Proximal alternating algorithms in dictionary learning

Bao Chenglong

Department of Mathematics

National University of Singapore

Joint work with Yuhui Quan, Hui Ji and Zuowei Shen.

Which kind of data we face?

Authorities are *only too aware* that Kashgar is 4,000 kilometres (2,500 miles) from Beijing but *only* a tenth of the distance from the Pakistani border, and are *desperate* to *ensure instability or militancy* does not leak over the frontiers.
Taiwan-made products *stood a good chance* of becoming *even more competitive thanks to* wider access to overseas markets and lower costs for material imports, he said.
"March *appears* to be a *more reasonable* estimate while earlier admission *cannot be entirely ruled out*," according to Chen, also Taiwan's chief WTO negotiator.
friday evening plans were great, but saturday's plans *didnt go as expected* – i went dancing & it was an *ok* club, but *terribly crowded :-(*
WHY THE *HELL* DO YOU GUYS ALL HAVE MRS. KENNEDY! SHES A FUCKING DOUCHE
AT&T was *okay* but whenever they do something *nice* in the name of customer service it seems like a favor, while T-Mobile makes that a *normal everyday thin*
obama should be *impeached* on *TREASON* charges. Our Nuclear arsenal was TOP Secret. Till HE told our enemies what we had. *#Coward #Traitor*
My graduation speech: "I'd like to *thanks* Google, Wikipedia and my computer! *:D* #iThingteens

Text data

Which kind of data we face?

Authorities are *only too aware* that Kashgar is 4,000 kilometres (2,500 miles) from Beijing but *only* a tenth of the distance from the Pakistani border, and are *desperate* to *ensure instability or militancy* does not leak over the frontiers.
Taiwan-made products *stood a good chance* of becoming *even more competitive thanks to* wider access to overseas markets and lower costs for material imports, he said.
"March *appears* to be a *more reasonable* estimate while earlier admission *cannot be entirely ruled out*," according to Chen, also Taiwan's chief WTO negotiator.
friday evening plans were great, but saturday's plans *didnt go as expected* – i went dancing & it was an *ok* club, but *terribly crowded :-(*
WHY THE *HELL* DO YOU GUYS ALL HAVE MRS. KENNEDY! SHES A FUCKING DOUCHE
AT&T was *okay* but whenever they do something *nice* in the name of customer service it seems like a favor, while T-Mobile makes that a *normal everyday thin*
obama should be *impeached* on *TREASON* charges. Our Nuclear arsenal was TOP Secret. Till HE told our enemies what we had. *#Coward #Traitor*
My graduation speech: "I'd like to *thanks* Google, Wikipedia and my computer! *:D* #iThingteens

Text data

Financial data

Which kind of data we face?

Authorities are *only too aware* that Kashgar is 4,000 kilometres (2,500 miles) from Beijing but *only* a tenth of the distance from the Pakistani border, and are *desperate* to *ensure instability or militancy* does not leak over the frontiers.
Taiwan-made products *stood a good chance* of becoming *even more competitive thanks to* wider access to overseas markets and lower costs for material imports, he said.
"March *appears* to be a *more reasonable* estimate while earlier admission *cannot be entirely ruled out*," according to Chen, also Taiwan's chief WTO negotiator.
friday evening plans were great, but saturday's plans *didnt go as expected* – i went dancing & it was an *ok* club, but *terribly crowded :-(*
WHY THE *HELL* DO YOU GUYS ALL HAVE MRS. KENNEDY! SHES A FUCKING DOUCHE
AT&T was *okay* but whenever they do something *nice* in the name of customer service it seems like a favor, while T-Mobile makes that a *normal everyday thin*
obama should be *impeached* on *TREASON* charges. Our Nuclear arsenal was TOP Secret. Till HE told our enemies what we had. *#Coward #Traitor*
My graduation speech: "I'd like to *thanks* Google, Wikipedia and my computer! *:D* #iThingteens

Text data

Healthy data

Financial data

Which kind of data we face?

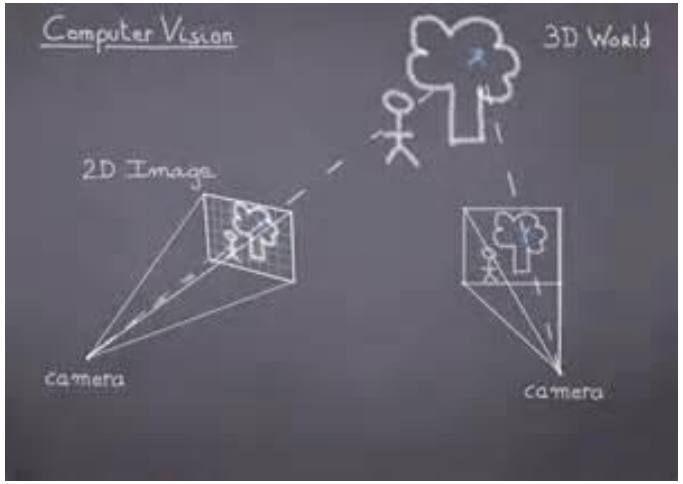
Authorities are *only too aware* that Kashgar is 4,000 kilometres (2,500 miles) from Beijing but *only* a tenth of the distance from the Pakistani border, and are *desperate* to *ensure instability or militancy* does not leak over the frontiers.
Taiwan-made products *stood a good chance* of becoming *even more competitive thanks to* wider access to overseas markets and lower costs for material imports, he said.
"March *appears* to be a *more reasonable* estimate while earlier admission *cannot be entirely ruled out*," according to Chen, also Taiwan's chief WTO negotiator.
friday evening plans were great, but saturday's plans *didnt go as expected* – i went dancing & it was an *ok* club, but *terribly crowded :-(*
WHY THE *HELL* DO YOU GUYS ALL HAVE MRS. KENNEDY! SHES A FUCKING DOUCHE
AT&T was *okay* but whenever they do something *nice* in the name of customer service it seems like a favor, while T-Mobile makes that a *normal everyday thin*
obama should be *impeached* on *TREASON* charges. Our Nuclear arsenal was TOP Secret. Till HE told our enemies what we had. *#Coward #Traitor*
My graduation speech: "I'd like to *thanks* Google, Wikipedia and my computer! *:D* #iThingteens

Text data

Healthy data

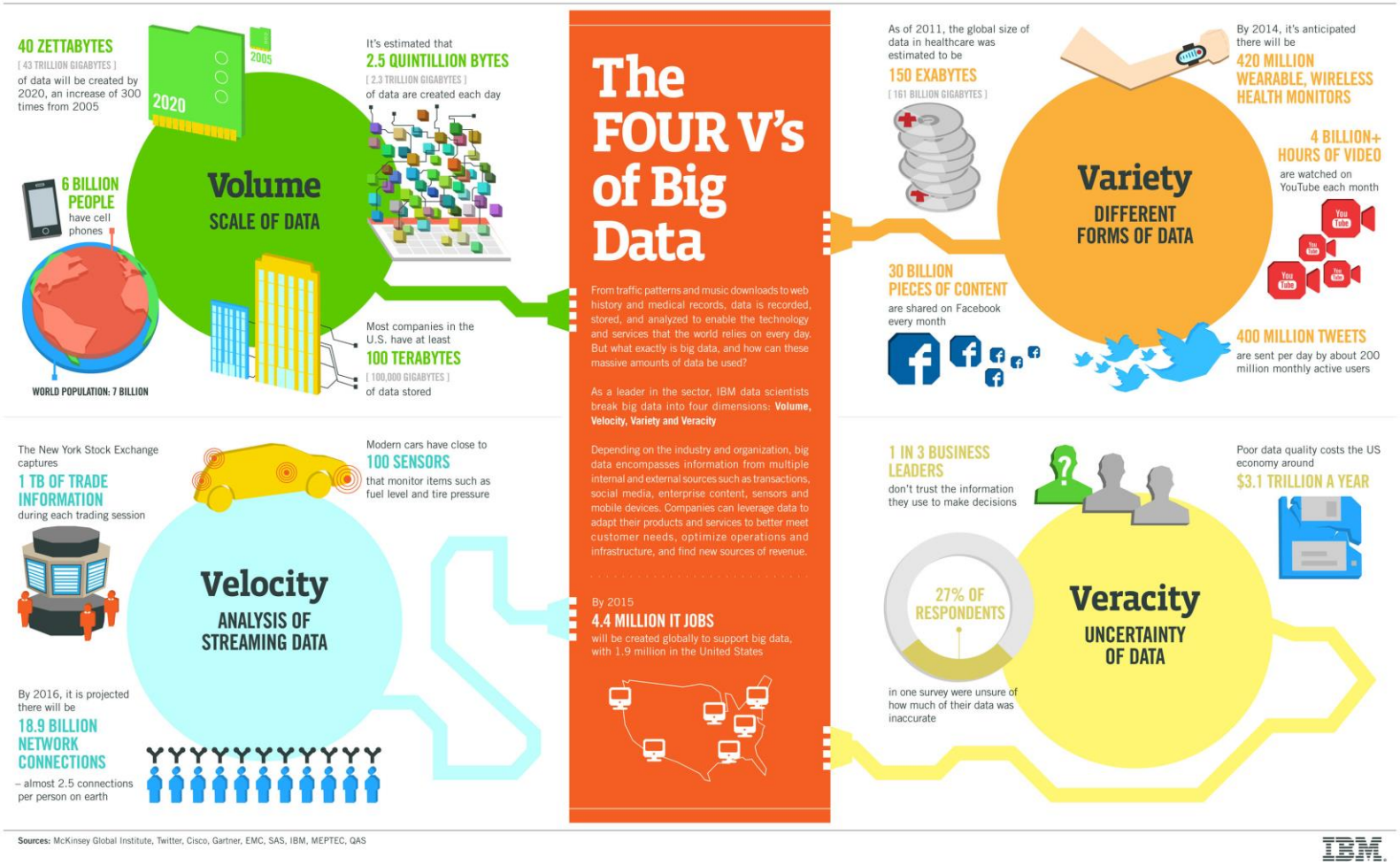
Data

Financial data



Vision data

Why focus on data representation?



- 4Vs in data science
  a) Volume
  b) Variety
  c) Velocity
  d) Veracity

- Key of the success
  - Efficient and effective data representation

Photo source: IBM

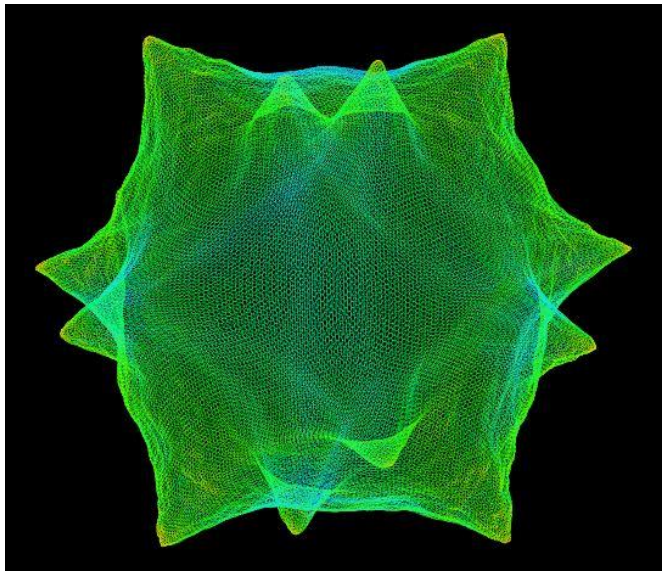
Data representation

- Main difficulty: curse of dimensionality (e.g. Hughes phenomenon)

- Sparsity prior
  - Regularity
  - Temporal and spatial coherence
  - Hierarchical organization of explanatory factors
  - **Dictionary learning**

→ Non-convex models

- Goal: “good” sparse representation
  - Efficient and convergent numerical schemes and effective models



Sparse matrices

Photo source: Gallery of the University of Florida

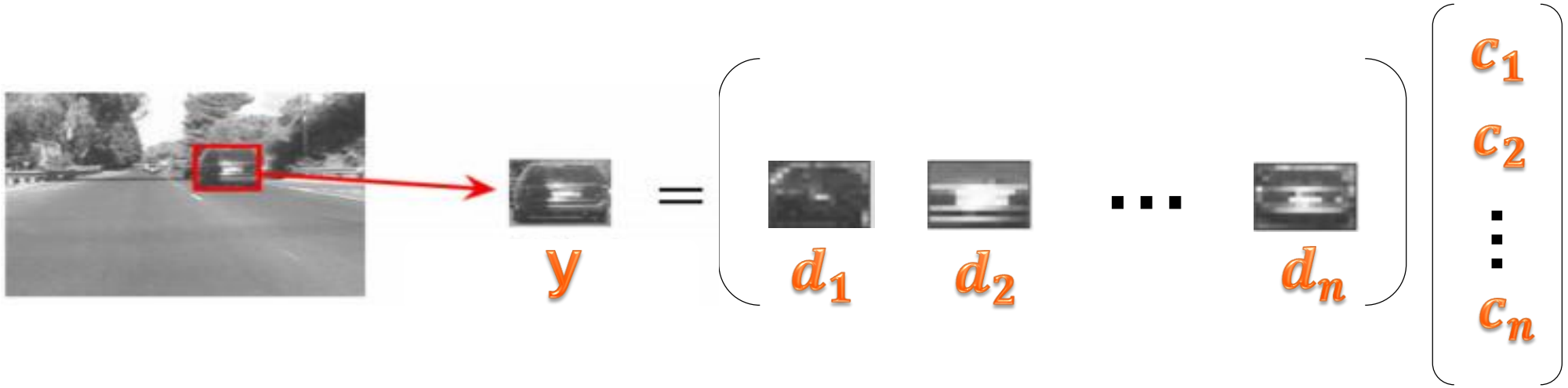
Sparsity based dictionary learning

- Initially proposed in [Olshausen et al, Nature, 1996]

- Many variants
  - Redundant dictionary learning [Elad et al, TIP, 2006]
  - Data-driven tight frame construction [Cai et al, ACHA, 2014]
  - Task-driven dictionary learning [Mairal et al, TPAMI, 2012]
  - Hierarchical dictionary learning [Jenatton et al, JMLR, 2011]

- Success in applications
  - Image denoising [Vincent et al, JMLR, 2010]
  - Abnormal event detection [Zhao et al, CVPR 2011]
  - Natural language processing [Bagnell et al, NIPS, 2009]
  - Visual tracking [Mei et al, TPAMI, 2011]

Dictionary learning

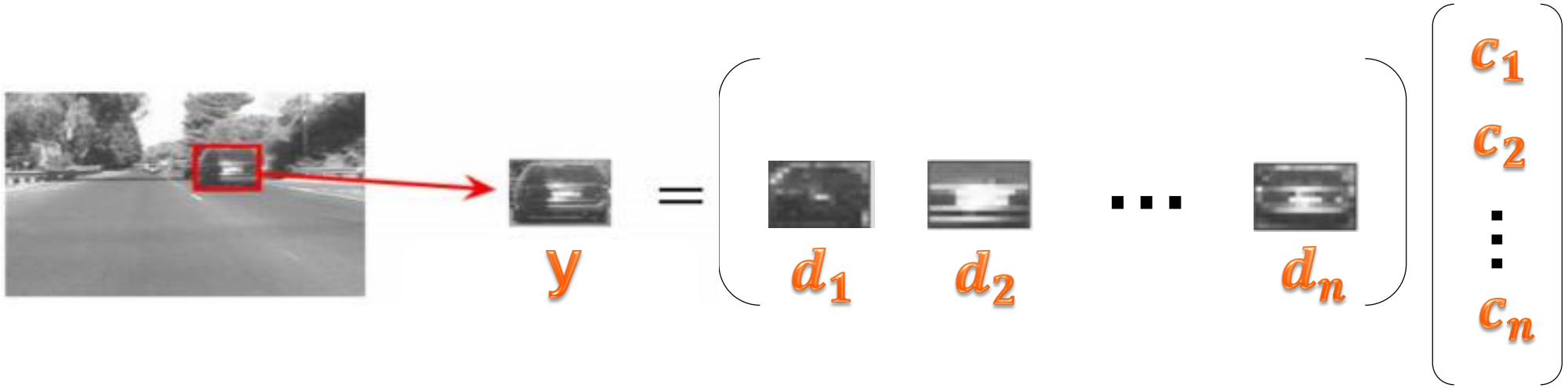


- Data-driven sparse representation
  - Dictionary learning: find an adaptive dictionary $D = (d_1, d_2, \ldots, d_n)$ such that

$$\boldsymbol{y = \sum_i c_i d_i},$$

and **most c_is are zero**.

Dictionary learning



- Data-driven sparse representation
  - Dictionary learning: find an adaptive dictionary $D = (d_1, d_2, \ldots, d_n)$ such that

$$y = \sum_i c_i d_i,$$

and **most c_is are zero**

How to solve the non-convex model?

The most popular dictionary learning method

- Data $Y = (y_1, y_2 \dots y_p)$ and $D = (d_1, d_2, \dots, d_m)$ and $C = (c_1, c_2, \dots, c_p)$

- K-SVD [Elad et al, TIP, 2006]

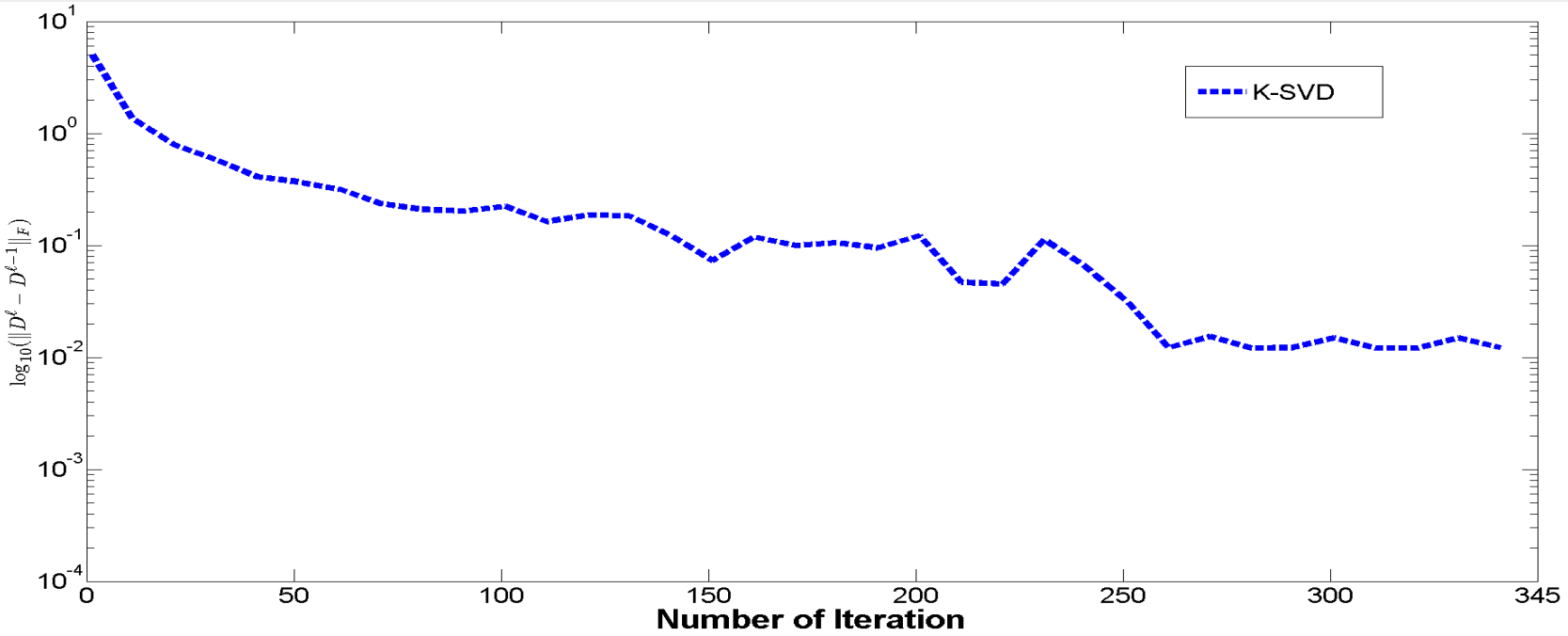
$$\min_{D,C} ||Y - DC||_F^2 , s.t. ||c_i||_0 \leq k, ||d_j||_2 = 1, \forall i \in [p], \forall j \in [m].$$

  - Alternating minimization: OMP and sequential SVD (No convergence guarantee and high computational complexity)

- K-SVD based applications
  - Image classification [Jiang et al, TPAMI, 2013]
  - Text corpora representation [Jenatton et al, JMLR, 2011]
  - Multi-task learning [Ruvolo et al, ICML, 2014]
  - Recommendation system [Gediminas et al, TKDE, 2012]

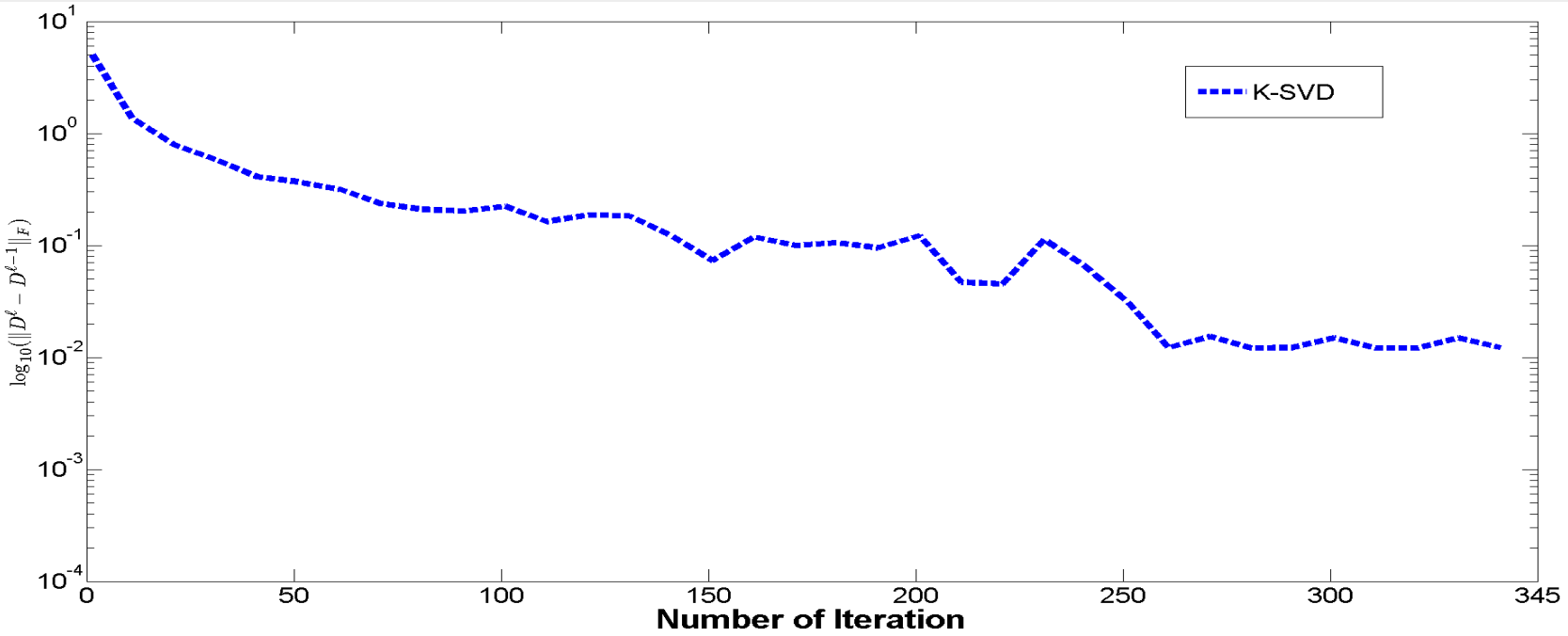
The numerical behavior of K-SVD

L2 norm of increments of the sequence generated by K-SVD



The numerical behavior of K-SVD

L2 norm of increments of the sequence generated by K-SVD



Goal: convergent and efficient numerical algorithm.

Mathematical formulation

- General dictionary learning model

$$\min_{D,C} \sum_{i=1}^{p} f(y_i, Dc_i) + r_1(C) + r_2(D)$$

  - f is the loss function, e.g. ℓ_2 loss,
  - r_1, r_2 are constraint functions on C, D, respectively.

Mathematical formulation

- General dictionary learning model

$$\min_{D,C} \underbrace{\sum_{i=1}^{p} f(y_i, Dc_i)}_{f} + \underbrace{r_1(C) + r_2(D)}_{\text{r}}$$

  - f is the loss function, e.g. ℓ_2 loss,
  - r_1, r_2 are constraint functions on C, D, respectively.

- Multi-block non-convex optimization

$$\min_{x=(x_1,\ldots,x_n)} H(x) \coloneqq f(x) + \sum_{i=1}^{n} r_i(x_i)$$

- Assumptions
  1. ∇f is Lipschitz on any bounded set and $\nabla_i f$ is L_i-Lipschitz
  2. $r_i, i \in [n]$ are proper and lower semi-continuous

Existing schemes

- Define $f_i^k(\cdot) \coloneqq f(x_1^{k+1}, \dots, x_{i-1}^{k+1}, \cdot, x_{i+1}^k, \dots, x_n^k)$

Alternating minimization

$$x_i^{k+1} \in \operatorname{argmin} f_i^k(x_i) + r_i(x_i)$$

Proximal alternating minimization

$$x_i^{k+1} \in \operatorname{argmin} f_i^k(x_i) + r_i(x_i) + \lambda_i ||x_i - x_i^k||_F^2$$

Proximal alternating linearized minimization

$$x_i^{k+1} \in \operatorname{argmin} \langle \nabla f_i^k(x_i^k), x_i - x_i^k \rangle + r_i(x_i) + \lambda_i ||x_i - x_i^k||_F^2$$

Literature review

- Alternating minimization (AM)
  - No convergence guarantee
- Proximal alternating minimization (PAM)
  - Global convergence property [Attouch et al, MOR, 2010]
- Proximal alternating linearized minimization (PALM)
  - Global convergence property [Bolte et al, MP, 2014]
- Hybrid method
  - Multi-convex case [Xu et al, SIIMS, 2013]

Literature review

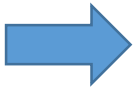
- Alternating minimization (AM)
  - No convergence guarantee
- Proximal alternating minimization (PAM)
  - Global convergence property [Attouch et al, MOR, 2010]
- Proximal alternating linearized minimization (PALM)
  - Global convergence property [Bolte et al, MP, 2014]
- Hybrid method
  - Multi-convex case [Xu et al, SIIMS, 2013]

Inner iter. No. : AM ≈ PAM > PALM
Outer iter. No. : AM < PAM < PALM

Literature review

- Alternating minimization (AM)
  - No convergence guarantee
- Proximal alternating minimization (PAM)
  - Global convergence property [Attouch et al, MOR, 2010]
- Proximal alternating linearized minimization (PALM)
  - Global convergence property [Bolte et al, MP, 2014]
- Hybrid method
  - Multi-convex case [Xu et al, SIIMS, 2013]

Inner iter. No. : AM ≈ PAM > PALM
Outer iter. No. : AM < PAM < PALM



Hybrid them for non-convex case.

Hybrid proximal alternating method

- HPAM [Bao et al, TPAMI, 2016]

  For $k = 1,2, \dots,$

  for $i = 1,2, \dots, n$

$$x_i^{k+1} \in \begin{cases} \operatorname{argmin} f_i^k(x_i) + r_i(x_i) + \lambda_i ||x_i - x_i^k||_F^2 \text{ or} & \text{PAM} \\ \operatorname{argmin} \langle \nabla f_i^k(x_i^k), x_i - x_i^k \rangle + r_i(x_i) + \lambda_i ||x_i - x_i^k||_F^2 & \text{PALM} \end{cases}$$

  where λ_i is the appropriate step size.

Hybrid proximal alternating method

- HPAM [Bao et al, TPAMI, 2016]

  For $k = 1,2, \dots,$

  for $i = 1,2, \dots, n$

$$x_i^{k+1} \in \begin{cases} \operatorname{argmin} f_i^k(x_i) + r_i(x_i) + \lambda_i ||x_i - x_i^k||_F^2 \text{ or} & \text{PAM} \\ \operatorname{argmin} \langle \nabla f_i^k(x_i^k), x_i - x_i^k \rangle + r_i(x_i) + \lambda_i ||x_i - x_i^k||_F^2 & \text{PALM} \end{cases}$$

  where λ_i is the appropriate step size.

Theorem (Global convergence): Let $\{x^k\}$ be the infinite sequence generated by the HPAM. If H is a KL function and $\{x^k\}$ is bounded. Then, the sequence $\{x^k\}$ converges to a point x^*, which is a critical point of H, i.e. $0 \in \partial H(x^*)$.

Remark: AM can be included under certain condition.

Sketch of the proof

- The proof is based on the result from [Attouch et al, MP, 2013].
- Four main steps:
  1. Sufficient decrease property

$$H\left(x^k\right) - H\left(x^{k+1}\right) \geq \rho_1 ||x^k - x^{k+1}||_F^2, \text{for some } \rho_1 > 0$$

  2. Bounded the subgradient:

$$\text{dist}\left(0, \partial H\left(x^k\right)\right) \leq \rho_2 ||x^k - x^{k-1}||, \text{for some } \rho_2 > 0$$

  3. Subsequence continuity: ∃ a subsequence $\{x^{k_j}\}$ such that

$$x^{k_j} \to \bar{x}, \qquad \text{and} \qquad H(x^{k_j}) \to H(\bar{x}).$$

  4. H satisfies Kurdyka-Lojasiwicz (KL) property.

HPAM in dictionary learning

- Plain dictionary learning model

$$\min_{D,C} ||Y - DC^{\mathsf{T}}||_F^2 + \lambda||C||_0, s.t. ||C||_\infty \leq M, ||d_j||_2 = 1, \forall j \in [m].$$

- Numerical schemes
  - C1 $(x_1, x_2, \dots, x_n) = (c_1, c_2, \dots, c_q, d_1, d_2, \dots, d_q)$
  - C2 $(x_1, x_2, \dots, x_n) = (C, D)$
  - C3 $(x_1, x_2, \dots, x_n) = (C, d_1, c_1, d_2, c_2 \dots, d_q, c_q)$

HPAM in dictionary learning

- Plain dictionary learning model

$$\min_{D,C} ||Y - DC^{\top}||_F^2 + \lambda||C||_0, s.t. ||C||_\infty \leq M, ||d_j||_2 = 1, \forall j \in [m].$$

- Numerical schemes
  - C1 $(x_1, x_2, \dots, x_n) = (c_1, c_2, \dots, c_q, d_1, d_2, \dots, d_q)$
  - C2 $(x_1, x_2, \dots, x_n) = (C, D)$
  - C3 $(x_1, x_2, \dots, x_n) = (C, d_1, c_1, d_2, c_2 \dots, d_q, c_q)$

SCHEMES	S1	S2	S3
Block choice	C1	C2	C3
Step I	PAM	PALM	PALM
Step II	PAM	PALM	PAM+AM

HPAM in dictionary learning

- Plain dictionary learning model

$$\min_{D,C} ||Y - DC^{\top}||_F^2 + \lambda||C||_0, s.t. ||C||_\infty \leq M, ||d_j||_2 = 1, \forall j \in [m].$$

- Numerical schemes
  - C1 $(x_1, x_2, \dots, x_n) = (c_1, c_2, \dots, c_q, d_1, d_2, \dots, d_q)$
  - C2 $(x_1, x_2, \dots, x_n) = (C, D)$
  - C3 $(x_1, x_2, \dots, x_n) = (C, d_1, c_1, d_2, c_2 \dots, d_q, c_q)$

SCHEMES	S1	S2	S3
Block choice	C1	C2	C3
Step I	PAM	PALM	PALM
Step II	PAM	PALM	PAM+AM

All the above schemes converge to a critical point.

Scheme 1 (S1)

Block choice: C1 $(x_1, x_2, \dots, x_n) = (c_1, c_2, \dots, c_q, d_1, d_2, \dots, d_q)$

Step 1. Sparse coding: for $i = 1, \dots, m$,

$$c_i^k \in \underset{||c||_\infty \leq M}{\operatorname{argmin}} \lambda ||c||_0 + ||J_i^k - d_i^k c||_F^2 + \lambda_c ||c - c_i^{k-1}||_F^2$$

where $J_i^k = Y - \sum_{j<i} d_j^k c_j^{k\mathsf{T}} + \sum_{j>i} d_j^k c_j^{k-1\mathsf{T}}$.

Step 2. Dictionary update: for $i = 1, \dots, m$,

$$d_i^{k+1} \in \underset{d}{\operatorname{argmin}} ||E_i^k - d c_i^{k\mathsf{T}}||_F^2 + \lambda_d ||d - d_i^k||_F^2 \; s.t. ||d|| = 1$$

where $E_i^k = Y - \sum_{j<i} d_j^{k+1} c_j^{k\mathsf{T}} + \sum_{j>i} d_j^k c_j^{k\mathsf{T}}$.

Scheme 2 (S2)

Block choice: C2 $(x_1, x_2, \dots, x_n) = (C, d_1, d_2, \dots, d_q)$

Step 1. Sparse coding:

$$C^k \in \underset{||C||_\infty \leq M}{\operatorname{argmin}} \lambda||C||_0 + \left\langle \nabla_C f\left(D^k, C^{k-1}\right), C - C^{k-1} \right\rangle + \lambda_C ||C - C^{k-1}||_F^2$$

Step 2. Dictionary update: for $i = 1, \dots, m$,

$$d_i^{k+1} \in \underset{d}{\operatorname{argmin}} ||E_i^k - d c_i^{k\mathsf{T}}||_F^2 + \lambda_d ||d - d_i^k||_F^2 \; s.t. ||d|| = 1$$

where $E_i^k = Y - \sum_{j<i} d_j^{k+1} c_j^{k\mathsf{T}} + \sum_{j>i} d_j^k c_j^{k\mathsf{T}}$.

Scheme 3 (S3)

Step 1. Sparse coding:

$$C^k \in \underset{||C||_\infty \le M}{\operatorname{argmin}} \lambda||C||_0 + \left\langle \nabla_C f\left(D^k, C^{k-1}\right), C - C^{k-1} \right\rangle + \lambda_C ||C - C^{k-1}||_F^2$$

Step 2. Dictionary update: for $i = 1, \ldots, m$,

1. Update the d_i

$$d_i^{k+1} \in \operatorname{argmin} ||E_i^k - d c_i^{k\top}||_F^2 + \lambda_d ||d - d_i^k||_F^2 \; s.t. ||d|| = 1$$

where $E_i^k \overset{d}{=} Y - \sum_{j<i} d_j^{k+1} c_j^{k\top} + \sum_{j>i} d_j^k c_j^{k\top}$.

2. Reupdate the non-zero coefficients c_i

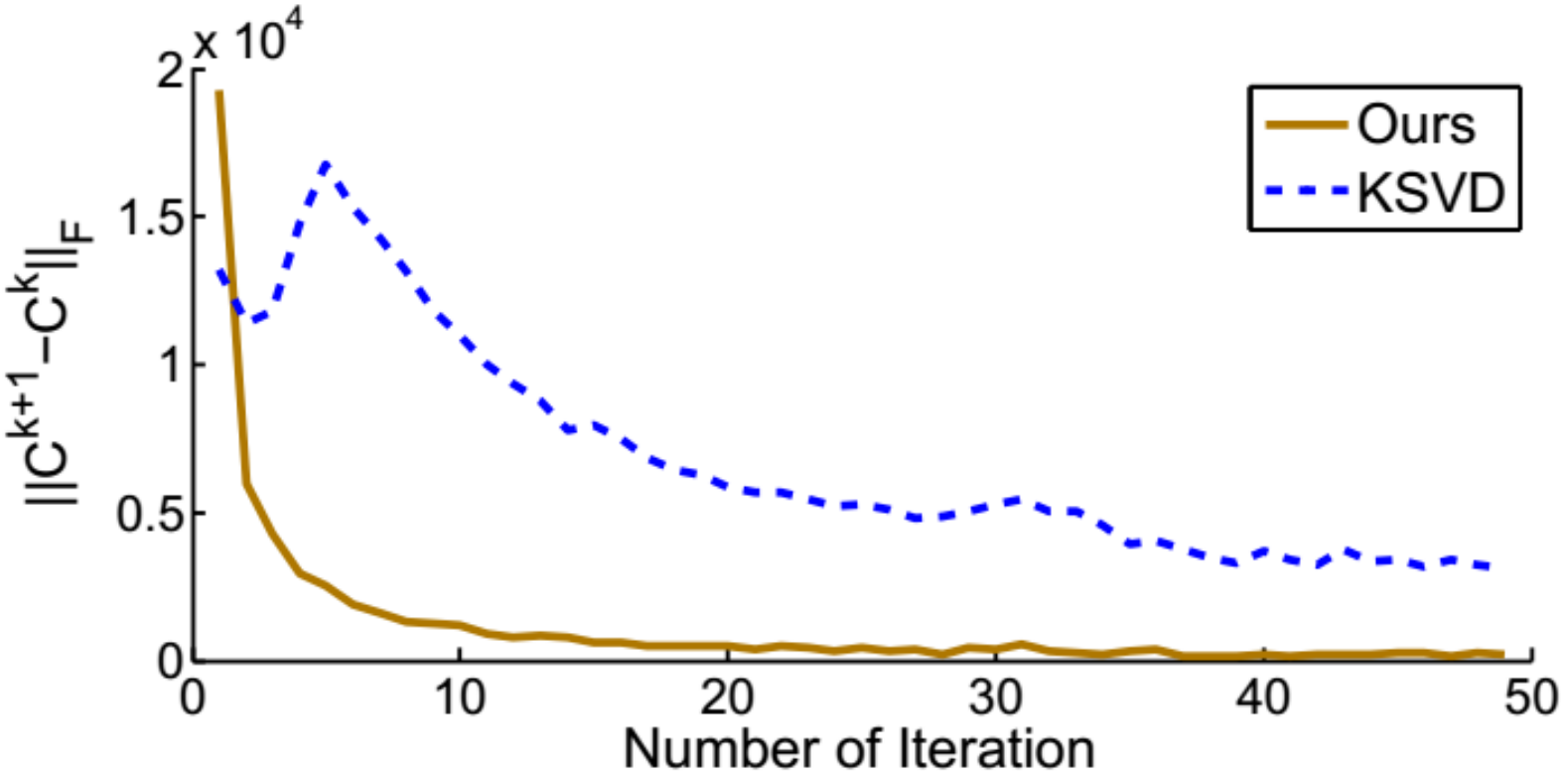
Further decrease the residual.

$$c_i^{k+1} \in \operatorname{argmin} ||E_i^k - d_i^{k+1} c||_F^2 \; s.t. \; c(j) = 0, \forall j \in I_j$$

where $I_j \overset{c}{=} \{q : c_i^k(q) = 0\}$.

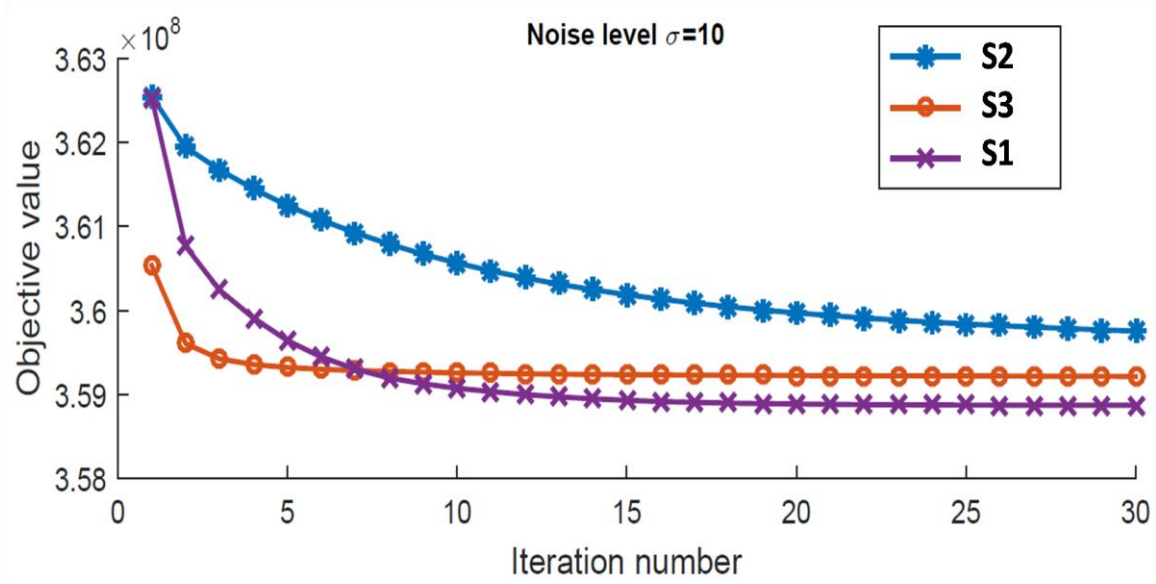
Numerical behavior

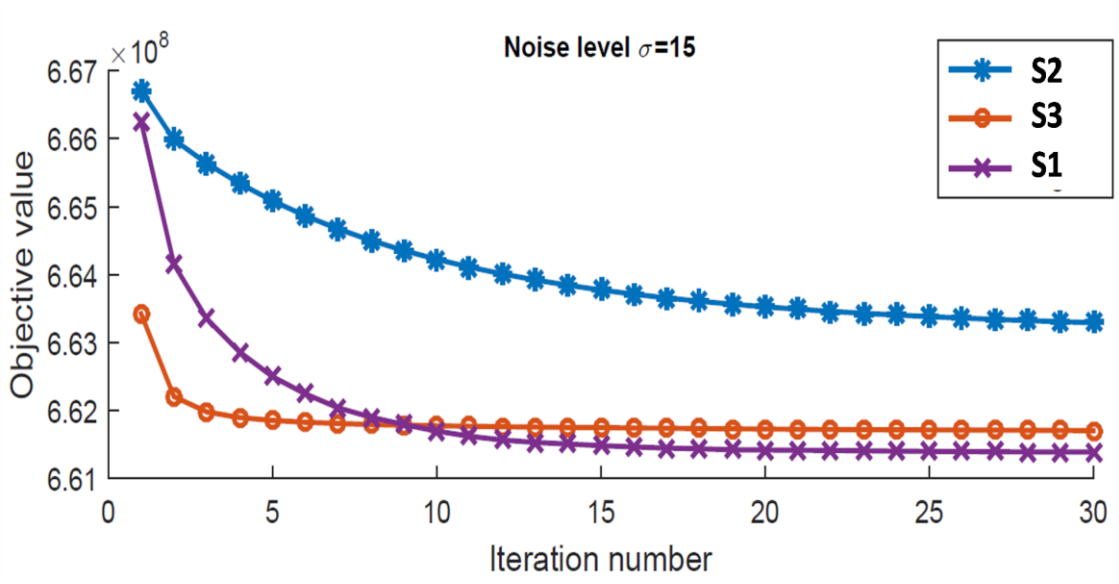
Increments of the sequence $\{C^k\}$



Numerical behavior

Objective value versus iteration





Different convergent algorithms achieve different critical points in DL.

Computational efficiency

Training time versus the atom dimensions

Atom Dim.	6*6	8*8	10*10	12*12	14*14	16*16
K-SVD	39	70	114	164	228	308
S1	71	217	465	1011	1848	3094
S2	9	16	28	42	60	86
S3	10	18	30	45	66	96

Efficiency: S2 ≈ S3 > K-SVD > S1

Image denoising

Image		**Fingerprint**			**Image**		**Lena**		
Noise	5	10	15	20	Noise	5	10	15	20
K-SVD	36.59	32.39	30.06	28.47	K-SVD	38.59	35.47	33.70	32.38
S1	36.58	32.27	29.87	28.24	S1	38.48	35.37	33.55	32.21
S2	36.50	32.21	29.84	28.18	S2	38.46	35.35	33.50	32.15
S3	36.59	32.35	30.03	28.44	S3	38.49	35.41	33.57	32.25

Performance: S3 ≈ S1 ≈ K-SVD > S2

Image denoising

Image		**Fingerprint**			**Image**		**Lena**		
Noise	5	10	15	20	Noise	5	10	15	20
K-SVD	36.59	32.39	30.06	28.47	K-SVD	38.59	35.47	33.70	32.38
S1	36.58	32.27	29.87	28.24	S1	38.48	35.37	33.55	32.21
S2	36.50	32.21	29.84	28.18	S2	38.46	35.35	33.50	32.15
S3	36.59	32.35	30.03	28.44	S3	38.49	35.41	33.57	32.25

Performance: S3 ≈ S1 ≈ K-SVD > S2

S3 is the most appropriate scheme.

Nonlinear dictionary learning

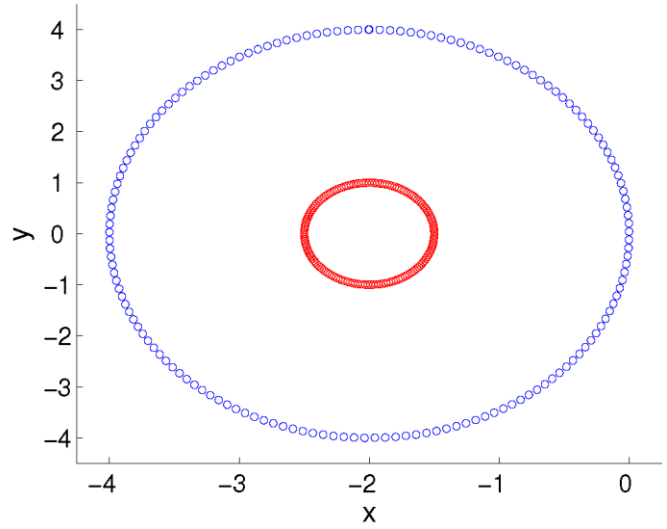
- Capturing the nonlinear data patterns

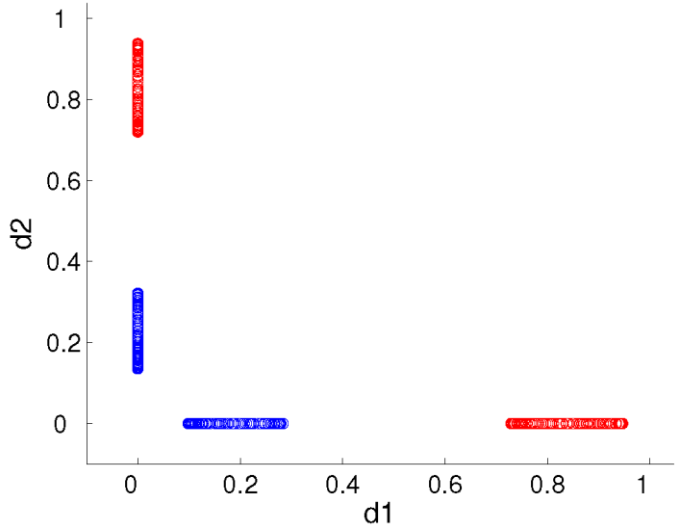
- Equiangular kernel learning [Quan et al, CVPR, 2016]

$$\min_{D,C} ||\Phi(Y) - \Phi(D)C||_F^2,.$$

$$s.t. ||c_i||_0 \leq k, \forall i, D^\top D = I$$

where Φ is a map associated with the kernel K.





Nonlinear dictionary learning

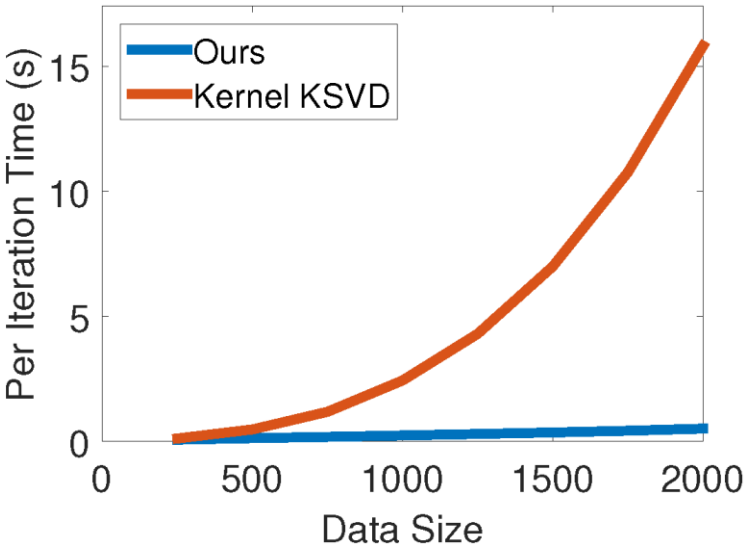
- Capturing the nonlinear data patterns

- Equiangular kernel learning [Quan et al, CVPR, 2016]

$$\min_{D,C} ||\Phi(Y) - \Phi(D)C||_F^2,.$$

$$s.t. ||c_i||_0 \leq k, \forall i, D^\top D = I$$

where Φ is a map associated with the kernel K.



- Properties:
  1. $\langle \Phi(d_m), \Phi(d_n) \rangle = \mu_0, \forall m \neq n$ if $K = \psi(||d_m - d_n||^2)$.
  2. More scalable than kernel K-SVD (free of $K(Y, Y)$)

Dynamic texture classification

- Recognizing the moving textures with certain stationary temporal changes
- Nonlinear data patterns

Fountain

Flag

Candle

Experiments

- UCLA-DT
  - 50 categories, 200 videos
- DynTex
  - 10 categories, 275 videos
- DynTex++
  - 36 categories, 3600 videos

Sample video

Dataset	DFS	DFS+	LBP-TOP	KGDL	OTF	Ours
UCLA	97.5	97.5	N.A.	N.A.	97.2	**98.6**
DynTex	74.5	74.8	72.0	75.1	73.5	**75.6**
DynTex++	89.9	91.7	89.2	92.8	89.8	**93.4**

Classification accuracy (%)

Discussion

- The hybrid proximal alternating scheme has certain convergence guarantee for multi-block nonconvex problems

- Application based alternating scheme can lead to more efficient and effective algorithm

Future work:

1. Find better initialization
2. Connection between hierarchical dictionary learning and convolutional neural networks

Thank you!