An Introduction to Sparse Coding and Dictionary Learning

Kai Cao

January 14, 2014

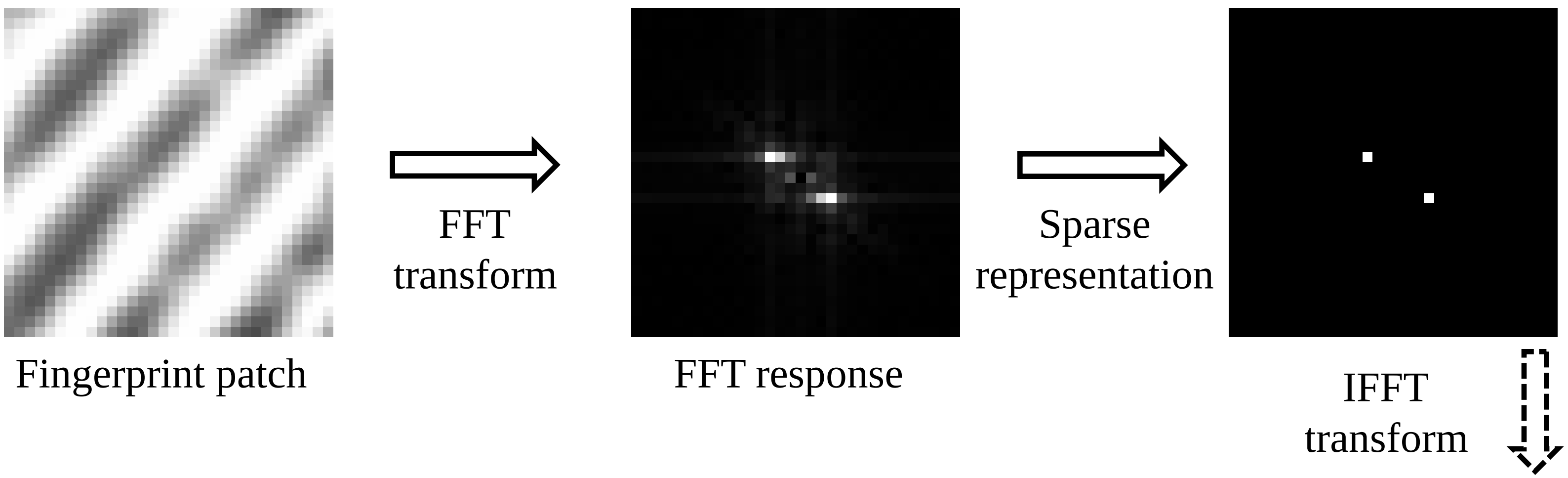
Outline

- Introduction
- Mathematical foundation
- Sparse coding
- Dictionary learning
- Summary

Introduction

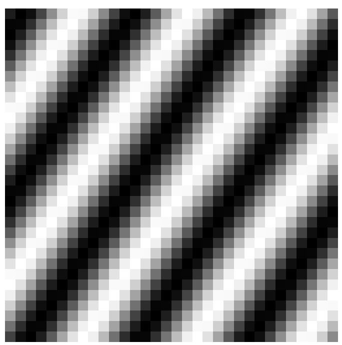
What is sparsity?

- Sparsity implies many zeros in a vector or a matrix



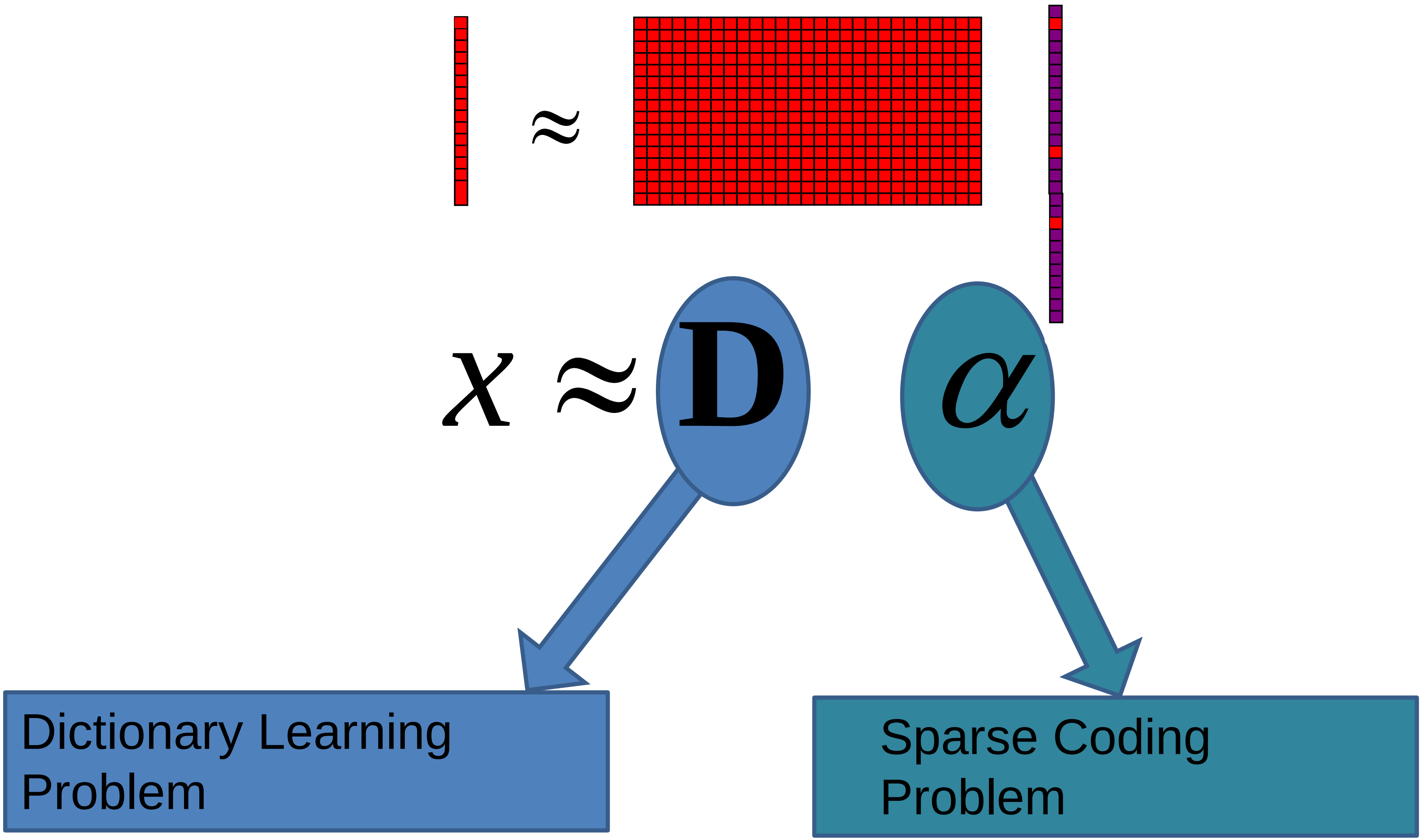
Usage:

- Compression
- Analysis
- Denoising
- ……



Reconstructed patch

Sparse Representation

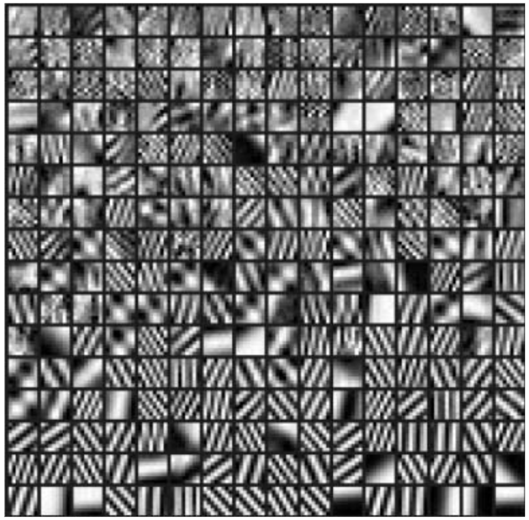


Application---Denoising

Source

Noisy image

Result 30.829dB

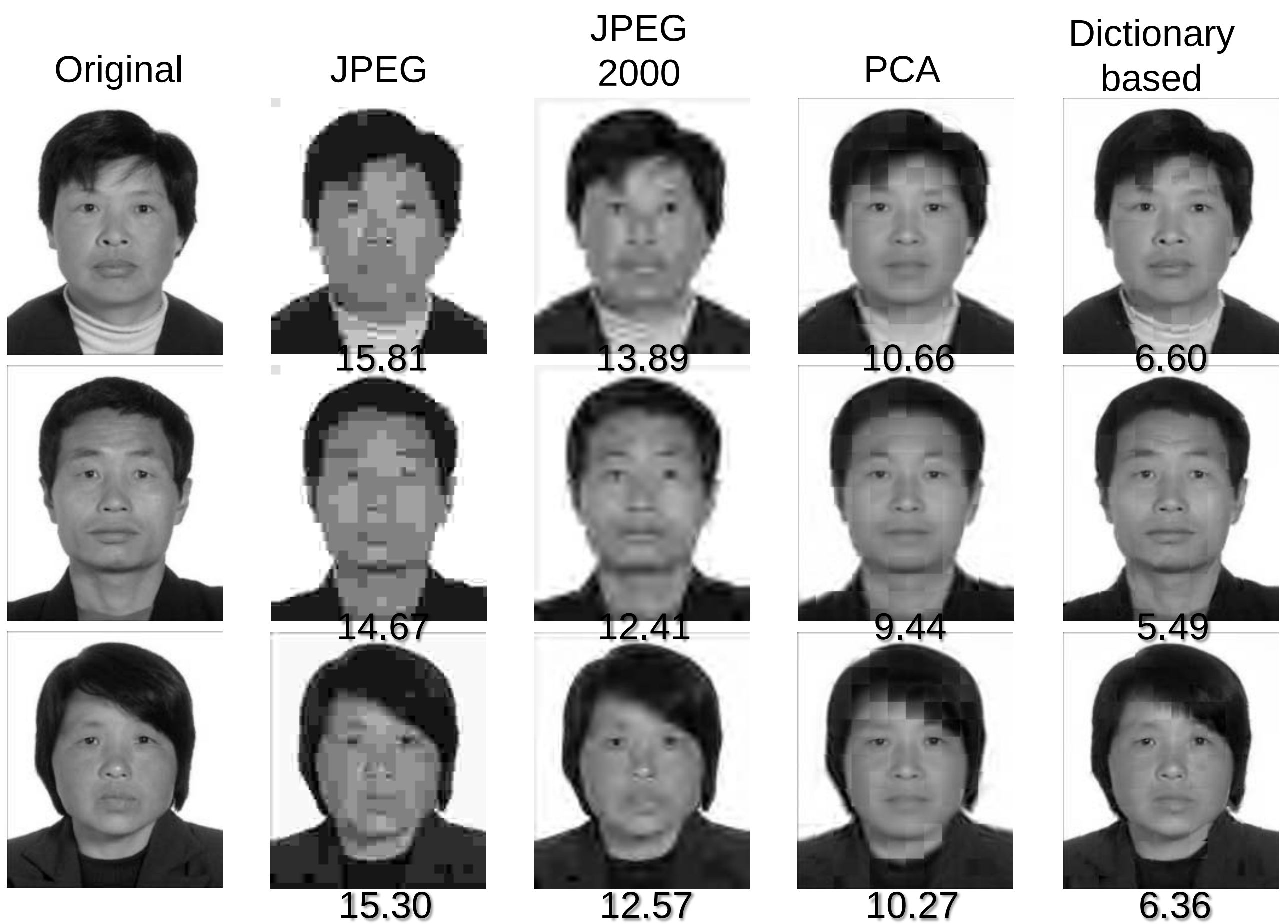


Dictionary

[M. Elad, Springer 2010]

Application---Compression

550 bytes per image



Bottom: RMSE values

[O. Bryta, M. Elad, 2008]

Mathematical foundation

Derivatives of vectors

- First order

$$\frac{\partial a^T x}{x} = \frac{\partial x^T a}{x} = a$$

- Second order

$$\frac{\partial x^T Bx}{\partial x} = (B + B^T)x$$

- Exercise

$$\min_{\alpha \in \mathbb{R}^m} \frac{1}{2} \| x - D\alpha \|_2^2 + \lambda \| \alpha \|_2^2, \quad x \in R^n, D \in R^{n \times m}$$

$$\Rightarrow \qquad \alpha = (D^T D + \lambda I)^{-1} D^T x$$

Trace of a Matrix

- Definition

$$Tr(A) = \sum_{i=1}^{n} a_{ii}, \qquad A = (a_{ij}) \in R^{n\times n}$$

- Properties

$$\| A \|_F^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij}^2 = Tr(A^T A),$$

$$Tr(A) = Tr(A^T),$$

$$Tr(A+B) = Tr(A+B), \qquad B \in R^{n\times n}$$

$$Tr(aA) = aTr(A), \qquad a \in R$$

$$Tr(AB) = Tr(BA), \qquad B \in R^{n\times n}$$

$$Tr(ABC) = Tr(BCA) = Tr(CAB), \qquad B, C \in R^{n\times n}$$

Derivatives of traces

- First order

$$\frac{\partial}{\partial X} Tr(XA) = A^T$$

$$\frac{\partial}{\partial X} Tr(X^T A) = A$$

- Derivatives of traces

$$\frac{\partial}{\partial X} Tr(X^T XA) = XA^T + XA$$

$$\frac{\partial}{\partial X} Tr(X^T BX) = B^T X + BX$$

- Exercise

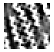
$$\min_{A \in R^{k \times m}} \| X - DA \|_F^2 + \lambda \| A \|_F^2, \quad X \in R^{n \times m}, D \in R^{n \times k}$$

$$\Longrightarrow A = (D^T D + \lambda I)^{-1} D^T X$$

Sparse coding

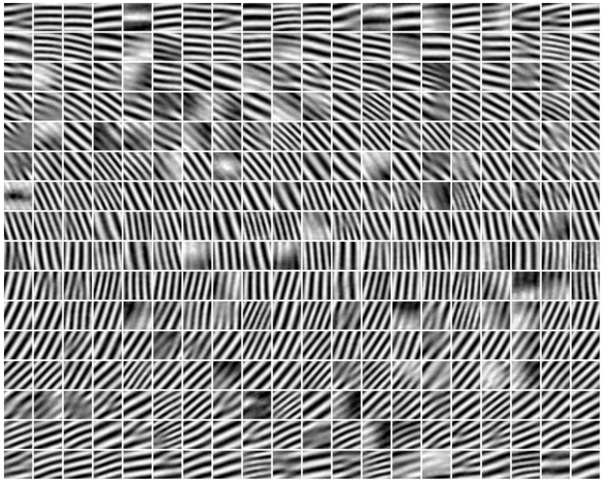
Sparse linear model

- Let $x \in R^n$ be a signal



x

- Let $D = [d_1, d_2, \ldots, d_m] \in R^{n \times m}$ be a set of normalized $(d_i^T d_i = 1)$ "basis vectors" (dictionary)



D

- Sparse representation is to find a sparse vector $\alpha \in R^m$ such that $x \approx D\alpha$, where α is regarded as sparse code

The sparse coding model

- Objective function

$$\underbrace{\min_{\alpha \in \mathbb{R}^m} \frac{1}{2} \| x - D\alpha \|_2^2}_{\text{Data fitting term}} \quad + \quad \underbrace{\lambda \varphi(\alpha)}_{\text{Regularization term}}$$

- The regularization term φ can be
  - the l_2 norm. $\| \alpha \|_2^2 \triangleq \sum_{i=1}^{m} \alpha_i^2$
  - the l_0 norm. $\| \alpha \|_0 \triangleq \#\{i \mid a_i \neq 0\}$
  - the l_1 norm. $\| \alpha \|_1 \triangleq \sum_{i=1}^{m} | \alpha_i |$
  - …

Sparsity inducing

Matching pursuit

$$\min_{\alpha \in \mathbb{R}^m} \frac{1}{2} \| x - D\alpha \|_2^2 \quad \text{s. t.} \quad \| \alpha \|_0 \leq L$$

1. Initialization: α = 0, residual r = x
2. while $||\alpha||_0 < L$
3. Select the element with maximum correlation with the residual

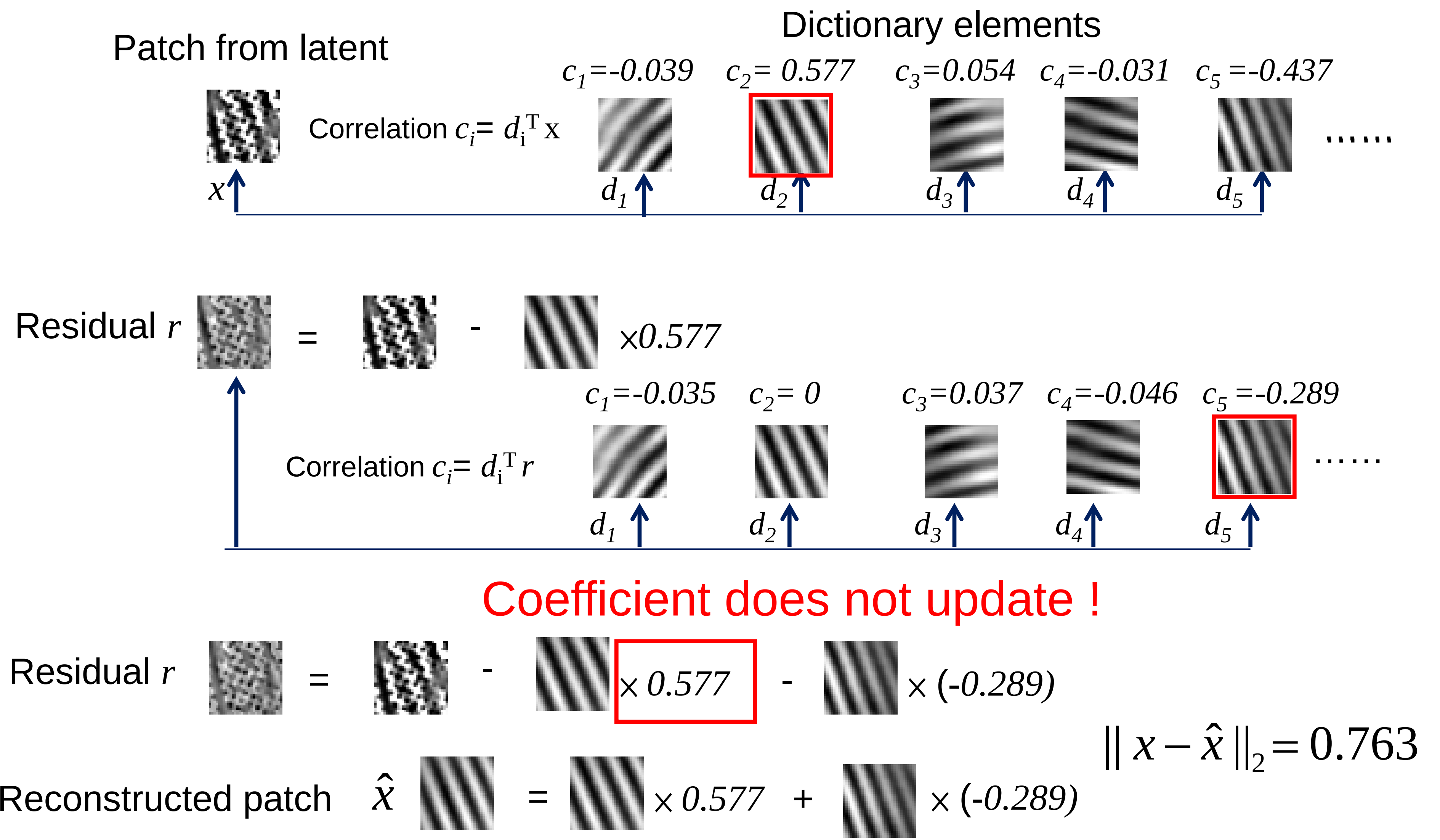
$$\hat{i} = \arg\max_{i=1,\ldots,m} | d_i^T r |$$

4. Update the coefficients and residual

$$\alpha_{\hat{i}} = \alpha_i + d_i^T r$$
$$r = r - (d_{\hat{i}}^T r) d_i$$

5. End while

An example for matching pursuit



Orthogonal matching pursuit

$$\min_{\alpha \in \mathbb{R}^m} \frac{1}{2} \| x - D\alpha \|_2^2 \quad \text{s. t.} \quad \| \alpha \|_0 \leq L$$

1. Initialization: $\alpha = 0$, residual $r = x$, active set $\Omega = \varnothing$
2. while $||\alpha||_0 < L$
3. Select the element with maximum correlation with the residual

$$\hat{i} = \arg\max_{i=1,\ldots,m} | d_i^T r |$$

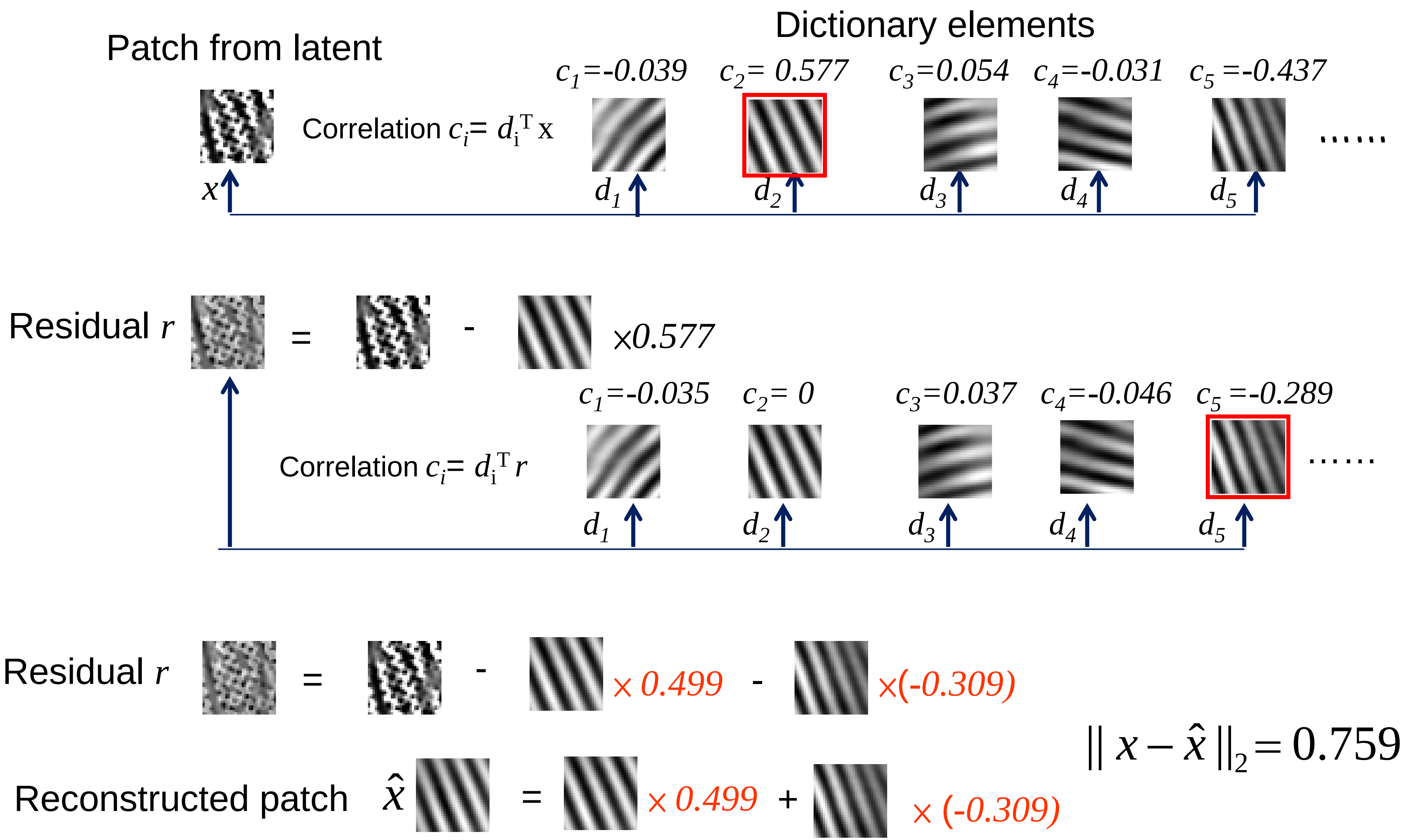
4. Update the active set, coefficients and residual

$$\Omega = \Omega \cup \hat{i}$$

$$\boxed{\begin{aligned} \alpha_\Omega &= (d_\Omega^T d_\Omega)^{-1} d_\Omega^T r \\ r &= x - d_\Omega \alpha_\Omega \end{aligned}}$$

5. End while

An example for orthogonal matching pursuit



Why does l_1-norm induce sparsity?

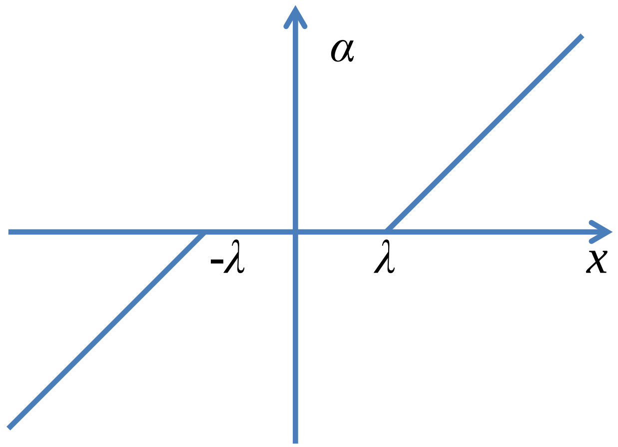
- Analysis in 1D (comparison with l_2)

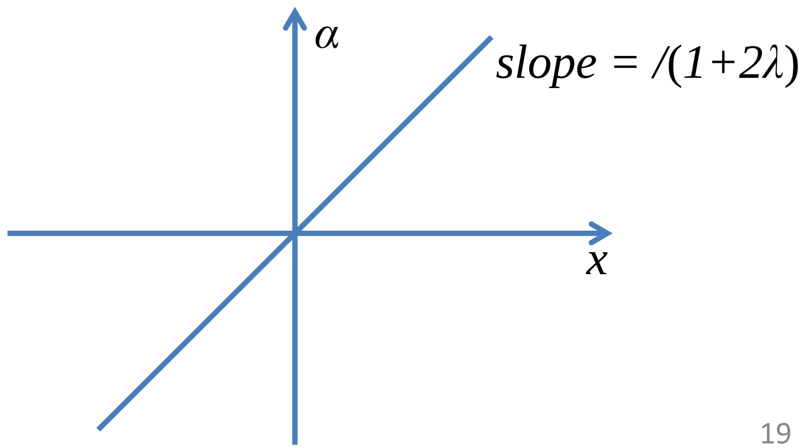
$$\min_{\alpha \in \mathbb{R}} \frac{1}{2}(x-\alpha)^2 + \lambda |\alpha|$$

$$\min_{\alpha \in \mathbb{R}} \frac{1}{2}(x-\alpha)^2 + \lambda \alpha^2$$

$$\Rightarrow \begin{array}{ll} \text{if } x \geq \lambda, & \alpha = x - \lambda \\ \text{if } x \leq -\lambda, & \alpha = x + \lambda \\ \text{else}, & \alpha = 0 \end{array}$$

$$\Rightarrow \alpha = x/(1+2\lambda)$$





Why does l_1-norm induce sparsity?

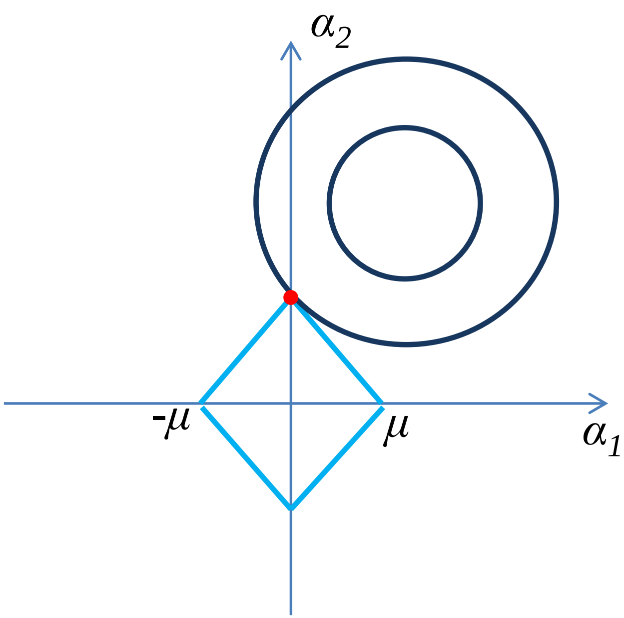
- Analysis in 2D (comparison with l_2)

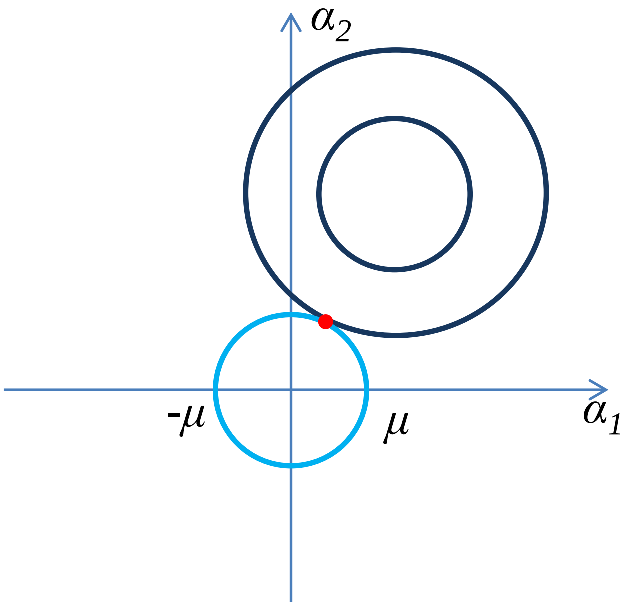
$$\min_{\alpha \in \mathbb{R}} \frac{1}{2} \| x - \alpha \|_2^2 + \lambda \| \alpha \|_1$$

$$\Leftrightarrow \min_{\alpha \in \mathbb{R}} \frac{1}{2} \| x - \alpha \|_2^2 \text{ s.t. } \| \alpha \|_1 \leq \mu$$

$$\min_{\alpha \in \mathbb{R}} \frac{1}{2} \| x - \alpha \|_2^2 + \lambda \| \alpha \|_2^2$$

$$\Leftrightarrow \min_{\alpha \in \mathbb{R}} \frac{1}{2} \| x - \alpha \|_2^2 \text{ s.t. } \| \alpha \|_2 \leq \mu$$





Optimality condition for l_1-norm regularization

$$\min_{\alpha \in \mathbb{R}^m} J(\alpha) = \frac{1}{2} \| x - D\alpha \|_2^2 + \lambda \| \alpha \|_1$$

- Directional derivative in the direction u at α

$$\nabla J(\alpha, u) = \lim_{t \to 0^+} \frac{J(\alpha + tu) - J(\alpha)}{t}$$

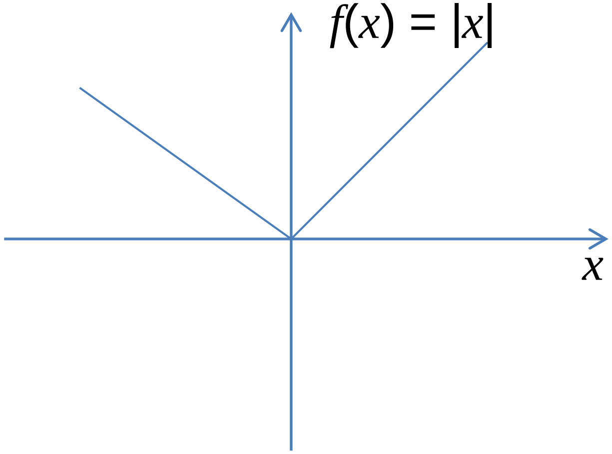
- g is subgradient of J at α if and only if

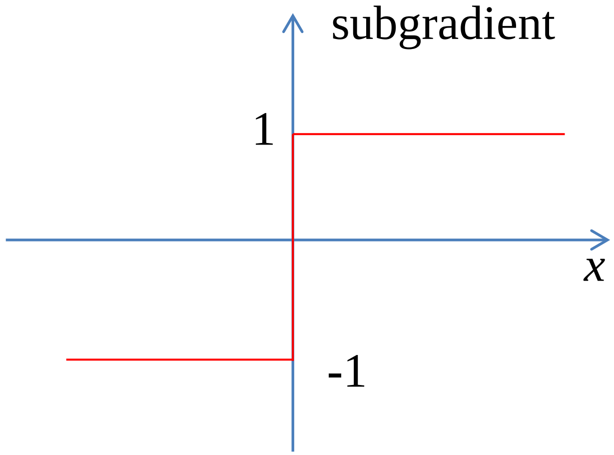
$$\forall t \in R^m, J(t) \geq J(\alpha) + g^T (t - \alpha)$$

- Proposition 1: g is a subgradient $\Leftrightarrow \forall u \in R^m, g^T u \leq \nabla J(\alpha, u)$
- Proposition 2: if J is differentiable at α, $\nabla J(\alpha, u) = \nabla J(\alpha)^T u$
- Proposition 3: α is optimal if and only if for all u, $\nabla J(\alpha, u) \geq 0$

Subgradient for l_1-norm regularization

- Example: $f(x) = |x|$





$$\nabla f(x,u) = \begin{cases} |u| & x = 0 \\ sign(x)u & x \neq 0 \end{cases}$$

Subgradient for l_1-norm regularization

$$\min_{\alpha \in \mathbb{R}^m} J(\alpha) = \frac{1}{2} \| x - D\alpha \|_2^2 + \lambda \| \alpha \|_1$$

$$\Rightarrow \quad \nabla J(\alpha, u) = -u^T D^T (x - D\alpha) + \lambda \sum_{i, a_i \neq 0} \operatorname{sign}(a_i) u_i + \lambda \sum_{i, a_i = 0} | u_i |$$

- g is a subgradient at α if and only if for all i

$$| g_i - d_i^T (x - D\alpha) | \leq \lambda \qquad \text{if} \qquad a_i = 0$$

$$g_i = d_i^T (x - D\alpha) + \lambda \operatorname{sign}(a_i) \qquad \text{if} \qquad a_i \neq 0$$

First order method for convex optimization

- Differentiable objective
  - Gradient descent: $\alpha_{t+1} = \alpha_t - \eta_t \nabla J(\alpha_t)$
  - With line search for a decent η_t
  - Diminishing step size: e.g., η_t=(t+t_0)$^{-1}$
- Non differentiable objective
  - Subgradient decent: $\alpha_{t+1} = \alpha_t - \eta_t g_t$, g_t is a subgradient
  - With line search
  - Diminishing step size

Reformulation as quadratic program

$$\min_{\alpha \in \mathbb{R}^m} \frac{1}{2} \| x - D\alpha \|_2^2 + \lambda \| \alpha \|_1$$

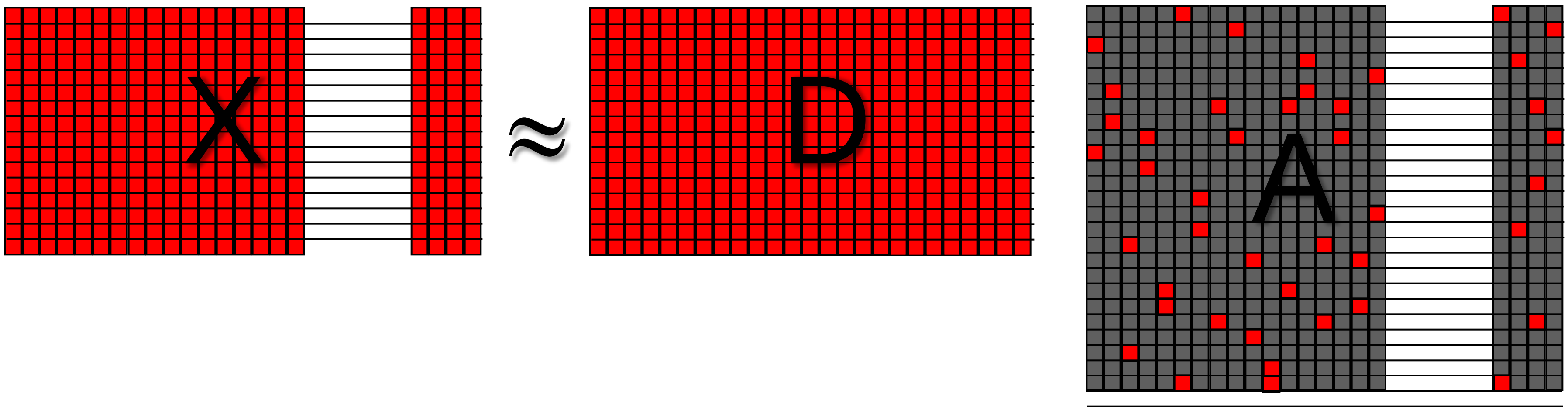
$$\Rightarrow \quad \min_{\alpha_+, \alpha_- \in \mathbb{R}_+^m} \frac{1}{2} \| x - D\alpha_+ + D\alpha_- \|_2^2 + \lambda (1^T \alpha_+ + 1^T \alpha_-)$$

Dictionary Learning

Dictionary selection

- Which D to use?
- A fixed set of basis:
  - Steerable wavelet
  - Contourlet
  - DCT Basis
  - ……
- Data adaptive dictionary – learn from data
  - K-SVD (l_0-norm)
  - On-line dictionary learning (l_1-norm)

The objective function for K-SVD



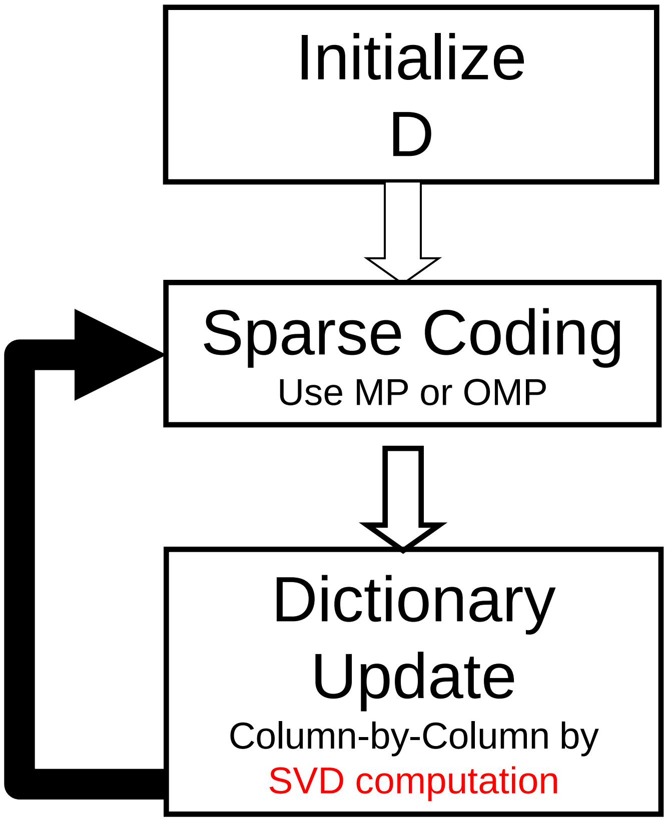
$$\min_{D,A} \| X - DA \|_F^2 \qquad \forall j, s.t. \quad \| \alpha_j \|_0 \leq L$$

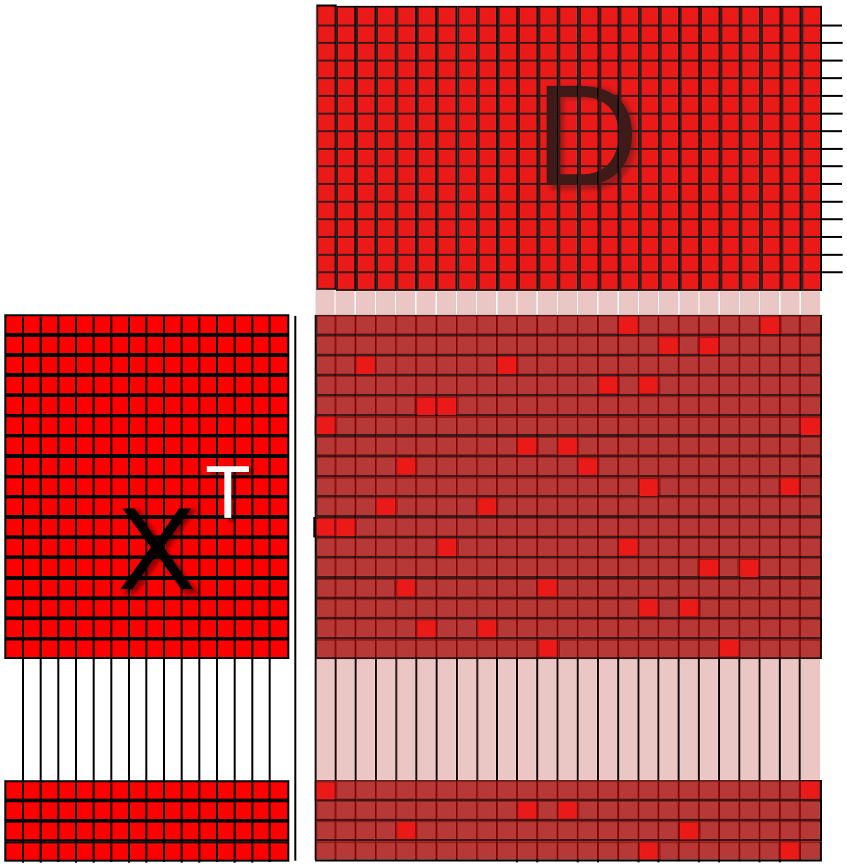
The examples are linear combinations of atoms from D

Each example has a sparse representation with no more than L atoms

www.cs.technion.ac.il/~ronrubin/Talks/**K-SVD.ppt**

K–SVD – An Overview





www.cs.technion.ac.il/~ronrubin/Talks/**K-SVD.ppt**

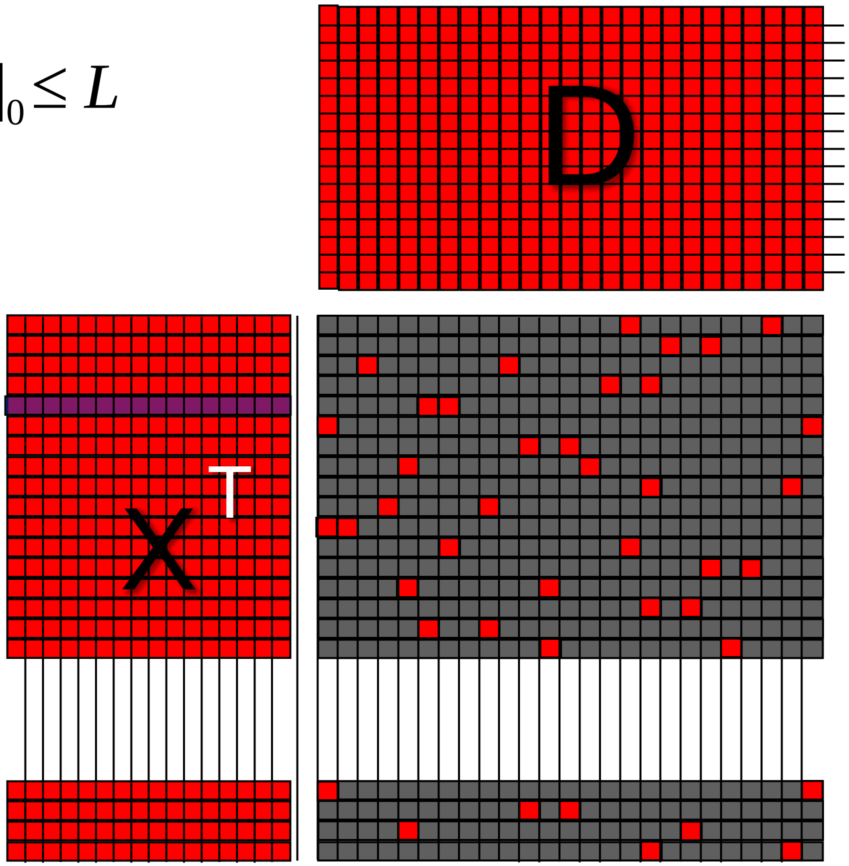
K–SVD: Sparse Coding Stage

$$\min_{A} \| X - DA \|_F^2 \qquad \forall j, s.t. \quad \| \alpha_j \|_0 \leq L$$

For the jth example we solve

$$\min_{\alpha} \quad \left\| \mathbf{D}\alpha - x_j \right\|_2^2 \quad s.t. \ \left\| \alpha \right\|_0 \leq L$$

Ordinary Sparse Coding !



www.cs.technion.ac.il/~ronrubin/Talks/**K-SVD.ppt**

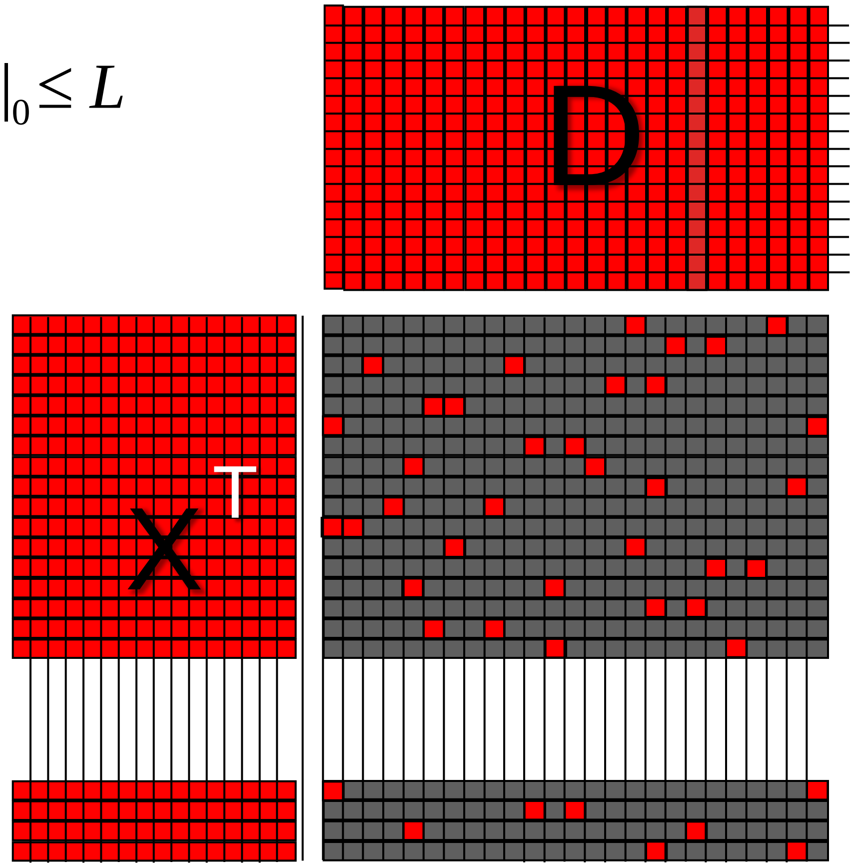
K–SVD: Dictionary Update Stage

$$\min_D \| X - DA \|_F^2 \qquad \forall j, s.t. \quad \| \alpha_j \|_0 \le L$$

For the k[th] atom we solve

$$\min_{d_k} \| d_k \alpha_T^k - E_k \|_F^2$$

$$E_k = \sum_{i \neq k} d_i \alpha_T^i - X$$ (the residual)



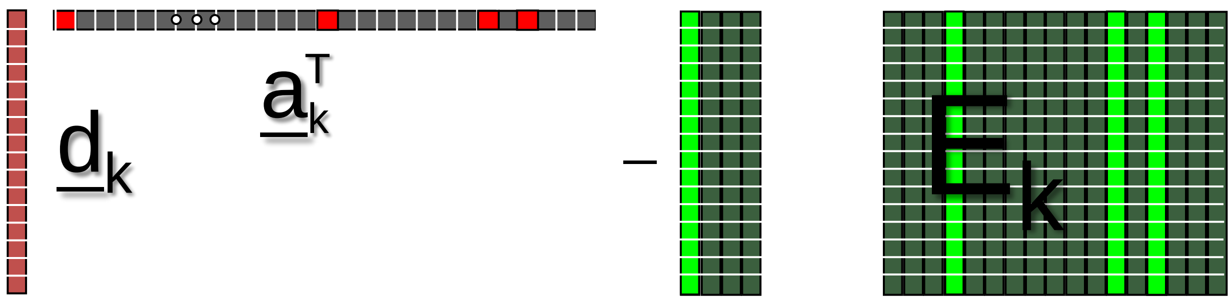
Solve with SVD

$$E_k = U \Lambda V^T \qquad d_k = u_1$$

www.cs.technion.ac.il/~ronrubin/Talks/**K-SVD.ppt**

K–SVD Dictionary Update Stage

We want to solve:



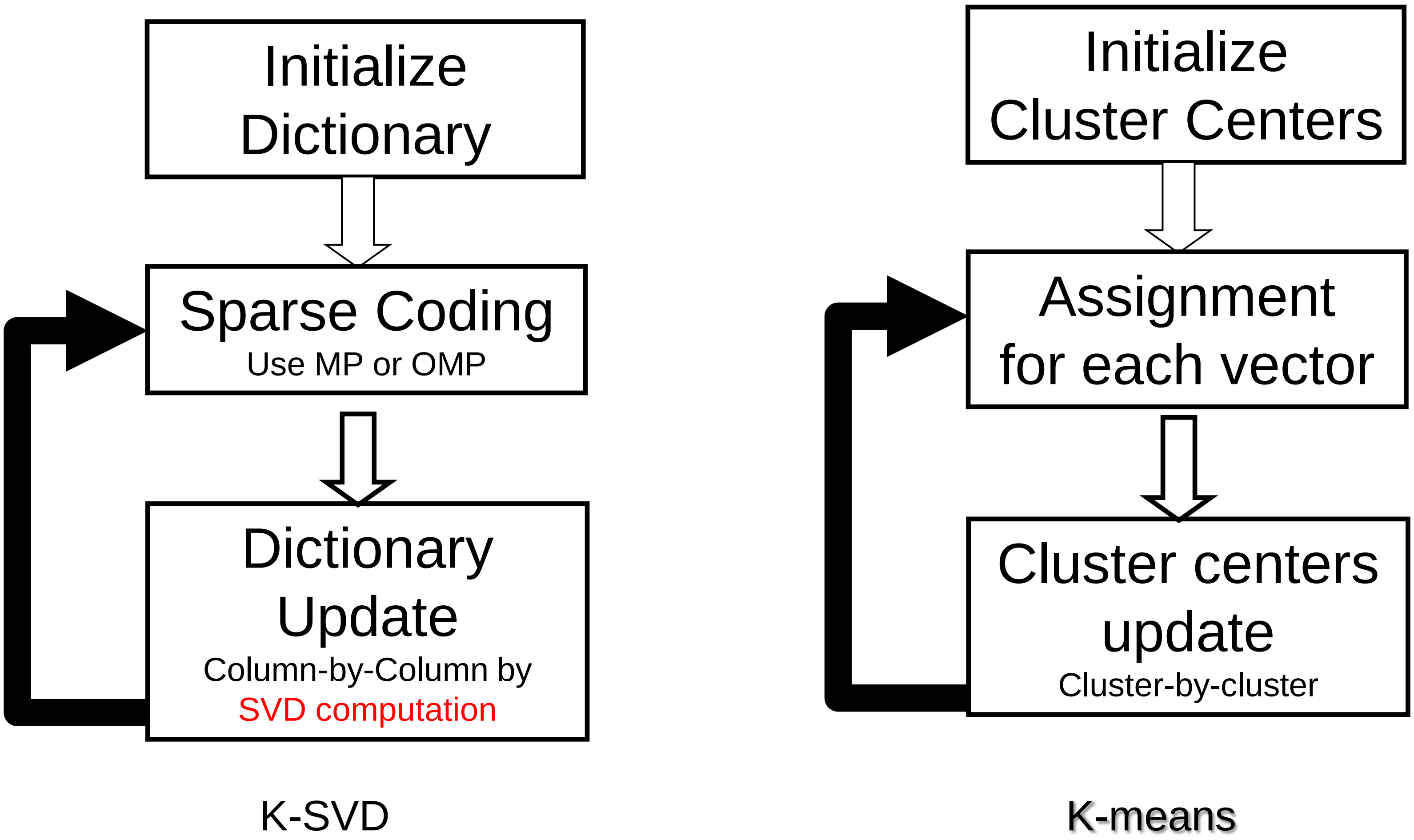
Only some of the examples use column $\underline{d}_k$!

When updating $\underline{a}_k$, only recompute the coefficients corresponding to those examples

Solve with SVD!

www.cs.technion.ac.il/~ronrubin/Talks/**K-SVD.ppt**

Compare K-SVD with K-means



dictionary learning with l_1-norm regularization

- Objective function for l_1-norm regularization

$$\min_{D} \frac{1}{t} \sum_{i=1}^{t} \frac{1}{2} \| x_i - D\alpha_i \|_2^2 + \lambda \| \alpha_i \|_1$$

where

$$\alpha_i \triangleq \arg\min_{\alpha \in R^m} \frac{1}{2} \| x_i - D\alpha \|_2^2 + \lambda \| \alpha \|_1$$

- Advantages of online learning:
  – Handle large and dynamic datasets,
  – Could be much faster than batch algorithms.

dictionary learning with l_1-norm regularization

$$F_t(D) = \frac{1}{t}\sum_{i=1}^{t}\frac{1}{2}\| x_i - D\alpha_i \|_2^2 + \lambda \| \alpha_i \|_1$$

$$= \frac{1}{t}(\frac{1}{2}Tr(D^T D A_t) - Tr(D^T B_t)) + \lambda \sum_{i=1}^{t} \| \alpha_i \|_1$$

where

$$A_t = \sum_{i=1}^{t} \alpha_i \alpha_i^T, \qquad B_t = \sum_{i=1}^{t} x_i \alpha_i^T$$

$$\Rightarrow \qquad \frac{\partial F_t(D)}{\partial D} = \frac{1}{t}(DA_t - B_t)$$

For a new x_{t+1}, $A_{t+1} = A_t + \alpha_{t+1}\alpha_{t+1}^T, \qquad B_{t+1} = B_t + x_{t+1}\alpha_{t+1}^T$

On-line dictionary learning

1) Initialization: $D_0 \in R^{n\times m}$; A_0=0; B_0=0;
2) For t=1,…,T
3) Draw x_t from the training data set
4) Get sparse code

$$\alpha_t = \arg\min_{\alpha \in R^m} \frac{1}{2} \| x_t - D_{t-1}\alpha \|_2^2 + \lambda \| \alpha \|_1$$

5) Aggregate sufficient statistics

$$A_t = A_{t-1} + \alpha_t \alpha_t^T, \quad B_t = B_{t-1} + x_t \alpha_t^T,$$

6) Dictionary update

$$D_t = D_{t-1} - \rho \frac{\partial F_t(D)}{\partial D}$$

7) End for

Toolbox - SPAMS

- SPArse Modeling Software:
  - Sparse coding
    - l_0-norm regularization
    - l_1-norm regularization
    - ……
  - Dictionary learning
    - K-SVD
    - Online dictionary learning
    - ……
- C++ implemented with Matlab interface
- http://spams-devel.gforge.inria.fr/

Summary

- Sparsity and sparse representation
- Sparse coding with l_0- and l_1-norm regularization
  - Orthogonal matching pursuit/matching pursuit
  - Subgradient and optimal condition
- Dictionary learning with l_0- and l_1-norm regularization
  - K-SVD
  - Online dictionary learning
- Try to use it !!

References

- T. T. Cai, Lie Wang ,Orthogonal Matching Pursuit for Sparse Signal Recovery With Noise, *IEEE Transactions on Information Theory*, 57(7): 4680-4688,2011
- Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Annals of statistics*, *32(2):407–499, 2004*.
- M. Aharon, M. Elad, and A. M. Bruckstein. The K-SVD: An algorithm for designing of overcomplete dictionaries for sparse representations. *IEEE Transactions on Signal Processing*, 54(11):4311-4322, November 2006.
- J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online dictionary learning for sparse coding. *In Proceedings of the International Conference on Machine Learning (ICML)*, 2009a.

Thank you for listening