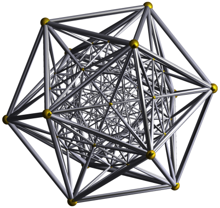
ICASSP 2022 Short Course One on
Low-Dimensional Models for High-Dimensional Data

Lecture 3: Learning Low-dimensional Models via Nonconvex Optimization

Sam Buchanan, Yi Ma, Qing Qu
John Wright, Yuqian Zhang, Zhihui Zhu

May 25, 2022

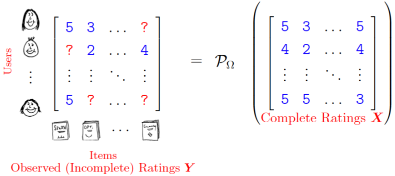


Outline

Example: Low-rank Matrix Completion

We observe:

$$\underset{\text{Observed ratings}}{\boldsymbol{Y}} = \mathcal{P}_\Omega \left[\underset{\text{Complete ratings}}{\boldsymbol{X}} \right].$$



Matrix completion

via bilinear low-rank factorization

$$\min_{\boldsymbol{U},\boldsymbol{V}} f(\boldsymbol{U},\boldsymbol{V}) = \sum_{(i,j)\in\Omega} [(\boldsymbol{U}\boldsymbol{V}^*)_{i,j} - \boldsymbol{Y}_{i,j}]^2 + \underbrace{\frac{\lambda}{2}\|\boldsymbol{U}\|_F^2 + \frac{\lambda}{2}\|\boldsymbol{V}\|_F^2}_{\mathrm{reg}(\boldsymbol{U},\boldsymbol{V})}.$$

$$\|\boldsymbol{M}\|_* = \min_{\boldsymbol{M}=\boldsymbol{U}\boldsymbol{V}^*} \frac{\lambda}{2}\|\boldsymbol{U}\|_F^2 + \frac{\lambda}{2}\|\boldsymbol{V}\|_F^2$$

Example: Dictionary for Image Representation

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

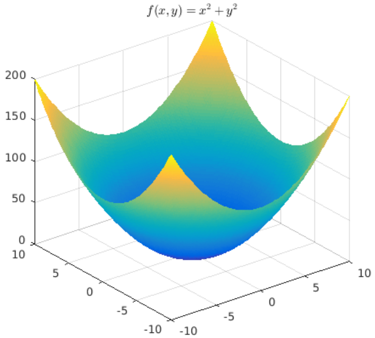
$$I_{\text{noisy}} = \underbrace{\boldsymbol{A}}_{\text{dictionary}} \times \underbrace{\boldsymbol{x}}_{\text{sparse}} + \underbrace{\boldsymbol{z}}_{\text{noise}}. \quad (1)$$

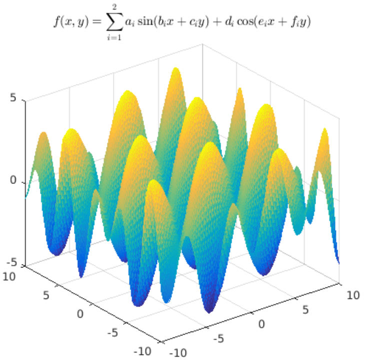
Dictionary learning: the motifs or atoms of the dictionary are unknown:

$$\underbrace{\boldsymbol{Y}}_{\textbf{data}} = \underbrace{\boldsymbol{A}}_{\textbf{dictionary}} \underbrace{\boldsymbol{X}}_{\textbf{sparse}}. \quad (2)$$

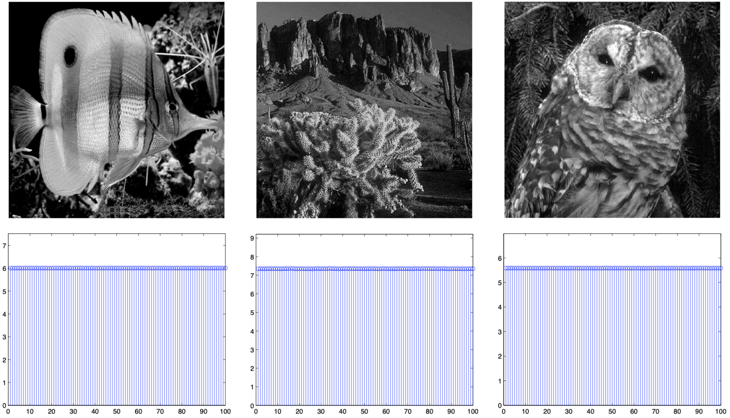
- Band-limited signals: $\boldsymbol{A} = \boldsymbol{F}$, the Fourier transform;
- Piecewise smooth signals: $\boldsymbol{A} = \boldsymbol{W}$, the wavelet transforms;
- Natural images $\boldsymbol{A} = ?$ (How to **learn** $\boldsymbol{A}$ from the data $\boldsymbol{Y}$?)

Convex and Nonconvex Optimization



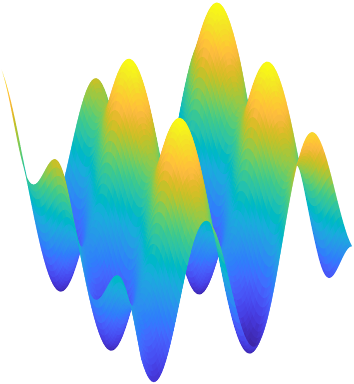


Dictionary Learning

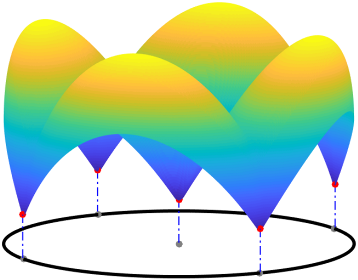


Recovered solutions always obtain the same objective value.

Benign Nonconvex Optimization Landscape



General Case



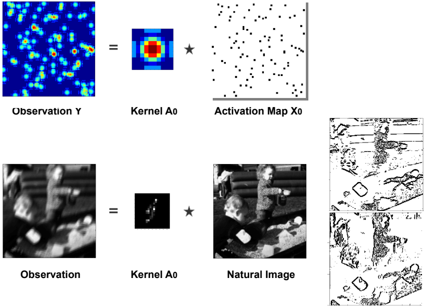
Structured Case

Example: Sparse Blind Deconvolution

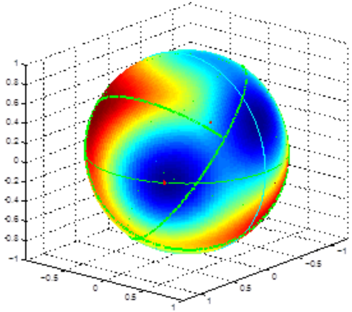
Sparse Blind Deconvolution: the convolutional motif or sparse activation signal are unknown:

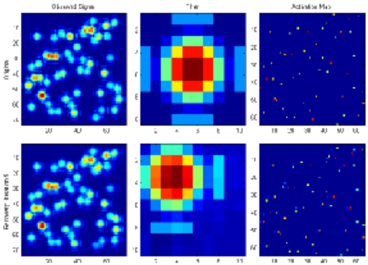
$$\underset{\text{data}}{\boldsymbol{Y}} = \underset{\text{motif}}{\boldsymbol{A}} * \underset{\text{sparse}}{\boldsymbol{X}}. \qquad (3)$$

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



Sparse Blind Deconvolution

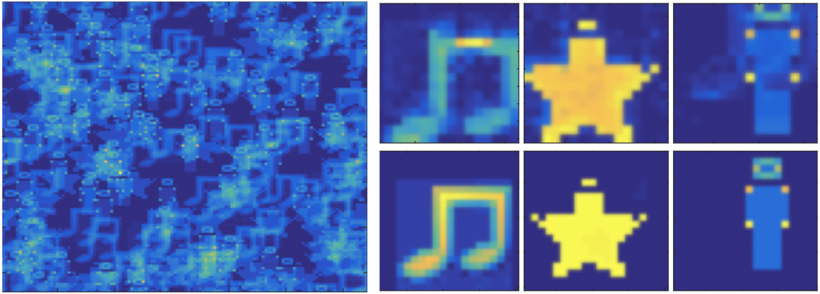




Recovered solutions are near signed shift-truncations of the ground truth.

Convoltional Dictionary learning

$$\underbrace{\boldsymbol{Y}}_{\text{data}} = \sum_i \underbrace{\boldsymbol{A}_i}_{\text{motif}} * \underbrace{\boldsymbol{X}_i}_{\text{sparse}}.$$



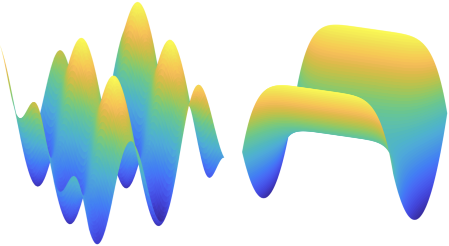
Recovered solutions are near signed shift-truncations of the ground truth.

Challenges of Nonconvex Optimization – Pessimistic Views

Consider the problem of minimizing a general nonlinear function:

$$\min_{\boldsymbol{z}} \varphi(\boldsymbol{z}), \quad \boldsymbol{z} \in \mathsf{C}. \tag{4}$$

In the worst case, even finding a *local* minimizer can be NP-hard[1].



Spurious local minimizers **Flat saddle points**

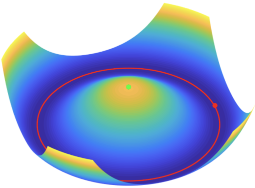
Hence typically people seek to work with relatively benign functions with benign guarantees:

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

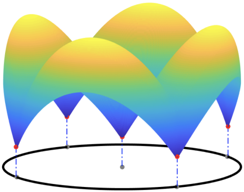
[1] Some NP-complete problems in quadratic and nonlinear programming, K.G Murty and S. N. Kabadi, 1987

Opportunities – Optimistic Views

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



Rotational symmetry



Discrete symmetry

The function φ is invariant under certain group action:

- for low rank matrix recovery, invariant under a continuous rotation:

$$\varphi((\boldsymbol{U}\boldsymbol{\Gamma}, \boldsymbol{V}\boldsymbol{\Gamma}^{-1})) = \varphi((\boldsymbol{U}, \boldsymbol{V})), \quad \forall \text{ invertible } \boldsymbol{\Gamma}.$$

- for dictionary learning, invariant under signed permutations:

$$\varphi((\boldsymbol{A}, \boldsymbol{X})) = \varphi((\boldsymbol{A}\boldsymbol{\Pi}, \boldsymbol{\Pi}^{*}\boldsymbol{X})), \quad \forall \boldsymbol{\Pi} \in \mathsf{SP}(n).$$

Nonlinearity and Symmetry

Intrinsic ambiguity against the uniqueness of the solution

- **low rank matrix recovery**

$$\boldsymbol{X} = \boldsymbol{U}_0 \boldsymbol{V}_0^T = \boldsymbol{U}_0 \boldsymbol{\Gamma} \boldsymbol{\Gamma}^{-1} \boldsymbol{V}_0^T$$

  for any invertible $\boldsymbol{\Gamma}$.

- **dictionary learning**

$$\boldsymbol{Y} = \boldsymbol{A}_0 \boldsymbol{X}_0 = \boldsymbol{A}_0 \boldsymbol{\Pi} \boldsymbol{\Pi}^* \boldsymbol{X}_0$$

  for any signed permutation $\boldsymbol{\Pi}$.

- **blind deconvolution**

$$\boldsymbol{y} = \boldsymbol{a}_0 * \boldsymbol{x}_0 = S_\tau[\boldsymbol{a}_0] * S_{-\tau}[\boldsymbol{x}_0]$$

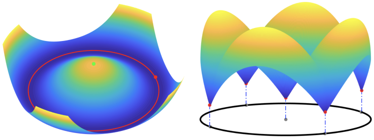
  for any signed shift τ.

Optimization under Symmetry

Definition (Symmetric Function)

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

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

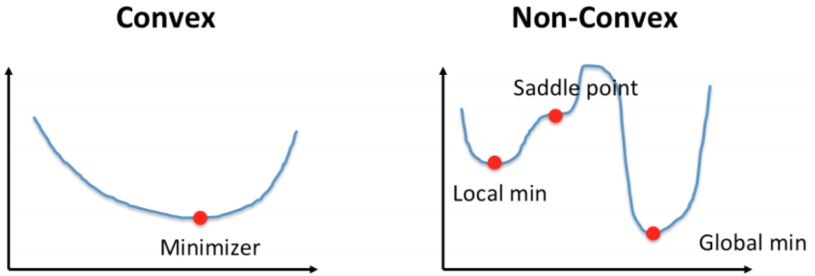


Rotational symmetry   Discrete symmetry

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.

Basic Calculus

Critical points or stationary points: gradient vanishes

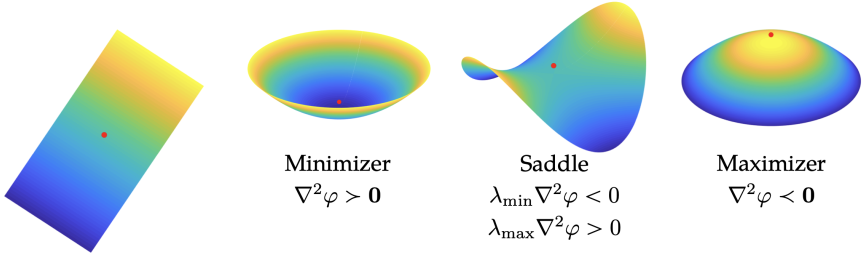


- convex function: critical point = minimizer
- nonconvex function: not all critical point are minimizer

Basic Calculus

Critical points with non-singular hessian

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



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

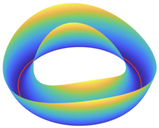
Outline

Problems with Rotational Symmetry

Nonconvex Problems with Rotational Symmetries

Eigenspace Computation

Compute the principal subspace of a symmetric matrix.



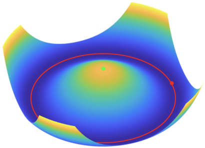
$$\min_{\boldsymbol{X}^*\boldsymbol{X}=\boldsymbol{I}} -\tfrac{1}{2}\mathrm{trace}\left[\boldsymbol{X}^*\boldsymbol{A}\boldsymbol{X}\right].$$

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

$\mathbb{G} = O(r)$

Generalized Phase Retrieval

Recover a complex vector $\boldsymbol{x}_0$ from magnitude measurements $\boldsymbol{y} = |\boldsymbol{A}\boldsymbol{x}_0|$.



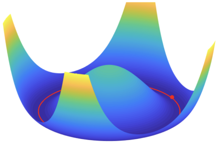
$$\min_{\boldsymbol{x}} \tfrac{1}{2}\|\boldsymbol{y}^2 - |\boldsymbol{A}\boldsymbol{x}|^2\|_2^2.$$

Symmetry: $\boldsymbol{x} \mapsto \boldsymbol{x}e^{\mathrm{i}\phi}$

$\mathbb{G} = \mathbb{S}^1 \cong O(2)$

Matrix Recovery

Recover a low-rank matrix $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{V}^*$ from incomplete / corrupted observations



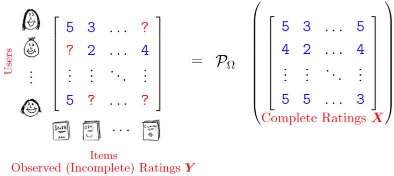
$$\min_{\boldsymbol{U},\boldsymbol{V}} \mathcal{L}(\boldsymbol{Y} - \mathcal{A}[\boldsymbol{U}\boldsymbol{V}^*]) \ + \ \rho(\boldsymbol{U},\boldsymbol{V}).$$

Symmetry: $(\boldsymbol{U},\boldsymbol{V}) \mapsto (\boldsymbol{U}\boldsymbol{\Gamma}, \boldsymbol{V}\boldsymbol{\Gamma}^{-*})$

$\mathbb{G} = \mathrm{GL}(r)$ or $\mathbb{G} = O(r)$

Low rank matrix recovery

Goal: Given $\boldsymbol{Y} = \mathcal{A}(\boldsymbol{X})$, recover low rank matrix $\boldsymbol{X} = \boldsymbol{U}_0 \boldsymbol{V}_0$



- Convex Formulation

$$\min_{\boldsymbol{X} \in \mathbb{R}^{m \times n}} \quad \|\boldsymbol{X}\|_\star \quad \text{s.t.} \quad \boldsymbol{Y} = \mathcal{A}(\boldsymbol{X})$$

- Nonconvex Formulation

$$\min_{\boldsymbol{U} \in \mathbb{R}^{m \times r}, \boldsymbol{V} \in \mathbb{R}^{n \times r}} \quad \left\|\boldsymbol{Y} - \mathcal{A}(\boldsymbol{U}\boldsymbol{V}^T)\right\|_F^2 + \text{reg}(\boldsymbol{U}, \boldsymbol{V})$$

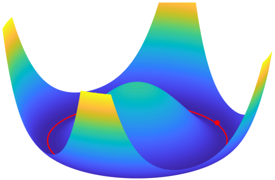
Low Rank Matrix Recovery

$$\min_{\boldsymbol{U},\boldsymbol{V}} \quad \frac{1}{2}\left\|\boldsymbol{Y} - \mathcal{A}(\boldsymbol{U}\boldsymbol{V}^T)\right\|_F^2 + \mathrm{reg}(\boldsymbol{U},\boldsymbol{V})$$

Inherent Symmetry:

$$\boldsymbol{X} = \boldsymbol{U}_0\boldsymbol{V}_0^T = \boldsymbol{U}_0\boldsymbol{\Gamma}\boldsymbol{\Gamma}^{-1}\boldsymbol{V}_0^T$$

for any invertible $\boldsymbol{\Gamma} \in \mathbb{R}^{r\times r}$.



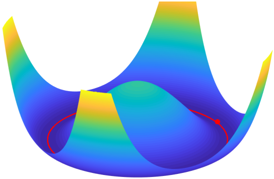
Low Rank Matrix Recovery

$$\min_{\boldsymbol{U},\boldsymbol{V}} \quad \frac{1}{2}\left\|\boldsymbol{Y} - \mathcal{A}(\boldsymbol{U}\boldsymbol{V}^T)\right\|_F^2 + \mathrm{reg}(\boldsymbol{U},\boldsymbol{V})$$

Inherent Symmetry:

$$\boldsymbol{X} = \boldsymbol{U}_0\boldsymbol{V}_0^T = \boldsymbol{U}_0\boldsymbol{\Gamma}\boldsymbol{\Gamma}^{-1}\boldsymbol{V}_0^T$$

for any invertible $\boldsymbol{\Gamma} \in \mathbb{R}^{r\times r}$.



- Are $\left(\boldsymbol{U}_0\boldsymbol{\Gamma}, \boldsymbol{V}_0\boldsymbol{\Gamma}^{-1}\right)$ the only local solutions?
- Does there exist flat stationary points?

Rank-1 Symmetric Matrix

Simplifications:

- $\boldsymbol{Y} = \mathcal{A}(\boldsymbol{X}) = \boldsymbol{X}$
- $\boldsymbol{X} = \boldsymbol{U}_0\boldsymbol{U}_0^T$ is symmetric and rank-1

$$\boldsymbol{X} = \boldsymbol{u}_0\boldsymbol{u}_0^T = (-\boldsymbol{u}_0)(-\boldsymbol{u}_0^T)$$

the rotational symmetry is reduced to sign symmetry.

Nonconvex formulation:

$$\min_{\boldsymbol{u}} \quad \phi(\boldsymbol{u}) \doteq \frac{1}{4}\left\|\boldsymbol{X} - \boldsymbol{u}\boldsymbol{u}^T\right\|_F^2 + \underbrace{\lambda\left\|\boldsymbol{u}\right\|_2^2}_{const}$$

Rank-1 Symmetric Matrix

$$\min_{\boldsymbol{u}} \quad \phi(\boldsymbol{u}) \doteq \frac{1}{4} \left\| \boldsymbol{X} - \boldsymbol{u}\boldsymbol{u}^T \right\|_F^2$$

Critical points have zero gradient

$$\begin{aligned} \nabla\phi &= (\boldsymbol{u}\boldsymbol{u}^T - \boldsymbol{X})\boldsymbol{u} \\ &= \|\boldsymbol{u}\|_2^2 \, \boldsymbol{u} - \boldsymbol{X}\boldsymbol{u} \\ &= \boldsymbol{0} \end{aligned}$$

therefore critical points must be one of the following

- $\boldsymbol{u} = \pm \boldsymbol{u}_0$
- $\boldsymbol{u} = \boldsymbol{0}$

Rank-1 Symmetric Matrix

$$\min_{\boldsymbol{u}} \quad \phi(\boldsymbol{u}) \doteq \frac{1}{4}\left\|\boldsymbol{X} - \boldsymbol{u}\boldsymbol{u}^T\right\|_F^2$$

with the second order derivative

$$\nabla^2 \phi = 2\boldsymbol{u}\boldsymbol{u}^T + \|\boldsymbol{u}\|_2^2 \boldsymbol{I} - \boldsymbol{X}.$$

Rank-1 Symmetric Matrix

$$\min_{\boldsymbol{u}} \quad \phi(\boldsymbol{u}) \doteq \frac{1}{4} \left\| \boldsymbol{X} - \boldsymbol{u}\boldsymbol{u}^T \right\|_F^2$$

with the second order derivative

$$\nabla^2 \phi = 2\boldsymbol{u}\boldsymbol{u}^T + \|\boldsymbol{u}\|_2^2 \boldsymbol{I} - \boldsymbol{X}.$$

Then the critical points can be grouped as

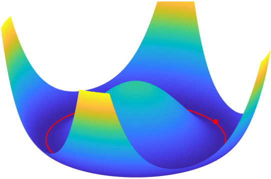
- Local minimizer $\boldsymbol{u} = \pm \boldsymbol{u}_0$ and $\boldsymbol{u}\boldsymbol{u}^T = \boldsymbol{X}$

$$\nabla^2 \phi = \boldsymbol{u}\boldsymbol{u}^T + \|\boldsymbol{u}\|_2^2 \boldsymbol{I}.$$

- Maximizer $\boldsymbol{u} = \mathbf{0}$

$$\nabla^2 \phi = -\boldsymbol{X}.$$

Low Rank Matrix Recovery



Symmetric low rank matrix

$$\min_{\boldsymbol{U}} \quad \phi(\boldsymbol{u}) \doteq \frac{1}{4} \left\| \boldsymbol{X} - \boldsymbol{U}\boldsymbol{U}^T \right\|_F^2 .$$

General low rank matrix recover

$$\min_{\boldsymbol{U},\boldsymbol{V}} \quad \phi(\boldsymbol{u}) \doteq \frac{1}{2} \left\| \boldsymbol{X} - \boldsymbol{U}\boldsymbol{V}^T \right\|_F^2 + \lambda \left\| \boldsymbol{U} \right\|_F^2 + \lambda \left\| \boldsymbol{V} \right\|_F^2 .$$

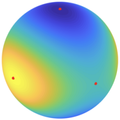
Local minimizers: *are ground truth* $\boldsymbol{U}_0$ *and* $\boldsymbol{V}_0$ *up to rotation;*
Negative curvature: *between multiple local minimizers.*

Problems with Discrete Symmetry

Nonconvex Problems with Discrete Symmetries

Eigenvector Computation

Maximize a quadratic form over the sphere.



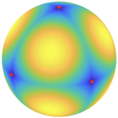
$\max_{\boldsymbol{x} \in \mathbb{S}^{n-1}} \frac{1}{2} \boldsymbol{x}^* \boldsymbol{A} \boldsymbol{x}.$

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

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

Dictionary Learning

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



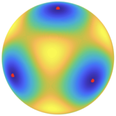
$\min_{\boldsymbol{A} \in \mathcal{A}, \boldsymbol{X}} \frac{1}{2} \|\boldsymbol{Y} - \boldsymbol{A}\boldsymbol{X}\|_F^2 + \lambda \|\boldsymbol{X}\|_1.$

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

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

Tensor Decomposition

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



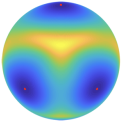
$\max_{\boldsymbol{X} \in O(n)} \sum_i \boldsymbol{T}(\boldsymbol{x}_i, \boldsymbol{x}_i, \boldsymbol{x}_i, \boldsymbol{x}_i).$

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

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

Short-and-Sparse Deconvolution

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



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

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

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

Dictionary Learning

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

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

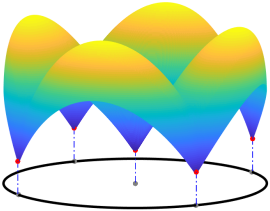
In presence of noise, the optimization problem can be rewritten as

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

Inherent Symmetry:

$$\boldsymbol{Y} = \boldsymbol{A}_0 \boldsymbol{\Gamma}\boldsymbol{\Gamma}^* \boldsymbol{X}_0,$$

for any signed permutation matrix $\boldsymbol{\Gamma}$.



Orthogonal Dictionary Learning

- Input: matrix $\boldsymbol{Y}$ which is the product of an orthogonal matrix $\boldsymbol{A}_0$ (called a dictionary) and a sparse matrix $\boldsymbol{X}_0$:

$$\boldsymbol{Y} = \boldsymbol{A}_0 \boldsymbol{X}_0, \quad \boldsymbol{A}_0 \boldsymbol{A}_0^* = \boldsymbol{I}, \boldsymbol{X}_0 \text{ sparse.}$$

- Optimization Formulation

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

- Given the optimization constraint, $\boldsymbol{X}$ is uniquely defined in terms of $\boldsymbol{A}$

$$\boldsymbol{X} = \boldsymbol{A}^* \boldsymbol{A} \boldsymbol{X} = \boldsymbol{A}^* \boldsymbol{Y}.$$

- Equivalent formulation

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

Orthogonal Dictionary Learning

Instead of aiming to solve the entire matrix $\boldsymbol{A} = [\boldsymbol{a}_1, \dots, \boldsymbol{a}_n]$ at once via

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

A simpler model problem solves for the columns $\boldsymbol{a}_i$ one at a time

$$\min_{\|\boldsymbol{a}\|_2 = 1} \quad \|\boldsymbol{a}^* \boldsymbol{Y}\|_1 .$$

More simplifications:

- orthogonal dictionary $\boldsymbol{A}_0 = \boldsymbol{I}$;
- sparse coefficients $\boldsymbol{X}_0 = \boldsymbol{I}$

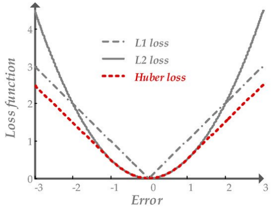
$$\min_{\|\boldsymbol{a}\|_2 = 1} \quad \|\boldsymbol{a}\|_1 .$$

Orthogonal Dictionary Learning

$$\min_{\|\boldsymbol{a}\|_2=1} \quad \|\boldsymbol{a}\|_1 .$$

To obtain the second order information for stationary points, we use a smoothed ℓ_1 penalty — Huber loss

$$h_\lambda(x) \;=\; \begin{cases} \lambda|x| - \lambda^2/2 & |x| > \lambda, \\ x^2/2 & |x| \leq \lambda. \end{cases}$$

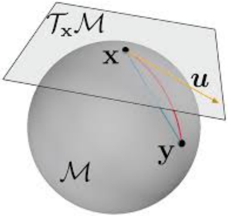


$$\min_{\|\boldsymbol{a}\|_2=1} \quad \phi(\boldsymbol{a}) \doteq h_\lambda(\boldsymbol{a}).$$

Orthogonal Dictionary Learning — Calculus

$$\min_{\|\boldsymbol{a}\|_2=1} \quad \phi(\boldsymbol{a}) = h_\lambda(\boldsymbol{a}),$$

$$h_\lambda(a_i) \; = \; \begin{cases} \lambda|a_i| - \lambda^2/2 & |a_i| > \lambda, \\ a_i^2/2 & |a_i| \le \lambda. \end{cases}$$



The Euclidean gradient

$$\nabla\phi = \lambda \mathrm{sign}(\boldsymbol{a}) \circ \mathbf{1}_{|\boldsymbol{a}|>\lambda} + \boldsymbol{a} \circ \mathbf{1}_{|\boldsymbol{a}|\le\lambda}.$$

With the sphere constraint, a critical point satisfies $\nabla\phi = \mathbf{0}$ or $\nabla\phi \propto \boldsymbol{a}$.

$$\boldsymbol{a} \propto \mathrm{sign}(\boldsymbol{a}).$$

Orthogonal Dictionary Learning — Calculus

Recall that

$$\nabla \phi = \lambda \mathrm{sign}(\boldsymbol{a}) \circ \mathbf{1}_{|\boldsymbol{a}|>\lambda} + \boldsymbol{a} \circ \mathbf{1}_{|\boldsymbol{a}|\leq\lambda}$$

has first-order critical points $\boldsymbol{a} \propto sign(\boldsymbol{a})$. Denote $I = supp(\boldsymbol{a})$, then the Riemannian Hessian over the sphere follows

$$\begin{aligned}\mathrm{Hess}[\phi] &= \boldsymbol{P}_{\boldsymbol{a}^\perp}\left[\underbrace{\nabla^2\phi}_{\text{curvature of }\phi} - \underbrace{\langle \nabla\phi, \boldsymbol{a}\rangle \boldsymbol{I}}_{\text{curvature of the sphere}}\right]\boldsymbol{P}_{\boldsymbol{a}^\perp} \\ &= \boldsymbol{P}_{\boldsymbol{a}^\perp}[\boldsymbol{D}_{\mathbf{1}_{|\boldsymbol{a}|\leq\lambda}} - \lambda\,|I|\,\boldsymbol{I}]\boldsymbol{P}_{\boldsymbol{a}^\perp}\end{aligned}$$

with $\boldsymbol{P}_{\boldsymbol{a}^\perp} = \boldsymbol{I} - \boldsymbol{a}\boldsymbol{a}^T$. The Hessian exhibits $|I| - 1$ negative eigenvalues and $n - |I|$ positive eigenvalues.

Orthogonal Dictionary Learning — Calculus

- $\boldsymbol{a} = \pm \boldsymbol{e}_i$, then the Hessian is positive definite

$$\mathrm{Hess}[\phi] = \boldsymbol{P}_{\boldsymbol{a}^\perp}[(1-\lambda)\boldsymbol{I} - \lambda \boldsymbol{D}_{\boldsymbol{e}_i}]\boldsymbol{P}_{\boldsymbol{a}^\perp} = \boldsymbol{P}_{\boldsymbol{a}^\perp}[(1-\lambda)\boldsymbol{I}]\boldsymbol{P}_{\boldsymbol{a}^\perp}$$

  with $\boldsymbol{P}_{\boldsymbol{a}^\perp} = \boldsymbol{I} - \boldsymbol{e}_i\boldsymbol{e}_i^T = \boldsymbol{I} - \boldsymbol{D}_{\boldsymbol{e}_i}$;

- $\boldsymbol{a} = \sum_{i \in I} \pm \frac{1}{\sqrt{|I|}} \boldsymbol{e}_i$, there exist negative curvatures alone $\boldsymbol{e}_i (i \in I)$

$$\mathrm{Hess}[\phi] = \boldsymbol{P}_{\boldsymbol{a}^\perp}\left[(1-\lambda\,|I|)\boldsymbol{D}_{\mathbf{1}_{|\boldsymbol{a}|\le\lambda}} - \lambda\,|I|\,\boldsymbol{D}_{\mathbf{1}_{|\boldsymbol{a}|>\lambda}}\right]\boldsymbol{P}_{\boldsymbol{a}^\perp}.$$

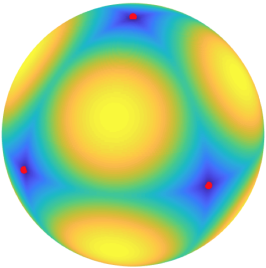
- $\boldsymbol{a} = \sum_{i \in [n]} \pm \frac{1}{\sqrt{n}} \boldsymbol{e}_i$, then $|I| = n$ and the Hessian is negative definite.

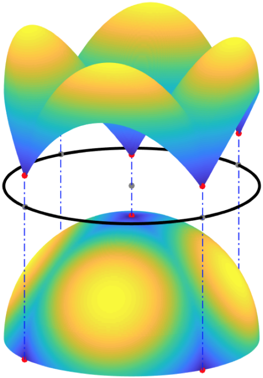
$$\mathrm{Hess}[\phi] = \boldsymbol{P}_{\boldsymbol{a}^\perp}[-\lambda n \boldsymbol{I}]\boldsymbol{P}_{\boldsymbol{a}^\perp}$$

  with $\boldsymbol{P}_{\boldsymbol{a}^\perp} = \boldsymbol{I} - \boldsymbol{a}\boldsymbol{a}^T = (1 - 1/n)\boldsymbol{I}$.

Orthogonal Dictionary Learning — Geometry

Local minimizers are ground truth e_i or $-e_i$.
Negative curvature between multiple local minimizers.



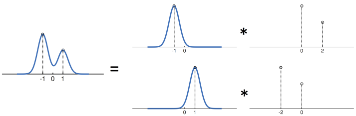


Short-and-Sparse Blind Deconvolution

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

Inherent Symmetry:

$$\boldsymbol{y} = \boldsymbol{a}_0 * \boldsymbol{x}_0 = \alpha s_l[\boldsymbol{a}_0] * \frac{1}{\alpha} s_{-l}[\boldsymbol{x}_0]$$

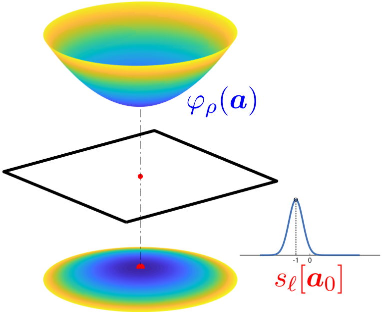


for any shift l and nonzero scaling.

The practical optimization problem can be written as

$$\min_{\|a\|_F^2 = 1, \boldsymbol{x}} \quad \frac{1}{2} \|\boldsymbol{y} - \boldsymbol{a} * \boldsymbol{x}\|_F^2 + \lambda \|\boldsymbol{x}\|_1 .$$

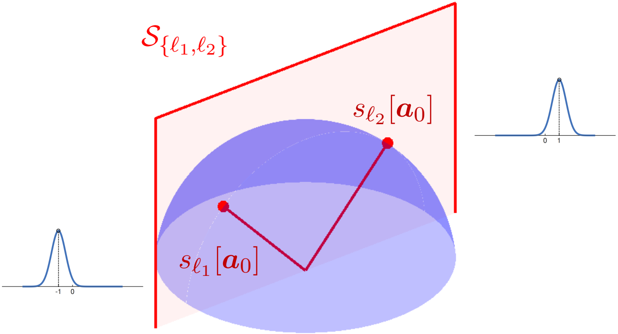
Objective Function – Near One Shift



$$\mathbb{S}^{p-1} \cap \{\boldsymbol{a} \in \mathbb{S}^{p-1} \mid \|\boldsymbol{a} - s_\ell[\boldsymbol{a}_0]\|_2 \leq r\}$$

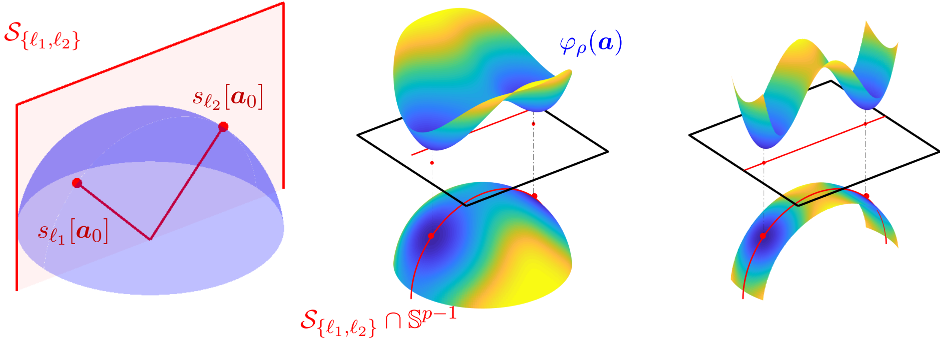
Objective function is **strongly convex** near a shift $s_\ell[\boldsymbol{a}_0]$ of the ground truth.

Objective Function – Linear Span of Two Shifts



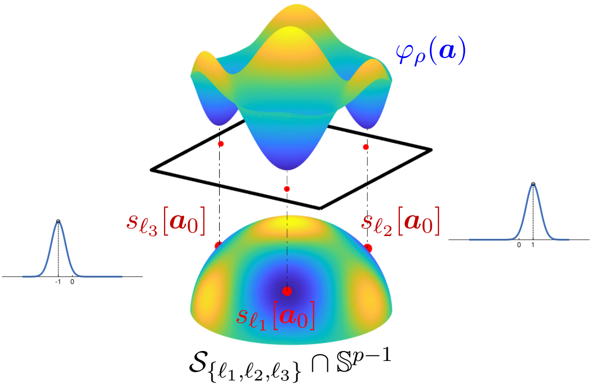
Subspace $\mathcal{S}_{\{\ell_1,\ell_2\}} = \{\alpha_{\ell_1} s_{\ell_1}[\boldsymbol{a}_0] + \alpha_{\ell_2} s_{\ell_2}[\boldsymbol{a}_0] \mid \alpha_{\ell_1}, \alpha_{\ell_2} \in \mathbb{R}\}$.

Objective Function – Linear Span of Two Shifts



Local minimizers are near signed shifts $\pm s_\ell[\boldsymbol{a}_0]$.
Negative curvature between two shifts $s_{\ell_1}[\boldsymbol{a}_0]$, $s_{\ell_2}[\boldsymbol{a}_0]$.

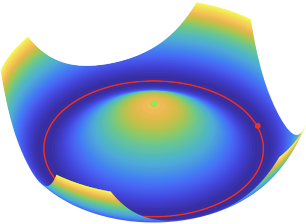
Objective Function – Multiple Shifts



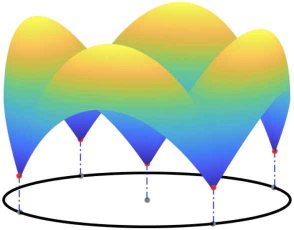
Objective φ_ρ over the linear span $\mathcal{S}_{\ell_1,\ell_2,\ell_3} = \{\sum_{i=1}^{3} \alpha_{\ell_i} s_{\ell_i}[\boldsymbol{a}_0]\}$
Local minimizers are near signed shifts $\pm s_{\ell_i}[\boldsymbol{a}_0]$.

Symmetry and Nonconvexity

- the (only!) local minimizers are symmetric versions of the ground truth.
- there is negative curvature in directions that break symmetry.



Rotational symmetry



Discrete symmetry

Outline

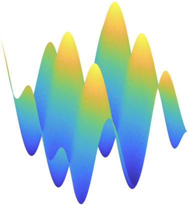
Nonconvex Optimization

Consider the problem of minimizing a general nonlinear function:

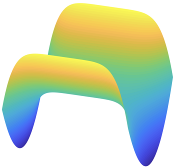
$$\min_{\boldsymbol{x}} f(\boldsymbol{x}), \quad \boldsymbol{x} \in \mathsf{C}. \tag{5}$$

In the worst case, even finding a *local* minimizer can be NP-hard[2].

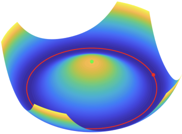
Nonconvex problems that arise from natural physical, geometrical, or statistical origins typically have nice structures, in terms of symmetries!



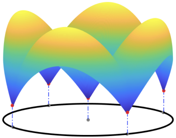
Spurious local minimizers



Flat saddle points



Rotational symmetry



Discrete symmetry

[2] Some NP-complete problems in quadratic and nonlinear programming, K.G Murty and S. N. Kabadi, 1987

Objectives

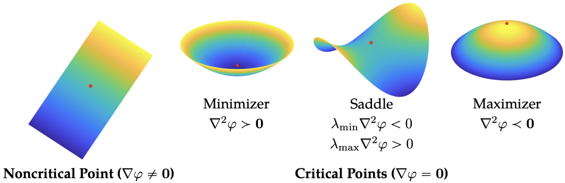
Hence typically people seek to work with relatively benign (gradient/Hessian Lipschitz continuous) functions:

$$\forall \boldsymbol{x}, \boldsymbol{y} \quad \|\nabla f(\boldsymbol{y}) - \nabla f(\boldsymbol{x})\|_2 \le L_1 \|\boldsymbol{y} - \boldsymbol{x}\|_2 \tag{6}$$

with benign objectives:

1. convergence to some critical point $\boldsymbol{x}_\star$ such that: $\nabla f(\boldsymbol{x}_\star) = \mathbf{0}$;
2. the critical point $\boldsymbol{x}_\star$ is second-order stationary: $\nabla^2 f(\boldsymbol{x}_\star) \succeq \mathbf{0}$.

Example: in general f could have irregular second-order stationary points:

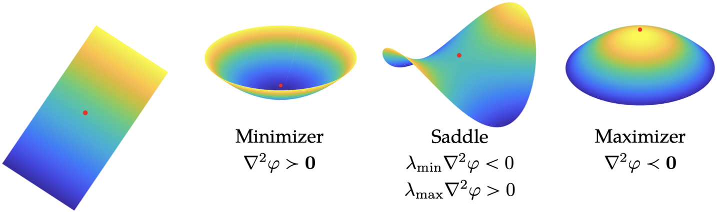


Objectives

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

1. convergence to some critical point $\boldsymbol{x}_\star$ such that: $\nabla f(\boldsymbol{x}_\star) = \mathbf{0}$;
2. the critical point $\boldsymbol{x}_\star$ is second-order stationary: $\nabla^2 f(\boldsymbol{x}_\star) \succeq \mathbf{0}$.

Example: a function φ with symmetry only has **regular** critical points:



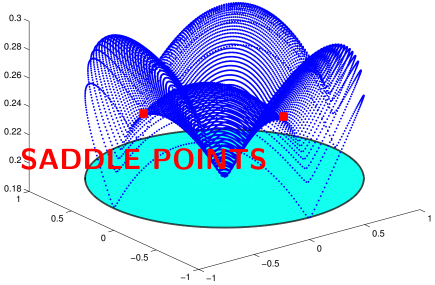
Noncritical Point ($\nabla\varphi \neq \mathbf{0}$) **Critical Points ($\nabla\varphi = \mathbf{0}$)**

"Any Reasonable Algorithm" Works

Key issue: using negative curvature

$$\lambda_{\min}(\mathrm{Hess} f) < 0$$

to escape saddles.

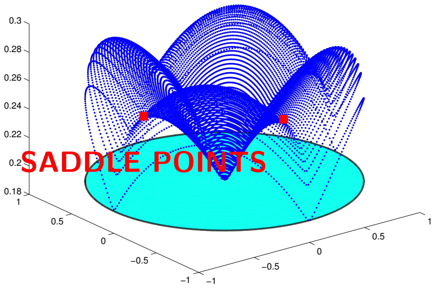


"Any Reasonable Algorithm" Works

Key issue: using negative curvature

$$\lambda_{\min}(\mathrm{Hess} f) < 0$$

to escape saddles.



Efficient (polynomial time) methods:

Trust region method, analyses in [Sun, Qu, W., '17]

Curvilinear search, [Goldfarb, Mu, W., Zhou, '16]

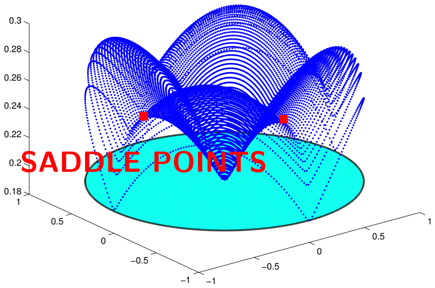
Noisy (stochastic) gradient descent, [Jin et. al. '17].

"Any Reasonable Algorithm" Works

Key issue: using negative curvature

$$\lambda_{\min}(\mathrm{Hess} f) < 0$$

to escape saddles.



Efficient (polynomial time) methods:

Trust region method, analyses in [Sun, Qu, W., '17]

Curvilinear search, [Goldfarb, Mu, W., Zhou, '16]

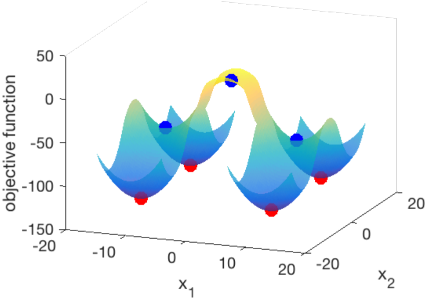
Noisy (stochastic) gradient descent, [Jin et. al. '17].

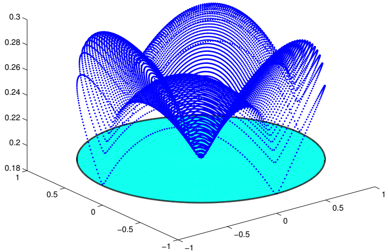
Randomly initialized gradient descent ....

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





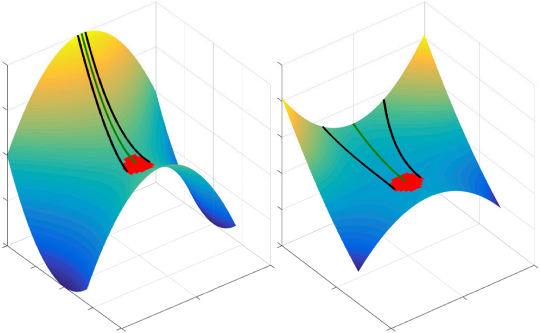
Worst Case

[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

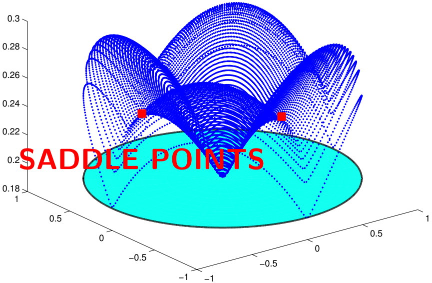


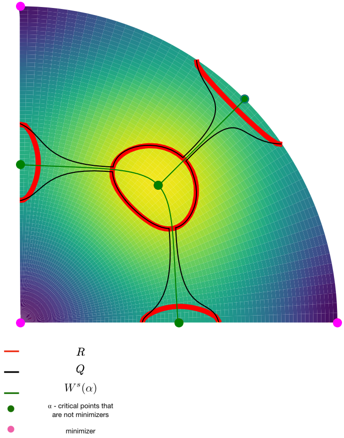
- Red: “slow region” of small gradient around a saddle point.
- Green: stable manifold associated with the saddle point.
- Black: points that flow to the slow region.

- 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.

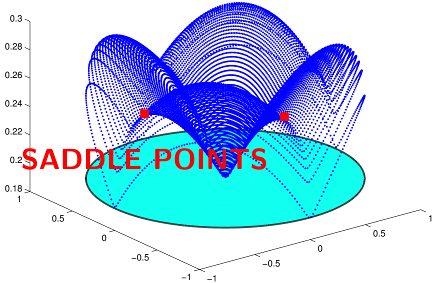
.

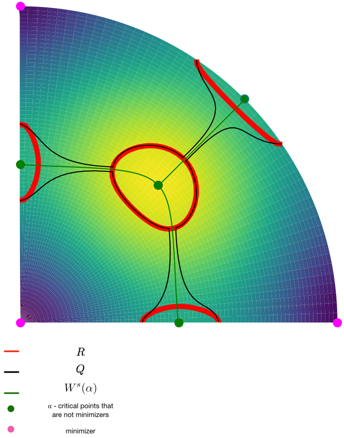
Gradient Descent Works for DL and Related Problems





Gradient Descent Works for DL and Related Problems





Dispersive structure: Negative curvature $\perp$ stable manifolds.

W.h.p. in random initialization $q^{(0)} \sim \mathrm{uni}(\mathbb{S}^{n-1})$, **convergence to a neighborhood of a minimizer in polynomial iterations.** [Gilboa,

Alternating Descent Method

$$\min_{\boldsymbol{a}\in\mathbb{S}^{n-1},\boldsymbol{x}} \underbrace{\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{a}*\boldsymbol{x}\|_F^2}_{\text{smooth } g} + \lambda \underbrace{\|\boldsymbol{x}\|_1}_{\text{nonsmooth } h}$$

- Fix $\boldsymbol{a}$ and take a **proximal** descent step on $\boldsymbol{x}$

$$\boldsymbol{x}^{(k+1)} \leftarrow \mathrm{prox}_h^{\lambda t}\left(\boldsymbol{x}^{(k)} - t\nabla g(\boldsymbol{a}^{(k)}, \boldsymbol{x}^{(k)})\right)$$

- Fix $\boldsymbol{x}$ and take a **projected** descent step on $\boldsymbol{a}$

$$\boldsymbol{a}^{(k+1)} \leftarrow \mathcal{P}_{\mathbb{S}^{n-1}}\left(\boldsymbol{a}^{(k)} - t'\mathrm{grad}_g(\boldsymbol{a}^{(k)}, \boldsymbol{x}^{(k)})\right)$$

Inertial Alternating Descent Method

Accelerating first-order descent with Momentum

- Fix $\boldsymbol{a}$ and take an **accelerated proximal** descent step on $\boldsymbol{x}$

$$\boldsymbol{w}^{(k)} = \boldsymbol{x}^{(k)} + \beta\left(\boldsymbol{x}^{(k)} - \boldsymbol{x}^{(k-1)}\right)$$
$$\boldsymbol{x}^{(k+1)} \leftarrow \mathrm{prox}_h^{\lambda t}\left(\boldsymbol{x}^{(k)} - t\nabla g(\boldsymbol{a}^{(k)}, \boldsymbol{w}^{(k)})\right)$$

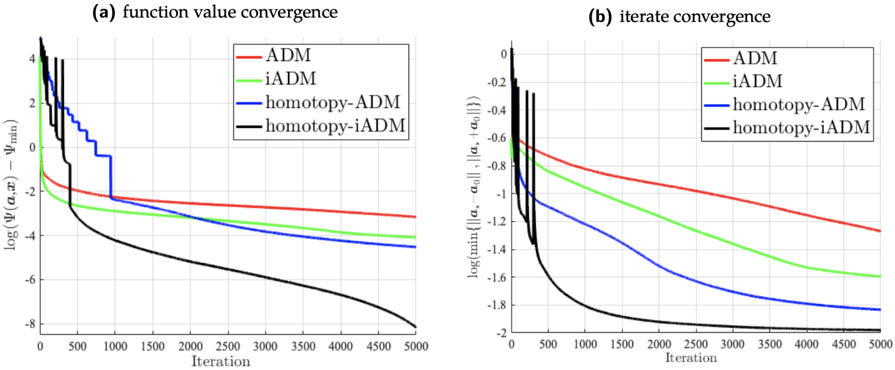
- Fix $\boldsymbol{x}$ and take an **accelerated projected** descent step on $\boldsymbol{a}$

$$\boldsymbol{z}^{(k)} = \boldsymbol{a}^{(k)} + \beta\left(\boldsymbol{a}^{(k)} - \boldsymbol{a}^{(k-1)}\right)$$
$$\boldsymbol{a}^{(k+1)} \leftarrow \mathcal{P}_{\mathbb{S}^{n-1}}\left(\boldsymbol{a}^{(k)} - t'\mathrm{grad}_g(\boldsymbol{z}^{(k)}, \boldsymbol{x}^{(k)})\right)$$

Convergence Comparison

For blind deconvolution problem

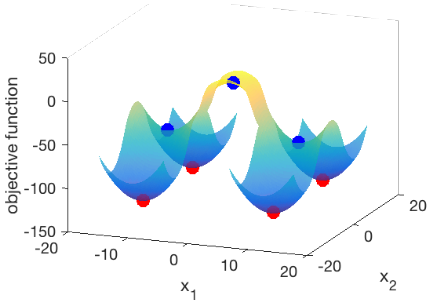
$$\min_{\boldsymbol{a}\in\mathbb{S}^{n-1},\boldsymbol{x}} \quad \frac{1}{2}\left\|\boldsymbol{y}-\boldsymbol{a}*\boldsymbol{x}\right\|_F^2+\lambda\left\|\boldsymbol{x}\right\|_1$$



3

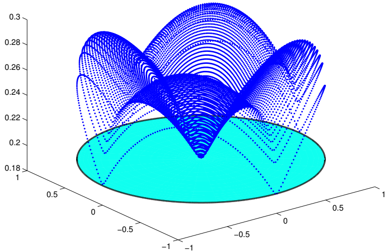
[3] The homotopy counterpart shrinks λ in every iteration.

Escaping Saddles in Worst Case Problems



Worst Case

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



Naturally Occuring

DL, Other sparsification problems
Dispersion away from stable manifold

Trust Region Method

Function class: f nonconvex.

The oracle: gradient $\nabla f(\boldsymbol{x})$, Hessian $\nabla^2 f(\boldsymbol{x})$, and the trusted region radius r

Trust region update:

$$\boldsymbol{x}^{(t+1)} = \boldsymbol{x}^{(t)} + \boldsymbol{\delta}$$

with

$$\boldsymbol{\delta} = \arg\min_{\|\boldsymbol{\delta}\| \leq r} \; f\left(\boldsymbol{x}^{(t)}\right) + \left\langle \nabla f(\boldsymbol{x}^{(k)}), \boldsymbol{\delta} \right\rangle + \frac{1}{2} \boldsymbol{\delta}^T \nabla^2 f\left(\boldsymbol{x}^{(k)}\right) \boldsymbol{\delta}$$

- At any stationary point, the gradient vanishes, and the above optimization problem boils down to the Hessian term;
- At an local solution with positive semi-definite Hessian, the above optimization problem renders $\boldsymbol{\delta} = \mathbf{0}$.

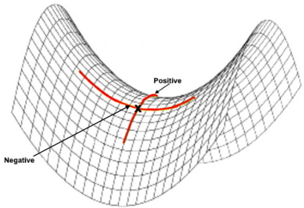
Gradient Descent with Small Random Noise

Function class: f nonconvex and Lips. continuous.

The oracle: gradient $\nabla f(\boldsymbol{x})$ and small random noise.

The updates for noisy gradient descent (Langevine dynamics):

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - t_1 \nabla f(\boldsymbol{x}^{(k)}) + t_2 \boldsymbol{n},$$



This avoids computing expensive Hessian.

Hybrid Noisy Gradient Descent

Function class: f nonconvex and Lips. continuous.

The oracle: gradient $\nabla f(\boldsymbol{x})$ and small noise $\boldsymbol{n}$.

Hybrid noisy gradient descent:

- **if** $\|\nabla f(\boldsymbol{x}_k)\|_2 \geq \epsilon_g$, **then** $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - t_1 \nabla f(\boldsymbol{x}_k)$;
- **else** $\boldsymbol{x}_k^0 = \boldsymbol{x}_k$, and negative curvature descent with noisy gradients: **for** $i = 0, 1, 2, \ldots, k_{\max} = O(\log n)$

$$\boldsymbol{x}_k^{i+1} = \boldsymbol{x}_k^i - t_1 \nabla f(\boldsymbol{x}_k^i) + t_2 \boldsymbol{n}^i,$$

where $\boldsymbol{n}^i \sim \mathcal{N}(0, \boldsymbol{I})$.

More saddle-escaping first-order optimization methods in book: Wright and Ma: https://book-wright-ma.github.io.

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

Next lecture: Exploiting Low-D Structures via Deep Networks.

Thank You! Questions?