The Sample Complexity of Dictionary Learning

Daniel Vainsencher[1]
Joint work with Shie Mannor[1] and Alfred M. Bruckstein[2]

[1]Department of Electrical Engineering

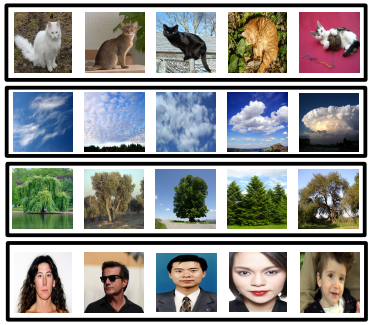
[2]Department of Computer Science
Technion, Israel

COLT 2011

Common ground

JPEG, JPEG 2000 use fixed bases (DCT, wavelets) for everything.

Varied but recurring classes of media suggest more adaptive solutions.

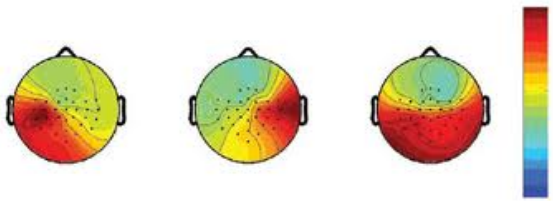


Applications

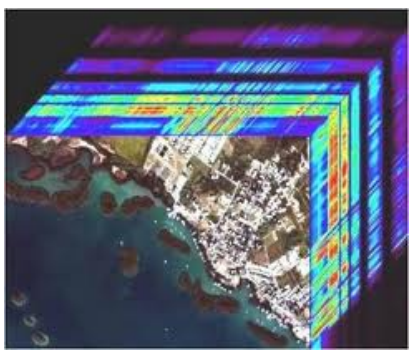
- Compression
- Restoration
- Acquisition

Not just images

- My phone should know and denoise *my* voice
- What is the right representation for EEG signals?



- For a hyperspectral pixel?



Problem definition

- Given $\mathbb{S}^{n-1} \supset \{x_i\}_{i=1}^m \sim \nu$
- For any $D \in \left(\mathbb{S}^{n-1}\right)^p$, denote $h_{A,D}(x) = \min_{a \in A} \|Da - x\|_2$
- A enforces "simplicity":
  - $R_\lambda = \{a : \|a\|_1 \leq \lambda\}$
  - $H_k = \{a : \|a\|_0 \leq k\}$ $(\|a\|_0 = \Sigma_{a_i \neq 0} 1)$
- Let $l : \mathbb{R}^+ \to \mathbb{R}^+$ be a loss function

Goal: learn $D \in \left(\mathbb{S}^{n-1}\right)^p$ minimizing $\mathbb{E}_{x \sim \nu} l(h_{A,D})$

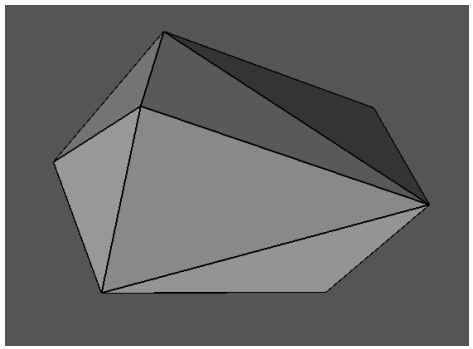
Problem definition

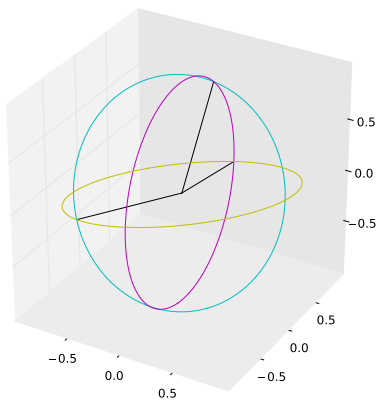
- Given $\mathbb{S}^{n-1} \supset \{x_i\}_{i=1}^m \sim \nu$
- For any $D \in \left(\mathbb{S}^{n-1}\right)^p$, denote $h_{A,D}(x) = \min_{a \in A} \|Da - x\|_2$
- A enforces "simplicity":
  - $R_\lambda = \{a : \|a\|_1 \leq \lambda\}$
  - $H_k = \{a : \|a\|_0 \leq k\}$ $(\|a\|_0 = \sum_{a_i \neq 0} 1)$
- Let $l : \mathbb{R}^+ \to \mathbb{R}^+$ be a loss function

Goal: learn $D \in \left(\mathbb{S}^{n-1}\right)^p$ minimizing $\mathbb{E}_{x \sim \nu} l(h_{A,D})$

Problem definition

- Given $\mathbb{S}^{n-1} \supset \{x_i\}_{i=1}^m \sim \nu$
- For any $D \in \left(\mathbb{S}^{n-1}\right)^p$, denote $h_{A,D}(x) = \min_{a \in A} \|Da - x\|_2$
- A enforces "simplicity":
  - $R_\lambda = \{a : \|a\|_1 \leq \lambda\}$
  - $H_k = \{a : \|a\|_0 \leq k\}$ $(\|a\|_0 = \sum_{a_i \neq 0} 1)$
- Let $l : \mathbb{R}^+ \to \mathbb{R}^+$ be a loss function

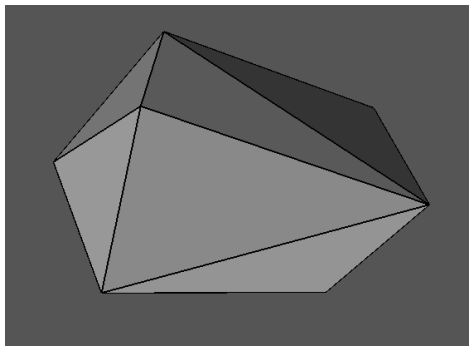


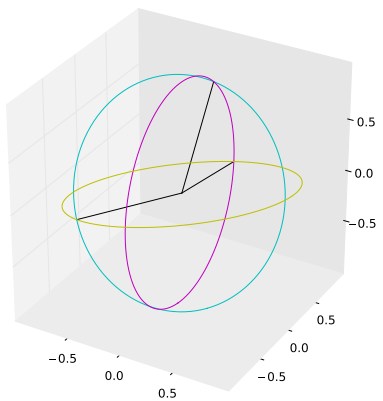


Goal: learn $D \in \left(\mathbb{S}^{n-1}\right)^p$ minimizing $\mathbb{E}_{x \sim \nu} l(h_{A,D})$

Problem definition

- Given $\mathbb{S}^{n-1} \supset \{x_i\}_{i=1}^m \sim \nu$
- For any $D \in \left(\mathbb{S}^{n-1}\right)^p$, denote $h_{A,D}(x) = \min_{a \in A} \|Da - x\|_2$
- A enforces "simplicity":
  - $R_\lambda = \{a : \|a\|_1 \leq \lambda\}$
  - $H_k = \{a : \|a\|_0 \leq k\}$ $(\|a\|_0 = \sum_{a_i \neq 0} 1)$
- Let $l : \mathbb{R}^+ \to \mathbb{R}^+$ be a loss function

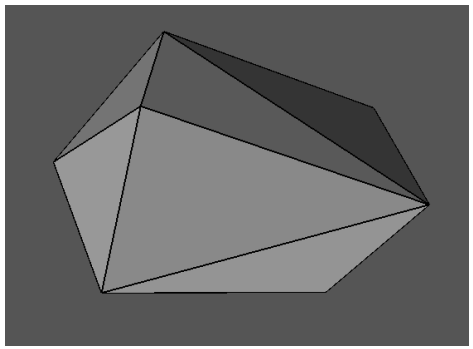


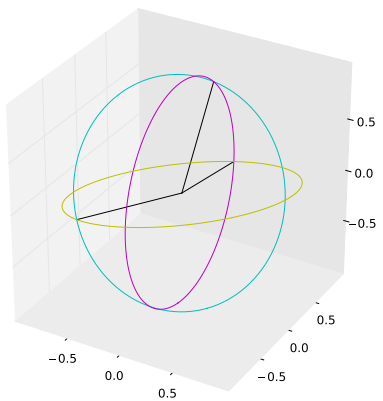


Goal: learn $D \in \left(\mathbb{S}^{n-1}\right)^p$ minimizing $\mathbb{E}_{x \sim \nu} l(h_{A,D})$

Problem definition

- Given $\mathbb{S}^{n-1} \supset \{x_i\}_{i=1}^m \sim \nu$
- For any $D \in \left(\mathbb{S}^{n-1}\right)^p$, denote $h_{A,D}(x) = \min_{a \in A} \|Da - x\|_2$
- A enforces "simplicity":
  - $R_\lambda = \{a : \|a\|_1 \leq \lambda\}$
  - $H_k = \{a : \|a\|_0 \leq k\}$ $(\|a\|_0 = \sum_{a_i \neq 0} 1)$
- Let $l : \mathbb{R}^+ \to \mathbb{R}^+$ be a loss function





Goal: learn $D \in \left(\mathbb{S}^{n-1}\right)^p$ minimizing $\mathbb{E}_{x \sim \nu} l(h_{A,D})$

Parameters in two problem domains

- Image restoration
- Denoising EEG (for a brain computer interface)

- $n = 64$ ($8x8$ pixel "patches" of images, or 64 EEG electrodes)
- $p = 100$ elements in the dictionary - overcomplete case
- $k = 10$
- $m \approx 3 \cdot 10^5$ (for EEG)

Parameters in two problem domains

- Image restoration
- Denoising EEG (for a brain computer interface)

- $n = 64$ ($8x8$ pixel "patches" of images, or 64 EEG electrodes)
- $p = 100$ elements in the dictionary - overcomplete case
- $k = 10$
- $m \approx 3 \cdot 10^5$ (for EEG)

Applications revisited

- Lossy compression
- Denoising: small non sparse noise is automatically removed by finding a sparse representation
- Compressed sensing: $\propto k \log n$ random measurements can be sufficient to allow reconstruction

Difficulties

- $A = H_1$ generalizes K means clustering (for $K = p$)
- Optimization
  - NP-hard subproblem in H_k: given x, D, find $\min_{a \in H_k} \|Da - x\|$
    - Solution: approximable when D is "almost orthogonal"
  - Space of dictionaries $\left(\mathbb{S}^{n-1}\right)^p$ is high dimensional (6300 in example)
  - The objective is non convex
- Generalization
  - Access to ν through samples only

Difficulties

- $A = H_1$ generalizes K means clustering (for $K = p$)
- Optimization
  - NP-hard subproblem in H_k: given x, D, find $\min_{a \in H_k} \|Da - x\|$
    - Solution: approximable when D is "almost orthogonal"
  - Space of dictionaries $\left(\mathbb{S}^{n-1}\right)^p$ is high dimensional (6300 in example)
  - The objective is non convex
- Generalization
  - Access to ν through samples only

Difficulties

- $A = H_1$ generalizes K means clustering (for $K = p$)
- Optimization
  - NP-hard subproblem in H_k: given x, D, find $\min_{a \in H_k} \|Da - x\|$
    - Solution: approximable when D is "almost orthogonal"
  - Space of dictionaries $\left(\mathbb{S}^{n-1}\right)^p$ is high dimensional (6300 in example)
  - The objective is non convex
- Generalization
  - Access to v through samples only

Difficulties

- $A = H_1$ generalizes K means clustering (for $K = p$)
- Optimization
  - NP-hard subproblem in H_k: given x, D, find $\min_{a \in H_k} \|Da - x\|$
    - Solution: approximable when D is "almost orthogonal"
  - Space of dictionaries $\left(\mathbb{S}^{n-1}\right)^p$ is high dimensional (6300 in example)
  - The objective is non convex
- Generalization
  - Access to ν through samples only

State of optimization problem

- Empirical algorithms from 1996 to 2010:
  - EM algorithms
  - Alternating optimization of D and a
  - Online optimization
- ... no global convergence/approximation guarantees
- Uniform convergence results needed

State of optimization problem

- Empirical algorithms from 1996 to 2010:
  - EM algorithms
  - Alternating optimization of D and a
  - Online optimization
- ... no global convergence/approximation guarantees
- Uniform convergence results needed

State of optimization problem

- Empirical algorithms from 1996 to 2010:
  - EM algorithms
  - Alternating optimization of D and a
  - Online optimization
- ... no global convergence/approximation guarantees
- Uniform convergence results needed

Generalization problem

- Fix the dimension n, dictionary size p, sample size m and coefficient set A
- Denote ν_m the empirical distribution for $\{x_i\}_{i=1}^m \sim \nu$
- Denote

$$E(D) = \mathbb{E}_{x\sim\nu} l(h_{A,D}) \quad E_m(D) = \mathbb{E}_{x\sim\nu_m} l(h_{A,D})$$

- Goal - uniform convergence bounds:

$$E(D) \leq (1+\eta) E_m(D) + \varepsilon(m, p, n, \dots)$$

- What is the correct dependence on problem parameters?

Compressability: $A = R_\lambda$

- Maurer & Pontil 2008 @$\eta = 0$:
  - $\varepsilon = \tilde{O}\left(\sqrt{p \min(p,n) \lambda^2 / m}\right)$
  - Dimension free!
  - Quadratic loss only

Under $\lambda = 1$, $E_m(D) = 0.01$, w.p. at least $1 - e^{-10}$:

- Our results:
  - Any Lipschitz loss l
  - @$\eta = 0$: $\varepsilon = \tilde{O}\left(\sqrt{pn \ln(\lambda) / m}\right)$
  - @$\eta = 0.1$: $\varepsilon = \tilde{O}\left(pn \ln(\lambda) / m\right)$

Even fast rates can be non-vacuous!

Compressability: $A = R_\lambda$

- Maurer & Pontil 2008 @$\eta = 0$:
  - $\varepsilon = \tilde{O}\left(\sqrt{p \min(p,n) \lambda^2 / m}\right)$
  - Dimension free!
  - Quadratic loss only
- Our results:
  - Any Lipschitz loss l
  - @$\eta = 0$: $\varepsilon = \tilde{O}\left(\sqrt{pn \ln(\lambda) / m}\right)$
  - @$\eta = 0.1$: $\varepsilon = \tilde{O}\left(pn \ln(\lambda) / m\right)$

Under $\lambda = 1$, $E_m(D) = 0.01$, w.p. at least $1 - e^{-10}$:

Even fast rates can be non-vacuous!

Compressability: $A = R_\lambda$

- Maurer & Pontil 2008 @$\eta = 0$:
  - $\varepsilon = \tilde{O}\left(\sqrt{p \min(p,n) \lambda^2 / m}\right)$
  - Dimension free!
  - Quadratic loss only
- Our results:
  - Any Lipschitz loss l
  - @$\eta = 0$: $\varepsilon = \tilde{O}\left(\sqrt{pn \ln(\lambda)/m}\right)$
  - @$\eta = 0.1$: $\varepsilon = \tilde{O}\left(pn \ln(\lambda)/m\right)$

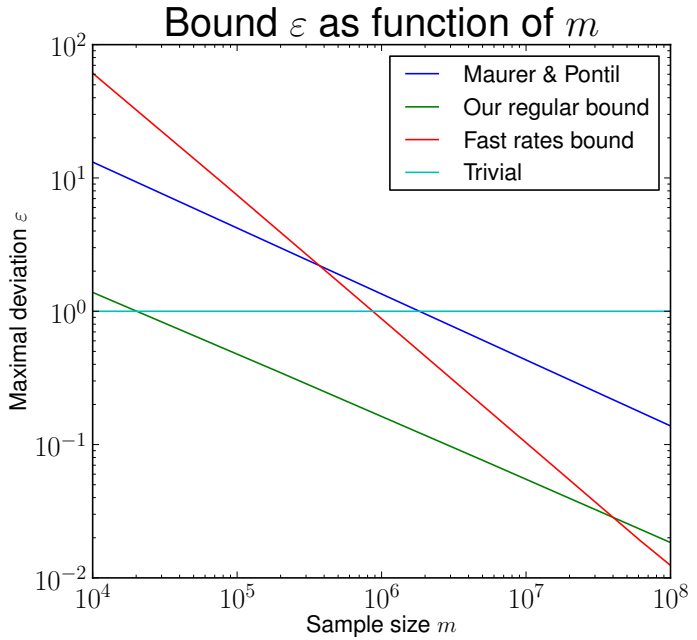
Under $\lambda = 1$, $E_m(D) = 0.01$, w.p. at least $1 - e^{-10}$:

Even fast rates can be non-vacuous!

Compressability: $A = R_\lambda$

- Maurer & Pontil 2008 @$\eta = 0$:
  - $\varepsilon = \tilde{O}\left(\sqrt{p \min(p,n) \lambda^2 / m}\right)$
  - Dimension free!
  - Quadratic loss only
- Our results:
  - Any Lipschitz loss l
  - @$\eta = 0$: $\varepsilon = \tilde{O}\left(\sqrt{pn \ln(\lambda)/m}\right)$
  - @$\eta = 0.1$: $\varepsilon = \tilde{O}\left(pn \ln(\lambda)/m\right)$

Under $\lambda = 1$, $E_m(D) = 0.01$, w.p. at least $1 - e^{-10}$:

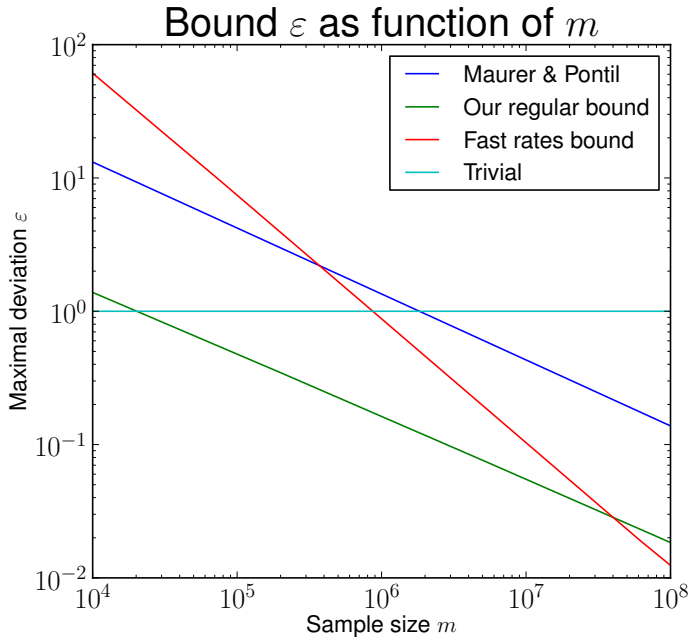


Even fast rates can be non-vacuous!

Compressability: $A = R_\lambda$

- Maurer & Pontil 2008 @$\eta = 0$:
  - $\varepsilon = \tilde{O}\left(\sqrt{p \min(p,n) \lambda^2 / m}\right)$
  - Dimension free!
  - Quadratic loss only
- Our results:
  - Any Lipschitz loss l
  - @$\eta = 0$: $\varepsilon = \tilde{O}\left(\sqrt{pn \ln(\lambda)/m}\right)$
  - @$\eta = 0.1$: $\varepsilon = \tilde{O}\left(pn \ln(\lambda)/m\right)$

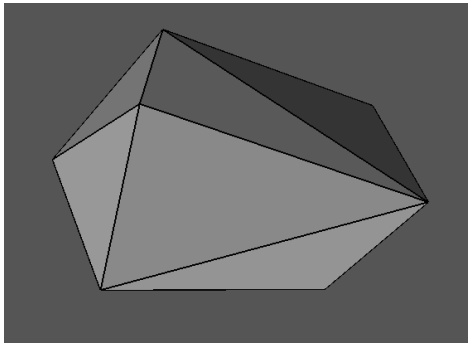
Under $\lambda = 1$, $E_m(D) = 0.01$, w.p. at least $1 - e^{-10}$:



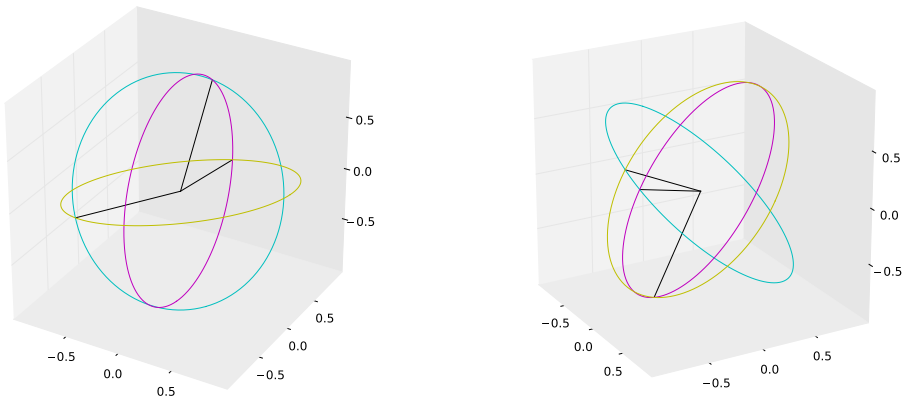
Even fast rates can be non-vacuous!

Proof sketch for R_λ

- $D \mapsto h_{R_\lambda, D} \in C\left(\mathbb{S}^{n-1}\right)$ is λ Lipschitz
- Covering number arguments



$D \mapsto h_{H_k,D}$ has singularities



- To be Lipschitz, stay away from singularities.

Stable dictionaries for H_k

Definition (The Babel function)

$$\mu_k(D) = \max_{i\in\{1,\dots,p\}} \max_{\{\Lambda | \Lambda\subset\{1,\dots,p\}\setminus i \text{ and } |\Lambda|=k\}} \sum_{j\in\Lambda} |\langle D_i, D_j\rangle|$$

- Leads to some classical results for $\arg\min_{a\in H_k} \|Da - x\|$:
  1. If $\mu_k(D) + \mu_{k-1}(D) < 1$ and $\exists a$ s.t. $Da = x$, finding a is easy.
  2. If $\mu_k(D) < \frac{1}{3}$, OMP has approximation ratio $\sqrt{1+6k}$.
- Allows compressed sensing with random projections
- Result: choose D uniformly at random. As $n \nearrow \infty$, for $k^2 \log p \lesssim n$, $P(\mu_{k-1}(D) < 1/2) \nearrow 1$.
- Most relevant is the next result.

Stable dictionaries for H_k

Definition (The Babel function)

$$\mu_k(D) = \max_{i \in \{1,\dots,p\}} \max_{\{\Lambda | \Lambda \subset \{1,\dots,p\} \setminus i \text{ and } |\Lambda| = k\}} \sum_{j \in \Lambda} |\langle D_i, D_j \rangle|$$

- Leads to some classical results for $\arg\min_{a \in H_k} \|Da - x\|$:
  1. If $\mu_k(D) + \mu_{k-1}(D) < 1$ and $\exists a$ s.t. $Da = x$, finding a is easy.
  2. If $\mu_k(D) < \frac{1}{3}$, OMP has approximation ratio $\sqrt{1+6k}$.
- Allows compressed sensing with random projections
- Result: choose D uniformly at random. As $n \nearrow \infty$, for $k^2 \log p \lesssim n$, $P(\mu_{k-1}(D) < 1/2) \nearrow 1$.
- Most relevant is the next result.

Stable dictionaries for H_k

Definition (The Babel function)

$$\mu_k(D) = \max_{i\in\{1,\dots,p\}} \max_{\{\Lambda | \Lambda\subset\{1,\dots,p\}\setminus i \text{ and } |\Lambda|=k\}} \sum_{j\in\Lambda} |\langle D_i, D_j\rangle|$$

- Leads to some classical results for $\arg\min_{a\in H_k} \|Da - x\|$:
  1. If $\mu_k(D) + \mu_{k-1}(D) < 1$ and $\exists a$ s.t. $Da = x$, finding a is easy.
  2. If $\mu_k(D) < \frac{1}{3}$, OMP has approximation ratio $\sqrt{1+6k}$.
- Allows compressed sensing with random projections
- Result: choose D uniformly at random. As $n \nearrow \infty$, for $k^2 \log p \lesssim n$, $P(\mu_{k-1}(D) < 1/2) \nearrow 1$.
- Most relevant is the next result.

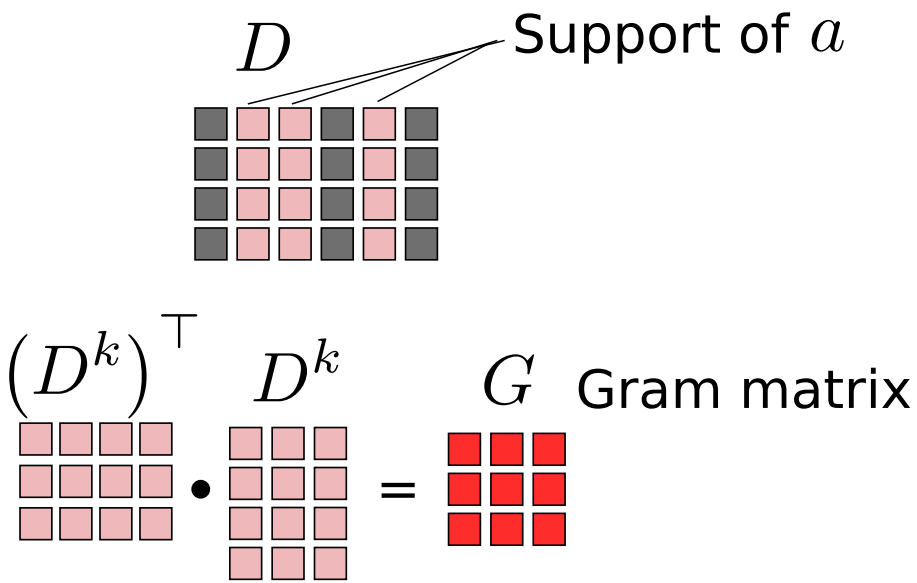
Sparsity: $A = H_k$

Theorem

Let $\mu_{k-1}(D) \leq \delta < 1$, *then* $\arg\min_{a \in H_k} \|Da - x\|_2 \in R_{\frac{k}{1-\delta}}$.

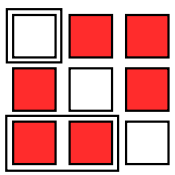
- For class $\{D : \mu_{k-1}(D) \leq \delta\}$, enables the results for R_λ
- Dependence on λ matters

Proof sketch for H_k: stability



Proof sketch for H_k: stability

$$G_{i,i} = \left\langle D_i^k, D_i^k \right\rangle = 1$$

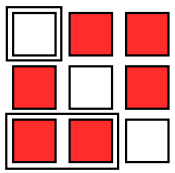


$$\sum_{j \neq i} |G_{j,i}| = \sum_{j \neq i} \left| \left\langle D_i^k, D_j^k \right\rangle \right| \leq \delta$$

- By Gersgorin's disc theorem, G has eigenvalues $v_i \in 1 + [-\delta, \delta]$
- Then $\left\| G^{-1} \right\|_{2,2} \leq (1 - \delta)^{-1}$

Proof sketch for H_k: stability

$$G_{i,i} = \left\langle D_i^k, D_i^k \right\rangle = 1$$



$$\sum_{j \neq i} |G_{j,i}| = \sum_{j \neq i} \left| \left\langle D_i^k, D_j^k \right\rangle \right| \leq \delta$$

- By Gersgorin's disc theorem, G has eigenvalues $v_i \in 1 + [-\delta, \delta]$
- Then $\left\| G^{-1} \right\|_{2,2} \leq (1 - \delta)^{-1}$

Proof sketch for H_k: connection

- Then

$$a = \arg\min_{a \in \mathbb{R}^k} \|Da - x\|_2 = \left(\left(D^k\right)^\top D^k\right)^{-1} \left(D^k\right)^\top x$$

$$\|a\|_2 \leq \left\|G^{-1}\right\|_{2,2} \left\|\left(D^k\right)^\top x\right\|_2$$

- x and each element of D^k have unit norm, then $\left\|\left(D^k\right)^\top x\right\|_\infty \leq 1 \Rightarrow \left\|\left(D^k\right)^\top x\right\|_2 \leq \sqrt{k}$
- Then $\|a\|_2 \leq \sqrt{k}/(1-\delta)$
- Thus $\|a\|_1 \leq k/(1-\delta)$ because a is supported on k elements

Proof sketch for H_k: connection

- Then

$$a = \arg\min_{a \in \mathbb{R}^k} \|Da - x\|_2 = \left(\left(D^k\right)^\top D^k\right)^{-1} \left(D^k\right)^\top x$$

$$\|a\|_2 \leq \left\|G^{-1}\right\|_{2,2} \left\|\left(D^k\right)^\top x\right\|_2$$

- x and each element of D^k have unit norm, then $\left\|\left(D^k\right)^\top x\right\|_\infty \leq 1 \Rightarrow \left\|\left(D^k\right)^\top x\right\|_2 \leq \sqrt{k}$

- Then $\|a\|_2 \leq \sqrt{k}/(1-\delta)$
- Thus $\|a\|_1 \leq k/(1-\delta)$ because a is supported on k elements

Proof sketch for H_k: connection

- Then

$$a = \arg\min_{a \in \mathbb{R}^k} \|Da - x\|_2 = \left(\left(D^k\right)^\top D^k\right)^{-1} \left(D^k\right)^\top x$$

$$\|a\|_2 \leq \left\|G^{-1}\right\|_{2,2} \left\|\left(D^k\right)^\top x\right\|_2$$

- x and each element of D^k have unit norm, then $\left\|\left(D^k\right)^\top x\right\|_\infty \leq 1 \Rightarrow \left\|\left(D^k\right)^\top x\right\|_2 \leq \sqrt{k}$
- Then $\|a\|_2 \leq \sqrt{k}/(1-\delta)$
- Thus $\|a\|_1 \leq k/(1-\delta)$ because a is supported on k elements

Summary

- Uniform convergence bounds for l_1 DL, weakly dependent on λ
- New bounds for k sparse DL, related to μ_{k-1} orthogonality
- Evidence that μ_{k-1} requirement is "reasonable"
- Initial study of dictionary learning in kernel spaces (in paper)

Summary

- Uniform convergence bounds for l_1 DL, weakly dependent on λ
- New bounds for k sparse DL, related to μ_{k-1} orthogonality
- Evidence that μ_{k-1} requirement is "reasonable"
- Initial study of dictionary learning in kernel spaces (in paper)

Summary

- Uniform convergence bounds for l_1 DL, weakly dependent on λ
- New bounds for k sparse DL, related to μ_{k-1} orthogonality
- Evidence that μ_{k-1} requirement is "reasonable"
- Initial study of dictionary learning in kernel spaces (in paper)

Open problems

- Is Babel the "real" parameter defining dictionary usability in H_k?
  - Gap between upper bounds and lower bounds
- Dictionary learning algorithms:
  - Optimality/approximation guarantees
  - Hardness results
- Babel function regularization in algorithms