UNIVERSITY OF MICHIGAN

EECS 559
Optimization Methods for SIPML

Lecture 11 – Intro on Nonconvex Problems in SIPML

Instructor: Prof. Qing Qu (qingqu@umich.edu)

Lecture Agenda

- Problem Introduction
- Applications in SIPML

Lecture Agenda

- **Problem Introduction**
  - General Nonconvex Problems
  - Global Optimization?
  - Strict Saddle Problems
- Applications in SIPML

Lecture Agenda

- **Problem Introduction**
  - **General Nonconvex Problems**
  - Global Optimization?
  - Strict Saddle Problems
- Applications in SIPML

Nonconvex Optimization

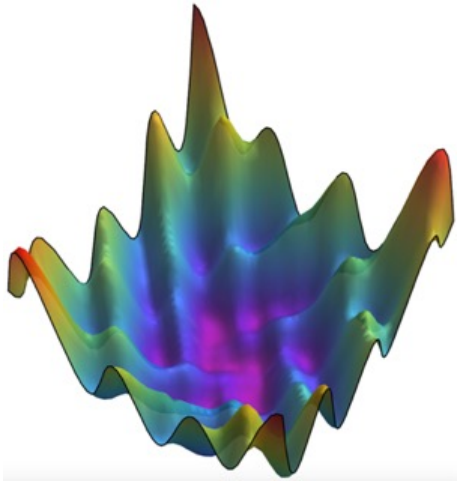
$$\min_{\boldsymbol{x}} \; f(\boldsymbol{x}), \;\; \text{s.t.} \;\; \boldsymbol{x} \in \mathcal{C}.$$

The problem is *nonconvex* if either

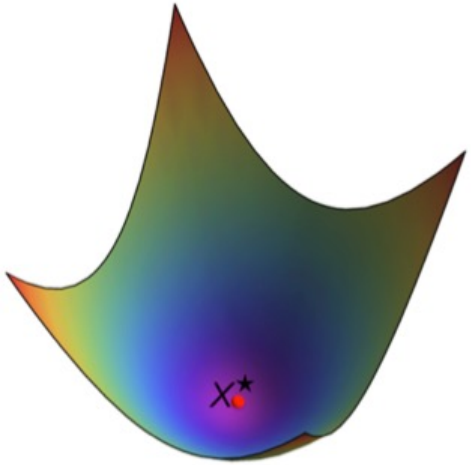
- the objective $f(\boldsymbol{x})$ is nonconvex,
- or the constraint set $\mathcal{C}$ is nonconvex.

Convex vs Nonconvex Problems

$$\min_{\boldsymbol{x}} \; f(\boldsymbol{x}), \;\; \text{s.t.} \;\; \boldsymbol{x} \in \mathbb{R}^n$$

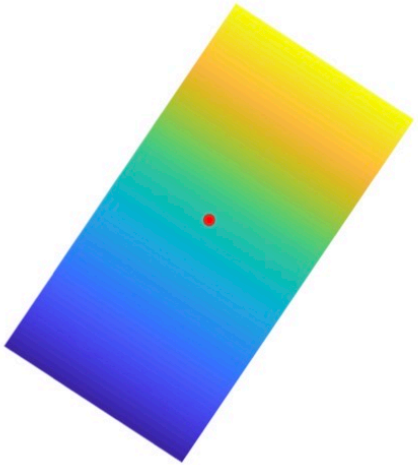


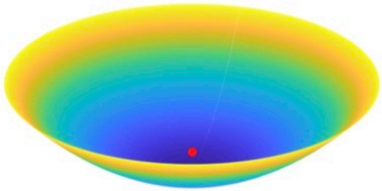
Nonconvex landscape



Convex landscape

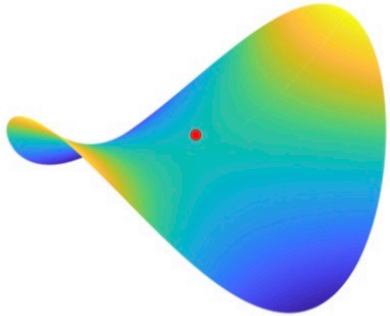
Classification of Critical Points





Minimizer

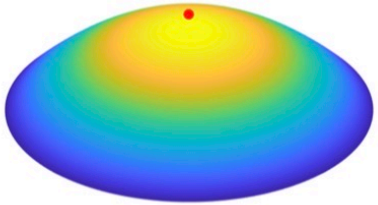
$\nabla^2\varphi \succ \mathbf{0}$



Saddle

$\lambda_{\min}\nabla^2\varphi < 0$

$\lambda_{\max}\nabla^2\varphi > 0$



Maximizer

$\nabla^2\varphi \prec \mathbf{0}$

Noncritical Point $(\nabla\varphi \neq \mathbf{0})$

Critical Points $(\nabla\varphi = \mathbf{0})$

Classification of Critical Points

Definition. (Critical points, local minimizers, & saddle points) Suppose $f : \mathbb{R}^n \mapsto \mathbb{R}$ is continuously differentiable, then

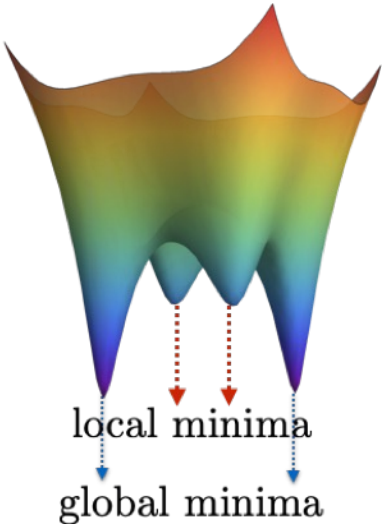
- A point $\boldsymbol{x}_\star$ is called a ***critical point*** of f if $\nabla f(\boldsymbol{x}_\star) = \mathbf{0}$;
- A point $\boldsymbol{x}_\star$ is a ***local minimizer*** if $\exists \delta > 0$ such that $f(\boldsymbol{x}_\star) \leq f(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathcal{B}(\boldsymbol{x}_\star, \delta)$;
- A point $\boldsymbol{x}_\star$ is a ***saddle point*** if it is a critical point but neither a local minimizer nor local maximizer.

A local minimizer is a critical point, but not vice versa.

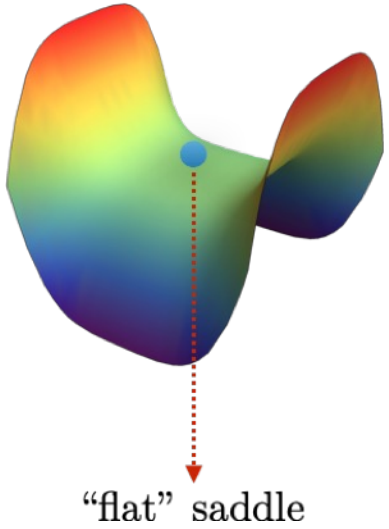
General Nonconvex Problems

$$\min_{\boldsymbol{x}} \; f(\boldsymbol{x}), \quad \text{s.t.} \quad \boldsymbol{x} \in \mathbb{R}^n$$

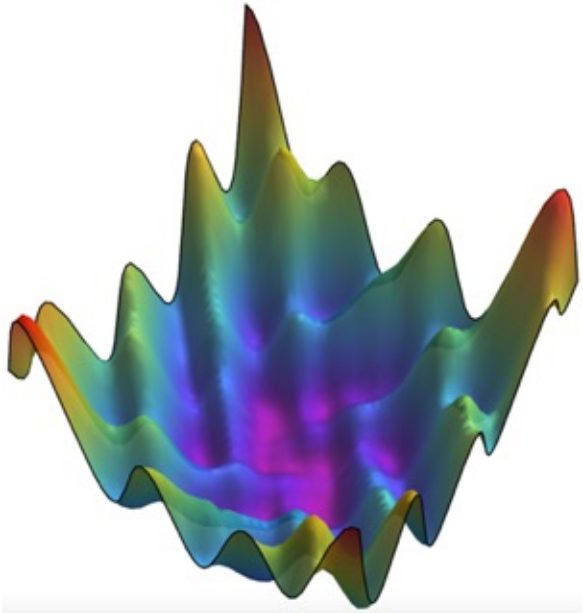
"bad" local minimizers



"flat" saddle points



General Nonconvex Problems



$$\min_{\boldsymbol{x}} \; f(\boldsymbol{x}), \;\; \text{s.t.} \;\; \boldsymbol{x} \in \mathbb{R}^n$$

In the **worst-case**, even finding a local minimizer is NP-hard (Murty et al. 1987)

Typical Convergence Guarantees

For nonconvex problems, in general we *cannot* hope for efficient global convergence to optimal solutions, but we may have

- guaranteed convergence to critical points (i.e., $\nabla f(\boldsymbol{x}) = \mathbf{0}$);
- convergence to local minimizers;
- local convergence to global minimizers (with proper initializations).

Convergence to Critical Points

Suppose we are satisfied with any critical point...
This means that our goal is merely to find a point $\boldsymbol{x}$ with

$$\|\nabla f(\boldsymbol{x})\|_2 \leq \varepsilon \quad \text{(called } \varepsilon\text{-approximate critical points)}$$

Question: can *vanilla* gradient descent

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \tau_k \cdot \nabla f(\boldsymbol{x}_k)$$

achieve this goal? If so, how fast?

Gradient Descent Converges to Critical Points

Theorem. (Convergence of gradient descent)
Let $f(\boldsymbol{x})$ be L-smooth function and $\tau_k = \tau \equiv 1/L$.

- In general, gradient descent obeys

$$\min_{0 \le i \le k} \|\nabla f(\boldsymbol{x}_i)\|_2 \le \sqrt{\frac{(f(\boldsymbol{x}_0) - f(\boldsymbol{x}_\star))}{k}}.$$

- If $f(\boldsymbol{x})$ is convex, then gradient descent obeys

$$\min_{k/2 \le i \le k} \|\nabla f(\boldsymbol{x}_i)\|_2 \le \frac{4L \|\boldsymbol{x}_0 - \boldsymbol{x}_\star\|}{k}.$$

- The rate cannot be improved for first-order method;
- It does not imply GD converges to critical points.

Lecture Agenda

- **Problem Introduction**
  - General Nonconvex Problems
  - **Global Optimization?**
  - Strict Saddle Problems
- Applications in SIPML

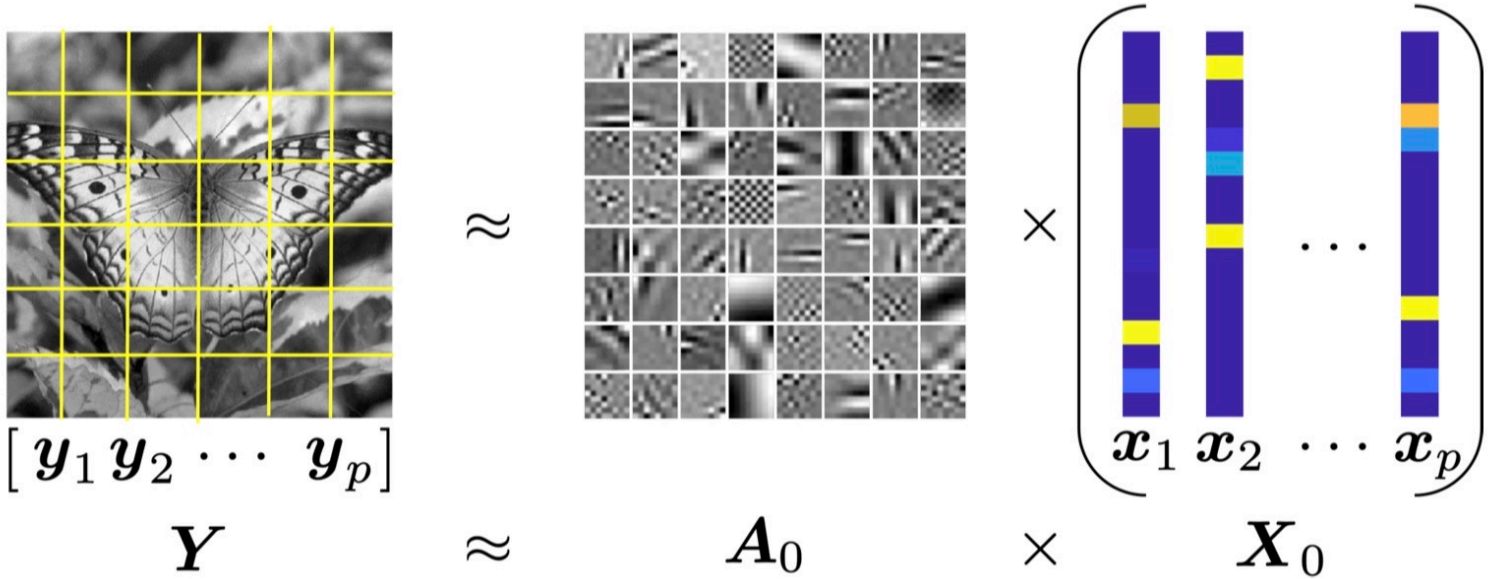
Global Optimization of Nonconvex Problems

- Solving general nonconvex problems to global optimality is NP-hard in the worst case;

- However, in practice we often observe many SIPML problems can be efficiently optimized to global solutions.

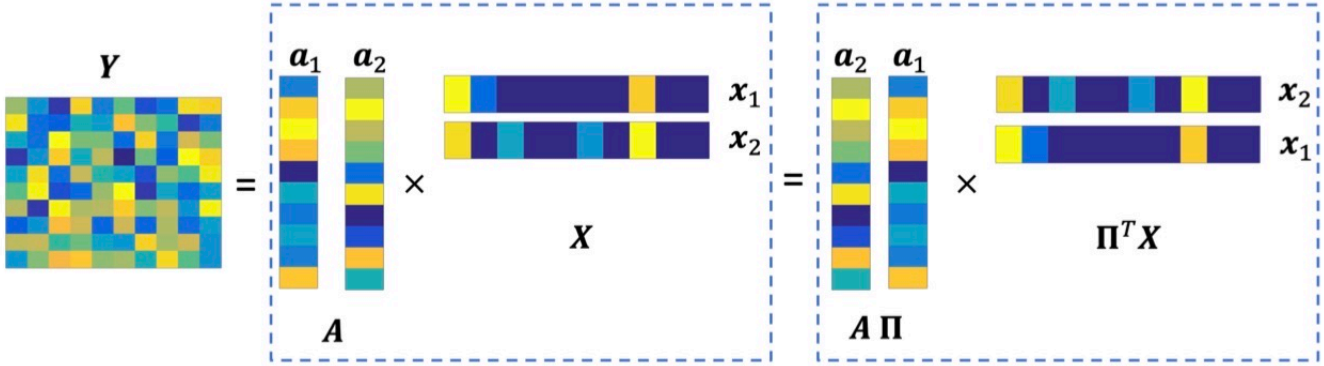
Example: Dictionary Learning

Given $\boldsymbol{Y}$, jointly learn a *compact* dictionary $\boldsymbol{A}_0$ and *sparse* $\boldsymbol{X}_0$?



Example: Dictionary Learning

Nonconvexity due to symmetry:

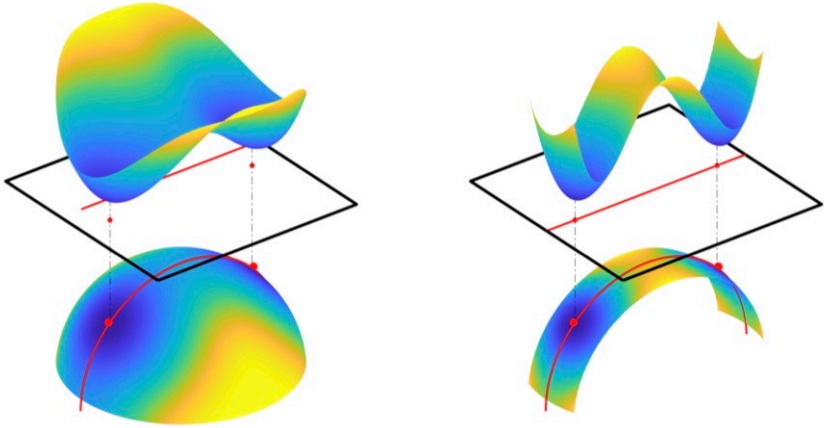


- **Permutation symmetry:** ($2^n n!$ signed permutation $\mathbf{\Pi}$)

$$\boldsymbol{Y} \;=\; \boldsymbol{A}_0\boldsymbol{X}_0 \;=\; (\boldsymbol{A}_0\boldsymbol{\Pi})\left(\boldsymbol{\Pi}^\top \boldsymbol{X}_0\right)$$

- **Equivalent solution pairs:** $(\boldsymbol{A}_0, \boldsymbol{X}_0) \Longleftrightarrow \left(\boldsymbol{A}_0\boldsymbol{\Pi}, \boldsymbol{\Pi}^\top \boldsymbol{X}_0\right).$

Example: Dictionary Learning



- **Permutation symmetry:** ($2^n n!$ signed permutation $\boldsymbol{\Pi}$)

$$\boldsymbol{Y} \;=\; \boldsymbol{A}_0\boldsymbol{X}_0 \;=\; (\boldsymbol{A}_0\boldsymbol{\Pi})\left(\boldsymbol{\Pi}^\top\boldsymbol{X}_0\right)$$

- **Equivalent solution pairs:** $(\boldsymbol{A}_0, \boldsymbol{X}_0) \Longleftrightarrow \left(\boldsymbol{A}_0\boldsymbol{\Pi}, \boldsymbol{\Pi}^\top\boldsymbol{X}_0\right).$

Example: Dictionary Learning

- A natural nonconvex formulation:

$$\min_{\boldsymbol{A},\boldsymbol{X}} \; f(\boldsymbol{A},\boldsymbol{X}) \;=\; \frac{1}{2}\left\|\boldsymbol{Y}-\boldsymbol{A}\boldsymbol{X}\right\|_F^2 + \lambda\cdot\left\|\boldsymbol{X}\right\|_1, \;\; \text{s.t.} \;\; \boldsymbol{A}\in\mathcal{C}.$$

- A natural algorithmic idea (alternating minimization):

1. optimize $\boldsymbol{X}$ with $\boldsymbol{A}_k$ fixed

$$\boldsymbol{X}_{k+1} \;=\; \arg\min_{\boldsymbol{X}} f(\boldsymbol{A}_k,\boldsymbol{X})$$

2. optimize $\boldsymbol{A}$ with $\boldsymbol{X}_{k+1}$ fixed

$$\boldsymbol{A}_{k+1} \;=\; \arg\min_{\boldsymbol{A}\in\mathcal{C}} f(\boldsymbol{A},\boldsymbol{X}_{k+1})$$

Example: Dictionary Learning

When the dictionary $\boldsymbol{A}_0 \in \mathbb{R}^{n\times n}$ is *orthogonal*, i.e.,

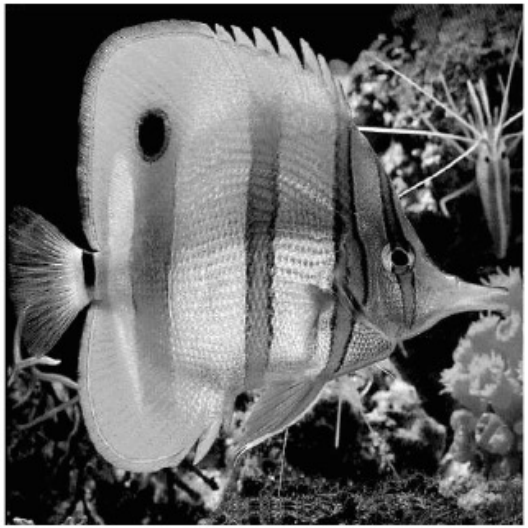
$$\boldsymbol{A}_0 \in \mathcal{O}(n) := \left\{\boldsymbol{Q} \in \mathbb{R}^{n\times n} \mid \boldsymbol{Q}^\top \boldsymbol{Q} = \boldsymbol{I}\right\}$$

Then we have

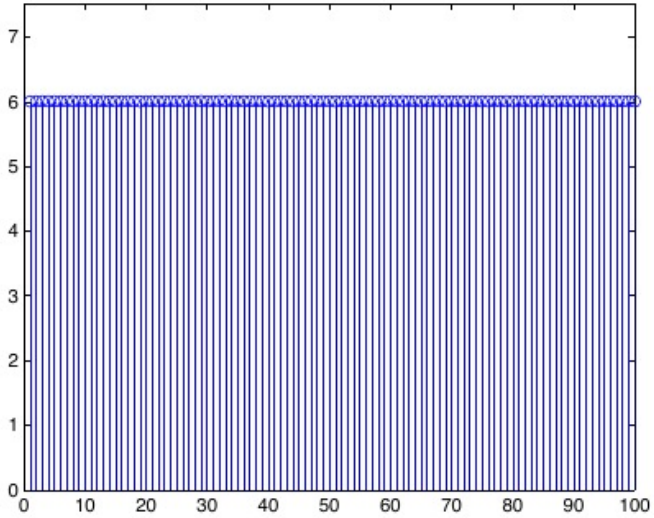
$$\begin{aligned}
\boldsymbol{X}_{k+1} &= \arg\min_{\boldsymbol{X}} f(\boldsymbol{A}_k, \boldsymbol{X}) = \mathrm{S}_\lambda\left[\boldsymbol{A}_k^\top \boldsymbol{Y}\right] \\
\boldsymbol{A}_{k+1} &= \arg\min_{\boldsymbol{A}\in\mathcal{C}} f(\boldsymbol{A}, \boldsymbol{X}_{k+1}) = \boldsymbol{U}\boldsymbol{V}^\top,
\end{aligned}$$

where $\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top = \mathrm{SVD}(\boldsymbol{Y}\boldsymbol{X}^\top)$. (Homework)

Example: Dictionary Learning

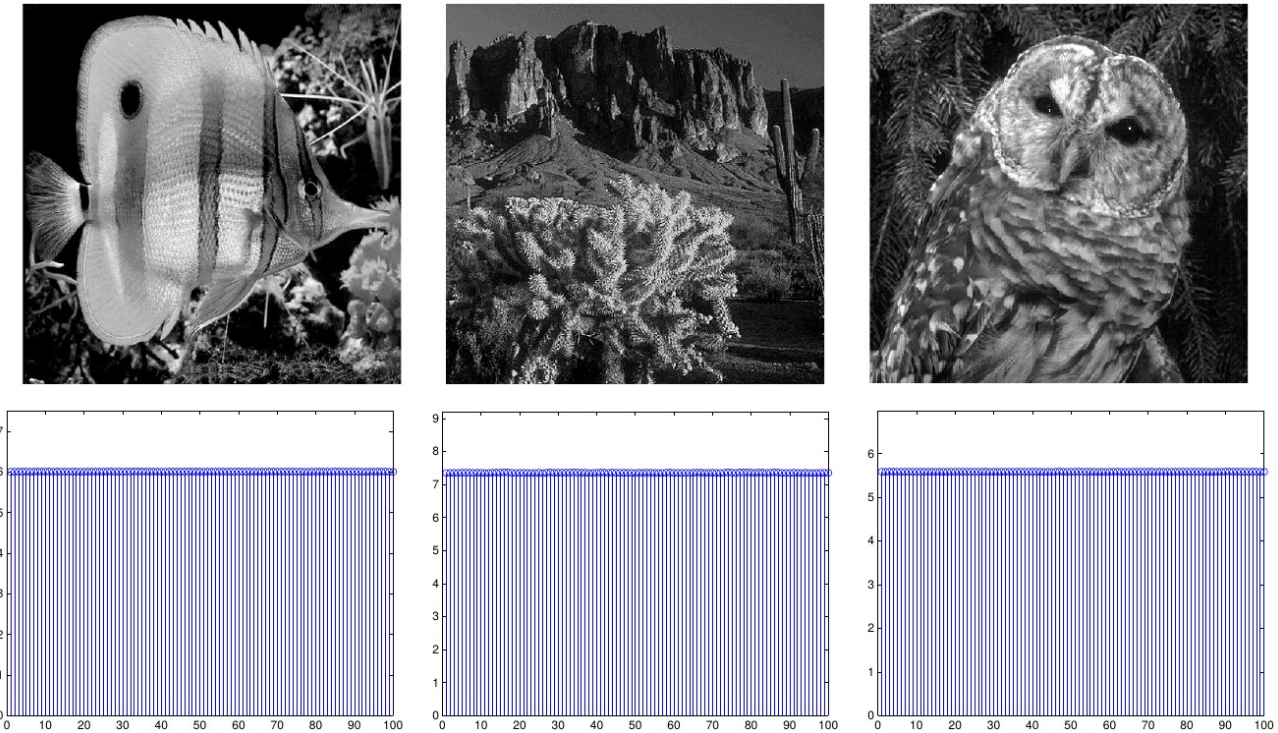


A natural image



$f(\boldsymbol{A}_\infty, \boldsymbol{X}_\infty)$ with random intial.

Example: Dictionary Learning



Further Readings

- *High-Dimensional Data Analysis with Low-Dimensional Models: Principles, Computation, and Applications*. John Wright, Yi Ma. (**Chapter 7**)

- *Complete Dictionary Recovery Over the Sphere I: Overview and the Geometric Picture.* Ju Sun, Qing Qu, and John Wright, IEEE Trans. Info. Theory, 2016.

Lecture Agenda

- **Problem Introduction**
  - General Nonconvex Problems
  - Global Optimization?
  - **Strict Saddle Problems**
- Applications in SIPML

Global Optimization of Nonconvex Problems

- Solving general nonconvex problems to global optimality is NP-hard in the worst case;

- We focus and study on a small class of nonconvex problems, which appears broadly in SIPML, that we can design efficient solvers to global solutions.

A Classical Example: the Rayleigh Quotient

For a given *symmetric* matrix $\boldsymbol{A}^{n\times n}$, solve

$$\min_{\boldsymbol{x}} \; -\frac{\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}}{\boldsymbol{x}^\top \boldsymbol{x}},$$

which is equivalent to

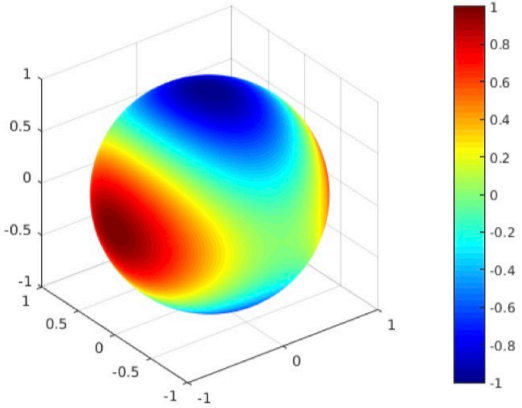
$$\boxed{\min_{\boldsymbol{x}} \; f(\boldsymbol{x}) \;=\; -\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}, \quad \text{s.t.} \quad \|\boldsymbol{x}\|_2 = 1.}$$

A Classical Example: the Rayleigh Quotient

$$\min_{\boldsymbol{x}} \; f(\boldsymbol{x}) \;=\; -\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}, \quad \text{s.t.} \quad \|\boldsymbol{x}\|_2 = 1.$$

Let $\boldsymbol{A} = \boldsymbol{V}^\top \boldsymbol{\Lambda} \boldsymbol{V}$, with $\boldsymbol{V} = \begin{bmatrix} \boldsymbol{v}_1 & \cdots & \boldsymbol{v}_n \end{bmatrix}$ being the eigenvectors, and $\boldsymbol{\Lambda} = \operatorname{diag}(\lambda_1, \cdots, \lambda_n)$ with $\lambda_1 > \lambda_2 \geq \cdots \geq \lambda_{n-1} > \lambda_n$.

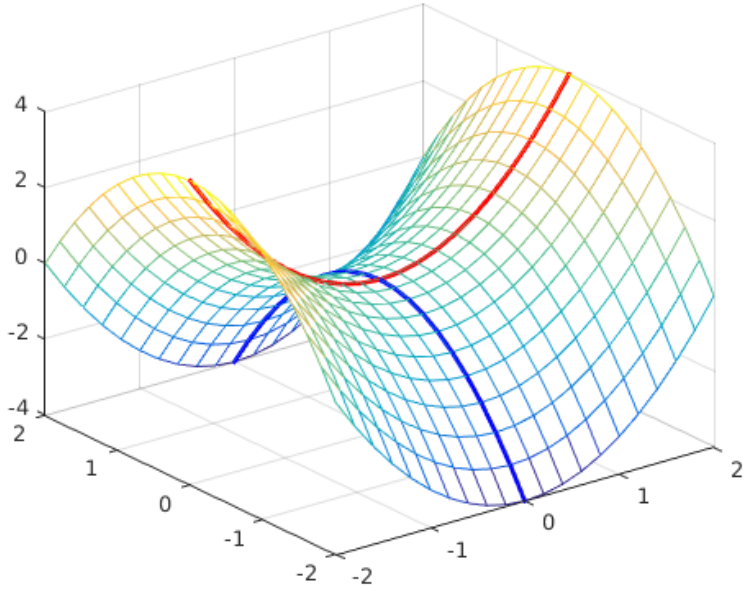
- Only *global minimizers* are $\pm \boldsymbol{v}_1$;
- Only *global maximizers* are $\pm \boldsymbol{v}_n$;
- All $\{\pm \boldsymbol{v}_i\}_{i=1}^n$ for $2 \leq i \leq n-1$ are *saddle points* with *directional negative curvatures.*



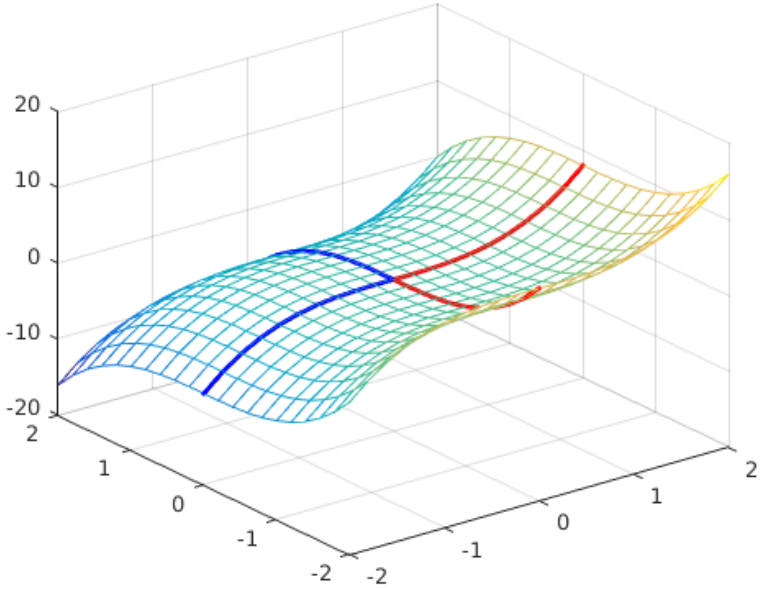
$\boldsymbol{A} = \operatorname{diag}(1, 0, -1)$

Saddle Points with Negative Curvature

$f(x, y) = x^2 - y^2$



$g(x, y) = x^3 - y^3$



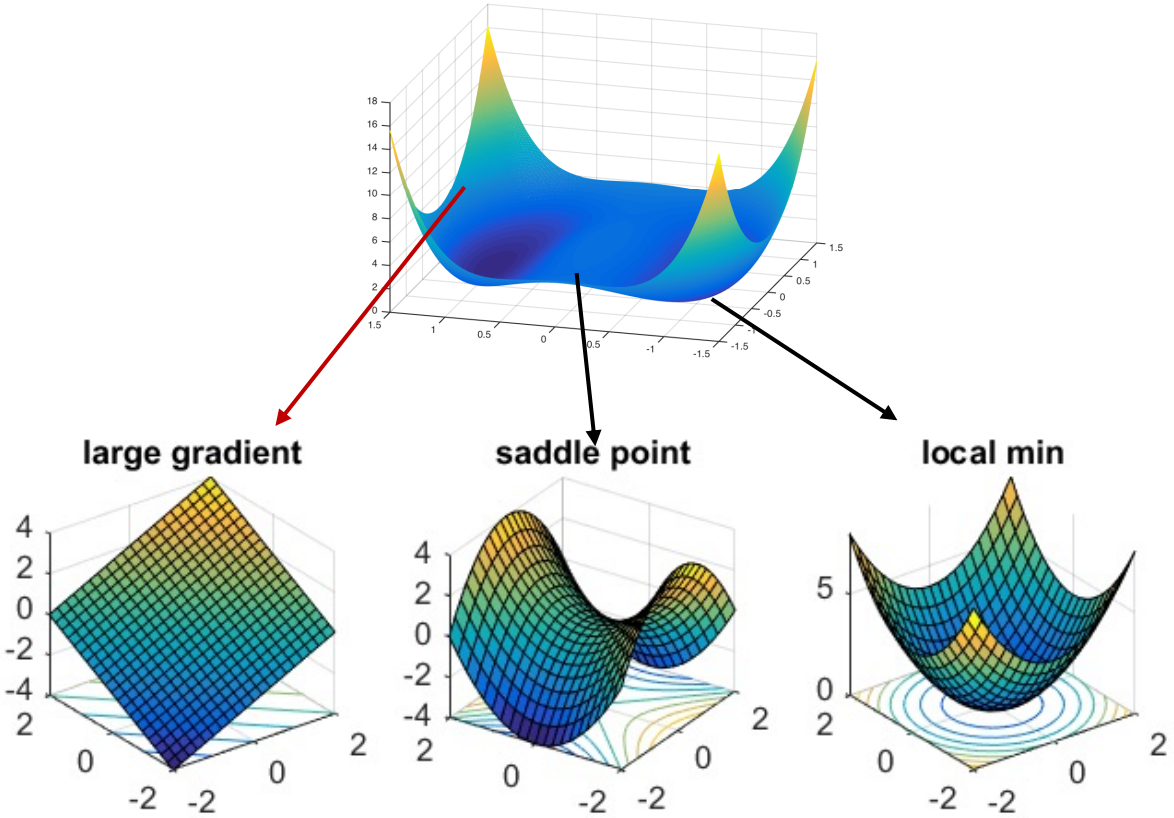
Strict Saddle Function on $\mathbb{R}^n$

Definition. (Strict Saddle Function in $\mathbb{R}^n$, Ge et al.'15) A function $f : \mathbb{R}^n \mapsto \mathbb{R}$ is $(\alpha, \beta, \gamma, \delta)$-strict saddle, if $\forall \boldsymbol{x} \in \mathbb{R}^n$ obeys *at least one* of the following:

- **[Large gradient]** $\|\nabla f(\boldsymbol{x})\|_2 \geq \beta$;
- **[Negative curvature]** $\exists\, \boldsymbol{v} \in \mathbb{S}^{n-1}$, such that
$$\boldsymbol{v}^\top \nabla^2 f(\boldsymbol{x}) \boldsymbol{v} \leq -\alpha;$$
- **[Strong convexity around minimizers]** $\exists \boldsymbol{x}_\star$ such that $\|\boldsymbol{x} - \boldsymbol{x}_\star\|_2 \leq \delta$, and for all $\boldsymbol{y} \in \mathcal{B}(\boldsymbol{x}_\star, 2\delta)$, we have $\nabla^2 f(\boldsymbol{y}) \succeq \gamma \boldsymbol{I}$.

Escaping From Saddle Points --- Online Stochastic Gradient for Tensor Decomposition. Rong Ge, Furong Huang, Chi Jin, Yang Yuan, COLT'15, 2015.

Strict Saddle Function on $\mathbb{R}^n$



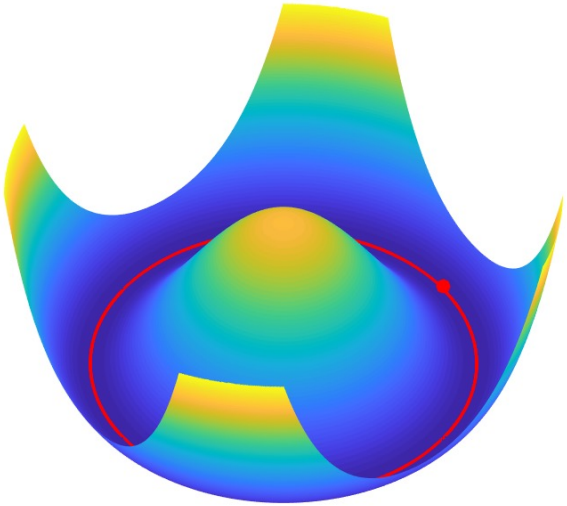
Strict Saddle Function on Manifold $\mathcal{M}$

Definition. (Strict Saddle Function in $\mathcal{M}$, Sun et al.'15) A function $f : \mathcal{M} \mapsto \mathbb{R}$ is $(\alpha, \beta, \gamma, \delta)$-strict saddle, if $\forall \boldsymbol{x} \in \mathcal{M}$ obeys *at least one* of the following:

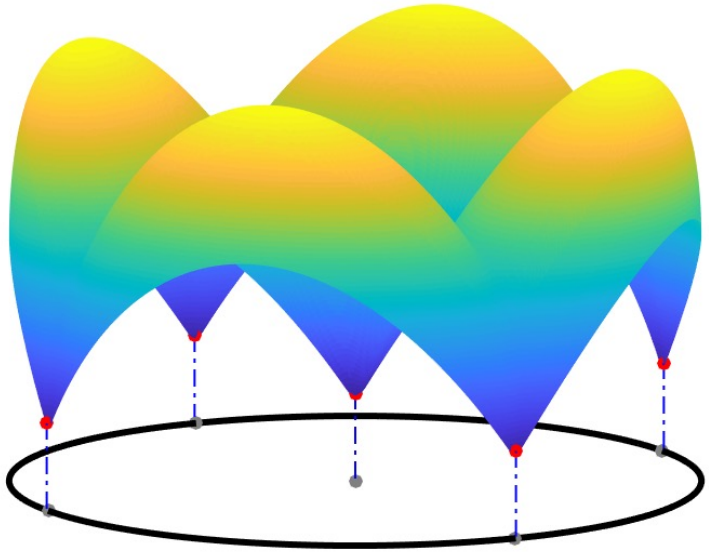
- **[Large gradient]** $\|\mathrm{grad}\, f(\boldsymbol{x})\|_2 \geq \beta$;
- **[Negative curvature]** $\exists\, \boldsymbol{v} \in T_{\boldsymbol{x}}\mathcal{M}$ with $\boldsymbol{v} \in \mathbb{S}^{n-1}$, such that
$$\langle \mathrm{Hess}\, f(\boldsymbol{x})[\boldsymbol{v}], \boldsymbol{v}\rangle \leq -\alpha\ ;$$
- **[Strong convexity around minimizers]** $\exists \boldsymbol{x}_\star$ such that $\|\boldsymbol{x} - \boldsymbol{x}_\star\|_2 \leq \delta$, and for all $\boldsymbol{y} \in \mathcal{B}(\boldsymbol{x}_\star, 2\delta) \cap \mathcal{M}$, we have $\mathrm{Hess}\, f(\boldsymbol{y}) \succeq \gamma \boldsymbol{I}$.

Symmetry Only Creates Equivalent Solutions

Continuous Symmetry

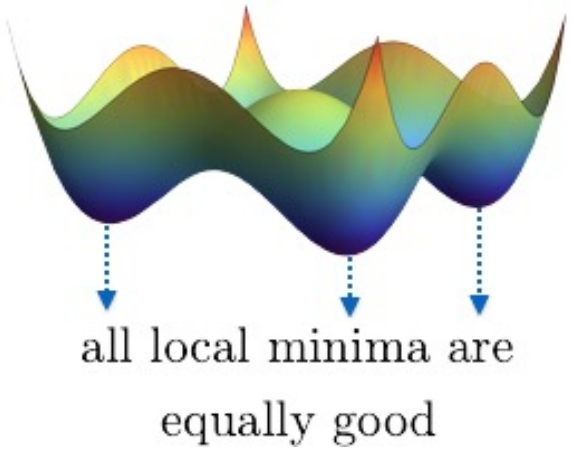


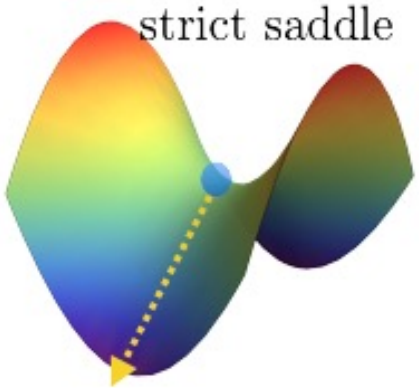
Discrete Symmetry

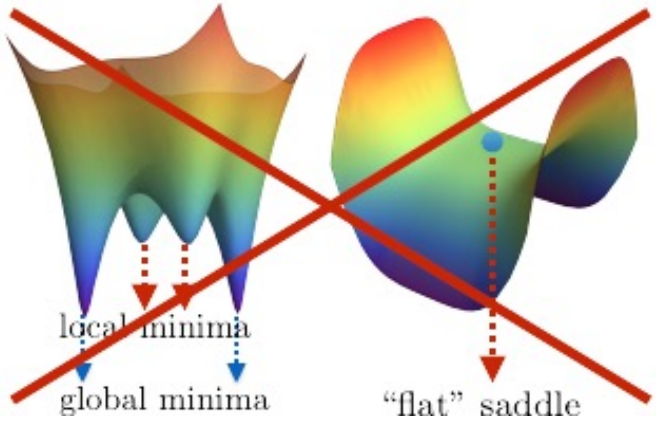


Symmetry creates *only* equivalent *global* minimizers,
but *not* bad local minima.

Optimizing Nonconvex Problems Globally







Benign nonconvex landscapes enable efficient global optimization!

Strict Saddle Problems in SIPML

- **Discrete Symmetry**
  - Generalized Phase Retrieval
  - Low-rank Matrix Recovery
  - ...
- **Continuous Symmetry**
  - (Convolutional) Sparse Dictionary Learning
  - Orthogonal Tensor Decomposition
  - Sparse Blind Deconvolution
  - ...

Further Readings

- *High-Dimensional Data Analysis with Low-Dimensional Models: Principles, Computation, and Applications.* John Wright, Yi Ma. (**Chapter 7**)
- *Escaping From Saddle Points --- Online Stochastic Gradient for Tensor Decomposition.* Rong Ge, Furong Huang, Chi Jin, Yang Yuan, COLT'15, 2015.
- *When Are Nonconvex Problems Not Scary?* Ju Sun, Qing Qu, John Wright, NeurIPS Workshop on Nonconvex Optimization, 2015.
- *From Symmetry to Geometry: Tractable Nonconvex Problems.* Yuqian Zhang, Qing Qu, John Wright, 2021.

Lecture Agenda

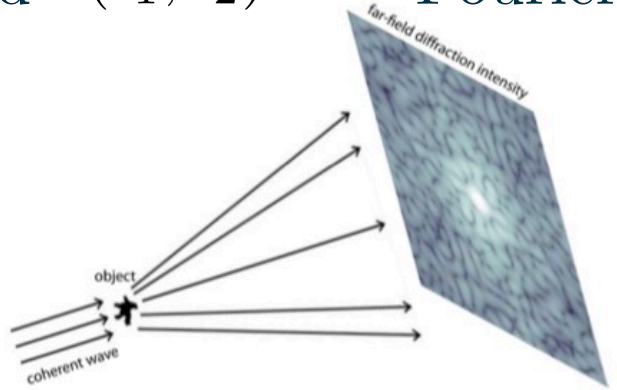
- Problem Introduction
- **Applications in SIPML**
  - Example I: Generalized Phase Retrieval
  - Example II: Low-Rank Matrix Recovery
  - Example III: Training Deep Neural Networks
  - Example IV: Sparse Dictionary Learning
  - Example V: Sparse Blind Deconvolution

Lecture Agenda

- Problem Introduction
- **Applications in SIPML**
  - **Example I: Generalized Phase Retrieval**
  - Example II: Low-Rank Matrix Recovery
  - Example III: Training Deep Neural Networks
  - Example IV: Sparse Dictionary Learning
  - Example V: Sparse Blind Deconvolution

Fourier Phase Retrieval

electric field $x(t_1, t_2)$ $\rightarrow$ Fourier transform $\widehat{x}(f_1, f_2)$



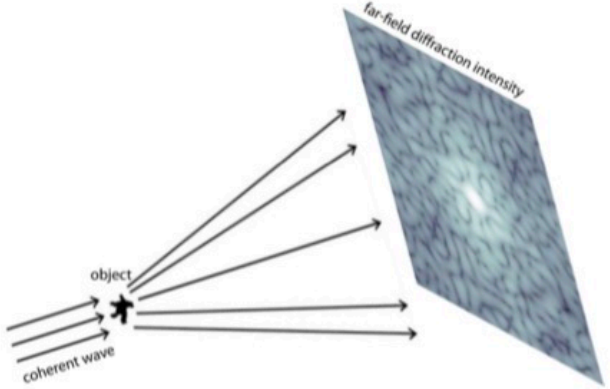
Coherent Diffraction Imaging[1]

Applications: X-ray crystallography, **diffraction imaging** (left), optics, astronomical imaging, and microscopy

Fourier phase retrieval: given intensity $\boldsymbol{y} = |\mathcal{F}(\boldsymbol{x}_\star)|$, recover $\boldsymbol{x}_\star$.

- Fourier phase retrieval is a very challenging ill-posed problem, which is notoriously difficult to solve.

Generalized Phase Retrieval



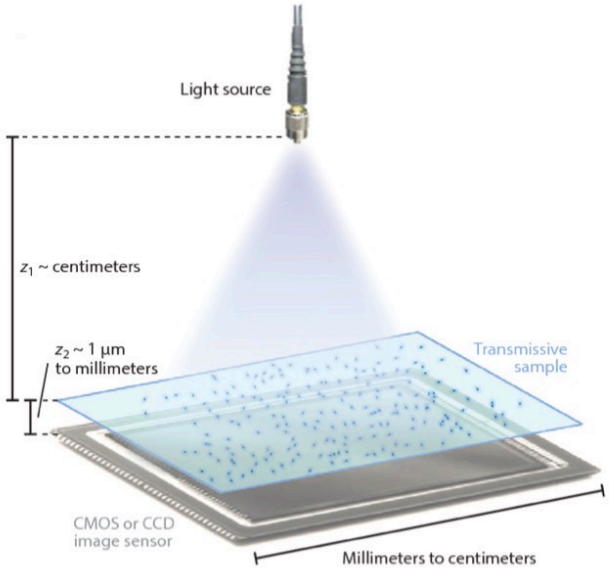
Coherent Diffraction Imaging[1]

Applications: X-ray crystallography, **diffraction imaging** (left), optics, astronomical imaging, and microscopy

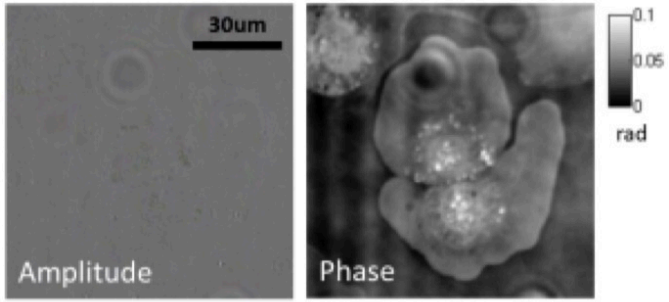
Generalized phase retrieval: given intensity $\boldsymbol{y} = |\boldsymbol{A}\boldsymbol{x}_\star|$, recover $\boldsymbol{x}_\star \in \mathbb{C}^m$.

- The sensing matrix $\boldsymbol{A}$ can be generic and less structured, making the problem easier to solve.

Applications in Microscopy Imaging



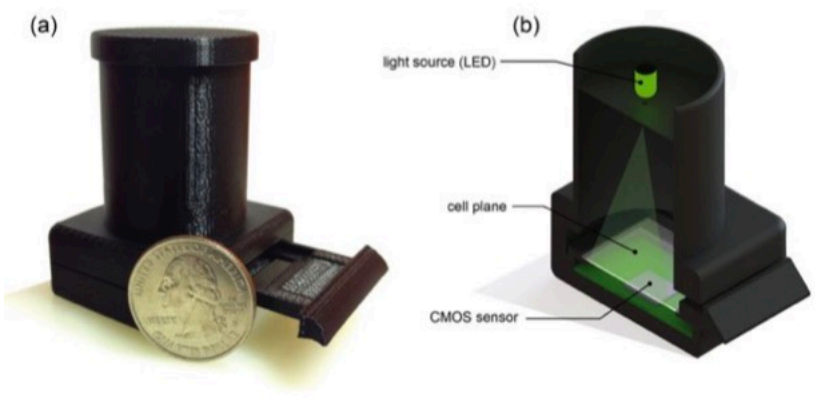
Lensfree microscopy

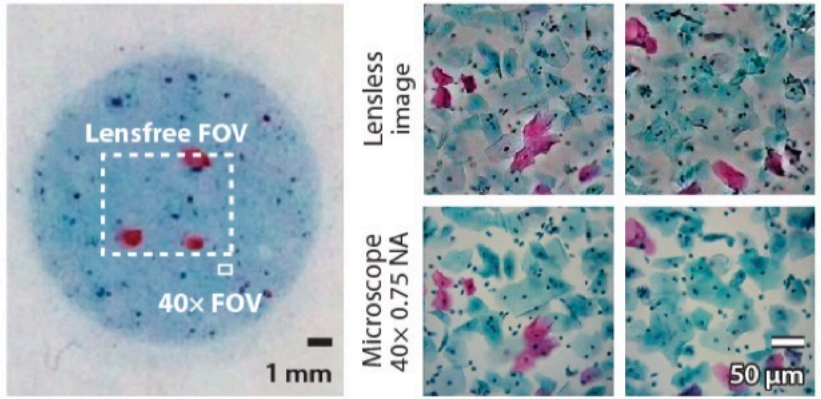


Phase contrast microscopy: phase is important[1].

Lensless Imaging and Sensing. Aydogan Ozcan and Euan McLeod. Annual Review of Biomedical Engineering, 2016.

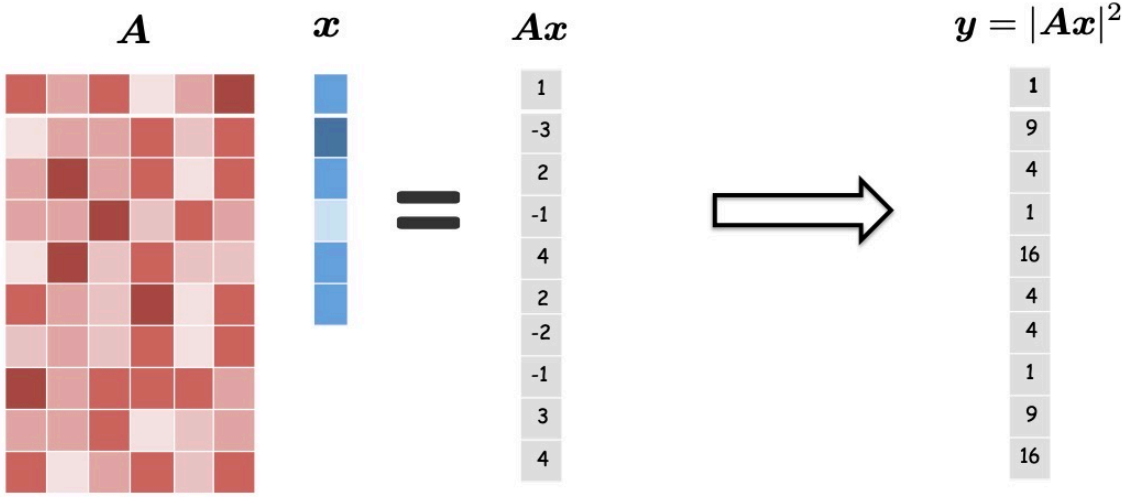
Applications in Microscopy Imaging





Lensfree microscopy imaging via phase retrieval.

Generalized Phase Retrieval



- Solve for $\boldsymbol{x} \in \mathbb{C}^n$ in m quadratic equations

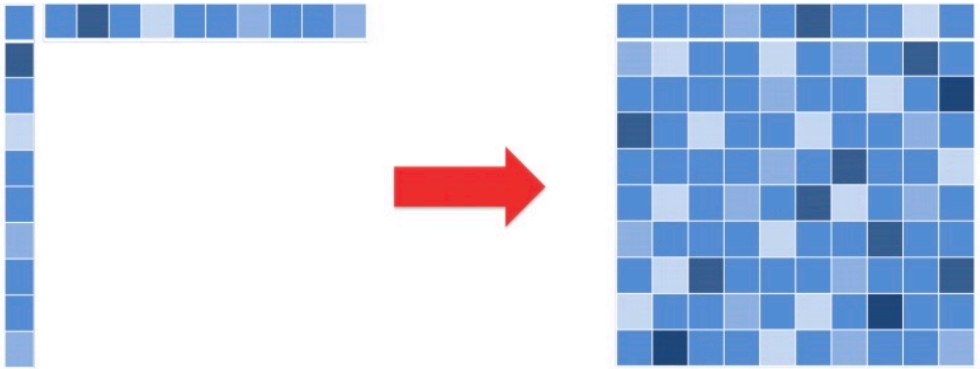
$$y_k = \left|\boldsymbol{a}_k^\top \boldsymbol{x}\right|^2, \quad k = 1, \cdots, m,$$

$$\text{or} \quad \boldsymbol{y} = |\boldsymbol{A}\boldsymbol{x}|^2, \quad \text{where } |\boldsymbol{z}|^2 := \left[|z_1|^2, \cdots, |z_m|^2\right]^\top.$$

Convex Relaxation (Lecture 6)

- **Lifting:** introduce $\boldsymbol{X} = \boldsymbol{x}\boldsymbol{x}^*$ to *linearize* the problem

$$y_k = \left|\boldsymbol{a}_k^* \boldsymbol{x}\right|^2 = \boldsymbol{a}_k^* \underbrace{(\boldsymbol{x}\boldsymbol{x}^*)}_{\boldsymbol{X}} \boldsymbol{a}_k \quad \Longrightarrow \quad y_k = \langle \boldsymbol{a}_k \boldsymbol{a}_k^*, \boldsymbol{X} \rangle$$



Convex Relaxation (Lecture 6)

$$\text{find } \boldsymbol{X} \succeq \mathbf{0}, \quad \text{s.t.} \quad y_k = \langle \boldsymbol{a}_k \boldsymbol{a}_k^*, \boldsymbol{X} \rangle, \quad k = 1, \cdots, m$$
$$\text{rank}(\boldsymbol{X}) = 1$$

- **Convex Relaxation:**

$$\min_{\boldsymbol{X}} \|\boldsymbol{X}\|_*, \quad \text{s.t.} \quad \boldsymbol{y} = \mathcal{A}(\boldsymbol{X}), \ \boldsymbol{X} \succeq \mathbf{0}.$$

Needs to optimize a problem of $O(n^2)$ variables instead of $O(n)$, and SDP problems are usually very expensive.

A Natural Nonconvex Formulation

Given $\boldsymbol{y} \in \mathbb{R}^m$ generated by $\boldsymbol{y} = |\boldsymbol{A}\boldsymbol{x}_\star|$, recover $\boldsymbol{x}_\star \in \mathbb{C}^n$.

- Optimizing the following *nonconvex* least-squares loss:

$$\min_{\boldsymbol{x}\in\mathbb{C}^n} \; f(\boldsymbol{x}) \;=\; \frac{1}{4m}\sum_{i=1}^{m}\left(y_i^2 - |\boldsymbol{a}_i^*\boldsymbol{x}|^2\right)^2$$

It only involves $O(n)$ variables, but the problem is highly *nonconvex* (**why?**)

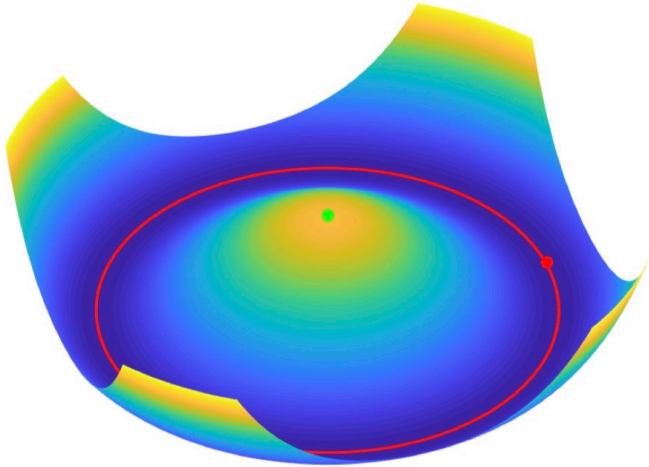
Symmetries in Generalized Phase Retrieval

For any $\phi \in [0, 2\pi)$, we observe

$$\boldsymbol{y} \;=\; |\boldsymbol{A}\boldsymbol{x}_\star| \;=\; \left|\boldsymbol{A}\boldsymbol{x}_\star e^{i\phi}\right|,$$

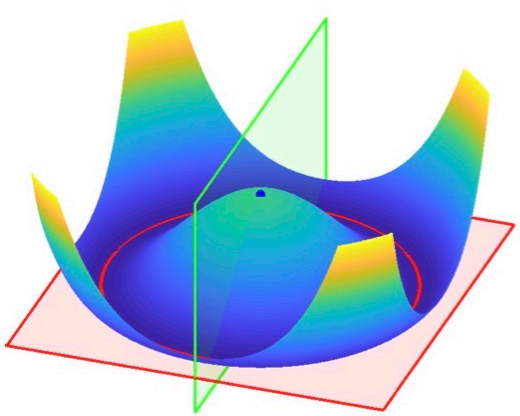
so that

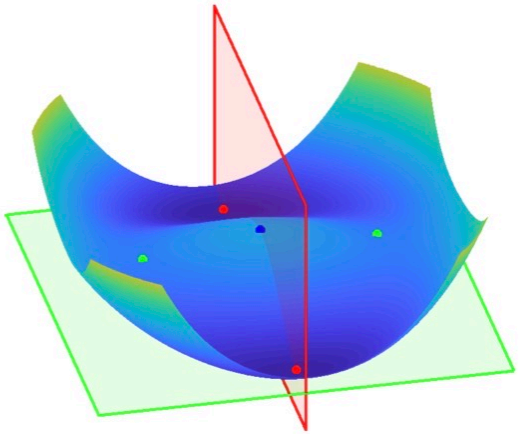
$$\mathcal{X} \;:=\; \left\{\boldsymbol{z} \in \mathbb{C}^n \;|\; \boldsymbol{z} = \boldsymbol{x}_\star e^{i\phi}, \phi \in [0, 2\pi)\right\}.$$

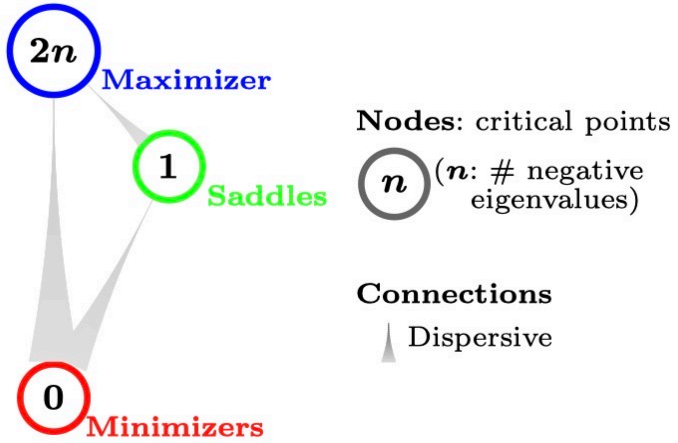


Plot of the landscape with a single unknown $\boldsymbol{y} = \boldsymbol{a}x_\star$.

A Natural Nonconvex Formulation



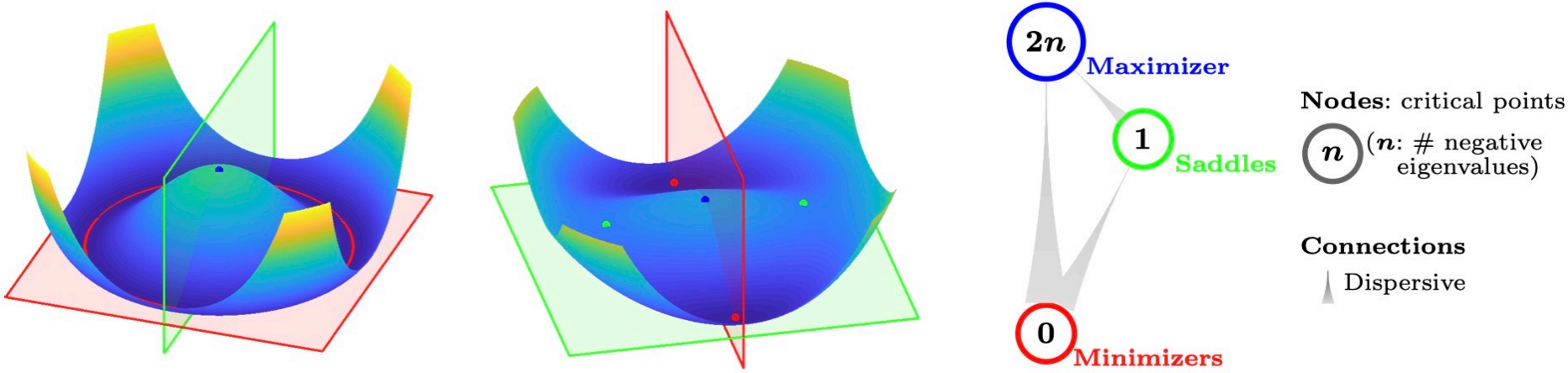




- Optimizing the following *nonconvex* least-squares loss:

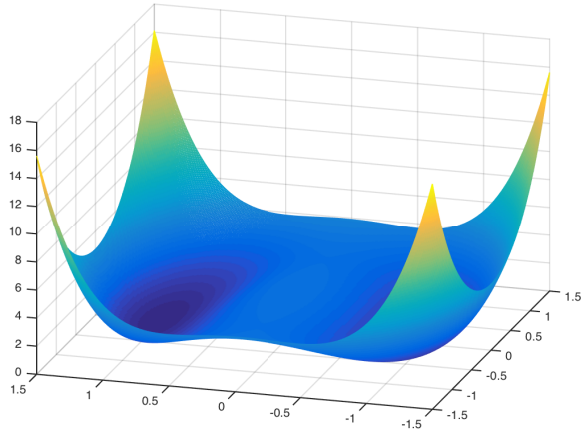
$$\min_{\boldsymbol{x}\in\mathbb{C}^n} \; f(\boldsymbol{x}) \;=\; \frac{1}{4m}\sum_{i=1}^{m}\left(y_i^2 - |\boldsymbol{a}_i^*\boldsymbol{x}|^2\right)^2$$

Generalized Phase Retrieval



- Symmetric copies of the ground truth are minimizers.
- Negative curvature in symmetry-breaking directions.
- Cascade of saddle points.

Generalized Phase Retrieval



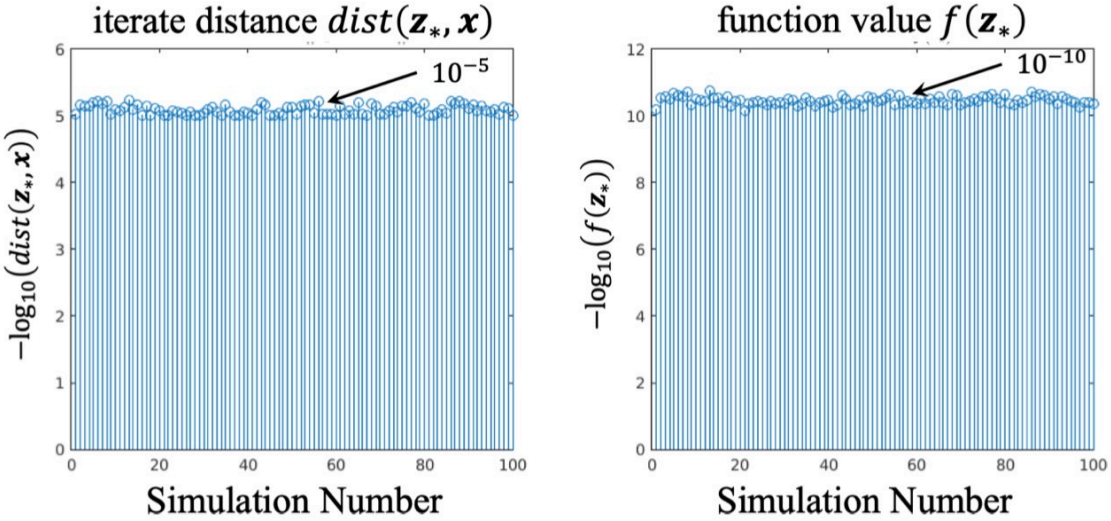
$$\min_{\boldsymbol{x}\in\mathbb{C}^n} \; f(\boldsymbol{x}) \;=\; \frac{1}{4m}\sum_{i=1}^{m}\left(y_i^2 - |\boldsymbol{a}_i^*\boldsymbol{x}|^2\right)^2$$

Theorem. (Informal, Sun et al.'18)

Let $\boldsymbol{a}_k \sim_{i.i.d.} \mathcal{CN}(\boldsymbol{0}, \boldsymbol{I})$. When $m \geq \Omega(n\log^3(n))$, w.h.p.

- All local (and global) minimizers are of the form $\boldsymbol{x}e^{i\phi}$, $\phi \in [0, 2\pi)$;
- $f(\boldsymbol{x})$ is $(c, c/(n\log m), c, c/(n\log m))$-strict saddle for some $c > 0$.

Generalized Phase Retrieval



Run vanilla gradient descent, with random initializations

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \tau_k \cdot \nabla f(\boldsymbol{x}_k)$$

Further Readings

- *High-Dimensional Data Analysis with Low-Dimensional Models: Principles, Computation, and Applications.* John Wright, Yi Ma. (**Chapter 7**)
- *Phase Retrieval via Wirtinger Flow: Theory and Algorithms.* Emmanuel Candes, Xiaodong Li, Mahdi Soltanolkotabi, IEEE Info. Theory, 2014.
- *A Geometric Analysis of Phase Retrieval.* Ju Sun, Qing Qu, John Wright, Foundations of Computational Math, 2018.
- *Phase Retrieval with Application to Optical Imaging: A Contemporary Overview.* Yoav Shechtman, Yonina C. Eldar, Oren Cohen, Henry N. Chapman, Jianwei Miao, Mordechai Segev, IEEE Signal Processing Magazine, 2015.

Lecture Agenda

- Problem Introduction
- **Applications in SIPML**
  - Example I: Generalized Phase Retrieval
  - **Example II: Low-Rank Matrix Recovery**
  - Example III: Training Deep Neural Networks
  - Example IV: Sparse Dictionary Learning
  - Example V: Sparse Blind Deconvolution

Low-rank Matrix Recovery

Given $\boldsymbol{y} = \mathcal{A}(\boldsymbol{X}_\star)$, recover a rank-$r$ (with $r \ll \min\{n_1, n_2\}$) matrix $\boldsymbol{X}_\star \in \mathbb{R}^{n_1 \times n_2}$ from $\boldsymbol{y} \in \mathbb{R}^m$.

- **Convex relaxation approaches:**

$$\min_{\boldsymbol{X}} \|\boldsymbol{X}\|_* \quad \text{s.t.} \quad \boldsymbol{y} = \mathcal{A}(\boldsymbol{X}),$$

or in the noisy setting:

$$\min_{\boldsymbol{X}} \frac{1}{2} \|\boldsymbol{y} - \mathcal{A}(\boldsymbol{X})\|_2^2 + \lambda \|\boldsymbol{X}\|_* .$$

The nuclear minimization problem with $O(n_1 n_2)$ variables could be very expensive (computing SVD $O(n_1^2 n_2)$).

Low-rank Matrix Recovery

Given $\boldsymbol{y} = \mathcal{A}(\boldsymbol{X}_\star)$, recover a rank-$r$ (with $r \ll \min\{n_1, n_2\}$) matrix $\boldsymbol{X}_\star \in \mathbb{R}^{n_1 \times n_2}$ from $\boldsymbol{y} \in \mathbb{R}^m$.

- **A nonconvex approach:** let $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{V}^\top$ with $\boldsymbol{U} \in \mathbb{R}^{n_1 \times r}$ and $\boldsymbol{V} \in \mathbb{R}^{n_2 \times r}$,

$$\min_{\boldsymbol{U},\boldsymbol{V}} \ \frac{1}{2}\left\|\boldsymbol{y} - \mathcal{A}(\boldsymbol{U}\boldsymbol{V}^\top)\right\|_2^2 + \rho(\boldsymbol{U},\boldsymbol{V})$$

  - The problem is nonconvex due to bilinearity $\boldsymbol{U}\boldsymbol{V}^\top$.
  - The number of optimization variables $O(\max\{n_1, n_2\}\, r)$ is much smaller than $O(n_1 n_2)$.

Symmetries in Matrix Recovery

For any *invertible* matrix $\boldsymbol{\Gamma} \in \mathbb{R}^{r\times,r}$

$$\boxed{\boldsymbol{X}_\star \;=\; \boldsymbol{U}_\star \boldsymbol{V}_\star^\top \;=\; \boldsymbol{U}_\star \boldsymbol{\Gamma}\boldsymbol{\Gamma}^{-1}\boldsymbol{V}_\star^\top \;=\; \left(\boldsymbol{U}_\star \boldsymbol{\Gamma}\right)\left(\boldsymbol{V}_\star \boldsymbol{\Gamma}^{-\top}\right)^\top.}$$

so that the problem possess a *general linear* symmetry

$$(\boldsymbol{U}_\star, \boldsymbol{V}_\star) \equiv \left(\boldsymbol{U}_\star \boldsymbol{\Gamma}, \boldsymbol{V}_\star \boldsymbol{\Gamma}^{-\top}\right),\ \forall \boldsymbol{\Gamma} \in \mathrm{GL}(r)$$

- The matrix $\boldsymbol{\Gamma}$ can be *arbitrarily ill-conditioned*, so that $\boldsymbol{U}$ and $\boldsymbol{V}$ can be *arbitrarily unbalanced*.
- It can be reduced to a simpler orthogonal symmetry $\mathcal{O}(r)$ by using a penalty such as $\rho(\boldsymbol{U}, \boldsymbol{V}) = \left\|\boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{V}\boldsymbol{V}^\top\right\|_F^2$
- The landscape is "almost" the same with factorization problem.

Geometry of Symmetric Matrix Factorization

- Study of the nonconvex landscape:

Let us start with a much simpler *symmetric* matrix factorization problem $\boldsymbol{Y} = \boldsymbol{X}_\star$, and we aim to factor it as $\boldsymbol{X}_\star = \boldsymbol{U}_\star \boldsymbol{U}_\star^\top$ by

$$\min_{\boldsymbol{U} \in \mathbb{R}^{n \times r}} \ \varphi(\boldsymbol{U}) \ := \ \frac{1}{4} \left\| \boldsymbol{Y} - \boldsymbol{U}\boldsymbol{U}^\top \right\|_F^2$$

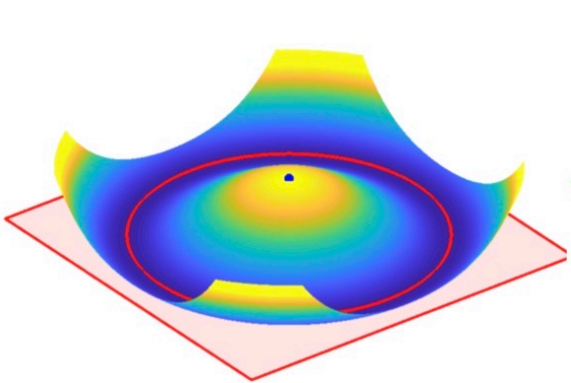
- The problem exhibits a simpler orthogonal symmetry

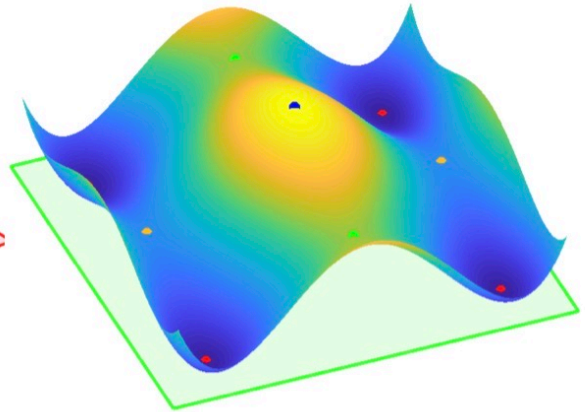
$$\boldsymbol{U}\boldsymbol{U}^\top \ = \ \boldsymbol{U}\boldsymbol{\Gamma}\boldsymbol{\Gamma}^\top\boldsymbol{U}^\top \ = \ (\boldsymbol{U}\boldsymbol{\Gamma})\,(\boldsymbol{U}\boldsymbol{\Gamma})^\top, \quad \forall\, \boldsymbol{\Gamma} \in \mathcal{O}(r),$$

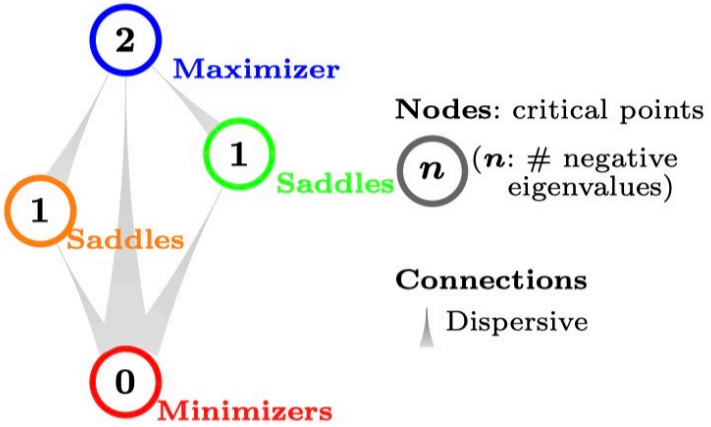
so that $\varphi(\boldsymbol{U}) \equiv \varphi(\boldsymbol{U}\boldsymbol{\Gamma})$.

Geometry of Symmetric Matrix Factorization

$$\min_{\boldsymbol{U}\in\mathbb{R}^{n\times r}} \ \varphi(\boldsymbol{U}) \ := \ \frac{1}{4}\left\|\boldsymbol{Y}-\boldsymbol{U}\boldsymbol{U}^{\top}\right\|_F^2$$

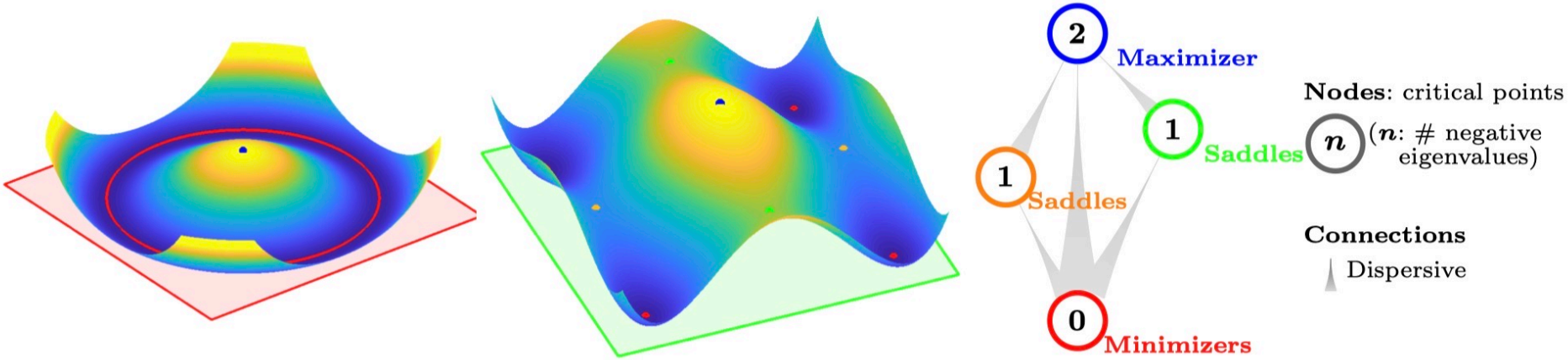






$\boldsymbol{X}_\star$ is a rank-2 symmetric matrix, with $\lambda_1 = \frac{3}{4}$ and $\lambda_2 = \frac{1}{2}$.

Geometry of Symmetric Matrix Factorization



- Symmetric copies of the ground truth are minimizers.
- Negative curvature in symmetry-breaking directions.
- Cascade of saddle points.

From Symmetric to Non-symmetric Factorization

This model geometry carries over to *non-symmetric* matrices,

$$\min_{\boldsymbol{U}\in\mathbb{R}^{n_1\times r},\boldsymbol{V}\in\mathbb{R}^{n_2\times r}} \frac{1}{2}\left\|\boldsymbol{Y}-\boldsymbol{U}\boldsymbol{V}^\top\right\|_F^2+\rho(\boldsymbol{U},\boldsymbol{V})$$

by penalization such as $\rho(\boldsymbol{U},\boldsymbol{V})=\left\|\boldsymbol{U}\boldsymbol{U}^\top-\boldsymbol{V}\boldsymbol{V}^\top\right\|_F^2$, that we can reduce $\mathrm{GL}(r)$ to $\mathcal{O}(r)$ symmetry.

From Factorization to Matrix Recovery & Completion

Given $\boldsymbol{y} = \mathcal{A}(\boldsymbol{X}_\star)$ with $y_i = \langle \boldsymbol{A}_i, \boldsymbol{X}_\star \rangle$ $(1 \le i \le m)$, optimize

$$\min_{\boldsymbol{U} \in \mathbb{R}^{n_1 \times r}, \boldsymbol{V} \in \mathbb{R}^{n_2 \times r}} \frac{1}{2} \left\| \boldsymbol{y} - \mathcal{A}(\boldsymbol{U}\boldsymbol{V}^\top) \right\|_2^2 + \rho(\boldsymbol{U}, \boldsymbol{V})$$

- **Matrix factorization:** $\mathcal{A} = \mathcal{I}$, with $\mathcal{I}$ being an identity mapping.
- **Matrix sensing:** $\mathcal{A}(\cdot)$ satisfies certain rank RIP property:

$$\left| \|\mathcal{A}(\boldsymbol{X}_r)\|_2^2 - \|\boldsymbol{X}_r\|_F^2 \right| \le \delta \|\boldsymbol{X}_r\|_F^2, \quad \forall \mathrm{rank}(\boldsymbol{X}_r) \le r.$$

It can be showed that the objective is strict saddle with benign global landscape.

From Factorization to Matrix Recovery & Completion

Given $\boldsymbol{y} = \mathcal{A}(\boldsymbol{X}_\star)$ with $y_i = \langle \boldsymbol{A}_i, \boldsymbol{X}_\star \rangle$ $(1 \le i \le m)$, optimize

$$\min_{\boldsymbol{U} \in \mathbb{R}^{n_1 \times r}, \boldsymbol{V} \in \mathbb{R}^{n_2 \times r}} \frac{1}{2} \left\| \boldsymbol{y} - \mathcal{A}(\boldsymbol{U}\boldsymbol{V}^\top) \right\|_2^2 + \rho(\boldsymbol{U}, \boldsymbol{V})$$

- **Matrix completion:** $\mathcal{A}(\cdot) = \mathcal{P}_\Omega(\cdot)$, where $\mathcal{P}_\Omega(\cdot)$ is not RIP in general , that extra assumption on $\boldsymbol{X}_\star$ (incoherence) and extra penalization is needed for global benign landscape.

$$\rho_{mc}(\boldsymbol{U}, \boldsymbol{V}) = \lambda_1 \sum_{i=1}^{n_1} \left(\left\| \boldsymbol{e}_i^\top \boldsymbol{U} \right\|_2 - \alpha_1 \right)_+^4 + \lambda_2 \sum_{i=1}^{n_2} \left(\left\| \boldsymbol{e}_i^\top \boldsymbol{U} \right\|_2 - \alpha_2 \right)_+^4 .$$

From Factorization to Robust Matrix Recovery

Given $\boldsymbol{Y} = \mathcal{A}(\boldsymbol{X}_\star) + \boldsymbol{S}_\star$ with *low-rank* $\boldsymbol{X}_\star$ and *sparse* $\boldsymbol{S}_\star$, optimize

$$\min_{\boldsymbol{U},\boldsymbol{V},\boldsymbol{S}} \ \frac{1}{2}\left\|\boldsymbol{Y} - \mathcal{A}(\boldsymbol{U}\boldsymbol{V}^\top) - \boldsymbol{S}\right\|_F^2 + \mu\left\|\boldsymbol{S}\right\|_1 + \rho_r(\boldsymbol{U},\boldsymbol{V})$$

- **Robust matrix recovery:** partial minimization w.r.t. $\boldsymbol{S}$ gives

$$\min_{\boldsymbol{U},\boldsymbol{V},\boldsymbol{S}} \ \frac{1}{2} H_\mu\left(\boldsymbol{Y} - \mathcal{A}(\boldsymbol{U}\boldsymbol{V}^\top)\right) + \rho_r(\boldsymbol{U},\boldsymbol{V}),$$

with $H_\mu(\cdot)$ being Huber loss. Thus, we can perform similar analysis to problems studied previously.

Further Readings

- *High-Dimensional Data Analysis with Low-Dimensional Models: Principles, Computation, and Applications*. John Wright, Yi Ma. (**Chapter 7**)

- *Matrix Completion has No Spurious Local Minimum.* Rong Ge, Jason D. Lee, Tengyu Ma, NeurIPS'16, 2016.

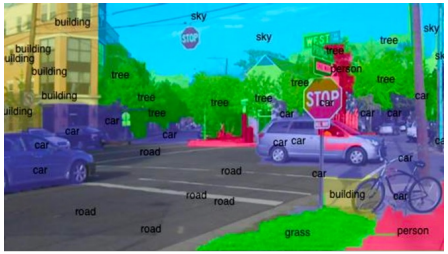
- *No Spurious Local Minima in Nonconvex Low Rank Problems: A Unified Geometric Analysis.* Rong Ge, Chi Jin, Yi Zheng, ICML 2017.

Lecture Agenda

- Problem Introduction
- **Applications in SIPML**
  - Example I: Generalized Phase Retrieval
  - Example II: Low-Rank Matrix Recovery
  - **Example III: Training Deep Neural Networks**
  - Example IV: Sparse Dictionary Learning
  - Example V: Sparse Blind Deconvolution

Deep Neural Networks

Deep learning has attained superior performances for many tasks in practice:

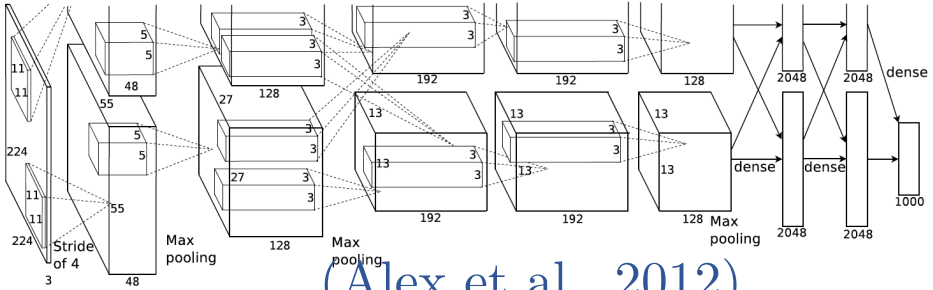


Computer vision

Natural language processing

Gameplay

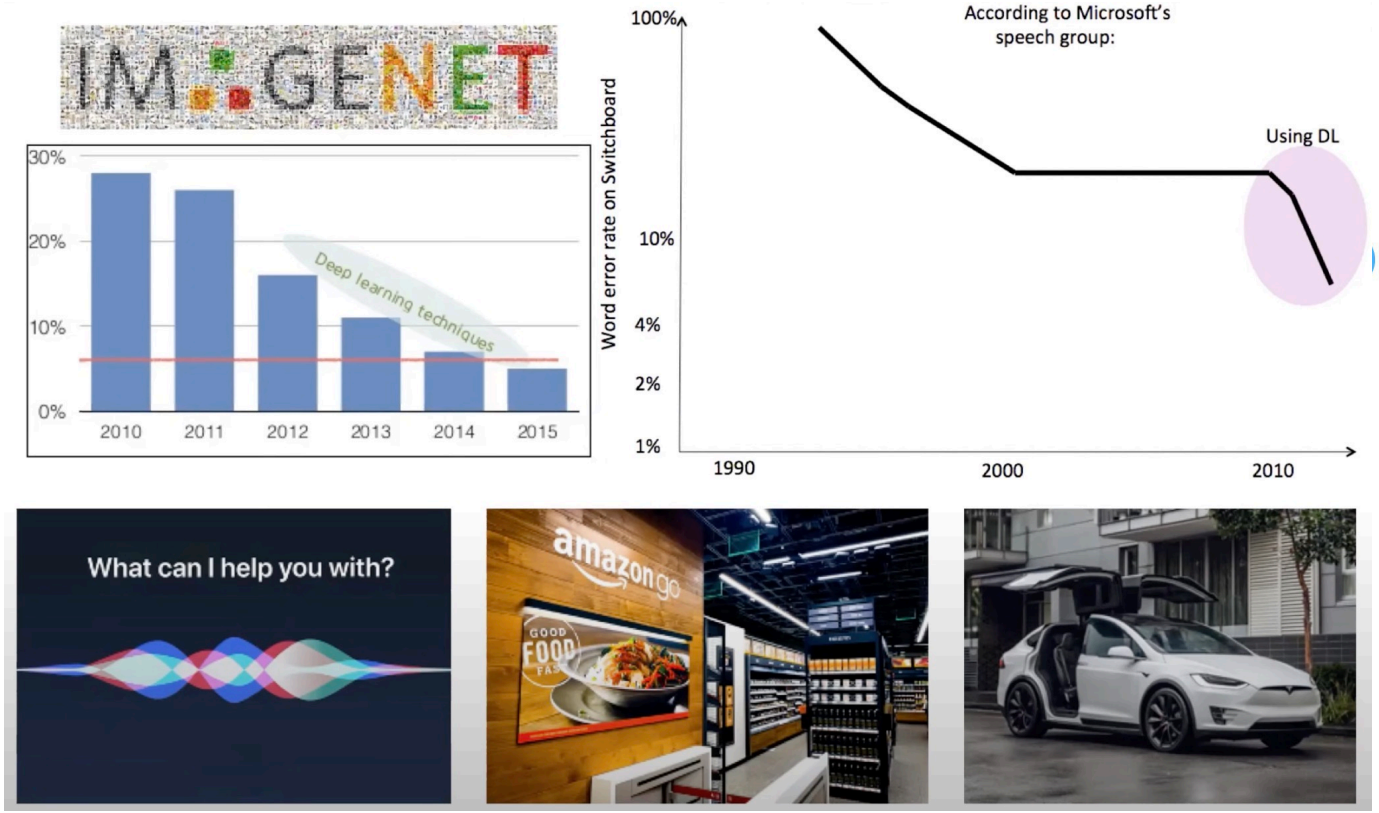
Protein modeling



"Cat"

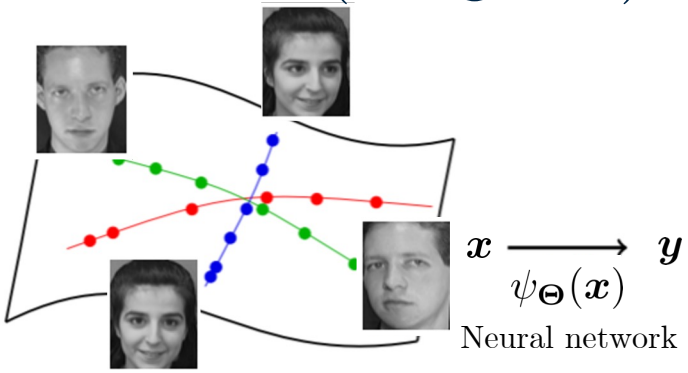
(Alex et al., 2012)

Impact of Deep Learning in ML



Multi-Class Classification Problem

- Learning predictor from **training set** $\{(x^{(i)}, \boldsymbol{y}^{(i)});\ i = 1, \cdots, n\}$
- Example: multi-class classification problem
  - $K = 10$ classes (MNIST, CIFAR10, etc.)
  - $K = 1000$ classes (ImageNet)



Data in the input space

$$\begin{bmatrix}1\\0\\\vdots\\0\end{bmatrix} \begin{bmatrix}0\\1\\\vdots\\0\end{bmatrix} \cdots \begin{bmatrix}0\\0\\\vdots\\1\end{bmatrix}$$

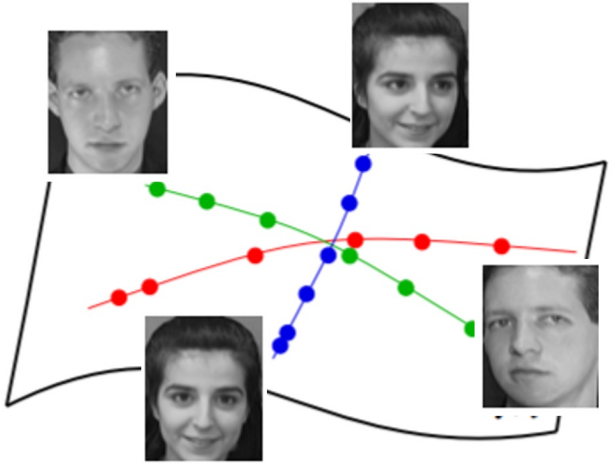
One-hot vectors in $\mathbb{R}^K$

Deep Neural Network Classifiers

$$\psi_{\Theta}(\boldsymbol{x}) = \boldsymbol{W}_L \underbrace{\sigma\left(\boldsymbol{W}_{L-1}\cdots\sigma(\boldsymbol{W}_1\boldsymbol{x}+\boldsymbol{b}_1)+\boldsymbol{b}_{L-1}\right)}_{\phi_{\boldsymbol{\theta}}(\boldsymbol{x})=:\boldsymbol{h}} + \boldsymbol{b}_L$$

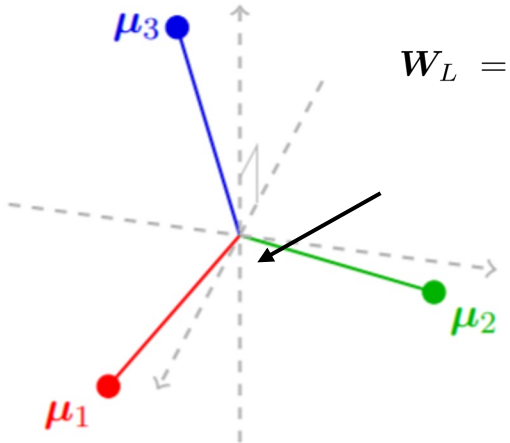
Last-layer classifier

Last-layer feature



Data in the Input Space

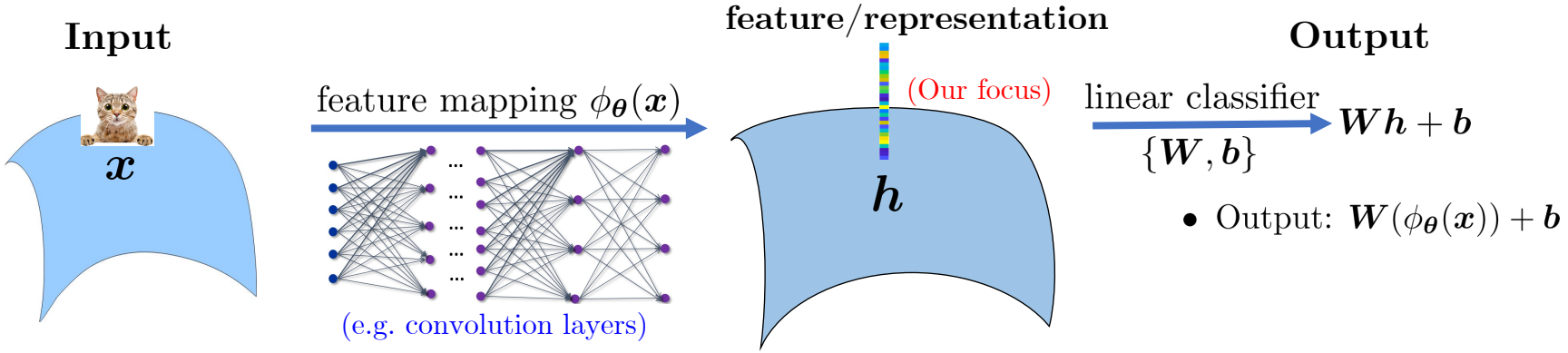
$\phi_{\boldsymbol{\theta}}(\cdot)$



$$\boldsymbol{W}_L = \begin{bmatrix}\boldsymbol{\mu}_1 & \cdots & \boldsymbol{\mu}_K\end{bmatrix}^\top.$$

Neural Collapse
in the Feature Space

Deep Neural Network Classifiers



- **Output:** $\boldsymbol{W} \cdot \phi_{\boldsymbol{\theta}}(\boldsymbol{x}) + \boldsymbol{b}$
- **Training problem:** assume balanced dataset with each class n training samples

$$\min_{\boldsymbol{\theta}, \boldsymbol{W}, \boldsymbol{b}} \frac{1}{Kn} \sum_{k=1}^{K} \sum_{i=1}^{n} \mathcal{L}_{\mathrm{CE}}\left(\boldsymbol{W}(\phi_{\boldsymbol{\theta}}(\boldsymbol{x}_{k,i})) + \boldsymbol{b}, \boldsymbol{y}_k\right) + \lambda \|(\boldsymbol{\theta}, \boldsymbol{W}, \boldsymbol{b})\|_F^2$$

cross-entropy loss

Weight decay

i-th input in the k-th class

One-hot vector for the k-th class

Neural Collapse in Classification

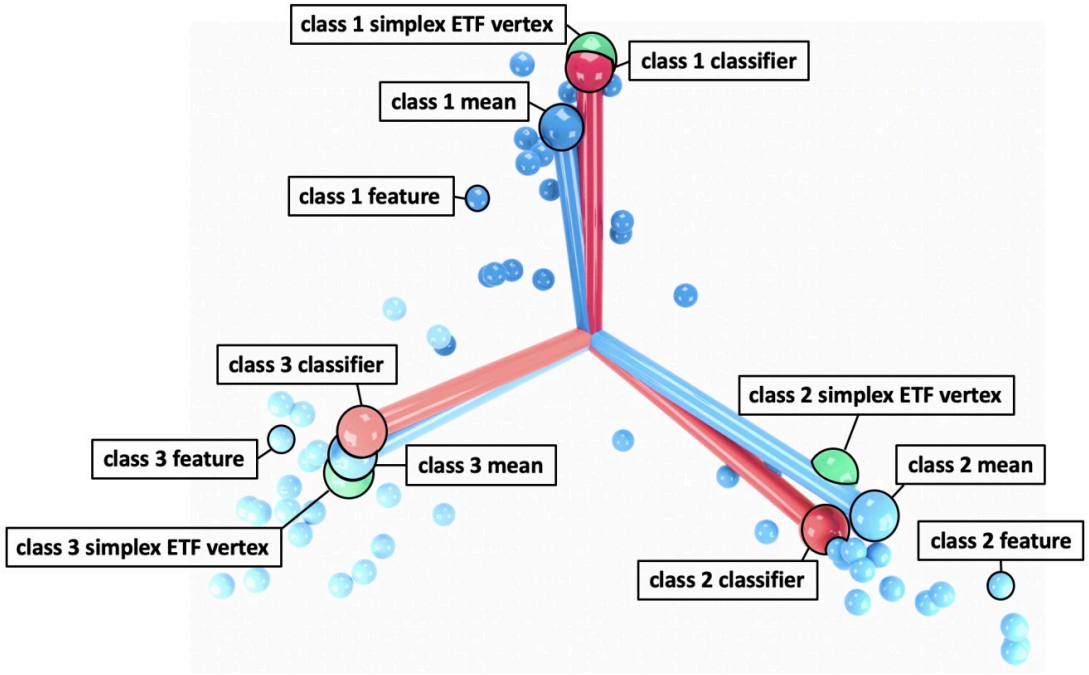


Image credited to Han et al. "Neural Collapse Under MSE Loss: Proximity to and Dynamics on the Central Path"

Neural Collapse in Classification

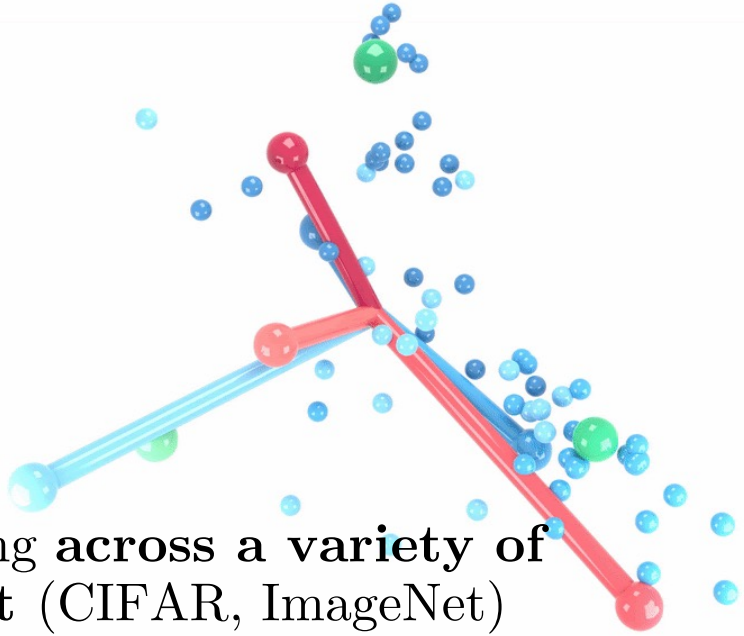
Prevalence of neural collapse during the terminal phase of deep learning training

Vardan Papyan, 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)



- Reveals common outcome of network training **across a variety of architectures** (ResNet, VGG) and **dataset** (CIFAR, ImageNet)
- **Precise mathematical structures** within the features and classifier

Neural Collapse: Symmetry and Structures

Balanced training dataset with $n = n_1 = n_2 = \cdots = n_K$, and

$$\boldsymbol{W} := \boldsymbol{W}_L, \quad \boldsymbol{H} := \begin{bmatrix} \boldsymbol{h}_{1,1} & \cdots & \boldsymbol{h}_{K,n} \end{bmatrix}.$$

Neural Collapse (NC) means that

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

$$\boldsymbol{h}_{k,i} \to \overline{\boldsymbol{h}}_k, \quad \forall k \in [K],\ i \in [n].$$

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

$$\overline{\boldsymbol{H}}^\top \overline{\boldsymbol{H}} \sim \boldsymbol{I}_K - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top, \qquad \overline{\boldsymbol{H}} = \begin{bmatrix} \overline{\boldsymbol{h}}_1 & \cdots & \overline{\boldsymbol{h}}_K \end{bmatrix}$$

Neural Collapse: Symmetry and Structures

Balanced training dataset with $n = n_1 = n_2 = \cdots = n_K$, and

$$\boldsymbol{W} := \boldsymbol{W}_L, \quad \boldsymbol{H} := \begin{bmatrix} \boldsymbol{h}_{1,1} & \cdots & \boldsymbol{h}_{K,n} \end{bmatrix}.$$

Neural Collapse (NC) means that

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

$$\boldsymbol{w}^k = \beta \overline{\boldsymbol{h}}_k, \quad \forall\, k \in [K].$$

4) ***Simple Decision Rule*** via Nearest Class-Center decision.

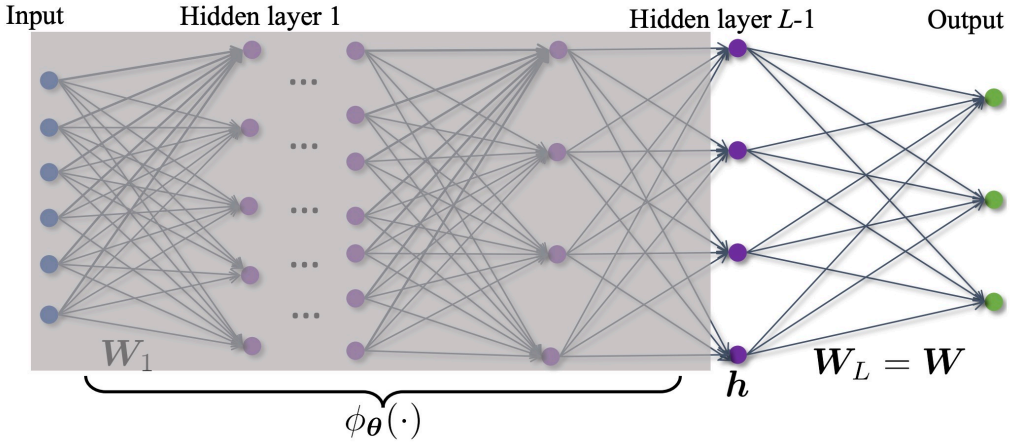
Simplification: Unconstrained Features

$$\psi_{\boldsymbol{\Theta}}(\boldsymbol{x}) = \boldsymbol{W}_L \underbrace{\sigma\left(\boldsymbol{W}_{L-1} \cdots \sigma(\boldsymbol{W}_1 \boldsymbol{x} + \boldsymbol{b}_1) + \boldsymbol{b}_{L-1}\right)}_{\phi_{\boldsymbol{\theta}}(\boldsymbol{x}) =: \boldsymbol{h}} + \boldsymbol{b}_L$$

Last-layer classifier

Last-layer feature

Treat $\boldsymbol{H} = \begin{bmatrix} \boldsymbol{h}_{1,1} & \cdots & \boldsymbol{h}_{K,n} \end{bmatrix}$ as a free optimization variable



Simplification: Unconstrained Features

$$\psi_{\boldsymbol{\Theta}}(\boldsymbol{x}) = \boldsymbol{W}_L \underbrace{\sigma\left(\boldsymbol{W}_{L-1}\cdots\sigma(\boldsymbol{W}_1\boldsymbol{x}+\boldsymbol{b}_1)+\boldsymbol{b}_{L-1}\right)}_{\phi_{\boldsymbol{\theta}}(\boldsymbol{x})=:\boldsymbol{h}} + \boldsymbol{b}_L$$

Last-layer classifier

Last-layer feature

Treat $\boldsymbol{H} = \begin{bmatrix}\boldsymbol{h}_{1,1} & \cdots & \boldsymbol{h}_{K,n}\end{bmatrix}$ as a free optimization variable

$$\min_{\boldsymbol{W},\boldsymbol{H},\boldsymbol{b}} \frac{1}{Kn}\sum_{k=1}^{K}\sum_{i=1}^{n}\mathcal{L}_{\mathrm{CE}}(\boldsymbol{W}\boldsymbol{h}_{k,i}+\boldsymbol{b},\boldsymbol{y}_k) + \frac{\lambda_{\boldsymbol{W}}}{2}\|\boldsymbol{W}\|_F^2 + \frac{\lambda_{\boldsymbol{H}}}{2}\|\boldsymbol{H}\|_F^2 + \frac{\lambda_{\boldsymbol{b}}}{2}\|\boldsymbol{b}\|_2^2$$

- **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.

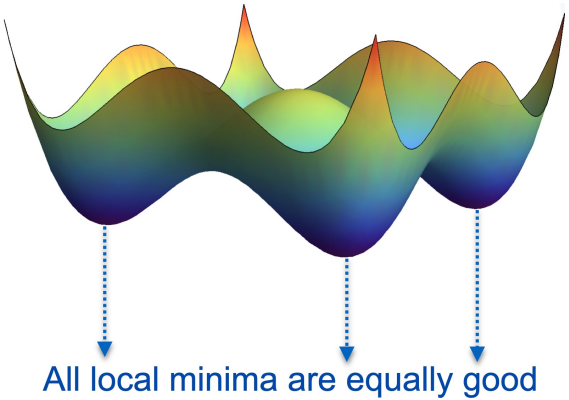
Main Theoretical Results

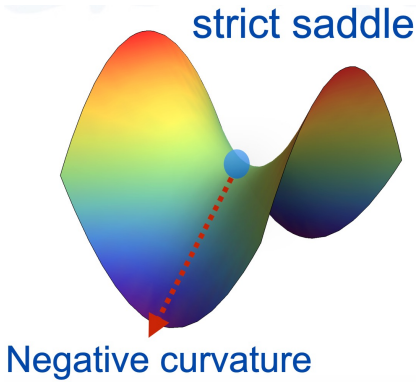
Theorem (Informal) *Consider the nonconvex loss with unconstrained feature model with* $K < d$ and balanced data

$$\min_{\boldsymbol{W},\boldsymbol{H},\boldsymbol{b}} \frac{1}{Kn} \sum_{k=1}^{K} \sum_{i=1}^{n} \mathcal{L}_{\mathrm{CE}}(\boldsymbol{W}\boldsymbol{h}_{k,i} + \boldsymbol{b}, \boldsymbol{y}_k) + \frac{\lambda_{\boldsymbol{W}}}{2} \|\boldsymbol{W}\|_F^2 + \frac{\lambda_{\boldsymbol{H}}}{2} \|\boldsymbol{H}\|_F^2 + \frac{\lambda_{\boldsymbol{b}}}{2} \|\boldsymbol{b}\|_2^2$$

- ***(Global Optimality)*** *Any global solution* $(\mathbf{W}_\star, \mathbf{H}_\star)$ *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





- Closely relates to **low-rank matrix factorization** problems [Burer et al'03, Bhojanapalli et al'16, Ge et al'16, Zhu et al'18,Li et al'19, Chi et al'19]
  - **Difference in tasks:** classification training vs recovery
  - **Difference in loss functions statistical properties :** cross-entropy vs least-squares; randomness or statistical properties of the sensing matrices
  - **Difference in global solutions**.

Further Readings

- *Prevalence of Neural Collapse during the terminal phase of deep learning training.* Vardan Papyan, X.Y. Han, David L. Donoho, Proceedings of the National Academy of Sciences, 2020.

- *A geometric analysis of Neural Collapse with unconstrained features.* Zhihui Zhu, Tianyu Ding, Jinxin Zhou, Xiao Li, Chong You, Jeremias Sulam, Qing Qu, NeurIPS, 2021.

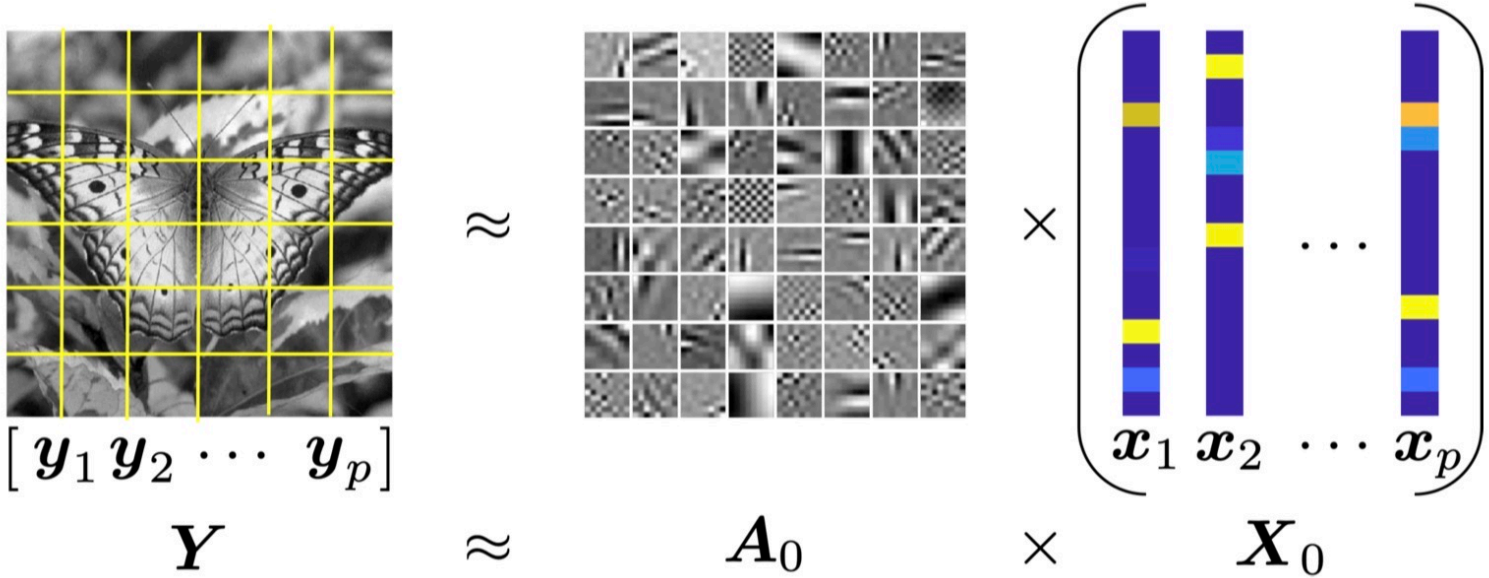
- *Exploring deep neural networks via layer-peeled model: Minority collapse in imbalanced training.* Cong Fang, Hangfeng He, Qi Long, Weijie J. Su, Proceedings of the National Academy of Sciences, 2021.

Lecture Agenda

- Problem Introduction
- **Applications in SIPML**
  - Example I: Generalized Phase Retrieval
  - Example II: Low-Rank Matrix Recovery
  - Example III: Training Deep Neural Networks
  - **Example IV: Sparse Dictionary Learning**
  - Example V: Sparse Blind Deconvolution

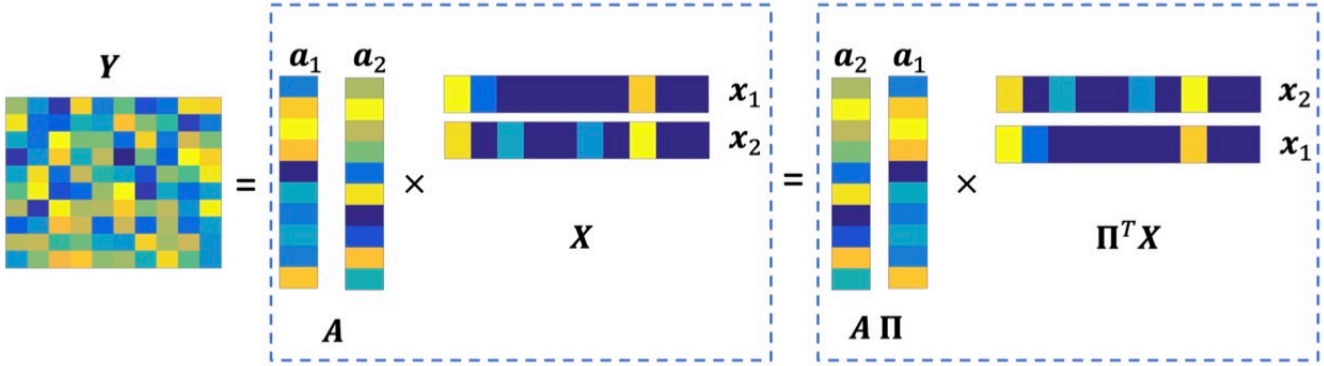
Dictionary Learning

Given $\boldsymbol{Y}$, jointly learn a *compact* dictionary $\boldsymbol{A}_0$ and *sparse* $\boldsymbol{X}_0$?



Symmetry in Dictionary Learning

Nonconvexity due to symmetry:



- **Permutation symmetry:** ($2^n n!$ signed permutation $\mathbf{\Pi}$)

$$\boldsymbol{Y} \;=\; \boldsymbol{A}_0\boldsymbol{X}_0 \;=\; (\boldsymbol{A}_0\boldsymbol{\Pi})\left(\boldsymbol{\Pi}^\top \boldsymbol{X}_0\right)$$

- **Equivalent solution pairs:** $(\boldsymbol{A}_0, \boldsymbol{X}_0) \Longleftrightarrow \left(\boldsymbol{A}_0\boldsymbol{\Pi}, \boldsymbol{\Pi}^\top \boldsymbol{X}_0\right).$

Dictionary Learning – Complete Case

We first study the problem when $\boldsymbol{A}_0 \in \mathbb{R}^{n\times n}$ is *complete*, that is, $\boldsymbol{A}_0$ is square and invertible.

$$\overline{\boldsymbol{Y}} \;=\; \left(\boldsymbol{Y}\boldsymbol{Y}^\top\right)^{-1/2}\boldsymbol{Y} \propto \overline{\boldsymbol{A}}_0\boldsymbol{X}_0, \quad \overline{\boldsymbol{A}}_0 \;\in\; O(n).$$

- We find **one column** of $\overline{\boldsymbol{A}}_0$ via

$$\min_{\boldsymbol{q}} \varphi(\boldsymbol{q}^\top\overline{\boldsymbol{Y}}), \quad \text{s.t.} \quad \boldsymbol{q} \in \mathbb{S}^{n-1}.$$

- $\varphi(\cdot)$ is a sparsity promoting function, $\varphi(\cdot) = \|\cdot\|_1, \varphi(\cdot) = -\|\cdot\|_4^4$
- assume that sparse $\boldsymbol{X}_0 \sim$ Bernoulli-Gaussian(θ).

Dictionary Learning – Complete Case

We find **one column** of $\overline{\boldsymbol{A}}_0$ via

$$\boxed{\min_{\boldsymbol{q}} \varphi(\boldsymbol{q}^\top \overline{\boldsymbol{Y}}), \quad \text{s.t.} \quad \boldsymbol{q} \in \mathbb{S}^{n-1}.}$$

- $\varphi(\cdot)$ is a sparsity promoting function, $\varphi(\cdot) = \|\cdot\|_1, \varphi(\cdot) = -\|\cdot\|_4^4$
- assume that sparse $\boldsymbol{X}_0 \sim \text{Bernoulli-Gaussian}(\theta)$.

High-level intuition: for any $\boldsymbol{q}_\star = \pm \overline{\boldsymbol{a}}_{0i}$

$$\mathbb{E}_{\boldsymbol{X}_0}\left[\varphi(\boldsymbol{q}_\star^\top \overline{\boldsymbol{Y}})\right] \propto \varphi(\boldsymbol{q}_\star^\top \overline{\boldsymbol{A}}_0) = \varphi(\boldsymbol{e}_i)$$

Dictionary Learning – Complete Case

We find **one column** of $\overline{\boldsymbol{A}}_0$ via

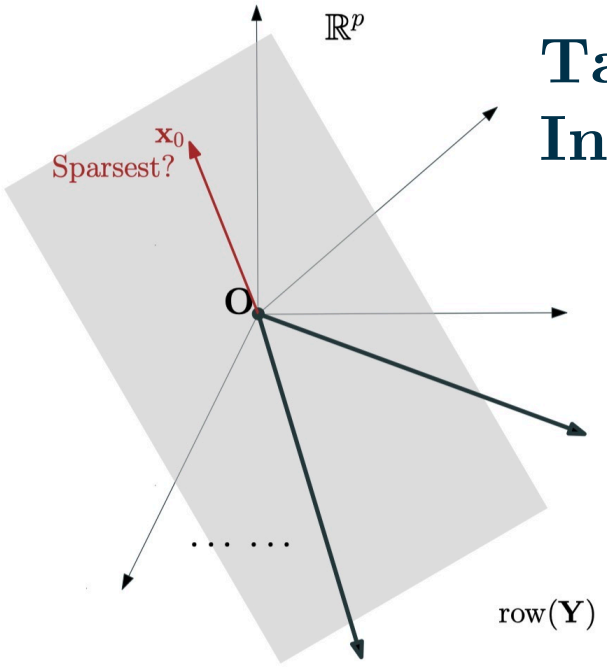
$$\boxed{\min_{\boldsymbol{q}} \left\| \boldsymbol{q}^\top \overline{\boldsymbol{Y}}) \right\|_1, \quad \text{s.t.} \quad \boldsymbol{q} \in \mathbb{S}^{n-1}.}$$

High-level intuition: for any $\boldsymbol{q}_\star = \pm \overline{\boldsymbol{a}}_{0i}$

$$\mathbb{E}_{\boldsymbol{X}_0} \left[\left\| \boldsymbol{q}_\star^\top \overline{\boldsymbol{Y}} \right\|_1 \right] \propto \left\| \boldsymbol{q}_\star^\top \overline{\boldsymbol{A}}_0 \right\|_1 = \left\| \boldsymbol{e}_i \right\|_1$$

The columns of $\overline{\boldsymbol{A}}_0$ are equivalent global solutions!

Alternative Interpretations: Finding the Sparsest Vector in a Subspace



Task: Given $\boldsymbol{Y} = \boldsymbol{A}_0 \boldsymbol{X}_0$, recover $\boldsymbol{A}_0$ and $\boldsymbol{X}_0$

Intuition: $\boldsymbol{A}_0$ is square, invertible,

$$\mathrm{row}(\boldsymbol{Y}) = \mathrm{row}(\boldsymbol{X}_0)$$

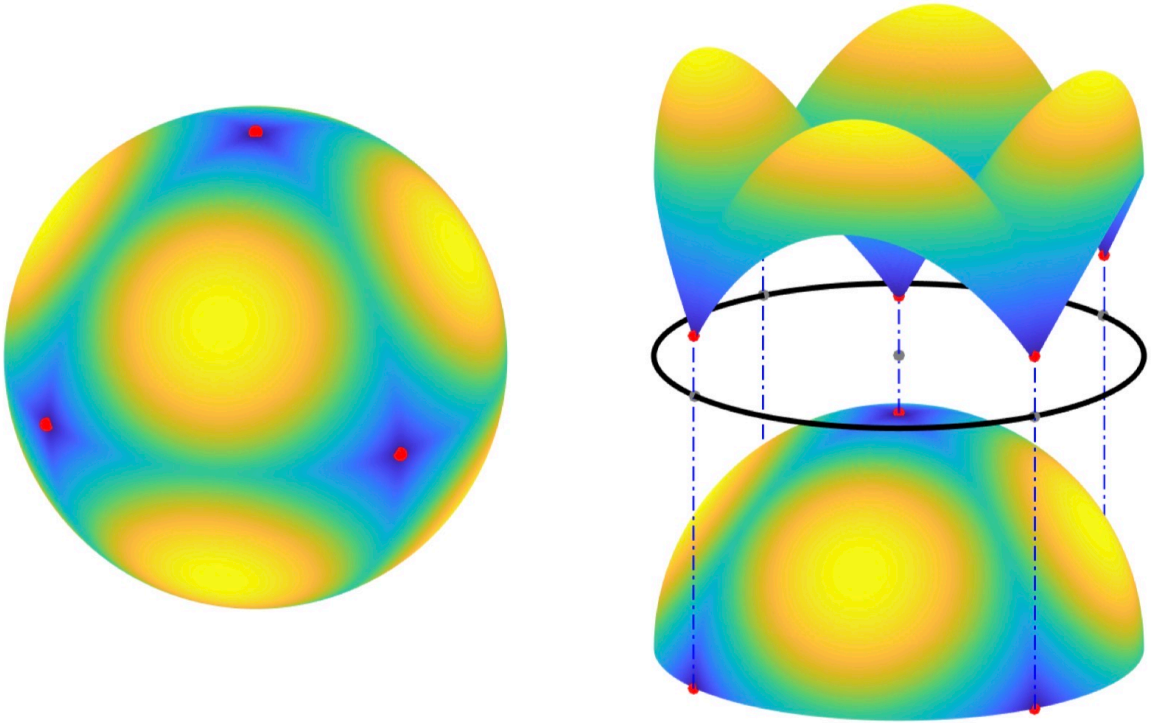
- Find the sparsest vectors in $\mathrm{row}(\boldsymbol{Y})$:

$$\min_{\boldsymbol{q}} \left\| \boldsymbol{q}^\top \boldsymbol{Y} \right\|_0, \quad \text{s.t.} \quad q \neq \mathbf{0}.$$

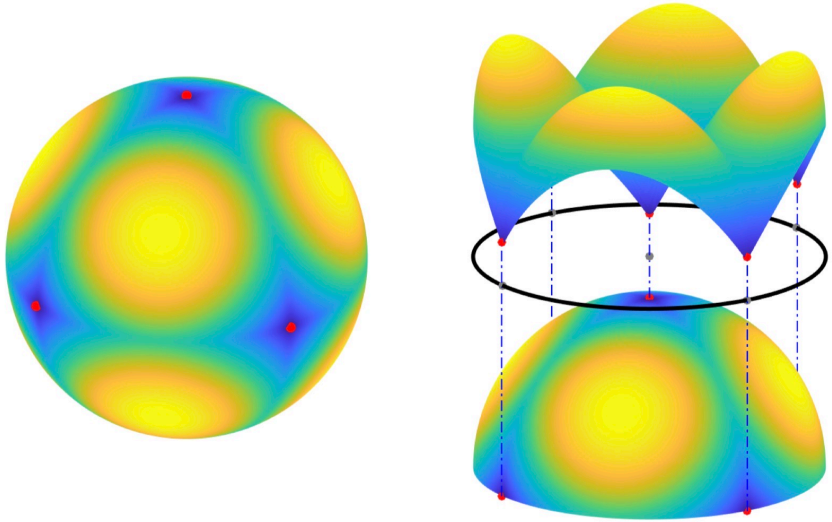
- Nonconvex "relaxation":

$$\min_{\boldsymbol{q}} \left\| \boldsymbol{q}^\top \boldsymbol{Y} \right\|_1, \quad \text{s.t.} \quad \|\boldsymbol{q}\|_2 = 1.$$

Dictionary Learning – Complete Case



Dictionary Learning – Complete Case



- Symmetric copies of the ground truth are minimizers.
- Negative curvature in symmetry-breaking directions.
- Cascade of saddle points.

From Complete Case to Overcomplete Case

We can extend the study to the *overcomplete* case $\boldsymbol{A}_0 \in \mathbb{R}^{n\times m}$ $(m > n)$ under the assumptions that

- Unit norm tight frame:

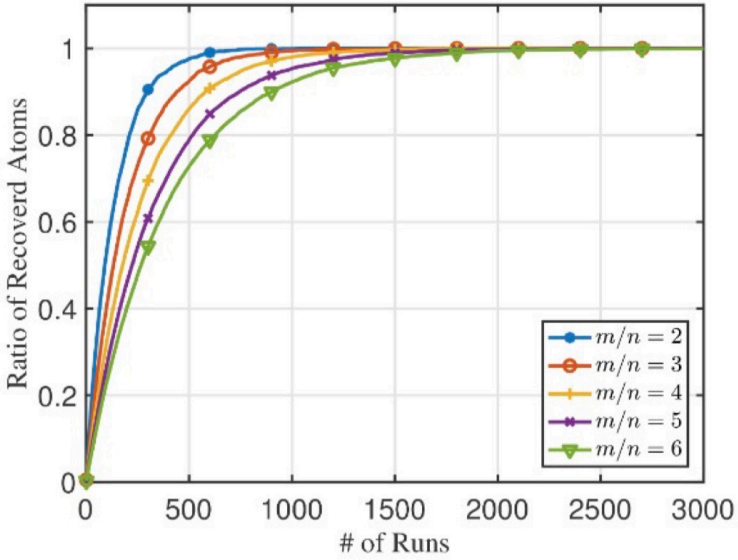
$$\sqrt{\frac{n}{m}}\boldsymbol{A}_0\boldsymbol{A}_0 \;=\; \boldsymbol{I}, \quad \|\boldsymbol{a}_{0i}\|_2 \;=\; 1.$$

- Column (near) orthogonal: incoherence

$$\max_{i\neq j} |\langle \boldsymbol{a}_{0i}, \boldsymbol{a}_{0j}\rangle| \;\leq\; \mu.$$

Recovery of All Dictionary Items

- **Complete case:** this can be done with a deflation process via orthogonal projections
- **Overcomplete case:** we can do repetitive random independent initializations



recover full A_0 via repeated independent trials

Further Readings

- *High-Dimensional Data Analysis with Low-Dimensional Models: Principles, Computation, and Applications*. John Wright, Yi Ma. (**Chapter 7**)

- *Complete Dictionary Recovery Over the Sphere I: Overview and the Geometric Picture.* Ju Sun, Qing Qu, and John Wright, IEEE Trans. Info. Theory, 2016.

- *Analysis of the Optimization Landscapes for Overcomplete Representation Learning.* Qing Qu, Yuexiang Zhai, Xiao Li, Yuqian Zhang, Zhihui Zhu, ICLR'20, 2020.

Lecture Agenda

- Problem Introduction
- **Applications in SIPML**
  - Example I: Generalized Phase Retrieval
  - Example II: Low-Rank Matrix Recovery
  - Example III: Training Deep Neural Networks
  - Example IV: Sparse Dictionary Learning
  - **Example V: Sparse Blind Deconvolution**

Sparse Blind Deconvolution with Multiple Inputs

Problem: Given multiple $\boldsymbol{y}_i$ of *circulant* convolutions

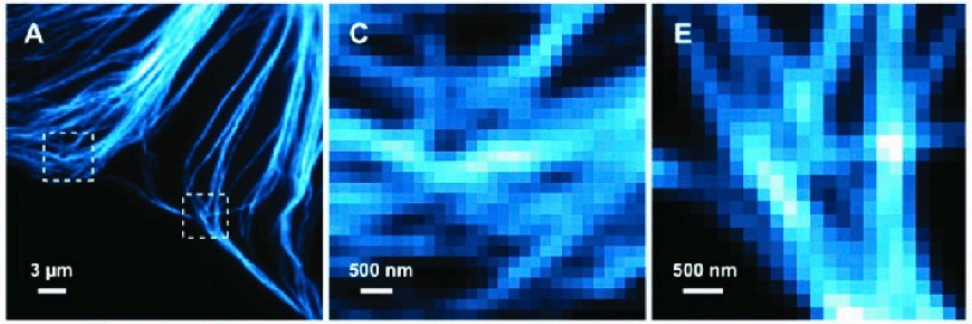
$$\boldsymbol{y}_i \;=\; \boldsymbol{a} \circledast \boldsymbol{x}_i, \quad 1 \le i \le p$$

jointly learn both $\boldsymbol{a}$ and *sparse* $\{\boldsymbol{x}_i\}_{i=1}^{p}$.

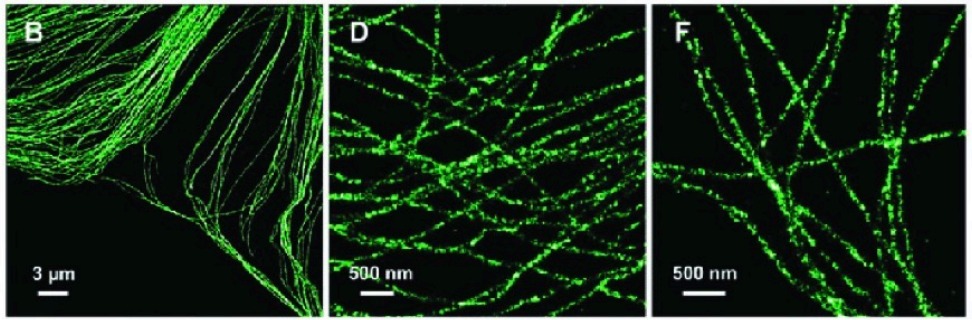
- Here, $\boldsymbol{y}_i, \boldsymbol{a}, \boldsymbol{x}_i \in \mathbb{R}^n$ $(1 \le i \le p)$.
- Sparse signal $\boldsymbol{x}_i$: $\boldsymbol{x}_i \sim_{i.i.d.}$ Bernoulli-Gaussian(θ).

Motivation: Super-resolution Microscopy Imaging

Conventional Fluorescent Optical Microscopy



Stochastic Optical Reconstruction Microscopy (STORM)

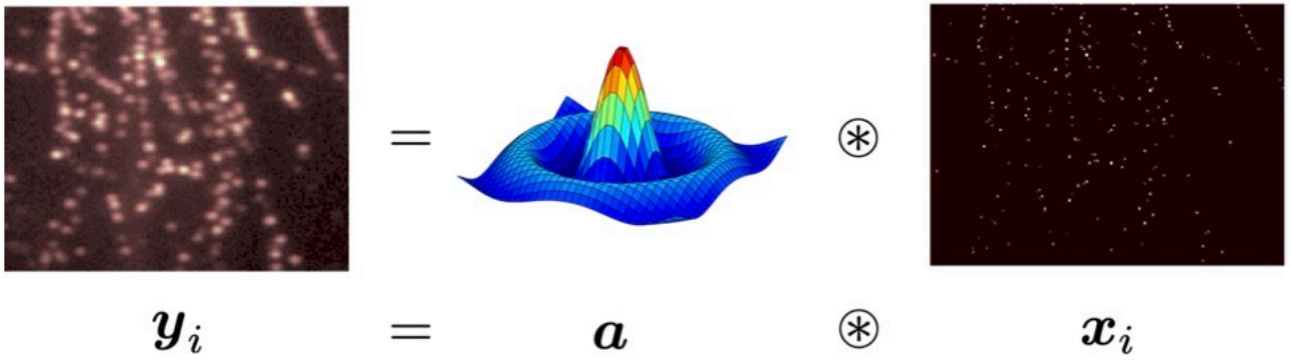


Motivation: Super-resolution Microscopy Imaging

Given multiple $\boldsymbol{y}_i$ of circulant convolution

$$\boldsymbol{y}_i = \boldsymbol{a} \circledast \boldsymbol{x}_i, \quad 1 \leq i \leq p$$

jointly learn both $\boldsymbol{a}$ and sparse $\{\boldsymbol{x}_i\}_{i=1}^{p}$

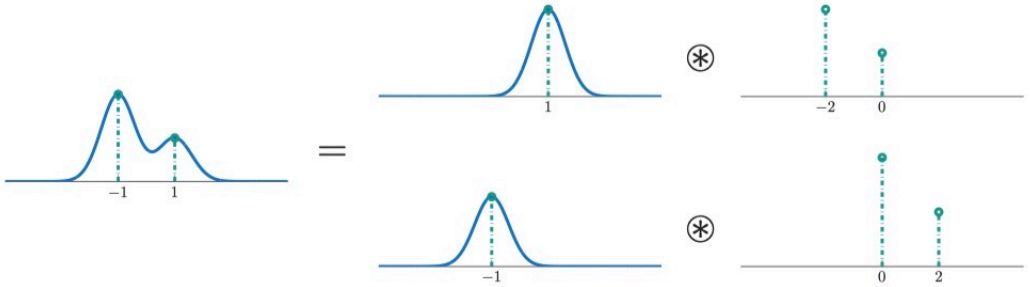


Symmetry of Sparse Blind Deconvolution

- **Scaling symmetry:** $\boldsymbol{y}_i = \boldsymbol{a} \circledast \boldsymbol{x}_i = \alpha \boldsymbol{a} \circledast \alpha^{-1} \boldsymbol{x}_i$

- **Shift symmetry:**

$$\left(\boldsymbol{a}, \{\boldsymbol{x}_i\}_{i=1}^p\right) \quad \Longleftrightarrow \quad \left(\mathrm{s}_\ell\left[\boldsymbol{a}\right], \{\mathrm{s}_{-\ell}\left[\boldsymbol{x}_i\right]\}_{i=1}^p\right)$$



$$\boldsymbol{y}_i \quad = \quad \alpha \; s_\ell[\boldsymbol{a}] \quad \circledast \quad (1/\alpha) \; s_{-\ell}[\boldsymbol{x}_i]$$

Symmetry of Sparse Blind Deconvolution

- **Scaling symmetry:** $\boldsymbol{y}_i = \boldsymbol{a} \circledast \boldsymbol{x}_i = \alpha \boldsymbol{a} \circledast \alpha^{-1} \boldsymbol{x}_i$

- **Shift symmetry:**

$$(\boldsymbol{a}, \{\boldsymbol{x}_i\}_{i=1}^p) \quad \Longleftrightarrow \quad (\mathrm{s}_\ell[\boldsymbol{a}], \{\mathrm{s}_{-\ell}[\boldsymbol{x}_i]\}_{i=1}^p)$$

  - can only be solved to a shift ambiguity;
  - natural formulations are nonconvex.

Reducing the Problem to Dictionary Learning

- For each $\boldsymbol{y}_i = \boldsymbol{a} \circledast \boldsymbol{x}_i$, equivalently

$$\boldsymbol{C}_{\boldsymbol{y}_i} = \boldsymbol{C}_{\boldsymbol{a}} \cdot \boldsymbol{C}_{\boldsymbol{x}_i}, \quad 1 \le i \le p.$$

- Rewrite the problem as

$$\begin{bmatrix} \boldsymbol{C}_{\boldsymbol{y}_1} & \cdots & \boldsymbol{C}_{\boldsymbol{y}_p} \end{bmatrix} = \boldsymbol{C}_{\boldsymbol{a}} \begin{bmatrix} \boldsymbol{C}_{\boldsymbol{x}_1} & \cdots & \boldsymbol{C}_{\boldsymbol{x}_p} \end{bmatrix}$$

data Y dictionary sparse code X_0

Reducing the Problem to Dictionary Learning

Given $\boldsymbol{Y} = \boldsymbol{C_a X}_0$, jointly learn the dictionary $\boldsymbol{C_a}$ and sparse $\boldsymbol{X}_0$?

$$\begin{bmatrix} \boldsymbol{C}_{\boldsymbol{y}_1} & \cdots & \boldsymbol{C}_{\boldsymbol{y}_p} \end{bmatrix} = \boldsymbol{C_a} \begin{bmatrix} \boldsymbol{C}_{\boldsymbol{x}_1} & \cdots & \boldsymbol{C}_{\boldsymbol{x}_p} \end{bmatrix}$$

data Y dictionary sparse code X_0

Reducing the Problem to Dictionary Learning

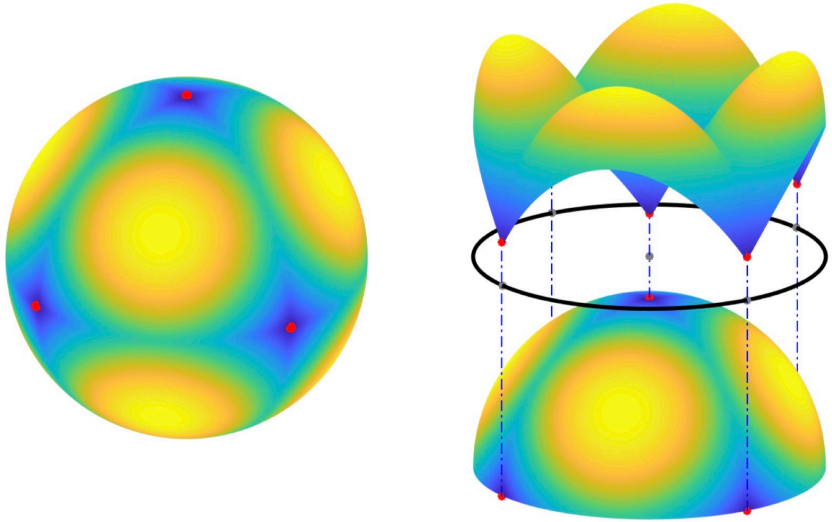
$$\overline{\boldsymbol{Y}} = \left(\boldsymbol{Y}\boldsymbol{Y}^\top\right)^{-1/2}\boldsymbol{Y} \propto \overline{\boldsymbol{A}}_0\boldsymbol{X}_0, \quad \overline{\boldsymbol{A}}_0 \in O(n).$$

We find **one column** of $\boldsymbol{C}_{\boldsymbol{a}}$ via

$$\boxed{\min_{\boldsymbol{q}} \varphi(\boldsymbol{q}^\top\overline{\boldsymbol{Y}}), \quad \text{s.t.} \quad \boldsymbol{q} \in \mathbb{S}^{n-1}.}$$

- $\varphi(\cdot)$ is a sparsity promoting function, $\varphi(\cdot) = \|\cdot\|_1$, $\varphi(\cdot) = -\|\cdot\|_4^4$
- assume that sparse $\boldsymbol{X}_0 \sim$ Bernoulli-Gaussian(θ).

Benign Global Nonconvex Landscape



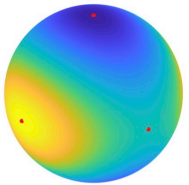
- Symmetric copies of the ground truth are minimizers.
- Negative curvature in symmetry-breaking directions.
- Cascade of saddle points.

Other Applications

Nonconvex Problems with Discrete Symmetries

Eigenvector Computation

Maximize a quadratic form over the sphere.



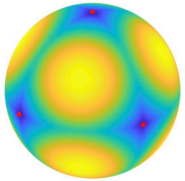
$\max_{\boldsymbol{x}\in\mathbb{S}^{n-1}} \frac{1}{2}\boldsymbol{x}^*\boldsymbol{A}\boldsymbol{x}.$

Symmetry: $\boldsymbol{x}\mapsto -\boldsymbol{x}$

$\mathbb{G}=\{\pm 1\}$

Dictionary Learning

Approximate a given matrix $\boldsymbol{Y}$ *as* $\boldsymbol{Y}=\boldsymbol{A}\boldsymbol{X}$*, with* $\boldsymbol{X}$ *sparse*



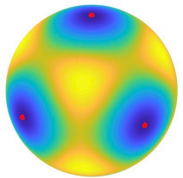
$\min_{\boldsymbol{A}\in\mathcal{A},\boldsymbol{X}} \frac{1}{2}\|\boldsymbol{Y}-\boldsymbol{A}\boldsymbol{X}\|_F^2+\lambda\|\boldsymbol{X}\|_1.$

Symmetry: $(\boldsymbol{A},\boldsymbol{X})\mapsto(\boldsymbol{A}\boldsymbol{\Gamma},\boldsymbol{X}\boldsymbol{\Gamma}^*)$

$\mathbb{G}=\mathsf{SP}(n)$

Tensor Decomposition

Determine components $\boldsymbol{a}_i$ *of an orthogonal decomposable tensor* $\boldsymbol{T}=\sum_i \boldsymbol{a}_i\otimes\boldsymbol{a}_i\otimes\boldsymbol{a}_i\otimes\boldsymbol{a}_i$



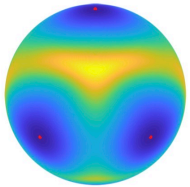
$\max_{\boldsymbol{X}\in\mathsf{O}(n)} \sum_i \boldsymbol{T}(\boldsymbol{x}_i,\boldsymbol{x}_i,\boldsymbol{x}_i,\boldsymbol{x}_i).$

Symmetry: $\boldsymbol{X}\mapsto\boldsymbol{X}\boldsymbol{\Gamma}$

$\mathbb{G}=\mathsf{P}(n)$

Short-and-Sparse Deconvolution

Recover a short $\boldsymbol{a}$ *and a sparse* $\boldsymbol{x}$ *from their convolution* $\boldsymbol{y}=\boldsymbol{a}\circledast\boldsymbol{x}$*.*



$\min_{\boldsymbol{a},\boldsymbol{x}} \frac{1}{2}\|\boldsymbol{y}-\boldsymbol{a}\circledast\boldsymbol{x}\|_2^2 \;+\; \lambda\|\boldsymbol{x}\|_1.$

Symmetry: $(\boldsymbol{a},\boldsymbol{x})\mapsto(\alpha s_\tau[\boldsymbol{a}],\alpha^{-1}s_{-\tau}[\boldsymbol{x}])$

$\mathbb{G}=\mathbb{Z}_n\times\mathbb{R}_*$ or $\mathbb{G}=\mathbb{Z}_n\times\{\pm 1\}$

Further Readings

- *High-Dimensional Data Analysis with Low-Dimensional Models: Principles, Computation, and Applications*. John Wright, Yi Ma. (**Chapter 7**)

- *A Nonconvex Approach for Exact and Efficient Multichannel Sparse Blind Deconvolution*. Qing Qu, Xiao Li, Zhihui Zhu, SIAM Imaging Sciences, 2020.

- *Short-and-Sparse Deconvolution -- A Geometric Approach*. Yenson Lau, Qing Qu, Han-Wen Kuo, Pengcheng Zhou, Yuqian Zhang, John Wright, ICLR'20, 2020.

Further Readings

- *Geometry and Symmetry in Short-and-Sparse Deconvolution.* Han-Wen Kuo, Yuqian Zhang, Yenson Lau, John Wright. SIAM Journal on Mathematics of Data Science, 2020.

- *Structured Local Optima in Sparse Blind Deconvolution.* Yuqian Zhang, Han-Wen Kuo, John Wright. IEEE Transaction on Information Theory, 2020.

- *On the Global Geometry of Sphere-Constrained Sparse Blind Deconvolution.* Yuqian Zhang, Yenson Lau, Han-Wen Kuo, Sky Cheung, Abhay Pasupathy, John Wright. EEE Transactions on Pattern Analysis and Machine Intelligence, 2019