Compressive Classification: A Guided Tour

Dr. Valerio Cambareri
Image and Signal Processing Group,
ICTEAM/ELEN, Université Catholique de Louvain

Image and Signal Processing Seminars,
Thursday, 24th September 2015

FREEDOM TO RESEARCH

UCL
Université catholique de Louvain

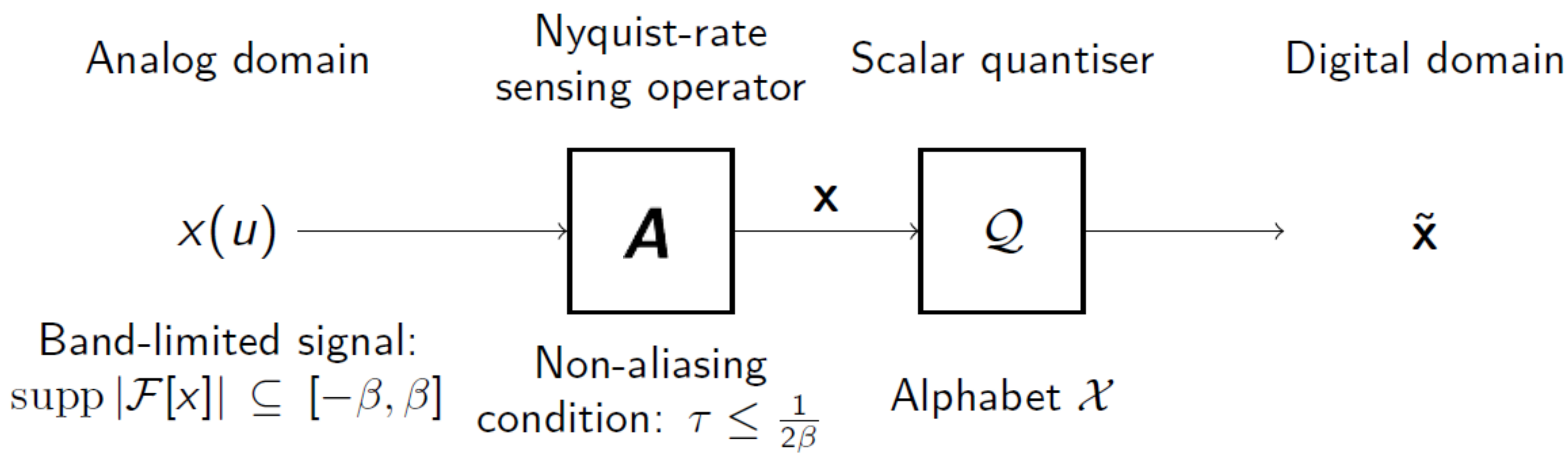
About me

- M.Sc. in Electronic Engineering (2011), Univ. of Bologna, IT.
- Ph.D. in Electronics, Telecommunications and Information Technologies (2012 - mid-2015; advisors: Prof. R. Rovatti, Prof. G. Setti), Univ. of Bologna, IT.
  - Thesis: *"Matrix Designs and Methods for Secure and Efficient Compressed Sensing"*.
  - Sensing matrix adaptation for Compressed Sensing of (wide-sense cyclostationary) correlated and compressible signals (*e.g.*, ECG).
  - Security analysis of Compressed Sensing: statistical and computational attacks; application-level analysis for private tele-monitoring (*e.g.*, ECG).
  - Design of a compressive hyperspectral imager (joint with IMEC and UCLouvain, BE).
- Postdoctoral researcher (mid-2015 - now; under FRS-FNRS Project "AlterSense"; P.I.: Prof. L. Jacques), UCLouvain, BE.
  - Blind calibration via non-convex optimisation (*e.g.*, for unmatched compressive sensor arrays).
  - Compressive classification with quantisation: theory and prospective applications.

Outline

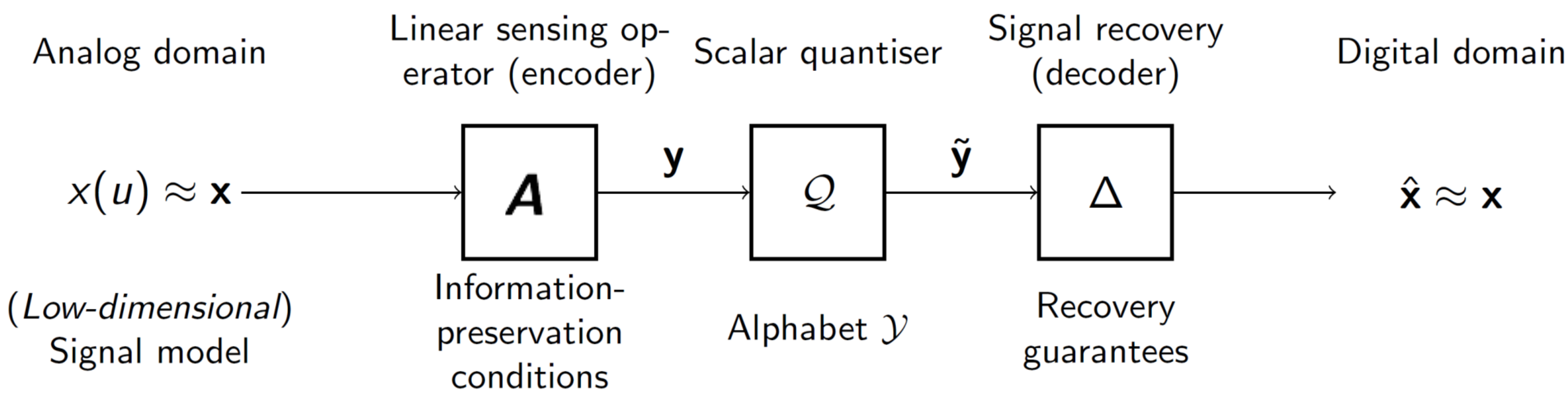
- **Introduction:** from Compressed Sensing to Compressive Classification
- **Compressive classification of finite sets** *(from M. Davenport* et al., *2010)*
  - Random matrices and stable embeddings
  - Compressive classification via p-ary hypothesis testing
- **Compressive classification of disjoint convex sets** *(from A. Bandeira* et al., *2014)*
  - Compressive classification of linearly separable classes
    - The Gaussian width of a set: a measure of "intrinsic complexity"
    - "Escape through a mesh": Gordon's theorem
    - Minimum projection rank (for linear separability)
  - The case of two disjoint Euclidean balls
  - The case of two disjoint ellipsoids
  - A comparison with PCA: adaptive *versus* non-adaptive dimensionality reduction
- **Conclusion**

Introduction

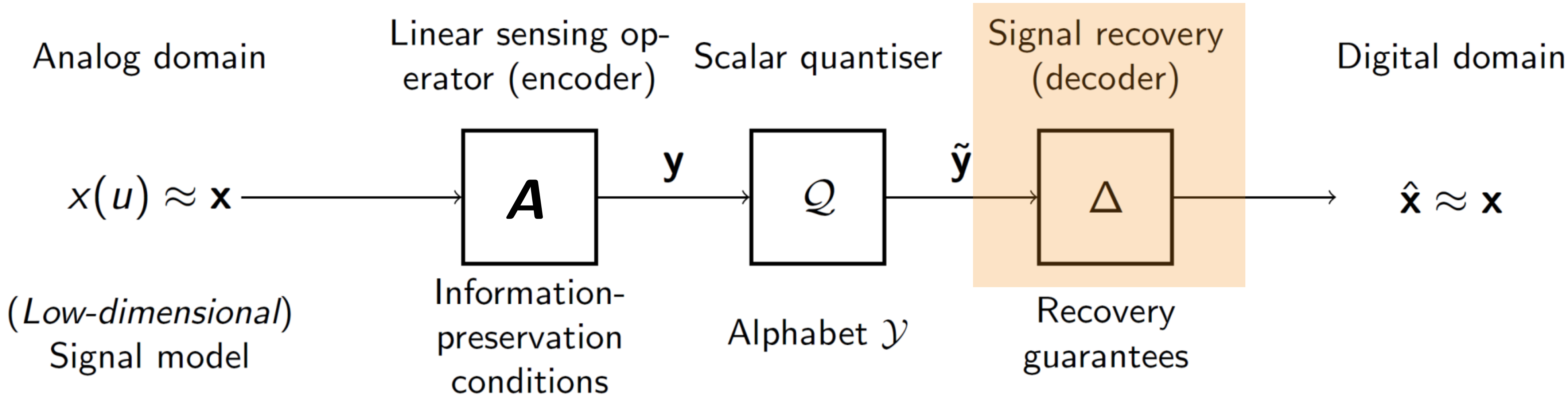


- **Compressed Sensing** (CS), ca. 2005: a mature framework [1] for non-adaptive acquisition of analog signals ("analog-to-information conversion").
  - The sensing interface is implemented and modelled as a *dimensionality reduction* w.r.t. the n-dim. Nyquist-rate representation $\boldsymbol{x}$ of an analog signal.
  - From the m-dim. *compressive measurements* $\boldsymbol{y} = \boldsymbol{A}\,\boldsymbol{x}$, recover an approximation by means of an optimisation algorithm enforcing a *low-complexity* (low-dimensional) *prior model,* $\boldsymbol{x} \in \mathcal{K}$. Example: k-sparse signals, $\mathcal{K} = \Sigma_k \subset \mathbb{R}^n$.
  - If $\boldsymbol{x}$ complies with such prior model (and its complexity is *sufficiently low*, *e.g.*, very sparse), then *exact signal recovery* is possible.

Introduction



- **Compressed Sensing** (CS), ca. 2005: a mature framework [1] for non-adaptive acquisition of analog signals ("analog-to-information conversion").
  - The sensing interface is implemented and modelled as a *dimensionality reduction* w.r.t. the n-dim. Nyquist-rate representation $\boldsymbol{x}$ of an analog signal.
  - From the m-dim. *compressive measurements* $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$, recover an approximation by means of an optimisation algorithm enforcing a *low-complexity* (low-dimensional) *prior model*, $\boldsymbol{x} \in \mathcal{K}$. Example: k-sparse signals, $\mathcal{K} = \Sigma_k \subset \mathbb{R}^n$.
  - If $\boldsymbol{x}$ complies with such prior model (and its complexity is *sufficiently low*, *e.g.*, very sparse), then *exact signal recovery* is possible.

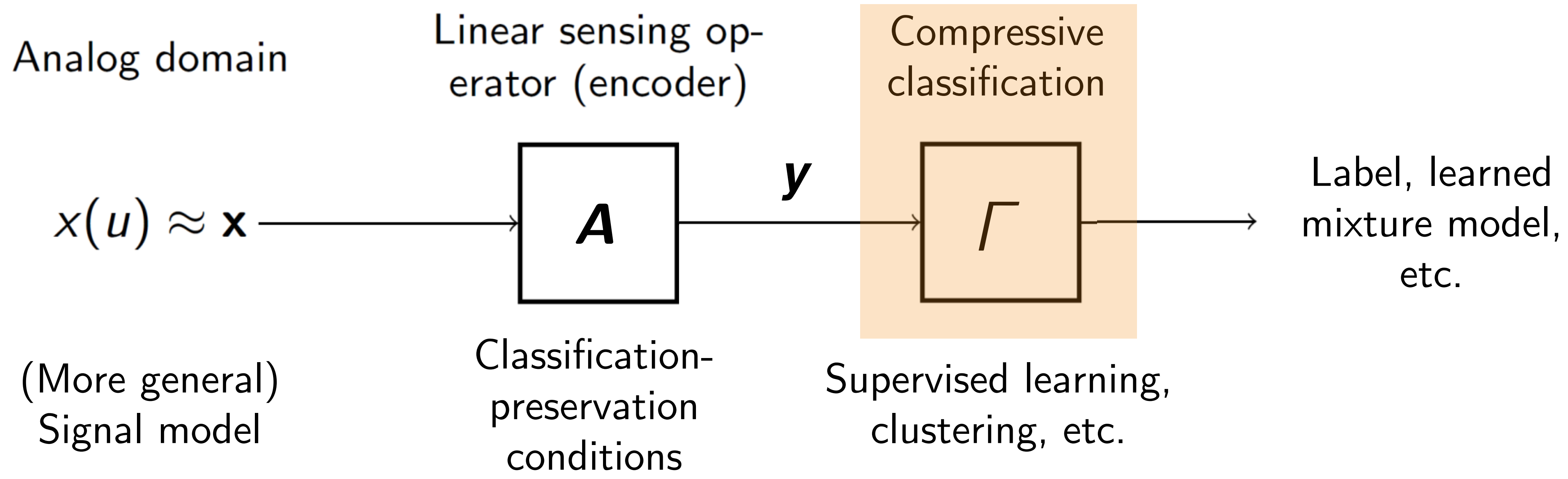


Proposition (Exact signal recovery, loosely based on Corollary 3.3 [2]).

Let $w(\mathcal{K})$ denote a *measure of complexity* of the prior model; let the random sensing matrix $\boldsymbol{A}_{m \times n} \sim \mathcal{D}$ follow a suitable distribution $\mathcal{D}$; then $\hat{\boldsymbol{x}} = \Delta(\boldsymbol{y}, \mathcal{K}) \equiv \boldsymbol{x}$ provided that $m \geq m^\star$, $m^\star = O(w(\mathcal{K})^2)$.

- Is (exact) signal recovery *really* required?
- If signal processing in the digital domain amounts to *detection*, *estimation*, *classification* or *filtering*, can we perform analogous operations on $\boldsymbol{y}$ rather than $\boldsymbol{x}$ with compatible results?

Compressive Classification



- With which accuracy can we perform *classification in the compressed domain* ? How does m affect the classification error?
- How does a random matrix $\boldsymbol{A}$ differ w.r.t. classical dimensionality reduction (*e.g.*, PCA)? ($\boldsymbol{A}$ is a non-adaptive, *universal* dimensionality reduction method.)
- What models of $\boldsymbol{x}$ can be **provably** classified with high probability in the compressed domain?
    - Finite sets [3,4], disjoint spheres and ellipsoids [5], mixtures of sufficiently separated Gaussians [6,7], . . .

Stable Embeddings and Random Matrices

- The classic framework of CS leverages *stable embeddings* to construct *distance-preserving mappings* w.r.t. the chosen signal set (*i.e.*, RIP for sparse signals [8]).
- The following Definition and Lemma summarise Johnson-Lindenstrauss [9], consequent proofs and applications [10,11] in the fashion of [3,8].

Definition (ε-stable embedding)**.**

Let $\varepsilon \in (0,1)$ and $\boldsymbol{x}', \boldsymbol{x}'' \in \mathcal{S} \subset \mathbb{R}^n$; $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ is a ε-stable embedding of $\mathcal{S}$ if, $\forall \boldsymbol{x}', \boldsymbol{x}'' \in \mathcal{S}$,

$$(1-\varepsilon)\left\|\boldsymbol{x}' - \boldsymbol{x}''\right\|_2^2 \leq \left\|\boldsymbol{A}(\boldsymbol{x}' - \boldsymbol{x}'')\right\|_2^2 \leq (1+\varepsilon)\left\|\boldsymbol{x}' - \boldsymbol{x}''\right\|_2^2$$

Lemma (Johnson-Lindenstrauss [9-11])**.**

Let $\mathcal{S} \subset \mathbb{R}^n$, $p = |\mathcal{S}|$; let $\boldsymbol{A}_{m \times n} \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \frac{1}{m}\right)$ and $\varepsilon, \eta \in (0,1)$. If $m \geq m^\star$,

$$m^\star = c\varepsilon^{-2}\left(\log(p) + \log\left(\frac{2}{\eta}\right)\right), c > 0$$

then $\boldsymbol{A}$ is a ε-stable embedding of $\mathcal{S}$ with probability $1 - \eta$.

Stable Embeddings and Random Matrices

- The classic framework of CS leverages *stable embeddings* to construct *distance-preserving mappings* w.r.t. the chosen signal set (*i.e.*, RIP for sparse signals [8]).
- The following Definition and Lemma summarise Johnson-Lindenstrauss [9], consequent proofs and applications [10,11] in the fashion of [3,8].

Definition (ε-stable embedding).

Let $\varepsilon \in (0,1)$ and $x', x'' \in \mathcal{S} \subset \mathbb{R}^n$; $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ is a ε-stable embedding of $\mathcal{S}$ if, $\forall x', x'' \in \mathcal{S}$,

$$(1-\varepsilon)\left\|x'-x''\right\|_2^2 \le \left\|\boldsymbol{A}(x'-x'')\right\|_2^2 \le (1+\varepsilon)\left\|x'-x''\right\|_2^2$$

Lemma (Johnson-Lindenstrauss [9-11]).

Let $\mathcal{S} \subset \mathbb{R}^n$, $p = |\mathcal{S}|$; let $\boldsymbol{A}_{m\times n} \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \frac{1}{m}\right)$ and $\varepsilon, \eta \in (0,1)$. If $m \ge m^\star$,

$$m^\star = c\varepsilon^{-2}\left(\log(p) + \log\left(\frac{2}{\eta}\right)\right), c > 0$$

then $\boldsymbol{A}$ is a ε-stable embedding of $\mathcal{S}$ with probability $1-\eta$.

Stable Embeddings and Random Matrices

- The classic framework of CS leverages *stable embeddings* to construct *distance-preserving mappings* w.r.t. the chosen signal set (*i.e.*, RIP for sparse signals [8]).
- The following Definition and Lemma summarise Johnson-Lindenstrauss [9], consequent proofs and applications [10,11] in the fashion of [3,8].

Definition (ε-stable embedding)**.**

We defined the property of $\boldsymbol{A}$ to preserve, within a small multiplicative distortion, the pairwise distances between all points in a finite set.

L… ding of $\mathcal{S}$ if, $\forall$…

$$(1-\varepsilon)\left\|x'-x''\right\|_2^2 \leq \left\|\boldsymbol{A}(x'-x'')\right\|_2^2 \leq (1+\varepsilon)\left\|x'-x''\right\|_2^2$$

Lemma (Johnson-Lindenstrauss [9-11])**.**

Let $\mathcal{S} \subset \mathbb{R}^n$, $p = |\mathcal{S}|$; let $\boldsymbol{A}_{m\times n} \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \frac{1}{m}\right)$ and $\varepsilon, \eta \in (0,1)$. If $m \geq m^\star$,

$$m^\star = c\varepsilon^{-2}\left(\log(p) + \log\left(\frac{2}{\eta}\right)\right), c > 0$$

then $\boldsymbol{A}$ is a ε-stable embedding of $\mathcal{S}$ with probability $1-\eta$.

Stable Embeddings and Random Matrices

- The classic framework of CS leverages *stable embeddings* to construct *distance-preserving mappings* w.r.t. the chosen signal set (*i.e.*, RIP for sparse signals [8]).
- The following Definition and Lemma summarise Johnson-Lindenstrauss [9], consequent proofs and applications [10,11] in the fashion of [3,8].

Definition (ε-stable embedding)**.**

We defined the property of $\boldsymbol{A}$ to preserve, within a small multiplicative distortion, the pairwise distances between all points in a finite set.

Let $\varepsilon \in (0,1)$ and $x', x'' \in \mathcal{S} \subset \mathbb{R}^n$; $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ is a ε-stable embedding of $\mathcal{S}$ if, $\forall x', x'' \in \mathcal{S}$

$$(1-\varepsilon)\left\|x'-x''\right\|_2^2 \le \left\|\boldsymbol{A}(x'-x'')\right\|_2^2 \le (1+\varepsilon)\left\|x'-x''\right\|_2^2$$

Lemma (Johnson-Lindenstrauss [9-11])**.**

There is *always* a rank $m = O(\log(p))$ linear transformation $\boldsymbol{A}$ (in particular, random matrices with i.i.d. (sub-)Gaussian entries) that is a stable embedding of a high-dimensional finite set of size p.

Let $\mathcal{S} \subset \mathbb{R}^n$, $p = |\mathcal{S}|$; let $\boldsymbol{A}_{m \times n} \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \frac{1}{m}\right)$ and $\varepsilon, \eta \in (0,1)$. If $m \ge m^\star$,

$$m^\star = c\varepsilon^{-2}\left(\log(p) + \log\left(\frac{2}{\eta}\right)\right), c > 0$$

then $\boldsymbol{A}$ is a ε-stable embedding of $\mathcal{S}$ with probability $1-\eta$.

Compressive Classification of Finite Sets

- Using stable embeddings and a standard p-ary hypothesis testing framework, Davenport *et al.* [3] establish a bound on the probability of error of a simple compressive classifier.
- Let $\mathcal{S} = \{\boldsymbol{s}_i\}_{i=1}^{p}$ be a set of reference vectors; let $\boldsymbol{x} = \boldsymbol{s}_i + \boldsymbol{\nu}, \boldsymbol{\nu} \sim \mathcal{N}(0_n, \sigma^2 \boldsymbol{I}_n)$
- For $\boldsymbol{y} = \boldsymbol{A}\ \boldsymbol{x}$ we form p equal-probability hypotheses:

$$H_i : \boldsymbol{y} = \boldsymbol{A}(\boldsymbol{s}_i + \boldsymbol{\nu}),\ i = 1, \ldots, p$$

$$f_{\boldsymbol{y}|H_i,\boldsymbol{A}}(\boldsymbol{y}) = \frac{1}{\sqrt{(2\pi)^m \det \sigma^2 \boldsymbol{A}\boldsymbol{A}^*}} e^{-\frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{A}\boldsymbol{s}_i)^*(\boldsymbol{A}\boldsymbol{A}^*)^{-1}(\boldsymbol{y}-\boldsymbol{A}\boldsymbol{s}_i)},\ i = 1, \ldots, p$$

- The *sufficient statistic* for our test is therefore:

$$t_i = (\boldsymbol{y} - \boldsymbol{A}\boldsymbol{s}_i)^*(\boldsymbol{A}\boldsymbol{A}^*)^{-1}(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{s}_i),\ i = 1, \ldots, p$$

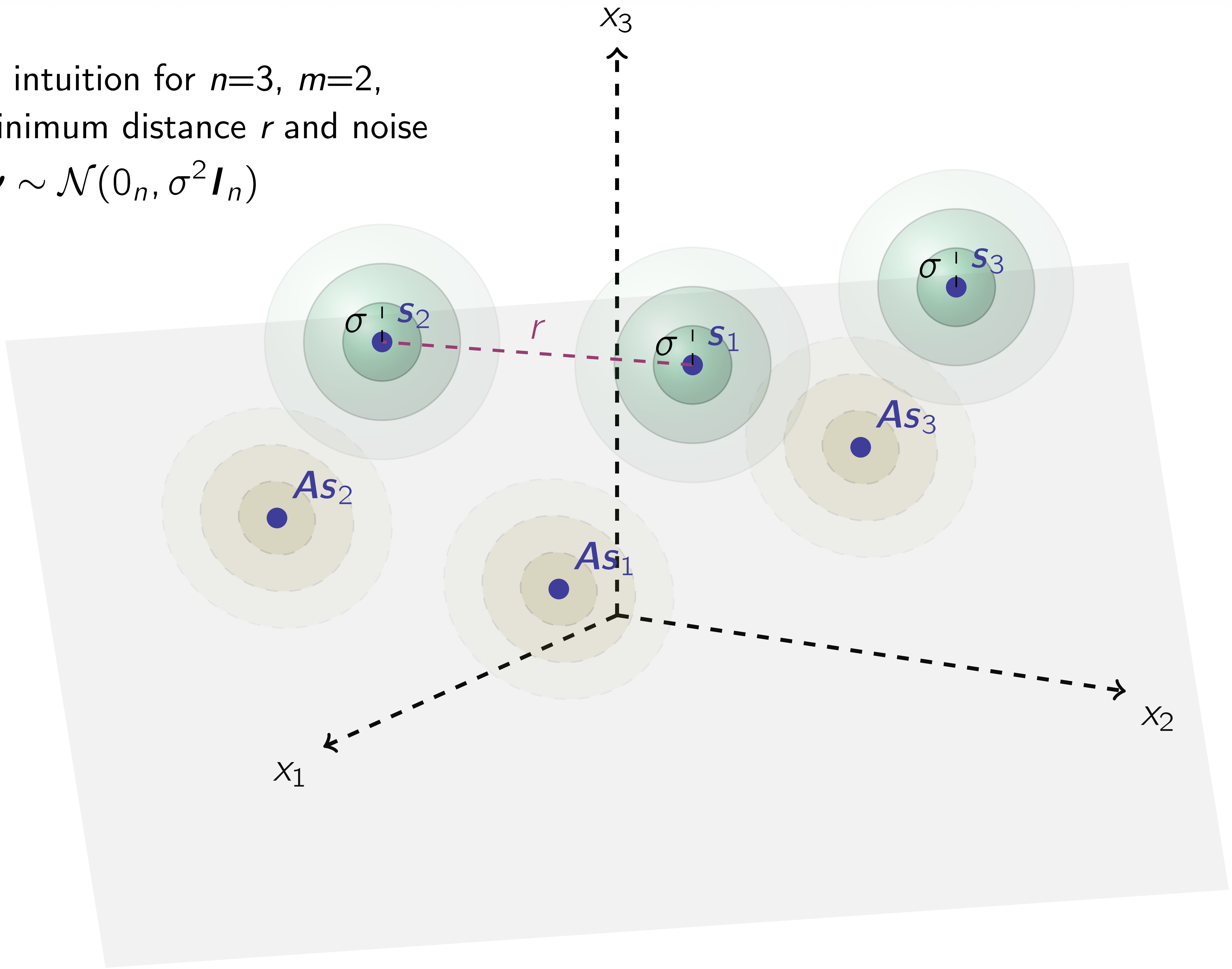
  so we classify $\boldsymbol{y}$ according to the maximum likelihood, *i.e.* (this case),

$$\hat{i} = \operatorname*{argmin}_{i \in [p]} t_i,\ i = 1, \ldots, p$$

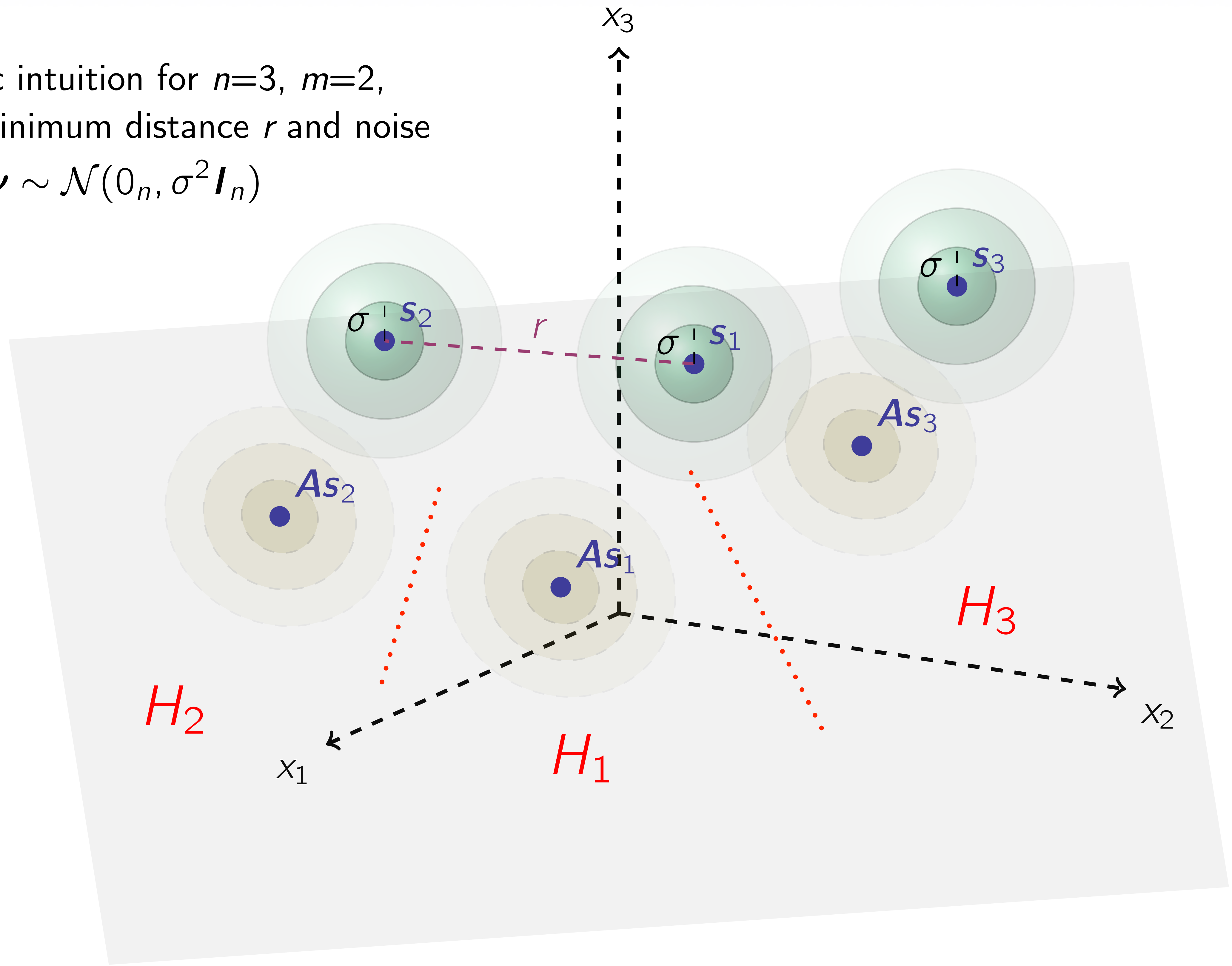
- Note that, asymptotically, $\boldsymbol{A}\boldsymbol{A}^* \simeq \boldsymbol{I}_m \Rightarrow t_i \simeq \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{s}_i\|_2^2 = \|\boldsymbol{A}(\boldsymbol{x} - \boldsymbol{s}_i)\|_2^2$

Compressive Classification of Finite Sets

- Geometric intuition for n=3, m=2, p=3 at minimum distance r and noise
$$\boldsymbol{\nu} \sim \mathcal{N}(0_n, \sigma^2 \boldsymbol{I}_n)$$



- Geometric intuition for n=3, m=2, p=3 at minimum distance r and noise $\boldsymbol{\nu} \sim \mathcal{N}(0_n, \sigma^2 \boldsymbol{I}_n)$



Classification Error in Finite Sets

Theorem (Compressive classification of finite sets, Theorem 3 in [3]).

Let $\boldsymbol{A}_{m\times n} \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \frac{1}{m}\right)$ be a ε-stable embedding of $\mathcal{S}$, $p = |\mathcal{S}|$. Define

$$r = \min_{i \neq j} \|\boldsymbol{s}_i - \boldsymbol{s}_j\|_2$$

and assume the measurements are produced by the $i^\star$-th hypothesis,

$$\boldsymbol{y} = \boldsymbol{A}(\boldsymbol{s}_{i^\star} + \boldsymbol{\nu}), \boldsymbol{\nu} \sim \mathcal{N}(0_n, \sigma^2 \boldsymbol{I}_n)$$

Then the classification error probability $P_e = \mathbb{P}[\hat{i} \neq i^\star]$ of the classifier

$$\hat{i} = \operatorname*{argmin}_{i \in [p]} (\boldsymbol{y} - \boldsymbol{A}\boldsymbol{s}_i)^* (\boldsymbol{A}\boldsymbol{A}^*)^{-1} (\boldsymbol{y} - \boldsymbol{A}\boldsymbol{s}_i)$$

is bounded by

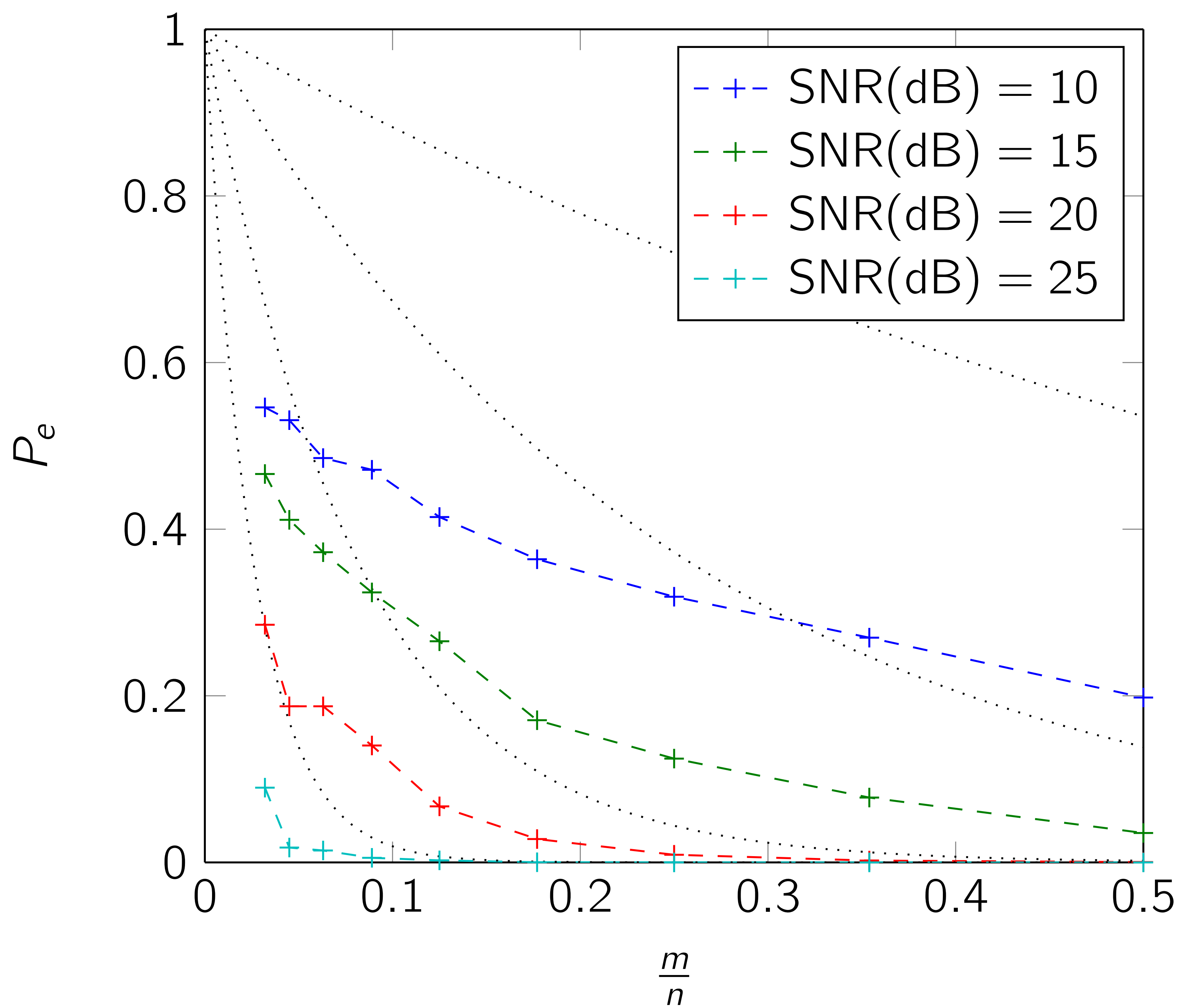
$$P_e \leq \frac{p-1}{2} e^{-\frac{r^2}{\sigma^2}\frac{m}{n}\frac{1-\varepsilon}{8}}$$

- The above bound is proved in [3] by simple inequalities using: the stable embedding assumption; the minimum distance r; a tail bound of the normal distribution; a union bound.

- Set $n = 1000$ and draw $p = 3$ random points in S at fixed minimum distance r.
- Generate random instances of $(S, \boldsymbol{x}, \boldsymbol{A}, \boldsymbol{y})$ according to the p-hypotheses model.
- Classify $\boldsymbol{y}$ with
$$\hat{i} = \operatorname*{argmin}_{i \in [p]} t_i$$
and evaluate $P_e(\frac{m}{n}, \sigma^2)$.
- Note that the noise variance fixes:

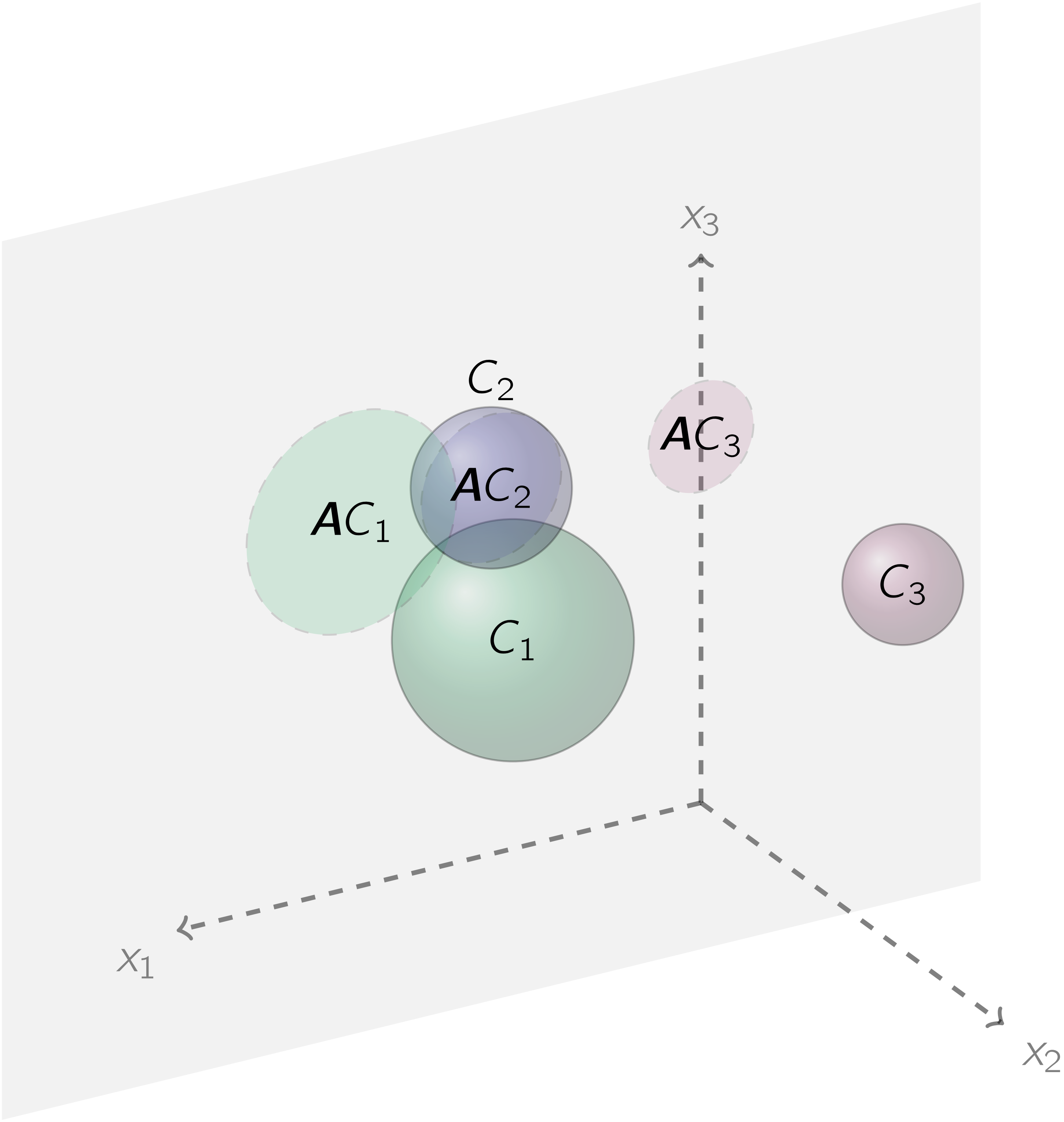
$$\mathrm{SNR(dB)} = 10 \log_{10} \frac{r^2}{\sigma^2}$$

(Depends on *separation* between the hypotheses.)

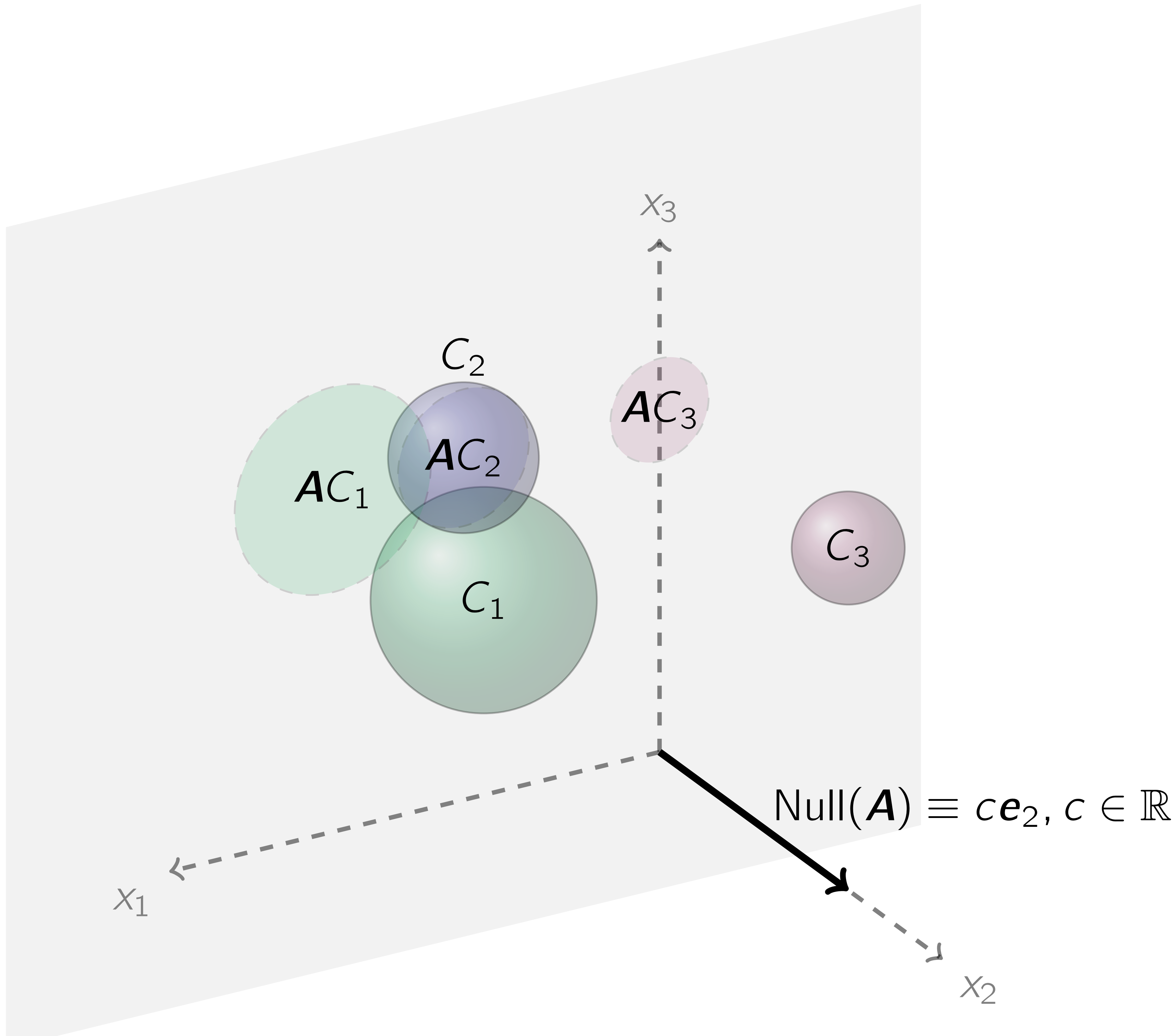


- Assume now that $\boldsymbol{x}$ is drawn from a mixture of classes $\{C_i\}_{i=1}^{p}$.
  - If the classes are not closed convex sets, we take as classes their *convex hulls:*
    $$\{S_i\}_{i=1}^{p}, S_i = \mathrm{Hull}(C_i), S_i \cap S_j = \emptyset$$
  - The classes (or their hulls) are assumed as disjoint closed convex sets, i.e., $\forall j \neq i, C_i \cap C_j = \emptyset$
- The classes are assumed *pairwise linearly separable* (by a hyperplane), *i.e.*,
  $$\forall j \neq i, \exists\, \boldsymbol{B} \in \mathbb{R}^{q \times n}, \boldsymbol{\beta} \in \mathbb{R}^{q}, q < n \ : \begin{cases} \boldsymbol{B}\mathbf{x} + \boldsymbol{\beta} \geq 0_q, & \forall \mathbf{x} \in C_i \\ \boldsymbol{B}\mathbf{x} + \boldsymbol{\beta} < 0_q, & \forall \mathbf{x} \in C_j \end{cases}$$
  (This setting strongly reminds *linear support vector machines*!)
- We want to assess whether a random projection is capable of preserving *linear separability*, that is assumed as critical event for compressive classification:
  $$\forall j \neq i, \boldsymbol{y} = \boldsymbol{A}\mathbf{x}, \exists\, \boldsymbol{B} \in \mathbb{R}^{q \times m}, \boldsymbol{\beta} \in \mathbb{R}^{q}, q < m : \begin{cases} \boldsymbol{B}\boldsymbol{y} + \boldsymbol{\beta} \geq 0_q, & \forall \mathbf{x} \in C_i \\ \boldsymbol{B}\boldsymbol{y} + \boldsymbol{\beta} < 0_q, & \forall \mathbf{x} \in C_j \end{cases}$$
- This notion is quite different (somewhat loose) w.r.t. stable embeddings. It does not measure *how* the distances between projected points of different classes are distorted *as long as their separability is maintained.*

The "Rare Eclipse" Problem

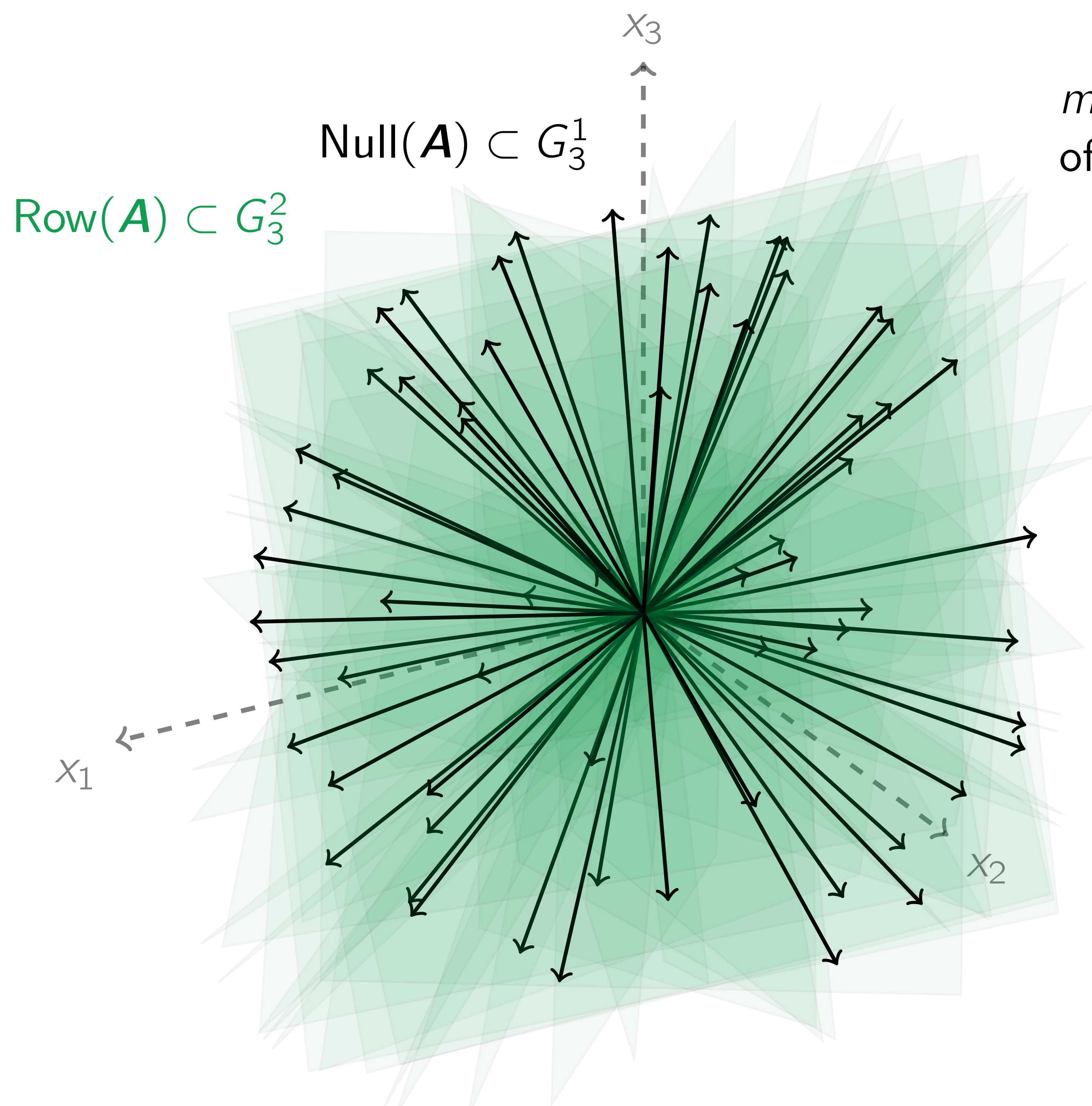


The "Rare Eclipse" Problem



The "Rare Eclipse" Problem

$$\boldsymbol{A}_{m\times n} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1) \Rightarrow \mathrm{Null}(\boldsymbol{A}) \sim \mathcal{U}(G_n^{n-m})$$



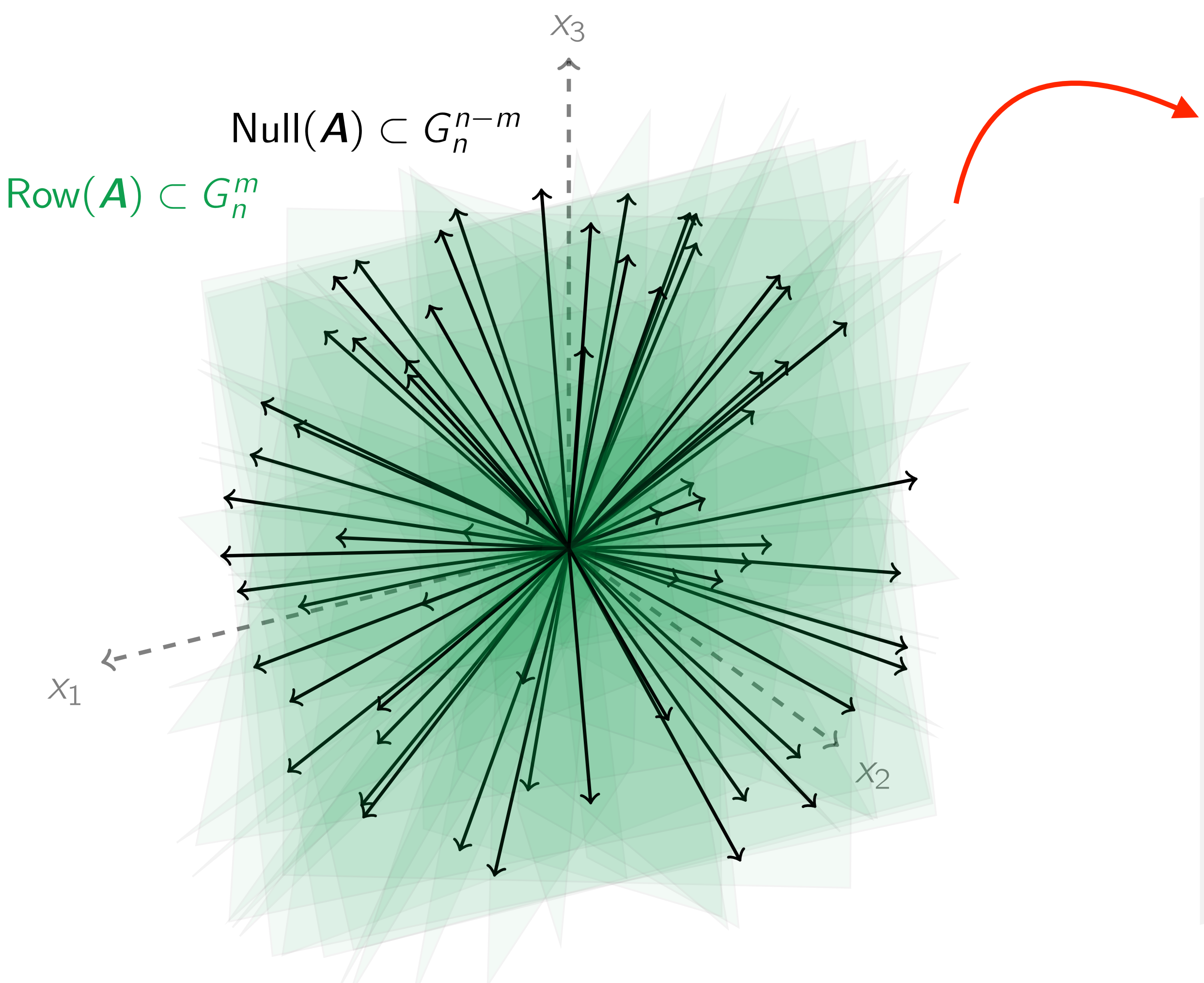
G_n^m: manifold of all m-dimensional subspaces of $\mathbb{R}^n$ (through the origin)

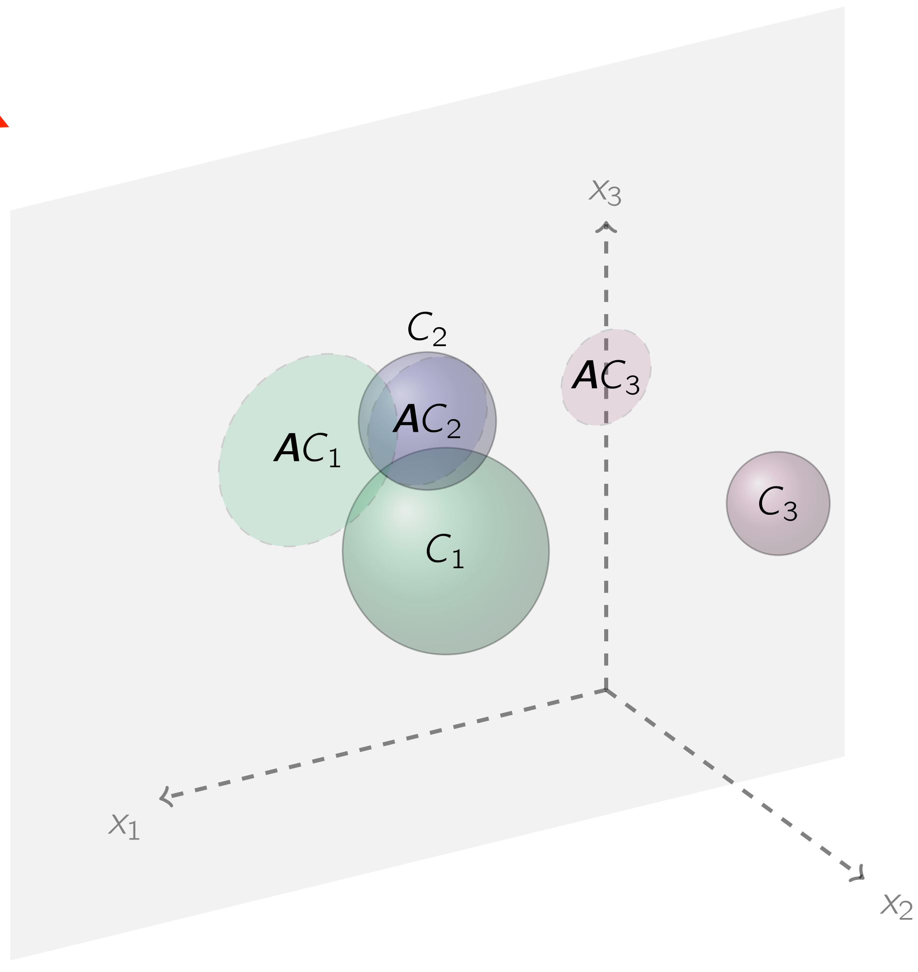
The "Rare Eclipse" Problem

Problem (Rare Eclipse [5]).

Let $\boldsymbol{A}_{m\times n} \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \frac{1}{m}\right)$ and two convex sets $C_i, C_j \subset \mathbb{R}^n, C_i \cap C_j = \emptyset$; find the smallest $m < n, \eta \in [0, 1)$ such that their images under $\boldsymbol{A}$ remain disjoint, *i.e.*,

$$\mathbb{P}[\boldsymbol{A}C_i \cap \boldsymbol{A}C_j = \emptyset] \geq 1 - \eta$$





The "Rare Eclipse" Problem

Problem (Rare Eclipse [5]).

Let $\boldsymbol{A}_{m\times n} \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \frac{1}{m}\right)$ and two convex sets $C_i, C_j \subset \mathbb{R}^n, C_i \cap C_j = \emptyset$; find the smallest $m < n, \eta \in [0, 1)$ such that their images under $\boldsymbol{A}$ remain disjoint, *i.e.*,

$$\mathbb{P}[\boldsymbol{A}C_i \cap \boldsymbol{A}C_j = \emptyset] \geq 1 - \eta$$

- Let's elaborate this requirement:

$$\boldsymbol{A}C_i \cap \boldsymbol{A}C_j = \emptyset \Leftrightarrow \forall \boldsymbol{x}' \in C_i, \boldsymbol{x}'' \in C_j, \boldsymbol{A}(\boldsymbol{x}'' - \boldsymbol{x}') \neq 0$$

- Define the Minkowski difference of the two sets:

$$C_i - C_j = \{\boldsymbol{x}' - \boldsymbol{x}'' \in \mathbb{R}^n : \boldsymbol{x}' \in C_i, \boldsymbol{x}'' \in C_j\}$$

- Thus:

$$\boldsymbol{A}C_i \cap \boldsymbol{A}C_j = \emptyset \Leftrightarrow \text{Null}(\boldsymbol{A}) \cap C_i - C_j = \emptyset$$

The "Rare Eclipse" Problem

Problem (Rare Eclipse [5]).

Let $\boldsymbol{A}_{m\times n} \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \frac{1}{m}\right)$ and two convex sets $C_i, C_j \subset \mathbb{R}^n, C_i \cap C_j = \emptyset$; find the smallest $m < n, \eta \in [0,1)$ such that their images under $\boldsymbol{A}$ remain disjoint, *i.e.*,

$$\mathbb{P}[\boldsymbol{A}C_i \cap \boldsymbol{A}C_j = \emptyset] \geq 1 - \eta$$

- Finally, since $\mathrm{Null}(\boldsymbol{A})$ is closed w.r.t. scalar multiplication, we can take the smallest cone that contains the Minkowski difference, *i.e.*,

$$C^- = \mathrm{Cone}(C_i - C_j), \bar{C}^- = \mathrm{Cone}(C_i - C_j) \cap \mathcal{S}^{n-1}$$

- Thus, the problem is mapped to evaluating the probability that

$$\boldsymbol{A}C_i \cap \boldsymbol{A}C_j = \emptyset \Leftrightarrow \mathrm{Null}(\boldsymbol{A}) \cap C^- = \emptyset$$

  that is the probability that $\mathrm{Null}(\boldsymbol{A})$ "avoids" the above cone.
  - We need a notion of "size" to measure the probability of this event.
  - Analogous concept in sparse signal recovery literature: the *null space property* [12].

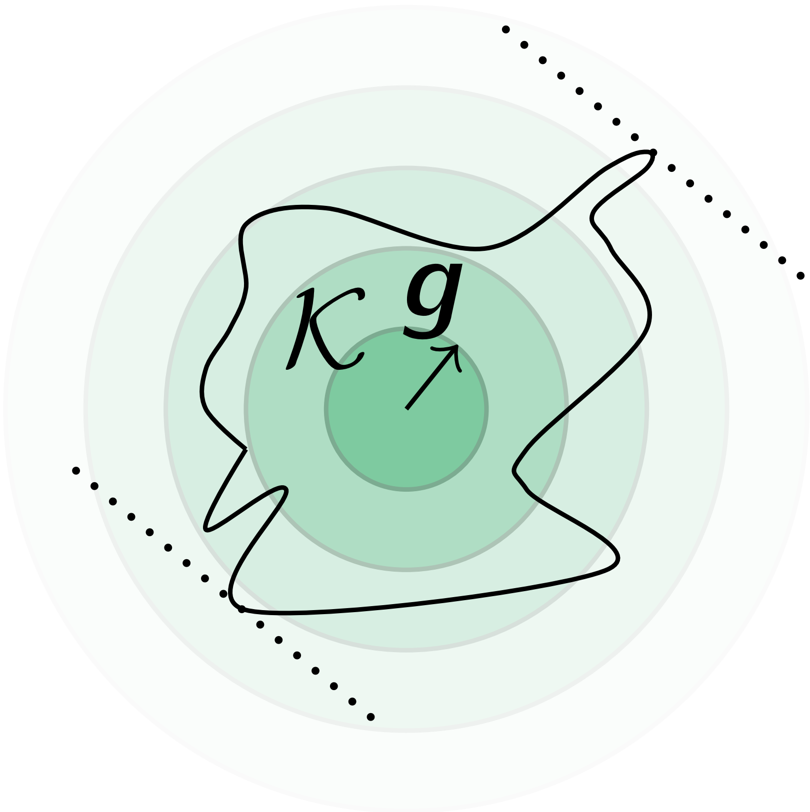
The Gaussian Width of a Set

Definition (Gaussian width of a set [2]).

Let $\boldsymbol{g} \sim \mathcal{N}(0_n, \boldsymbol{I}_n)$; the Gaussian (mean) width of a set $\mathcal{K} \subset \mathbb{R}^n$ is

$$w(\mathcal{K}) = \mathbb{E}_{\boldsymbol{g}} \left[\max_{\boldsymbol{x} \in \mathcal{K}} \langle \boldsymbol{x}, \boldsymbol{g} \rangle \right]$$

$w(\mathcal{K})$ is invariant under translations, orthogonal transformations, and convex hulls of the set $\mathcal{K}$.



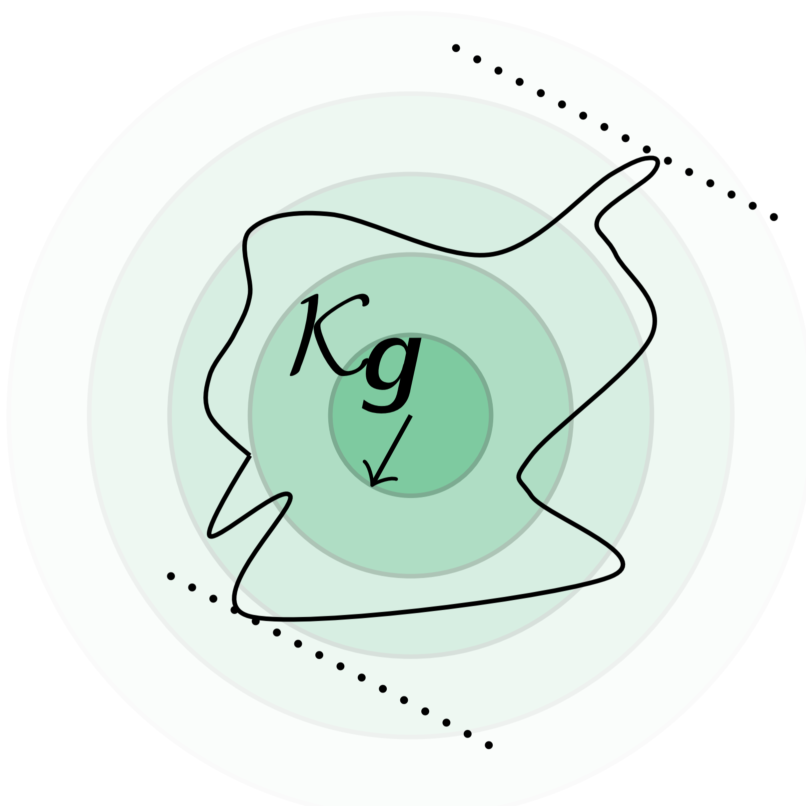
The Gaussian Width of a Set

Definition (Gaussian width of a set [2]).

Let $\boldsymbol{g} \sim \mathcal{N}(0_n, \boldsymbol{I}_n)$; the Gaussian (mean) width of a set $\mathcal{K} \subset \mathbb{R}^n$ is

$$w(\mathcal{K}) = \mathbb{E}_{\boldsymbol{g}} \left[\max_{\boldsymbol{x} \in \mathcal{K}} \langle \boldsymbol{x}, \boldsymbol{g} \rangle \right]$$

$w(\mathcal{K})$ is invariant under translations, orthogonal transformations, and convex hulls of the set $\mathcal{K}$.



The Gaussian Width of a Set

Definition (Gaussian width of a set [2]).

Let $\boldsymbol{g} \sim \mathcal{N}(0_n, \boldsymbol{I}_n)$; the Gaussian (mean) width of a set $\mathcal{K} \subset \mathbb{R}^n$ is

$$w(\mathcal{K}) = \mathbb{E}_{\boldsymbol{g}} \left[\max_{\boldsymbol{x} \in \mathcal{K}} \langle \boldsymbol{x}, \boldsymbol{g} \rangle \right]$$

$w(\mathcal{K})$ is invariant under translations, orthogonal transformations, and convex hulls of the set $\mathcal{K}$.

- Some relevant examples in the literature [14]:
  - k-sparse signals:

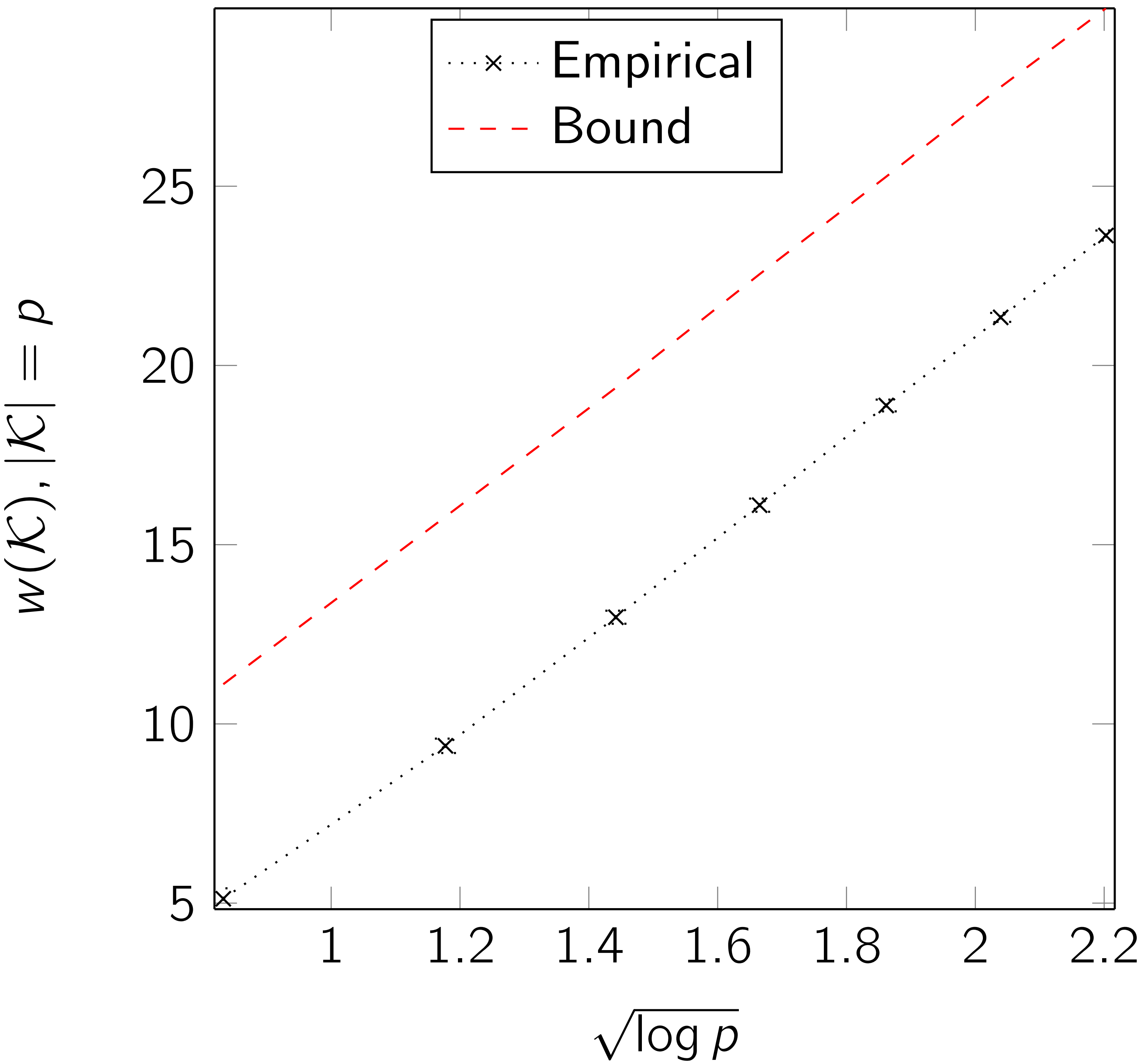
$$\mathcal{K} = \Sigma_k, \;\; w(\Sigma_k) \lesssim \sqrt{k \log \tfrac{n}{k}}$$

  - p-cardinality sets of vectors:

$$\mathcal{K} = \{\boldsymbol{s}_i\}_{i=1}^{p}, \;\; w(\mathcal{K}) \leq \sqrt{2 \log p} \max_{i \in [p]} \|\boldsymbol{s}_i\|_2$$

The Gaussian Width of a Set

$$\mathcal{K} = \{\boldsymbol{s}_i\}_{i=1}^{p}, \quad w(\mathcal{K}) \leq \sqrt{2\log p}\, \max_{i\in[p]} \|\boldsymbol{s}_i\|_2$$

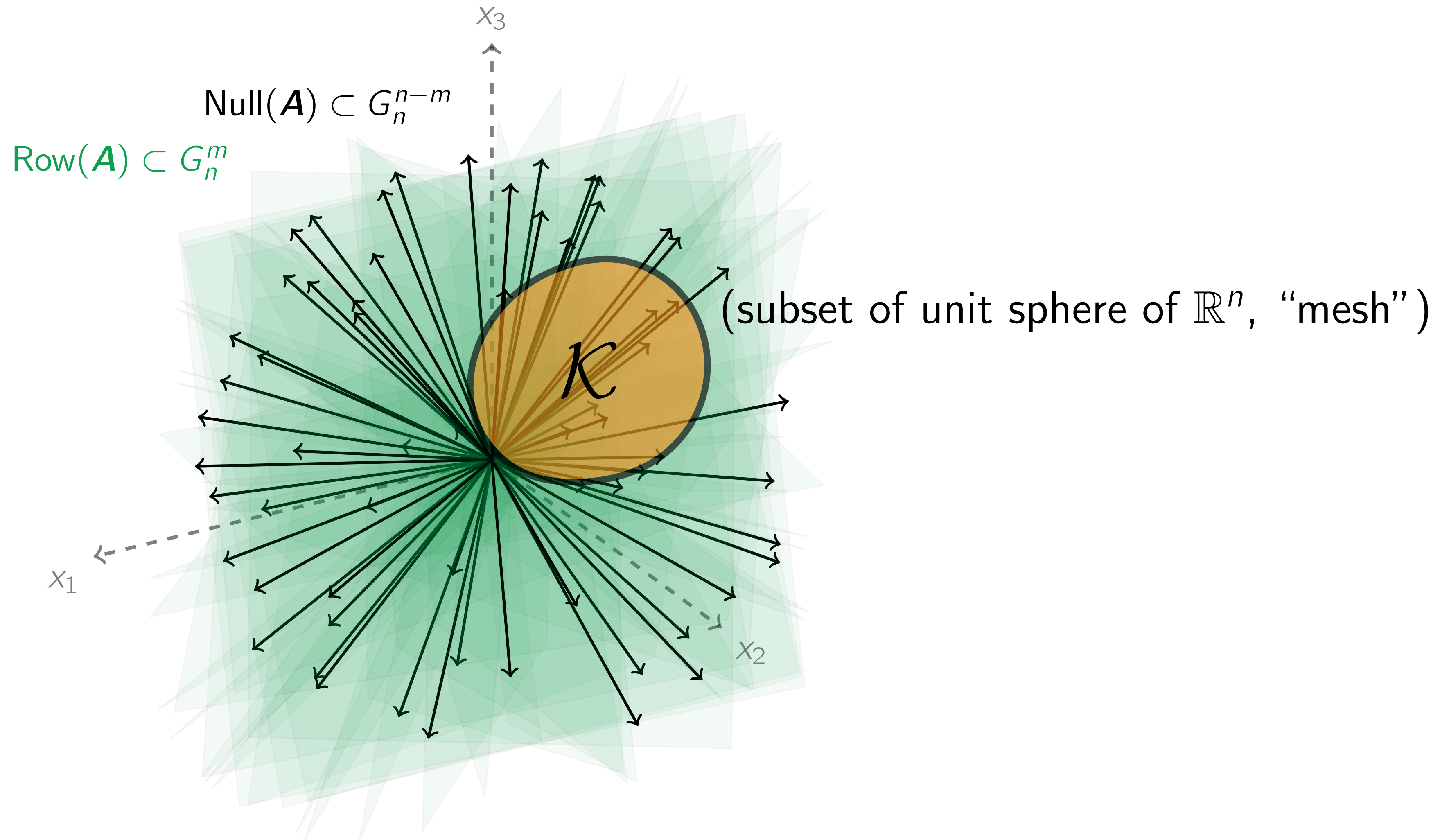


"Escape through a mesh"

Theorem (Gordon's Escape through a Mesh Theorem [13]).

Let $\mathcal{K} \subset \mathcal{S}^{n-1}$ and $\boldsymbol{g} \sim \mathcal{N}(0_m, \boldsymbol{I}_m)$; denote $\lambda_m = \mathbb{E}[\|\boldsymbol{g}\|_2]$. If $w(\mathcal{K}) \leq \lambda_m$, then any uniformly drawn $Y \in G_m^{n-m}$ satisfies

$$\mathbb{P}\left[Y \cap \mathcal{K} = \emptyset\right] \geq 1 - e^{\left(-\frac{1}{2}(\lambda_m - w(\mathcal{K}))^2\right)}$$



Minimum Projection Rank (for linear separability)

Corollary (“Escape through a Minkowski difference”, Corollary 3.1 in [5]).

Let $C_i, C_j \subseteq \mathbb{R}^n$ be disjoint convex sets; $\bar{C}^- = \text{Cone}(C_i - C_j) \cap \mathcal{S}^{n-1}$; $w_\cap = w(\bar{C}^-)$. Let $\boldsymbol{A}_{m\times n} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ and $\eta \in (0,1)$. Then for $m \geq m^\star$,

$$m^\star = \left(w_\cap + \sqrt{2\log\frac{1}{\eta}}\right)^2 + 1 \Rightarrow \mathbb{P}\left[\boldsymbol{A}C_i \cap \boldsymbol{A}C_j = \emptyset\right] \geq 1 - \eta$$

- This corollary is simply obtained by taking in Gordon’s Theorem:

$$Y := \text{Null}(\boldsymbol{A}), \mathcal{K} := \bar{C}^-, \lambda_m \leq \sqrt{m}$$

$$\eta = e^{\left(-\frac{1}{2}(\sqrt{m} - w(\bar{C}^-))^2\right)}$$

  and by fixing the last quantity to an arbitrary probability value.
- The rest of Bandeira *et al.* [5] is simply concerned with *finding closed-form expressions for the Gaussian width* of the cone that encloses the Minkowski difference of special convex sets.
- Spheres (simple) and ellipsoids (much harder) lend themselves to this calculation.

The Case of Disjoint Euclidean Balls

Lemma (C^- of two balls is a circular cone, Lemma 3.3 in [5]).

Let $i = 1, 2$, $C_i = \rho_i \boldsymbol{x} + \boldsymbol{s}_i : \boldsymbol{x} \in \mathcal{B}_{\ell_2}(\mathbb{R}^n)$ of centers $\boldsymbol{s}_i \in \mathbb{R}^n$ and radii $\rho_i > 0$; assume that $\rho_1 + \rho_2 < \|\boldsymbol{s}_1 - \boldsymbol{s}_2\|_2$. Then the $\mathrm{Cone}(C_1 - C_2) = \mathrm{Circ}(\alpha)$, that is the *circular cone* of aperture α

$$\mathrm{Circ}(\alpha) = \left\{ \boldsymbol{z} \in \mathbb{R}^n : \frac{\langle \boldsymbol{z}, \boldsymbol{s}_1 - \boldsymbol{s}_2 \rangle}{\|\boldsymbol{z}\|_2 \|\boldsymbol{s}_1 - \boldsymbol{s}_2\|_2} \geq \cos \alpha \right\}$$

for $\alpha \in (0, \frac{\pi}{2})$, $\sin \alpha = \frac{\rho_1 + \rho_2}{\|\boldsymbol{s}_1 - \boldsymbol{s}_2\|_2}$.

- The proof entails showing:

$$\mathrm{Cone}(C_1 - C_2) \subseteq \mathrm{Circ}(\alpha) \text{ and } \mathrm{Circ}(\alpha) \subseteq \mathrm{Cone}(C_1 - C_2)$$

- Since the Gaussian width of the circular cone is known,

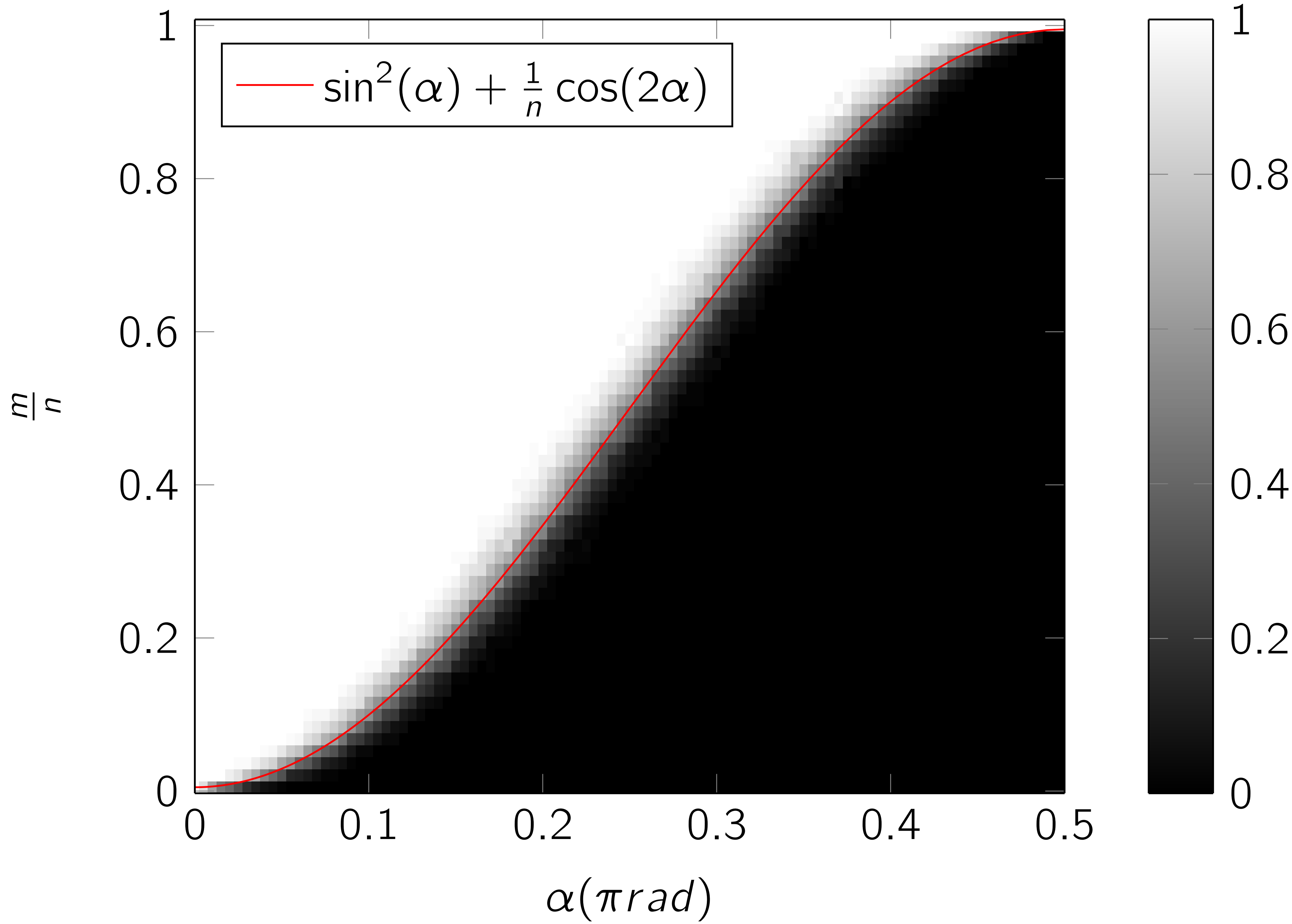
$$w_\cap^2 = w(\mathrm{Circ}(\alpha) \cap \mathcal{S}^{n-1})^2 = n \sin^2 \alpha + O(1)$$

plugging this into the previous Corollary yields:

$$m^\star = n \left(\frac{\rho_1 + \rho_2}{\|\boldsymbol{s}_1 - \boldsymbol{s}_2\|_2} \right)^2 + O(\sqrt{n})$$

The Case of Disjoint Euclidean Balls

$\mathbb{P}(\mathrm{Null}(\boldsymbol{A}) \cap C^- = \emptyset)$



The Case of Disjoint Ellipsoids

- A naive approach would be taking the *radii* as the largest semi-axes (*i.e.*, maximum singular values) of the symmetric PSD matrices defining the ellipsoids, *i.e.*, taking the smallest balls that enclose them.
  - Implicitly assumes that the bounding balls do not intersect.
  - This would lead to an *extremely loose* bound.
- Bandeira *et al.* [5] take a step further and arrive to the following statement (proof is less intuitive):

Theorem (Gaussian width of C^- of two ellipses, Theorem 3.5 in [5]).

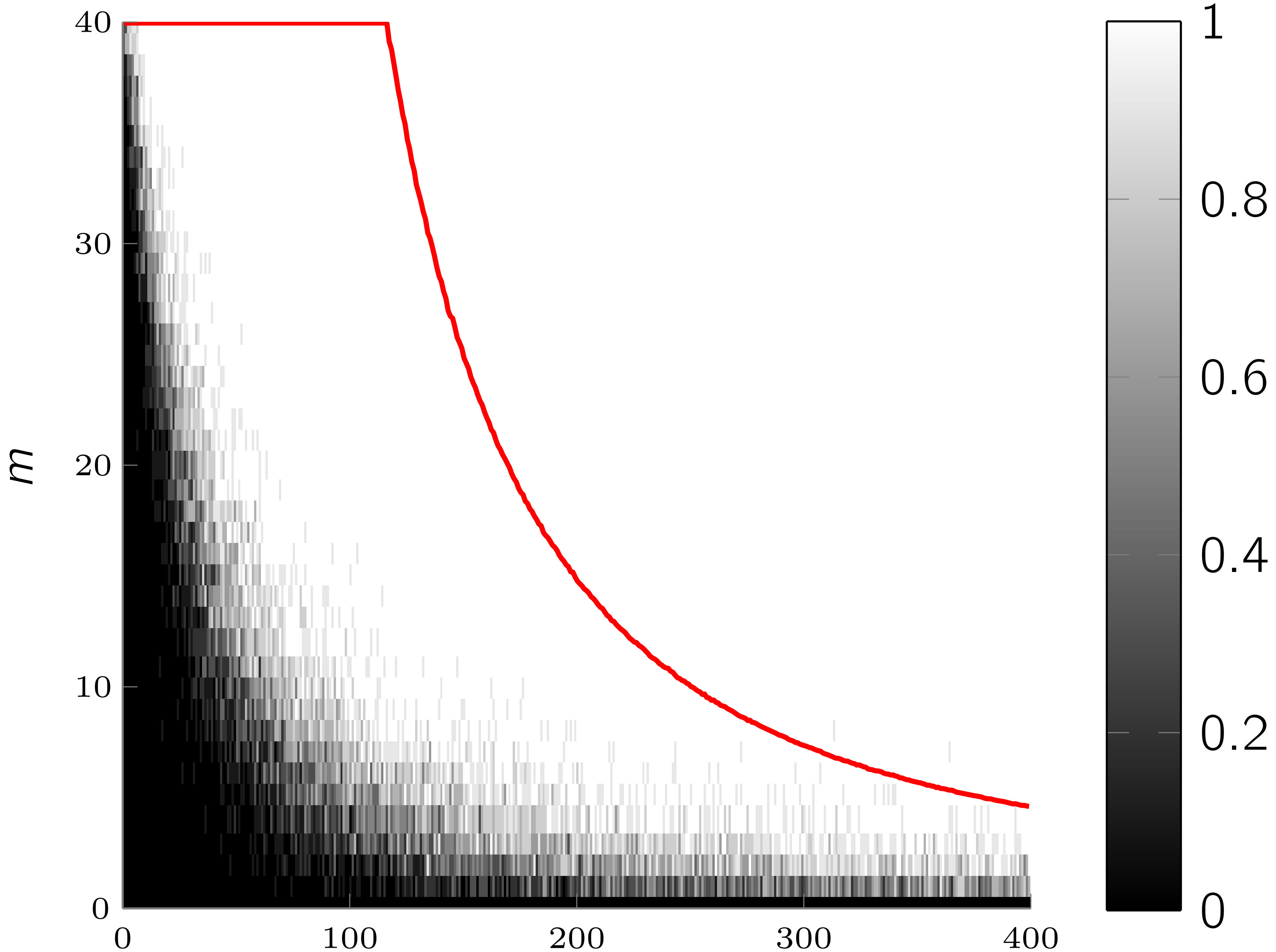
Let $i = 1, 2, \Gamma_i \in \mathbb{R}^{n \times n}$ symmetric PSD, $C_i = \{\Gamma_i \boldsymbol{x} + \boldsymbol{s}_i : \boldsymbol{x} \in \mathcal{B}_{\ell_2}(\mathbb{R}^n)\}$ of centers $\boldsymbol{s}_i \in \mathbb{R}^n$. Then

$$w_\cap \leq \frac{\|\Gamma_1\|_F + \|\Gamma_2\|_F}{\zeta - (\|\Gamma_1 \boldsymbol{\xi}\|_2 + \|\Gamma_2 \boldsymbol{\xi}\|_2)}$$

where $\boldsymbol{\xi} = \frac{\boldsymbol{s}_1 - \boldsymbol{s}_2}{\|\boldsymbol{s}_1 - \boldsymbol{s}_2\|_2}, \zeta = \|\boldsymbol{s}_1 - \boldsymbol{s}_2\|_2 > \|\Gamma_1 \boldsymbol{\xi}\|_2 + \|\Gamma_2 \boldsymbol{\xi}\|_2$.

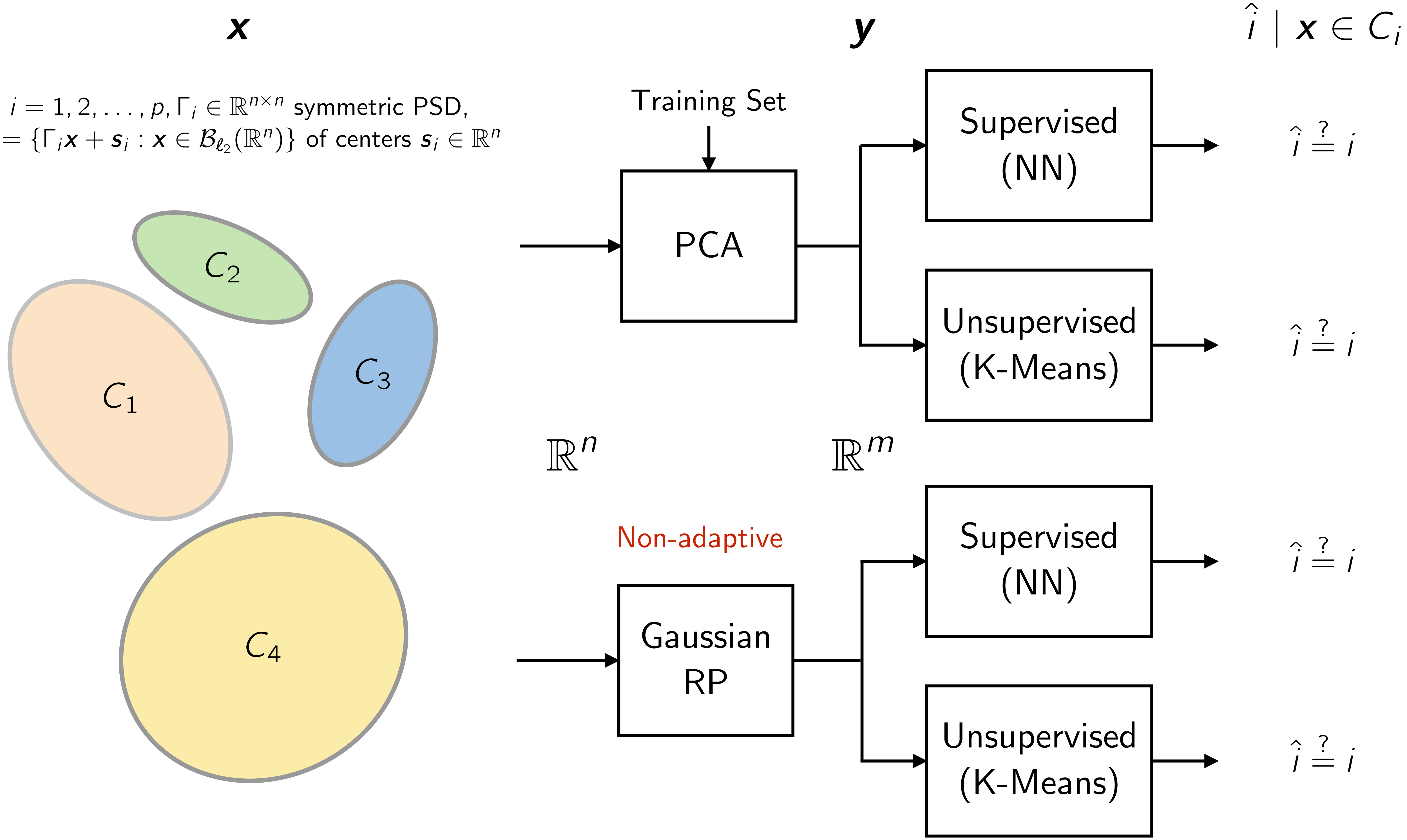
The Case of Disjoint Ellipsoids

$\mathbb{P}(\mathrm{Null}(\boldsymbol{A}) \cap C^- = \emptyset)$

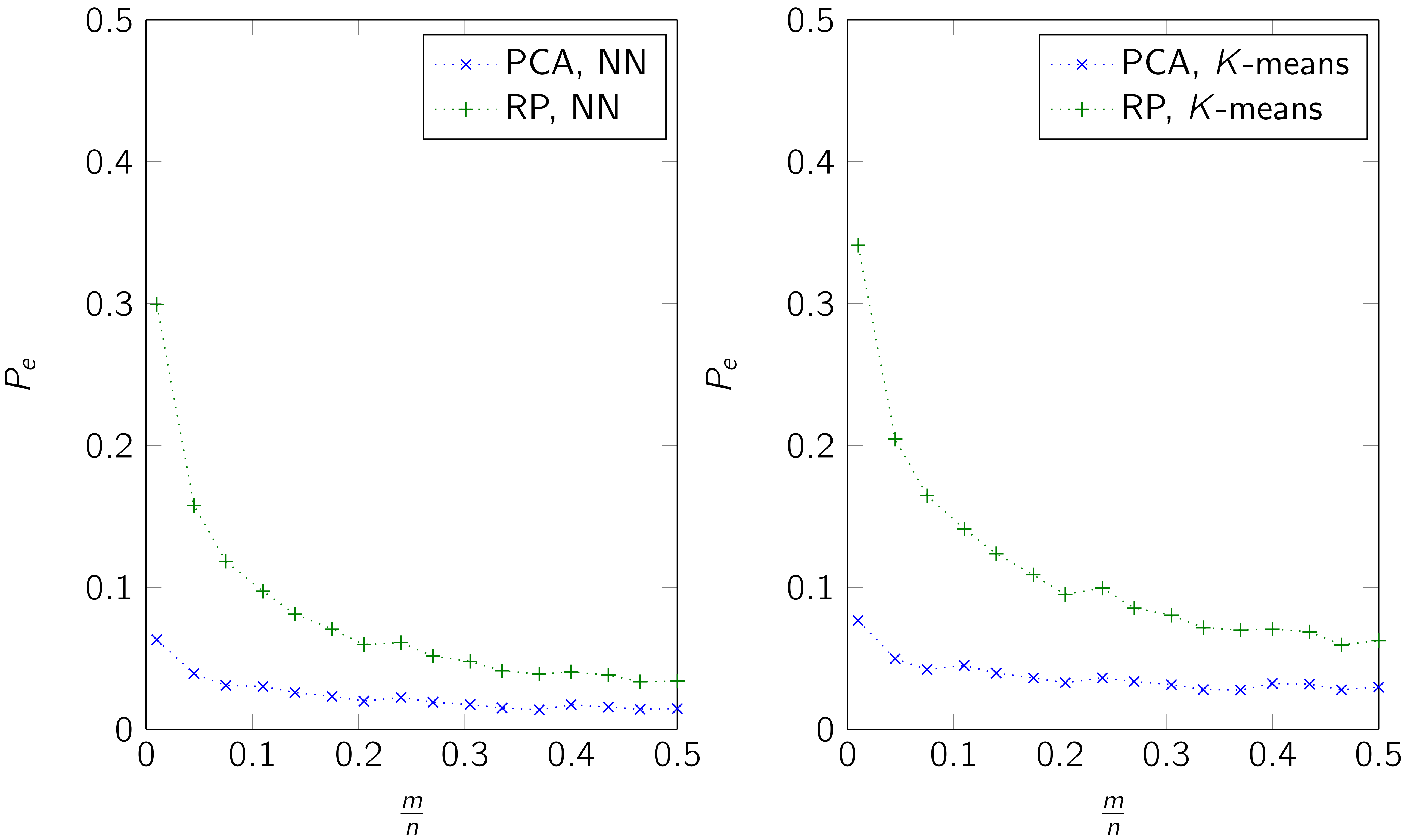


ζ: Distance between ellipsoid centers

$i = 1, 2, \ldots, p$, $\Gamma_i \in \mathbb{R}^{n \times n}$ symmetric PSD,
$C_i = \{\Gamma_i \boldsymbol{x} + \boldsymbol{s}_i : \boldsymbol{x} \in \mathcal{B}_{\ell_2}(\mathbb{R}^n)\}$ of centers $\boldsymbol{s}_i \in \mathbb{R}^n$



A comparison with PCA for mixtures of ellipses



Conclusion

- Emphasis of this talk was on assessing whether it is (theoretically) possible to distinguish *linearly separable classes* after random projection.
  - This ensures that even the simplest classification algorithm will succeed.
  - Ideally, *unsupervised* learning will yield separated clusters after a non-adaptive dimensionality reduction.
  - Application: compressive classification "right after" the sensing interface, with minimum computational and hardware complexity requirements.
- Open questions:
  - How does (1 to q)-bit quantisation affect compressive classification? By how much the requirements on m will be increased? Can we characterise:

$$\mathbb{P}\left[\mathcal{Q}_q\left(\boldsymbol{A}C_i\right) \cap \mathcal{Q}_q\left(\boldsymbol{A}C_j\right) = \emptyset\right] \geq 1 - \eta$$

  - Study of different models for other disjoint convex sets of interest
  - Application to classification of very high-dimensional (*e.g.*, volumetric) data

Bibliography

- [1] Donoho, D. L. (2006). Compressed sensing. Information Theory, IEEE Transactions on, 52(4), 1289-1306.
- [2] Chandrasekaran, V., Recht, B., Parrilo, P. A., & Willsky, A. S. (2012). The convex geometry of linear inverse problems. Foundations of Computational mathematics, 12(6), 805-849.
- [3] Davenport, M., Boufounos, P. T., Wakin, M. B., & Baraniuk, R. G. (2010). Signal processing with compressive measurements. Selected Topics in Signal Processing, IEEE Journal of, 4(2), 445-460.
- [4] Haupt, J., Castro, R., Nowak, R., Fudge, G., & Yeh, A. (2006, October). Compressive sampling for signal classification. In Signals, Systems and Computers, 2006. ACSSC'06. Fortieth Asilomar Conference on (pp. 1430-1434). IEEE.
- [5] Bandeira, A. S., Mixon, D. G., & Recht, B. (2014). Compressive classification and the rare eclipse problem. arXiv preprint arXiv:1404.3203.
- [6] Dasgupta, S. (1999). Learning mixtures of Gaussians. In Foundations of Computer Science, 1999. 40th Annual Symposium on (pp. 634-644). IEEE.
- [7] Reboredo, H., Renna, F., Calderbank, R., & Rodrigues, M. R. (2013, July). Compressive classification. In Information Theory Proceedings (ISIT), 2013 IEEE International Symposium on (pp. 674-678). IEEE.
- [8] Baraniuk, R., Davenport, M., DeVore, R., & Wakin, M. (2008). A simple proof of the restricted isometry property for random matrices. Constructive Approximation, 28(3), 253-263.
- [9] Johnson, W. B., & Lindenstrauss, J. (1984). Extensions of Lipschitz mappings into a Hilbert space. Contemporary mathematics, 26(189-206), 1.
- [10] Achlioptas, D. (2003). Database-friendly random projections: Johnson-Lindenstrauss with binary coins. Journal of computer and System Sciences, 66(4), 671-687.
- [11] Dasgupta, S., & Gupta, A. (2003). An elementary proof of a theorem of Johnson and Lindenstrauss. Random structures and algorithms, 22(1), 60-65.
- [12] Cohen, A., Dahmen, W., & DeVore, R. (2009). Compressed sensing and best k-term approximation. Journal of the American mathematical society, 22(1), 211-231.
- [13] Gordon, Y. (1988). On Milman's inequality and random subspaces which escape through a mesh in $\mathbb{R}$ n (pp. 84-106). Springer Berlin Heidelberg.
- [14] Amelunxen, D., Lotz, M., McCoy, M. B., & Tropp, J. A. (2014). Living on the edge: Phase transitions in convex programs with random data. Information and Inference, iau005.

Thank you for your attention.

Further questions:

valerio.cambareri@uclouvain.be

fnrs FREEDOM TO RESEARCH

UCL Université catholique de Louvain