

THE UNIVERSITY OF  
**SYDNEY**

# Advanced Machine Learning

(COMP 5328)

## Sparse Coding and Regularisation

Tongliang Liu



THE UNIVERSITY OF  
**SYDNEY**

# Quiz results

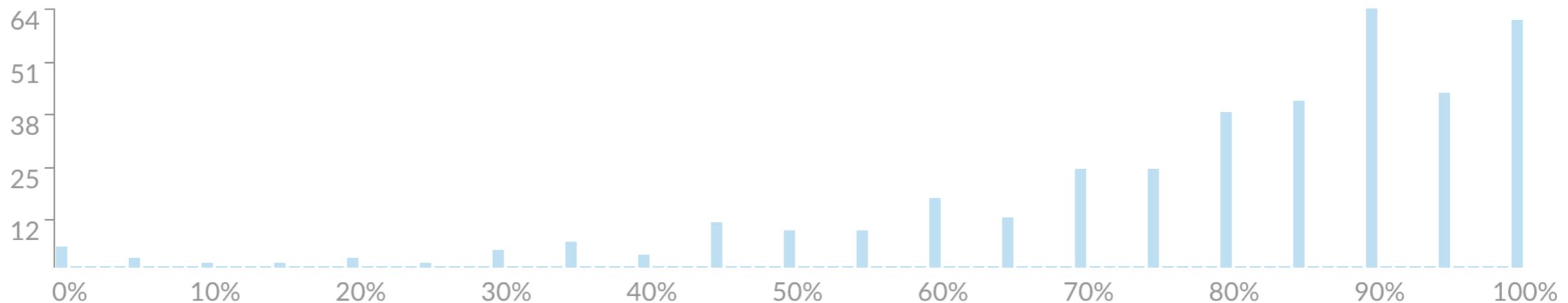
μ Average Score  
**79%**

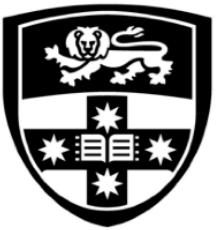
↗ High Score  
100%

↘ Low Score  
0%

σ Standard Deviation  
21.21

⌚ Average Time  
38:28





THE UNIVERSITY OF  
**SYDNEY**

# Review

# Dictionary learning

What is a dictionary in machine learning?

Let  $x \in \mathbb{R}^d$ ,  $D \in \mathbb{R}^{d \times k}$

$$\alpha^* = \arg \min_{\alpha \in \mathbb{R}^k} \|x - D\alpha\|^2.$$

Note that  $\|x\| = \sqrt{x^\top x}$  is the ell 2 norm.

Given  $x_1, \dots, x_n \in \mathbb{R}^d$

$$\{D^*, \alpha_1^*, \dots, \alpha_n^*\} = \arg \min_{D \in \mathbb{R}^{d \times k}, \alpha_1, \dots, \alpha_n \in \mathbb{R}^k} \frac{1}{n} \sum_{i=1}^n \|x_i - D\alpha_i\|^2.$$

# Dictionary learning

Note that

$$\frac{1}{n} \sum_{i=1}^n \|x_i - D\alpha_i\|^2 = \frac{1}{n} \|X - DR\|_F^2,$$

where  $X = [x_1, x_2, \dots, x_n] \in \mathbb{R}^{d \times n}$ ,

$R = [\alpha_1, \alpha_2, \dots, \alpha_n] \in \mathbb{R}^{k \times n}$ ,

$\|X\|_F = \sqrt{\text{trace}(X^\top X)} = \sqrt{\sum_{i=1}^d \sum_{j=1}^n X_{i,j}^2}$  is the Frobenius norm of  $X$ .

# Dictionary learning

Note that

$$\arg \min_{D \in \mathcal{D}, R \in \mathcal{R}} \|X - DR\|_F^2,$$

where  $\mathcal{D}$  and  $\mathcal{R}$  are some specific domains for  $D$  and  $R$ .

# Optimisation

**Objective:**

$$\min_{D \in \mathcal{D}, R \in \mathcal{R}} \|X - DR\|_F^2$$

The objective is convex with respect to either  $R$  or  $D$  but not to both.

**Fix  $R$ , solve for  $D$**

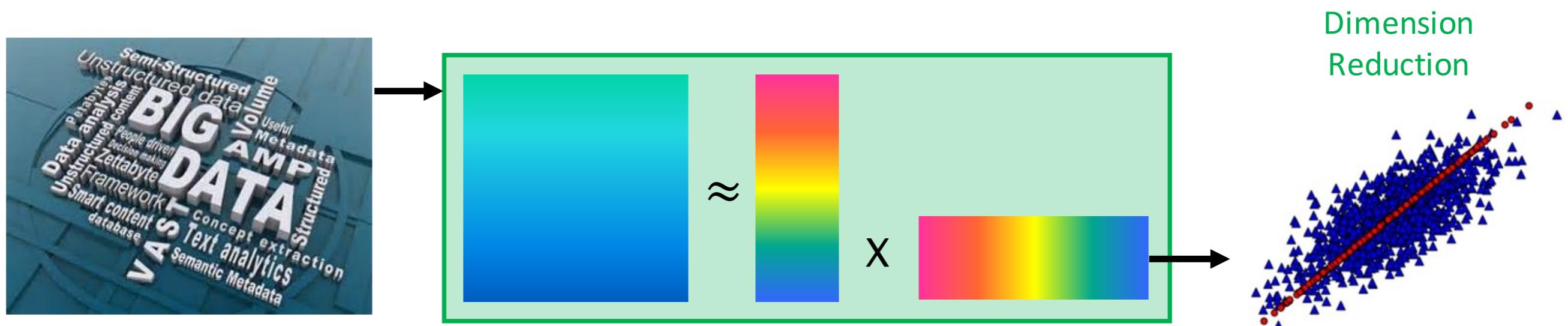
$$\min_{D \in \mathcal{D}} \|X - DR\|_F^2$$

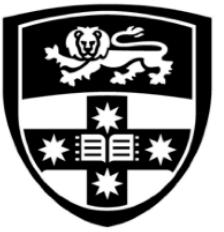
**Fix  $D$ , solve for  $R$**

$$\min_{R \in \mathcal{R}} \|X - DR\|_F^2$$

Engan, Kjersti, Sven Ole Aase, and J. Hakon Husoy. "Method of optimal directions for frame design." Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on. Vol. 5. IEEE, 1999.

# Dictionary learning application



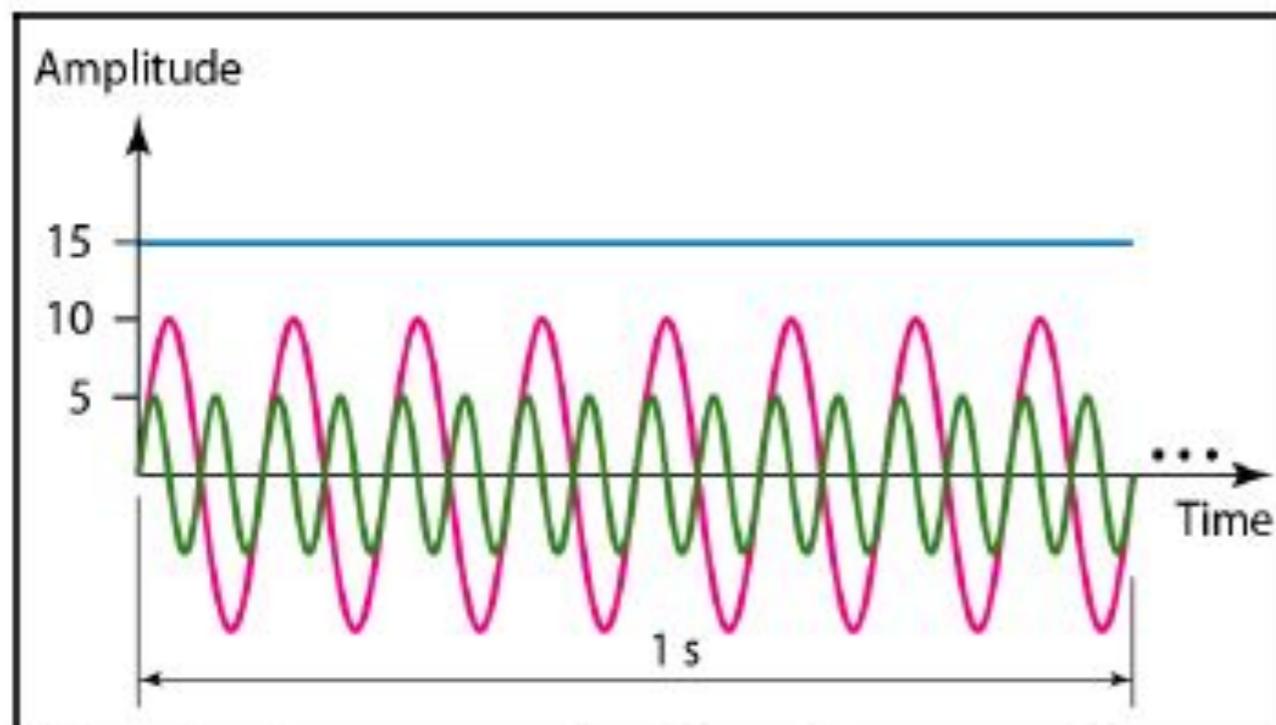


THE UNIVERSITY OF  
**SYDNEY**

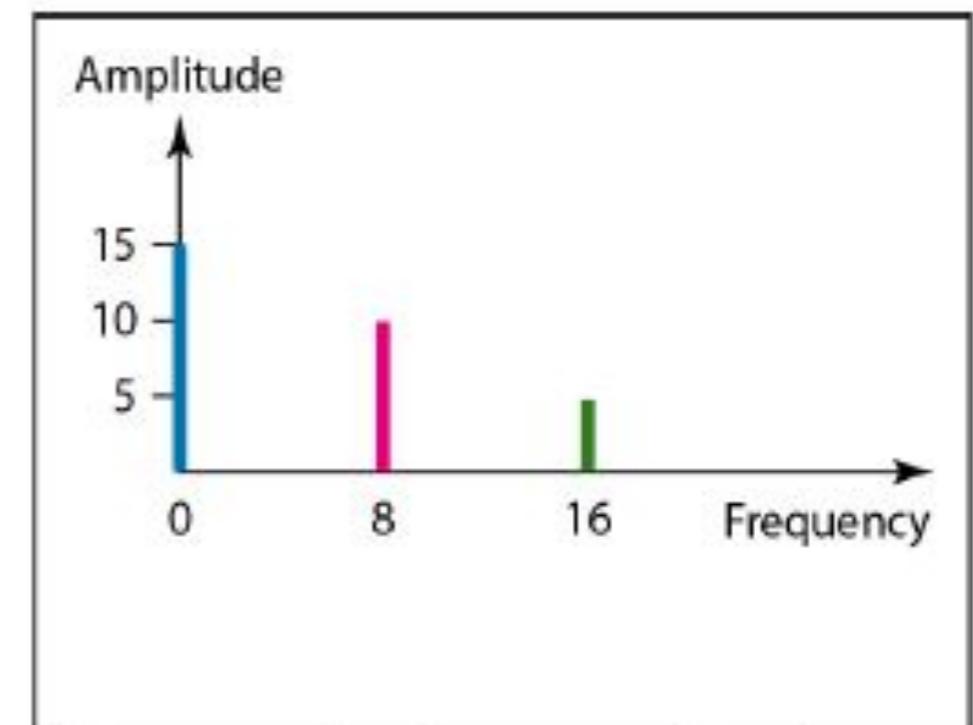
# Sparse Coding

# Why sparse?

Many signals are sparse in some transform domain.



a. Time-domain representation of three sine waves with frequencies 0, 8, and 16

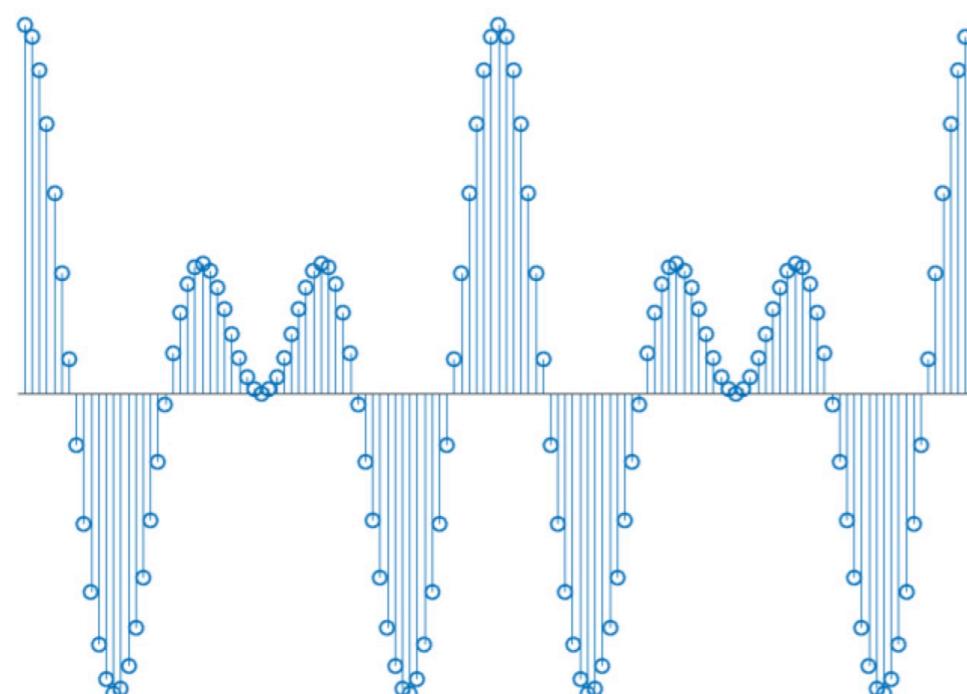


b. Frequency-domain representation of the same three signals

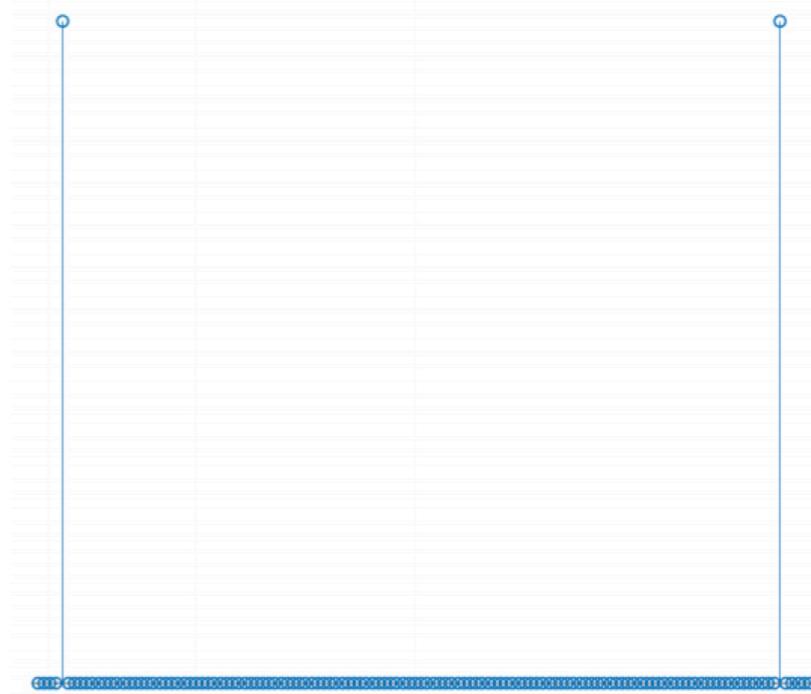
Image credit: <https://datacommandnet.blogspot.com/p/periodic-analog-signals.html>

# Why sparse?

Many signals are sparse in some transform domain.



time-representation of  $y_1$

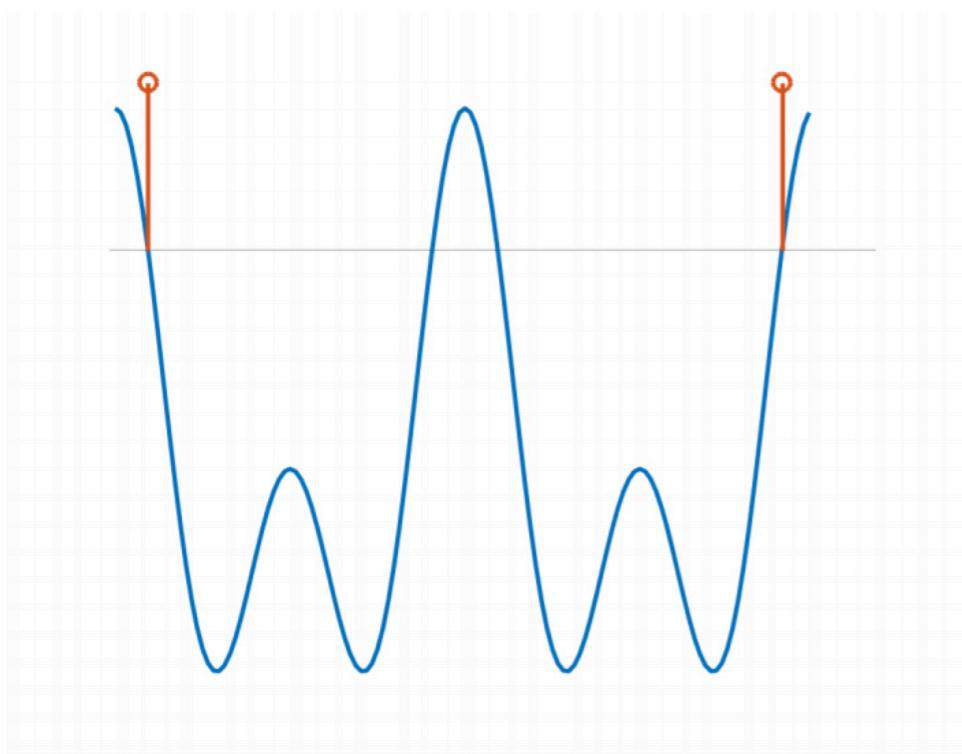


frequency-representation of  $y_1$

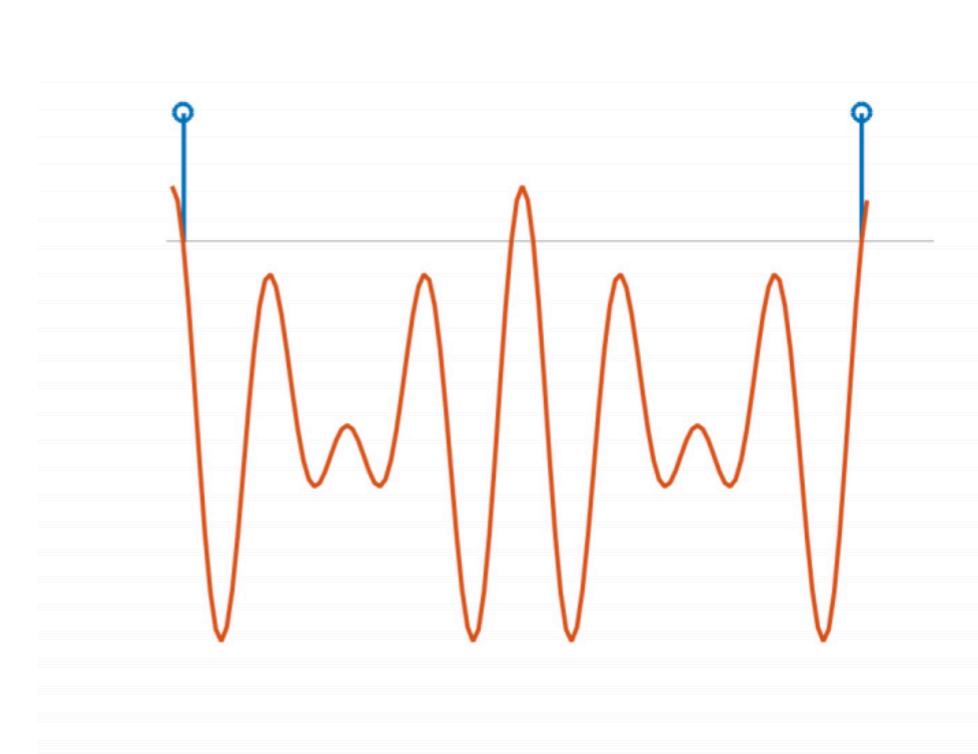
Image credit: [http://www.princeton.edu/~yc5/ele538b\\_sparsity/lectures/sparse\\_representation.pdf](http://www.princeton.edu/~yc5/ele538b_sparsity/lectures/sparse_representation.pdf)

# Why sparse?

Many signals are sparse in some transform domain.



time representation of  $y_2$



frequency representation of  $y_2$

Image credit: [http://www.princeton.edu/~yc5/ele538b\\_sparsity/lectures/sparse\\_representation.pdf](http://www.princeton.edu/~yc5/ele538b_sparsity/lectures/sparse_representation.pdf)

# Why sparse?

Many signals are sparse in some transform domain.

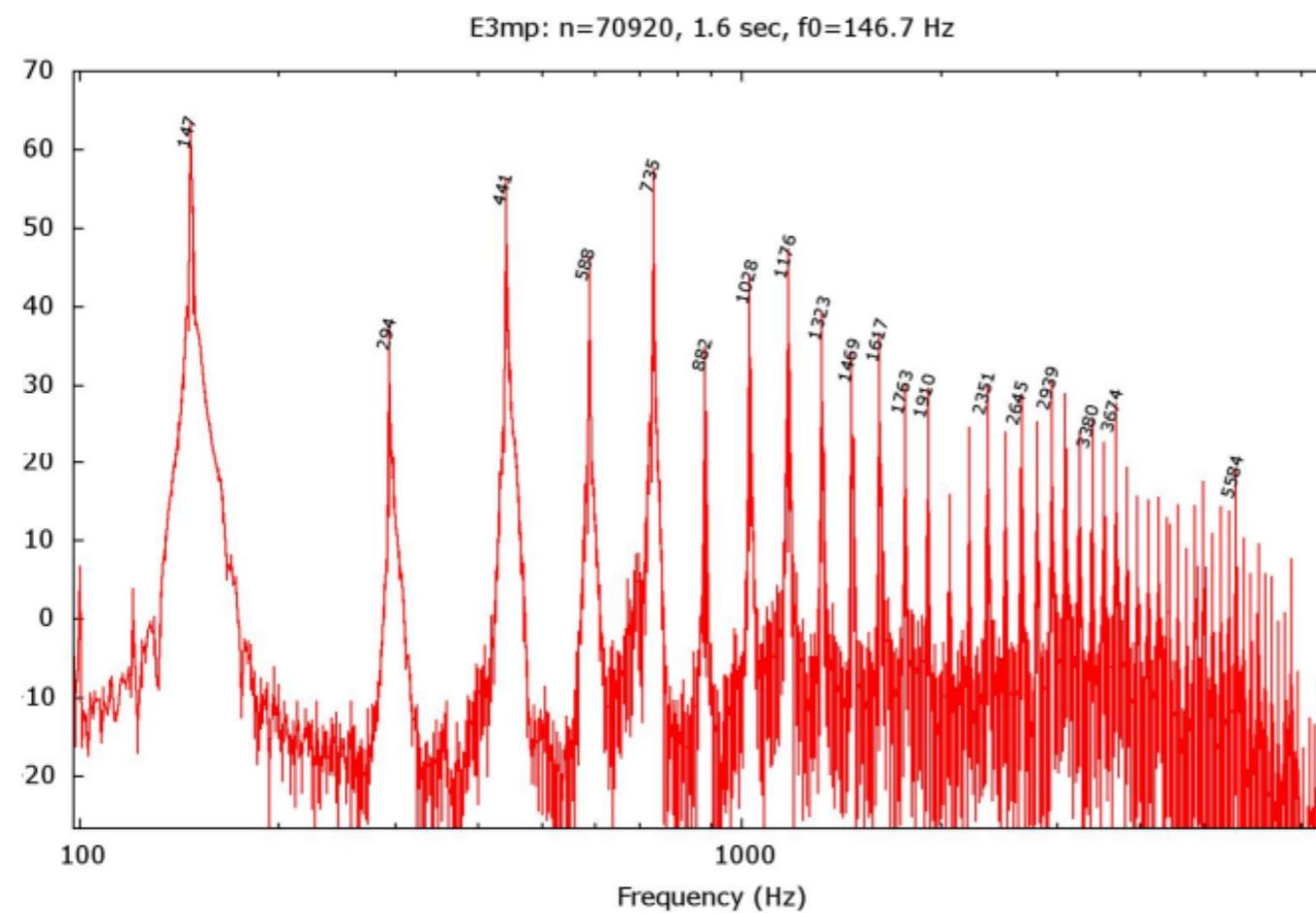


representation of  $y_2$  in overcomplete basis (time + frequency)

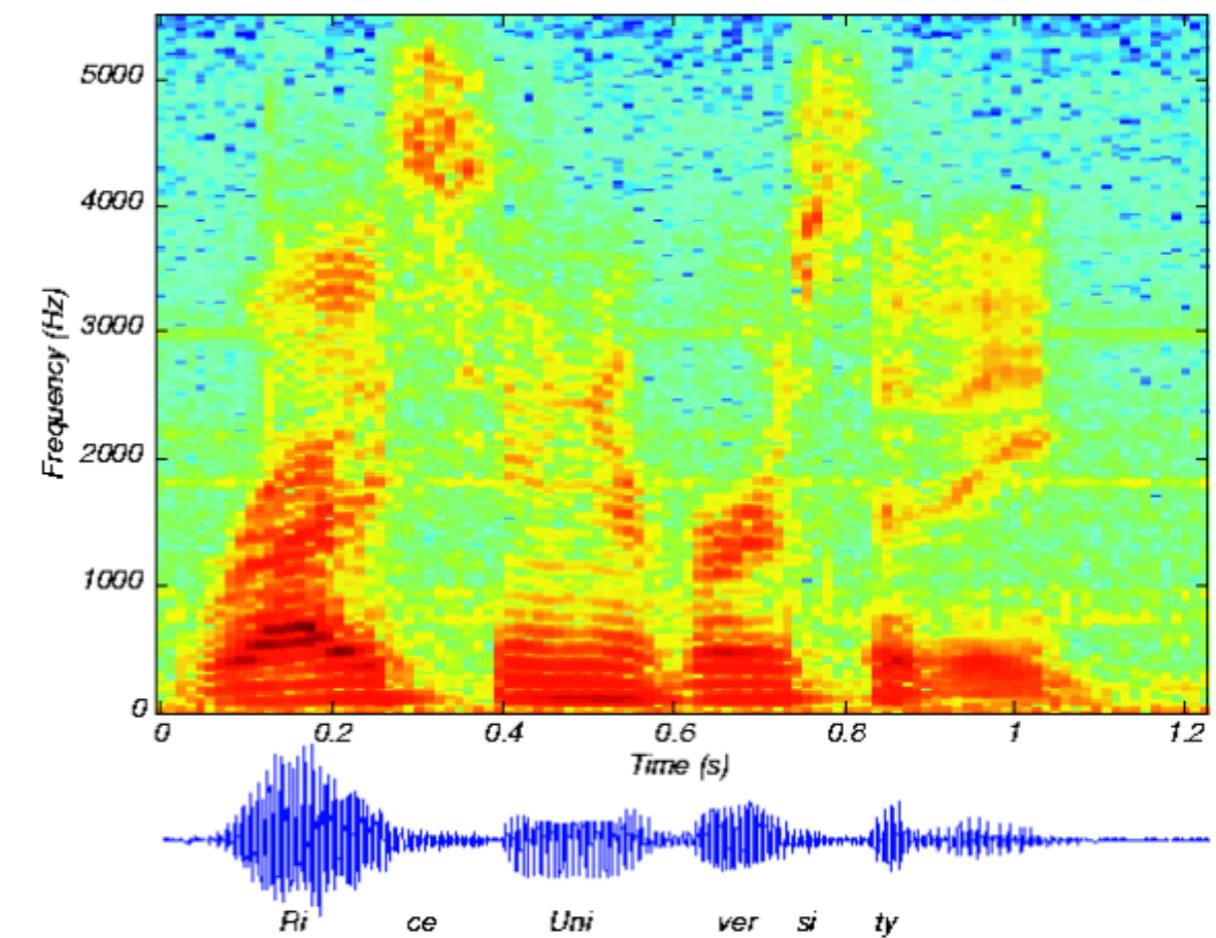
Image credit: [http://www.princeton.edu/~yc5/ele538b\\_sparsity/lectures/sparse\\_representation.pdf](http://www.princeton.edu/~yc5/ele538b_sparsity/lectures/sparse_representation.pdf)

# Why sparse?

Many signals are sparse in some transform domain.



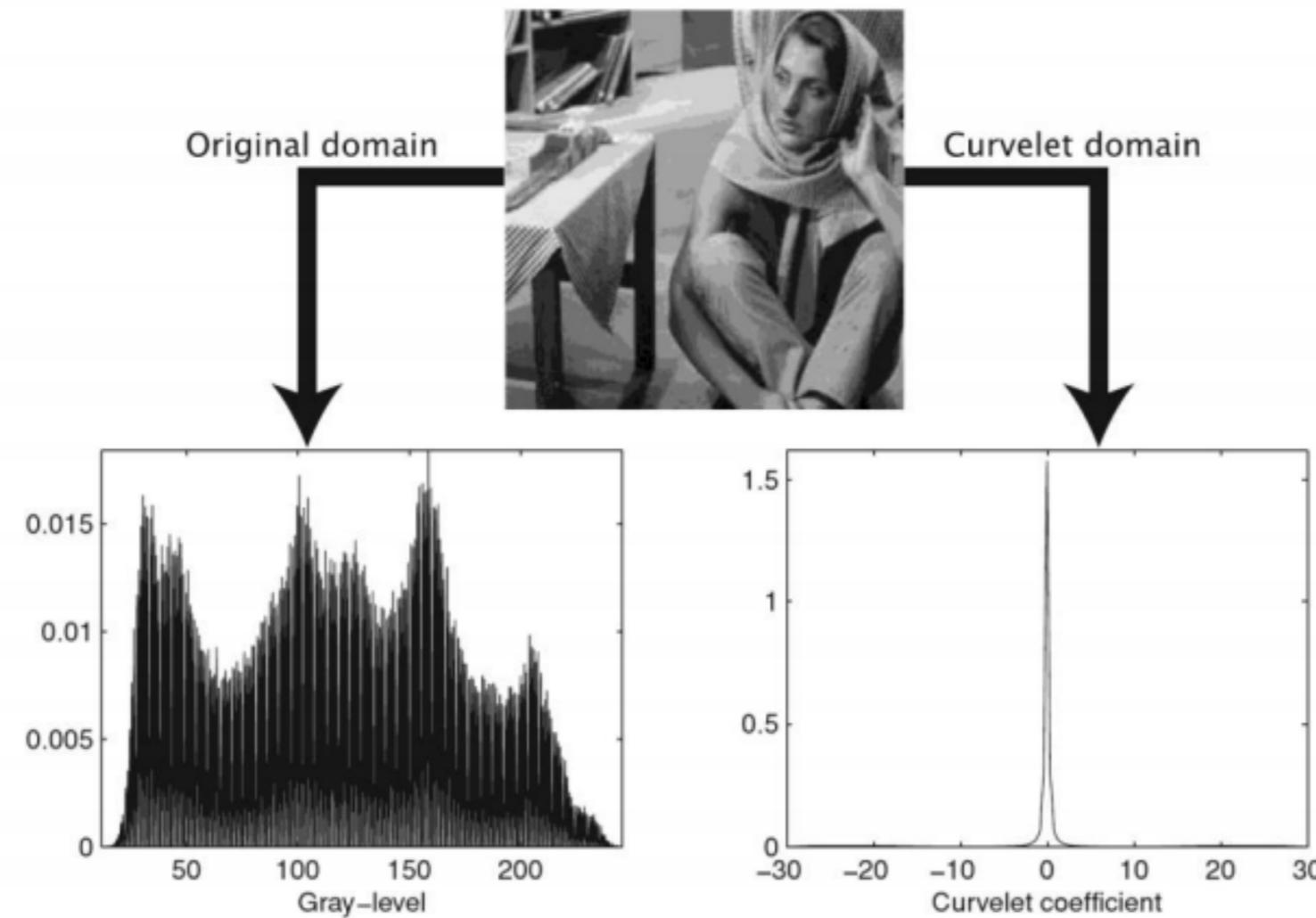
Musical instrument spectrum



Speech spectrogram

# Why sparse?

Many signals are sparse in some transform domain.



Natural image

Image credit: Zaouali, Bouzidi, and Zagrouba: Review of multiscale geometric decompositions in a remote sensing context

# Dictionary learning

Note that

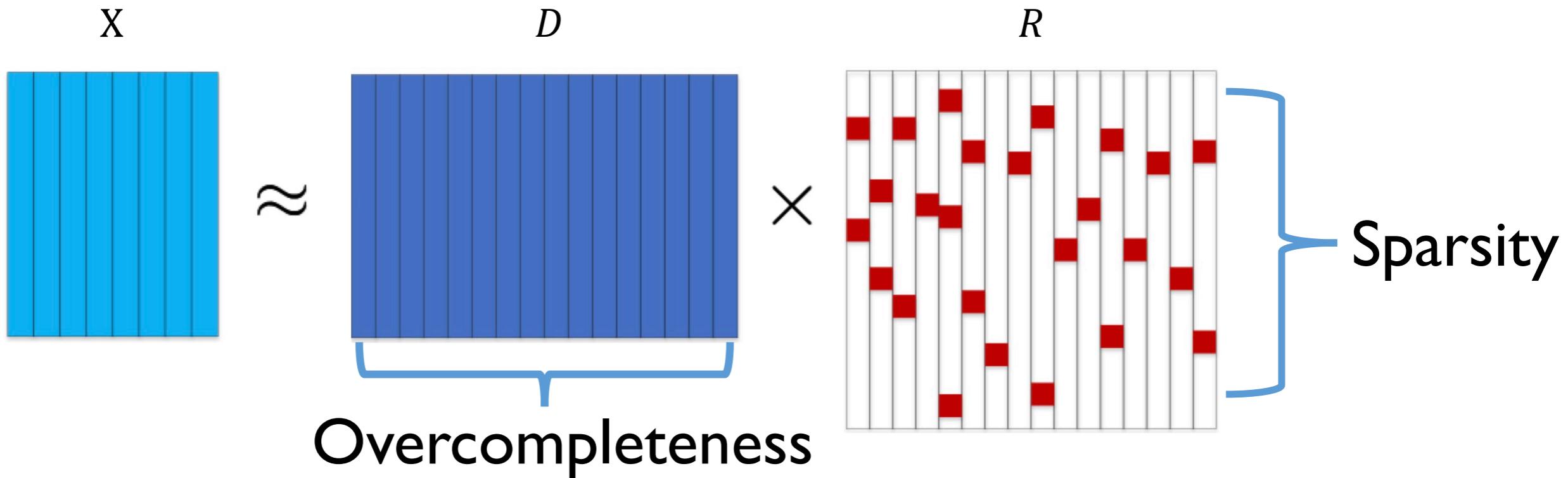
$$\arg \min_{D \in \mathcal{D}, R \in \mathcal{R}} \|X - DR\|_F^2,$$

where  $\mathcal{D}$  and  $\mathcal{R}$  are some specific domains for  $D$  and  $R$ .

# Sparse coding

Note that

$$\arg \min_{D \in \mathcal{D}, R \in \mathcal{R}} \|X - DR\|_F^2$$



# $\ell_p$ norm

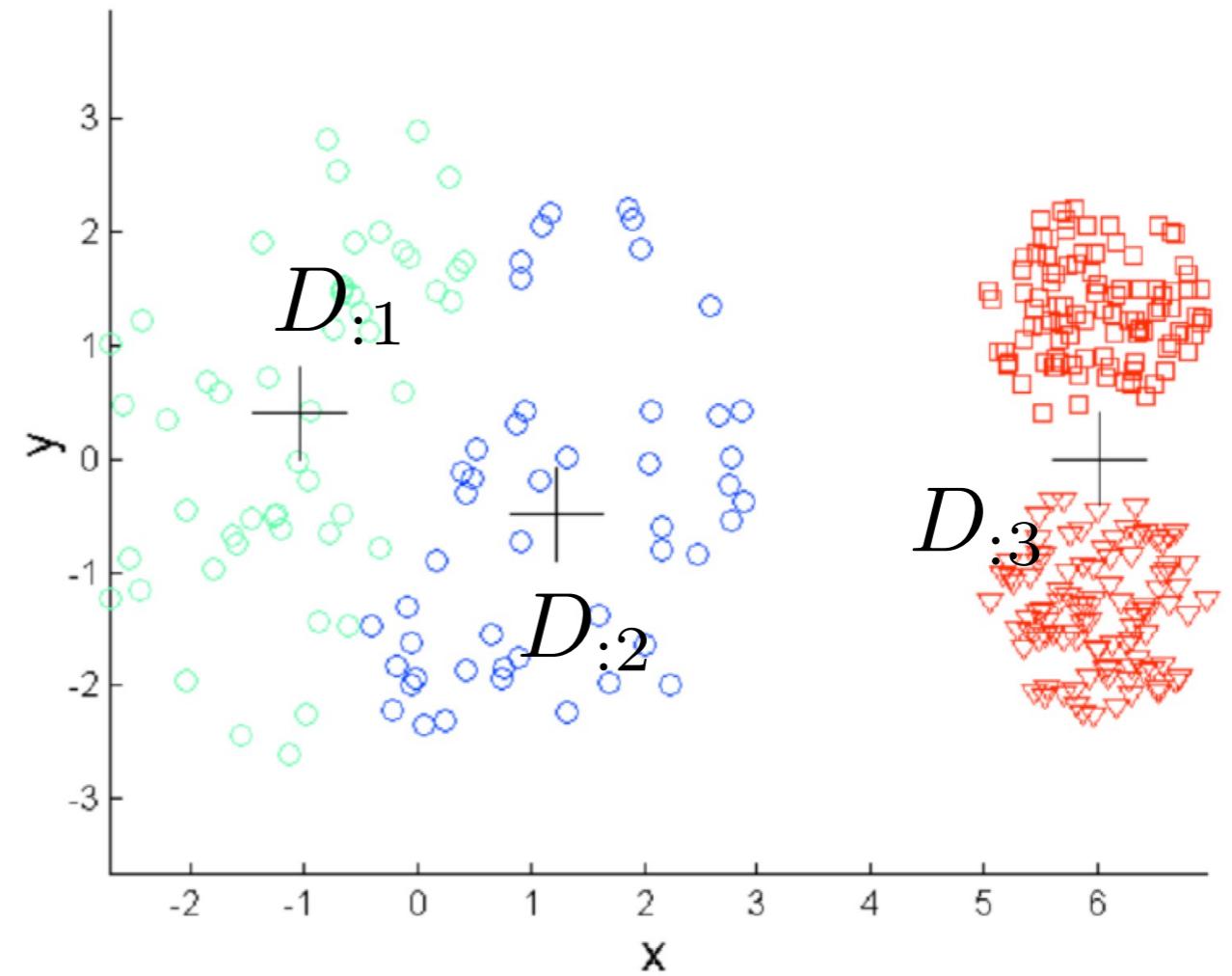
$\ell_p$  norm:  $\|\alpha\|_p = \left( \sum_{j=1}^k |\alpha_j|^p \right)^{1/p}$ , where  $\alpha \in \mathbb{R}^k$ .

In other words,  $\|\alpha\|_p^p = \sum_{j=1}^k |\alpha_j|^p$ .

# K-means

K-means clustering:

$$\min_{D \in \mathcal{D}, R \in \mathcal{R}} \|X - DR\|_F^2$$



Special requirement: each column of R is an one-hot vector, i.e.,  $\|R_i\|_0 = 1$  &  $\|R_i\|_1 = 1$ .

# K-SVD

K-SVD:

$$\min_{D \in \mathcal{D}, R \in \mathcal{R}} \|X - DR\|^2$$

**Special requirement:** each column of R is a sparse vector,  
i.e.,  $\|R_i\|_0 \leq k' \ll k$ .

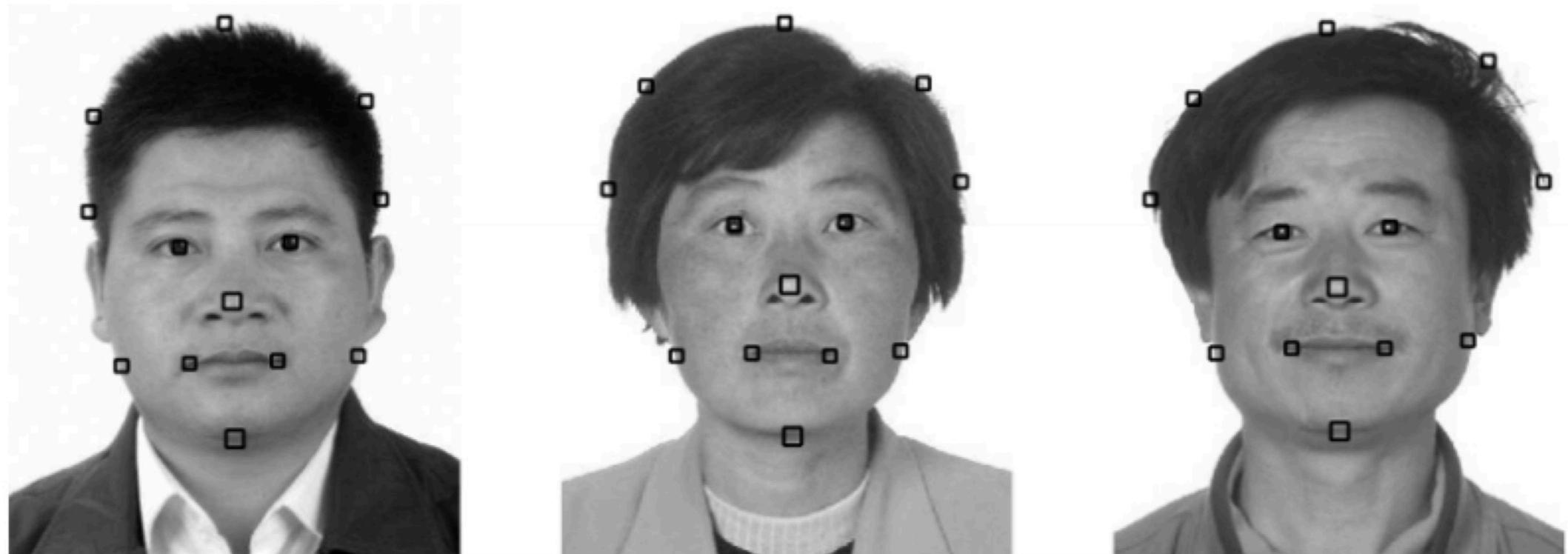
# Sparse coding applications

Image Compression: Results for 820 bytes per image



Image credit: Compression of facial images using the K-SVD algorithm, J. Vis. Comun. Image Represent.

# Data pre-processing for K-SVD



Facial feature points.

Image credit: CLow Bit-Rate Compression of Facial Images, IEEE Transactions on Image Processing

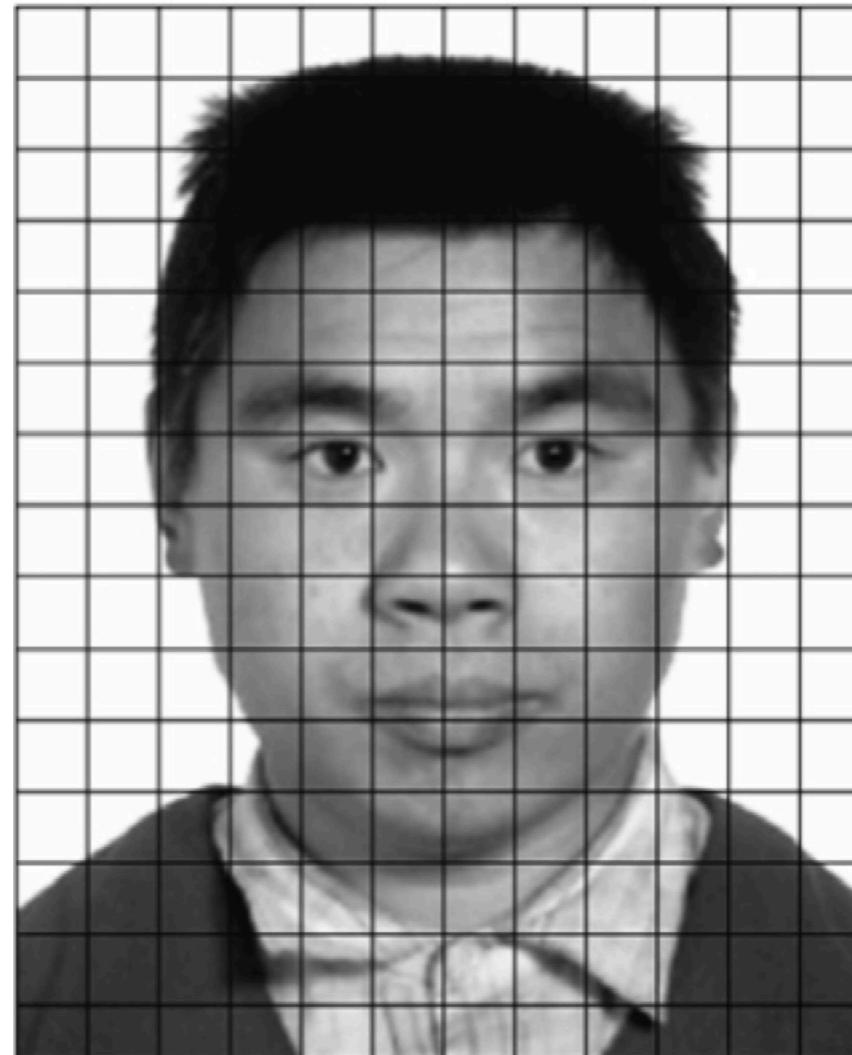
# Data pre-processing for K-SVD



(Top) Input images and their (canonical) aligned (bottom) versions.

Image credit: CLow Bit-Rate Compression of Facial Images, IEEE Transactions on Image Processing

# Data pre-processing for K-SVD



**Uniform slicing**

Image credit: Compression of facial images using the K-SVD algorithm, J. Vis. Comun. Image Represent.

# Sparse coding applications

## Inpainting:

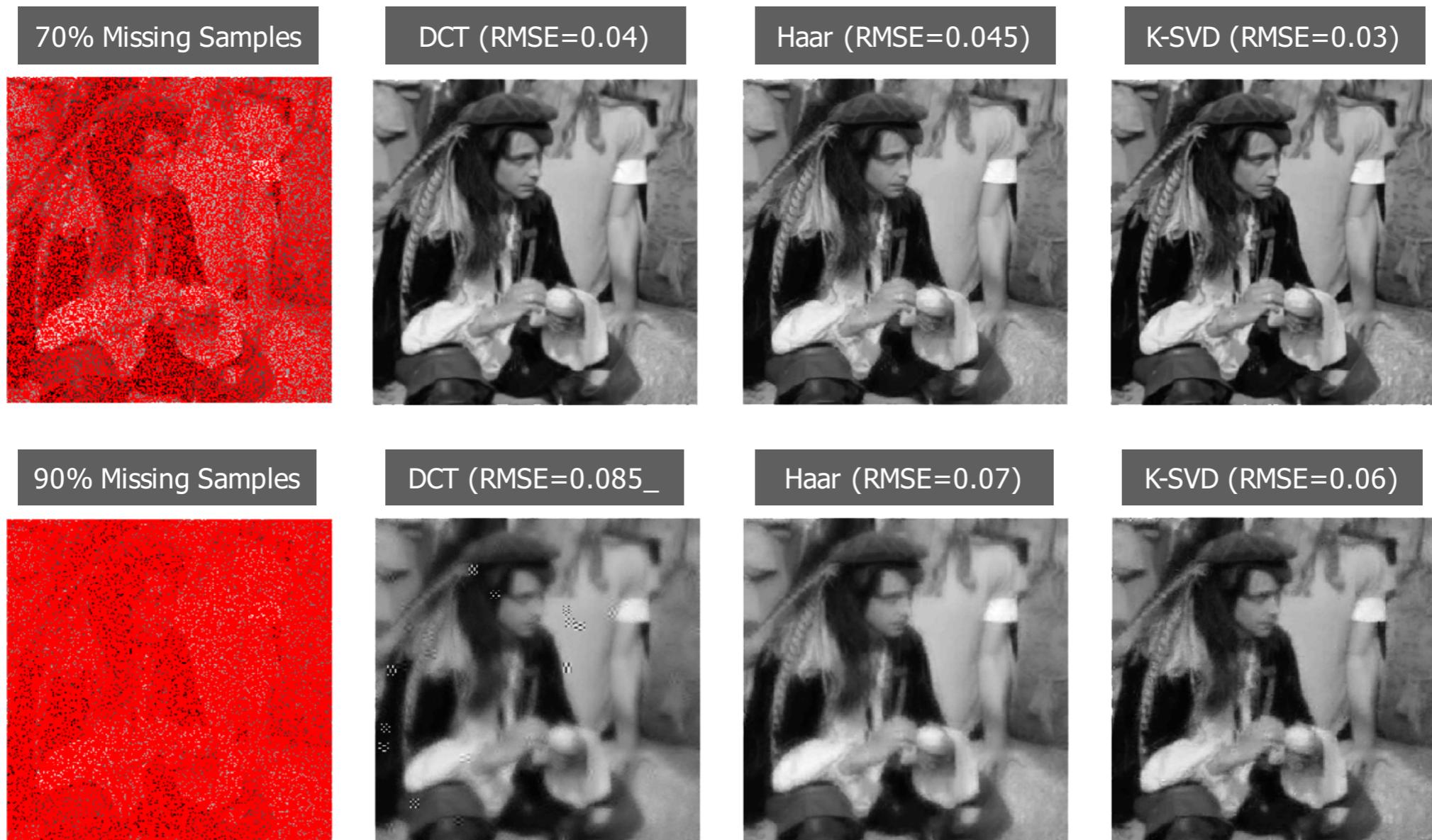


Image credit: <http://www.cs.technion.ac.il/~ronrubin/Talks/K-SVD.ppt>

# Sparse coding applications

Inpainting:



Image credit: <http://www.cs.technion.ac.il/~ronrubin/Talks/K-SVD.ppt>

# Sparse coding applications

Inpainting for text removal:



Image credit: Sparse representation for color image restoration. IEEE Transactions on Image Processing.

# Sparse coding applications

Inpainting for text removal:



Image credit: Sparse representation for color image restoration. IEEE Transactions on Image Processing.

# Measure of Sparsity

The objective function:

$$\min_{D \in \mathcal{D}, R \in \mathcal{R}} \|X - DR\|_F^2 + \lambda \psi(R)$$

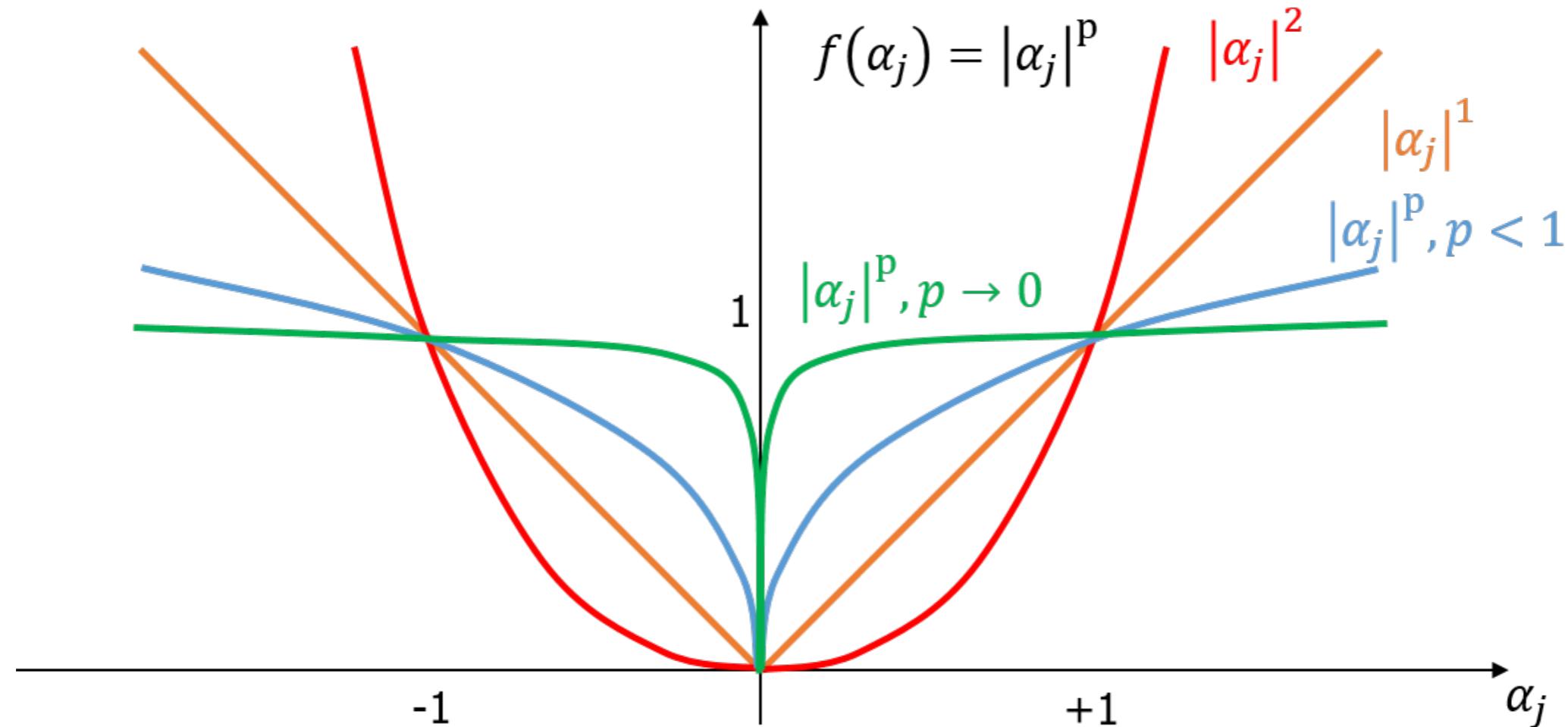

Data fitting

Sparse regularisation

Question: how can we design the regularisation to make  $R$  to be sparse?

# Measure of Sparsity: $\ell_0$ norm

$$\|\alpha\|_p^p = \sum_{j=1}^k |\alpha_j|^p.$$



As  $p \rightarrow 0$ , we get a count of the non-zeros in the vector.  
So we can employ  $\|\alpha\|_0$  to measure sparsity.

# Measure of Sparsity: $\ell_0$ norm

As  $p \rightarrow 0$ , we get a count of the non-zeros in the vector.  
So we can employ  $\|\alpha\|_0$  to measure sparsity.

However, the  $\ell_0$  minimisation is not easy. How to do?

# Measure of Sparsity $\ell_1$ norm

2D example (compared with  $\ell_2$ -norm):

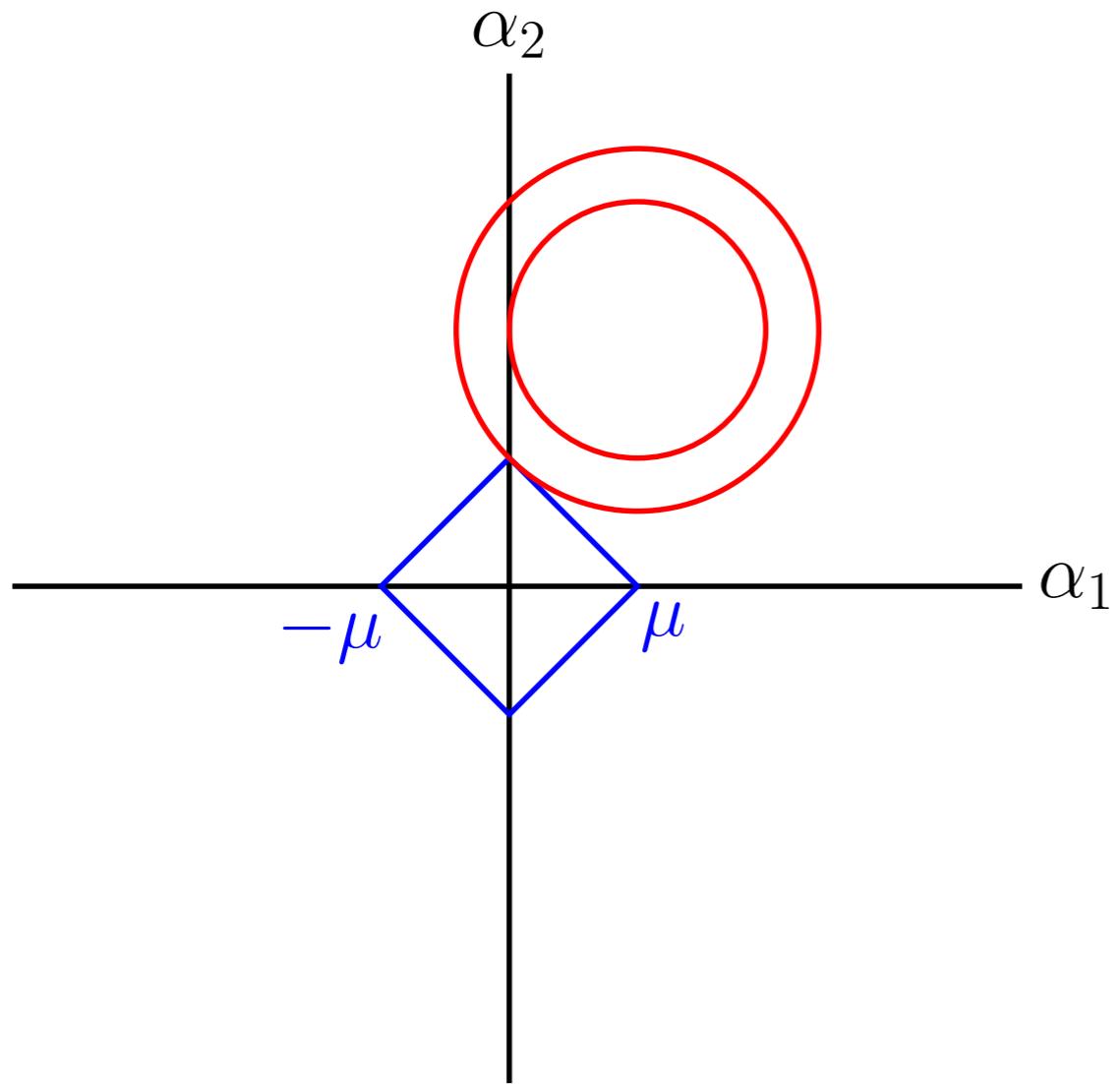
$$\min_{\alpha} \frac{1}{2} \|x - \alpha\|_2^2 \text{ s.t. } \|\alpha\|_1 \leq \mu$$

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

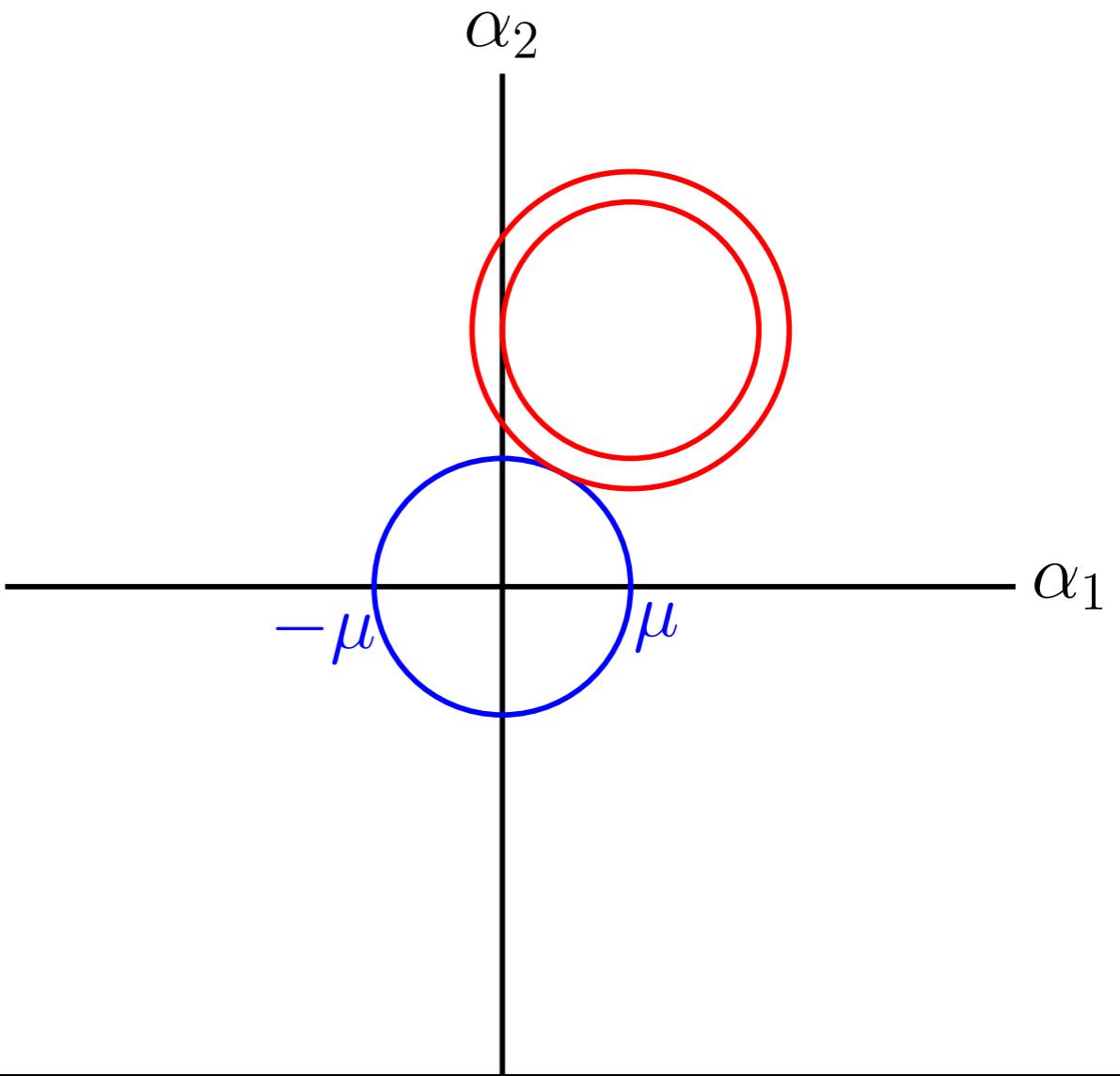
# Measure of Sparsity $\ell_1$ norm

2D example (compared with  $\ell_2$ -norm):

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



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



# Sparse coding learning algorithms

The  $\ell_0$  norm based approaches:

- $\min_{\alpha} \|X - D\alpha\|_F^2 \text{ s.t. } \forall i, \|\alpha\|_0 < L.$
- $\min_{\alpha} \|\alpha\|_0 \text{ s.t. } \|X - D\alpha\|_F^2 \leq \epsilon.$

Greedy algorithms:

- OMP (Y. Pati, et al. 1993; J. Tropp 2004).
- Subspace pursuit (SP) (W. Dai and O. Milenkovic 2009),  
CosaMP (D. Needell and J. Tropp 2009).
- IHT (T. Blumensath and M. Davies 2009).

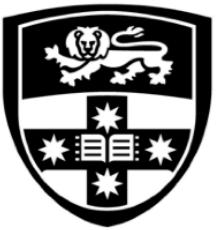
# Sparse coding learning algorithms

The  $\ell_1$  norm based approaches:

- $\min_{\alpha} \|\alpha\|_1 \text{ s.t. } \|X - D\alpha\|_F^2 \leq \epsilon.$
- $\min_{\alpha} \|X - D\alpha\|_F^2 + \lambda \|\alpha\|_1.$

Bayesian approach:

- Relevance vector machine (RVM) (M.Tipping 2001 ).
- Bayesian compressed sensing (BCS) (S.Ji, et al. 2008).



THE UNIVERSITY OF  
SYDNEY

# Regularisation and algorithmic stability



# No-Free-Lunch Theorem

- Sparse algorithms are not stable!
- A learning algorithm is said to be stable if slight perturbations in the training data result in small changes in the output of the algorithm, and these changes vanish as the data set grows bigger and bigger.

Xu, H., Caramanis, C., & Mannor, S. (2012). Sparse algorithms are not stable: A no-free-lunch theorem. *IEEE transactions on pattern analysis and machine intelligence*, 34(1), 187-193.



# Algorithmic Stability

We have two different training samples:

$$S = \{(X_1, Y_1), \dots, (X_{i-1}, Y_{i-1}), (X_i, Y_i), (X_{i+1}, Y_{i+1}), \dots, (X_n, Y_n)\}$$

$$S^i = \{(X_1, Y_1), \dots, (X_{i-1}, Y_{i-1}), (X'_i, Y'_i), (X_{i+1}, Y_{i+1}), \dots, (X_n, Y_n)\}$$

They are different because of only one training example.

An algorithm is uniformly stable if for any example  $(X, Y)$

$$|\ell(X, Y, h_S) - \ell(X, Y, h_{S^i})| \leq \epsilon(n).$$

Note that  $\epsilon(n)$  will vanish as  $n$  goes to infinity.

# Generalisation error

$$\begin{aligned} R(h_S) - \min_{h \in H} R(h) &= R(h_S) - R(h^*) \\ &= R(h_S) - R_S(h_S) + R_S(h_S) - R_S(h^*) + R_S(h^*) - R(h^*) \\ &\leq R(h_S) - R_S(h_S) + R_S(h^*) - R(h^*) \\ &\leq |R(h_S) - R_S(h_S)| + |R(h^*) - R_S(h^*)| \\ &\leq \sup_{h \in H} |R(h) - R_S(h)| + \sup_{h \in H} |R(h) - R_S(h)| \\ &= 2 \sup_{h \in H} |R(h) - R_S(h)|. \end{aligned}$$

$$R(h_S) - R_S(h_S) \leq \sup_{h \in H} |R(h) - R_S(h)|.$$



# Algorithmic Stability

A good stable algorithm will have a good generalisation ability:

$$\mathbb{E}[R(h_S) - R_S(h_S)]$$

$$= \mathbb{E}_S \left[ \mathbb{E}_{X,Y}[\ell(X, Y, h_S)] - \frac{1}{n} \sum_{i=1}^n \ell(X_i, Y_i, h_S) \right]$$

$$= \mathbb{E}_S \left[ \mathbb{E}_{S'} \left[ \frac{1}{n} \sum_{i=1}^n \ell(X'_i, Y'_i, h_S) \right] - \frac{1}{n} \sum_{i=1}^n \ell(X_i, Y_i, h_S) \right]$$

$$= \mathbb{E}_{S,S'} \left[ \frac{1}{n} \sum_{i=1}^n (\ell(X'_i, Y'_i, h_S) - \ell(X_i, Y_i, h_S)) \right]$$

$$= \mathbb{E}_{S,S'} \left[ \frac{1}{n} \sum_{i=1}^n (\ell(X'_i, Y'_i, h_S) - \ell(X'_i, Y'_i, h_{S^i})) \right]$$

$$\leq \epsilon'(n)$$

$\epsilon'(n)$  would be small if the learning algorithm is stable (because  $h_S$  and  $h_{S^i}$  are similar). This implies that stable algorithms will have small expected generalisation errors.

where  $S' = \{(X'_1, Y'_1), \dots, (X'_n, Y'_n)\}$ .



THE UNIVERSITY OF  
SYDNEY

# Regularisation and Stability

$\ell_2$  norm regularisation will make learning algorithms stable if the employed surrogate loss function is convex.

$$h_S = \arg \min_{h \in H} \frac{1}{n} \sum_{i=1}^n \ell(X_i, Y_i, h) + \lambda \|h\|_2^2$$

If the convex surrogate loss function is  $L$ -Lipschitz continuous w.r.t.  $h$ ,  $\|X\|_2 \leq B$ , we have

$$|\ell(X, Y, h_S) - \ell(X, Y, h_{S^i})| \leq \frac{2L^2 B^2}{\lambda n}.$$

# Proof (optional)

A surrogate loss function is  $L$ -Lipschitz continuous w.r.t.  $h$  if  
for any  $(X, Y)$  in the domain and any  $h, h' \in H$

$$|\ell(X, Y, h) - \ell(X, Y, h')| \leq L|h(X) - h'(X)|.$$

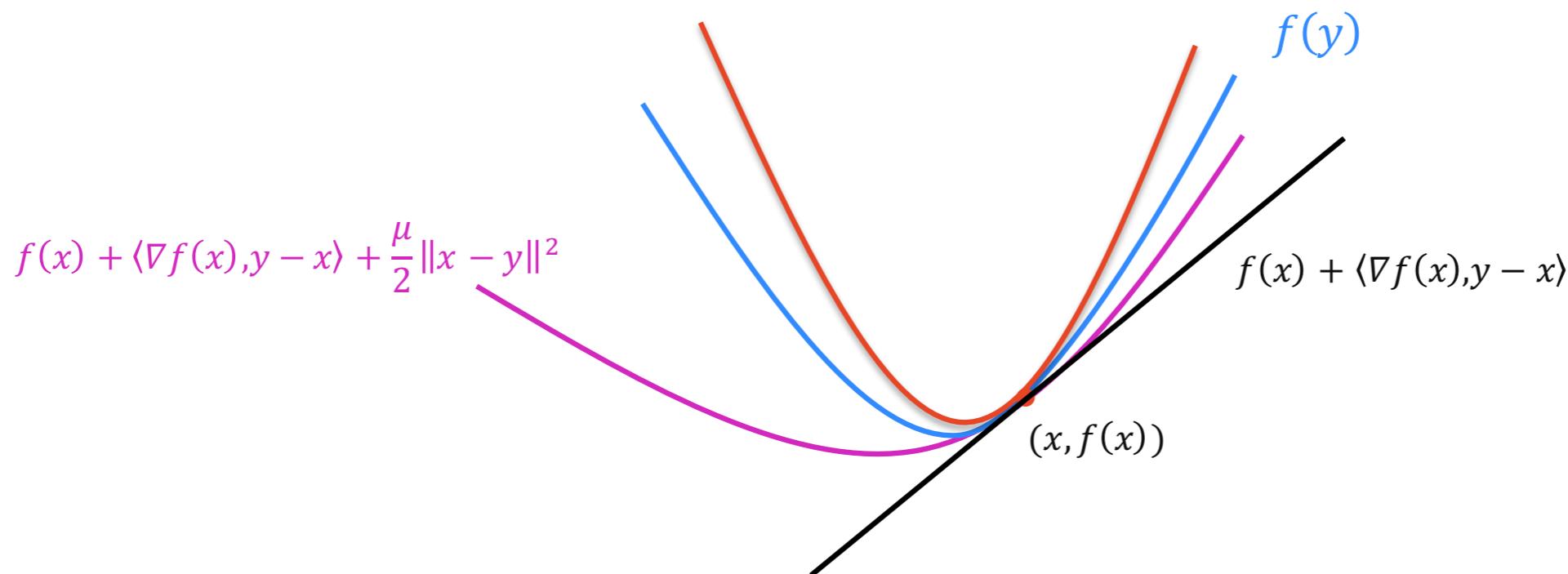
# Proof (optional)

A function is  $\mu$ -strongly convex:

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2} \|x - y\|^2, \forall x, y,$$

$\Leftrightarrow$

$$\mu I \leq \nabla^2 f(x), \forall x$$



# Proof (optional)

Note that the function  $\ell(X, Y, h) + \lambda\|h\|^2$  is always strongly convex w.r.t.  $h$ .

Let

$$R_{S,\lambda}(h) = \frac{1}{n} \sum_{i=1}^n \ell(X_i, Y_i, h) + \lambda\|h\|^2.$$

We have

$$R_{S,\lambda}(h_{S^i}) \geq R_{S,\lambda}(h_S) + \langle \nabla R_{S,\lambda}(h_S), h_{S^i} - h_S \rangle + \frac{\lambda}{2} \|h_{S^i} - h_S\|^2$$

$$= R_{S,\lambda}(h_S) + \frac{\lambda}{2} \|h_{S^i} - h_S\|^2$$

$$R_{S^i,\lambda}(h_S) \geq R_{S^i,\lambda}(h_{S^i}) + \langle \nabla R_{S^i,\lambda}(h_{S^i}), h_S - h_{S^i} \rangle + \frac{\lambda}{2} \|h_S - h_{S^i}\|^2$$

$$= R_{S^i,\lambda}(h_{S^i}) + \frac{\lambda}{2} \|h_S - h_{S^i}\|^2.$$

# Proof (optional)

We have

$$\begin{aligned}\lambda \|h_{S^i} - h_S\|^2 &\leq R_{S^i, \lambda}(h_S) - R_{S, \lambda}(h_S) + R_{S, \lambda}(h_{S^i}) - R_{S^i, \lambda}(h_{S^i}) \\ &\leq \frac{1}{n} |\ell(X'_i, Y'_i, h_S) - \ell(X'_i, Y'_i, h_{S^i})| + \frac{1}{n} |\ell(X_i, Y_i, h_S) - \ell(X_i, Y_i, h_{S^i})| \\ &\leq \frac{L}{n} |h_S(X'_i) - h_{S^i}(X'_i)| + \frac{L}{n} |h_S(X_i) - h_{S^i}(X_i)|\end{aligned}$$

aumme  $h(X) = \langle h, X \rangle$

$$\begin{aligned}&= \frac{L}{n} |\langle h_S - h_{S^i}, X'_i \rangle| + \frac{L}{n} |\langle h_S - h_{S^i}, X_i \rangle| \\ &\leq \frac{L}{n} \|h_S - h_{S^i}\| \|X'_i\| + \frac{L}{n} \|h_S - h_{S^i}\| \|X_i\| \\ &\leq \frac{2LB}{n} \|h_S - h_{S^i}\|,\end{aligned}$$

where the fifth inequality holds because of Cauchy-Schwartz inequality:  $\langle a, b \rangle \leq \|a\| \|b\|$ .

# Proof (optional)

Thus,

$$\|h_{S^i} - h_S\| \leq \frac{2LB}{\lambda n}.$$

We then have

$$\begin{aligned} |\ell(X, Y, h_S) - \ell(X, Y, h_{S^i})| &\leq L|h_S(X) - h_{S^i}(X)| \\ &\leq L\|h_S - h_{S^i}\| \|X\| \\ &\leq \frac{2L^2B^2}{\lambda n}. \end{aligned}$$