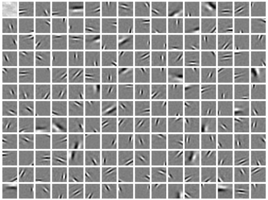
Computational Principles for High-dim Data Analysis

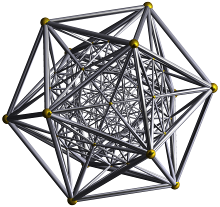
(Lecture Fifteen)

Yi Ma

EECS Department, UC Berkeley

October 21, 2021





Nonconvex Methods for Low-Dimensional Models
Dictionary Learning

"The mathematical sciences particularly exhibit order, symmetry, and limitations; and these are the greatest forms of the beautiful."

– *Aristotle*, Metaphysica

Example: Magnetic Resonance Imaging

Simplified linear measurement model for MRI:

$$y = \mathcal{F}[I](\boldsymbol{u}) = \int_{\boldsymbol{v}} I(\boldsymbol{v}) \exp(-\mathrm{i}\, 2\pi\, \boldsymbol{u}^* \boldsymbol{v})\, d\boldsymbol{v} \;\in \mathbb{C}. \quad (1)$$

Real physical measurements as modulus:

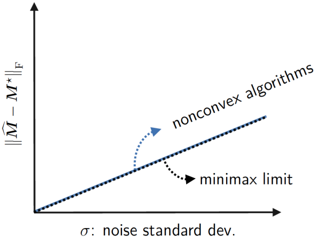
$$y = |\mathcal{F}[I](\boldsymbol{u})| \;\in \mathbb{R}_+. \quad (2)$$

Fourier phase retrieval from multiple nonlinear real measurements:

$$\underbrace{\boldsymbol{y}}_{\text{observation}} = \Big| \mathcal{F}\Big(\underbrace{\boldsymbol{x}}_{\text{unknown signal}} \Big) \Big| \;\in \mathbb{R}_+^m. \quad (3)$$

Example: Low-rank Matrix Completion

minimax limit	$\sigma\sqrt{n/p}$
nonconvex algorithms	$\sigma\sqrt{n/p}$ (optimal!)



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[1]:

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

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

[1]figure courtesy from the lecture by Prof. Yuxin Chen of Princeton

Example: Dictionary for Image Representation

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

$$I_{\text{noisy}} = \underset{\text{dictionary}}{\boldsymbol{A}} \times \underset{\text{sparse}}{\boldsymbol{x}} + \underset{\text{noise}}{\boldsymbol{z}}. \quad (4)$$

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

$$\underset{\textbf{data}}{\boldsymbol{Y}} = \underset{\textbf{dictionary}}{\boldsymbol{A}} \underset{\textbf{sparse}}{\boldsymbol{X}}. \quad (5)$$

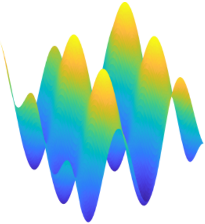
- Band-limited signals: $\boldsymbol{A} = \boldsymbol{F}$, the Fourier transform;
- Piecewise smooth signals: $\boldsymbol{A} = \boldsymbol{W}$, the wavelet transforms;
- Natural images $\boldsymbol{A} = ?$ (How to **learn** $\boldsymbol{A}$ from the data $\boldsymbol{Y}$?)

Challenges of Nonconvex Optimization – Pessimistic Views

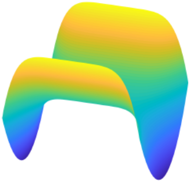
Consider the problem of minimizing a general nonlinear function:

$$\min_{\boldsymbol{z}} \varphi(\boldsymbol{z}), \quad \boldsymbol{z} \in \mathsf{C}. \qquad (6)$$

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



Spurious local minimizers



Flat saddle points

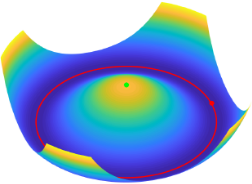
Hence typically people seek to work with relatively benign functions with benign guarantees (Chapter 9):

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 $\nabla^2\varphi(\bar{\boldsymbol{z}}) \succeq \mathbf{0}$.

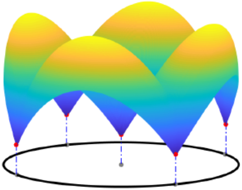
[2]Some NP-complete problems in quadratic and nonlinear programming, K.G Murty and S. N. Kabadi, 1987

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 phase recovery, invariant under a continuous rotation:

$$\varphi(e^{i\theta}\boldsymbol{x}) = \varphi(\boldsymbol{x}), \quad \forall \theta \in [0, 2\pi) = \mathbb{S}^1,$$

- for dictionary learning, invariant under signed permulations:

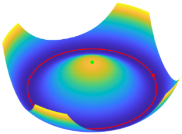
$$\varphi((\boldsymbol{A}, \boldsymbol{X})) = \varphi((\boldsymbol{A}\boldsymbol{\Pi}, \boldsymbol{\Pi}^* \boldsymbol{X})), \quad \forall \boldsymbol{\Pi} \in \mathsf{SP}(n),$$

Optimization under Symmetry

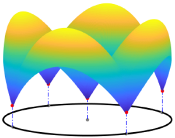
Definition (Symmetric Function)

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

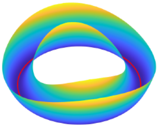
Slogan 1: the (only!) local minimizers are symmetric versions of the ground truth.
Slogan 2: any local critical point has negative curvature in directions that break symmetry.

Taxonomy of Symmetric Nonconvex Problems

Nonconvex Problems with Rotational Symmetries

Eigenspace Computation

Compute the principal subspace of a symmetric matrix.



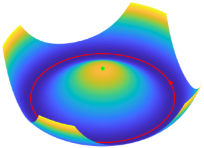
$\min_{X^* X = I} -\frac{1}{2}\mathrm{trace}\left[X^* A X\right].$

Symmetry: $X \mapsto XR$

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

Generalized Phase Retrieval

Recover a complex vector x_o *from magnitude measurements* $y = |Ax_o|$.



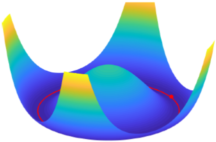
$\min_{x} \frac{1}{2}\|y^2 - |Ax|^2\|_2^2.$

Symmetry: $x \mapsto xe^{i\phi}$

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

Matrix Recovery

Recover a low-rank matrix $X = UV^*$ *from incomplete/corrupted observations*



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

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

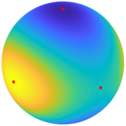
$\mathbb{G} = \mathsf{GL}(r)$ or $\mathbb{G} = \mathsf{O}(r)$

Taxonomy of Symmetric Nonconvex Problems

Nonconvex Problems with Discrete Symmetries

Eigenvector Computation

Maximize a quadratic form over the sphere.



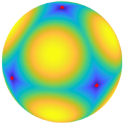
$\max_{\boldsymbol{x}\in\mathbb{S}^{n-1}} \frac{1}{2}\boldsymbol{x}^* \boldsymbol{A}\boldsymbol{x}.$

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*



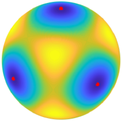
$\min_{\boldsymbol{A}\in\mathcal{A},\boldsymbol{X}} \frac{1}{2}\|\boldsymbol{Y} - \boldsymbol{A}\boldsymbol{X}\|_F^2 + \lambda\|\boldsymbol{X}\|_1.$

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* $\boldsymbol{T} = \sum_i \boldsymbol{a}_i \otimes \boldsymbol{a}_i \otimes \boldsymbol{a}_i \otimes \boldsymbol{a}_i$



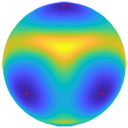
$\max_{\boldsymbol{X}\in\mathsf{O}(n)} \sum_i \boldsymbol{T}(\boldsymbol{x}_i, \boldsymbol{x}_i, \boldsymbol{x}_i, \boldsymbol{x}_i).$

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 $\mathbb{G} = \mathbb{Z}_n \times \{\pm 1\}$

Dictionary Learning: the Minimal Case

Dictionary Learning with **one sparsity:**

$$\underbrace{\boldsymbol{Y}}_{\text{data}} = \underbrace{\boldsymbol{A}_o}_{\text{orthogonal dictionary}} \underbrace{\boldsymbol{X}_o}_{\text{1-sparse coefficients}}. \tag{7}$$

Signed permutation symmetry:

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

Search for an orthogonal $\boldsymbol{A}$ such that $\boldsymbol{A}^* \boldsymbol{Y}$ is *as sparse as possible*:

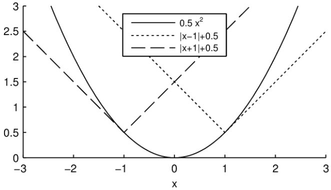
$$\min \; h(\boldsymbol{A}^* \boldsymbol{Y}) \quad \text{such that} \quad \boldsymbol{A} \in \mathsf{O}(m), \tag{8}$$

where $h(\boldsymbol{X}) = \sum_{ij} h(\boldsymbol{X}_{ij})$ is a function that promotes sparsity.

Find One Atom at a Time

Take h to be **the Huber function**:

$$h_\lambda(x) = \begin{cases} \lambda|x| - \lambda^2/2 & |x| > \lambda, \\ x^2/2 & |x| \le \lambda. \end{cases} \quad (9)$$



This can be viewed as a differentiable surrogate for the ℓ^1 norm.

For the dictionary $\boldsymbol{A} = [\boldsymbol{a}_1, \ldots, \boldsymbol{a}_m]$, find the columns $\boldsymbol{a}_i$ one at a time:

$$\min \ \varphi(\boldsymbol{a}) \doteq h_\lambda(\boldsymbol{a}^* \boldsymbol{Y}) \quad \text{such that} \quad \boldsymbol{a} \in \mathbb{S}^{m-1}. \quad (10)$$

Dictionary Learning: the Simplest Case

WLOG, assume $\boldsymbol{A}_o = \boldsymbol{I}$, and $\boldsymbol{X}_o = \boldsymbol{I}$ (uniformly random sampling).

$$\min \ \varphi(\boldsymbol{a}) \doteq h_\lambda(\boldsymbol{a}) \quad \text{such that} \quad \boldsymbol{a} \in \mathbb{S}^{m-1}. \tag{11}$$

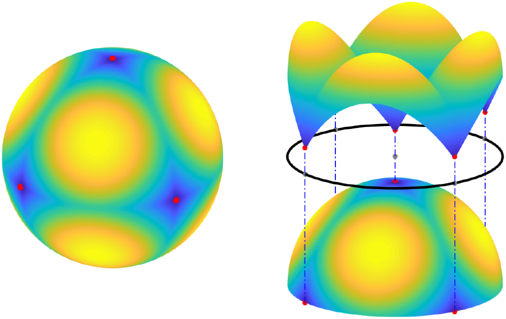


Figure: $h_\lambda(\boldsymbol{u})$ as a function on the sphere $\mathbb{S}^2$.

First Order Characteristics of The Simplest Case

Critical Points of φ.

The gradient of φ:

$$\nabla\varphi(\boldsymbol{a}) = \lambda\,\mathrm{sign}(\boldsymbol{a}) \odot \mathbb{1}_{|\boldsymbol{a}|>\lambda} \; + \; \boldsymbol{a} \odot \mathbb{1}_{|\boldsymbol{a}|\le\lambda}, \tag{12}$$

where $\odot$ denotes element-wise multiplication.

The Riemannian gradient is (tangent to the sphere $\mathbb{S}^{m-1}$):

$$\mathrm{grad}[\varphi](\boldsymbol{a}) = \boldsymbol{P}_{\boldsymbol{a}^\perp}\nabla\varphi(\boldsymbol{a}). \tag{13}$$

The Riemannian gradient vanishes iff $\nabla\varphi(\boldsymbol{a}) \;\propto\; \boldsymbol{a}$, which occurs whenever

$$\boldsymbol{a} \;\propto\; \mathrm{sign}(\boldsymbol{a}). \tag{14}$$

Second Order Characteristics of the Simplest Case

Hessian at Critical Points of φ.

The *Riemannian Hessian* is given by[3]

$$\begin{aligned} \mathrm{Hess}[\varphi](\boldsymbol{a}) &= \boldsymbol{P}_{\boldsymbol{a}^\perp}\Big(\underbrace{\nabla^2\varphi(\boldsymbol{a})}_{\text{curvature of } \varphi} - \underbrace{\langle \nabla\varphi(\boldsymbol{a}), \boldsymbol{a}\rangle \boldsymbol{I}}_{\text{curvature of the sphere}} \Big) \boldsymbol{P}_{\boldsymbol{a}^\perp} \\ &= \boldsymbol{P}_{\boldsymbol{a}_{\mathsf{I},\boldsymbol{\sigma}}^\perp} \left(\boldsymbol{P}_{|\boldsymbol{a}_{\mathsf{I},\sigma}| \le \lambda} - \lambda |\mathsf{I}| \boldsymbol{I} \right) \boldsymbol{P}_{\boldsymbol{a}_{\mathsf{I},\boldsymbol{\sigma}}^\perp}. \end{aligned}$$

At critical points $\boldsymbol{a}_{\mathsf{I},\sigma}$ the Hessian exhibits $(|\mathsf{I}| - 1)$ negative eigenvalues, and $m - |\mathsf{I}|$ positive eigenvalues.

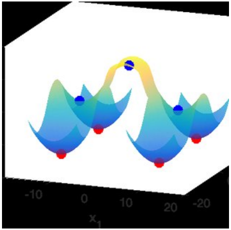
[3] can be derived by calculating $\frac{d^2}{dt^2}\Big|_{t=0} \varphi\Big(\boldsymbol{a}\cos t + \boldsymbol{\delta}\sin t\Big)$, with any direction $\boldsymbol{\delta} \in T_{\boldsymbol{a}}\mathbb{S}^{m-1}$ and $\|\boldsymbol{\delta}\| = 1$.

General Messages from the Simplest Case

Symmetric copies of the ground truth are minimizers. The objective function is strongly convex in the vicinity of local minimizers $\boldsymbol{a} = \pm \boldsymbol{e}_i$.

Negative curvature in symmetry breaking directions. Saddle points are balanced superpositions of target solutions: $\boldsymbol{a}_{\mathsf{I},\sigma} = \frac{1}{\sqrt{|\mathsf{I}|}} \sum_{i \in \mathsf{I}} \sigma_i \boldsymbol{e}_i$ with I and signs $\sigma_i \in \{\pm 1\}$. There is negative curvature in directions $\boldsymbol{\delta} \in \mathrm{span}(\{\boldsymbol{e}_i \mid i \in \mathsf{I}\})$ that break the balance between target solutions.

Cascade of saddle points. Downstream negative curvature directions are the image of upstream negative curvature directions under gradient flow. Worst case, such as the "octopus function" shown in the Figure[4], never occurs!



[4] Gradient Descent Can Take Exponential Time to Escape Saddle Points, S. Du et. al, NeurIPS 2017.

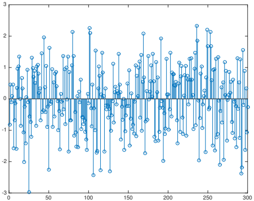
Dictionary Learning: General Case

A Fundamental Problems in Data Analysis:

Given an n-dimensional signal: $\boldsymbol{y} \in \mathbb{R}^n$, find a transformation $\mathcal{T} : \mathbb{R}^n \to \mathbb{R}^m$ or its "inverse" $\boldsymbol{D} : \mathbb{R}^m \to \mathbb{R}^n$, such that

$$\boldsymbol{x} = \mathcal{T}[\boldsymbol{y}], \quad \text{or} \quad \boldsymbol{y} = \boldsymbol{D}\boldsymbol{x}$$

where $\boldsymbol{x}$ highly compressible or the sparsest possible.



$\boldsymbol{y} \Rightarrow \boldsymbol{x}$

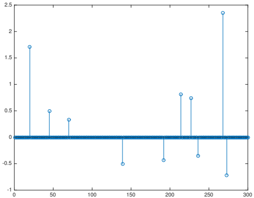


Figure: **Sparse Representation** Left: a *generic* vector $\boldsymbol{y} \in \mathbb{R}^n$, Right: a *sparse* representation $\boldsymbol{x} = \mathcal{T}[\boldsymbol{y}]$, after a proper transformation $\mathcal{T}$.

Introduction: History of Finding Good Transform

Figure: Joseph Fourier, $1768 - 1830$

- **Fourier Transform** $\boldsymbol{D} = \boldsymbol{F}$
- Wavelet Transform $\boldsymbol{D} = \boldsymbol{W}$
- Dictionary Learning

Introduction: Fourier Transform

Assumption:
The signal $\boldsymbol{y}$ is **band-limited and sparse** in frequency domain: $y_k = \sum_{l=0}^{n-1} x_l \cdot e^{-\frac{i2\pi}{n}kl}$ ($\boldsymbol{y} = \boldsymbol{F}\boldsymbol{x}$.)

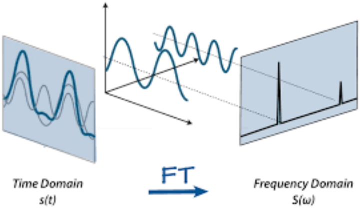


Figure: Fourier Transform

Figure: Lena Compression using Discrete Cosine Transform (JPEG) [pip18]

Introduction: History of Finding Good Transform

Figure: Alfred Haar, $1855 - 1933$

- Fourier Transform $\boldsymbol{D} = \boldsymbol{F}$
- **Wavelet Transform** $\boldsymbol{D} = \boldsymbol{W}$
- Dictionary Learning

Introduction: Wavelet Transform

Assumption:
Signal $\boldsymbol{y}$ is piece-wise smooth, scale-invariant, etc: $\boldsymbol{y} = \boldsymbol{W}\boldsymbol{x}$, $\boldsymbol{W}^*\boldsymbol{W} = \boldsymbol{I}$.

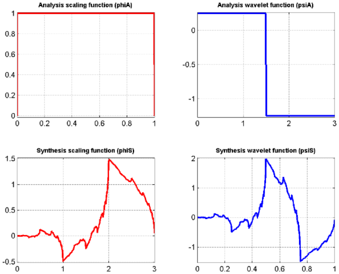


Figure: Haar & Daubechies Wavelets

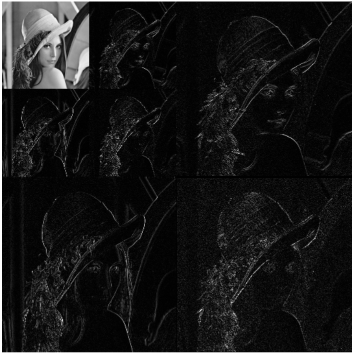


Figure: Lena Compression using Wavelet Transform (JPEG2000) [Jor06]

Why Dictionary Learning?

Limitations of Traditional "By Design" Methods

- A transform is not optimal for signals that do not satisfy the conditions under which the transform is designed (e.g. DCT not ideal for images).
- For different classes of signals, we need to design different transforms (e.g. all the x-lets), which may not even be possible if the properties are not clear.

Why Dictionary Learning?

Limitations of Traditional "By Design" Methods

- A transform is not optimal for signals that do not satisfy the conditions under which the transform is designed (e.g. DCT not ideal for images).
- For different classes of signals, we need to design different transforms (e.g. all the x-lets), which may not even be possible if the properties are not clear.

For a given class of signals, can we directly "learn" the corresponding optimal transform, from its samples?

Dictionary Learning: General Case

Given n-dimensional input data: $\{\boldsymbol{y}_1, \dots, \boldsymbol{y}_p\}$, $\forall i \in [p], \boldsymbol{y}_i \in \mathbb{R}^n$, find a dictionary $\boldsymbol{D} \in \mathbb{R}^{n \times m}$ and its corresponding coefficients $\{\boldsymbol{x}_1, \dots, \boldsymbol{x}_p\}$, $\boldsymbol{x}_i \in \mathbb{R}^m$, such that

$$\boldsymbol{y}_i = \boldsymbol{D}\boldsymbol{x}_i, \quad \forall i \in [p], \tag{15}$$

and $\boldsymbol{x}_i$ is sufficiently sparse. That is to factor the data matrix $\boldsymbol{Y}$ into **two structured unknowns**: a matrix $\boldsymbol{D}$ and a sparse matrix $\boldsymbol{X}$:

$$\boldsymbol{Y} = \underbrace{\begin{pmatrix} | & & | \\ \boldsymbol{y}_1 & \dots & \boldsymbol{y}_p \\ | & & | \end{pmatrix}}_{\text{Observations}} = \underbrace{\begin{pmatrix} d_{1,1} & \dots & d_{1,m} \\ \vdots & \ddots & \vdots \\ d_{n,1} & \dots & d_{n,m} \end{pmatrix}}_{\text{Dictionary } \boldsymbol{D}} \underbrace{\begin{pmatrix} | & & | \\ \boldsymbol{x}_1 & \dots & \boldsymbol{x}_p \\ | & & | \end{pmatrix}}_{\boldsymbol{X} \text{ is sparse, } \|\boldsymbol{x}_i\|_0 \ll m} = \boldsymbol{D}\boldsymbol{X}.$$

Dictionary Learning: General Case

Challenges

- **Computational Complexity**
  Optimizing a nonconvex bilinear problem is NP-hard.
- **Sample Complexity**
  Combinatorial possible outcomes for $k-$sparse $\boldsymbol{x}$.
- **Signed Permutation Ambiguities**
  $\forall \boldsymbol{P} \in \mathsf{SP}(m)$, [5] $(\boldsymbol{D}_\star \boldsymbol{P}, \boldsymbol{P}^* \boldsymbol{X}_\star)$ and $(\boldsymbol{D}_\star, \boldsymbol{X}_\star)$ are equally sparse.

[5] $\mathsf{SP}(m)$ denote m dimensional signed permutation group, a group of orthogonal matrices whose entries contain only $0, \pm 1$.

Dictionary Learning: General Case

Challenges

- **Computational Complexity**
  Optimizing a nonconvex bilinear problem is NP-hard.
- **Sample Complexity**
  Combinatorial possible outcomes for $k-$sparse $\boldsymbol{x}$.
- **Signed Permutation Ambiguities**
  $\forall \boldsymbol{P} \in \mathsf{SP}(m)$, [5] $(\boldsymbol{D}_\star \boldsymbol{P}, \boldsymbol{P}^* \boldsymbol{X}_\star)$ and $(\boldsymbol{D}_\star, \boldsymbol{X}_\star)$ are equally sparse.

Some heuristic algorithms

- K-SVD [AEB$^+$06]
- Alternative Direction Methods [SQW17]

[5] $\mathsf{SP}(m)$ denote m dimensional signed permutation group, a group of orthogonal matrices whose entries contain only $0, \pm 1$.

Dictionary Learning: General Case

Challenges

- **Computational Complexity**
  Optimizing a nonconvex bilinear problem is NP-hard.
- **Sample Complexity**
  Combinatorial possible outcomes for $k-$sparse $\boldsymbol{x}$.
- **Signed Permutation Ambiguities**
  $\forall \boldsymbol{P} \in \mathsf{SP}(m)$, [5] $(\boldsymbol{D}_\star \boldsymbol{P}, \boldsymbol{P}^* \boldsymbol{X}_\star)$ and $(\boldsymbol{D}_\star, \boldsymbol{X}_\star)$ are equally sparse.

Some heuristic algorithms

- K-SVD [AEB$^+$06]
- Alternative Direction Methods [SQW17]

Learn the dictionary with tractable algorithms and sample size?

[5] $\mathsf{SP}(m)$ denote m dimensional signed permutation group, a group of orthogonal matrices whose entries contain only $0, \pm 1$.

Complete Dictionary Learning – Prior Arts

A Random Model:

For complete dictionary learning, [SWW12] assumes data $\boldsymbol{Y}$ is generated by a complete[6] dictionary $\boldsymbol{D}_o$ and sparse coefficients $\boldsymbol{X}_o$:

$$\boldsymbol{Y} = \boldsymbol{D}_o \boldsymbol{X}_o,$$

where $\boldsymbol{X}_o$ follows a Bernoulli Gaussian model:

$$\boldsymbol{X}_o = \boldsymbol{\Omega} \circ \boldsymbol{G}^{7}, \quad \Omega_{i,j} \sim_{iid} \text{Ber}(\theta), G_{i,j} \sim_{iid} \mathcal{N}(0,1).$$

[6] square and invertible

[7] $\circ$ denote element-wise product: $\forall \boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{n \times m}$, $\{\boldsymbol{A} \circ \boldsymbol{B}\}_{i,j} = a_{i,j} b_{i,j}$

Complete Dictionary Learning – Prior Arts

A Random Model:

For complete dictionary learning, [SWW12] assumes data $\boldsymbol{Y}$ is generated by a complete[6] dictionary $\boldsymbol{D}_o$ and sparse coefficients $\boldsymbol{X}_o$:

$$\boldsymbol{Y} = \boldsymbol{D}_o \boldsymbol{X}_o,$$

where $\boldsymbol{X}_o$ follows a Bernoulli Gaussian model:

$$\boldsymbol{X}_o = \boldsymbol{\Omega} \circ \boldsymbol{G}^7, \quad \Omega_{i,j} \sim_{iid} \mathsf{Ber}(\theta), G_{i,j} \sim_{iid} \mathcal{N}(0,1).$$

Preconditioning:

[SQW17] shows that learning a complete dictionary is equivalent with learning an orthogonal one through preconditioning

$$\bar{\boldsymbol{Y}} \leftarrow \left(\frac{1}{p\theta}\boldsymbol{Y}\boldsymbol{Y}^*\right)^{-\frac{1}{2}}\boldsymbol{Y} = \boldsymbol{D}_o\boldsymbol{X}_o, \quad \text{with} \quad \boldsymbol{D}_o \in \mathsf{O}(n).$$

[6] square and invertible

[7] $\circ$ denote element-wise product: $\forall \boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{n\times m}$, $\{\boldsymbol{A} \circ \boldsymbol{B}\}_{i,j} = a_{i,j} b_{i,j}$

Complete Dictionary Learning – Prior Arts

Complete dictionary learning can be reduced to find the sparsest direction in a subspace:

1. $\boldsymbol{D}_o$ is complete $\implies$ $\boxed{\mathrm{row}(\boldsymbol{Y}) = \mathrm{row}(\boldsymbol{X}_o)}$
2. Rows of $\boldsymbol{X}_o$ form *a sparse basis* of $\mathrm{row}(\boldsymbol{Y})$.
3. Find $\boldsymbol{x}_1$, *the sparsest vector* in the subspace $\mathrm{row}(\boldsymbol{Y})$.
4. Find $\boldsymbol{x}_i$, *the sparsest vector* in $\mathrm{row}(\boldsymbol{Y})\backslash\{\boldsymbol{x}_1, , \ldots, \boldsymbol{x}_{i-1}\}$.
5. Recover $\boldsymbol{D}_o$ by: $\boldsymbol{D}_o = \boldsymbol{Y}\boldsymbol{X}_o^*(\boldsymbol{X}_o\boldsymbol{X}_o^*)^{-1}$.

Complete Dictionary Learning – Prior Arts

Finding the sparsest vector in $\mathrm{row}(\boldsymbol{Y})$ can be naïvely formulated as

$$\min_{\boldsymbol{q}} \|\boldsymbol{q}^* \boldsymbol{Y}\|_0 \,, \quad \text{such that} \quad \boldsymbol{q} \neq \boldsymbol{0},$$

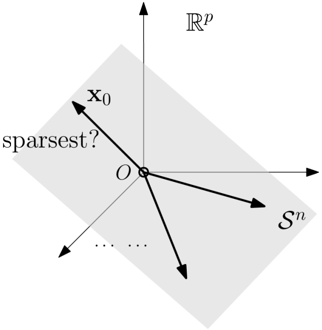


Figure: The sparsest direction in a subspace. Credit: Prof. Qing Qu.

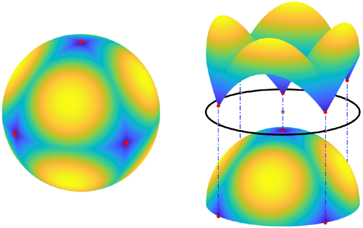
Related Works in Finding the Sparsest Direction

- Linear Programming [SWW12]:

$$\min_{\boldsymbol{q}} \|\boldsymbol{q}^* \boldsymbol{Y}\|_1, \quad \text{such that} \quad \|\boldsymbol{q}^* \boldsymbol{Y}\|_\infty = 1.$$

- Nonconvex Optimization on a Sphere [SQW17, BJS18]:

$$\min_{\boldsymbol{q}} \|\boldsymbol{q}^* \boldsymbol{Y}\|_1, \quad \text{such that} \quad \|\boldsymbol{q}\|_2 = 1.$$



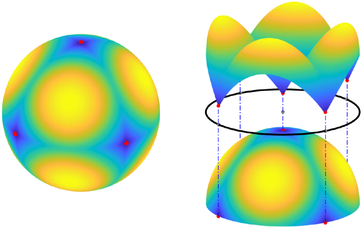
Related Works in Finding the Sparsest Direction

- Linear Programming [SWW12]:

$$\min_{\boldsymbol{q}} \|\boldsymbol{q}^* \boldsymbol{Y}\|_1 , \quad \text{such that} \quad \|\boldsymbol{q}^* \boldsymbol{Y}\|_\infty = 1.$$

- Nonconvex Optimization on a Sphere [SQW17, BJS18]:

$$\min_{\boldsymbol{q}} \|\boldsymbol{q}^* \boldsymbol{Y}\|_1 , \quad \text{such that} \quad \|\boldsymbol{q}\|_2 = 1.$$



Solving the same optimization n times (high computational cost)!

Assignments

- Reading: Section 7.1 - 7.3 of Chapter 7.
- Programming Homework #3.

References I

Michal Aharon, Michael Elad, Alfred Bruckstein, et al.
K-svd: An algorithm for designing overcomplete dictionaries for sparse representation.
IEEE Transactions on signal processing, 54(11):4311, 2006.

Yu Bai, Qijia Jiang, and Ju Sun.
Subgradient descent learns orthogonal dictionaries.
arXiv preprint arXiv:1810.10702, 2018.

Palle Jorgensen.
http://homepage.divms.uiowa.edu/~jorgen/Haar.html, 2006.

pipo1995_2.
https://www.taringa.net/+info/como-una-foto-de-una-playboy-se-convirtio-en-el-formato-jpg_1ejzk6, 2018.

Ju Sun, Qing Qu, and John Wright.
Complete dictionary recovery over the sphere i: Overview and the geometric picture.
IEEE Transactions on Information Theory, 63(2):853–884, 2017.

Daniel A Spielman, Huan Wang, and John Wright.
Exact recovery of sparsely-used dictionaries.
In *Conference on Learning Theory*, pages 37–1, 2012.