Complete Dictionary Learning via ℓp-norm Maximization

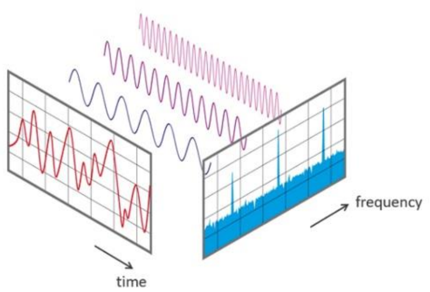
Ye XUE

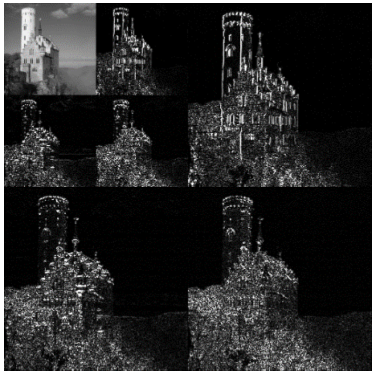
香港科技大學
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

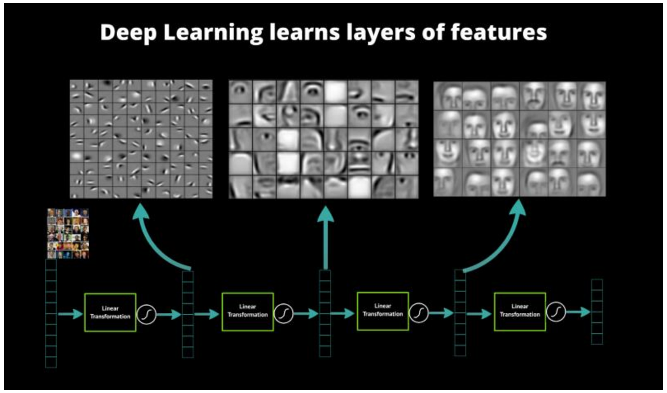
Yifei Shen	Jun Zhang	Khaled B. Letaief	Vincent Lau
HKUST	PolyU HK	HKUST	HKUST

Representation of Data

- Finding a good representation of data is what we pursuit for centuries
  - Fourier Transform (Fourier ~1800)
  - Wavelet Transform (Haar ~1900)
  - Deep Learning (Now)







What representation should we use?

"It is human nature to prefer simplicity"

Dictionary Learning – Finding the most sparse representation

$$\underbrace{\boldsymbol{Y}}_{\text{Known Data}} = \underbrace{\boldsymbol{D}}_{\text{Unknown Transform}} \cdot \underbrace{\boldsymbol{X}}_{\text{Unknown Sparse Representation}}$$

$$\begin{aligned} \underset{\boldsymbol{D},\boldsymbol{X}}{\text{minimize}} \quad & \|\boldsymbol{X}\|_0 \\ \text{subject to} \quad & \boldsymbol{Y} = \boldsymbol{D}\boldsymbol{X} \end{aligned}$$

Dictionary Learning Challenges

- Computational challenges
  - The signals are usually high dimension
  - NP-hard in general due to non-convexity
- Sample complexity challenges
  - We would like to learn from as few as possible samples

- Goal: Design an efficient algorithm to learn the optimal dictionary with tractable sample complexity

Preliminary

- Complete dictionary learning can be converted to an orthogonal dictionary learning through preconditioning [SQW15]

$$\bar{\boldsymbol{Y}} = \left(\frac{1}{p\theta} \boldsymbol{Y} \boldsymbol{Y}^* \right)^{-\frac{1}{2}} \boldsymbol{Y}$$

- Assumptions for tractable analysis
  - [SWW12] assumes the data $\boldsymbol{Y}$ is generated by an orthogonal dictionary $\boldsymbol{D}_0$ and sparse coefficients $\boldsymbol{X}_0$ with $\boldsymbol{Y} = \boldsymbol{D}_0 \boldsymbol{X}_0$
  - $\boldsymbol{D}_0 \in O(n)$, i.e., $\boldsymbol{D}_0^* \boldsymbol{D}_0 = \boldsymbol{D}_0 \boldsymbol{D}^* = \boldsymbol{I}$

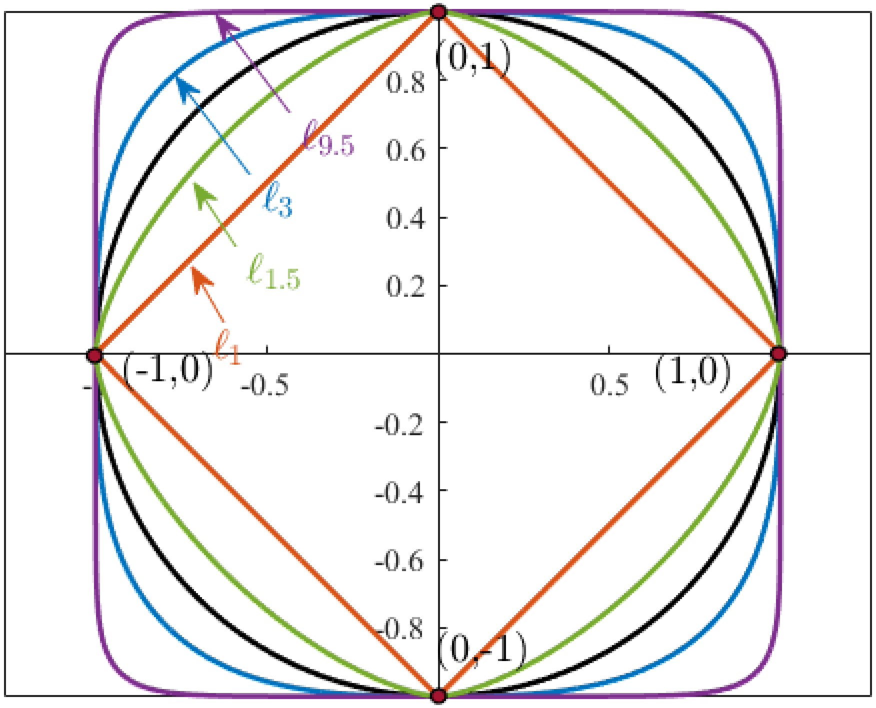
Existing Works - ℓ1-norm Minimization

$$\begin{aligned} \min_{\boldsymbol{D},\boldsymbol{X}} \quad & \|\boldsymbol{X}\|_0 \\ \text{s.t.} \quad & \boldsymbol{Y} = \boldsymbol{D}\boldsymbol{X} \\ & \boldsymbol{D} \in O(n) \end{aligned} \Rightarrow \begin{aligned} \min_{\boldsymbol{D}} \quad & \|\boldsymbol{D}^*\boldsymbol{Y}\|_0 \\ \text{s.t.} \quad & \boldsymbol{D} \in O(n) \end{aligned} \Rightarrow \begin{aligned} \min_{\boldsymbol{D}} \quad & \|\boldsymbol{D}^*\boldsymbol{Y}\|_1 \\ \text{s.t.} \quad & \boldsymbol{D} \in O(n) \end{aligned}$$

- Pros: Can achieve optimal solution [SQW15]
- Cons: Algorithmic challenges
  - Slow convergence due to non-smoothness
  - Difficult to tune parameters
- Question: Can we use other norm for relaxation?

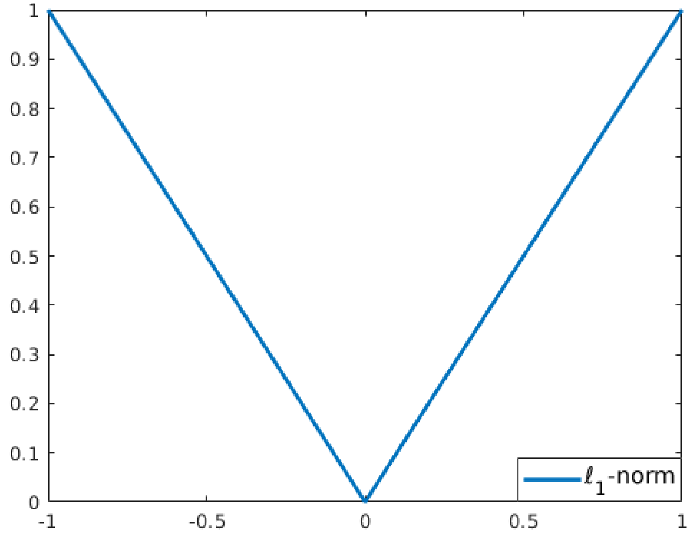
ℓp-norm Maximization - Observations

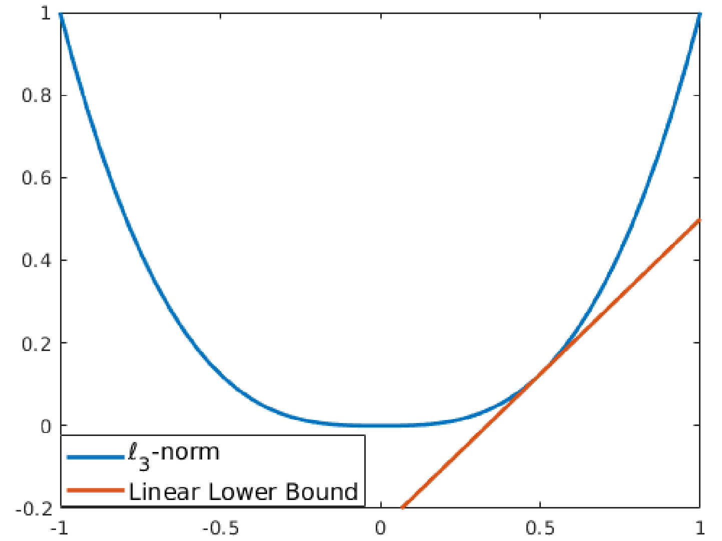
- Observation I: Over an ℓ2-norm constraint, minimizing any ℓq-norm (0<=q<2) or maximizing any ℓp-norm (p>2) are equivalent
  - Maximizing ℓp-norm (p>2) induces sparsity



ℓp-norm Maximization - Observations

- Observation II: ℓp-norms (p>2) is smooth and majorized by linear functions
  – ℓ1-norm: non-smooth, hard to optimize over a non-convex constraint
  – ℓp-norms (p>2): smooth, efficient algorithm can be designed by maximize linear lower bound at each iteration





ℓp-based Formulation

- Instead of minimizing ℓ1, we maximize ℓp (p>2)

$$\begin{aligned} \underset{\boldsymbol{D}}{\text{minimize}} \quad & \|\boldsymbol{D}^*\boldsymbol{Y}\|_1 \\ \text{subject to} \quad & \boldsymbol{D} \in O(n) \end{aligned} \quad \Rightarrow \quad \begin{aligned} \underset{\boldsymbol{D}}{\text{maximize}} \quad & \|\boldsymbol{D}^*\boldsymbol{Y}\|_p \\ \text{subject to} \quad & \boldsymbol{D} \in O(n) \end{aligned}$$

An Efficient Parameter-free Algorithm

- Maximize a linear lower bound at each time

$$f(\boldsymbol{A}) = \|\boldsymbol{A}^*\boldsymbol{Y}\|_p^p$$

$$\boldsymbol{A}^{(t+1)} = \underset{\boldsymbol{s} \in O(n)}{\operatorname{argmax}} \langle \boldsymbol{s}, f'(\boldsymbol{A}^{(t)}) \rangle$$

Algorithm 2 The GPM algorithm for ℓ_p-based dictionary learning

1: Initialize $\boldsymbol{A}^{(0)*} \in \mathrm{St}(n, m)$.
2: **for** $t = 0...T$ **do**
3: $\quad \nabla f(\boldsymbol{A}^{(t)}) = \left(|(\boldsymbol{A}^{(t)}\boldsymbol{Y})^{\circ(p-1)}| \circ \mathrm{sign}(\boldsymbol{A}^{(t)}\boldsymbol{Y}) \right) \boldsymbol{Y}^*$
4: $\quad \boldsymbol{A}^{(t+1)} = \mathrm{Polar}(\nabla^* f(\boldsymbol{A}^{(t)}))^*$
5: **end for**

ℓp-norm Minimization

- Can ℓp formulations provably recover the true dictionary?
  - $\boldsymbol{X}_0 \sim \boldsymbol{\Omega} \cdot \boldsymbol{G}, \Omega_{i,j} \sim \mathrm{Ber}(\theta), G_{i,j} \sim \mathcal{N}(0,1)$
  - All ℓp(p>2) can recover the dictionary when the sample size is large and ℓ3 achieves the best sample complexity

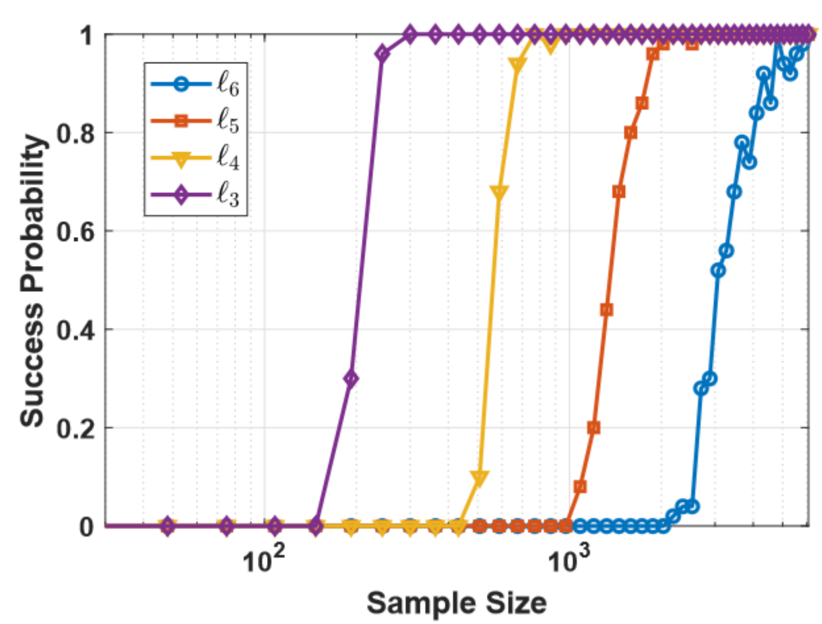
Theorem 2.1. *Let $\boldsymbol{X} \in \mathbb{R}^{n\times r}, x_{i,j} \sim \mathcal{BG}(\theta)$ with $\theta \in (0,1)$, $\boldsymbol{D}_0 \in \mathbb{O}(n)$ be an orthogonal dictionary, and $\boldsymbol{Y} = \boldsymbol{D}_0\boldsymbol{X}$. Suppose $\hat{\mathbf{A}}$ is a global maximizer to*

$$\underset{\boldsymbol{A}}{\text{maximize}} \quad \|\boldsymbol{A}\boldsymbol{Y}\|_p^p \qquad \text{subject to } \boldsymbol{A} \in \mathbb{O}(n).$$

Provided that the sample size $r = \Omega\left(\delta^{-2} n \log(n/\delta)(\theta n \log^2 n)^{\frac{p}{2}}\right)$, then for $\delta > 0$, there exists a signed permutation $\boldsymbol{\Pi}$, such that

$$\frac{1}{n}\left\|\hat{\mathbf{A}}^* - \boldsymbol{D}_0\boldsymbol{\Pi}\right\|_F^2 \leq C_\theta \delta$$

with probability at least $1 - r^{-1}$ and C_θ is a constant that depends on θ.



Experiments - Scalability

- Benchmarks
  - K-SVD [AEB06]: A classic algorithm for dictionary learning

Table 1: The performance of different algorithms for noiseless objectives. Since the dictionary recovery is up to some signed permutations, we adopt the error metric $1 - \|\boldsymbol{A}\boldsymbol{D}_0\|_4^4/n$ in [26], which gives 0% error for a perfect recovery.

Settings			ℓ_3-based		ℓ_4-based [26]		ℓ_5-based		K-SVD [18]	
n	θ	p($\times 10^4$)	Time	Error	Time	Error	Time	Error	Time	Error
100	0.1	4	**0.8s**	**0.056%**	1.8s	0.21%	1.7s	0.50%	61s	1.45%
200	0.1	8	**4.1s**	**0.056%**	9.3s	0.21%	8.0s	0.51%	131s	3.03%
400	0.1	16	**35s**	**0.056%**	50s	0.21%	41s	0.50%	315s	6.45%
100	0.3	4	**1.2s**	**0.094%**	3.4s	0.34%	3.1s	0.84%	98s	2.60%
200	0.3	8	**10s**	**0.094%**	18s	0.35%	15s	0.85%	215s	6.41%
400	0.3	16	**91s**	**0.096%**	122s	0.35%	146s	1.00%	589s	8.25%

Smaller p: faster, more accurate!

Experiments - Robustness

- Benchmarks
  - K-SVD [AEB06]: A classic algorithm for dictionary learning
  - RTR [SQW15]: ℓ1-based formulation and Riemannian Trust Region
  - ℓ4 [ZMLZM19]: ℓ4-based formulation and specialized *matching, stretching, pursuit* algorithm

Table 2: The performance of different algorithms under Gaussian noise. We set sparsity level $\theta = 0.3$.

Settings			ℓ_3-based		ℓ_4-based [26]		RTR [21]		K-SVD [18]	
n	p($\times 10^4$)	σ	Time	Error	Time	Error	Time	Error	Time	Error
32	1	0	**0.05s**	0.10%	0.24s	0.4%	100s	**0.05%**	25s	0.2%
32	1	0.2	**0.05s**	**0.27%**	0.24s	0.6%	250s	0.5%	25s	0.37%
32	1	0.4	**0.1s**	**0.79%**	0.36s	1.2%	577s	4.27%	25s	2.0%
32	1	0.6	**0.2s**	**2.3%**	0.7s	3.4%	823s	57.4%	25s	57.4%
100	4	0	**1.2s**	0.1%	3.4s	0.35%	863s	**0.05%**	98s	2.60%
100	4	0.2	**2.2s**	**0.2%**	4.2s	0.5%	1643s	0.3%	104s	3.46%
100	4	0.4	**3.5s**	**0.6%**	6.1s	1.1%	3796s	5.26%	105s	3.56%
100	4	0.6	**8.4s**	**1.95%**	13.5s	2.63%	5412s	50.5%	104s	51.26%

Beyond the work region of existing works!

Conclusion

- ℓp norm maximization enables efficient and robust solution for complete dictionary learning.
- Simple parameter-free algorithm can be used to solve ℓp norm maximization, which is tractable and optimal-achievable.
- Among all the choices of p ($p>2$), $\ell 3$ is the best.

References

- [SWW12] Daniel A Spielman, Huan Wang, and John Wright. Exact recovery of sparsely-used dictionaries. In Conference on Learning Theory, pages1–37,2012.
- [SQW15] 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.
- [AEB06] Michal Aharon, Michael Elad, and Alfred Bruckstein. K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. IEEE Transactions on Signal Processing, 54(11):4311–4322, 2006.
- [ZMLZM19] Yuexiang Zhai, Zitong Yang, Zhenyu Liao, John Wright, and Yi Ma. Complete dictionary learning via $\ell 4$-norm maximization over the orthogonal group. arXiv preprint arXiv:1906.02435, 2019.