Concentration of measure in probability and high-dimensional statistical learning

Guillaume Aubrun, Aurélien Garivier, Rémi Gribonval

remi.gribonval@inria.fr

http://perso.ens-lyon.fr/remi.gribonval

Why dimension reduction ?

- High dimensions imply high costs
  - ✓ storage
  - ✓ computation
    - ✦ e.g.: computing inner products
- Dimension reduction: choose a « good » mapping

$$E : \mathbb{R}^d \to \mathbb{R}^k, \ k \ll d$$

  - ✓ How to characterize its « goodness » ?
  - ✓ How to « choose » it ?

Table of contents

- Dimension reduction with PCA and its limitations
- Notion of sparsity
- Dimension reduction with random projections

Dimension reduction with PCA

Principal component analysis

- **Goal**: project on *k*-dimensional subspace *V* with min error
  - ✓ mapping = coordinates of an orthoprojection in an orthobasis of *V*

$$E : \mathbb{R}^d \to \mathbb{R}^k, \;\; k \ll d$$

- **Expression as an (unsupervised) learning problem**
  - ✓ training samples $x_1, \ldots, x_n \in \mathbb{R}^d$
  - ✓ hypothesis class $\mathcal{H} = \{V \text{ subspace of } \mathbb{R}^d, \dim(V) = k\}$
  - ✓ loss function $\ell(x, h) = \|x - P_V(x)\|_2^2$

- **Empirical risk**

$$\hat{\mathcal{R}}_n(V) = \frac{1}{n} \sum_i \|x_i - P_V x_i\|_2^2 = \frac{1}{n} \sum_i \|x_i\|_2^2 - \frac{1}{n} \sum_i \|P_V x_i\|_2^2$$

PCA via eigenvalue decomposition

- **ERM equivalent to** $\max_V \sum_{i=1}^n \|P_V x_i\|_2^2$

$$\sum_{i=1}^n \|P_V x_i\|_2^2 = \sum_{i=1}^n \langle P_V x_i, P_V x_i \rangle = \sum_{i=1}^n \mathtt{trace}(x_i^\top P_V^\top P_V x_i)$$
$$= \sum_{i=1}^n \mathtt{trace}(x_i x_i^\top P_V^\top P_V) = \mathtt{trace}\left(\left(\sum_i x_i x_i^\top \right) P_V^\top P_V \right)$$

- **Eigenvectors of empirical autocorrelation matrix**

$$\hat{\mathrm{C}} := \frac{1}{n} \sum_i x_i x_i^\top \qquad = \sum_i \lambda_i \mathbf{u}_i \mathbf{u}_i^\top ; \lambda_1 \geq \ldots \geq \lambda_n \geq 0$$

  - for centered data, this is also the empirical covariance matrix

- **Finding *an* optimum subspace**

$$\hat{V}_n := \mathrm{span}(\mathbf{u}_i, 1 \leq i \leq k)$$

How « good » is PCA ?

- **« Encoder » = dimension reducing map**

$$E : \mathbb{R}^d \to \mathbb{R}^k, \ k \ll d$$
$$x \mapsto (\langle x, \mathbf{u}_i \rangle)_{i=1}^k$$

- **« Decoder » = reconstruction map**

$$D : \mathbb{R}^k \to \mathbb{R}^d$$
$$y \mapsto \sum_{i=1}^{k} y_i \mathbf{u}_i$$

- **Encoding + decoding = lossy process**
  - perfect reconstruction iff x belongs to V
  - good reconstruction iff x is close to V
  - however if *x* orthogonal to *V* then $\|D[E(x)] - x\|_2 = \|x\|_2$

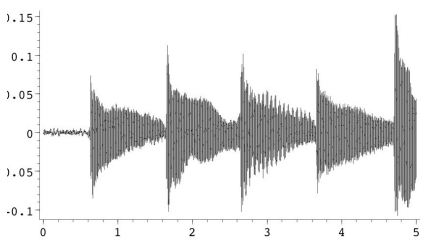
Cases where PCA is not adapted

- **Data with i.i.d. white Gaussian centered entries**
  - $x \sim \mathcal{N}(0, \sigma^2 \mathbf{Id})$
  - Autocorrelation $\hat{C} \approx C = \sigma^2 \mathbf{Id}$
  - Relative expected error can be shown to be $\frac{\mathbb{E}\|x - P_{\hat{V}_n} x\|_2^2}{\mathbb{E}\|x\|_2^2} \approx 1 - \frac{k}{d}$
- **Data with sparse entries, such as**
  - text data (e.g. x = histogram of word appearances in a document)
  - wavelet transform of images, time-frequency representation of audio
  - …

Notion of sparsity

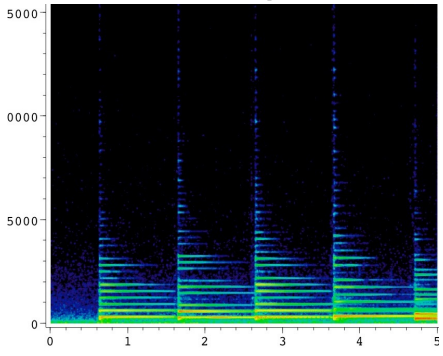
Notion of sparse representation

- Audio : time-frequency representations (MP3 etc.)



ANALYSIS

SYNTHESIS

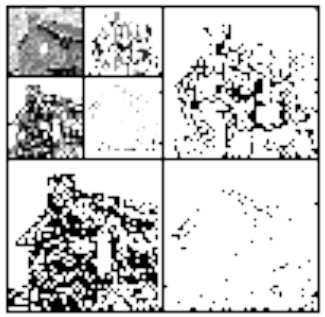


Black = zero

- Images : wavelet transform

ANALYSIS

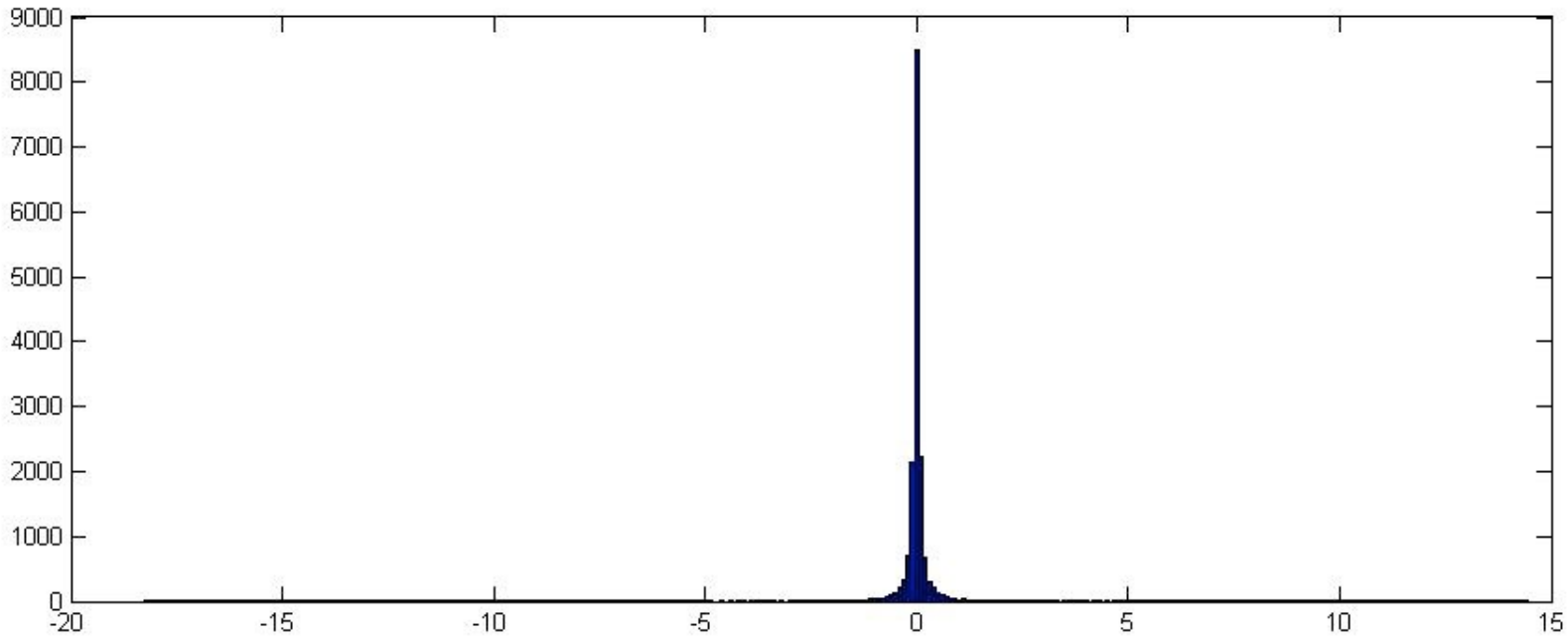
SYNTHESIS



White = zero

Evidence of sparsity

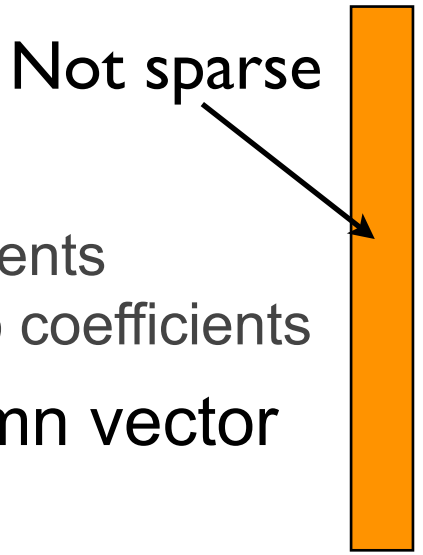
- Histogram of MDCT coefficients of a musical



Sparsity: definitions

- A vector is
  - ✓ **sparse** if it has (many) zero coefficients
  - ✓ ***k*-sparse** if it has *at most k* nonzero coefficients

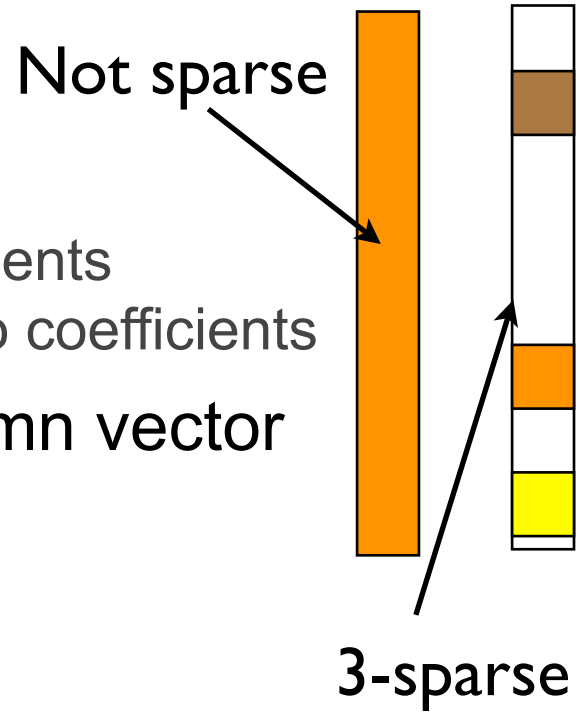
Sparsity: definitions



- A vector is
  - ✓ **sparse** if it has (many) zero coefficients
  - ✓ ***k*-sparse** if it has *at most k* nonzero coefficients
- Symbolic representation as column vector

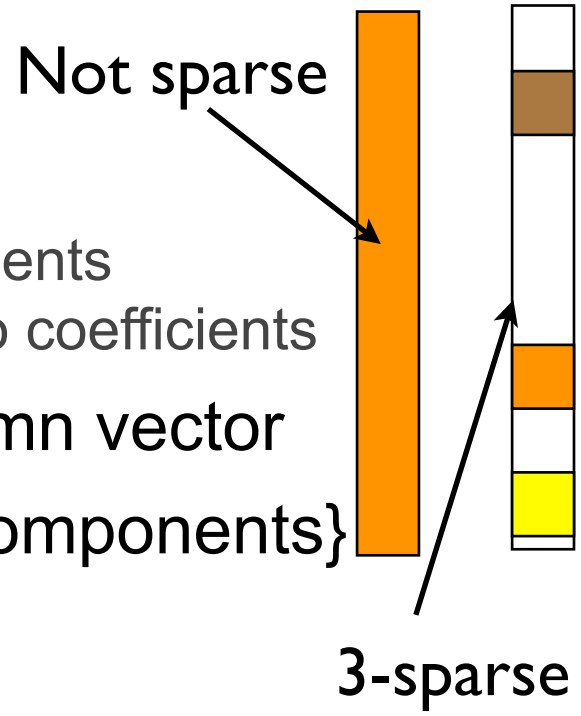
Sparsity: definitions

- A vector is
  - ✓ **sparse** if it has (many) zero coefficients
  - ✓ ***k*-sparse** if it has *at most k* nonzero coefficients
- Symbolic representation as column vector



Sparsity: definitions

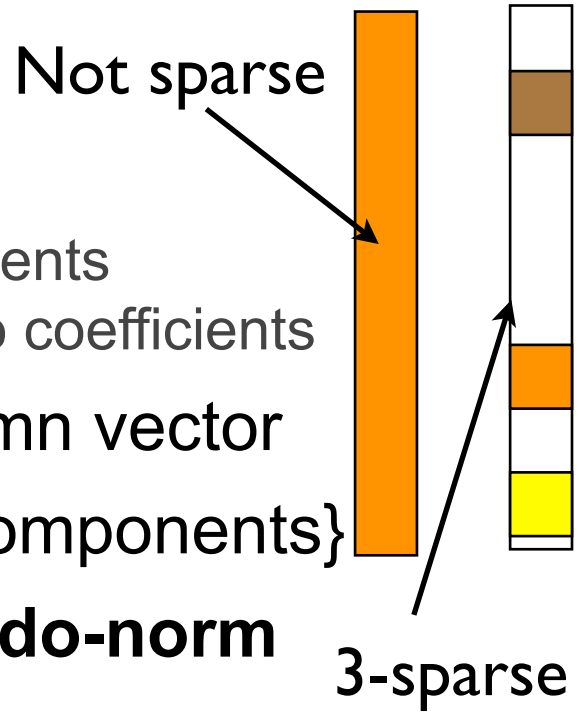
- A vector is
  - ✓ **sparse** if it has (many) zero coefficients
  - ✓ ***k*-sparse** if it has *at most k* nonzero coefficients
- Symbolic representation as column vector
- **Support** = {indices of nonzero components}



Sparsity: definitions

- A vector is
  - ✓ **sparse** if it has (many) zero coefficients
  - ✓ ***k*-sparse** if it has *at most k* nonzero coefficients
- Symbolic representation as column vector
- **Support** = {indices of nonzero components}
- Sparsity measured with **L0 pseudo-norm**

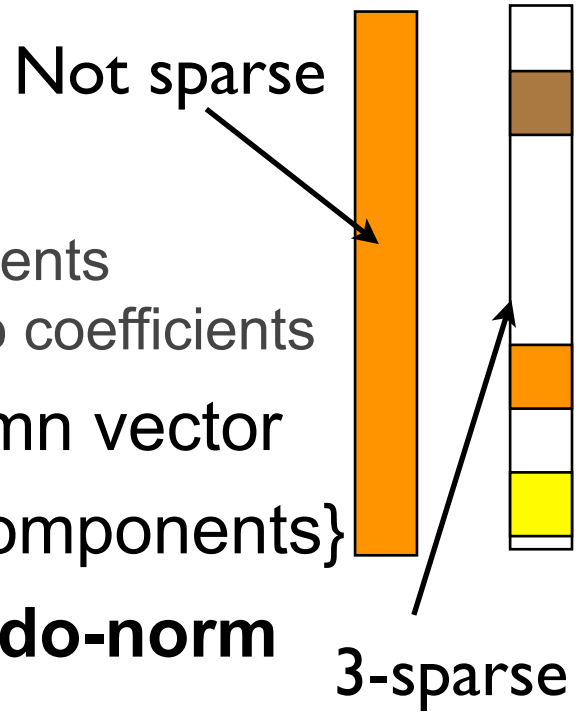
$$\|x\|_0 := \sharp\{j,\ x_j \neq 0\} = \sum_j |x_j|^0$$



Convention here

$$a^0 = 1(a > 0); 0^0 = 0$$

Sparsity: definitions



- A vector is
  - ✓ **sparse** if it has (many) zero coefficients
  - ✓ ***k*-sparse** if it has *at most k* nonzero coefficients
- Symbolic representation as column vector
- **Support** = {indices of nonzero components}
- Sparsity measured with **L0 pseudo-norm**

$$\|x\|_0 := \sharp\{j,\ x_j \neq 0\} = \sum_j |x_j|^0$$

Convention here

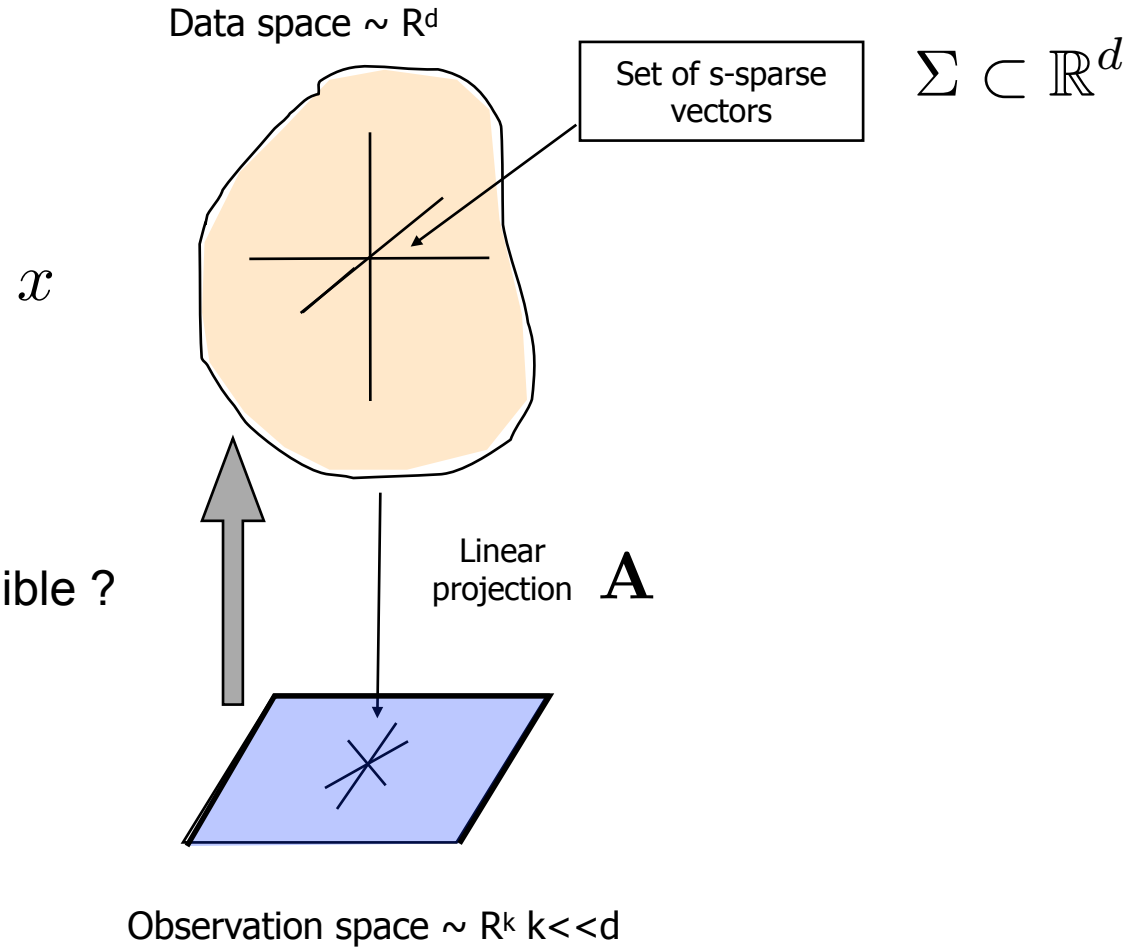
$$a^0 = 1(a > 0); 0^0 = 0$$

- *In french:*
  - ✦ sparse -> *«creux», «parcimonieux»*
  - ✦ sparsity, sparseness -> *«parcimonie»,* ~~*«sparsité»*~~

Dimension reduction and sparsity

- **Nonlinear dimension reduction**
  - ✓ Encoder: keep *k* largest entries *and their indices*
  - ✓ *Decoder: fill with zeroes unknown entries*
  - ✓ *Pros:*
    - ✦ *perfect reconstruction of sparse vectors*
    - ✦ *good approximation of « compressible » vectors (close to sparse)*
  - ✓ *Cons: nonlinear*
- **Is *linear* dimension reduction possible ?**
  - ✓ Encoder: some *kxd* matrix **A** to be chosen
  - ✓ Decoder: ???

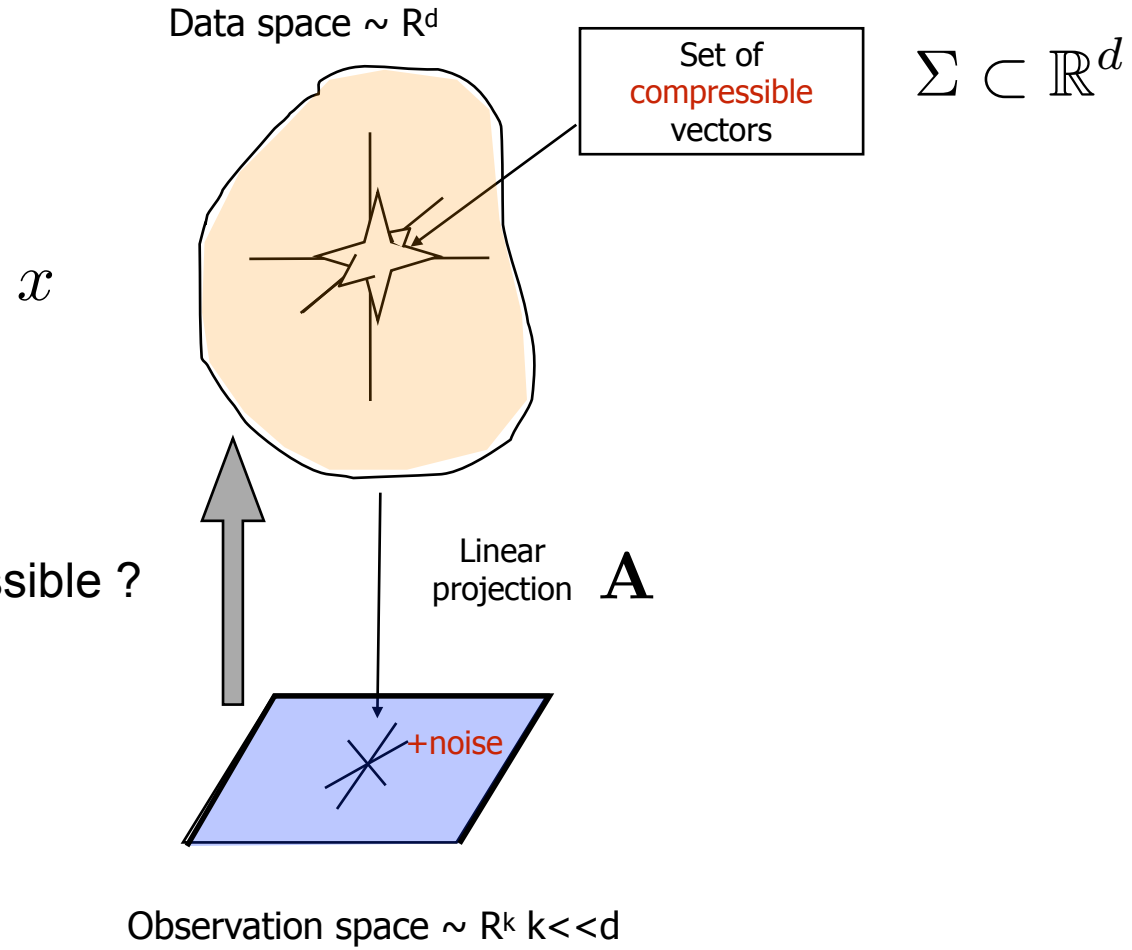
Inverse problems with sparse vectors



Is reconstruction possible ?

Courtesy: M. Davies, U. Edinburgh

Inverse problems with sparse vectors



Is stable reconstruction possible ?

Courtesy: M. Davies, U. Edinburgh

Inverse problems with sparse vectors

Instance optimality

- **Setting**
  - ✓ *X, Y*: normed spaces equipped with $\|\cdot\|_X, \|\cdot\|_Y$
  - ✓ d: metric on *X*
  - ✓ Model set $\Sigma \subset X$
  - ✓ Encoder: $E : X \to Y$
  - ✓ Decoder: $D : Y \to X$
- **Definition**: D is instance optimal wrt (X, Y, d, Σ, E) if there exists finite constants C,C' such that

$$\forall x \in X, \forall y \in Y, \|x - D[E(x) + y)]\|_X \leq C d(x, \Sigma) + C' \|y\|_Y$$

Lower Restricted Isometry Property

- **Theorem**
  - ✓ if there exists an instance optimal decoder *D* wrt (X, Y, d, Σ, E) then a ``lower restricted isometry property'' (LRIP) holds

$$\forall x, x' \in \Sigma, \|x - x'\|_X \leq C' \|E(x) - E(x')\|_Y$$

  - ✓ the LRIP implies the instance optimality of the decoder

$$D(y) := \arg\min_{x \in \Sigma} \|y - E(x)\|_Y$$

    - ✦ with the following metric $\forall x, x' \in X,\ \tilde{d}(x, x') := \|x - x'\|_X + 2C' \|E(x) - E(x')\|_Y$
    - ✦ assuming argmin is achievable, otherwise adaptations exist

Proof

Linear encoders and the LRIP

- **The case of linear encoders:**
  - ✓ the LRIP then reads

$$\forall z \in \Sigma - \Sigma, \|z\|_X \leq C' \|E(z)\|_Y$$

    - ✦ with the *secant set* $\Sigma - \Sigma = \{x - x' : \ x, x' \in \Sigma\}$
- **The case of s-sparse vectors**

$$\Sigma = \{x \in \mathbb{R}^d : \ \|x\|_0 \leq s\} \qquad \Sigma - \Sigma = \{z \in \mathbb{R}^d : \ \|z\|_0 \leq 2s\}$$

    - ✦ with Euclidean norms, a *two-sided version* of the LRIP is **the RIP**

$$\forall z \text{ s.t. } \|z\|_0 \leq 2s, 1 - \delta \leq \frac{\|E(z)\|_2^2}{\|z\|_2^2} \leq 1 + \delta$$

    - ✦ with small enough δ it implies instance optimality of tractable decoders

RIP constant

- **Best RIP constant of a linear encoder**
  - ✓ Given a model set

$$\delta^\star = \delta^\star(E, \Sigma) = \sup_{z \in \Sigma - \Sigma} \left| \frac{\|E(z)\|_2^2}{\|z\|_2^2} - 1 \right| = \sup_{u \in \mathcal{S}} \left| \|E(u)\|_2^2 - 1 \right|$$

    - ✦ with *normalized secant set* $\mathcal{S} = \left\{ \frac{z}{\|z\|_2} : \; z \in \Sigma - \Sigma \right\} \subset \mathbb{S}^{d-1}$
  - ✓ Straightforward computation for linear model sets
    - ✦ model set = linear subspace *V* with orthonormal *dxs* basis matrix ***B***
  - ✓ NP-hard to compute for *s*-sparse model set
    - ✦ naive algorithm = combinatorial search among all 2s-sparse supports
  - ✓ Tractable bounds available in some cases

Limits of dimension reduction

- Given model set $\Sigma \subset \mathbb{R}^d$, for which dimension k does there exist an encoder $E : \mathbb{R}^d \to \mathbb{R}^k$ with

$$\delta^\star(E, \Sigma) < 1$$

- Examples of model sets
  - ✓ linear model set — $k \geq \dim(\Sigma)$
  - ✓ finite set — $k \geq C \log \sharp\Sigma$
  - ✓ {*s*-sparse vectors} — $k \geq Cs \log(d/s)$
  - ✓ {rank-*r* matrices} $d = p \times p$ — $k \geq Crp$

  - ✦ How to build such encoders ? Where does the « dimensions » come from ?

Limits of dimension reduction

- Given model set $\Sigma \subset \mathbb{R}^d$, for which dimension k does there exist an encoder $E : \mathbb{R}^d \to \mathbb{R}^k$ with

$$\delta^\star(E, \Sigma) < 1$$

- Examples of model sets
  - ✓ linear model set — $k \geq \dim(\Sigma)$
  - ✓ finite set — Johnson-Lindenstrauss lemma — $k \geq C \log \sharp \Sigma$
  - ✓ {*s*-sparse vectors} — $k \geq Cs \log(d/s)$
  - ✓ {rank-*r* matrices} $d = p \times p$ — $k \geq Crp$

    - ✦ How to build such encoders ? Where does the « dimensions » come from ?

Limits of dimension reduction

- Given model set $\Sigma \subset \mathbb{R}^d$, for which dimension k does there exist an encoder $E : \mathbb{R}^d \to \mathbb{R}^k$ with

$$\delta^\star(E, \Sigma) < 1$$

- Examples of model sets
  - ✓ linear model set — $k \geq \dim(\Sigma)$
  - ✓ finite set — Johnson-Lindenstrauss lemma — $k \geq C \log \sharp \Sigma$
  - ✓ {*s*-sparse vectors} — $k \geq Cs \log(d/s)$
  - ✓ {rank-*r* matrices} $d = p \times p$ — $k \geq Crp$

  - ✦ How to build such encoders ? Where does the « dimensions » come from ?

    random projections

Limits of dimension reduction

- Given model set $\Sigma \subset \mathbb{R}^d$, for which dimension k does there exist an encoder $E : \mathbb{R}^d \to \mathbb{R}^k$ with

$$\delta^\star(E, \Sigma) < 1$$

- Examples of model sets
  - ✓ linear model set — $k \geq \dim(\Sigma)$
  - ✓ finite set — Johnson-Lindenstrauss lemma — $k \geq C \log \sharp \Sigma$
  - ✓ {*s*-sparse vectors} — $k \geq Cs \log(d/s)$
  - ✓ {rank-*r* matrices} $d = p \times p$ — $k \geq Crp$

  - ✦ How to build such encoders ? Where does the « dimensions » come from ?

random projections — covering numbers

That’s all folks !