary are unknown:

\[ Y_{\text{data}} = A_{\text{dictionary}} \times X_{\text{sparse}}. \] (3)
Example II: Dictionary for Image Representation

Image processing
(e.g. denoising or super-resolution)
against a known sparsifying dictionary:

\[ I_{\text{noisy}} = A_{\text{dictionary}} \times x_{\text{sparse}} + z_{\text{noise}}. \]  \hspace{1cm} (2)

**Dictionary learning**: the motifs or atoms of the dictionary are unknown:

\[ Y_{\text{data}} = A_{\text{dictionary}} \times X_{\text{sparse}}. \]  \hspace{1cm} (3)

- Band-limited signals: \( A = F \), the Fourier transform;
- Piecewise smooth signals: \( A = W \), the wavelet transforms;
- Natural images \( A = ? \) (How to learn \( A \) from the data \( Y \)?)
Dictionary Learning

Recovered solutions always obtain the same objective value.
Example: Sparse Blind Deconvolution

Sparse Blind Deconvolution: the convolutional motif or sparse activation signal are unknown:

\[ Y_{\text{data}} = A_{\text{motif}} \ast X_{\text{sparse}} \]  (4)

- Scientific signals: activation signals are sparse
- Image deblurring: natural images are sparse in the gradient domain
Sparse Blind Deconvolution

Recovered solutions are near signed shift-truncations of the ground truth.
**Convolutional Dictionary learning**

\[ Y_{\text{data}} = \sum_i A_i \otimes X_i. \]

Recovered solutions are near signed shift-truncations of the ground truth.
Opportunities – Optimistic Views

Nonconvex problems that arise in machine learning typically have **benign** data structures, in terms of **symmetries**!
Opportunities – Optimistic Views

Nonconvex problems that arise in machine learning typically have benign data structures, in terms of symmetries!

The function $\varphi$ is invariant under certain group action:

- **low rank matrix recovery**: invariant under a continuous rotation:
  \[
  \varphi((U\Gamma, V\Gamma^{-1})) = \varphi((U, V)), \quad \forall \text{ invertible } \Gamma.
  \]

- **dictionary learning**: invariant under signed permutations:
Opportunities – Optimistic Views

Nonconvex problems that arise in machine learning typically have **benign** data structures, in terms of **symmetries**!

The function $\varphi$ is **invariant** under certain group action:

- **low rank matrix recovery**: invariant under a continuous rotation:
  \[ \varphi((U\Gamma, V\Gamma^{-1})) = \varphi((U, V)), \quad \forall \text{ invertible } \Gamma. \]

- **dictionary learning**: invariant under signed permutations:
  \[ \varphi((A, X)) = \varphi((A\Pi, \Pi^* X)), \quad \forall \Pi \in \text{SP}(n). \]
Nonlinearity and Symmetry

Intrinsic ambiguity against the uniqueness of the solution

• **low rank matrix recovery**

\[ X = U_0 V_0^T = U_0 \Gamma \Gamma^{-1} V_0^T \]

for any invertible \( \Gamma \).

• **dictionary learning**

\[ Y = A_0 X_0 = A_0 \Pi \Pi^* X_0 \]

for any signed permutation \( \Pi \).

• **blind deconvolution**

\[ y = a_0 \ast x_0 = S_\tau[a_0] \ast S_{-\tau}[x_0] \]

for any signed shift \( \tau \).
Optimization under Symmetry

Definition (Symmetric Function)

Let $G$ be a group acting on $\mathbb{R}^n$. A function $\varphi : \mathbb{R}^n \rightarrow \mathbb{R}^{n'}$ is $G$-symmetric if for all $z \in \mathbb{R}^n$, $g \in G$, $\varphi(g \circ z) = \varphi(z)$.

Most symmetric objective functions that arise in structured signal recovery do not have spurious local minimizers or flat saddles.
Optimization under Symmetry

**Definition (Symmetric Function)**

Let $G$ be a group acting on $\mathbb{R}^n$. A function $\varphi : \mathbb{R}^n \to \mathbb{R}^{n'}$ is $G$-symmetric if for all $z \in \mathbb{R}^n$, $g \in G$, $\varphi(g \circ z) = \varphi(z)$.

Most symmetric objective functions that arise in structured signal recovery do not have spurious local minimizers or flat saddles.

**Slogan 1:** the (only!) local minimizers are symmetric versions of the ground truth.

**Slogan 2:** any local critical point has negative curvature in directions that break symmetry.
Outline

1. Introduction

2. Symmetry & Geometry for Nonconvex Problems in Practice
   - Problems with Rotational Symmetry
   - Problems with Discrete Symmetry

3. Efficient Nonconvex Optimization

4. Conclusion
Problems with Rotational Symmetry

Nonconvex Problems with Rotational Symmetries

**Eigenspace Computation**

Compute the principal subspace of a symmetric matrix.

\[ \min_{X^*X=I} \frac{1}{2} \text{trace} [X^*AX]. \]

*Symmetry:* \( X \mapsto XR \)

\( G = O(r) \)

**Generalized Phase Retrieval**

Recover a complex vector \( x_0 \) from magnitude measurements \( y = |Ax_0| \).

\[ \min_x \frac{1}{2} \| y^2 - |Ax|^2 \|^2_2. \]

*Symmetry:* \( x \mapsto xe^{i\phi} \)

\( G = S^1 \cong O(2) \)

**Matrix Recovery**

Recover a low-rank matrix \( X = UV^* \) from incomplete/corrupted observations.

\[ \min_{U,V} \mathcal{L}(Y - A[UV^*]) + \rho(U,V). \]

*Symmetry:* \( (U,V) \mapsto (U^*,V^T) \)

\( G = GL(r) \) or \( G = O(r) \)
Low Rank Matrix Recovery

Goal: Given $Y = \mathcal{A}(X)$, recover low rank matrix $X = U_0V_0$
Low Rank Matrix Recovery

Goal: Given $Y = A(X)$, recover low rank matrix $X = U_0V_0$

- Convex formulation:

$$\min_{X \in \mathbb{R}^{m \times n}} \|X\|_* \quad \text{s.t.} \quad Y = A(X)$$
Low Rank Matrix Recovery

**Goal:** Given $Y = A(X)$, recover low rank matrix $X = U_0V_0$

- **Convex formulation:**
  $$\min_{X \in \mathbb{R}^{m \times n}} \|X\|_* \quad \text{s.t.} \quad Y = A(X)$$

- **Nonconvex formulation:**
  $$\min_{U \in \mathbb{R}^{m \times r}, V \in \mathbb{R}^{n \times r}} \|Y - A(UV^T)\|_F^2 + \text{reg}(U, V)$$
Low Rank Matrix Recovery

\[
\min_{U,V} \frac{1}{2} \| Y - A(UV^T) \|_F^2 + \text{reg}(U, V)
\]

**Inherent Symmetry:**

\[
X = U_0 V_0^T = U_0 \Gamma \Gamma^{-1} V_0^T
\]

for any invertible \( \Gamma \in \mathbb{R}^{r \times r} \).
Low Rank Matrix Recovery

\[
\min_{U,V} \quad \frac{1}{2} \left\| Y - A(UV^T) \right\|_F^2 + \text{reg}(U, V)
\]

Inherent Symmetry:

\[
X = U_0V_0^T = U_0\Gamma\Gamma^{-1}V_0^T
\]

for any invertible \( \Gamma \in \mathbb{R}^{r\times r} \).

- Are \((U_0\Gamma, V_0\Gamma^{-1})\) the only local solutions?
- Does there exist any flat stationary point?
Simple Setting: Rank-1 Symmetric Matrix

- **Simplifications:**
  - \( Y = \mathcal{A}(X) = X \)
  - \( X = U_0 U_0^T \) is symmetric and rank-1

\[
X = u_0 u_0^T = (-u_0 Q)(-Q^T u_0^T)
\]

the signed rotational symmetry.
Simple Setting: Rank-1 Symmetric Matrix

- **Simplifications:**
  - \( Y = A(X) = X \)
  - \( X = U_0 U_0^T \) is symmetric and rank-1

\[
X = u_0 u_0^T = (-u_0 Q)(-Q^T u_0^T)
\]

the signed rotational symmetry.

- **Nonconvex formulation:**

\[
\min_u \phi(u) = \frac{1}{4} \| X - uu^T \|_F^2 + \lambda \| u \|_2^2
\]
Rank-1 Symmetric Matrix

\[
\min_u \phi(u) = \frac{1}{4} \|X - uu^T\|_F^2
\]

- Critical points have zero gradient

\[
\nabla \phi = (uu^T - X)u = \|u\|_2^2 u - Xu = 0
\]
Rank-1 Symmetric Matrix

\[
\min_{u} \phi(u) = \frac{1}{4} \| X - uu^T \|_F^2
\]

- Critical points have zero gradient

\[
\nabla \phi = (uu^T - X)u = \|u\|^2_2 u - Xu = 0
\]

- Therefore, critical points must be one of the following
  - \( u = \pm Qu_0 \)
  - \( u = 0 \)
Rank-1 Symmetric Matrix

\[ \min_{u} \phi(u) = \frac{1}{4} \| X - uu^T \|_F^2 \]

with the second-order derivative

\[ \nabla^2 \phi = 2uu^T + \| u \|_2^2 I - X. \]
Rank-1 Symmetric Matrix

\[
\min_u \phi(u) = \frac{1}{4} \| X - uu^T \|_F^2
\]

with the second-order derivative

\[
\nabla^2 \phi = 2uu^T + \|u\|_2^2 I - X.
\]

Then the stationary points can be grouped as

- **Local minimizer** \( u = \pm Qu_0 \):
  \[
  \nabla^2 \phi = uu^T + \|u\|_2^2 I \succeq 0
  \]

- **Maximizer** \( u = 0 \)
  \[
  \nabla^2 \phi = -X < 0.
  \]
Low Rank Matrix Recovery

- Symmetric low rank matrix recovery:
  \[
  \min_U \phi(u) = \frac{1}{4} \| X - UU^T \|_F^2.
  \]

- General low rank matrix recovery:
  \[
  \min_{U, V} \phi(u) = \frac{1}{2} \| X - UV^T \|_F^2 + \lambda \| U \|_F^2 + \lambda \| V \|_F^2.
  \]

Local minimizers: are ground truth \( U_0 \) and \( V_0 \) up to rotation;
Negative curvature: between multiple local minimizers.
Problems with Discrete Symmetry

Nonconvex Problems with Discrete Symmetries

**Eigenvector Computation**
Maximize a quadratic form over the sphere.

$$\max_{x \in \mathbb{S}^{n-1}} \frac{1}{2} x^* A x.$$  
*Symmetry: $x \mapsto -x$  
$\mathcal{G} = \{ \pm 1 \}$

**Dictionary Learning**
Approximate a given matrix $Y$ as $Y \approx AX$, with $X$ sparse

$$\min_{A, X} \frac{1}{2} \| Y - AX \|_F^2 + \lambda \| X \|_1.$$  
*Symmetry: $(A, X) \mapsto (A\Gamma, X\Gamma^*)$  
$\mathcal{G} = \text{SP}(n)$

**Tensor Decomposition**
Determine components $a_i$ of an orthogonal decomposable tensor $T = \sum_i a_i \otimes a_i \otimes a_i \otimes a_i$

$$\max_{X \in O(n)} \sum_i T(x_i, a_i, a_i, a_i).$$  
*Symmetry: $X \mapsto X\Gamma$  
$\mathcal{G} = P(n)$

**Short-and-Sparse Deconvolution**
Recover a short $a$ and a sparse $x$ from their convolution $y = a * x$.

$$\min_{a, x} \frac{1}{2} \| y - a * x \|_2^2 + \lambda \| x \|_1.$$  
*Symmetry: $(a, x) \mapsto (a_{s^+} [a], a_{s^-}^{-1} [x])$  
$\mathcal{G} = \mathbb{Z}_n \times \mathbb{R}^*$ or $\mathcal{G} = \mathbb{Z}_n \times \{ \pm 1 \}$
Dictionary Learning

Goal: Given dataset $Y$, find the optimal dictionary $A$ that renders the sparsest coefficient $X$

$$\min_{A, X} \| X \|_1 \quad \text{s.t.} \quad Y = AX.$$ 

In presence of noise, the optimization problem can be rewritten as

$$\min_{A, X} \frac{1}{2} \| Y - AX \|_F^2 + \lambda \| X \|_1.$$
**Dictionary Learning**

Goal: Given dataset $Y$, find the optimal dictionary $A$ that renders the sparsest coefficient $X$

$$\min_{A,X} \|X\|_1 \quad \text{s.t.} \quad Y = AX.$$ 

In presence of noise, the optimization problem can be rewritten as

$$\min_{A,X} \frac{1}{2} \|Y - AX\|_F^2 + \lambda \|X\|_1.$$ 

**Inherent Symmetry:**

$$Y = A_0 \Gamma \Gamma^* X_0,$$

for any signed permutation matrix $\Gamma$. 

Qing Qu (EECS, University of Michigan)

Nonconvex Optimization

June 10th, 2023 35 / 53
Orthogonal Dictionary Learning

- Input: matrix $Y$ which is the product of an orthogonal matrix $A_0$ (called a dictionary) and a sparse matrix $X_0$:

$$Y = A_0X_0, \quad A_0A_0^* = I, \quad X_0 \text{ sparse.}$$

- Optimization formulation:

$$\min_{A,X} \|X\|_1 \quad \text{s.t.} \quad Y = AX, \quad AA^* = I.$$
Orthogonal Dictionary Learning

- Input: matrix $Y$ which is the product of an orthogonal matrix $A_0$ (called a dictionary) and a sparse matrix $X_0$:

$$Y = A_0 X_0, \quad A_0 A_0^* = I, \quad X_0 \text{ sparse.}$$

- Optimization formulation:

$$\min_{A,X} \|X\|_1 \quad \text{s.t.} \quad Y = AX, \quad AA^* = I.$$  

- Given the constraint, $X$ is uniquely defined in terms of $A$

$$X = A^* AX = A^* Y.$$
Orthogonal Dictionary Learning

- Input: matrix $Y$ which is the product of an orthogonal matrix $A_0$ (called a dictionary) and a sparse matrix $X_0$:
  \[ Y = A_0 X_0, \quad A_0 A_0^* = I, \quad X_0 \text{ sparse}. \]

- Optimization formulation:
  \[
  \min_{A,X} \|X\|_1 \quad \text{s.t.} \quad Y = AX, \quad AA^* = I.
  \]

- Given the constraint, $X$ is uniquely defined in terms of $A$
  \[ X = A^* AX = A^* Y. \]

- Equivalent formulation:
  \[
  \min_{A \in O(n)} \|A^* Y\|_1.
  \]
Orthogonal Dictionary Learning

Instead of aiming to solve the entire matrix $A = [a_1, \ldots, a_n]$ at once via

$$\min_{A \in O(n)} \| A^* Y \|_1.$$ 

A simpler model problem solves for the columns $a_i$ one at a time

$$\min_{\|a\|_2 = 1} \| a^* Y \|_1.$$
Orthogonal Dictionary Learning

Instead of aiming to solve the entire matrix \( A = [a_1, \ldots, a_n] \) at once via

\[
\min_{A \in O(n)} \|A^*Y\|_1.
\]

A simpler model problem solves for the columns \( a_i \) one at a time

\[
\min_{\|a\|_2=1} \|a^*Y\|_1.
\]

Stationary Points:

- \( a = \pm a_i \), then the Hessian is positive definite
- \( a = \sum_{i \in I} \pm \frac{1}{\sqrt{|I|}} a_i \), there exist negative curvatures alone \( a_i (i \in I) \)
Orthogonal Dictionary Learning — Geometry

**Local minimizers** are ground truth $a_i$ or $-a_i$.

**Negative curvature** between multiple local minimizers.
Short-and-Sparse Blind Deconvolution

Goal: Given convolutional data $y$, find the **short** signal $a$ and the **sparse** signal $x$ such that $y = a \ast x$.

**Inherent Symmetry:**

$$y = a_0 \ast x_0 = \alpha s_l[a_0] \ast \frac{1}{\alpha} s_{-l}[x_0]$$

for any shift $l$ and nonzero scaling.
Short-and-Sparse Blind Deconvolution

Goal: Given convolutional data $y$, find the short signal $a$ and the sparse signal $x$ such that $y = a * x$.

**Inherent Symmetry:**

$$y = a_0 * x_0 = \alpha s_l[a_0] * \frac{1}{\alpha} s_{-l}[x_0]$$

for any shift $l$ and nonzero scaling.

The practical optimization problem can be written as

$$\min_{\|a\|_F = 1, x} \frac{1}{2} \|y - a * x\|_F^2 + \lambda \|x\|_1.$$
Objective Function – Near One Shift

\[ \mathbb{S}^{p-1} \cap \{ \mathbf{a} \in \mathbb{S}^{p-1} \mid \| \mathbf{a} - s_\ell[a_0] \|_2 \leq r \} \]

Objective function is **strongly convex** near a shift \( s_\ell[a_0] \) of the ground truth.
Objective Function – Linear Span of Two Shifts

Subspace \( S_{\{\ell_1, \ell_2\}} = \{ \alpha_1 s_{\ell_1}[a_0] + \alpha_2 s_{\ell_2}[a_0] \mid \alpha_1, \alpha_2 \in \mathbb{R} \} \).
Objective Function – Linear Span of Two Shifts

Local minimizers are near signed shifts $\pm s\ell[a_0]$. Negative curvature between two shifts $s\ell_1[a_0], s\ell_2[a_0]$. 
Objective Function – Multiple Shifts

Objective $\varphi_{\rho}$ over the linear span $S_{\ell_1,\ell_2,\ell_3} = \{ \sum_{i=1}^{3} \alpha_{\ell_i} s_{\ell_i}[a_0] \}$

Local minimizers are near signed shifts $\pm s_{\ell_i}[a_0]$. 
Symmetry and Nonconvexity

**Slogan 1:** the (only!) local minimizers are symmetric versions of the ground truth.

**Slogan 2:** any local critical point has negative curvature in directions that break symmetry.
Outline

1 Introduction

2 Symmetry & Geometry for Nonconvex Problems in Practice
   Problems with Rotational Symmetry
   Problems with Discrete Symmetry

3 Efficient Nonconvex Optimization

4 Conclusion
Nonconvex Optimization in Generic Setting

Consider the problem of minimizing a general nonconvex function:

\[
\min_x f(x), \quad x \in C. \tag{5}
\]

In the worst case, even finding a local minimizer can be NP-hard\(^2\).

Nonconvex problems that arise from natural physical, geometrical, or statistical origins typically have nice structures, in terms of symmetries!

\[^2\text{Some NP-complete problems in quadratic and nonlinear programming, K.G Murty and S. N. Kabadi, 1987}\]
Nonconvex Optimization in Generic Setting

Hence typically people seek to work with relatively benign (gradient/Hessian Lipschitz continuous) functions:

\[ \forall x, y \quad \| \nabla f(y) - \nabla f(x) \|_2 \leq L_1 \| y - x \|_2 \quad (6) \]

with benign objectives:

1. convergence to some critical point \( x_\star \) such that: \( \nabla f(x_\star) = 0 \);
2. the critical point \( x_\star \) is second-order stationary: \( \nabla^2 f(x_\star) \succeq 0 \).
Nonconvex Optimization in Generic Setting

Hence typically people seek to work with relatively benign (gradient/Hessian Lipschitz continuous) functions:

\[ \forall x, y \quad \| \nabla f(y) - \nabla f(x) \|_2 \leq L_1 \| y - x \|_2 \]  

(6)

with benign objectives:

1. convergence to some critical point \( x^\ast \) such that: \( \nabla f(x^\ast) = 0 \);
2. the critical point \( x^\ast \) is second-order stationary: \( \nabla^2 f(x^\ast) \succeq 0 \).

Example: a function \( f \) with symmetry only has regular critical points, while general \( f \) could have irregular second-order stationary points:
Benign Nonconvexity: “Any Reasonable Algorithm” Works

**Key issue:** using negative curvature

\[ \lambda_{\min}(\text{Hess} f) < 0 \]

to escape saddles.
Benign Nonconvexity: “Any Reasonable Algorithm” Works

**Key issue:** using negative curvature
\[ \lambda_{\min}(\text{Hess} f) < 0 \]
to escape saddles.

- **Efficient (polynomial time) methods:**
  Trust region method, analyses in [Sun, Qu, W., ’17]
  Curvilinear search, [Goldfarb, Mu, W., Zhou, ’16]
  Noisy (stochastic) gradient descent, [Jin et. al. ’17].
Key issue: using negative curvature
\[ \lambda_{\text{min}}(\text{Hess} f) < 0 \]
to escape saddles.

- Efficient (polynomial time) methods:
  - Trust region method, analyses in [Sun, Qu, W., ’17]
  - Curvilinear search, [Goldfarb, Mu, W., Zhou, ’16]
  - Noisy (stochastic) gradient descent, [Jin et. al. ’17].

- Randomly initialized gradient descent ....
  Obtains a minimizer almost surely [Lee et. al. ’16].
  Efficient for matrix completion, dictionary learning, ... not efficient in general.
Worst Case vs. Naturally Occurring Strict Saddle Functions

Worst Case

[Du, Jin, Lee, Jordan, Poczos, Singh ’17]
Concentration around stable manifold

Naturally Occuring

DL, Other sparsification problems
Dispersion away from stable manifold
Worst Case vs. Naturally Occurring Strict Saddle Functions

- Red: “slow region” of small gradient around a saddle point.
- Green: stable manifold associated with the saddle point.
- Black: points that flow to the slow region.

- Left: global negative curvature normal to the stable manifold
- Right: positive curvature normal to the stable manifold – randomly initialized gradient descent is more likely to encounter the slow region.
**Dispersive structure:** Negative curvature \( \perp \) stable manifolds.

W.h.p. in random initialization \( q^{(0)} \sim \text{uni}(S^{n-1}) \), convergence to a neighborhood of a minimizer in polynomial iterations. [Gilboa, Buchanan, W. '18]
Outline

1 Introduction

2 Symmetry & Geometry for Nonconvex Problems in Practice
   Problems with Rotational Symmetry
   Problems with Discrete Symmetry

3 Efficient Nonconvex Optimization

4 Conclusion
References


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

Conclusion and Coming Attractions

For Nonconvex, Sparse and Low-rank problems

- **Benign Geometry**:  
  - The only local minimizers are symmetric copies of the ground truth  
  - There exist negative curvatures breaking symmetry

- **Efficient Algorithms**:  
  - gradient descent algorithms always suffice  
  - proximal, projection, acceleration steps can be transferred over

Thank You! Questions?
Call for Papers

- IEEE JSTSP Special Issue on Seeking Low-dimensionality in Deep Neural Networks (SLowDNN) Manuscript Due: Nov. 30, 2023.