



Sparse dictionary learning in the presence of noise & outliers

Rémi Gribonval

INRIA Rennes - Bretagne Atlantique, France

remi.gribonval@inria.fr



projection, learning and sparsity for efficient data processing

Overview

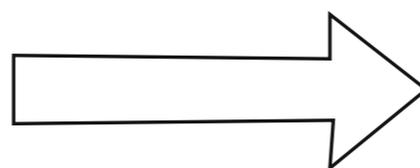
- Context: sparse signal processing
- Dictionary learning
- Statistical guarantees
- Flavor of the proof
- Conclusion

Sparse signal processing

Sparse Signal / Image Processing

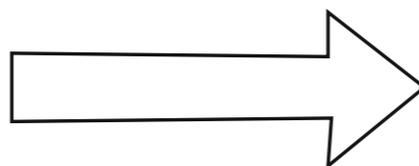


denoising



(Reuters) - Eating bacon, sausage, hot dogs and other
Eating unprocessed beef, pork or lamb appeared not t
The study, an analysis of other research called a meta
"To lower risk of heart attacks and diabetes, people sh
"Processed meats such as bacon, salami, sausages, ho
Based on her findings, she said people who eat one ser
The American Meat Institute objected to the findings, s
"At best, this hypothesis merits further study. It is cer
Most dietary guidelines recommend eating less meat.
But studies rarely look for differences in risk between
She and colleagues did a systematic review of nearly 1
They defined processed meat as any meat preserved b

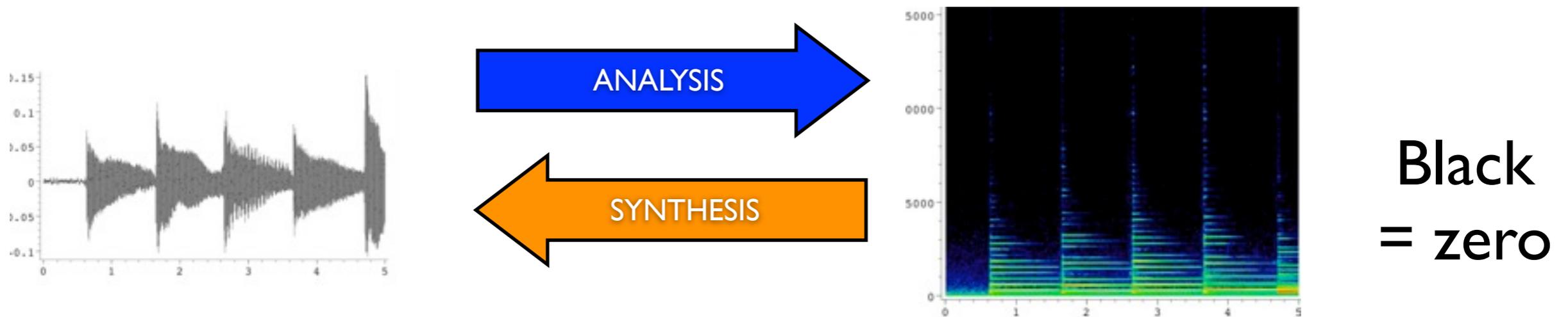
inpainting



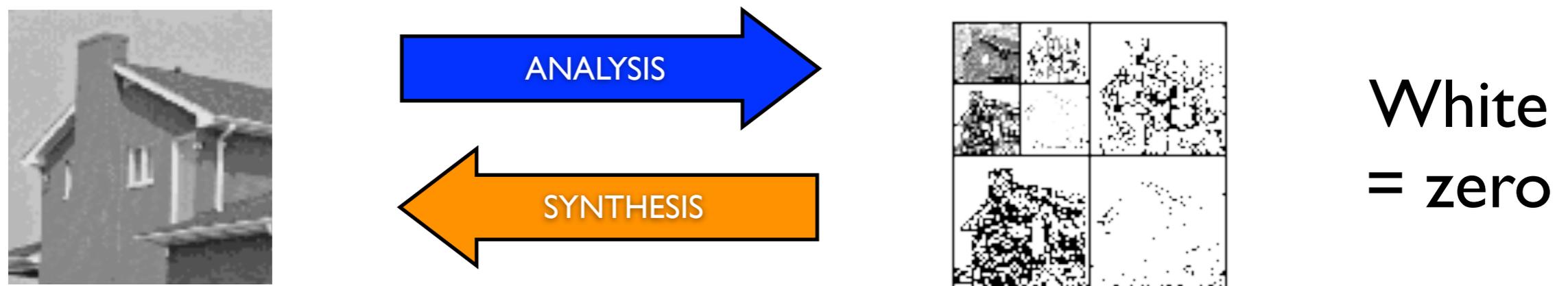
*+ Compression,
Source Localization, Separation,
Compressed Sensing ...*

Typical Sparse Models

- Audio : time-frequency representations (MP3)



- Images : wavelet transform (JPEG2000)



Mathematical expression

- Signal / image = high dimensional vector

$$\mathbf{x} \in \mathbb{R}^d$$

- **Model** = linear combination of basis vectors
(ex: *time-frequency atoms, wavelets*)

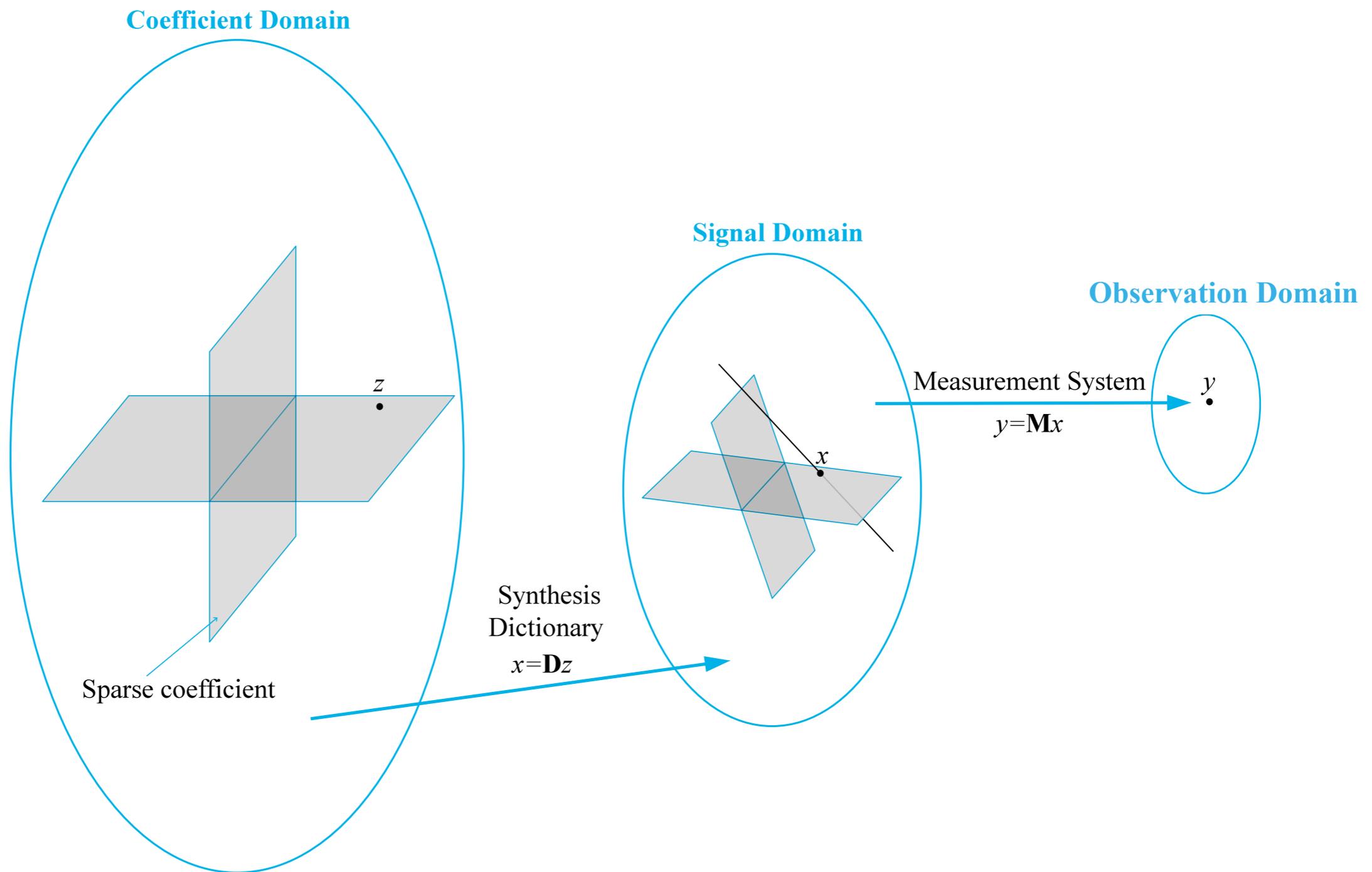
$$\mathbf{x} \approx \sum_k z_k \mathbf{d}_k = \mathbf{Dz}$$

*Dictionary of atoms
(Mallat & Zhang 93)*

- **Sparsity** = small L0 (quasi)-norm

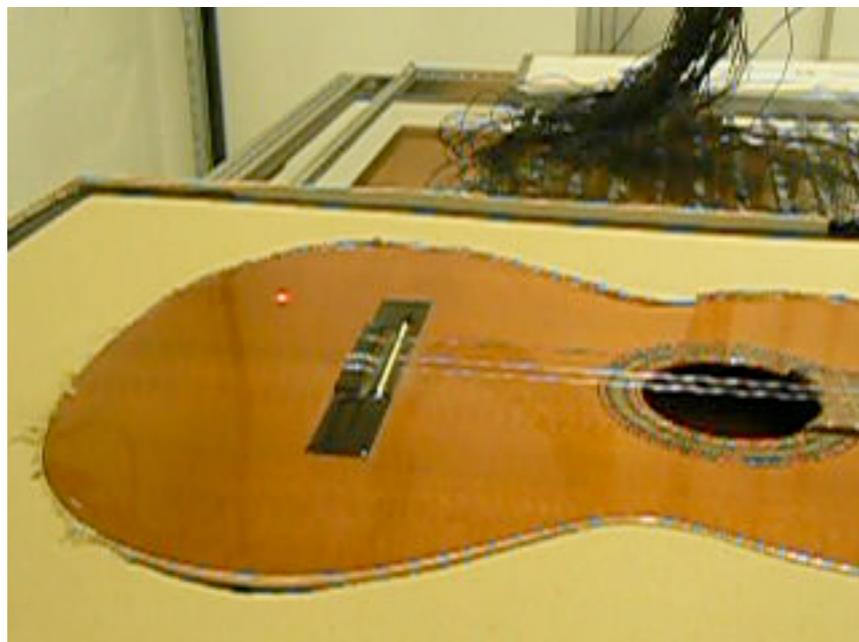
$$\|\mathbf{z}\|_0 = \sum_k |z_k|^0 = \text{card}\{k, z_k \neq 0\}$$

Sparse models and inverse problems



Acoustic Imaging

- Ground truth: laser vibrometry
 - ✓ direct optical measures
 - ✓ sequential
 - ✓ 2000 measures



- Nearfield Acoustic Holography
 - ✓ indirect acoustic measures
 - ✓ 120 microphones at a time
 - ✓ $120 \times 16 = 1920$ measures
 - ✓ *Tikhonov regularization*



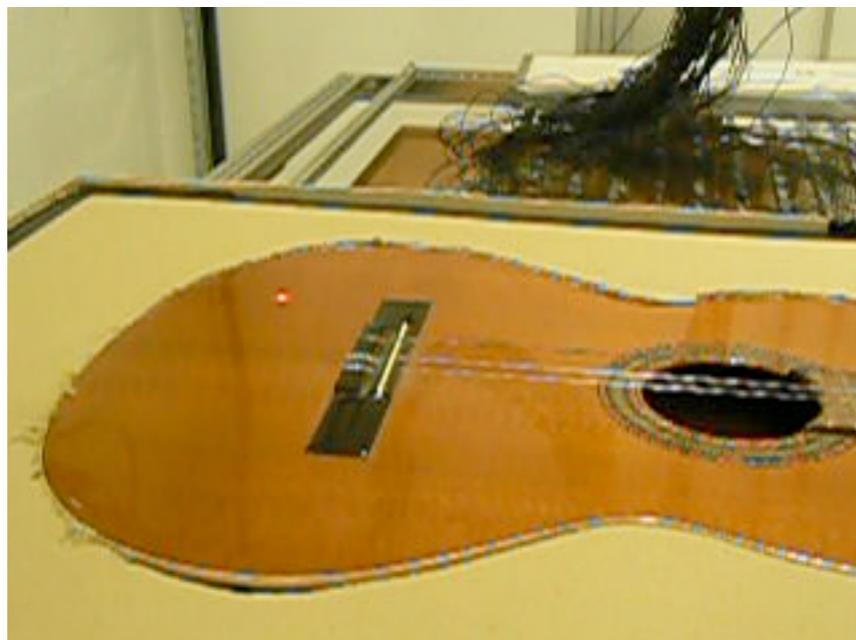
echange.inria.fr

echange

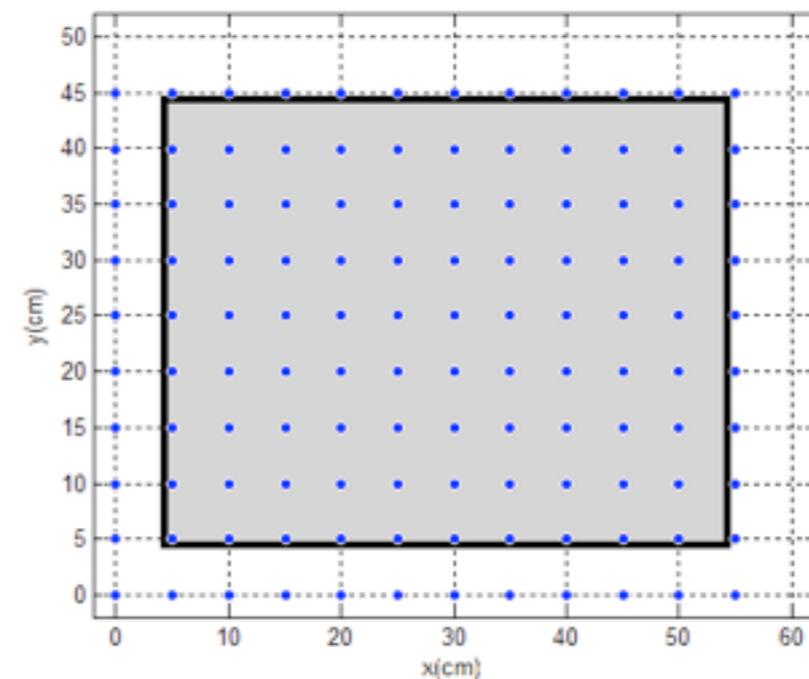
SUPPORTED BY
ANR

Acoustic Imaging

- Ground truth: laser vibrometry
 - ✓ direct optical measures
 - ✓ sequential
 - ✓ 2000 measures



- Nearfield Acoustic Holography
 - ✓ indirect acoustic measures
 - ✓ 120 microphones at a time
 - ✓ $120 \times 16 = 1920$ measures
 - ✓ *Tikhonov regularization*



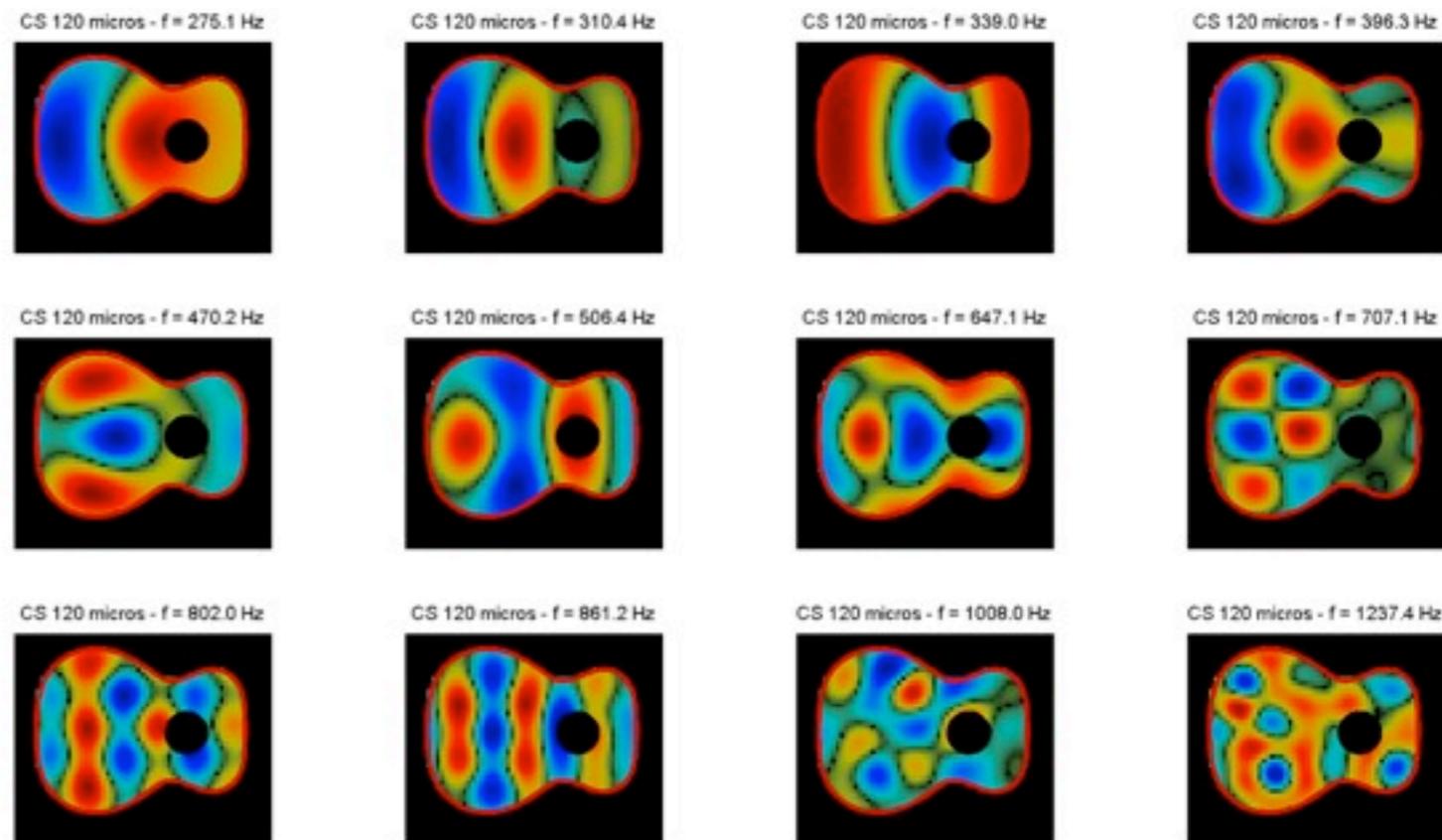
echange.inria.fr

echange

SUPPORTED BY
ANR

Compressive Nearfield Acoustic Holography

- One shot with 120 micros
- Sparse regularization



echange.inria.fr

echange

SUPPORTED BY
ANR

Dictionary learning

with K. Schnass, F. Bach, R. Jenatton



small-project.eu



Sparse Atomic Decompositions

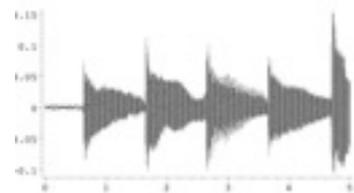
$$\mathbf{x} \approx \mathbf{Dz}$$

Signal
Image

(Overcomplete)
dictionary of atoms

Sparse
Representation
Coefficients

Data Deluge



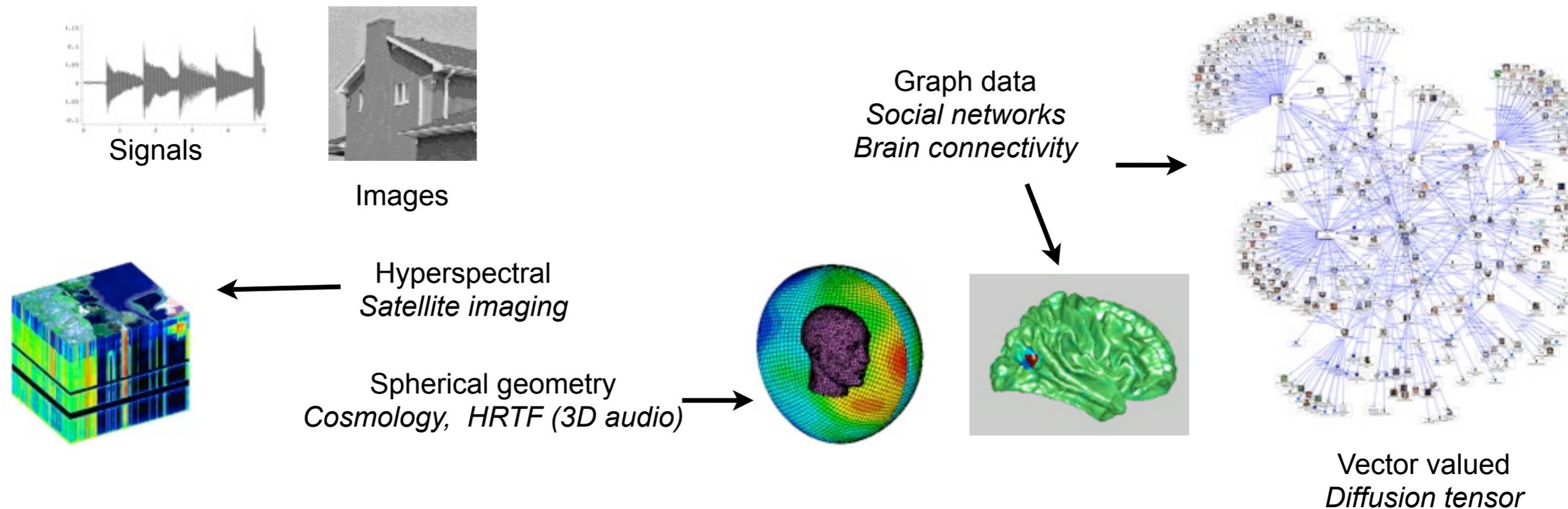
Signals



Images

Data Deluge + Jungle

- Sparsity: historically for signals & images
 - ✓ bottleneck = **large-scale** algorithms



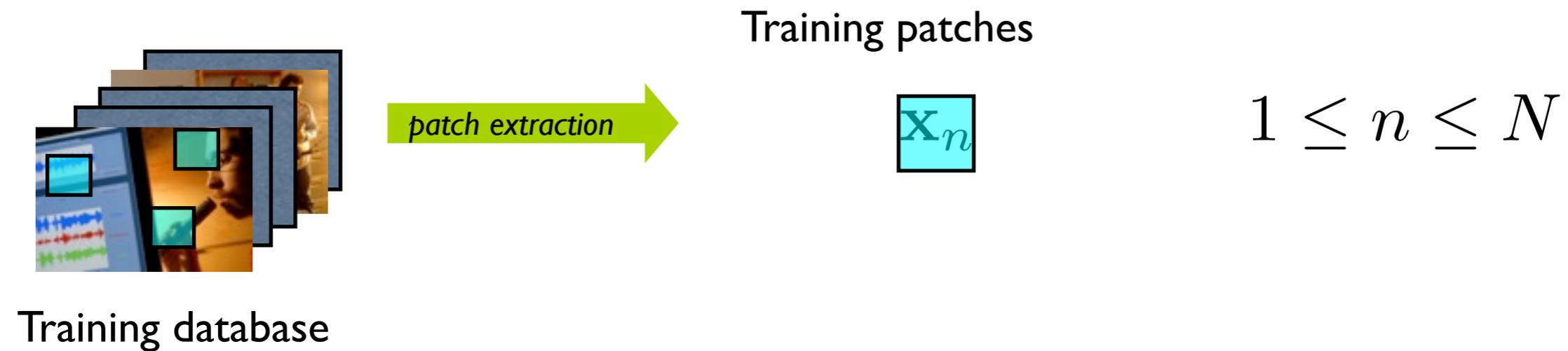
- New “exotic” or composite data
 - ✓ bottleneck = **dictionary/operator design/learning**

A quest for the perfect sparse model



Training database

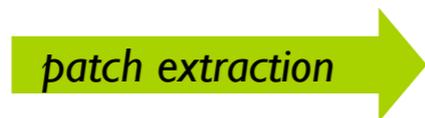
A quest for the perfect sparse model



A quest for the perfect sparse model



Training database



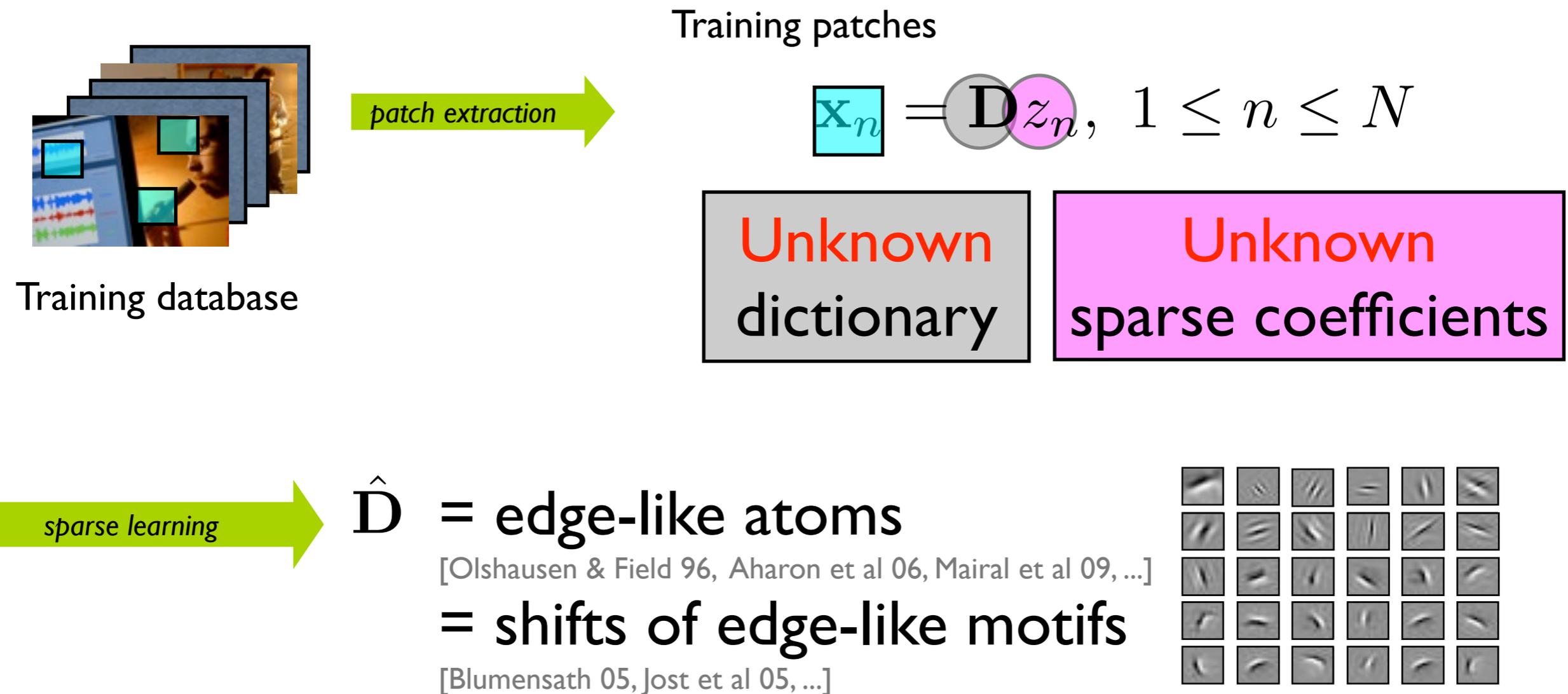
Training patches

$$\mathbf{x}_n = \mathbf{D} \mathbf{z}_n, \quad 1 \leq n \leq N$$

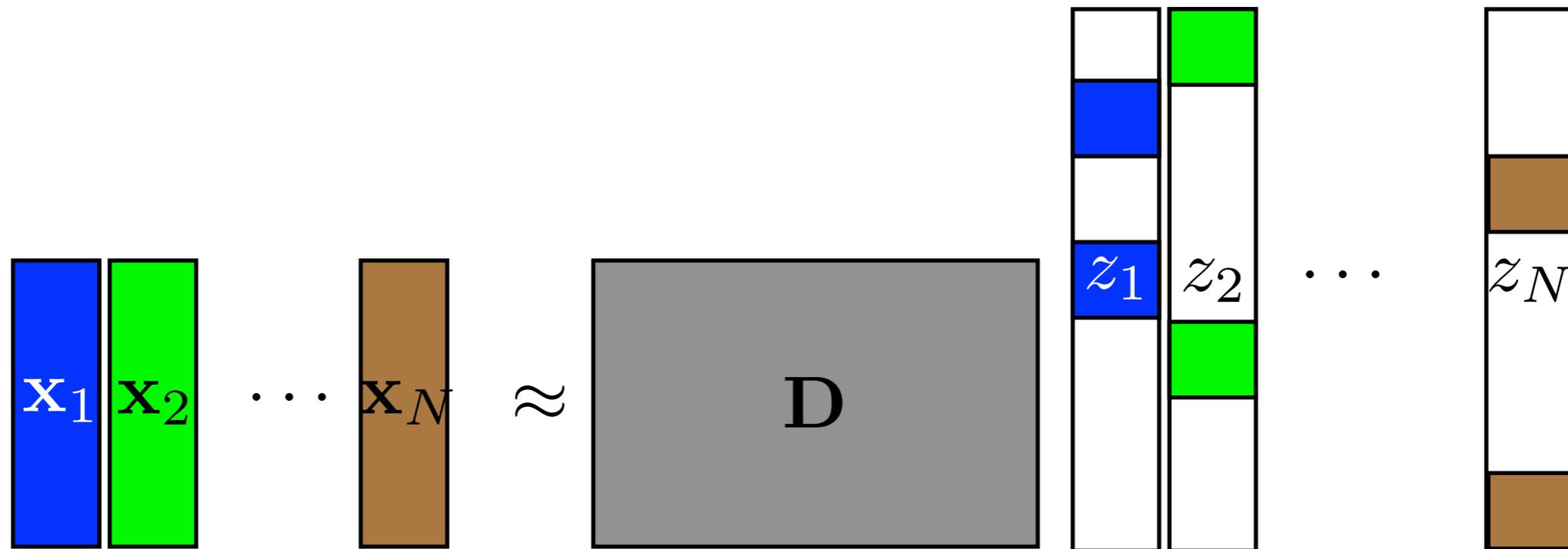
Unknown
dictionary

Unknown
sparse coefficients

A quest for the perfect sparse model



Dictionary Learning = Sparse Matrix Factorization



$$\mathbf{X} \approx \mathbf{D} \mathbf{Z}$$

$d \times N$ $d \times K$ $K \times N$ with s-sparse columns

Many approaches

- Independent component analysis
 - ◆ [see e.g. book by Comon & Jutten 2011]
- Convex
 - ◆ [Bach et al., 2008; Bradley and Bagnell, 2009]
- Submodular
 - ◆ [Krause and Cevher, 2010]
- Bayesian
 - ◆ [Zhou et al., 2009]
- **Non-convex matrix-factorization**
 - ◆ [Olshausen and Field, 1997; Pearlmutter & Zibulevsky 2001, Aharon et al. 2006; Lee et al., 2007; Mairal et al., 2010 (... and many other authors)]

Sparse coding objective function

- **Given one training sample: Basis Pursuit / LASSO**

$$f_{\mathbf{x}_n}(\mathbf{D}) = \min_{z_n} \frac{1}{2} \|\mathbf{x}_n - \mathbf{D}z_n\|_2^2 + \lambda \|z_n\|_1$$

- **Given N training samples**

$$F_{\mathbf{X}}(\mathbf{D}) = \frac{1}{N} \sum_{n=1}^N f_{\mathbf{x}_n}(\mathbf{D})$$

$$\propto \min_Z \frac{1}{2} \|\mathbf{X} - \mathbf{D}Z\|_F^2 + \lambda \|Z\|_1$$

Learning = constrained minimization

$$\hat{\mathbf{D}} = \arg \min_{\mathbf{D} \in \mathcal{D}} F_{\mathbf{X}}(\mathbf{D})$$

- ✓ Online learning with SPAMS library (Mairal & al)
- ✓ Constraint = dictionary with unit columns

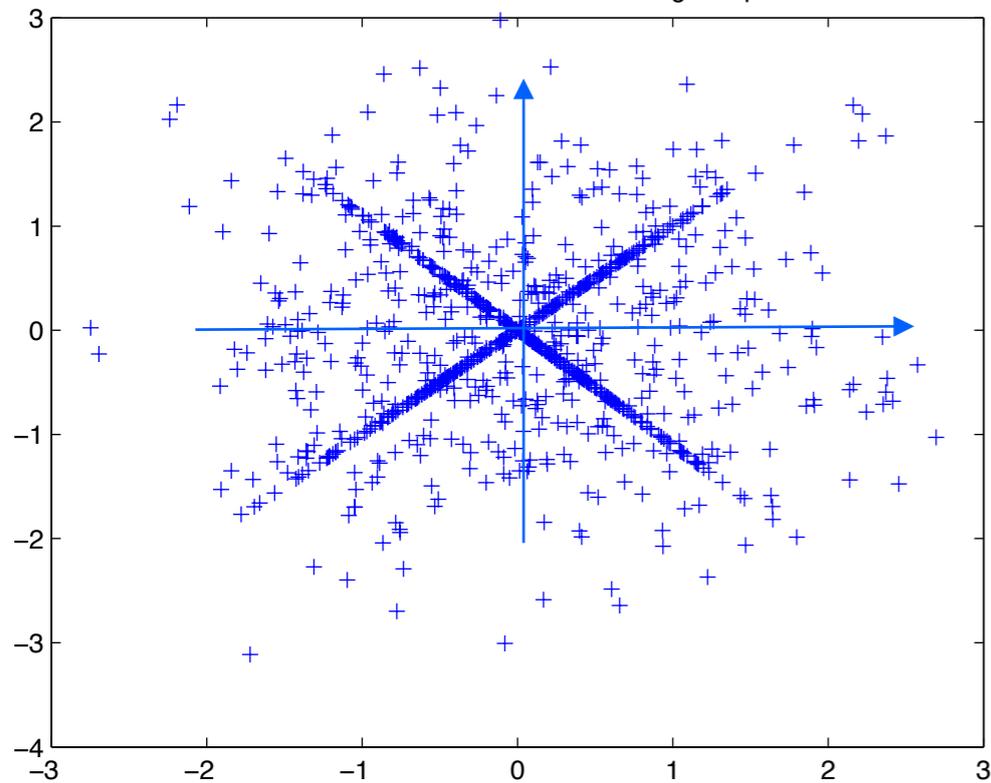
$$\mathcal{D} = \{ \mathbf{D} = [\mathbf{d}_1, \dots, \mathbf{d}_D], \forall k \|\mathbf{d}_k\|_2 = 1 \}$$

Empirical findings

Numerical example (2D)

$$\mathbf{X} = \mathbf{D}_0 \mathbf{Z}_0$$

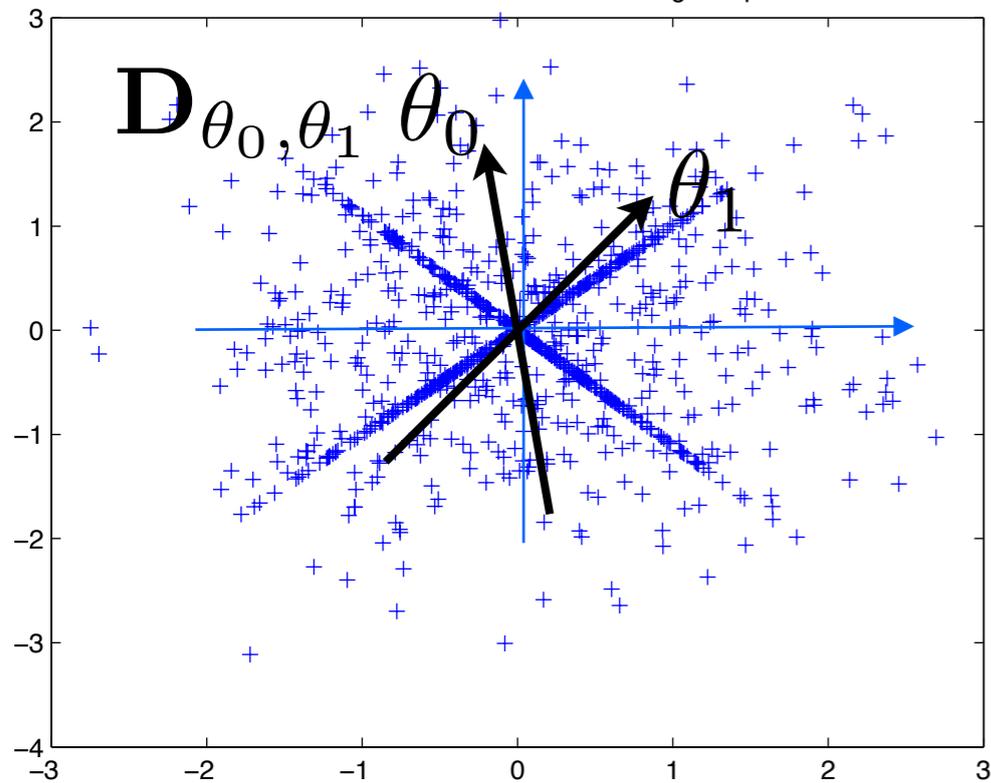
N = 1000 Bernoulli-Gaussian training samples



Numerical example (2D)

$$\mathbf{X} = \mathbf{D}_0 \mathbf{Z}_0$$

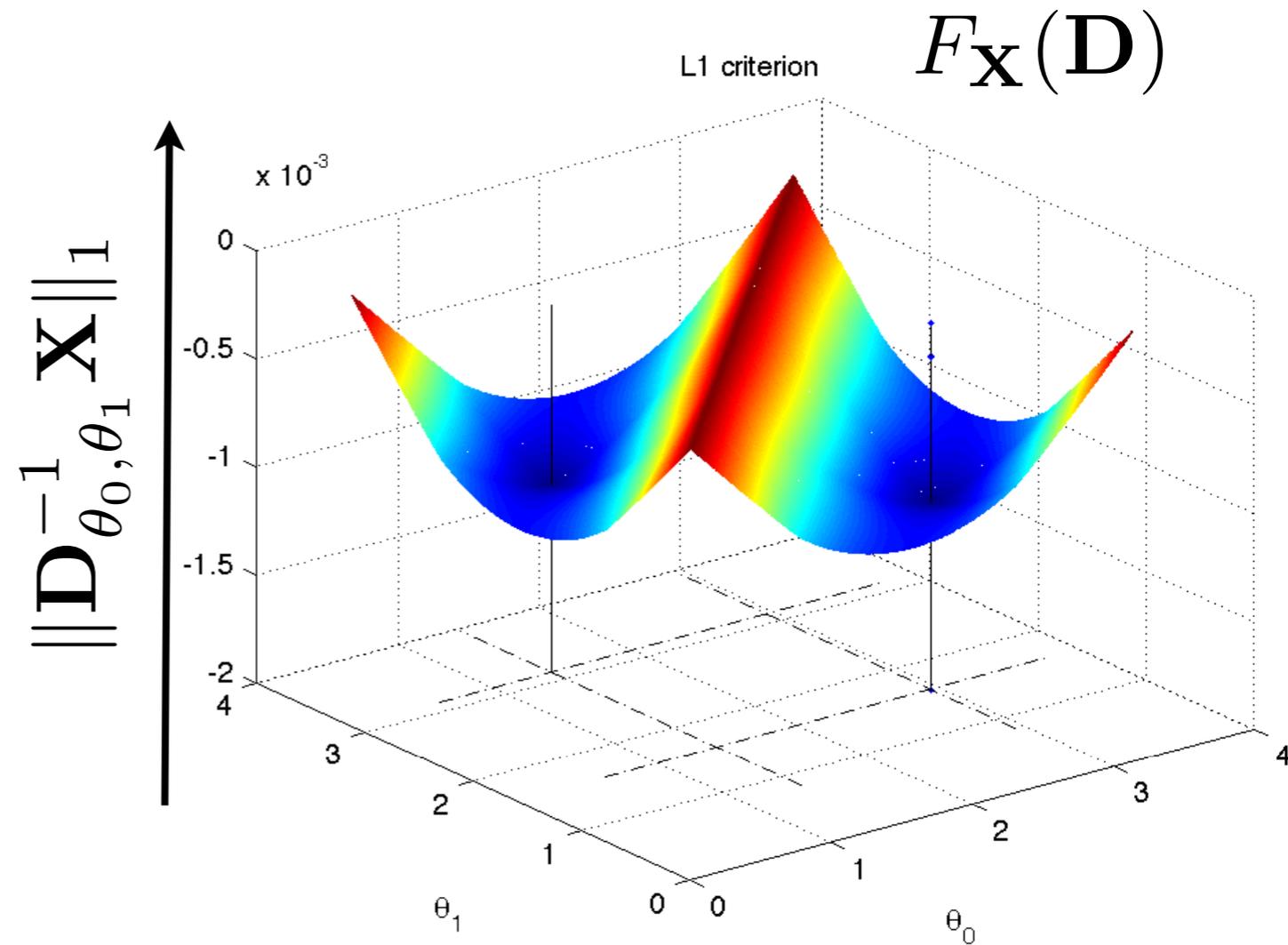
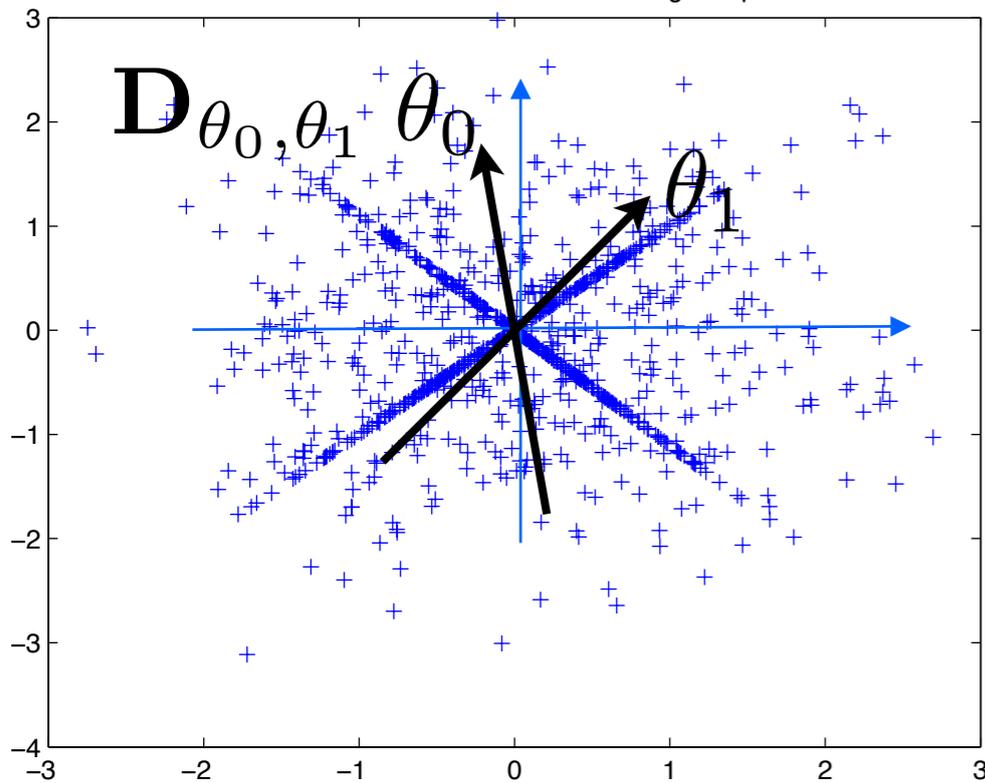
N = 1000 Bernoulli-Gaussian training samples



Numerical example (2D)

$$\mathbf{X} = \mathbf{D}_0 \mathbf{Z}_0$$

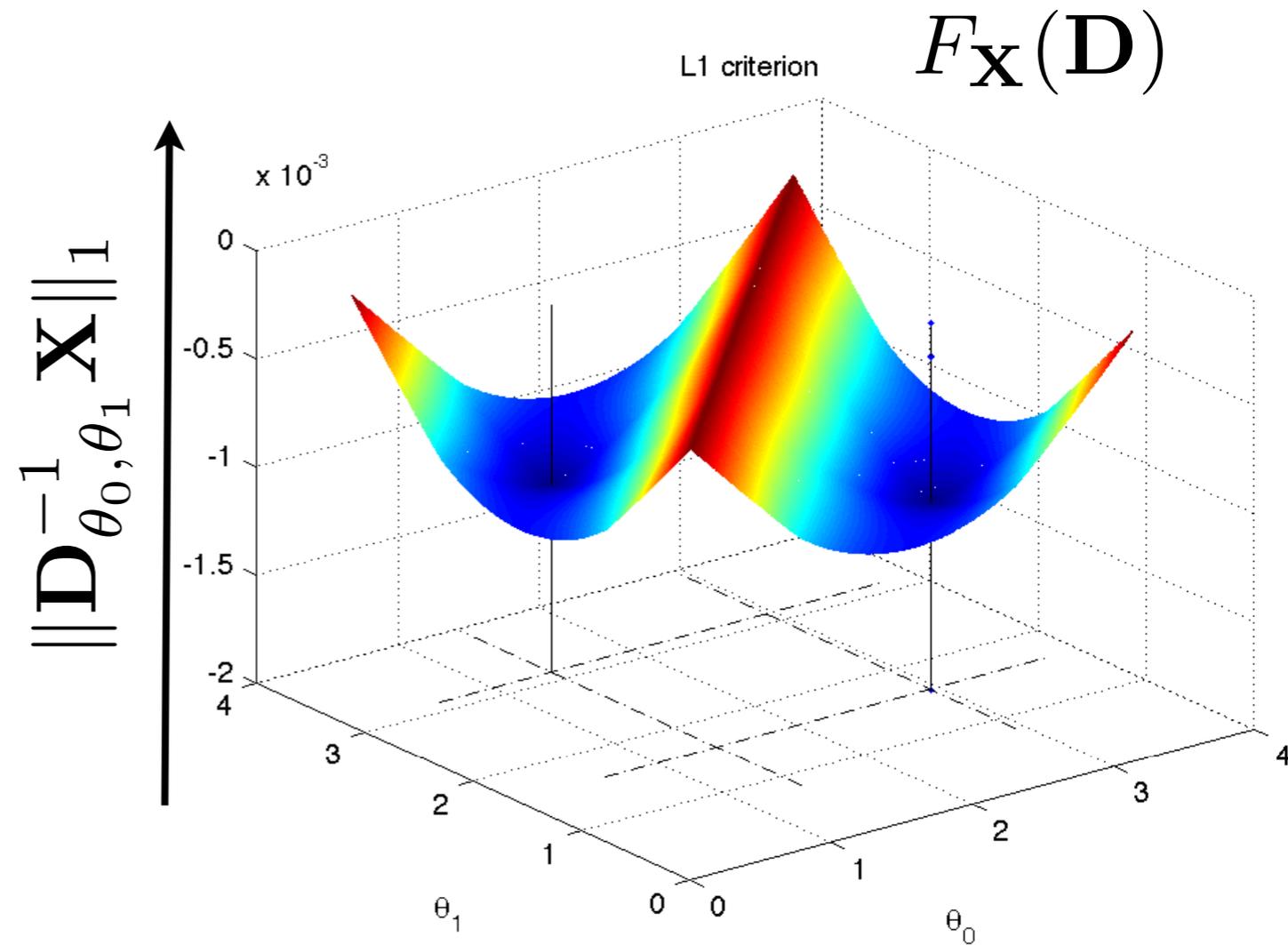
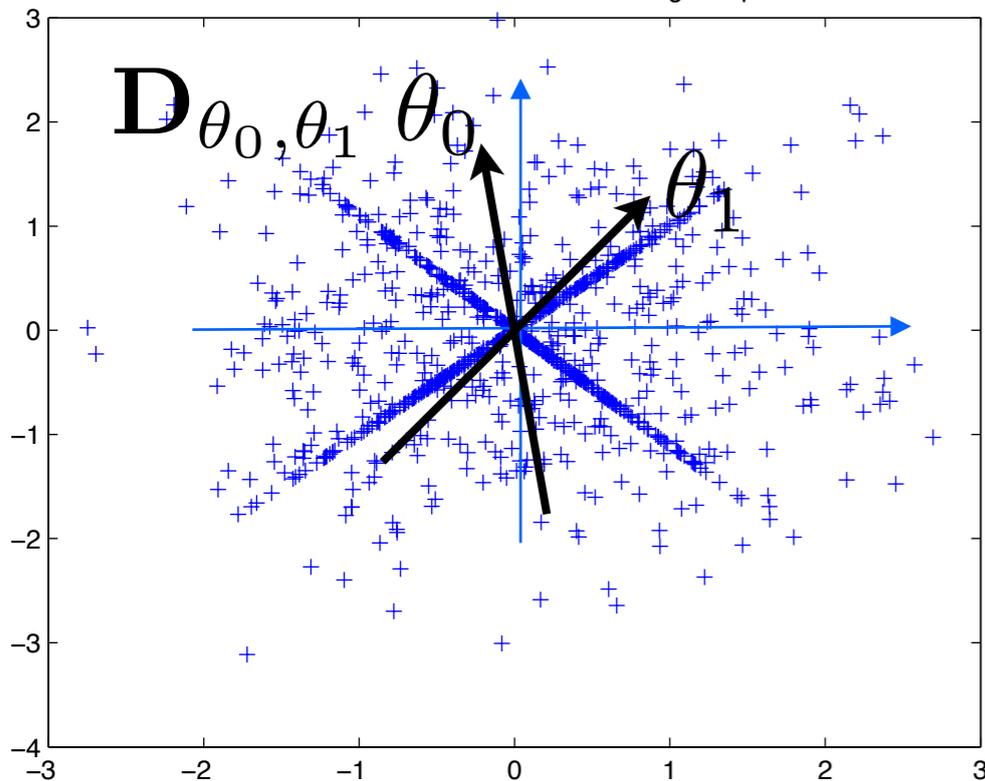
N = 1000 Bernoulli-Gaussian training samples



Numerical example (2D)

$$\mathbf{X} = \mathbf{D}_0 \mathbf{Z}_0$$

N = 1000 Bernoulli-Gaussian training samples

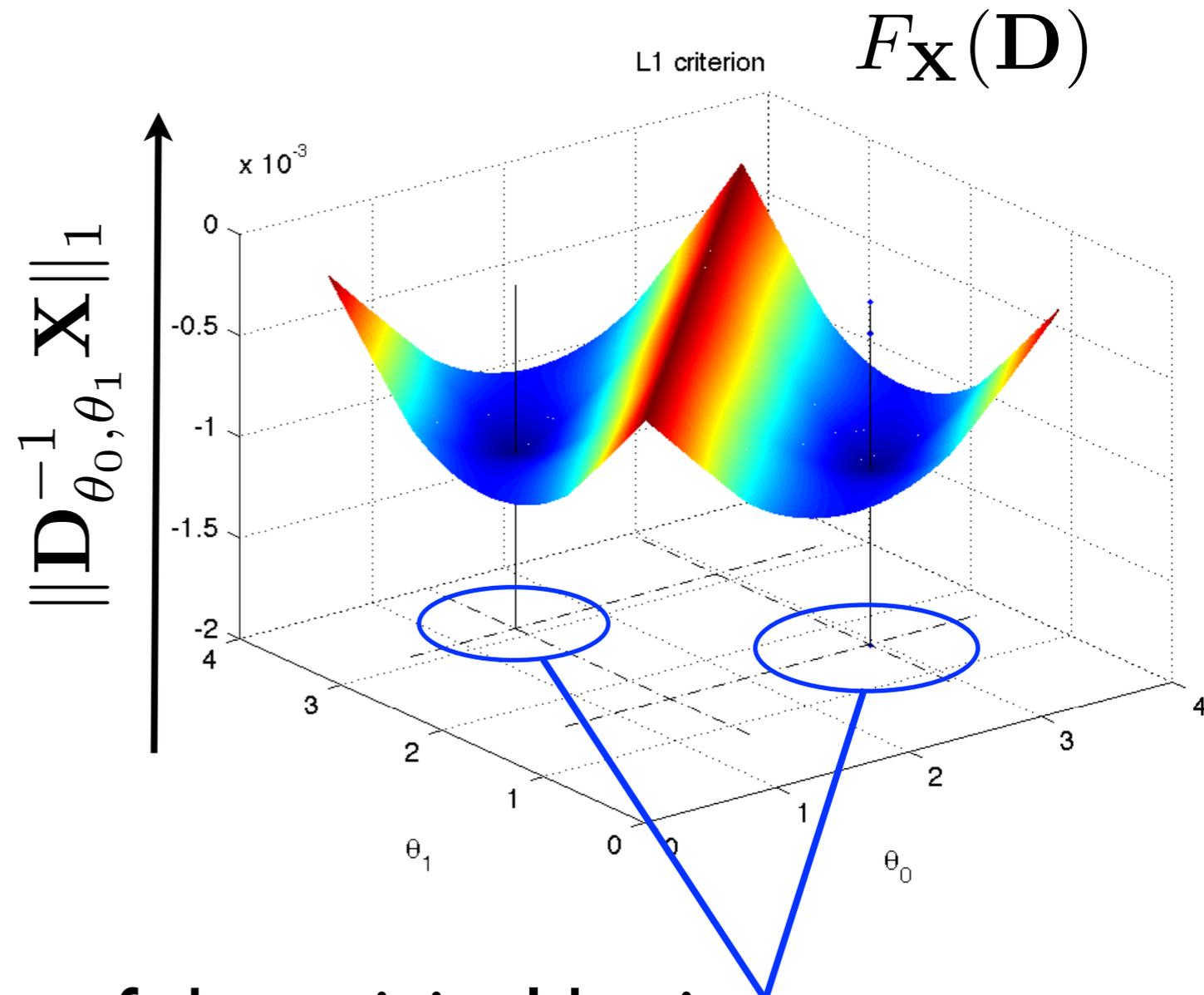
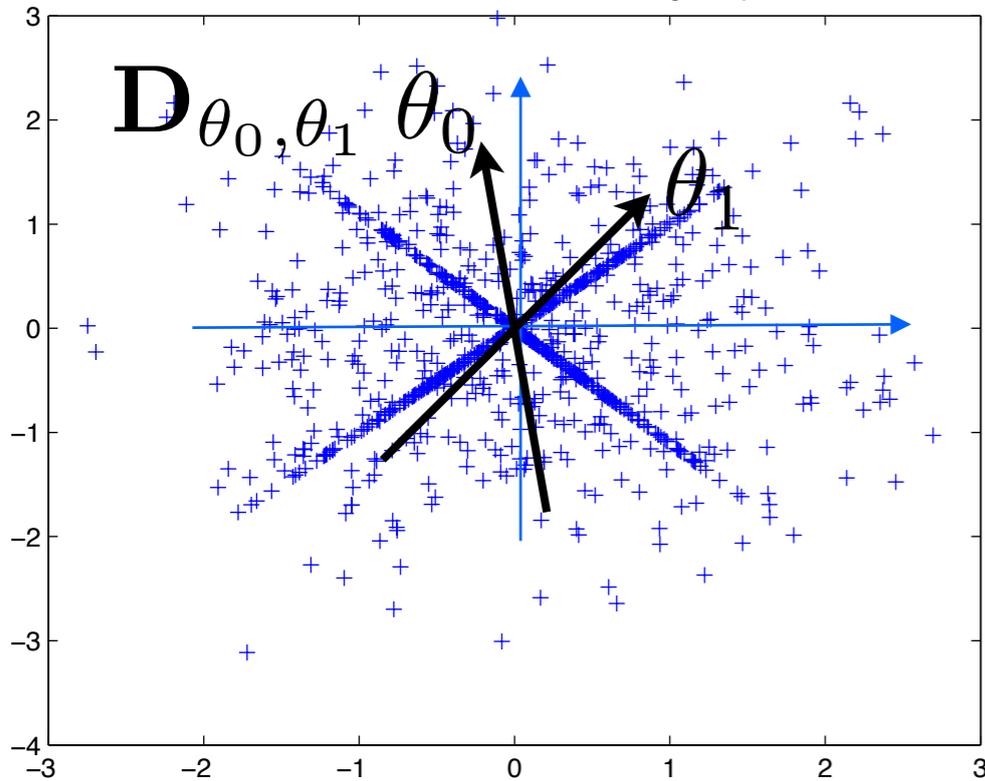


**Symmetry =
permutation ambiguity**

Numerical example (2D)

$$\mathbf{X} = \mathbf{D}_0 \mathbf{Z}_0$$

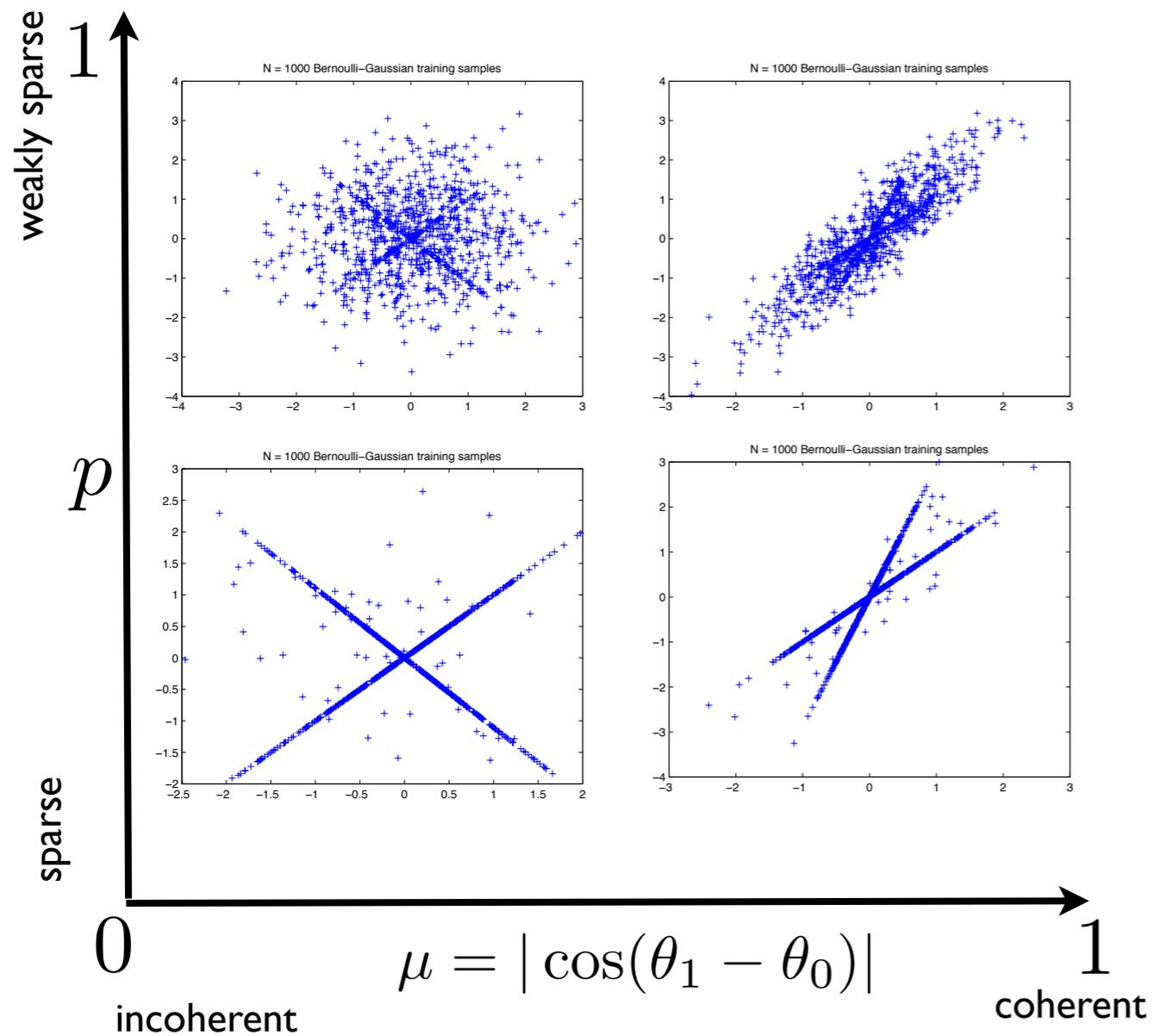
N = 1000 Bernoulli-Gaussian training samples



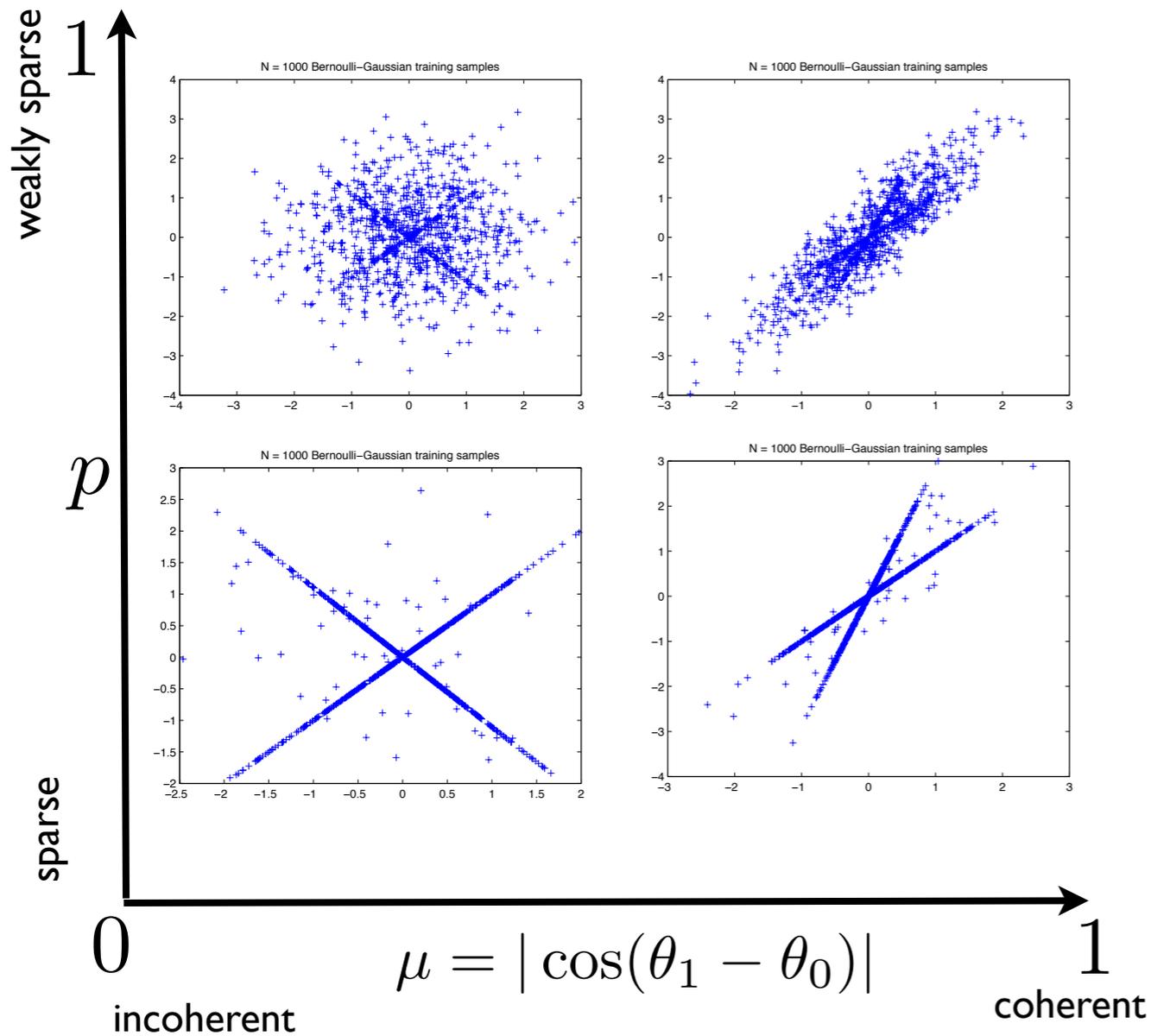
Empirical observations

- Global minima match angles of the original basis
- There is no other local minimum.

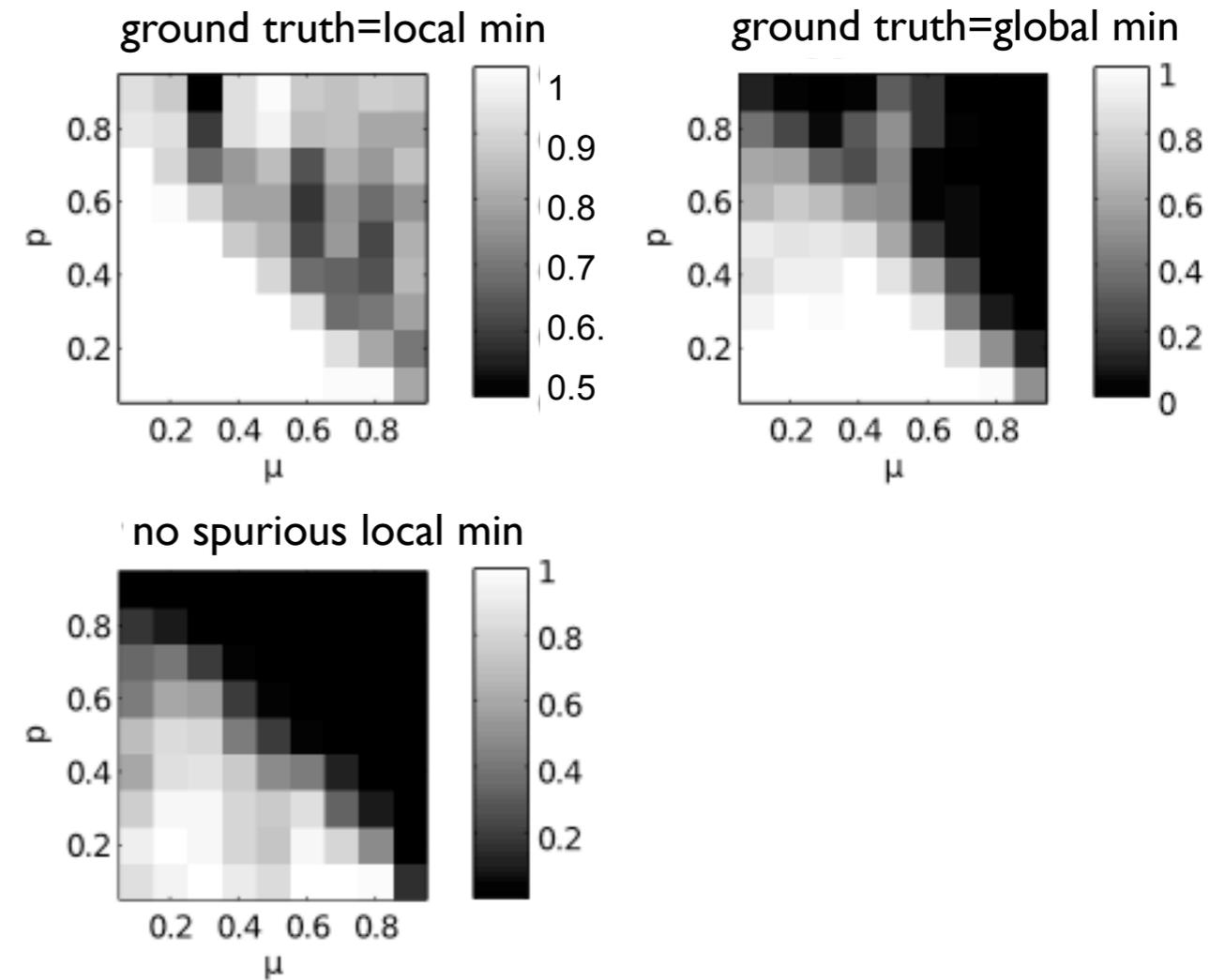
Sparsity vs coherence (2D)



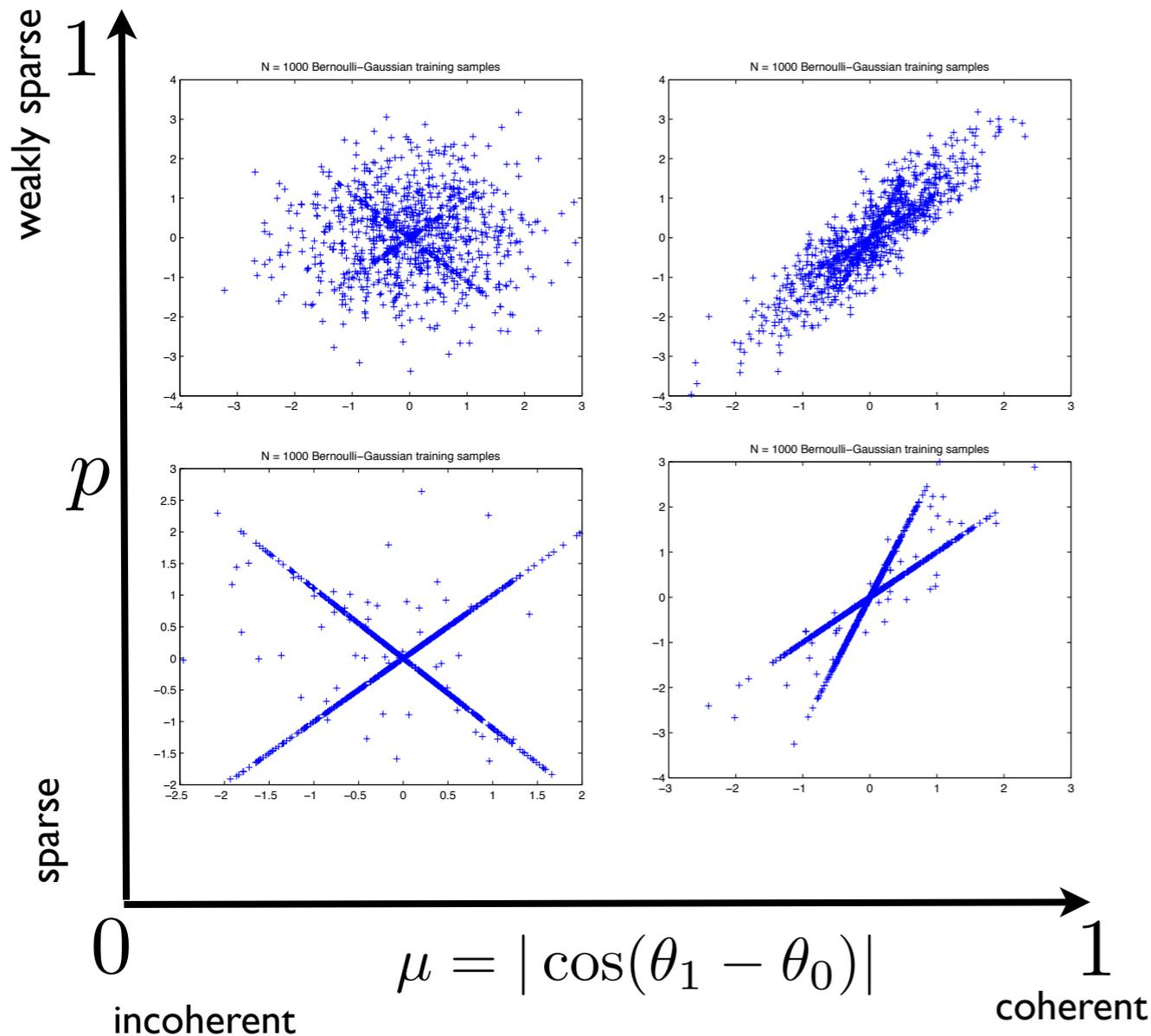
Sparsity vs coherence (2D)



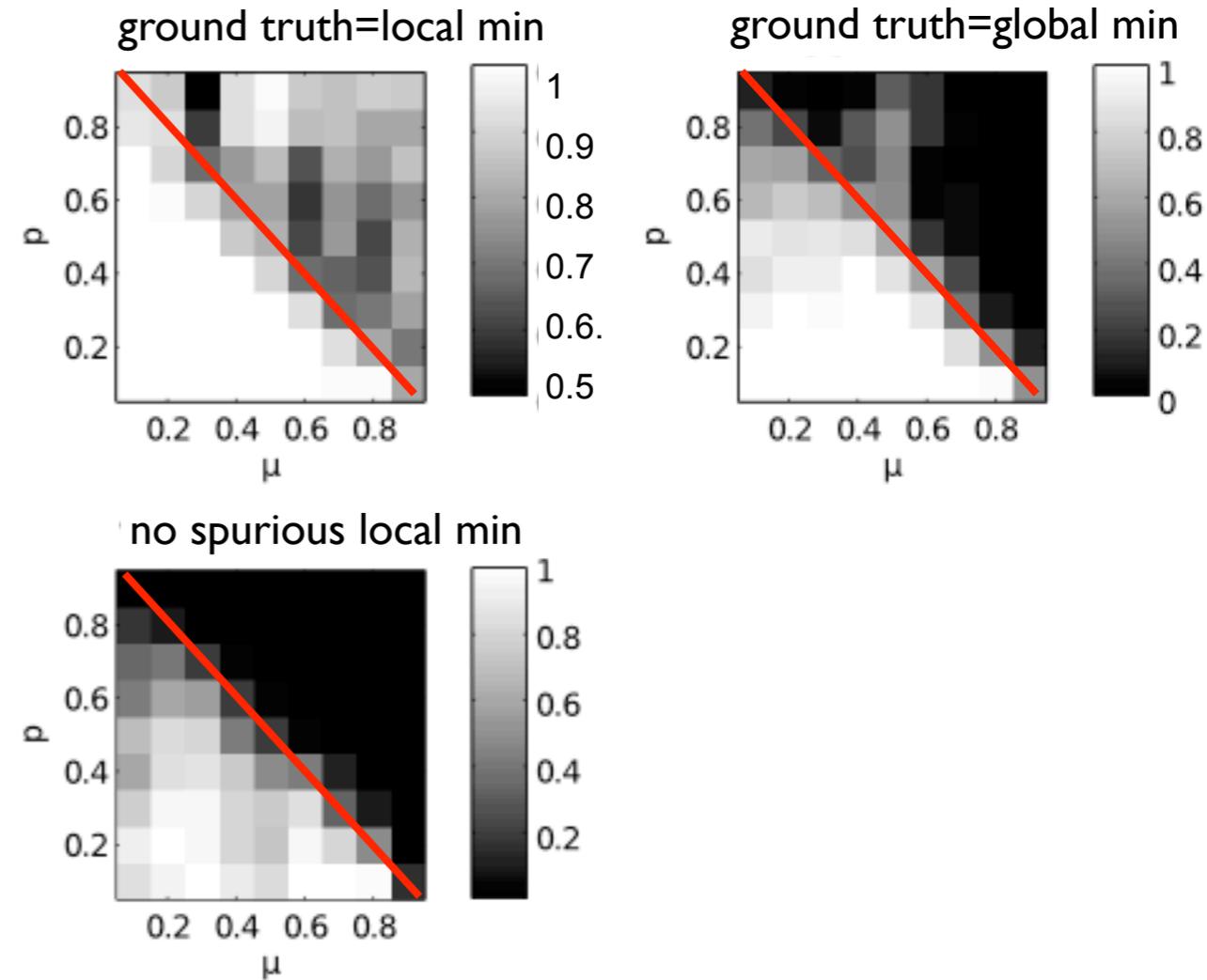
Empirical probability of success



Sparsity vs coherence (2D)



Empirical probability of success



- Rule of thumb:** perfect recovery if:
- Incoherence $\mu < 1 - p$
 - Enough training samples (N large enough)

Empirical findings

- **Stable & robust dictionary identification**

- ✓ Global minima often match ground truth
- ✓ Often, there is no spurious local minimum

- **Role of parameters ?**

- ✓ *sparsity* of Z ?
- ✓ *incoherence* of \mathbf{D} ?
- ✓ *noise* level ?
- ✓ presence / nature of *outliers* ?
- ✓ *sample complexity* (number of training samples) ?

Theoretical guarantees

Theoretical guarantees

Theoretical guarantees

- **Excess risk analysis (~Machine Learning)**

- ◆ [Maurer and Pontil, 2010; Vainsencher et al., 2010; Mehta and Gray, 2012]

$$F_{\mathbf{X}}(\hat{\mathbf{D}}) - \min_{\mathbf{D}} \mathbb{E}_{\mathbf{X}} F_{\mathbf{X}}(\mathbf{D})$$

Theoretical guarantees

- **Excess risk analysis** (~Machine Learning)

- ◆ [Maurer and Pontil, 2010; Vainsencher et al., 2010; Mehta and Gray, 2012]

$$F_{\mathbf{X}}(\hat{\mathbf{D}}) - \min_{\mathbf{D}} \mathbb{E}_{\mathbf{X}} F_{\mathbf{X}}(\mathbf{D})$$

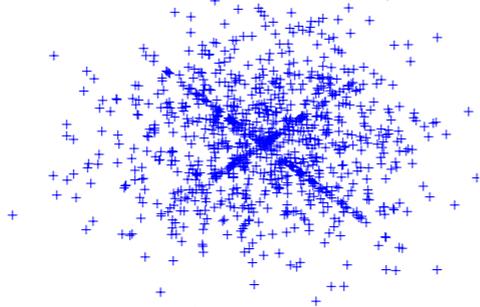
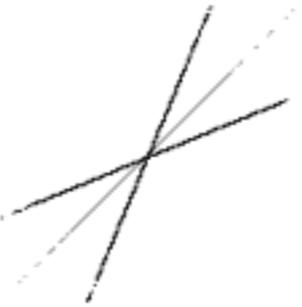
- **Identifiability analysis** (~Signal Processing)

- ◆ [Independent Component Analysis, e.g. book Comon & Jutten 2011]

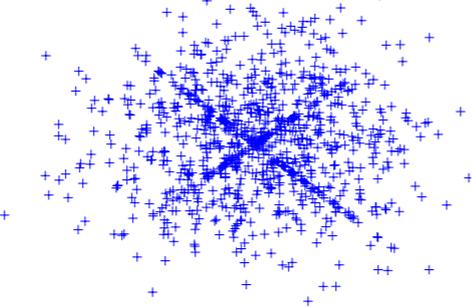
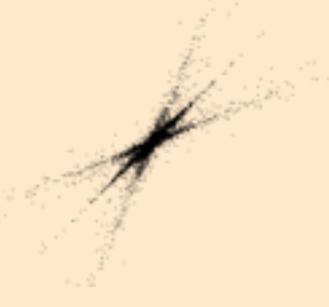
$$\|\hat{\mathbf{D}} - \mathbf{D}_0\|_F$$

- ✓ Array processing perspective
 - ◆ Dictionary ~ directions of arrival
 - ◆ Identification ~ source localization
- ✓ Neural coding perspective:
 - ◆ Dictionaries ~ receptive fields

Theoretical guarantees: overview

| | <i>[G. & Schnass 2010]</i> | <i>[Geng & al 2011]</i> |
|--------------------------|--|---|
| signal model |  |  |
| overcomplete ($d < K$) | no | yes |
| outliers | yes | no |
| noise | no | |
| cost function | $\min_{\mathbf{D}, \mathbf{Z}} \ \mathbf{Z}\ _1 \text{ s.t. } \mathbf{D}\mathbf{Z} = \mathbf{X}$ | |

Theoretical guarantees: overview

| | [G. & Schnass 2010] | [Geng & al 2011] | [Jenatton, Bach & G.] |
|--------------------|--|---|---|
| signal model |  |  |  |
| overcomplete (d<K) | no | yes | yes |
| outliers | yes | no | yes |
| noise | no | | yes |
| cost function | $\min_{\mathbf{D}, \mathbf{Z}} \ \mathbf{Z}\ _1 \text{ s.t. } \mathbf{D}\mathbf{Z} = \mathbf{X}$ | | $\min F_{\mathbf{X}}(\mathbf{D})$ |

Sparse Signal Model

- **Random support**

$$J \subset [1, K], \#J = s$$

- **Sub-Gaussian iid coefficients, bounded below**

$$P(|z_i| < \underline{z}) = 0$$

- **Sub-Gaussian additive noise**

$$\mathbf{x} = \sum_{i \in J} z_i \mathbf{d}_i + \varepsilon = \mathbf{D}_J \mathbf{z}_J + \varepsilon$$

Local stability & robustness

- **Theorem 1: local stability** [Jenatton, Bach & G. 2012]

- ✓ Assumptions:

- ◆ overcomplete *incoherent* dictionary \mathbf{D}_0
 - ◆ s -sparse sub-Gaussian coefficient model (no outlier)
- $$s\mu(\mathbf{D}_0) \ll 1$$

- ✓ Conclusion:

- ◆ with high probability there exists a local minimum of $F_{\mathbf{X}}(\mathbf{D})$ such that

$$\|\mathbf{D} - \mathbf{D}_0\|_F \leq C \sqrt{sdK^3 \cdot \frac{\log N}{N}}$$

- **Theorem 2: robustness to noise**

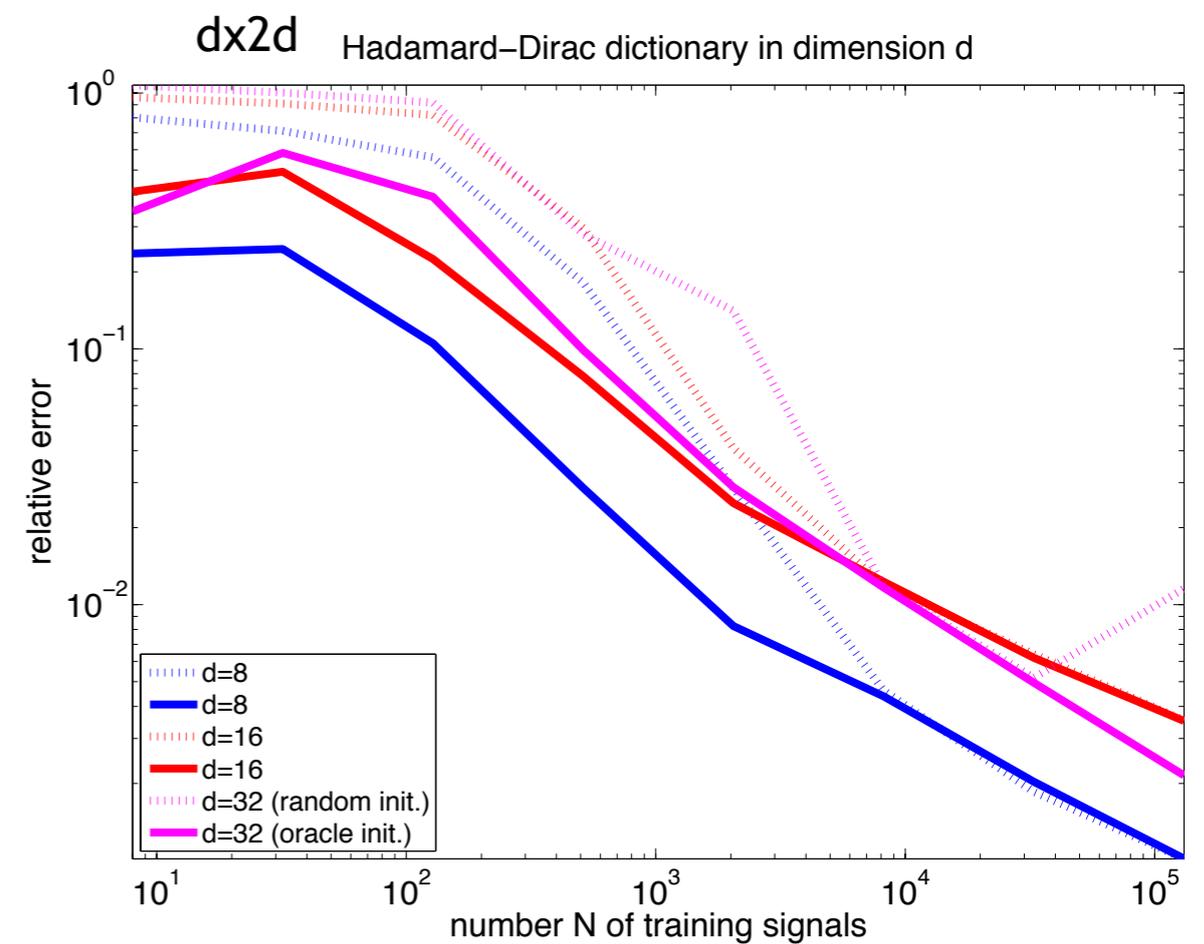
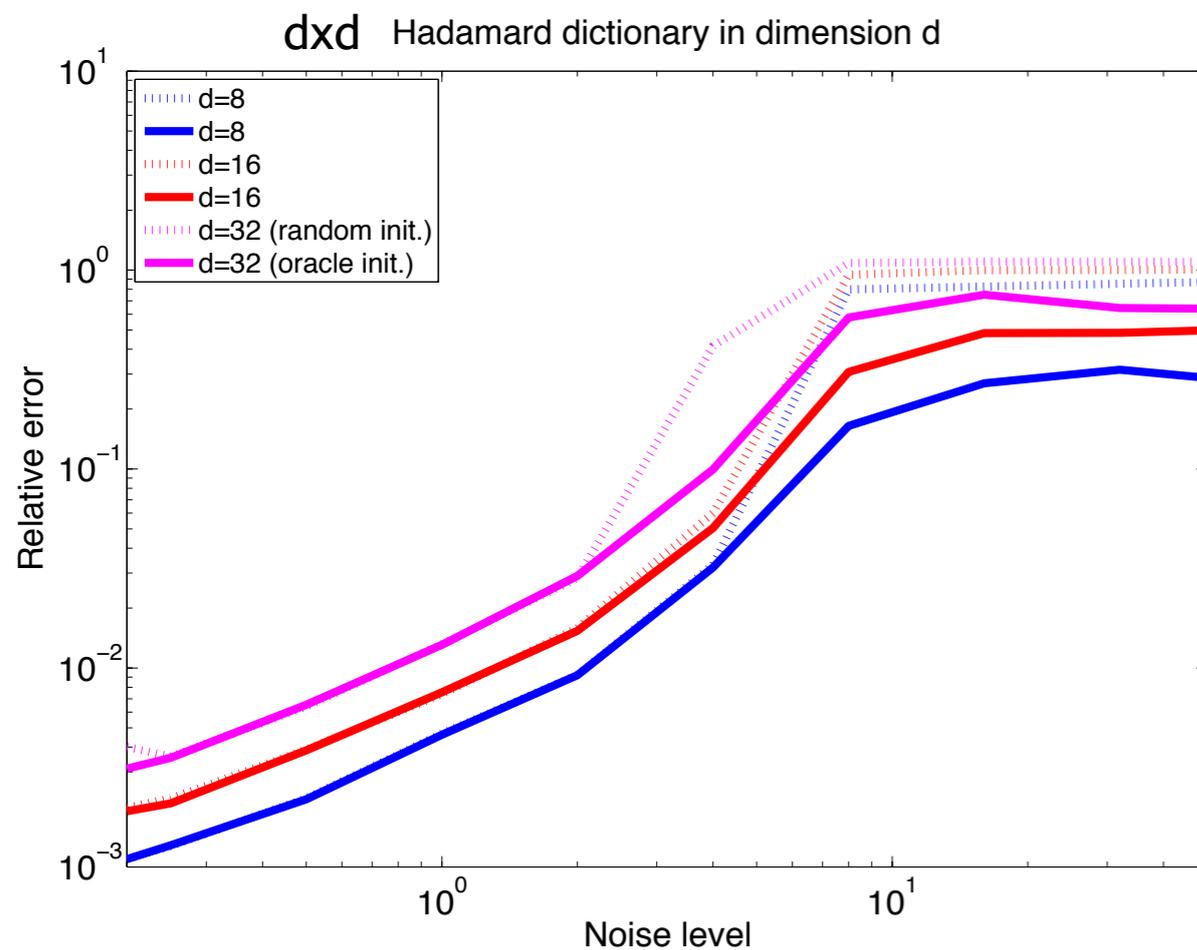
- ✓ technical assumption: bounded coefficient model

- **Theorem 3: robustness to outliers**

Learning Guarantees vs Empirical Findings

- **Robustness to noise**

- **Sample complexity**

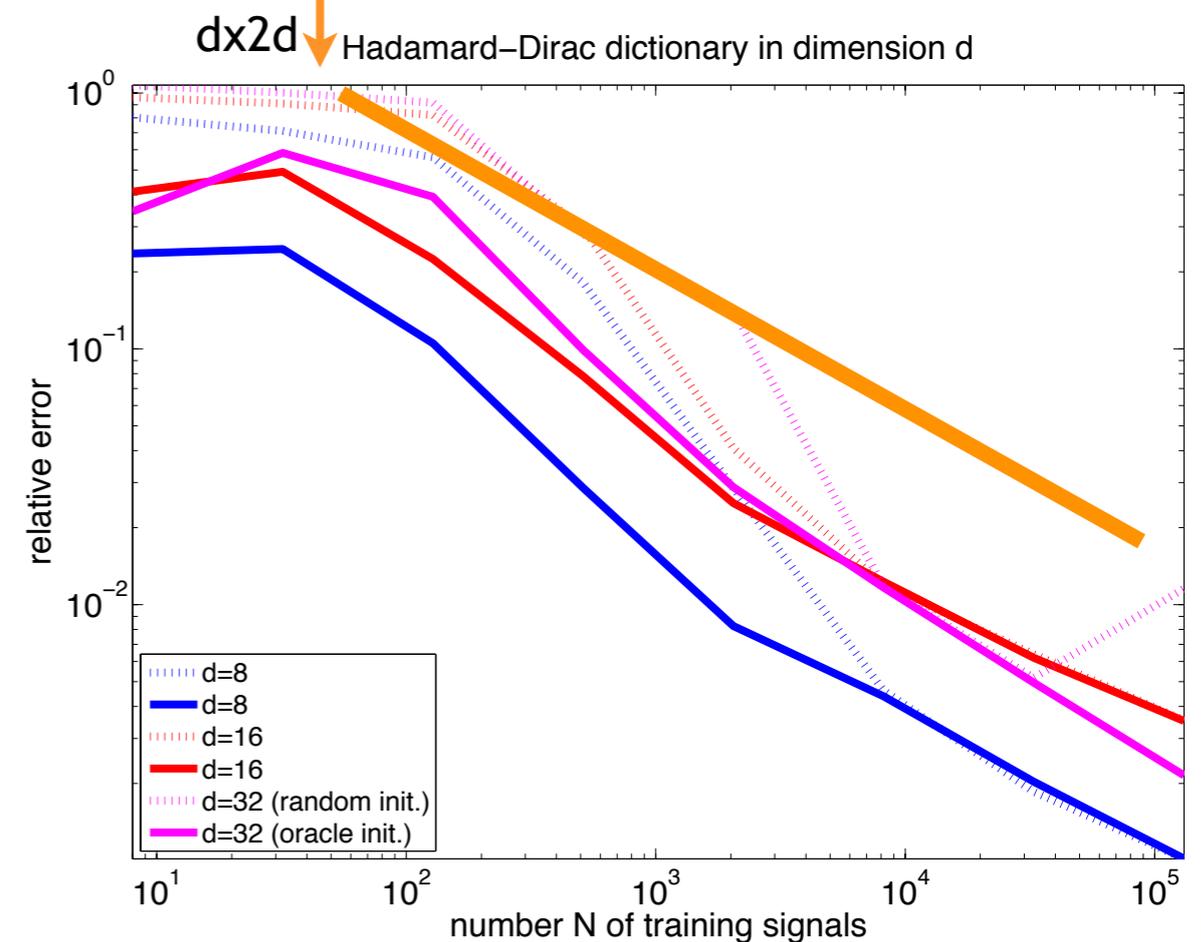
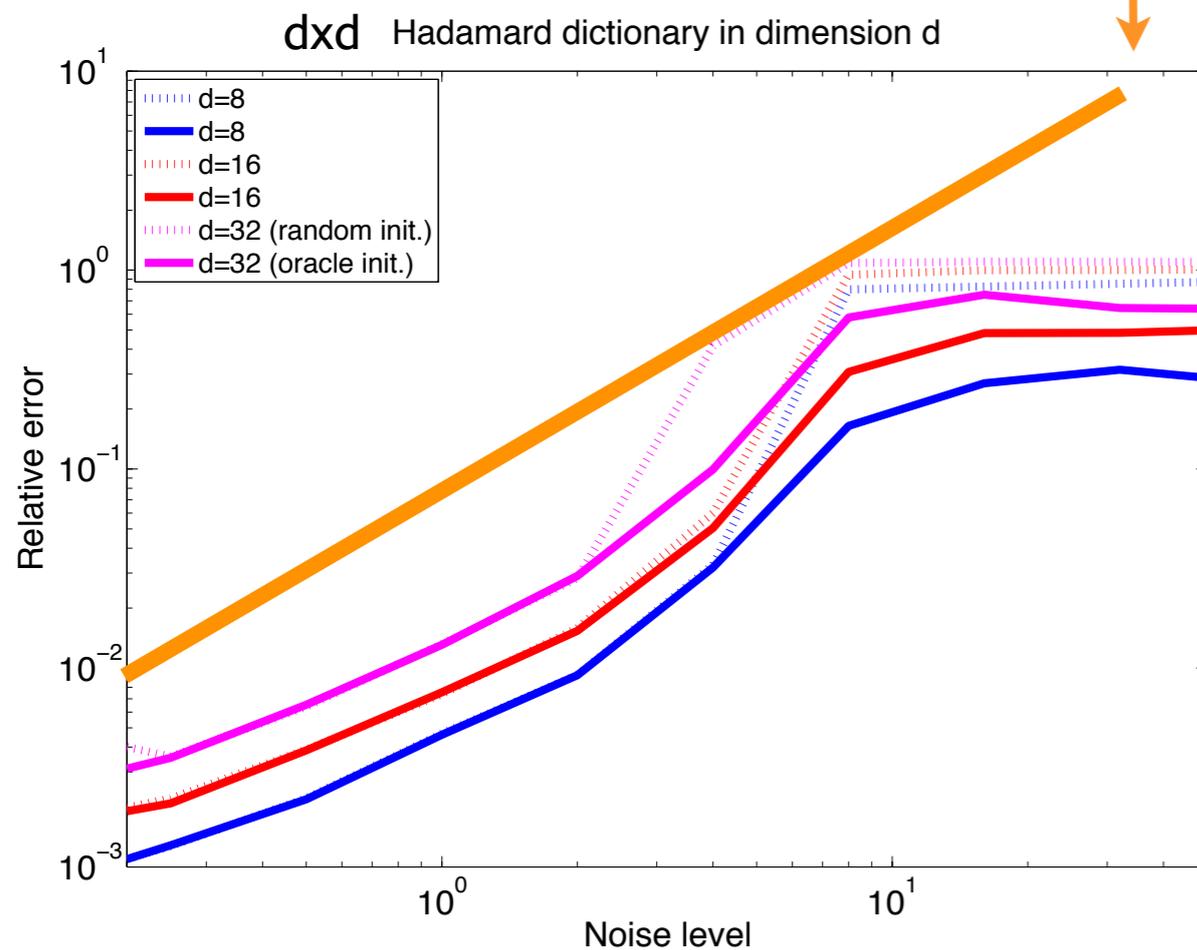


Learning Guarantees vs Empirical Findings

- Robustness to noise

- Sample complexity

Predicted slope



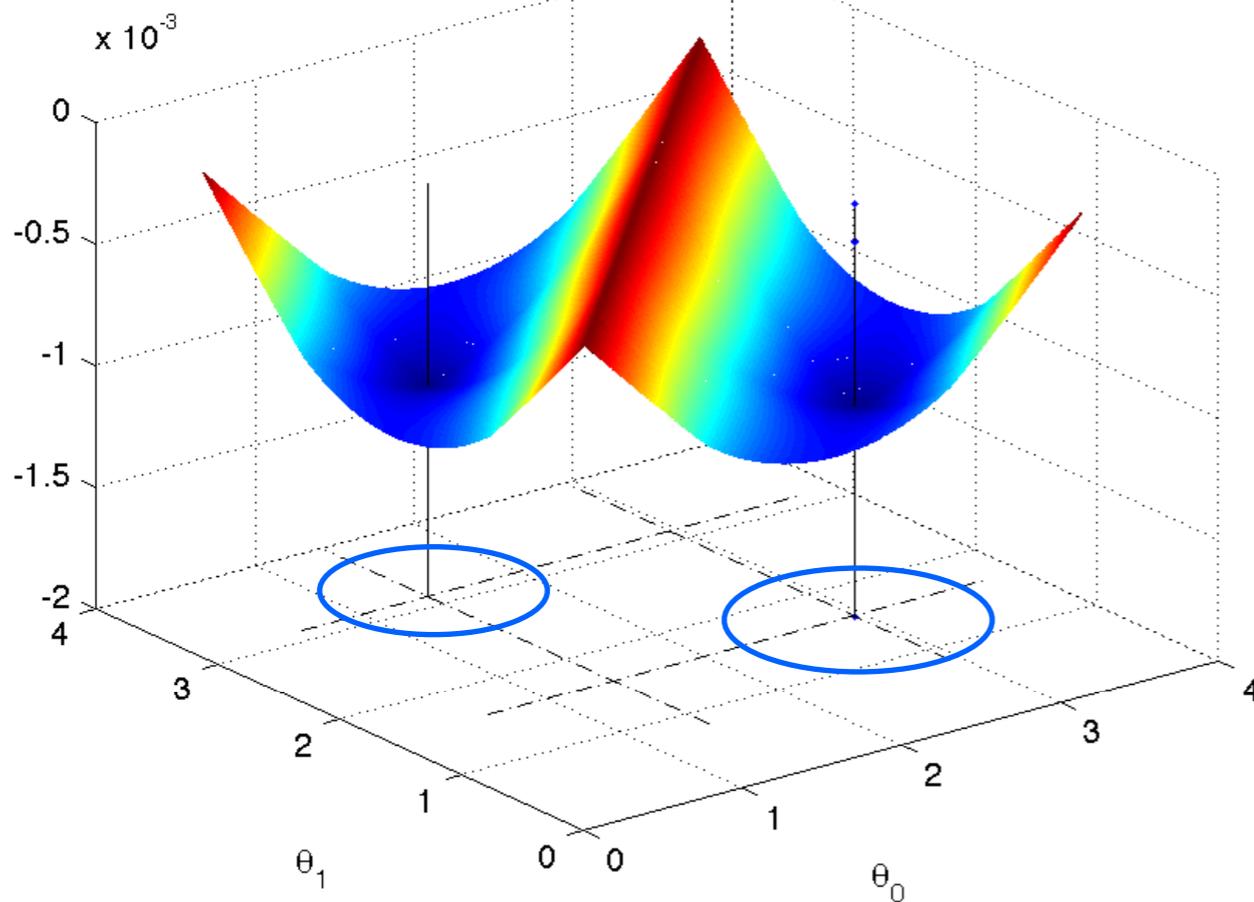
Flavor of the proof

Characterizing local minima (1)

- **Noiseless setting**

- ✓ Minimum *exactly* at ground truth

$$F_{\mathbf{X}}(\mathbf{D}) - F_{\mathbf{X}}(\mathbf{D}_0)$$



- **Noisy setting**

- ✓ Minimum *close to* ground truth

- ✓ Zero at ground truth

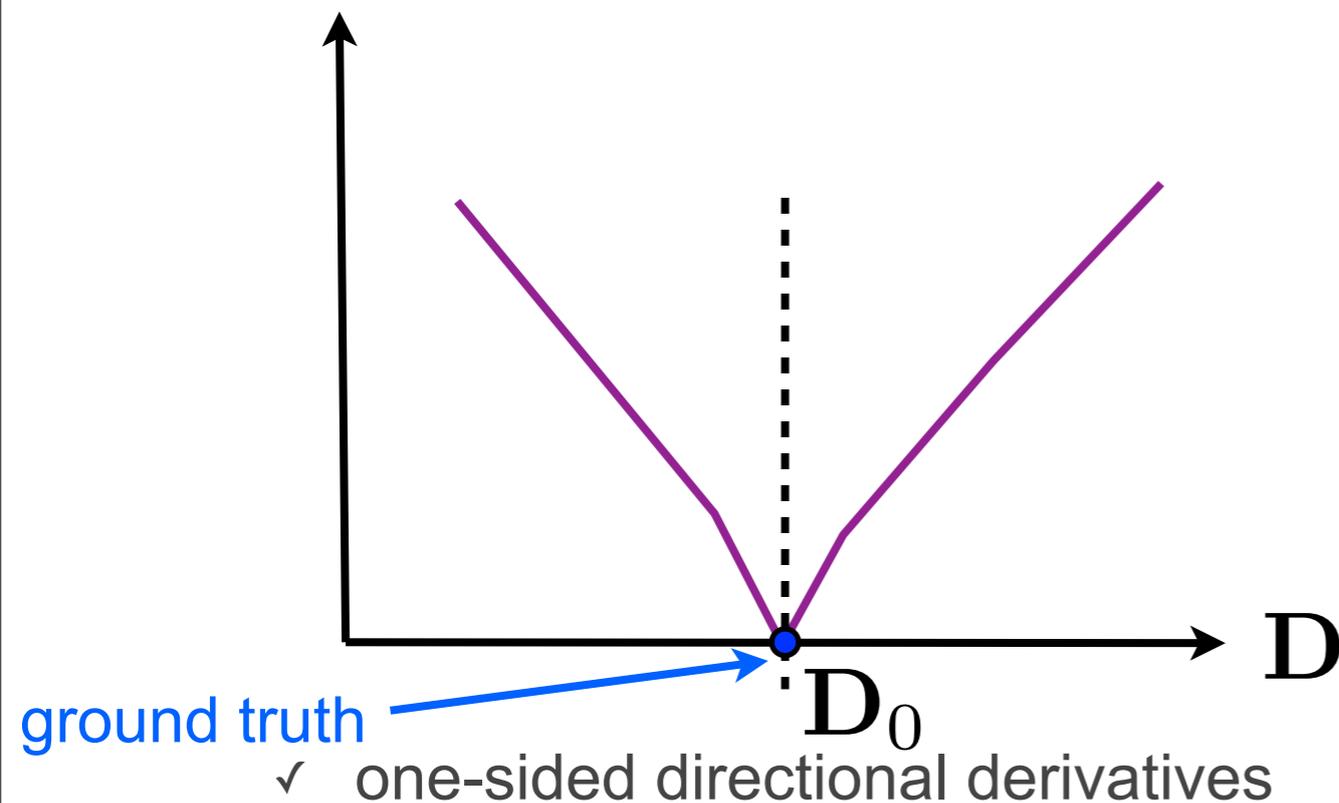
- ✓ Lower bound at radius r

Characterizing local minima (1)

- **Noiseless setting**

- ✓ Minimum *exactly* at ground truth

$$F_{\mathbf{X}}(\mathbf{D}) - F_{\mathbf{X}}(\mathbf{D}_0)$$



- **Noisy setting**

- ✓ Minimum *close to* ground truth

- ✓ Zero at ground truth

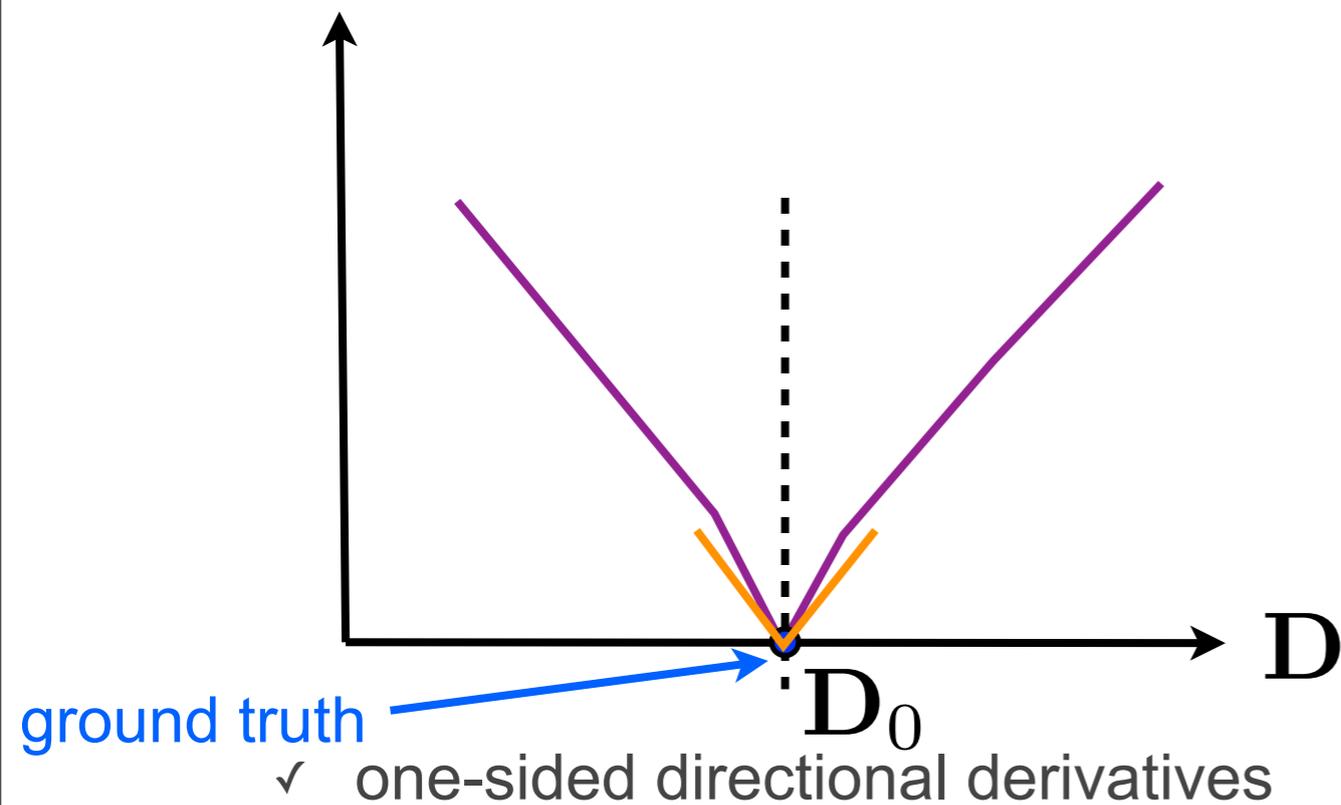
- ✓ Lower bound at radius r

Characterizing local minima (1)

- **Noiseless setting**

- ✓ Minimum *exactly* at ground truth

$$F_{\mathbf{X}}(\mathbf{D}) - F_{\mathbf{X}}(\mathbf{D}_0)$$

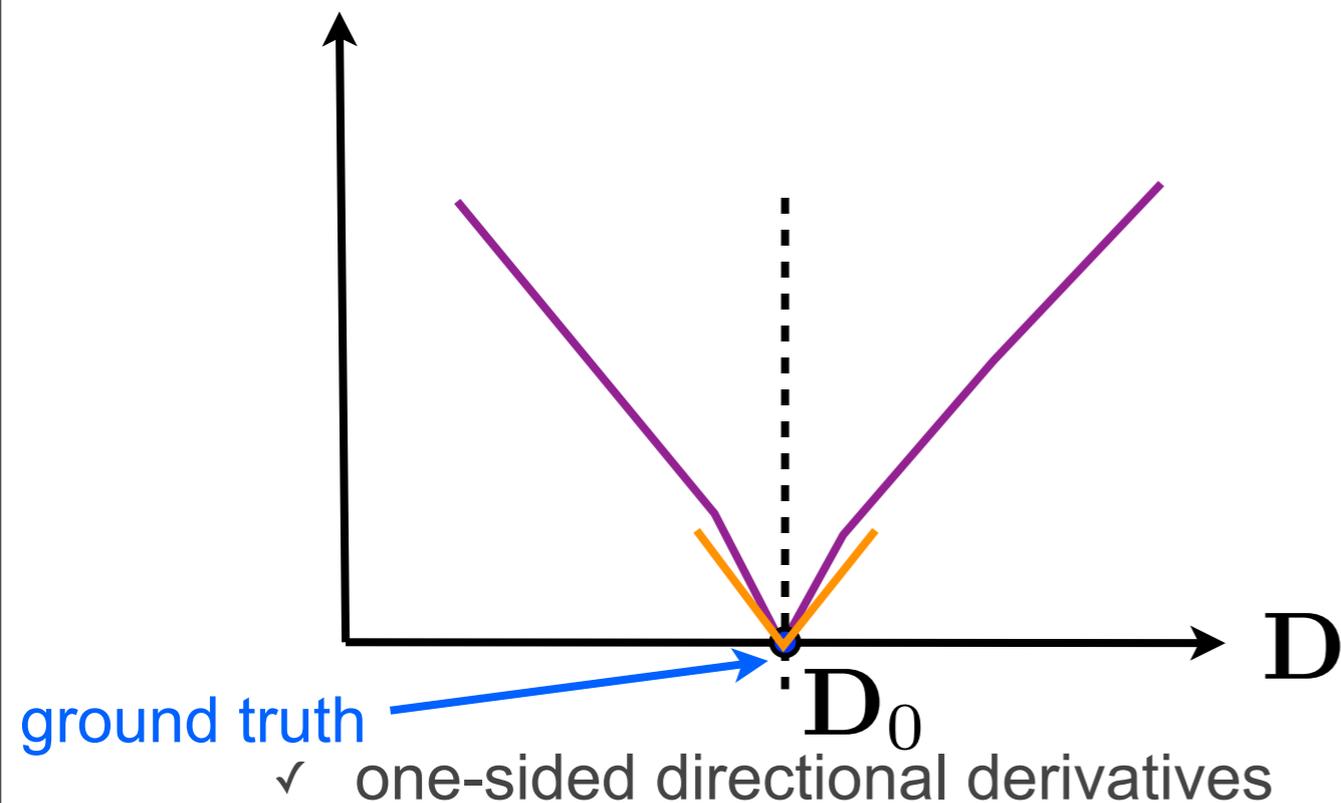


Characterizing local minima (1)

- **Noiseless setting**

- ✓ Minimum *exactly* at ground truth

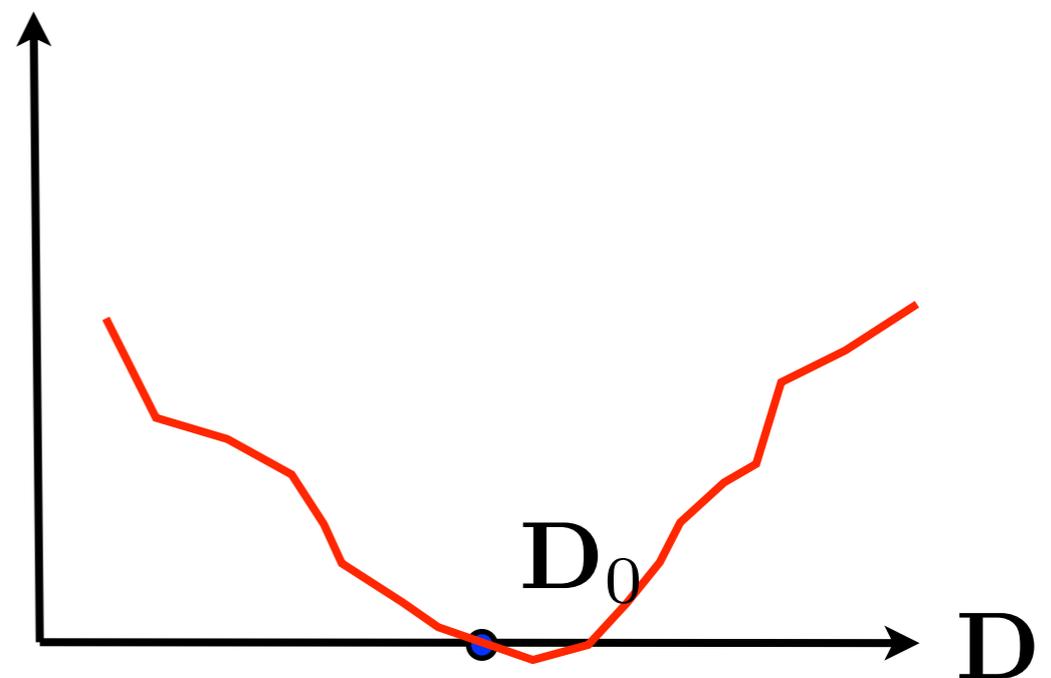
$$F_{\mathbf{X}}(\mathbf{D}) - F_{\mathbf{X}}(\mathbf{D}_0)$$



- **Noisy setting**

- ✓ Minimum *close to* ground truth

$$F_{\mathbf{X}}(\mathbf{D}) - F_{\mathbf{X}}(\mathbf{D}_0)$$

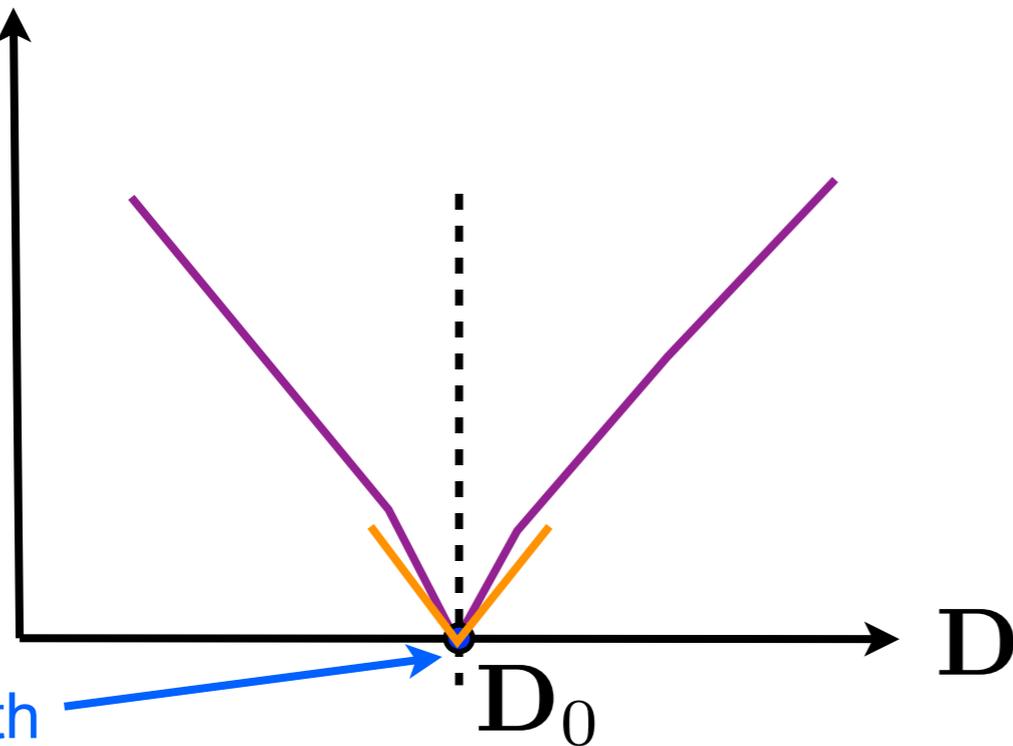


Characterizing local minima (1)

- **Noiseless setting**

- ✓ Minimum *exactly* at ground truth

$$F_{\mathbf{X}}(\mathbf{D}) - F_{\mathbf{X}}(\mathbf{D}_0)$$

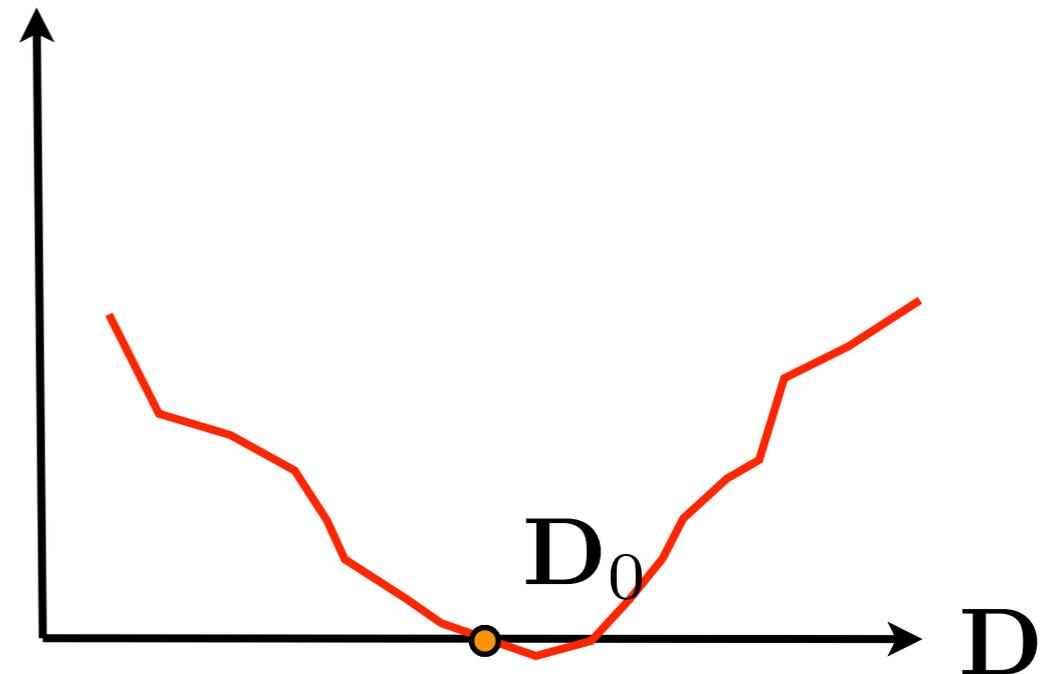


- ✓ one-sided directional derivatives

- **Noisy setting**

- ✓ Minimum *close to* ground truth

$$F_{\mathbf{X}}(\mathbf{D}) - F_{\mathbf{X}}(\mathbf{D}_0)$$



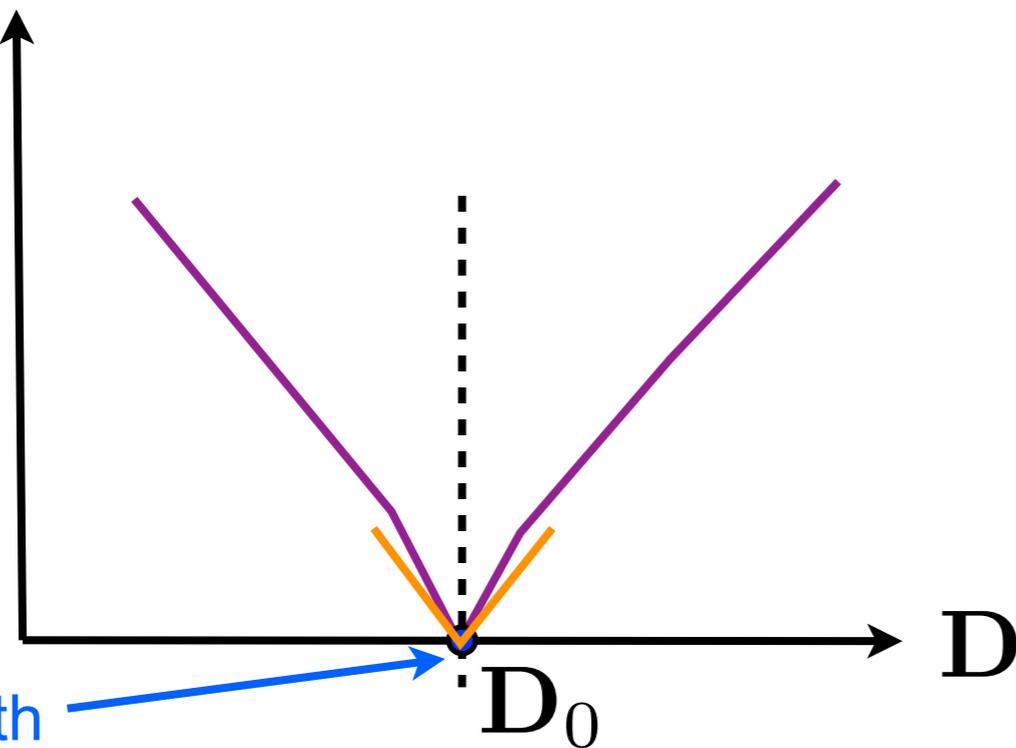
- ✓ Zero at ground truth

Characterizing local minima (1)

- **Noiseless setting**

- ✓ Minimum *exactly* at ground truth

$$F_{\mathbf{X}}(\mathbf{D}) - F_{\mathbf{X}}(\mathbf{D}_0)$$

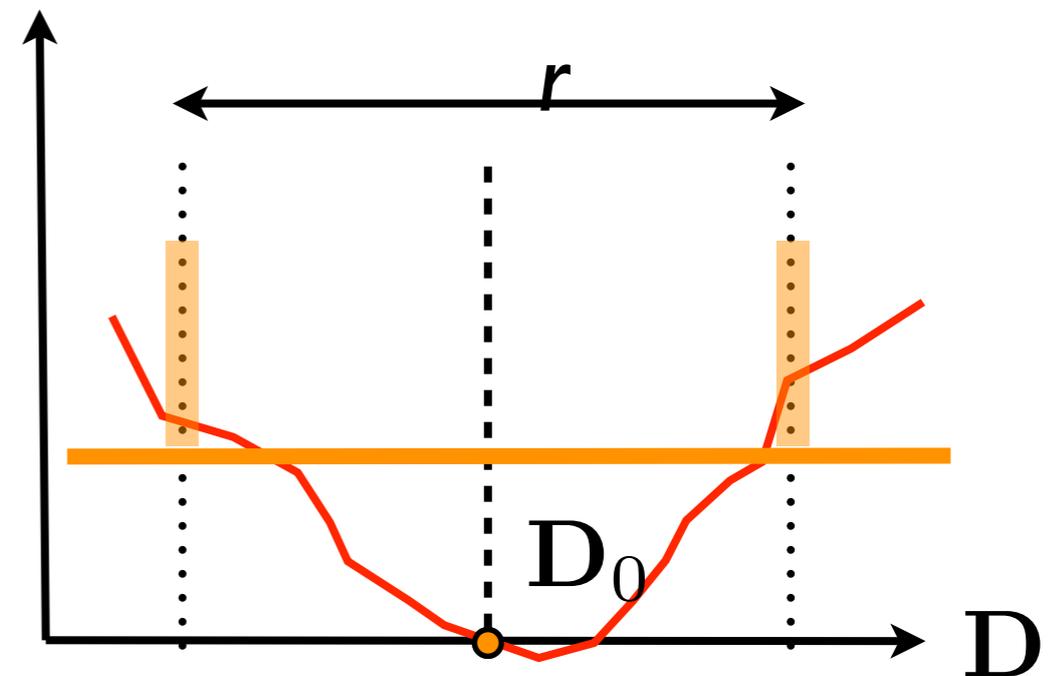


- ✓ one-sided directional derivatives

- **Noisy setting**

- ✓ Minimum *close to* ground truth

$$F_{\mathbf{X}}(\mathbf{D}) - F_{\mathbf{X}}(\mathbf{D}_0)$$



- ✓ Zero at ground truth

- ✓ Lower bound at radius r

Controlling the cost function

- **Problem:** $F_{\mathbf{X}}(\mathbf{D})$ sum of complicated functions!

$$f_{\mathbf{x}_n}(\mathbf{D}) = \min_{z_n} \frac{1}{2} \|\mathbf{x}_n - \mathbf{D}z_n\|_2^2 + \lambda \|z_n\|_1$$

- **Solution:** simplified expression if sparse recovery
 - ◆ adaptation from [Fuchs, 2005; Zhao and Yu, 2006; Wainwright, 2009]

$$f_{\mathbf{x}}(\mathbf{D}) = \phi_{\mathbf{x}}(\mathbf{D}|\text{sign}(z_0)) \quad \mathbf{x} = \mathbf{D}_0\mathbf{z}_0 + \varepsilon$$

- ✓ Approximate cost function $\Phi_{\mathbf{X}}(\mathbf{D}) \approx F_{\mathbf{X}}(\mathbf{D})$

Controlling the *approximate* cost function

- **Problem:**

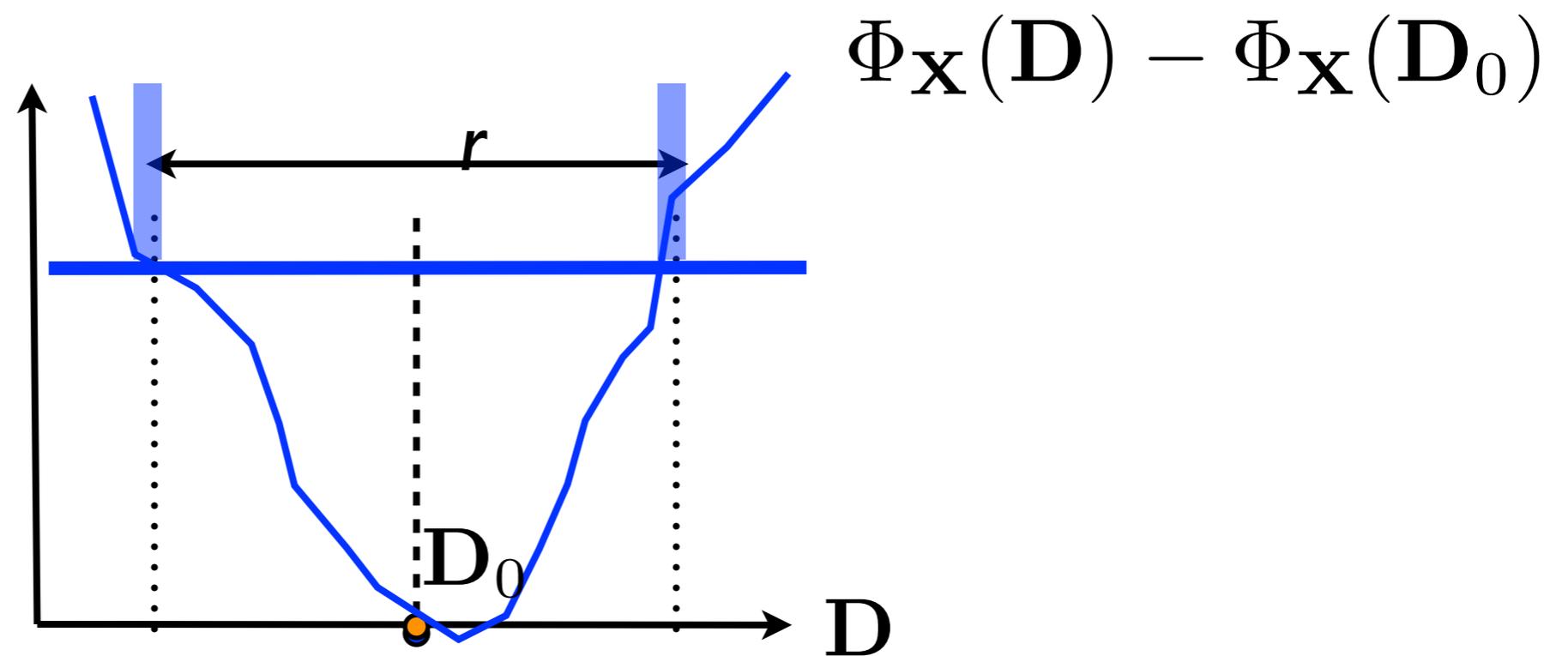
- ✓ Need *uniform* lower bound on the sphere $\|\mathbf{D} - \mathbf{D}_0\|_F = r$ of the *random* function

$$\Phi_{\mathbf{X}}(\mathbf{D}) - \Phi_{\mathbf{X}}(\mathbf{D}_0)$$

- **Solution:**

- ✓ Lower bound expectation for a given \mathbf{D}
- ✓ Control Lipschitz constant (with high probability)
- ✓ Conclude with epsilon-net argument

Putting the pieces together

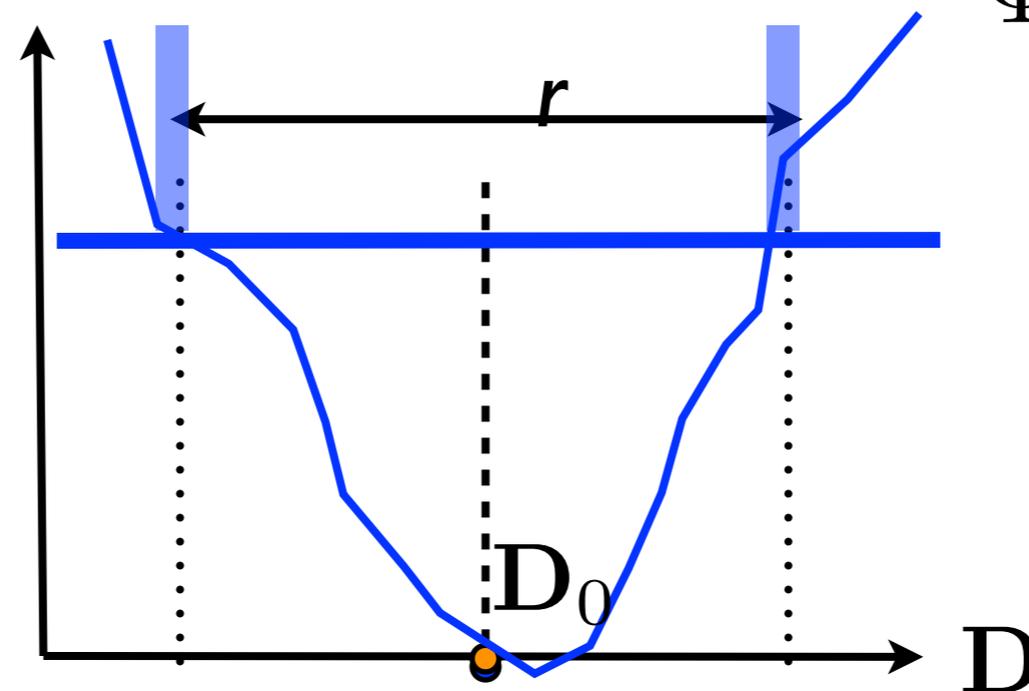


Putting the pieces together

- **With high probability:**

- ✓ lower-bound on approximate cost function

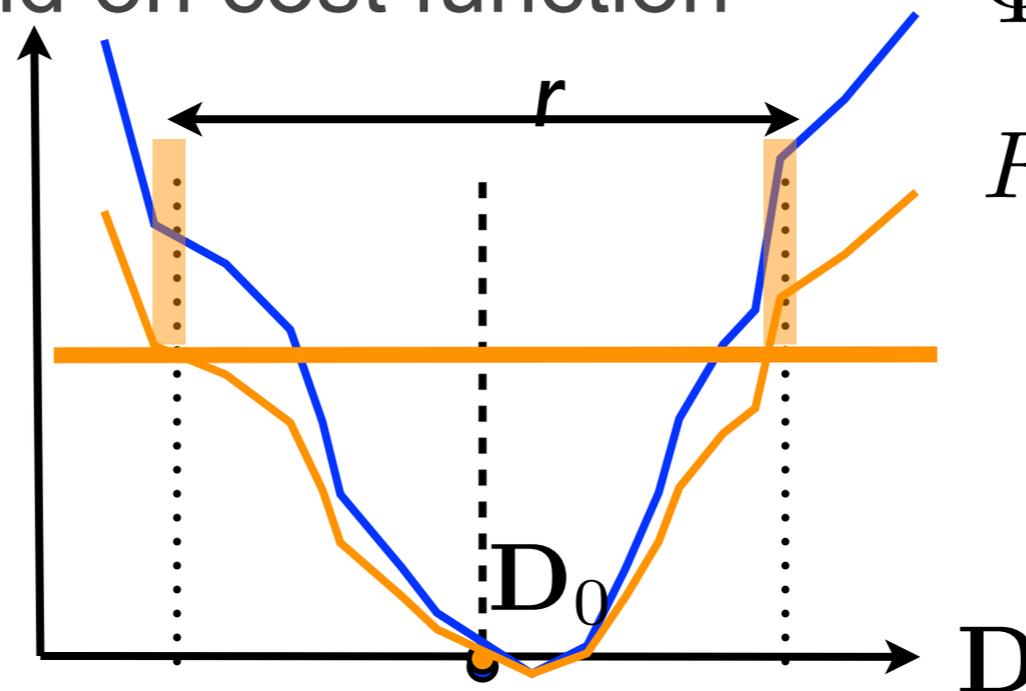
$$\Phi_{\mathbf{X}}(\mathbf{D}) - \Phi_{\mathbf{X}}(\mathbf{D}_0)$$



Putting the pieces together

- **With high probability:**

- ✓ lower-bound on approximate cost function
- ✓ lower-bound on cost function



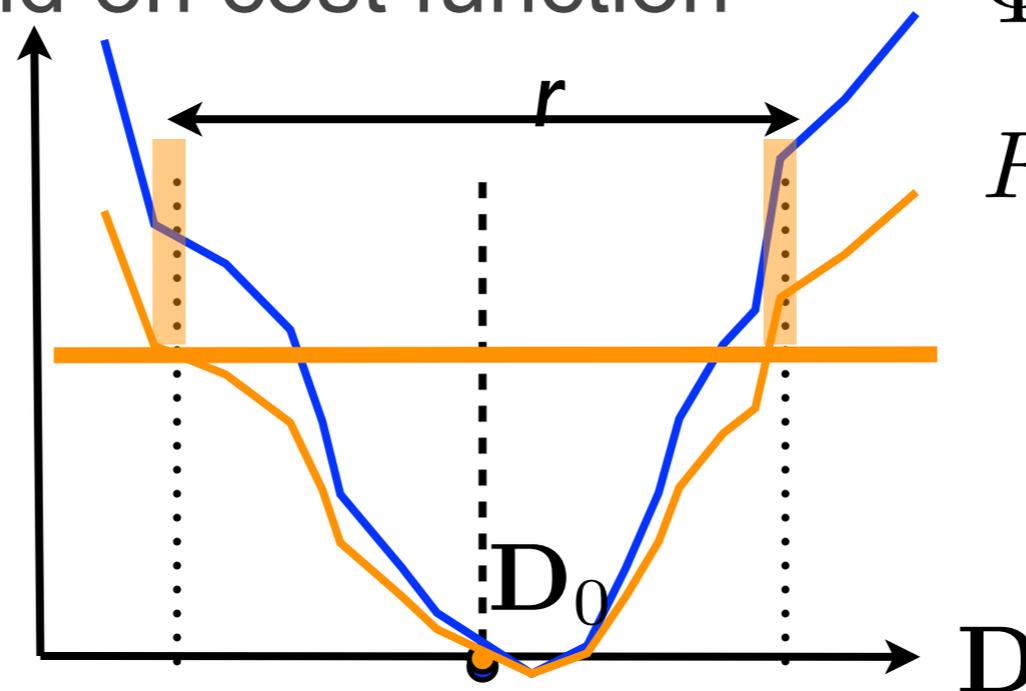
$$\Phi_{\mathbf{X}}(\mathbf{D}) - \Phi_{\mathbf{X}}(\mathbf{D}_0)$$

$$F_{\mathbf{X}}(\mathbf{D}) - F_{\mathbf{X}}(\mathbf{D}_0)$$

Putting the pieces together

- **With high probability:**

- ✓ lower-bound on approximate cost function
- ✓ lower-bound on cost function



$$\Phi_{\mathbf{X}}(\mathbf{D}) - \Phi_{\mathbf{X}}(\mathbf{D}_0)$$

$$F_{\mathbf{X}}(\mathbf{D}) - F_{\mathbf{X}}(\mathbf{D}_0)$$

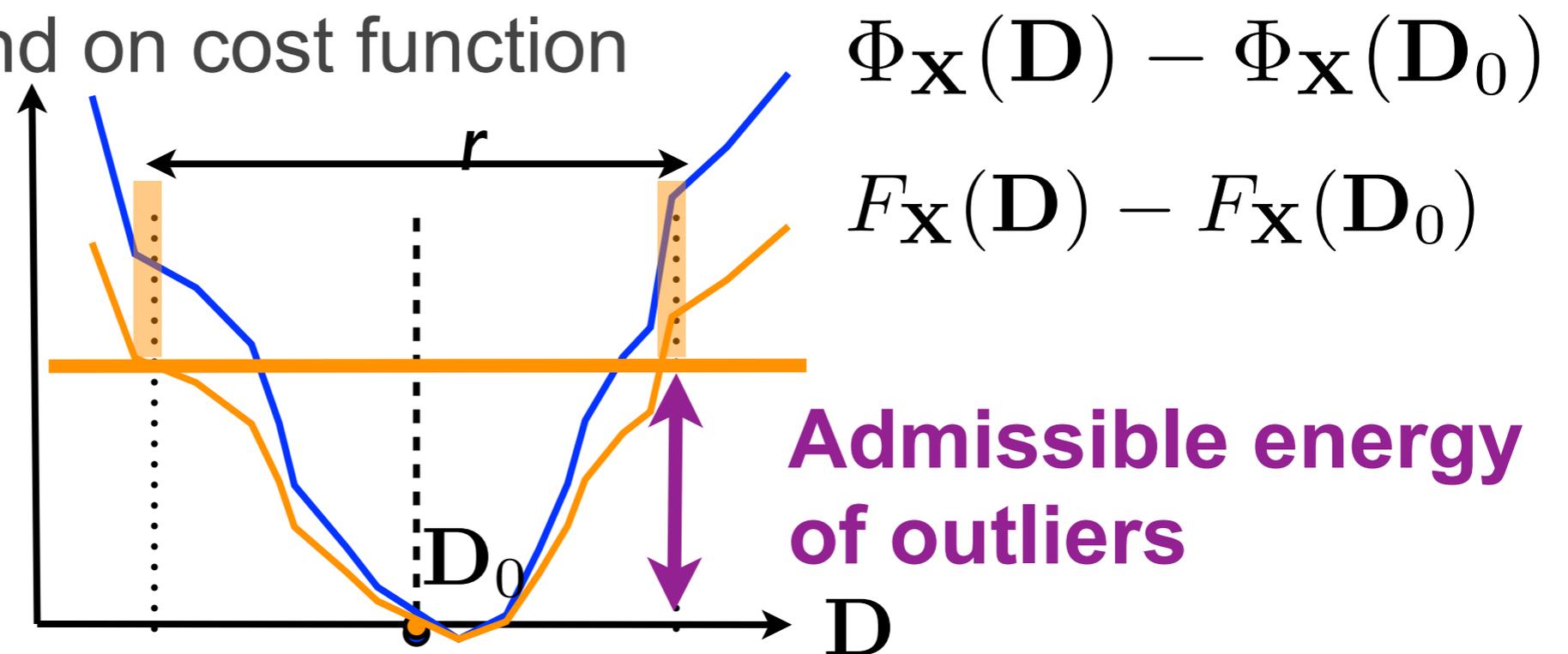
- **Outliers:** «no model» but total energy bounded

$$\frac{1}{N} \sum_{n \in \text{outlier}} \|\mathbf{x}_n\|_2^2 \leq c$$

Putting the pieces together

- **With high probability:**

- ✓ lower-bound on approximate cost function
- ✓ lower-bound on cost function

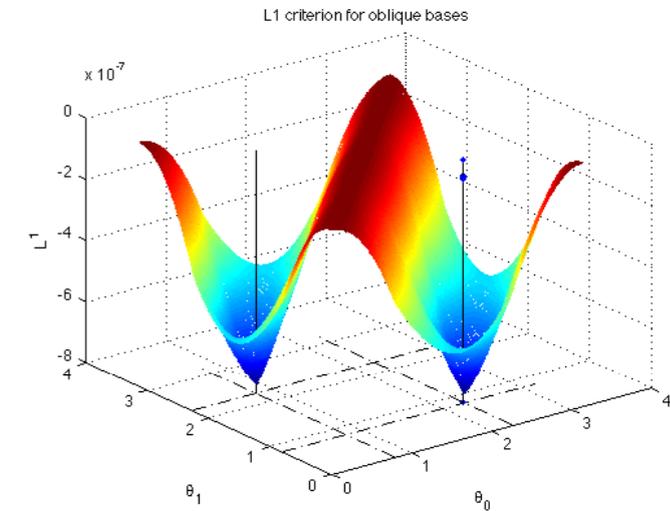


- **Outliers:** «no model» but total energy bounded

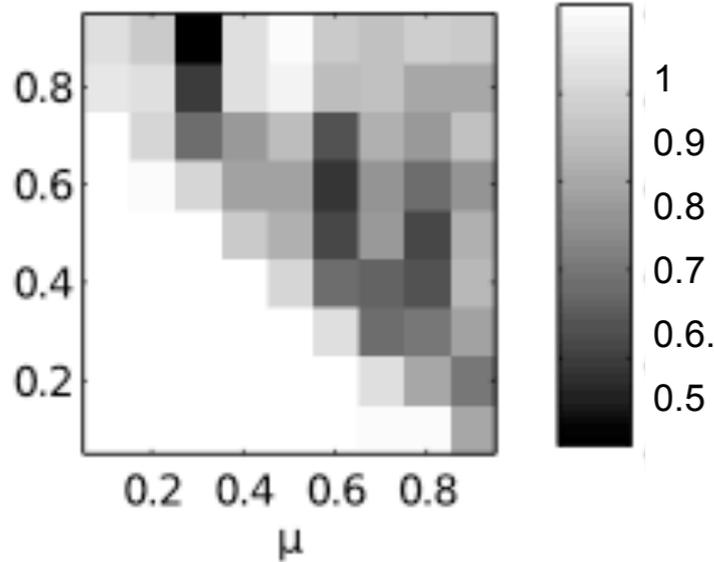
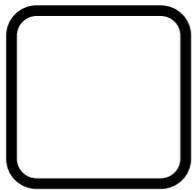
$$\frac{1}{N} \sum_{n \in \text{outlier}} \|\mathbf{x}_n\|_2^2 \leq c$$

From local to global guarantees ?

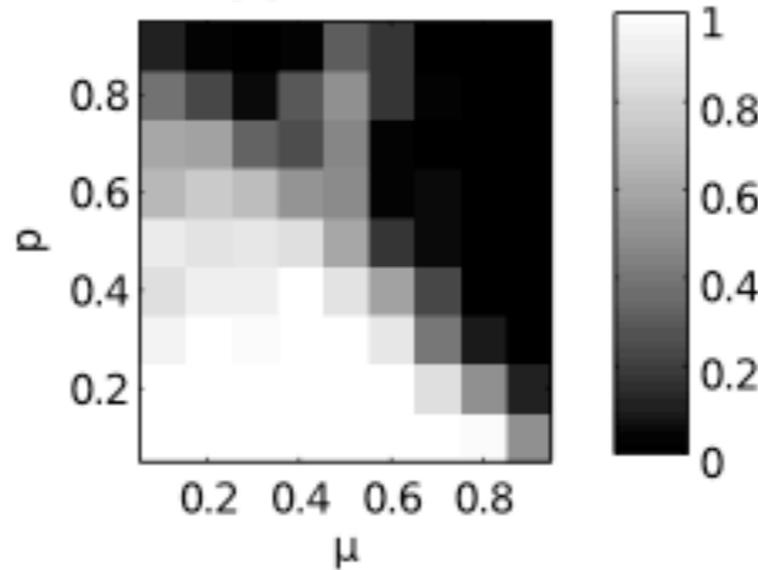
$$\hat{\mathbf{D}} = \arg \min_{\mathbf{D} \in \mathcal{D}} F_{\mathbf{X}}(\mathbf{D})$$



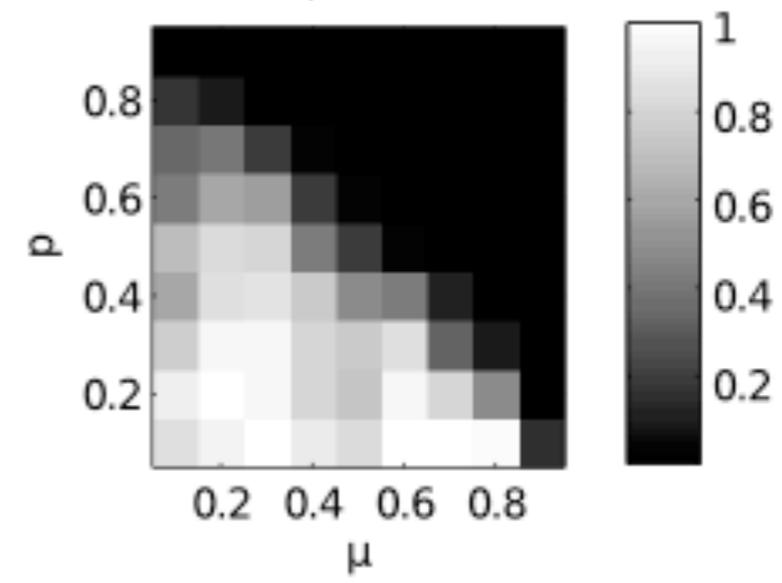
ground truth=local min



ground truth=global min

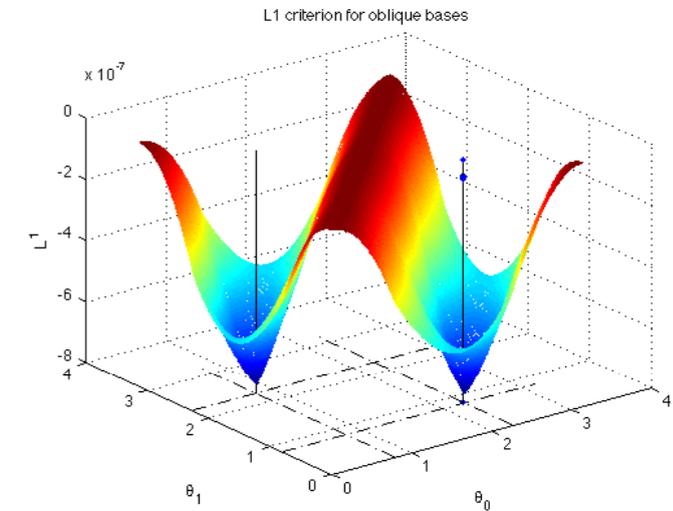


no spurious local min

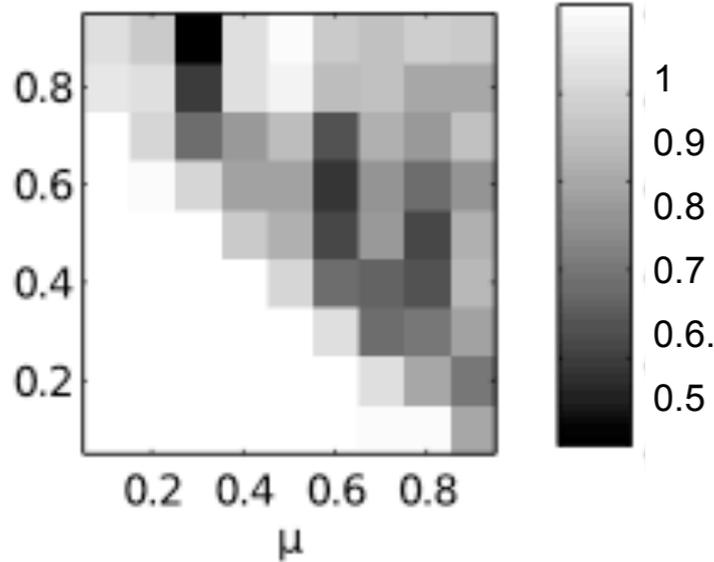
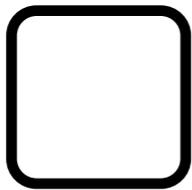


From local to global guarantees ?

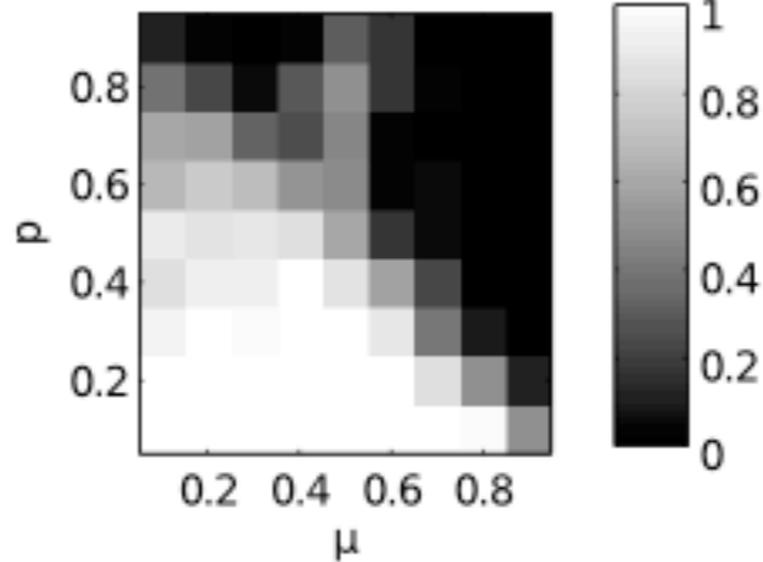
$$\hat{\mathbf{D}} = \arg \min_{\mathbf{D} \in \mathcal{D}} F_{\mathbf{X}}(\mathbf{D})$$



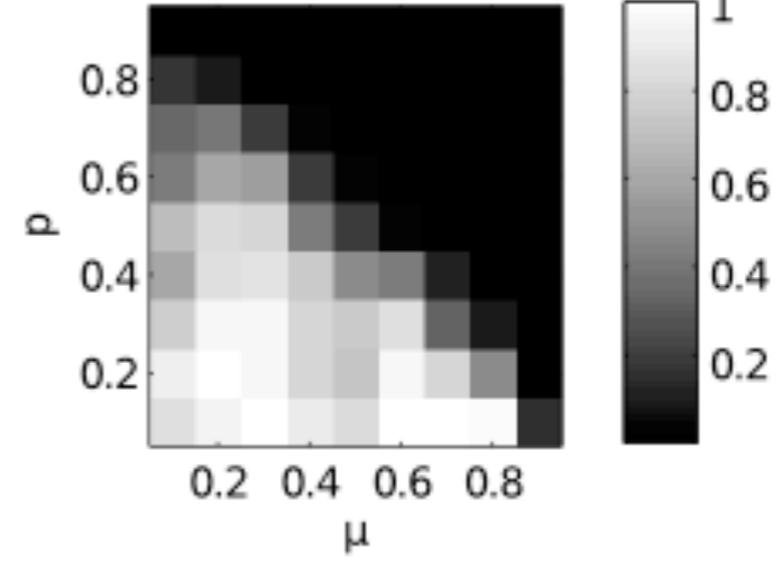
ground truth=local min



ground truth=global min



no spurious local min



To conclude ...

Summary

- **Sparse Dictionary Learning**

- ✓ widely used in image processing and machine learning
- ✓ from **heuristics** ...
 - ◆ online algorithms, empirically successful
- ✓ ... to **statistics**
 - ◆ local stability and robustness guarantees
 - ◆ <http://hal.inria.fr/hal-00737152> [Jenatton, G. & Bach, Local stability and robustness of sparse dictionary learning in the presence of noise, Oct 2012]

What's next ?

- **Immediate challenges**
 - ✓ global guarantees ? empirically yes
 - ✓ sharp sample complexity
 - ✓ guarantees from cost functions to *algorithms*
- **Sparse learning beyond dictionaries**
 - ✓ synthesis / analysis flavor (e.g. TV-like)
 - ✓ structured models (shift-invariance, etc.)
 - ✓ structured sparsity (e.g. trees, graphs)
- **More examples = less work to learn ?**

THANKS



PLEASE

projection, learning and sparsity for efficient data processing