Geometry-preserving Embeddings: Dimensionality Reduction Techniques for Information Representation

Petros Boufounos and Laurent Jacques

ICIP'18, Athens, Greece

MERL
Mitsubishi Electric Research Laboratories

Institute for Information
and Communication Technologies,
Electronics and Applied Mathematics

Outline

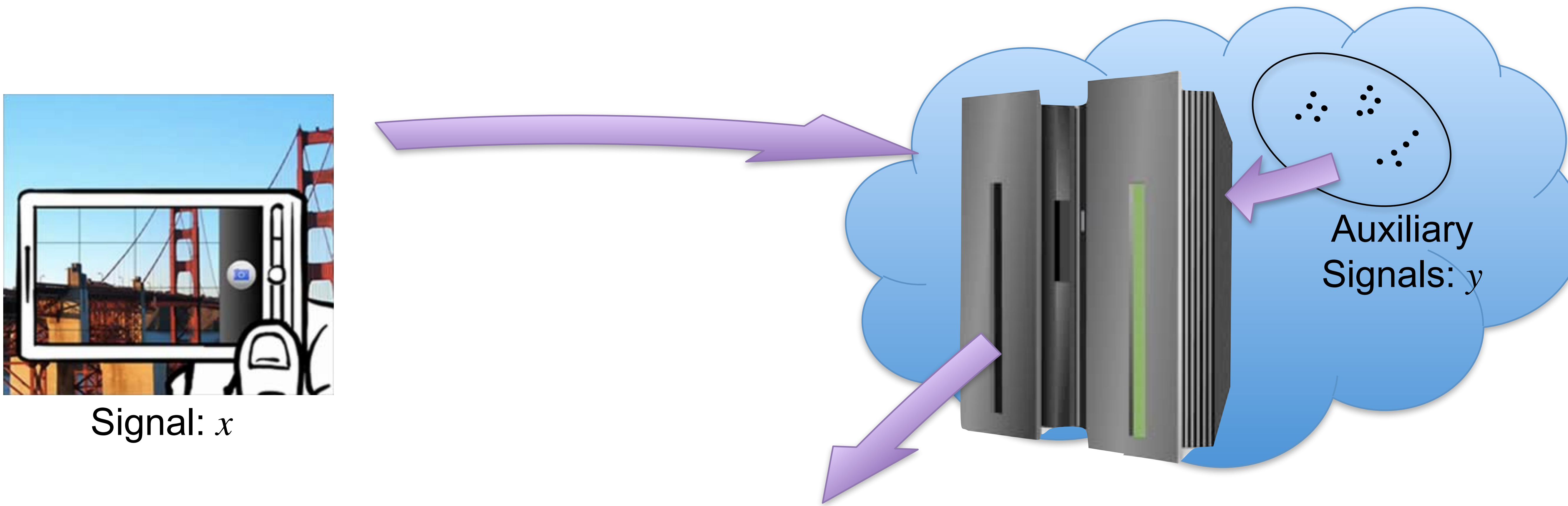
Outline

1. Introduction
2. Fundamentals of embeddings and embedology
3. Quantized embeddings

   Coffee/Tea break

4. Embedding Design
5. Embeddings of Alternative Metrics
6. Learning Embeddings
7. Conclusions and open problems

Motivation: The Big Picture



Output: $g(x,y)$

Information Scalable Processing:

How to only represent and process information required by $g(\cdot,\cdot)$?

Main goals:

Rate- and computation-**efficient representation**

Accurate and **efficient computation** of $g(\cdot,\cdot)$

Fruitful interaction of representation and computation

Cloud-based Signal Processing: The Big Picture

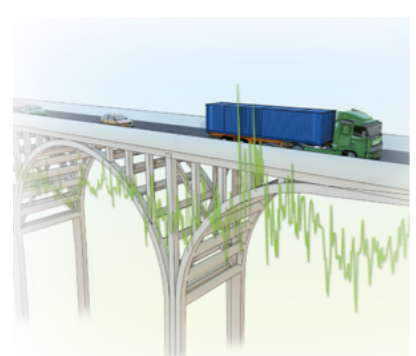
Signal: x

Output: $g(x,y)$

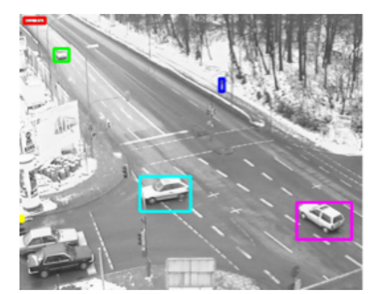
Augmented Reality

Remote Medicine

Surveillance



Infrastructure Monitoring

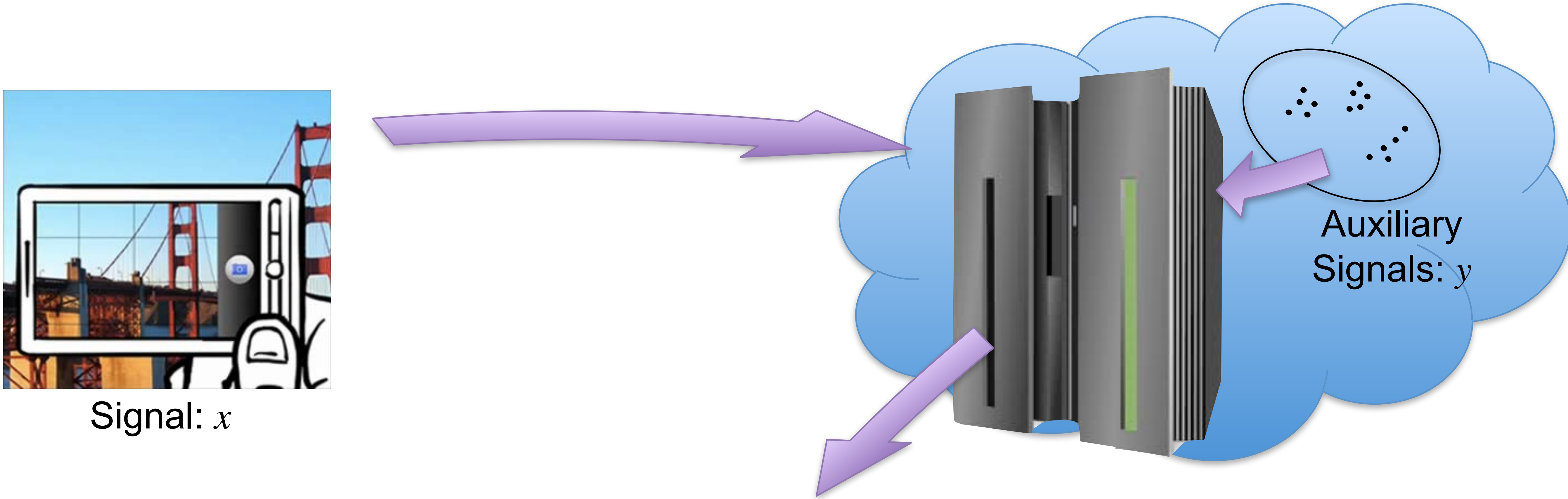


Traffic Monitoring

Biometric Authentication

Image/Video Retrieval

The Big Picture: Distances



Output: $g(x,y)=g(\|x-y\|)$

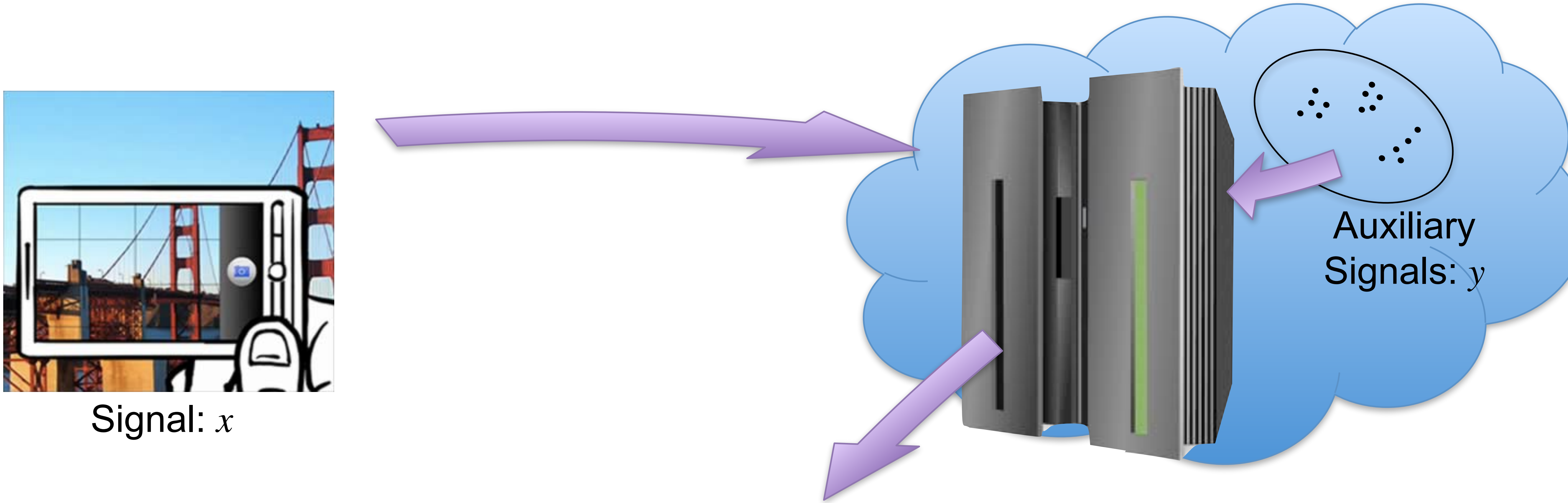
Function computes **functions of signal distances**,

Auxiliary **information: other signals**

⇒ Representations of signal distances

Main tool: **Embeddings**

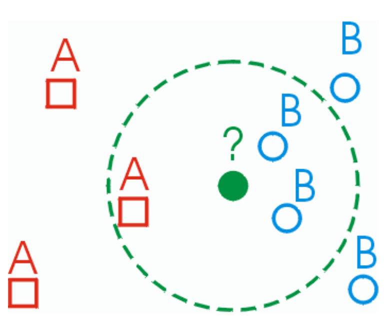
The Big Picture: Distances



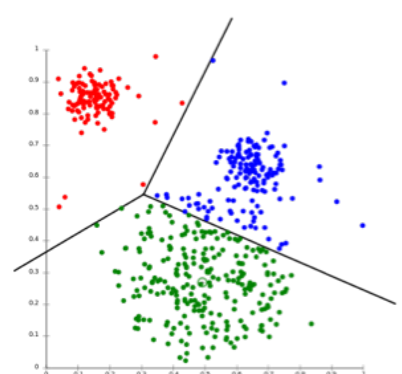
Output: $g(x,y)=g(\|x-y\|)$

Why distances?

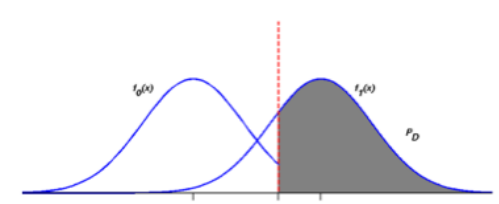
Fundamental primitive for a large number of methods



Classification



Clustering

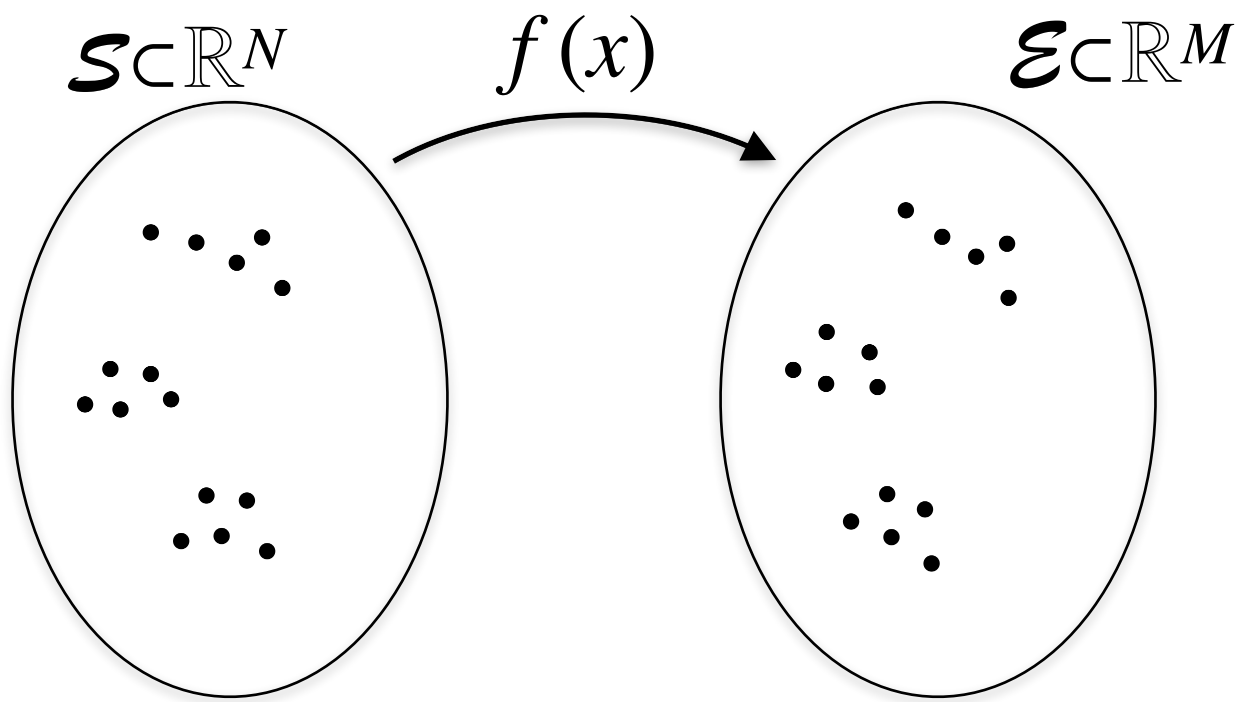


Detection/
Estimation

Embeddings in Words and Pictures

Original space

Distance metric: $d_{\mathcal{S}}$



Embedding space

Embed in $\mathcal{E}$

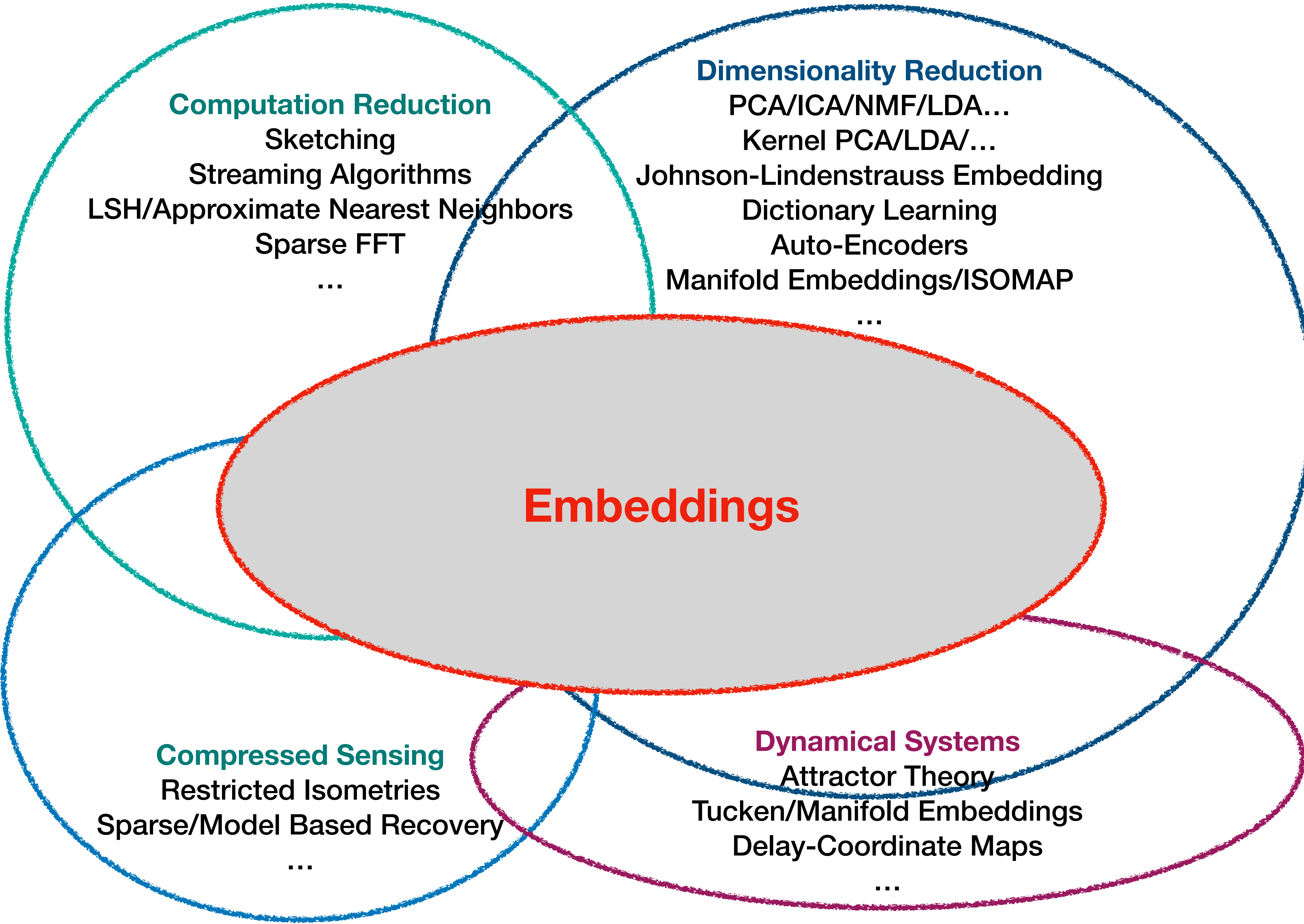
Distance metric: $d_{\mathcal{E}}$

An embedding is a **function**
from an **original space** to an **embedding space**
that **preserves** aspects of the **geometry** of the original space

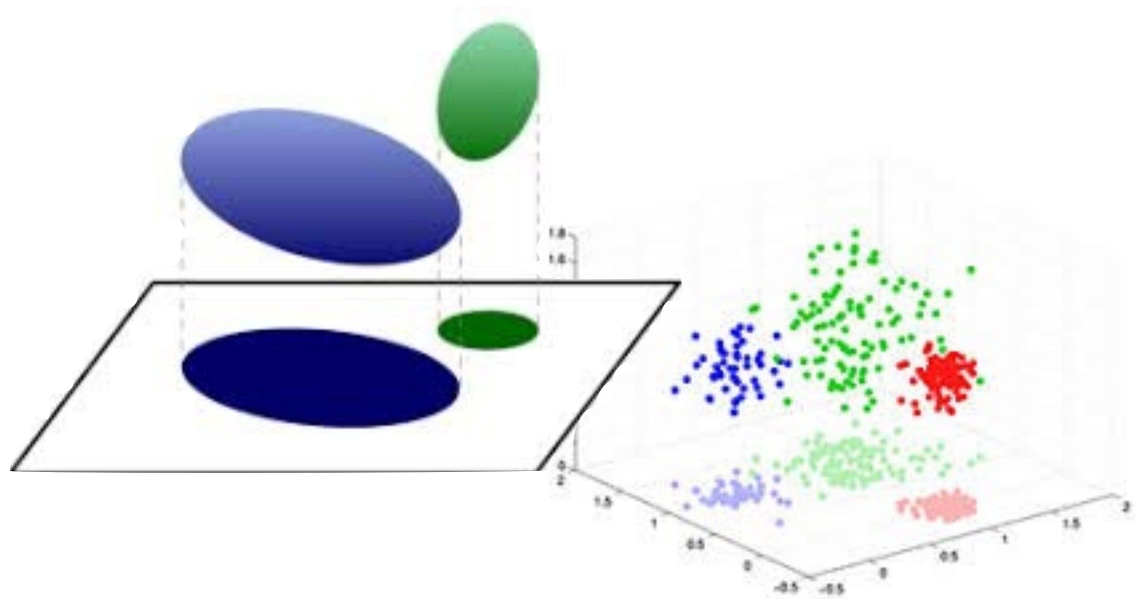
Why?

It hopefully **makes life simpler** in the embedding space

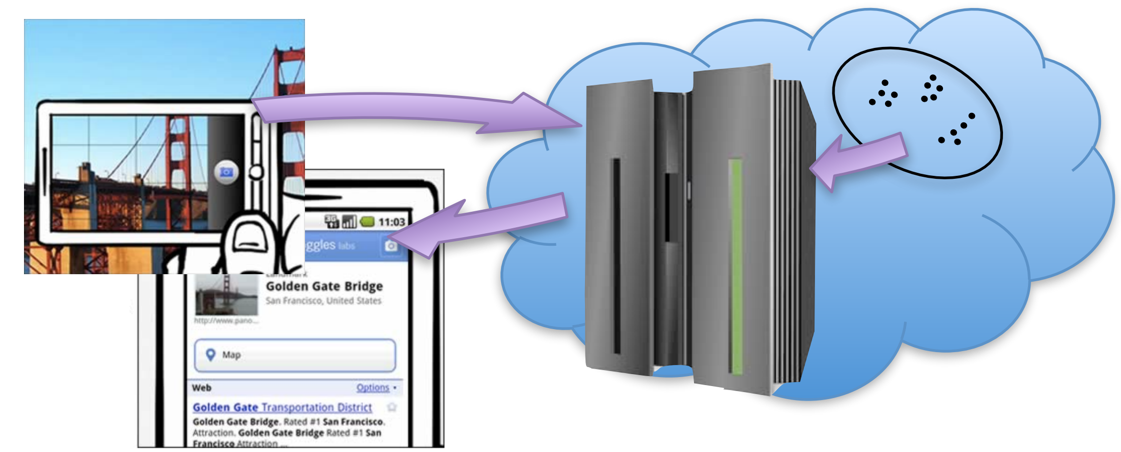
Embeddings In Context



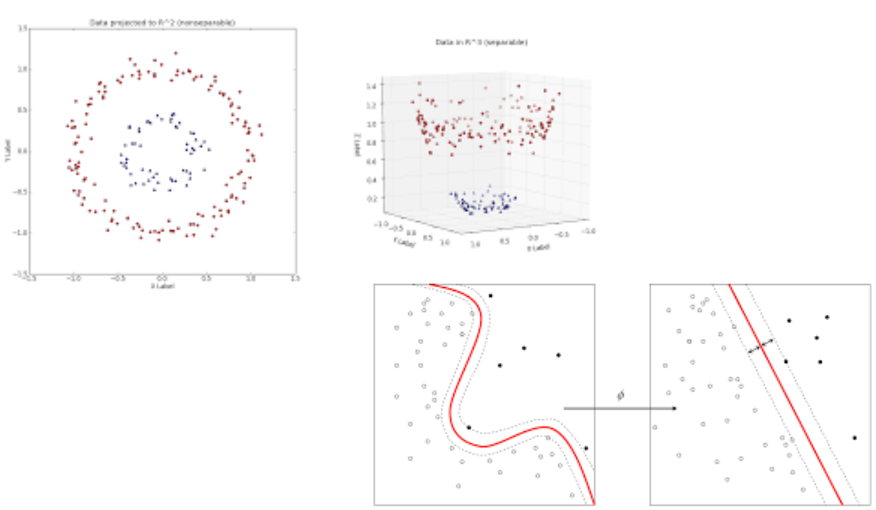
Applications



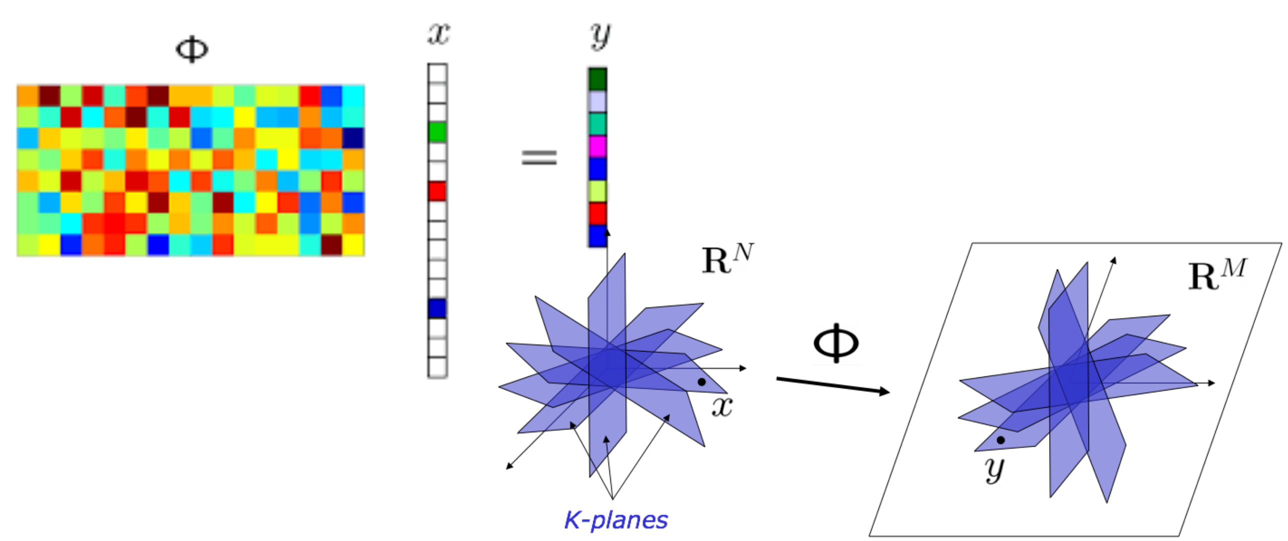
Dimensionality Reduction



Signal Retrieval



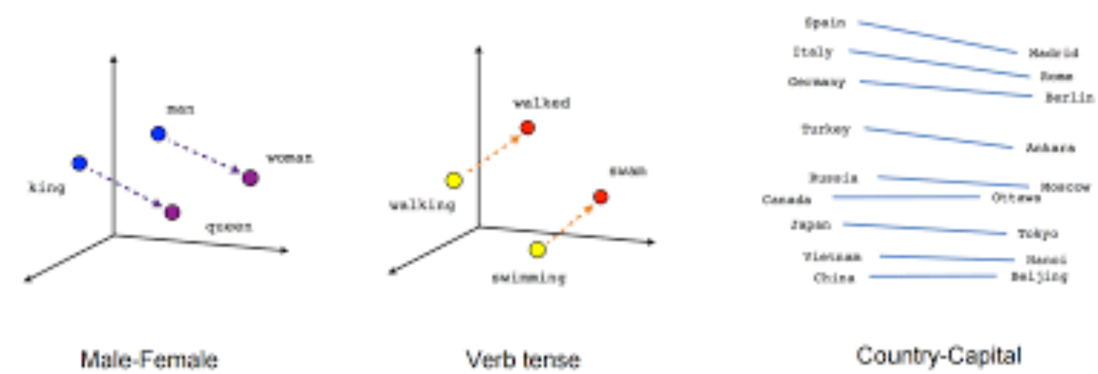
Kernel Methods

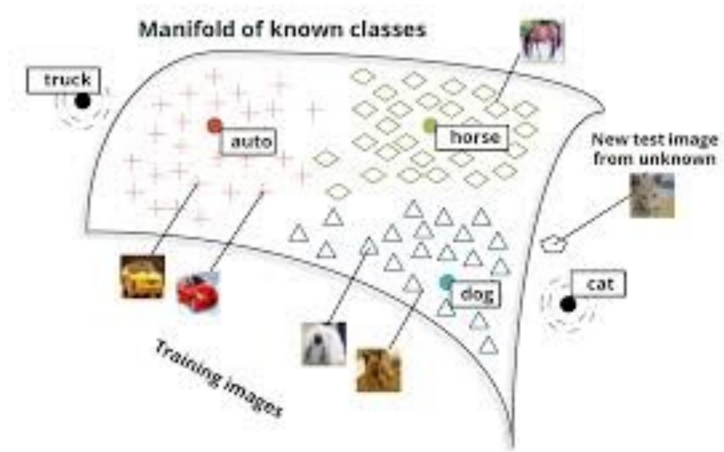


Compressed Sensing

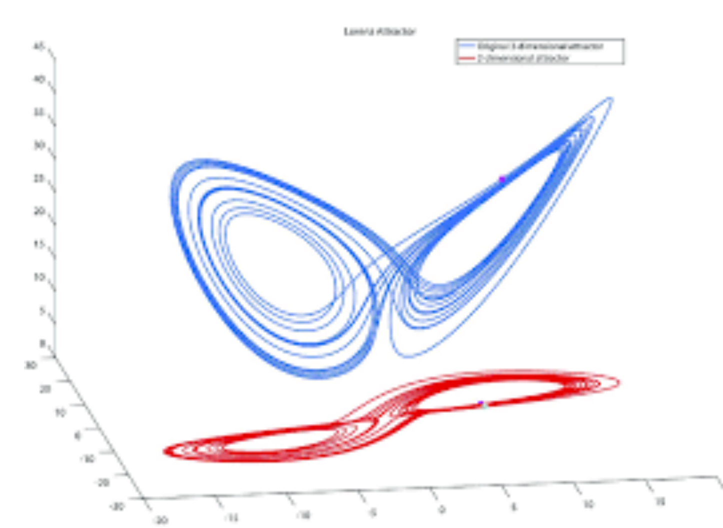
Not in this tutorial

- Embeddings Using Deep Learning
- Word embeddings
  - Term “embedding” is only used qualitatively in this literature
  - Not many guarantees
  - We touch on some of the similarities and differences





- Embeddings of Dynamical Systems
  - A lot of past work and theory; could be tutorial by itself (e.g., **[Eftekhari et al, ’17]**)
  - Different focus
  - We mention some of the results



Outline

1. Introduction
2. Fundamentals of embeddings and embedology
3. Quantized embeddings

   Coffee/Tea break ☕

4. Embedding Design
5. Embeddings of Alternative Metrics
6. Learning Embeddings
7. Conclusions and open problems

2. Fundamentals of embeddings and *embedology*

- The Big Picture
- General Embedding Definition
- The Johnson-Lindenstrauss Lemma & variants
- The restricted isometry property (RIP)
  - Principles and definition
  - Market of RIP matrices
  - RIP of more general signal sets & manifolds
  - Proving the RIP with JL Lemma

The Big Picture

- High dimensional signals in Sciences & Technology
  (e.g., images, video, hyperspectral data, dynamic medical data volumes, data on manifolds, dynamical system ...)

- Big Data & high dimension are obstacles for:
  - acquisition, storage, processing,
  - data classification, data learning, ...

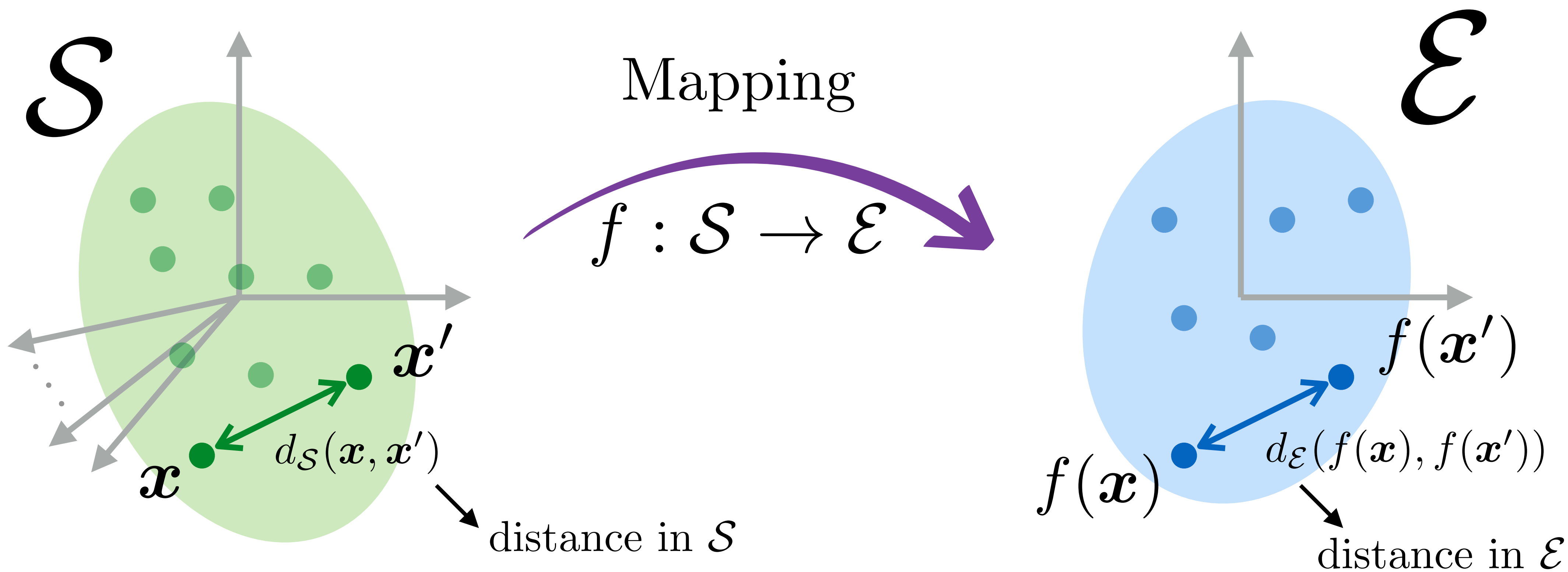
- Possible solution: Dimensionality Reduction

- Crucial questions:
  - Trade-offs btw embedding dimensions,
    number of bits, accuracy
  - DR preserving geometry of individual signal/set of signals
  - Design of the *embedding*, e.g., preserve close signals only

General Embedding Definition

High-dimensional signals
in a signal space $\mathcal{S}$

Embedding space
(e.g., low-dimension,
small number of bits)



Embedding relation of $\mathcal{S}$ in $\mathcal{E}$:

$$d_{\mathcal{E}}(f(\boldsymbol{x}), f(\boldsymbol{x}')) \approx g\big(d_{\mathcal{S}}(\boldsymbol{x}, \boldsymbol{x}')\big) \text{ for all } \boldsymbol{x}, \boldsymbol{x}' \in \mathcal{S}.$$

(possible distortions) (possible distance alteration)

Formal definition

Given some "distortions" $\epsilon, \epsilon' > 0$ and a possible alteration $g : \mathbb{R}_+ \to \mathbb{R}_+$,

$$(1-\epsilon)g\big(d_{\mathcal{S}}(\boldsymbol{x},\boldsymbol{x}')\big) - \epsilon' \leqslant d_{\mathcal{E}}(f(\boldsymbol{x}), f(\boldsymbol{x}')) \leqslant (1+\epsilon)g\big(d_{\mathcal{S}}(\boldsymbol{x},\boldsymbol{x}')\big) + \epsilon',$$

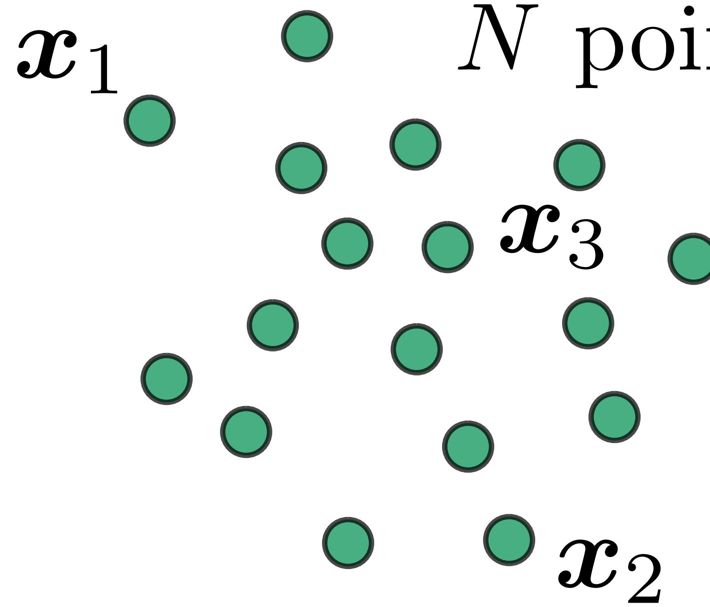
for all $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{S}$, with high probability.

As will be clearer later:

- f can be linear, quantized, periodic, non-linear, ...
- Tradeoffs expected between ϵ, ϵ', $\dim \mathcal{S}$, and $\dim \mathcal{E}$.
- Random constructions (hence probability)
- ϵ' can be zero (*e.g.*, for linear f)
- $g \neq \mathrm{Id}$ for periodic and non-linear f

Let's study a few examples ...

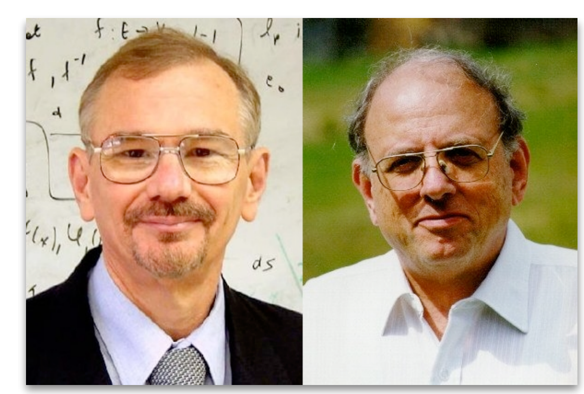
Johnson-Lindenstrauss embedding (1984)



N points

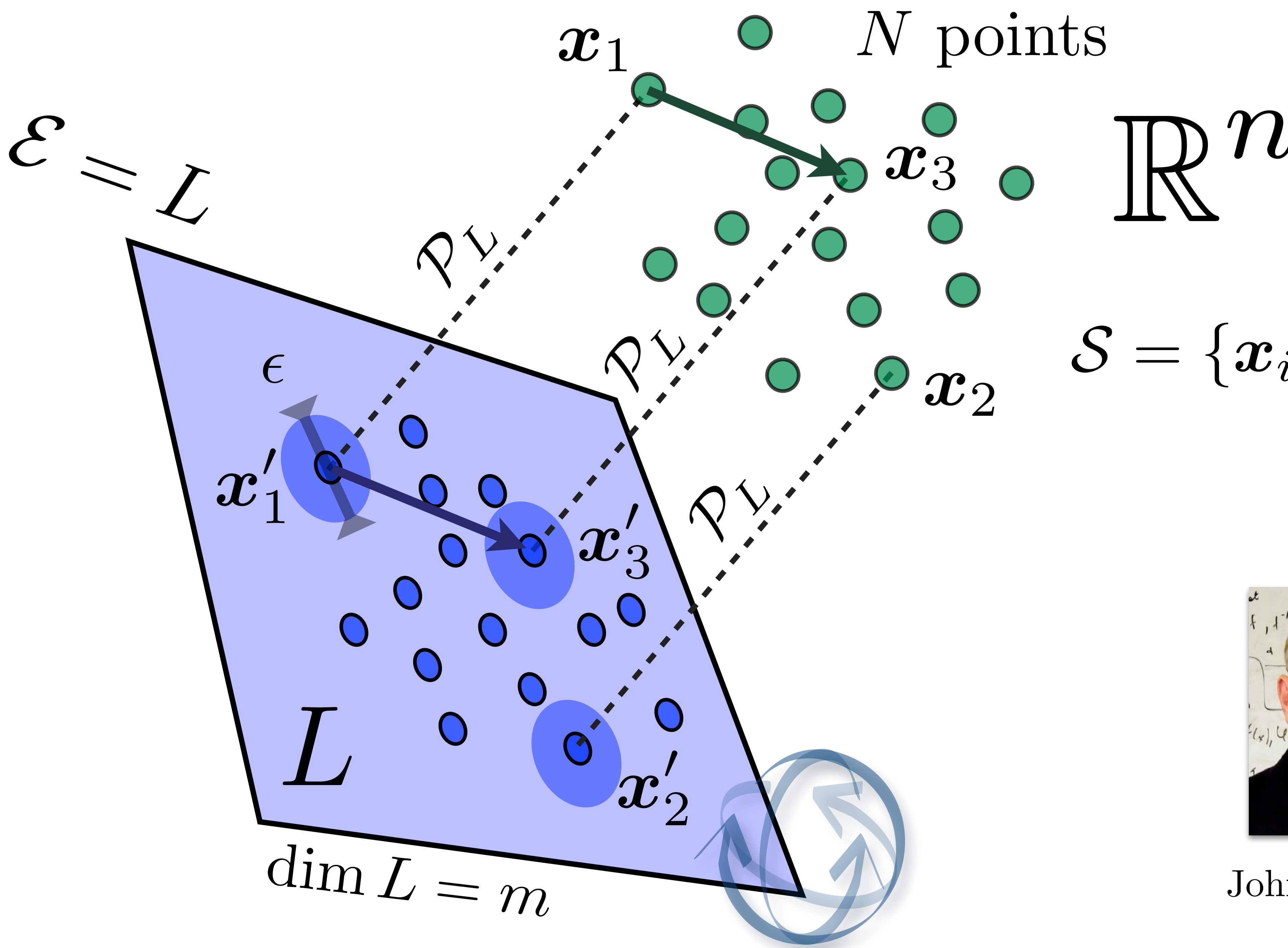
$$\mathbb{R}^n$$

$$\mathcal{S} = \{\boldsymbol{x}_i : 1 \leqslant i \leqslant N\}$$

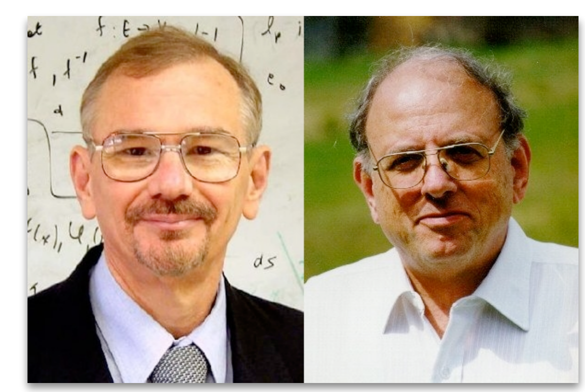


Johnson & Lindenstrauss

Johnson-Lindenstrauss embedding (1984)



$\mathcal{S} = \{x_i : 1 \leqslant i \leqslant N\}$



Johnson & Lindenstrauss

Random linear subspace $L \subset \mathbb{R}^n$

(amongst all possible linear subspaces with dimension m)

Johnson-Lindenstrauss embedding (1984)

For any $0 < \epsilon < 1$, provided

$$m \geqslant 4(\epsilon^2/2 - \epsilon^3/3)^{-1} \log N, \quad \text{"the tradeoff"}$$

there is a (linear) map $f : \mathbb{R}^n \to \mathbb{R}^m$ such that, for all $1 \leqslant i, j \leqslant N$,

$$(1-\epsilon)\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2 \leqslant \|f(\boldsymbol{x}_i) - f(\boldsymbol{x}_j)\|^2 \leqslant (1+\epsilon)\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2.$$

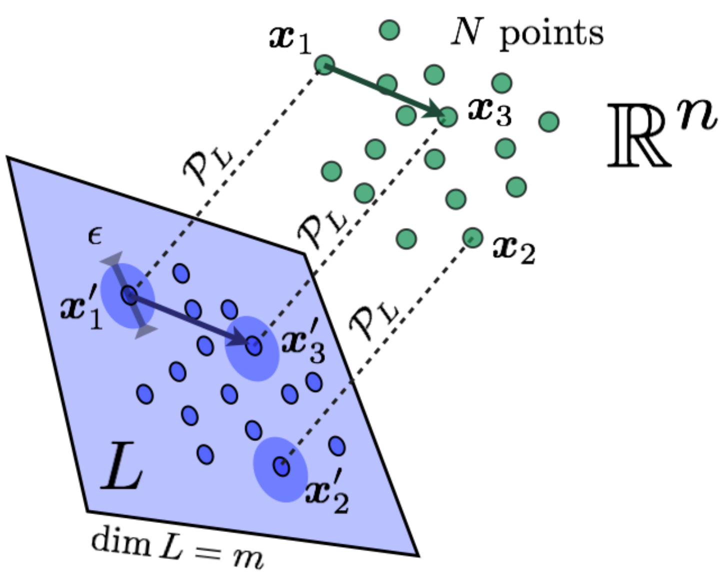
For instance, with high probability, $f(\boldsymbol{u}) = \sqrt{\frac{n}{m}}\mathcal{P}_L \boldsymbol{u}$ works with $L \sim_{\text{unif. randomly}}$ all m-dim. subspaces.

<u>Here:</u>
- $\mathcal{S} = \{\boldsymbol{x}_i : 1 \leqslant i \leqslant N\} \subset \mathbb{R}^N$, $\mathcal{E} = \mathbb{R}^m$,
- $d_{\mathcal{S}} \equiv d_{\mathcal{E}} \equiv$ Euclidean distance

<u>Remark:</u>

$\log N \simeq$ "dimension" of $\{\boldsymbol{x}_i : 1 \leqslant i \leqslant N\}$

(more on this after)



Random linear subspace $L \subset \mathbb{R}^n$

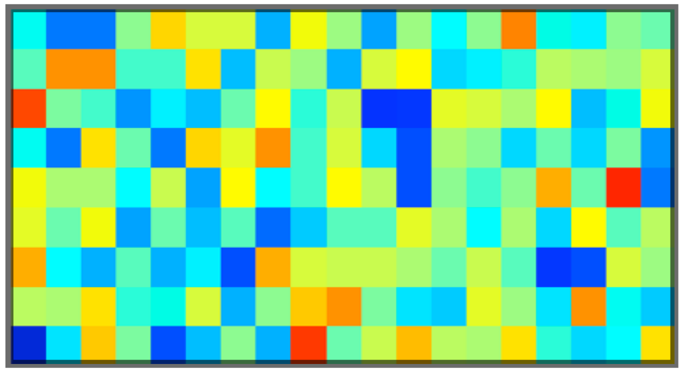
Johnson-Lindenstrauss embedding (variants)

Provided*

$$m \geqslant C\epsilon^{-2}\log N,$$

if $\mathbf{\Phi} \in \mathbb{R}^{m\times n}$ with $\Phi_{ij} \sim_{\text{iid}} \mathcal{N}(0,1)$, then, with probability exceeding $1 - C\exp(-c\epsilon^2 m)$, for all $1 \leqslant i,j \leqslant N$,

$$(1-\epsilon)\|\boldsymbol{x}_i - \boldsymbol{x}_j\| \leqslant \sqrt{\tfrac{1}{m}}\|\mathbf{\Phi}\boldsymbol{x}_i - \mathbf{\Phi}\boldsymbol{x}_j\| \leqslant (1+\epsilon)\|\boldsymbol{x}_i - \boldsymbol{x}_j\|.$$

$\mathbf{\Phi} =$ 

Matrix-vector multiplication:

$O(mn)$ operations (heavy!)

Johnson-Lindenstrauss embedding (variants)

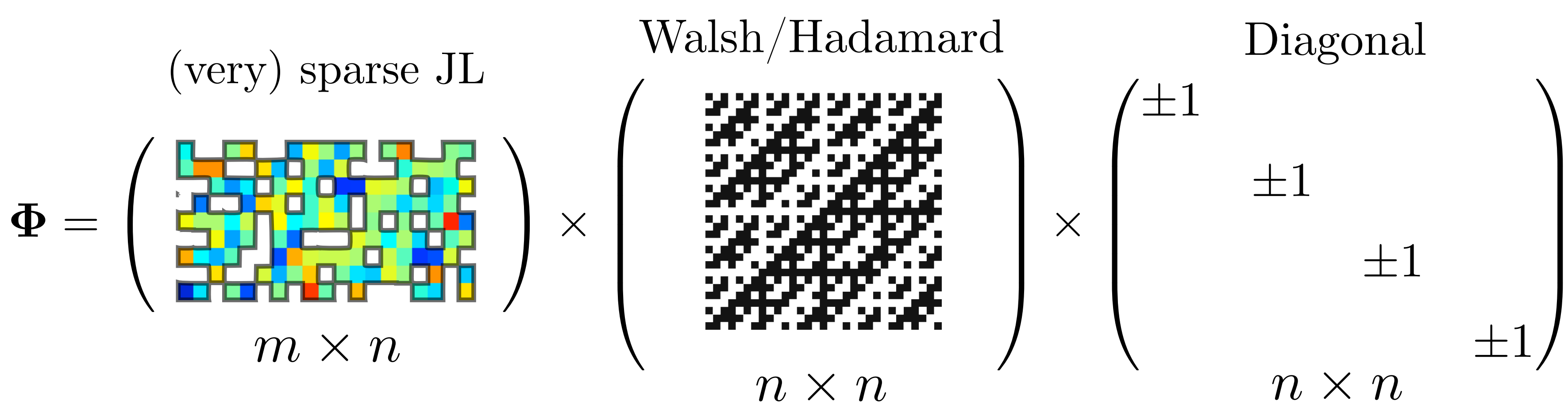
Other variants:

- Sparse JL (e.g., **[Achlioptas, '03]**):

$$\Phi_{ij} \sim_{\text{iid}} \begin{cases} 1/\sqrt{3} & \text{with } p = 1/6, \\ 0 & \text{with } p = 2/3 \\ -1/\sqrt{3}, & \text{with } p = 1/6. \end{cases}$$

- Structured matrices:

  Fast Johnson Lindenstrauss Transform **[Ailon, Chazelle, '09]**

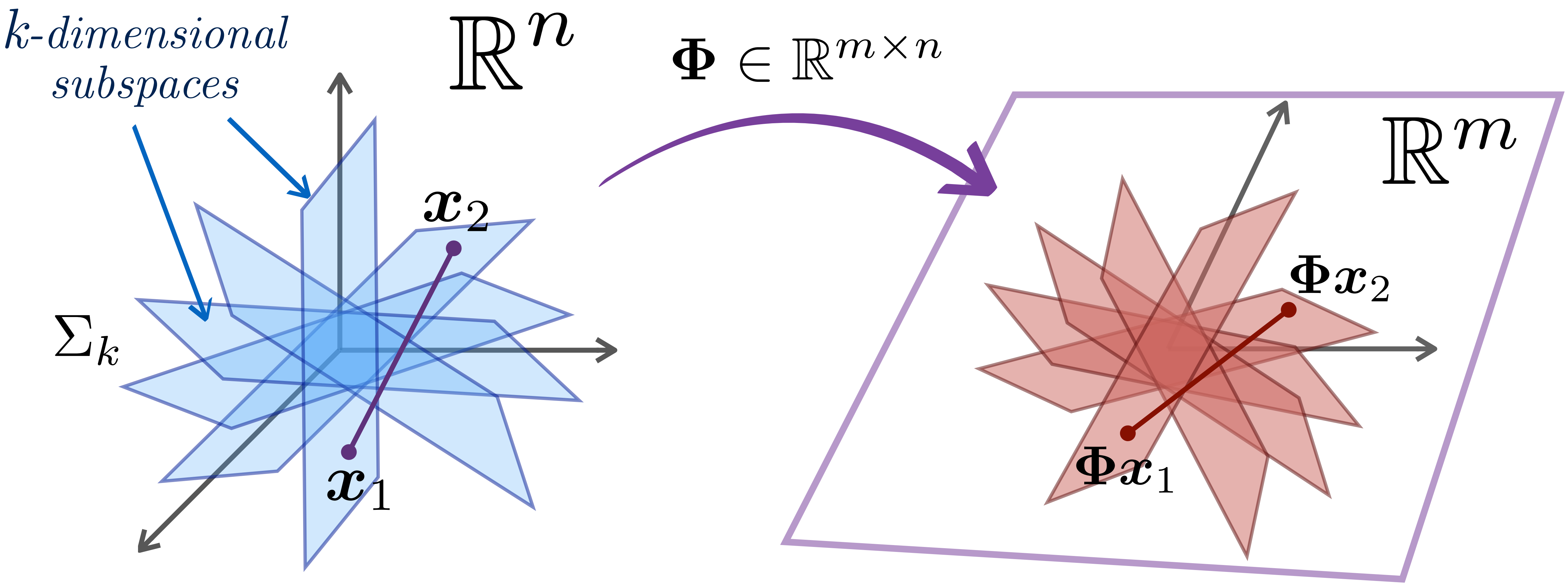


Complexity: $\quad O(n \log n + \epsilon^{-2} \min(n \log N, \log^3 N))$

Remark: FJL ok for $d_{\mathcal{E}}(\boldsymbol{y}, \boldsymbol{y}') = \|\boldsymbol{y} - \boldsymbol{y}'\|_1 = \sum_j |y_j - y'_j|$ (the ℓ_1-norm)

Restricted Isometry Property (RIP)

- 1^{st} example of embedding of continuous sets
- Preserving geometry of *sparse* vectors

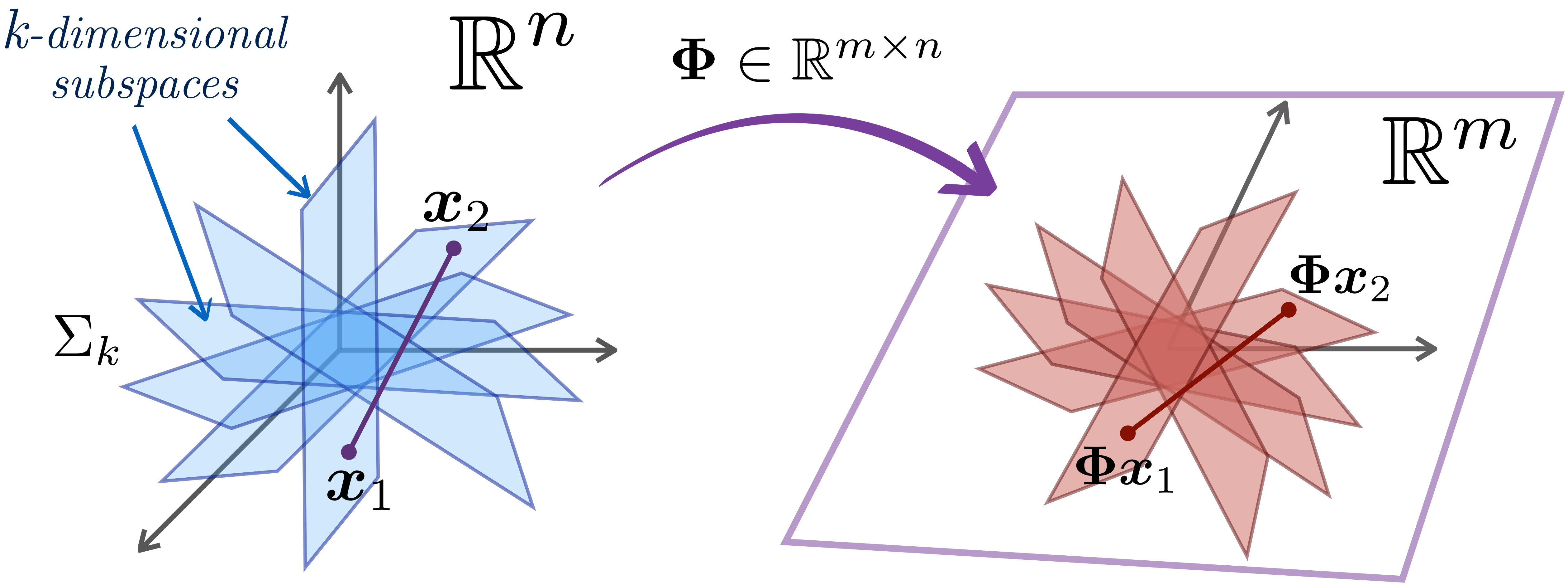


Restricted Isometry Property (RIP)

RIP over $\Sigma_k - \Sigma_k = \Sigma_{2k}$: $\mathrm{RIP}(\Sigma_{2k}, \epsilon)$

For all $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \Sigma_k := \{\boldsymbol{u} \in \mathbb{R}^n : \|\boldsymbol{u}\|_0 := |\operatorname{supp} \boldsymbol{u}| \leqslant k\}$,

$$(1-\epsilon)\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|^2 \leqslant \|\boldsymbol{\Phi}\boldsymbol{x}_1 - \boldsymbol{\Phi}\boldsymbol{x}_2\|^2 \leqslant (1+\epsilon)\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|^2.$$

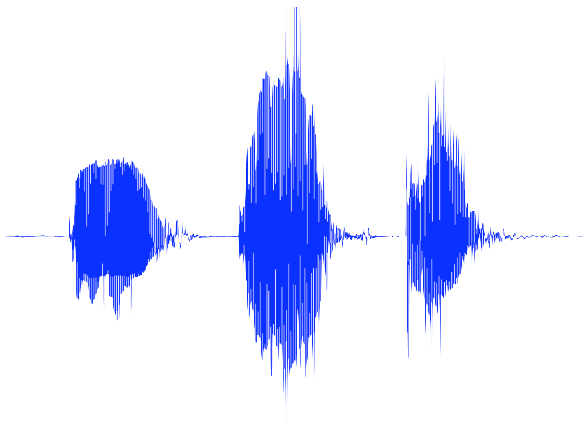
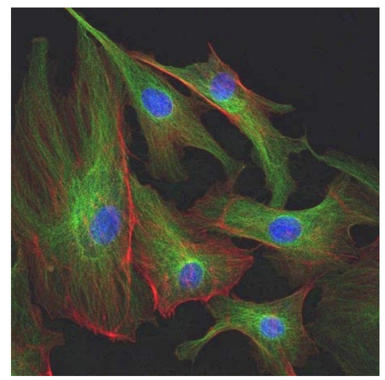
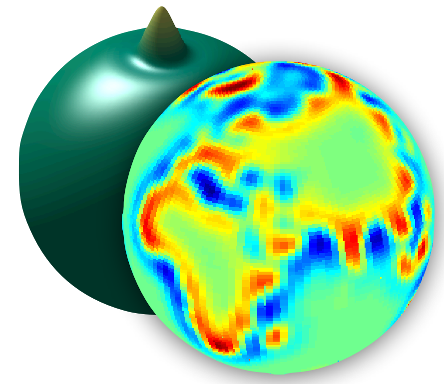
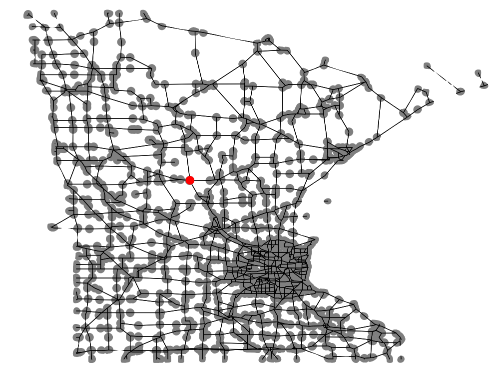


$$\mathcal{S} = \Sigma_k,\ \mathcal{E} = \boldsymbol{\Phi}\mathcal{S} \subset \mathbb{R}^m,\ f \equiv \boldsymbol{\Phi}$$

$$d_{\mathcal{S}} \equiv d_{\mathcal{E}} \equiv \text{Euclidean distance}$$

Restricted Isometry Property (RIP)

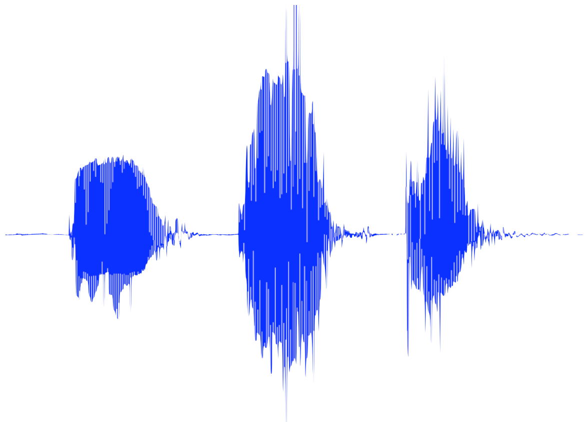
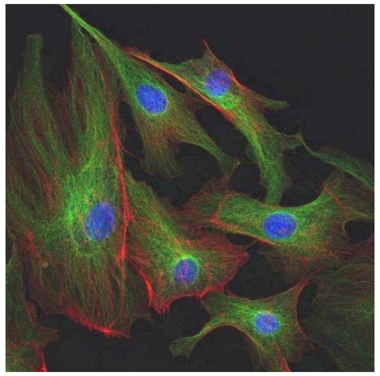
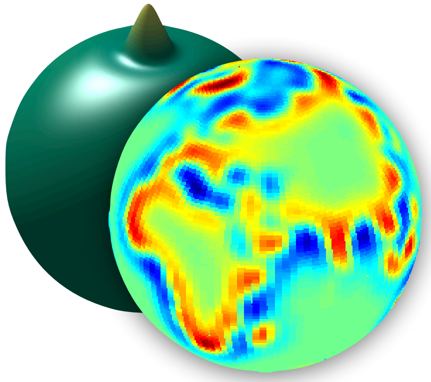
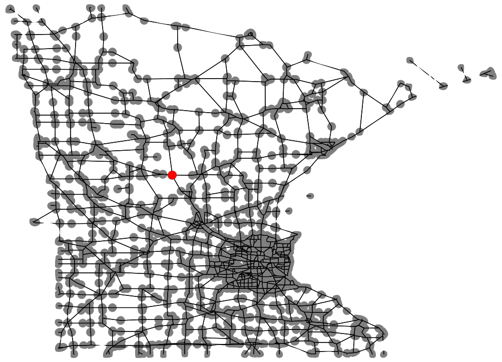
What about sparsity in a non-trivial basis?

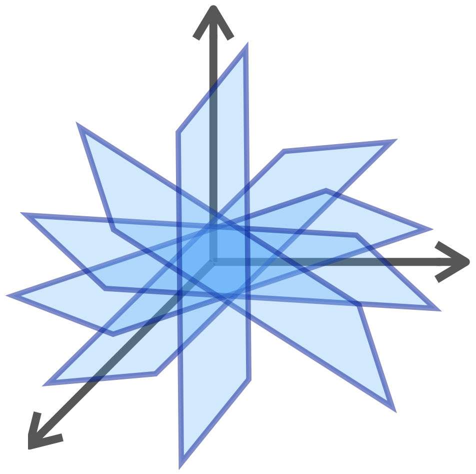
$$\boldsymbol{x} = \boldsymbol{\Psi}\boldsymbol{\alpha} = \sum_{i=1}^{N} \boldsymbol{\Psi}_i \alpha_i \text{ with } \|\boldsymbol{\alpha}\|_0 \leqslant k$$

Restricted Isometry Property (RIP)

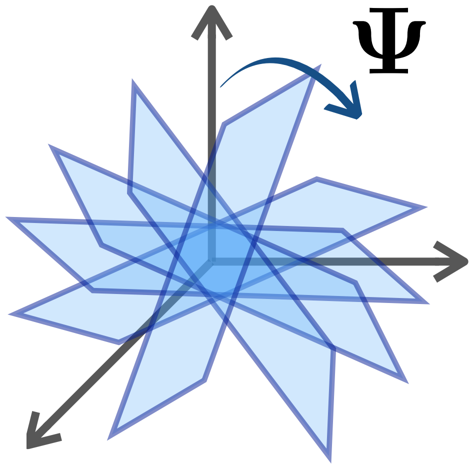
What about sparsity in a non-trivial basis?

$$\boldsymbol{x} = \boldsymbol{\Psi\alpha} = \sum_{i=1}^{N} \boldsymbol{\Psi}_i \alpha_i \text{ with } \|\boldsymbol{\alpha}\|_0 \leqslant k$$



$$\Sigma_k \to \boldsymbol{\Psi}\Sigma_k$$



$$\Leftrightarrow$$

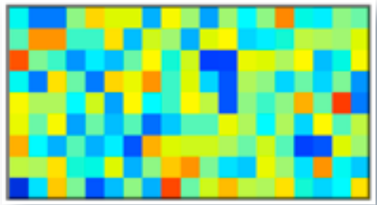
Embedding of $\boldsymbol{\Psi}\Sigma_k$ if $\boldsymbol{\Phi}' = \boldsymbol{\Phi\Psi}$ is $\mathrm{RIP}(\Sigma_k, \epsilon)$

Market of RIP matrices?

- Dense & unstructured sensing matrices (initial constructions):
  - random sub-Gaussian ensembles (e.g., Gaussian, Bernoulli)

e.g., Gaussian: $\mathbf{\Phi} \in \mathbb{R}^{m \times n}$, with $\Phi_{ij} \sim_{\text{iid}} \mathcal{N}(0,1)$

or $\Phi_{ij} \sim_{\text{iid}} \pm 1$ (eq. prob), $\cdots$



Sample complexity:

$$m \gtrsim \epsilon^{-2} k \log(n/k)$$

Universal sensing matrices:

They can be RIP$(\mathbf{\Psi}\Sigma_k, \epsilon)$ for any ONB $\mathbf{\Psi} \in \mathbb{R}^{n \times n}$.

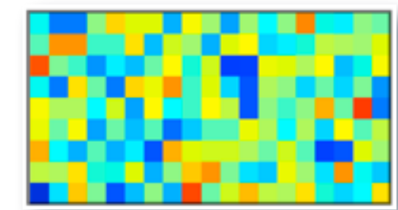
Matrix-vector multiplication:

$O(mn)$ operations (heavy!)

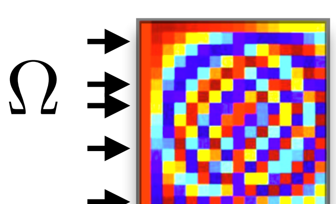
Market of RIP matrices?

- Dense & unstructured sensing matrices (initial constructions):
  - random sub-Gaussian ensembles (e.g., Gaussian, Bernoulli)

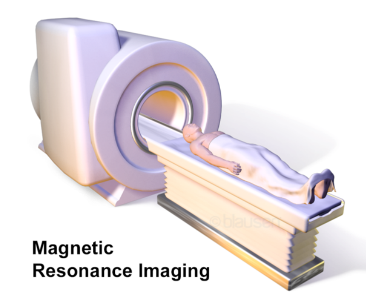
    e.g., Gaussian: $\boldsymbol{\Phi} \in \mathbb{R}^{m \times n}$, with $\Phi_{ij} \sim_{\rm iid} \mathcal{N}(0,1)$
    or $\Phi_{ij} \sim_{\rm iid} \pm 1$ (eq. prob), $\cdots$



- Structured sensing matrices (less memory, fast computations):
  - random Fourier/Hadamard ensembles (e.g., for CT, MRI, astron.);



$e.g.$, $\boldsymbol{\Phi} = \boldsymbol{F}_\Omega$, with $\boldsymbol{F} \in \mathbb{C}^{n \times n}$
and random $\Omega \subset \{1, \cdots, n\}$, $|\Omega| = m$



  - random convolutions, spread-spectrum (e.g., for imaging), ...

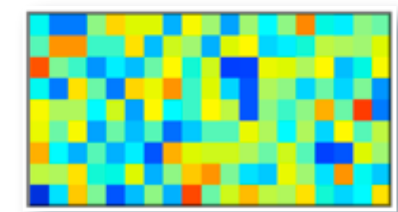
(see, $e.g.$, **[Foucart, Rauhut, 2013]**)

Market of RIP matrices?

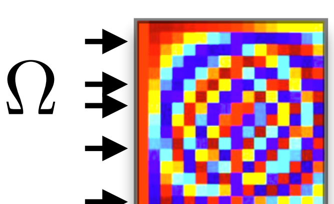
- Dense & unstructured sensing matrices (initial constructions):
  - random sub-Gaussian ensembles (e.g., Gaussian, Bernoulli)

    e.g., Gaussian: $\boldsymbol{\Phi} \in \mathbb{R}^{m \times n}$, with $\Phi_{ij} \sim_{\rm iid} \mathcal{N}(0,1)$

    or $\Phi_{ij} \sim_{\rm iid} \pm 1$ (eq. prob), $\cdots$

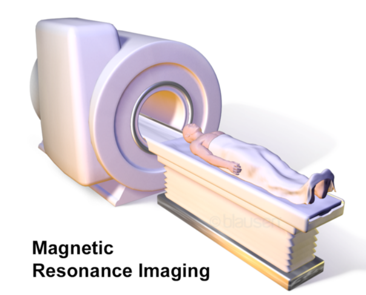


- Structured sensing matrices (less memory, fast computations):
  - random Fourier/Hadamard ensembles (e.g., for CT, MRI, astron.);



e.g., $\boldsymbol{\Phi} = \boldsymbol{F}_\Omega$, with $\boldsymbol{F} \in \mathbb{C}^{n \times n}$

and random $\Omega \subset \{1, \cdots, n\}$, $|\Omega| = m$



  - random convolutions, spread-spectrum (e.g., for imaging), ...

    (see, *e.g.*, **[Foucart, Rauhut, 2013]**)

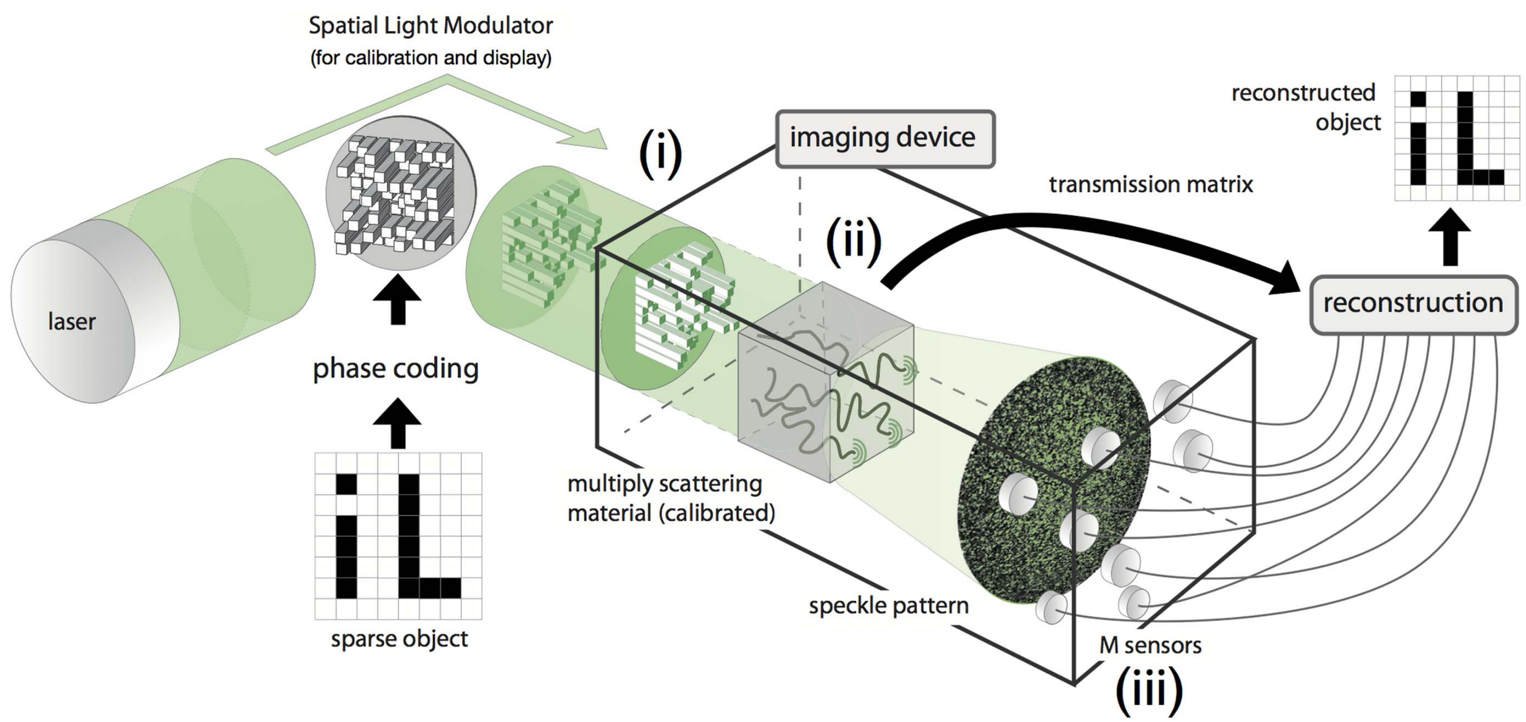
Sample complexity: $m \gtrsim \epsilon^{-2} k \,\mathrm{polylog}(\mathrm{dims}, \epsilon^{-1}, (\mathrm{prob.\ failure})^{-1})$

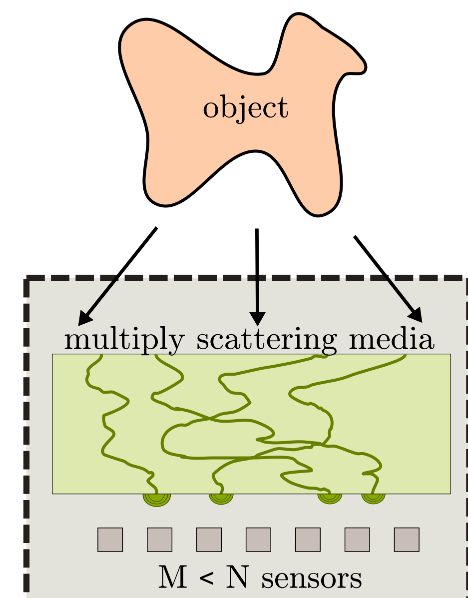
Less universal matrices; but complexity often reduced to $O(n \log n)$!

Market of RIP matrices?

[Liutkus et al., 14]

- Nature!





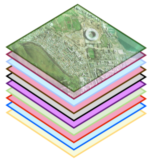
Pros:

- Random "for free"
- massively parallel/super fast
- allows random projections, imaging, classifications, ...

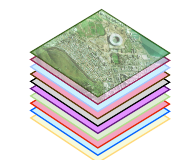
Cons:

- stable on a limited time (about 10')
- hard to characterize (but not always needed)

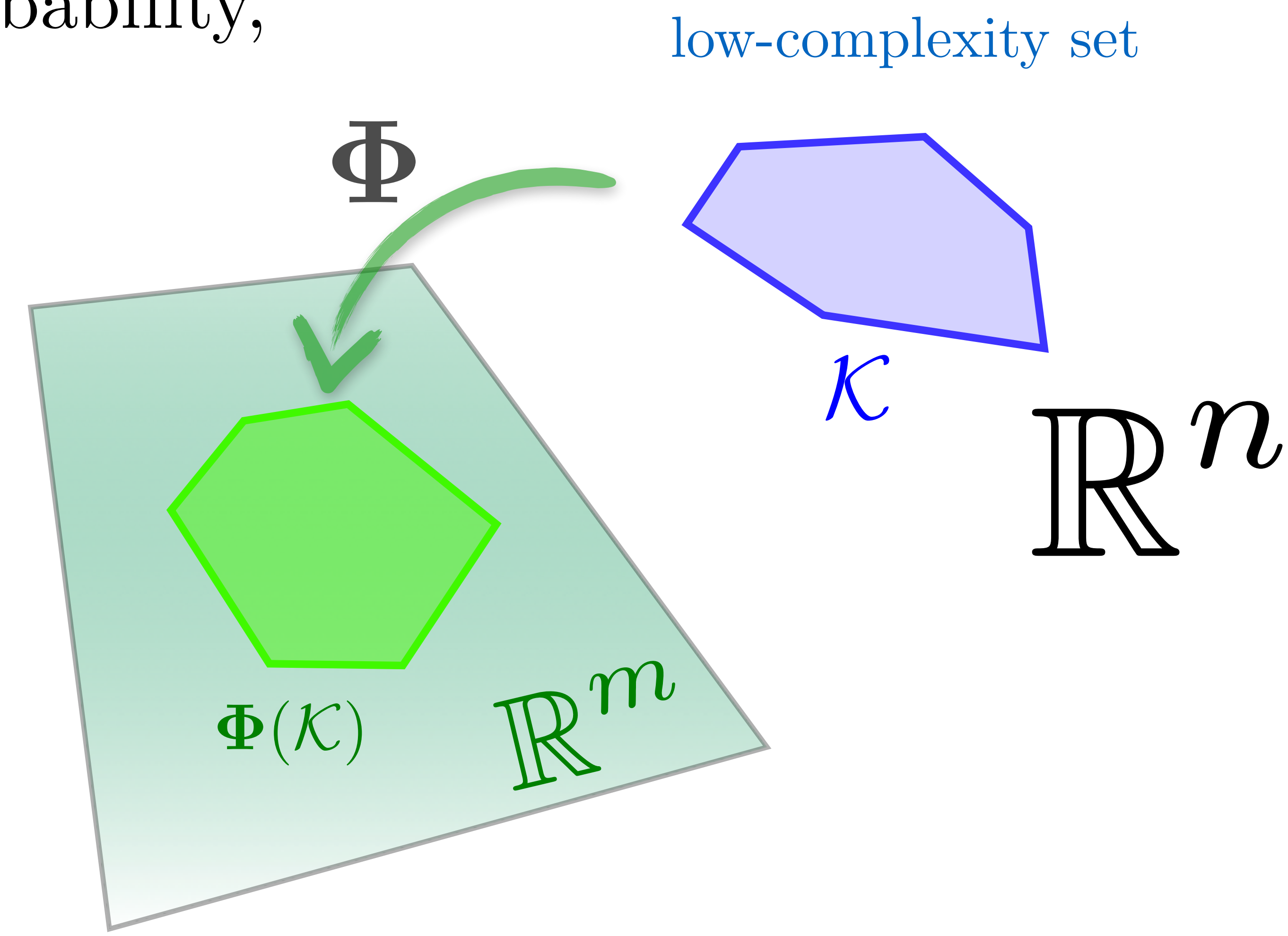
Beyond sparse signals ...

Two low-complexity (l.c.) signals $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{K}$ (e.g., low-rank data )

Beyond sparse signals ...

Two low-complexity (l.c.) signals $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{K}$ (e.g., low-rank data )

For many random constructions of $\boldsymbol{\Phi}$ (e.g., Gaussian, Bernoulli, structured) and "$m \gtrsim C_{\mathcal{K}}$", with high probability,



Geometry of $\boldsymbol{\Phi}(\mathcal{K})$
$\approx$ Geometry of $\mathcal{K}$

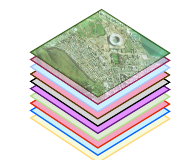
$$\boldsymbol{\Phi x} \approx \boldsymbol{\Phi x}' \;\Leftrightarrow\; \boldsymbol{x} \approx \boldsymbol{x}'$$

[see, e.g., Johnson, Lindenstrauss, Schechtman, Bourgain, Dirksen, Mendelson, Vershynin, Plan, Chandrasekaran, Puy, Gribonval, ...]

For all $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{K}$ and $0 < \epsilon < 1$,

$$(1-\epsilon)\|\boldsymbol{x}-\boldsymbol{x}'\|^2 \le \|\boldsymbol{\Phi x} - \boldsymbol{\Phi x}'\|^2 \le (1+\epsilon)\|\boldsymbol{x}-\boldsymbol{x}'\|^2$$

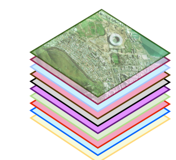
Beyond sparse signals ...

Two low-complexity (l.c.) signals $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{K}$ (e.g., low-rank data )

For many random constructions of $\boldsymbol{\Phi}$ (e.g., Gaussian, Bernoulli, structured)

If $m \gtrsim \epsilon^{-2} w(\mathcal{K})^2$ polylog(dimensions, ϵ^{-1}, (prob. of failure)$^{-1}$),
then $\boldsymbol{\Phi}$ is RIP$(\mathcal{K}, \epsilon)$ with high probability.

Beyond sparse signals ...

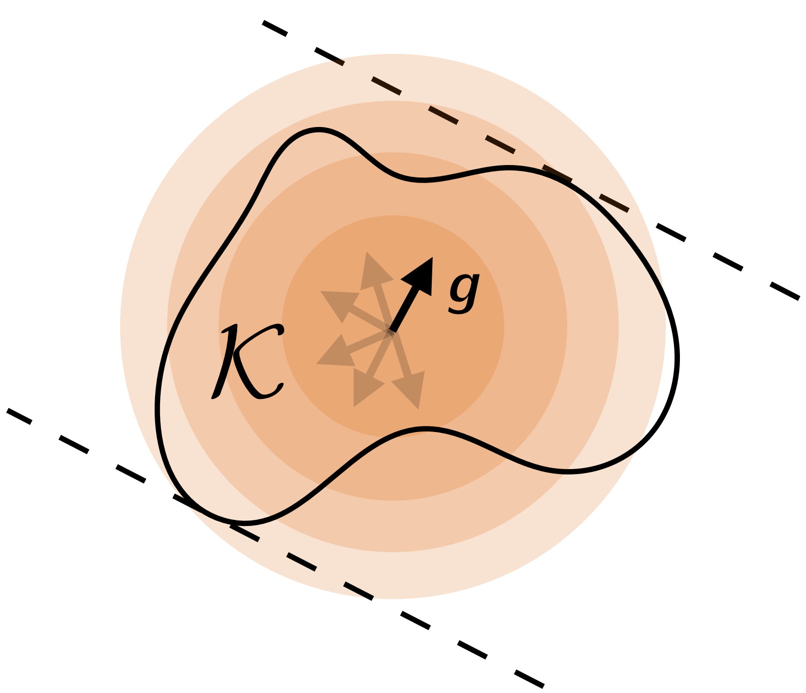
Two low-complexity (l.c.) signals $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{K}$ (e.g., low-rank data )

For many random constructions of $\boldsymbol{\Phi}$ (e.g., Gaussian, Bernoulli, structured)

If $m \gtrsim \epsilon^{-2} w(\mathcal{K})^2$ polylog(dimensions, ϵ^{-1}, (prob. of failure)$^{-1}$),
then $\boldsymbol{\Phi}$ is RIP$(\mathcal{K}, \epsilon)$ with high probability.

Let $\mathcal{K} \subset \mathbb{R}^n$, $\boldsymbol{g} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$,

$$w(\mathcal{K}) = \mathbb{E}_{\boldsymbol{g}} \sup_{\boldsymbol{x} \in \mathcal{K}} |\langle \boldsymbol{x}, \boldsymbol{g} \rangle|$$



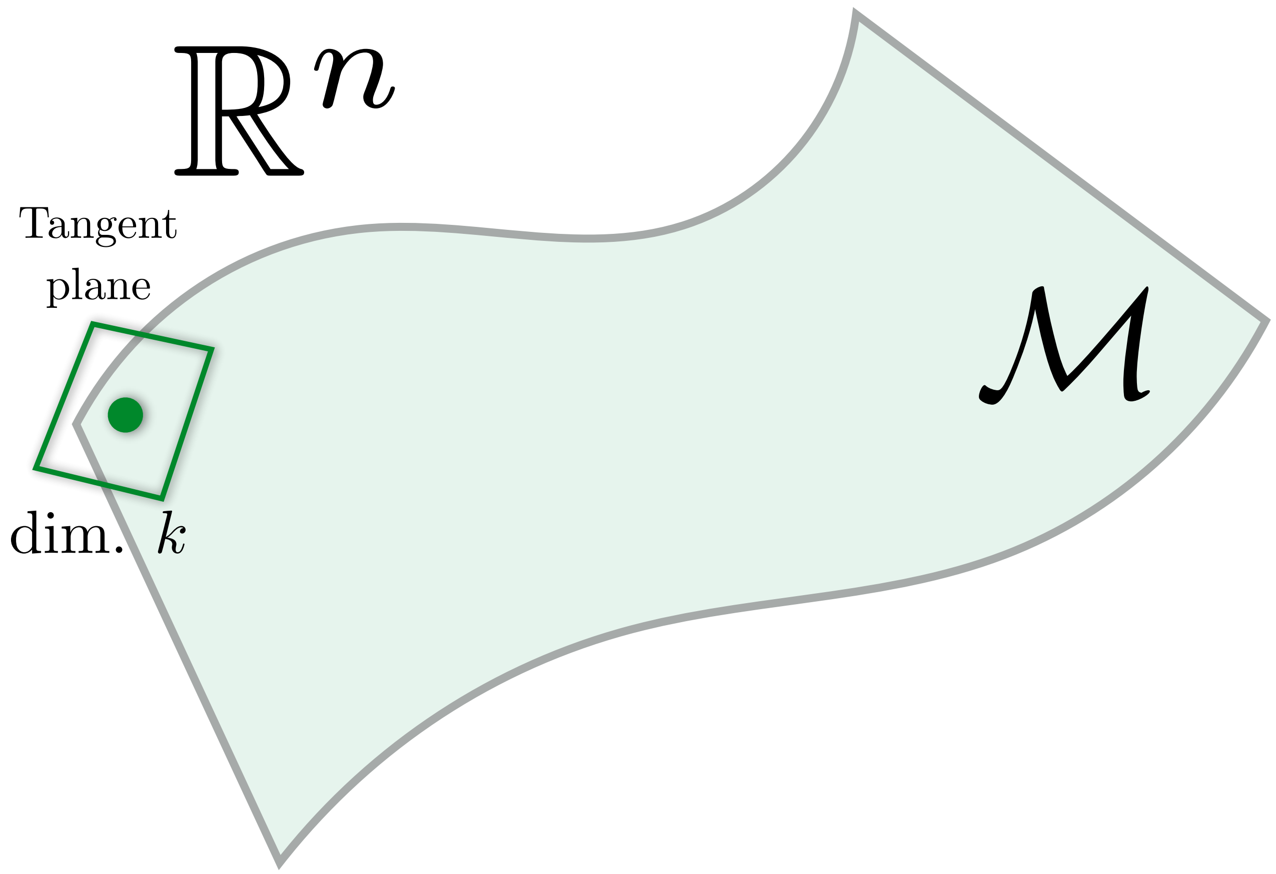
Examples:

$$w^2(\mathcal{K}) \lesssim \log |\mathcal{K}|$$
$$w^2(\mathbb{B}^n) \lesssim n$$
$$w^2(\Sigma_k^n \cap \mathbb{B}^n) \lesssim k \log(n/k)$$
$$w^2(\mathcal{M}_r \cap \mathbb{B}_F^{n \times n}) \lesssim rn$$
$$w^2(\cup_{i=1}^T \mathcal{K}_i) \lesssim \log T + \max_i w^2(\mathcal{K}_i)$$
$$\vdots$$

We met them before!

RIP for more general spaces

- Embedding a manifold (k-dim, smooth & compact):



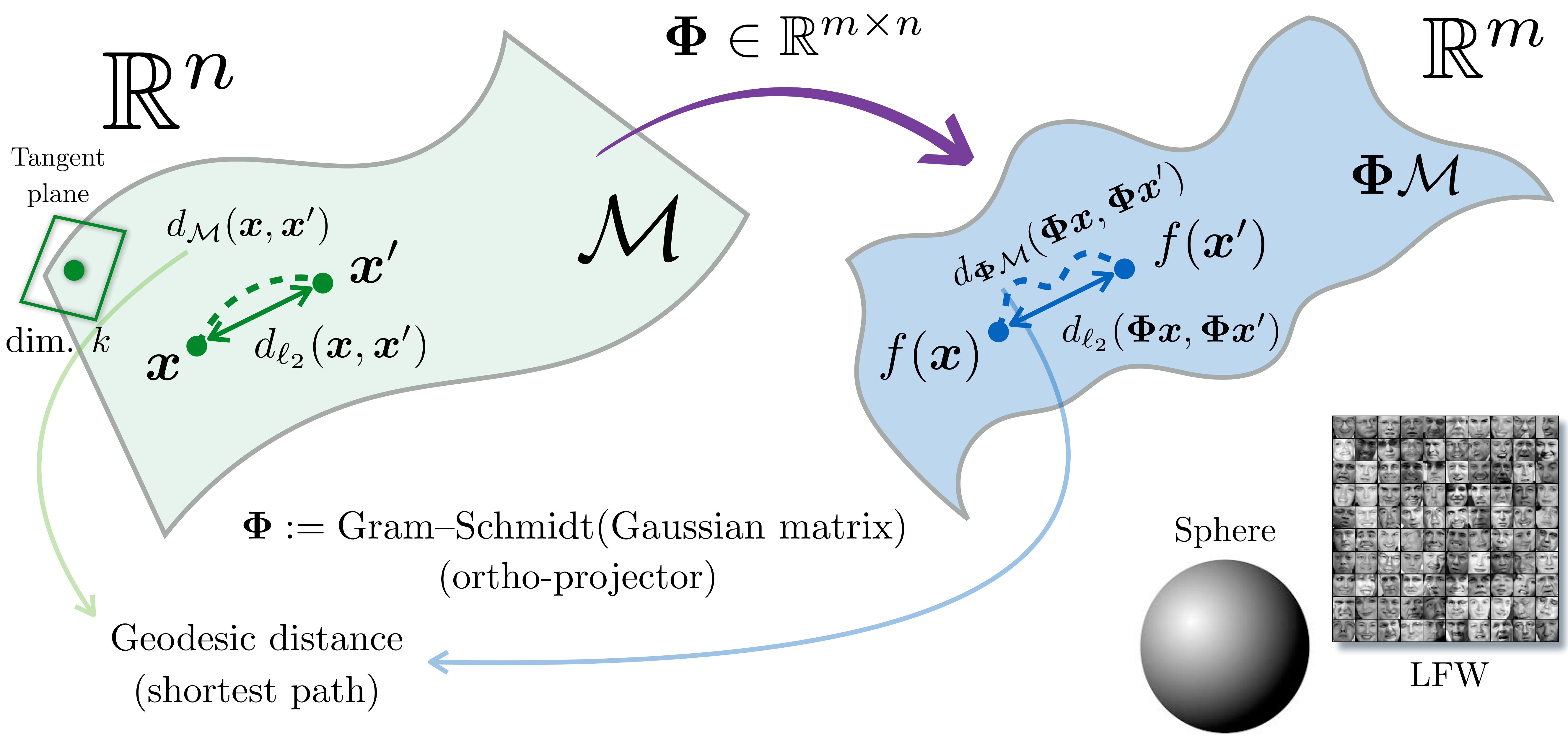
Sphere

LFW

Examples: sphere (e.g., the Earth), time-delay of a signal, phase valued data (e.g., in optics), appearance of a parametric image/models, ...

RIP for more general spaces

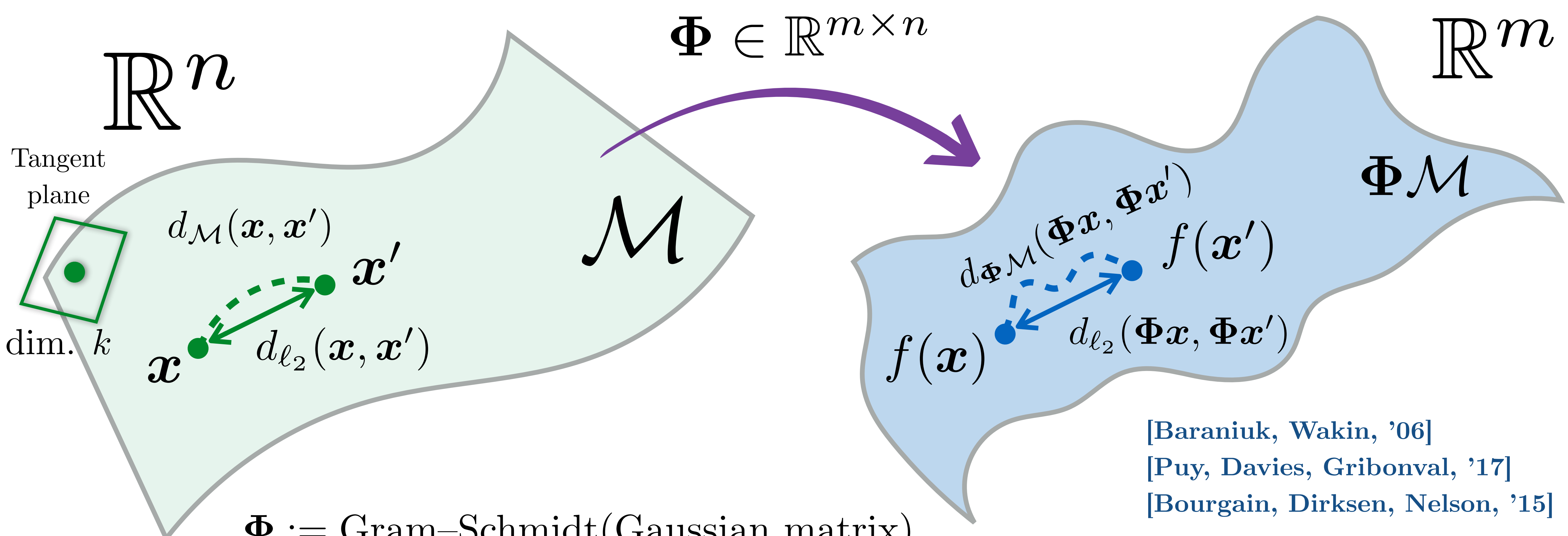
- Embedding a manifold (k-dim, smooth & compact):



<u>Examples</u>: sphere (e.g., the Earth), time-delay of a signal, phase valued data (e.g., in optics), appearance of a parametric image/models, ...

RIP for more general spaces

- Embedding a manifold (k-dim, smooth & compact):



[Baraniuk, Wakin, '06]
[Puy, Davies, Gribonval, '17]
[Bourgain, Dirksen, Nelson, '15]
...

$\boldsymbol{\Phi} :=$ Gram–Schmidt(Gaussian matrix)
(ortho-projector)

Smoothness & covering regularity

Provided that

$$m \gtrsim \epsilon^{-2} k \log(n \,\mathrm{Vol}(\mathcal{M})\, C_{\mathcal{M}})$$

Then, w.h.p.,

$$d_{\ell_2}(\boldsymbol{\Phi x}, \boldsymbol{\Phi x}') \approx_\epsilon d_{\ell_2}(\boldsymbol{x}, \boldsymbol{x}')$$

$$d_{\boldsymbol{\Phi}\mathcal{M}}(\boldsymbol{\Phi x}, \boldsymbol{\Phi x}') \approx_\epsilon d_{\mathcal{M}}(\boldsymbol{x}, \boldsymbol{x}')$$

Exercise: JL involves RIP! (for sparse signals)

Global idea: tightly *sample* Σ_k, extends JL lemma by continuity!

Exercise: JL involves RIP! (for sparse signals)

Global idea: tightly *sample* Σ_k, extends JL lemma by continuity!

- Sparse signals belong to a *union of (k-dim) subspaces*

$$\Sigma_k = \bigcup_{T\subset\{1,\cdots,N\}:\,|T|=k} \Sigma_T, \quad \Sigma_T := \{\boldsymbol{u} : \operatorname{supp} \boldsymbol{u} = T\}$$

Exercise: JL involves RIP! (for sparse signals)

Global idea: tightly *sample* Σ_k, extends JL lemma by continuity!

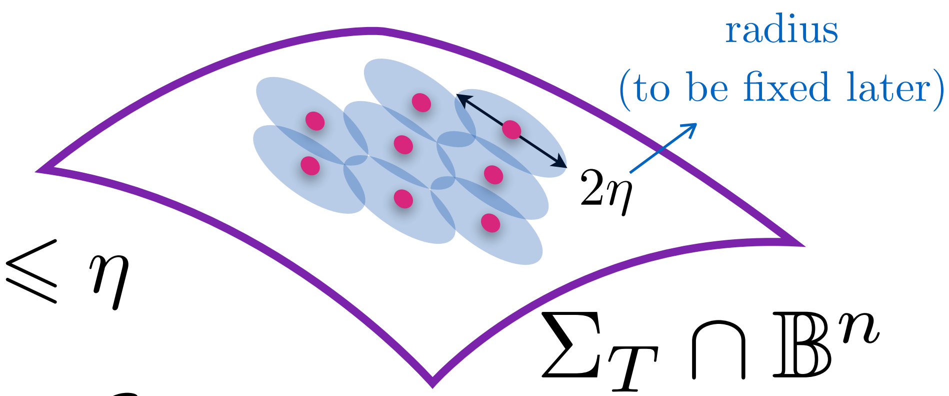
- Sparse signals belong to a *union of (k-dim) subspaces*

$$\Sigma_k = \bigcup_{T \subset \{1,\cdots,N\}:\, |T|=k} \Sigma_T, \qquad \Sigma_T := \{\boldsymbol{u} : \operatorname{supp} \boldsymbol{u} = T\}$$

- Each subspace, restricted to a ball, can be *covered* (i.e., sampled)

Optimal η-covering $\mathcal{G}_{\eta,T}$ of Σ_T: (radius ↑ η)

$\forall \boldsymbol{x} \in \Sigma_T \cap \mathbb{B}^n, \exists \boldsymbol{q} \in \mathcal{G}_{\eta,T}$ s.t. $\|\boldsymbol{x} - \boldsymbol{q}\| \leqslant \eta$



η-Covering of $\Sigma_k \cap \mathbb{B}^n$: $\mathcal{G}_\eta := \bigcup_{T:|T|=k} \mathcal{G}_{\eta,T}$

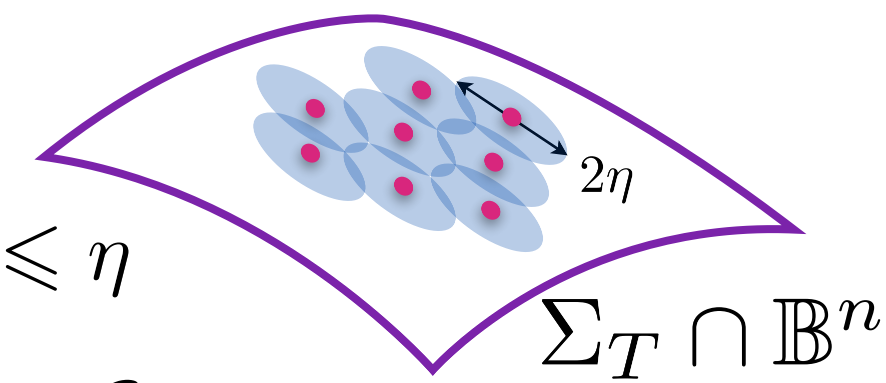
Exercise: JL involves RIP! (for sparse signals)

Global idea: tightly *sample* Σ_k, extends JL lemma by continuity!

- Sparse signals belong to a *union of (k-dim) subspaces*

$$\Sigma_k = \bigcup_{T\subset\{1,\cdots,N\}:\,|T|=k} \Sigma_T, \qquad \Sigma_T := \{\boldsymbol u : \operatorname{supp}\boldsymbol u = T\}$$

- Each subspace, restricted to a ball, can be *covered* (i.e., sampled)



Optimal η-covering $\mathcal{G}_{\eta,T}$ of Σ_T:

$\forall \boldsymbol x \in \Sigma_T \cap \mathbb{B}^n, \exists \boldsymbol q \in \mathcal{G}_{\eta,T}$ s.t. $\|\boldsymbol x - \boldsymbol q\| \leqslant \eta$

η-Covering of $\Sigma_k \cap \mathbb{B}^n$: $\mathcal{G}_\eta := \bigcup_{T:|T|=k} \mathcal{G}_{\eta,T}$

- Covering cardinality is bounded:

$$\left.\begin{array}{l}\Sigma_T \cap \mathbb{B}^n \simeq \mathbb{B}^k \Rightarrow |\mathcal{G}_{\eta,T}| \leqslant (1+2/\eta)^k \\ \text{No more than } \binom{n}{k} \text{ supports } T\end{array}\right\} \Rightarrow |\mathcal{G}_\eta| \leqslant \binom{n}{k}(1+2/\eta)^k \underset{\text{Stirling}}{\leqslant} \left(\tfrac{3en}{k\eta}\right)^k$$

(other bounds exist for, e.g., low-rank matrices, and other conic spaces)

Exercise: JL involves RIP! (for sparse signals)

Global idea: tightly *sample* Σ_k, extends JL lemma by continuity!

- Apply JL lemma to the covering:

  Given $\epsilon > 0$, provided

  $$m \geqslant C\epsilon^{-2} \log|\mathcal{G}_\eta| \simeq C\epsilon^{-2} k \log(\tfrac{n}{\eta k}),$$

  if $\boldsymbol{\Phi} \in \mathbb{R}^{m \times n}$ with $\Phi_{ij} \sim_{\text{iid}} \mathcal{N}(0,1)$, then, with probability exceeding $1 - C\exp(-c\epsilon^2 m)$, for all $\boldsymbol{q} \in \mathcal{G}_\eta$

  $$(1-\epsilon)\|\boldsymbol{q}\| \leqslant \sqrt{\tfrac{1}{m}}\|\boldsymbol{\Phi q}\| \leqslant (1+\epsilon)\|\boldsymbol{q}\|.$$

Exercise: JL involves RIP! (for sparse signals)

Global idea: tightly *sample* Σ_k, extends JL lemma by continuity!

- Apply JL lemma to the covering:

  Given $\epsilon > 0$, provided

$$m \geqslant C\epsilon^{-2} \log |\mathcal{G}_\eta| \simeq C\epsilon^{-2} k \log(\tfrac{n}{\eta k}),$$

  if $\boldsymbol{\Phi} \in \mathbb{R}^{m\times n}$ with $\Phi_{ij} \sim_{\rm iid} \mathcal{N}(0,1)$, then, with probability exceeding $1 - C\exp(-c\epsilon^2 m)$, for all $\boldsymbol{q} \in \mathcal{G}_\eta$

$$(1-\epsilon)\|\boldsymbol{q}\| \leqslant \sqrt{\tfrac{1}{m}}\|\boldsymbol{\Phi q}\| \leqslant (1+\epsilon)\|\boldsymbol{q}\|.$$

- Continuity extension: Let $\boldsymbol{\Phi}' = \frac{1}{\sqrt{m}}\boldsymbol{\Phi}$, $\boldsymbol{x} \in \Sigma_k$ with $\|\boldsymbol{x}\| = 1$ (WLOG), and $\boldsymbol{q} \in \mathcal{G}_\eta$ with $\|\boldsymbol{x} - \boldsymbol{q}\| \leqslant \eta$ & $\operatorname{supp} \boldsymbol{x} = \operatorname{supp} \boldsymbol{q}$.

$$\begin{aligned}\|\boldsymbol{\Phi}'\boldsymbol{x}\| &\leqslant \|\boldsymbol{\Phi}'\boldsymbol{q}\| + \|\boldsymbol{\Phi}'(\boldsymbol{x}-\boldsymbol{q})\| \leqslant (1+\epsilon)\|\boldsymbol{q}\| + \|\boldsymbol{\Phi}'(\tfrac{\boldsymbol{x}-\boldsymbol{q}}{\|\boldsymbol{x}-\boldsymbol{q}\|})\|\|\boldsymbol{x}-\boldsymbol{q}\| \\ &\leqslant (1+\epsilon)\|\boldsymbol{x}\| + (1+\epsilon)\|\boldsymbol{q}-\boldsymbol{x}\| + \|\boldsymbol{x}-\boldsymbol{q}\|\|\boldsymbol{\Phi}'(\tfrac{\boldsymbol{x}-\boldsymbol{q}}{\|\boldsymbol{x}-\boldsymbol{q}\|})\| \\ &\leqslant (1+\epsilon)(1+\eta) + \eta\,\|\boldsymbol{\Phi}'(\underbrace{\tfrac{\boldsymbol{x}-\boldsymbol{q}}{\|\boldsymbol{x}-\boldsymbol{q}\|}}_{\boldsymbol{r}^{(1)} \in \Sigma_k,\ \text{with } \|\boldsymbol{r}^{(1)}\| = 1})\|\end{aligned}$$

Exercise: JL involves RIP! (for sparse signals)

Global idea: tightly *sample* Σ_k, extends JL lemma by continuity!

- Apply JL lemma to the covering:

  Given $\epsilon > 0$, provided

$$m \geqslant C\epsilon^{-2} \log |\mathcal{G}_\eta| \simeq C\epsilon^{-2} k \log(\tfrac{n}{\eta k}),$$

  if $\boldsymbol{\Phi} \in \mathbb{R}^{m\times n}$ with $\Phi_{ij} \sim_{\rm iid} \mathcal{N}(0,1)$, then, with probability exceeding $1 - C\exp(-c\epsilon^2 m)$, for all $\boldsymbol{q} \in \mathcal{G}_\eta$

$$(1-\epsilon)\|\boldsymbol{q}\| \leqslant \sqrt{\tfrac{1}{m}}\|\boldsymbol{\Phi q}\| \leqslant (1+\epsilon)\|\boldsymbol{q}\|.$$

- Continuity extension: Let $\boldsymbol{\Phi}' = \frac{1}{\sqrt{m}}\boldsymbol{\Phi}$, $\boldsymbol{x} \in \Sigma_k$ with $\|\boldsymbol{x}\| = 1$ (WLOG), and $\boldsymbol{q} \in \mathcal{G}_\eta$ with $\|\boldsymbol{x} - \boldsymbol{q}\| \leqslant \eta$ & $\operatorname{supp}\boldsymbol{x} = \operatorname{supp}\boldsymbol{q}$.

$$\|\boldsymbol{\Phi}'\boldsymbol{x}\| \leqslant (1+\epsilon)(1+\eta) + \eta\,\|\boldsymbol{\Phi}'\boldsymbol{r}^{(1)}\|$$

Exercise: JL involves RIP! (for sparse signals)

Global idea: tightly *sample* Σ_k, extends JL lemma by continuity!

- Apply JL lemma to the covering:

  Given $\epsilon > 0$, provided

  $$m \geqslant C\epsilon^{-2} \log |\mathcal{G}_\eta| \simeq C\epsilon^{-2} k \log(\tfrac{n}{\eta k}),$$

  if $\boldsymbol{\Phi} \in \mathbb{R}^{m\times n}$ with $\Phi_{ij} \sim_{\rm iid} \mathcal{N}(0,1)$, then, with probability exceeding $1 - C\exp(-c\epsilon^2 m)$, for all $\boldsymbol{q} \in \mathcal{G}_\eta$

  $$(1-\epsilon)\|\boldsymbol{q}\| \leqslant \sqrt{\tfrac{1}{m}}\|\boldsymbol{\Phi q}\| \leqslant (1+\epsilon)\|\boldsymbol{q}\|.$$

- Continuity extension: Let $\boldsymbol{\Phi}' = \frac{1}{\sqrt{m}}\boldsymbol{\Phi}$, $\boldsymbol{x} \in \Sigma_k$ with $\|\boldsymbol{x}\| = 1$ (WLOG), and $\boldsymbol{q} \in \mathcal{G}_\eta$ with $\|\boldsymbol{x} - \boldsymbol{q}\| \leqslant \eta$ & $\operatorname{supp} \boldsymbol{x} = \operatorname{supp} \boldsymbol{q}$.

$$\begin{aligned}\|\boldsymbol{\Phi}'\boldsymbol{x}\| &\leqslant (1+\epsilon)(1+\eta) + \eta\,\|\boldsymbol{\Phi}'\boldsymbol{r}^{(1)}\| \\ &\leqslant (1+\epsilon)(1+\eta) + \eta\,\big((1+\epsilon)(1+\eta) + \eta\,\|\boldsymbol{\Phi}'\underbrace{\boldsymbol{r}^{(2)}}_{\in \Sigma_k,\ \text{with } \|\boldsymbol{r}^{(2)}\| = 1}\|\big) \\ &\cdots \\ &\leqslant (1+\epsilon)(1+\eta)\textstyle\sum_{i=0}^{+\infty} \eta^i = (1+\epsilon)\frac{(1+\eta)}{1-\eta}\end{aligned}$$

Exercise: JL involves RIP! (for sparse signals)

2/2

Global idea: tightly *sample* Σ_k, extends JL lemma by continuity!

- Apply JL lemma to the covering:

  Given $\epsilon > 0$, provided

  $$m \geqslant C\epsilon^{-2} \log |\mathcal{G}_\eta| \simeq C\epsilon^{-2} k \log(\tfrac{n}{\eta k}),$$

  if $\boldsymbol{\Phi} \in \mathbb{R}^{m\times n}$ with $\Phi_{ij} \sim_{\rm iid} \mathcal{N}(0,1)$, then, with probability exceeding $1 - C\exp(-c\epsilon^2 m)$, for all $\boldsymbol{q} \in \mathcal{G}_\eta$

  $$(1-\epsilon)\|\boldsymbol{q}\| \leqslant \sqrt{\tfrac{1}{m}}\|\boldsymbol{\Phi q}\| \leqslant (1+\epsilon)\|\boldsymbol{q}\|.$$

- Continuity extension: Let $\boldsymbol{\Phi}' = \frac{1}{\sqrt{m}}\boldsymbol{\Phi}$, $\boldsymbol{x} \in \Sigma_k$ with $\|\boldsymbol{x}\| = 1$ (WLOG), and $\boldsymbol{q} \in \mathcal{G}_\eta$ with $\|\boldsymbol{x} - \boldsymbol{q}\| \leqslant \eta$ & $\operatorname{supp} \boldsymbol{x} = \operatorname{supp} \boldsymbol{q}$.

  Setting $\eta = \epsilon/2$ with $0 < \epsilon < 1$: $\|\boldsymbol{\Phi}'\boldsymbol{x}\| \leqslant (1+\epsilon)\frac{(1+\eta)}{1-\eta} \leqslant (1+5\epsilon)$.

  Similarly: $\|\boldsymbol{\Phi}'\boldsymbol{x}\| \geqslant (1-5\epsilon)$.

  A rescaling of ϵ gives the RIP.

  $\square$

Exercise: JL involves RIP! (for sparse signals)

Global idea: tightly *sample* Σ_k, extends JL lemma by continuity!

- Apply JL lemma to the covering:

  Given $\epsilon > 0$, provided

  $$m \geqslant C\epsilon^{-2} \log |\mathcal{G}_\eta| \simeq C\epsilon^{-2} k \log(\tfrac{n}{\eta k}),$$

  if $\boldsymbol{\Phi} \in \mathbb{R}^{m \times n}$ with $\Phi_{ij} \sim_{\rm iid} \mathcal{N}(0,1)$, then, with probability exceeding $1 - C\exp(-c\epsilon^2 m)$, for all $\boldsymbol{q} \in \mathcal{G}_\eta$

  $$(1-\epsilon)\|\boldsymbol{q}\| \leqslant \sqrt{\tfrac{1}{m}}\|\boldsymbol{\Phi q}\| \leqslant (1+\epsilon)\|\boldsymbol{q}\|.$$

- Continuity extension: Let $\boldsymbol{\Phi}' = \frac{1}{\sqrt{m}}\boldsymbol{\Phi}$, $\boldsymbol{x} \in \Sigma_k$ with $\|\boldsymbol{x}\| = 1$ (WLOG), and $\boldsymbol{q} \in \mathcal{G}_\eta$ with $\|\boldsymbol{x} - \boldsymbol{q}\| \leqslant \eta$ & $\operatorname{supp} \boldsymbol{x} = \operatorname{supp} \boldsymbol{q}$.

  Setting $\eta = \epsilon/2$ with $0 < \epsilon < 1$: $\|\boldsymbol{\Phi}'\boldsymbol{x}\| \leqslant (1+\epsilon)\frac{(1+\eta)}{1-\eta} \leqslant (1+5\epsilon)$.

  Similarly: $\|\boldsymbol{\Phi}'\boldsymbol{x}\| \geqslant (1-5\epsilon)$.

  A rescaling of ϵ gives the RIP.

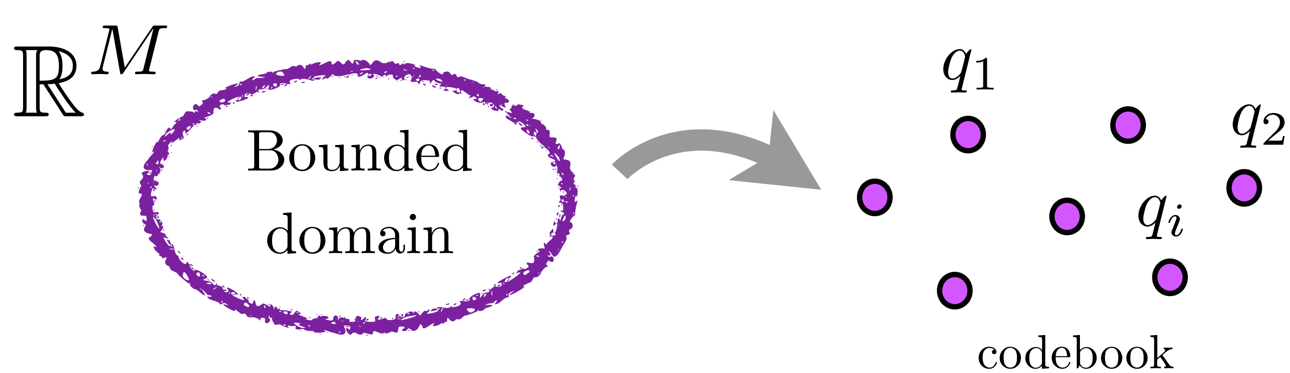
  $\square$

Note: RIP can involve JL !! **[Krahmer, Ward, '11]**

3. Quantized embeddings

- Quantized embeddings with regular, scalar quantizers
- The power of dithering
  - Diversion: Buffon's needle
  - Quantized RIP property and Consistency width
- Binary embeddings: Constructions and Properties
- Universal quantization and locally-preserved geometry

Quantization?

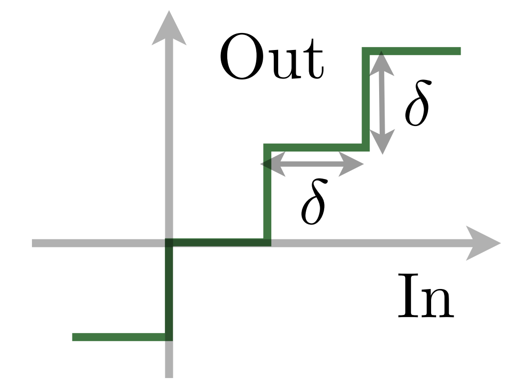


$$\mathcal{Q}[\boldsymbol{u}] \in \{\boldsymbol{q}_1, \boldsymbol{q}_2, \cdots\}$$

Quantization?

- Simple example: rounding/flooring

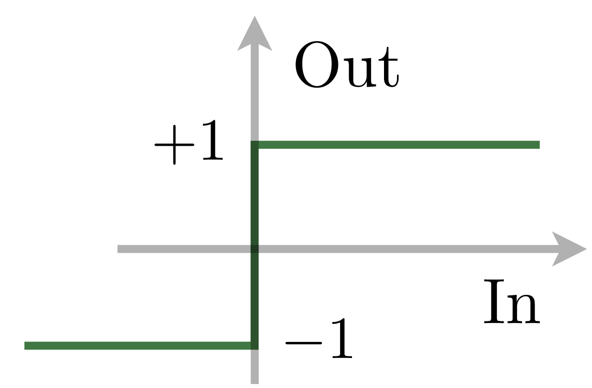
$$\mathcal{Q}[\lambda] = \delta\lfloor \tfrac{\lambda}{\delta} \rfloor \in \delta\mathbb{Z}$$



for some resolution $\delta > 0$ and $\mathcal{Q}(\boldsymbol{u}) = (\mathcal{Q}(u_1), \mathcal{Q}(u_2), \cdots)$.

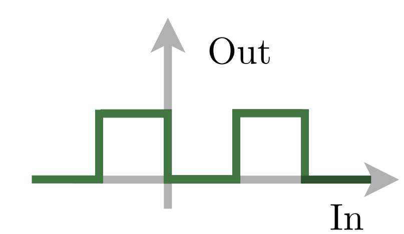
- Even simpler: 1-bit quantizer

$$\mathcal{Q}[\lambda] = \operatorname{sign}\lambda \in \pm 1$$



- Non-regular, e.g., square wave (or LSB)

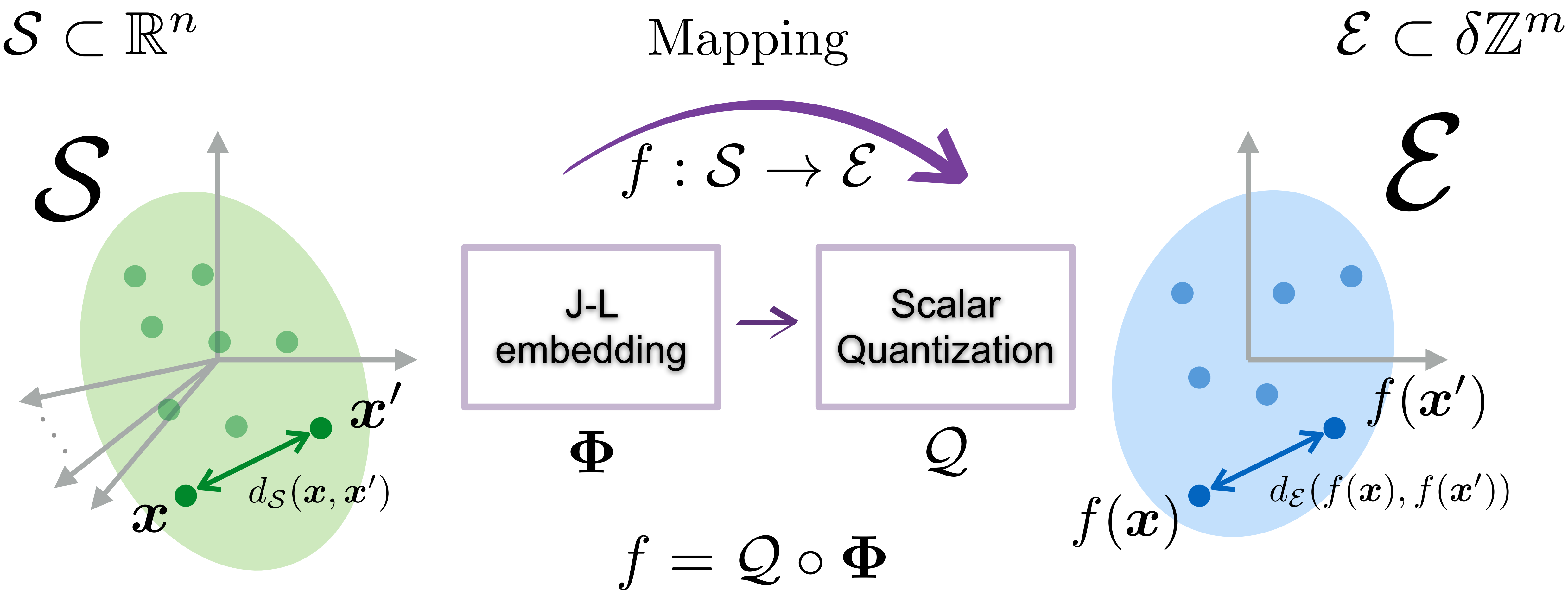
$$\mathcal{Q}[\lambda] := \delta(\lfloor \tfrac{\lambda}{\delta} \rfloor \bmod 2)$$



<u>Not covered here:</u> Non-uniform scalar quantizer, vector quantizer, $\Sigma\Delta$ quantizer, noise shaping, ...

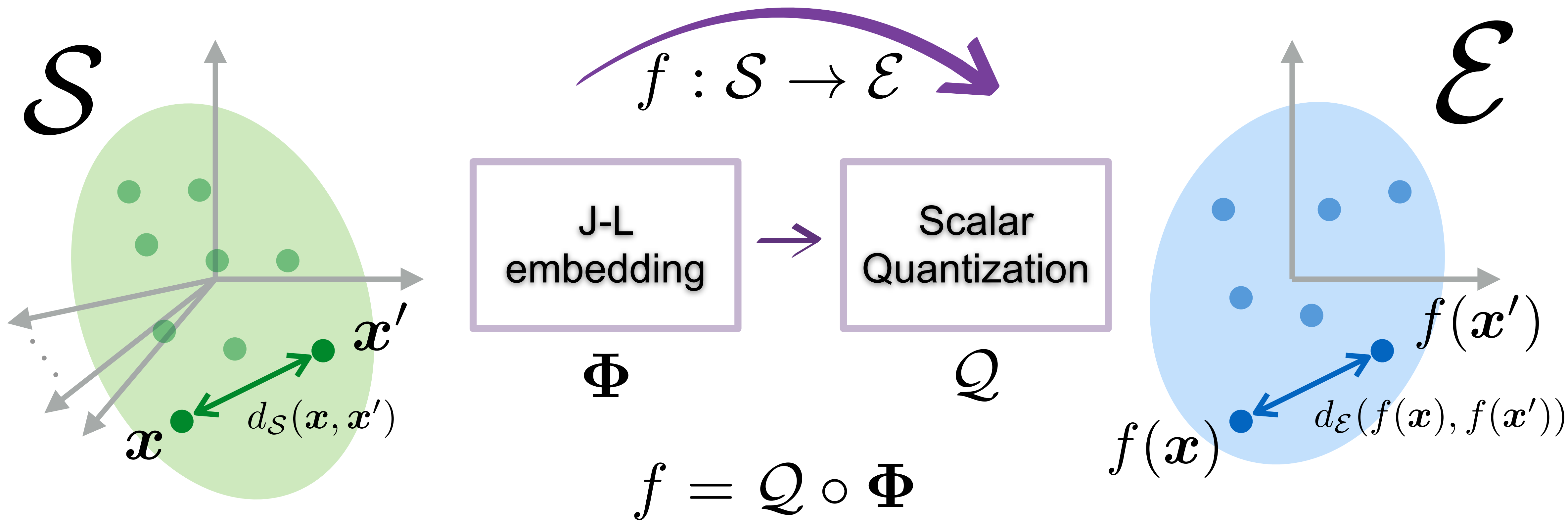
(see the works of, e.g., **[Gunturk, Lammers, Powell, Saab, Yilmaz, Goyal]**)

Naive quantized JL embedding



Naive quantized JL embedding

$\mathcal{S} \subset \mathbb{R}^n$ Mapping $\mathcal{E} \subset \delta\mathbb{Z}^m$



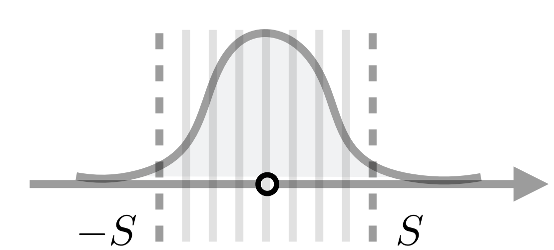
$$f = \mathcal{Q} \circ \boldsymbol{\Phi}$$

Let's use: (deterministic, always true fact)

$$|\mathcal{Q}(\lambda) - \mathcal{Q}(\lambda')| \lesseqgtr |\lambda - \lambda'| \pm (|\mathcal{Q}(\lambda) - \lambda| + |\mathcal{Q}(\lambda') - \lambda'|) \lesseqgtr |\lambda - \lambda'| \pm \delta, \quad \forall \lambda, \lambda' \in \mathbb{R}$$

Moreover, for B bits quantizer and dynamic range S:

$$\delta = \tfrac{2S}{2^B} \;\; (e.g., \; S = \|\boldsymbol{\Phi x}\|_\infty)$$



Naive quantized JL embedding [PB, Li, Rane]

For $|\mathcal{S}| = N$ points, f provides this quantized embedding in $\delta\mathbb{Z}^m$:

$$\forall \boldsymbol{x}, \boldsymbol{x}' \in \mathcal{S}$$

$$(1-\epsilon)\|\boldsymbol{x}-\boldsymbol{x}'\| - 2^{-B+1}S \leqslant \|f(\boldsymbol{x}) - f(\boldsymbol{x}')\| \leqslant (1+\epsilon)\|\boldsymbol{x}-\boldsymbol{x}'\| + 2^{-B+1}S,$$

Using only $m = O(\frac{\log N}{\epsilon^2})$ dimensions!

and B bits per dimension

(with appropriate normalizations & saturation levels)

Naive quantized JL embedding [PB, Li, Rane]

For $|\mathcal{S}| = N$ points, f provides this quantized embedding in $\delta\mathbb{Z}^m$:

$$\forall \boldsymbol{x}, \boldsymbol{x}' \in \mathcal{S}$$

$$(1-\epsilon)\|\boldsymbol{x}-\boldsymbol{x}'\| - 2^{-\frac{R}{m}+1}S \leqslant \|f(\boldsymbol{x}) - f(\boldsymbol{x}')\| \leqslant (1+\epsilon)\|\boldsymbol{x}-\boldsymbol{x}'\| + 2^{-\frac{R}{m}+1}S,$$

Using only $m = O(\frac{\log N}{\epsilon^2})$ dimensions!

and B bits per dimension

(with appropriate normalizations & saturation levels)

for a constant rate $R = mB$!

Naive quantized JL embedding [PB, Li, Rane]

For $|\mathcal{S}| = N$ points, f provides this quantized embedding in $\delta\mathbb{Z}^m$:

Larger B, less quantization distortion

$2^{-B+1}S$

$\forall \boldsymbol{x}, \boldsymbol{x}' \in \mathcal{S}$

$$(1-\epsilon)\|\boldsymbol{x}-\boldsymbol{x}'\| - 2^{-\frac{R}{m}+1}S \leqslant \|f(\boldsymbol{x}) - f(\boldsymbol{x}')\| \leqslant (1+\epsilon)\|\boldsymbol{x}-\boldsymbol{x}'\| + 2^{-\frac{R}{m}+1}S,$$

Larger m, less J-L type distortion

$\epsilon = O(1/\sqrt{m})$

Using only $m = O(\frac{\log N}{\epsilon^2})$ dimensions!

and B bits per dimension

(with appropriate normalizations & saturation levels)

for a constant rate $R = mB$!

Naive quantized JL embedding [PB, Li, Rane]

For $|\mathcal{S}| = N$ points, f provides this quantized embedding in $\delta\mathbb{Z}^m$:

Larger B, less quantization distortion

$$2^{-B+1}S$$

$$\forall \boldsymbol{x}, \boldsymbol{x}' \in \mathcal{S}$$

$$(1-\epsilon)\|\boldsymbol{x}-\boldsymbol{x}'\| - 2^{-\frac{R}{m}+1}S$$
$$\leqslant \|f(\boldsymbol{x}) - f(\boldsymbol{x}')\|$$
$$\leqslant (1+\epsilon)\|\boldsymbol{x}-\boldsymbol{x}'\| + 2^{-\frac{R}{m}+1}S,$$

Larger m, less J-L type distortion

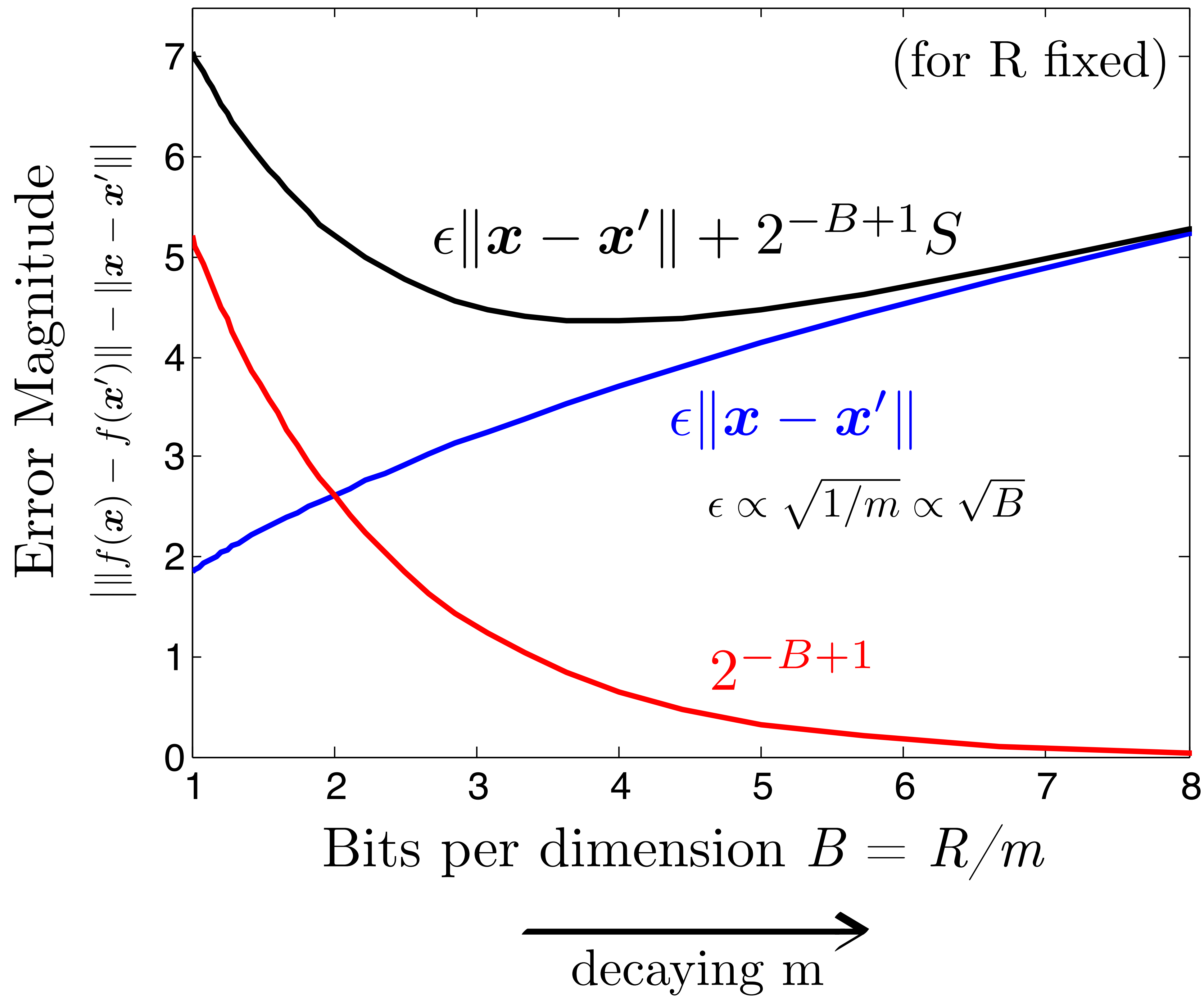
$$\epsilon = O(1/\sqrt{m})$$

Given total rate $R = mB$, how to assign B and m?

More m or more B?

Design tradeoff: Number of projections vs. bits per projection

Exploring the Design Trade-off



Limitation

- Additive distortion not decaying with m
- But distortion is required!

Counterexample:

Take $\boldsymbol{\Phi} \in \{\pm 1\}^{m \times n}$ (an admissible JL embedding)

$$\boldsymbol{x} = \boldsymbol{e}_1 = (1, 0, \cdots, 0)^\top \in \mathbb{R}^n$$
$$\boldsymbol{x}' = \boldsymbol{e}_1 + \lambda \boldsymbol{e}_2 \text{ with } 0 < |\lambda| < 1.$$

We have: $\|\boldsymbol{x} - \boldsymbol{x}'\| = \lambda > 0$

However, $\boldsymbol{\Phi x} \equiv (1^{\text{st}} \text{ col. of } \boldsymbol{\Phi}) \in \{\pm 1\}^m$

$$\boldsymbol{\Phi x}' \equiv (1^{\text{st}} \text{ col. of } \boldsymbol{\Phi} + \lambda \times 2^{\text{nd}} \text{ col. of } \boldsymbol{\Phi}) \in \{\pm 1 \pm \lambda\}^m$$

Therefore: For the rounding operator $\mathcal{Q}(\cdot) := \lfloor \cdot + 1/2 \rfloor$ (if $|\lambda| < 1/2$), or with $\mathcal{Q}(\cdot) := \operatorname{sign}(\cdot)$, **[Plan, Vershynin]**

$$\mathcal{Q}(\boldsymbol{\Phi x}) = \boldsymbol{\Phi x} = \mathcal{Q}(\boldsymbol{\Phi x}') \;\Leftrightarrow\; \|f(\boldsymbol{x}) - f(\boldsymbol{x}')\| = 0$$

Limitation

- Additive distortion not decaying with m
- But distortion is required!

Counterexample:

Take $\boldsymbol{\Phi} \in \{\pm 1\}^{m \times n}$ (an admissible JL embedding)

$$\boldsymbol{x} = \boldsymbol{e}_1 = (1, 0, \cdots, 0)^\top \in \mathbb{R}^n$$

$$\boldsymbol{x}' = \boldsymbol{e}_1 + \lambda \boldsymbol{e}_2 \text{ with } 0 < |\lambda| < 1.$$

We have: $\|\boldsymbol{x} - \boldsymbol{x}'\| = \lambda > 0$

$\|f(\boldsymbol{x}) - f(\boldsymbol{x}')\| \not\propto \|\boldsymbol{x} - \boldsymbol{x}'\|$

However, $\boldsymbol{\Phi x} \equiv (1^{\text{st}} \text{ col. of } \boldsymbol{\Phi}) \in \{\pm 1\}^m$

$$\boldsymbol{\Phi x}' \equiv (1^{\text{st}} \text{ col. of } \boldsymbol{\Phi} + \lambda \times 2^{\text{nd}} \text{ col. of } \boldsymbol{\Phi}) \in \{\pm 1 \pm \lambda\}^m$$

Therefore: For the rounding operator $\mathcal{Q}(\cdot) := \lfloor \cdot + 1/2 \rfloor$ (if $|\lambda| < 1/2$), or with $\mathcal{Q}(\cdot) := \text{sign}\,(\cdot)$, **[Plan, Vershynin]**

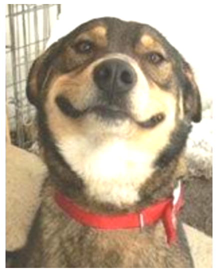
$$\mathcal{Q}(\boldsymbol{\Phi x}) = \boldsymbol{\Phi x} = \mathcal{Q}(\boldsymbol{\Phi x}') \;\Leftrightarrow\; \|f(\boldsymbol{x}) - f(\boldsymbol{x}')\| = 0$$

The power of dithering (an old trick revisited*)

- Inject a pre-quantization, uniform "noise":

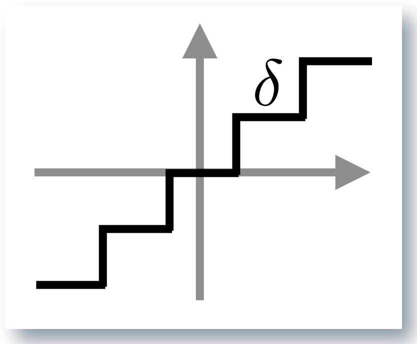
  i.e., a *dithering* $\boldsymbol{\xi} \in \mathbb{R}^m$ with $\xi_j \sim_{\text{iid}} \mathcal{U}([0, \delta])$

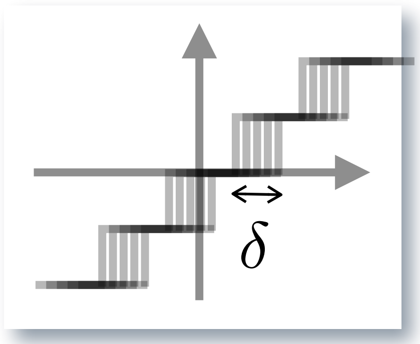
The good boy!



(QDRM) $$\mathsf{A}(\boldsymbol{x}) := \mathcal{Q}(\boldsymbol{\Phi x} + \boldsymbol{\xi})$$

$\mathcal{Q}(\cdot)$





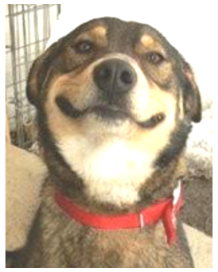
$\mathcal{Q}(\cdot + \xi)$

*: See, e.g., Gray & Neuhoff in Q theory, and P. Boufounos, A. Powell, ... in CS

The power of dithering (an old trick revisited)

- Inject a pre-quantization, uniform "noise":

  i.e., a *dithering* $\boldsymbol{\xi} \in \mathbb{R}^m$ with $\xi_j \sim_{\rm iid} \mathcal{U}([0, \delta])$ → **The good boy!**

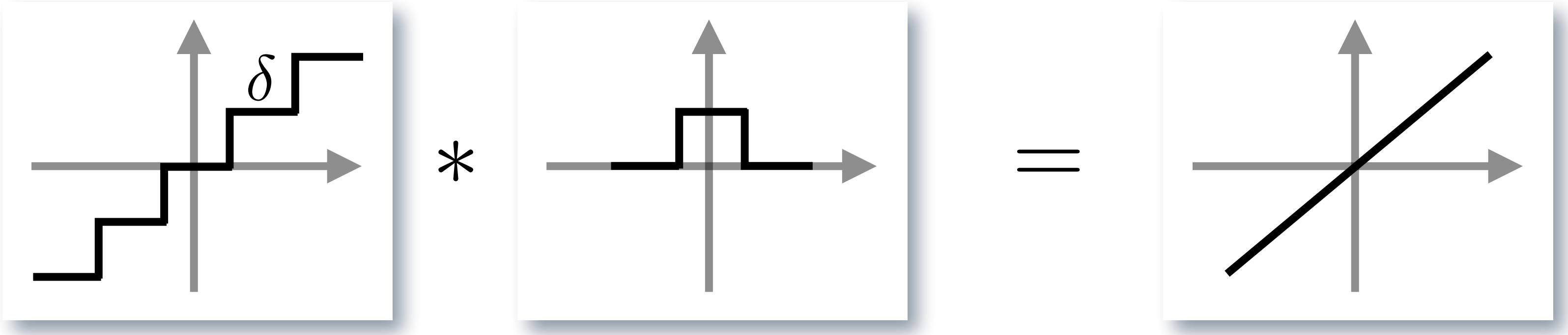


(QDRM) $$\mathsf{A}(\boldsymbol{x}) := \mathcal{Q}(\boldsymbol{\Phi}\boldsymbol{x} + \boldsymbol{\xi})$$

- Motivation? $\mathbb{E}_{\boldsymbol{\xi}} \mathcal{Q}(\boldsymbol{u} + \boldsymbol{\xi}) = \boldsymbol{u}$

  $\Rightarrow \mathsf{A}(\boldsymbol{x}) \approx \boldsymbol{\Phi}\boldsymbol{x}$ if M large

$\mathcal{Q}(\cdot)$

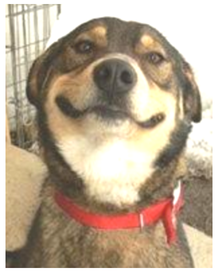


The power of dithering (an old trick revisited)

- Inject a pre-quantization, uniform "noise":

  i.e., a *dithering* $\boldsymbol{\xi} \in \mathbb{R}^m$ with $\xi_j \sim_{\rm iid} \mathcal{U}([0, \delta])$

The good boy!



$$\text{(QDRM)} \qquad \boxed{\mathsf{A}(\boldsymbol{x}) := \mathcal{Q}(\boldsymbol{\Phi}\boldsymbol{x} + \boldsymbol{\xi})}$$

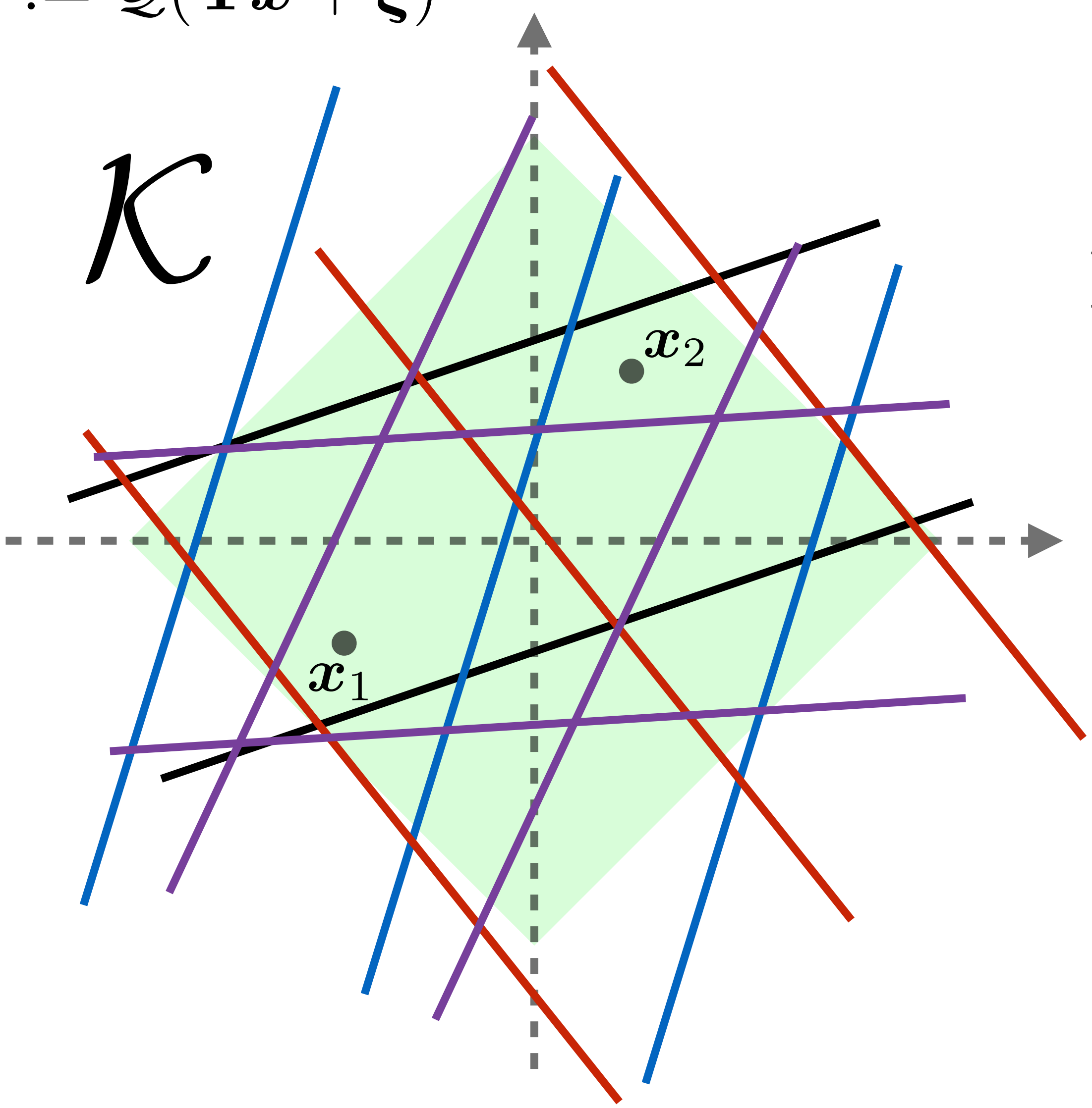
- Motivation? $\mathbb{E}_{\boldsymbol{\xi}} \mathcal{Q}(\boldsymbol{u} + \boldsymbol{\xi}) = \boldsymbol{u}$

  $\Rightarrow \mathsf{A}(\boldsymbol{x}) \approx \boldsymbol{\Phi}\boldsymbol{x}$ if M large

- Possibility to define quantized dimensionality reduction/embedding!

Quantizing the RIP (approximate consistency)

$\mathsf{A}(\boldsymbol{x}) := \mathcal{Q}(\boldsymbol{\Phi x} + \boldsymbol{\xi})$



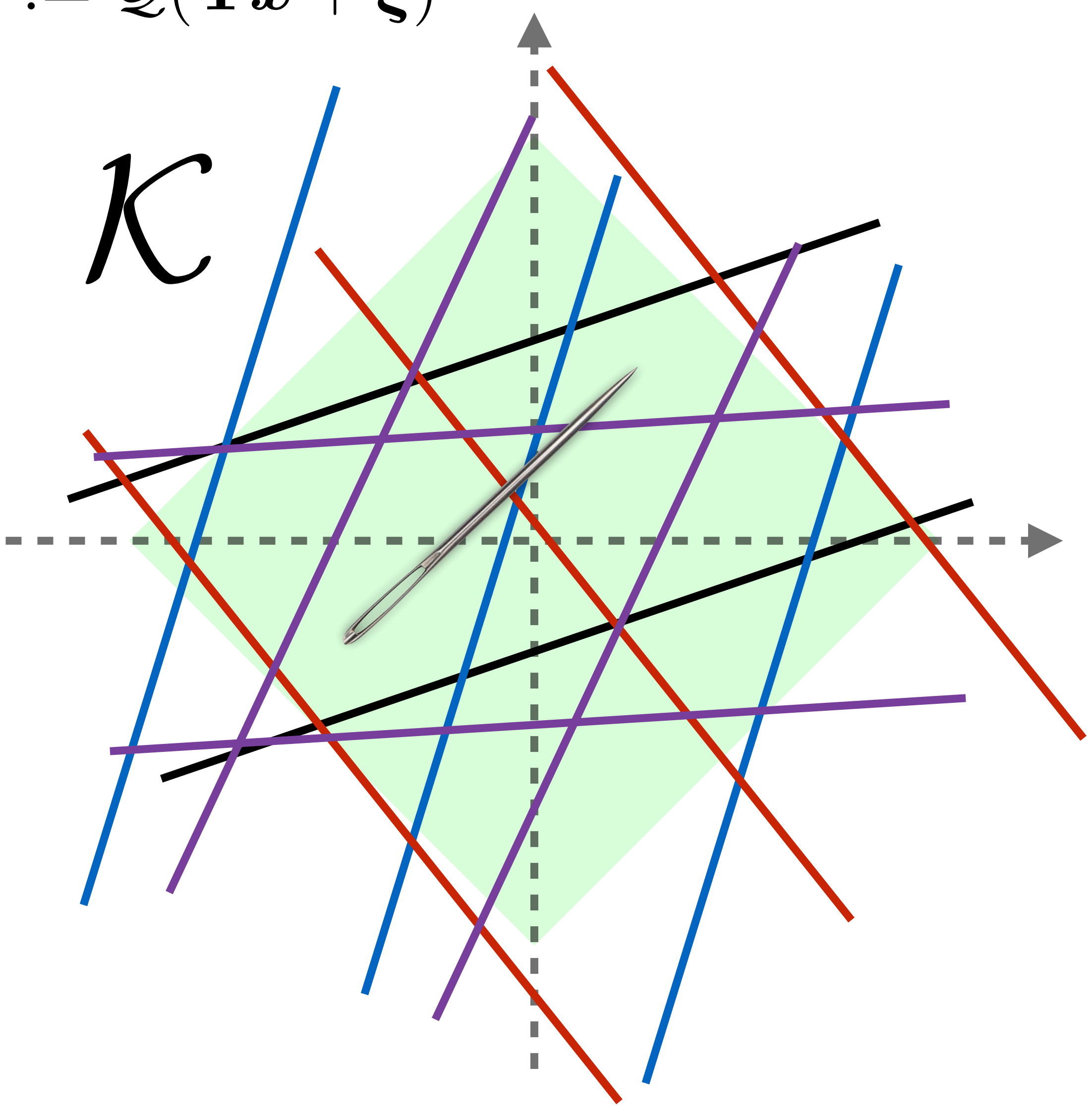
Distance between the two points

$\approx$

Number of quantization frontiers between the two points?

Quantizing the RIP (approximate consistency)

$\mathsf{A}(\boldsymbol{x}) := \mathcal{Q}(\boldsymbol{\Phi x} + \boldsymbol{\xi})$



$\equiv$

(thanks to the dithering)

Buffon's **needle** problem

http://www.buffon.cnrs.fr

(In 1733)

Length of  $= \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|$

(short diversion)

http://www.buffon.cnrs.fr

Buffon's needle problem

[Buffon's problem 1733, Buffon's solution 1777]

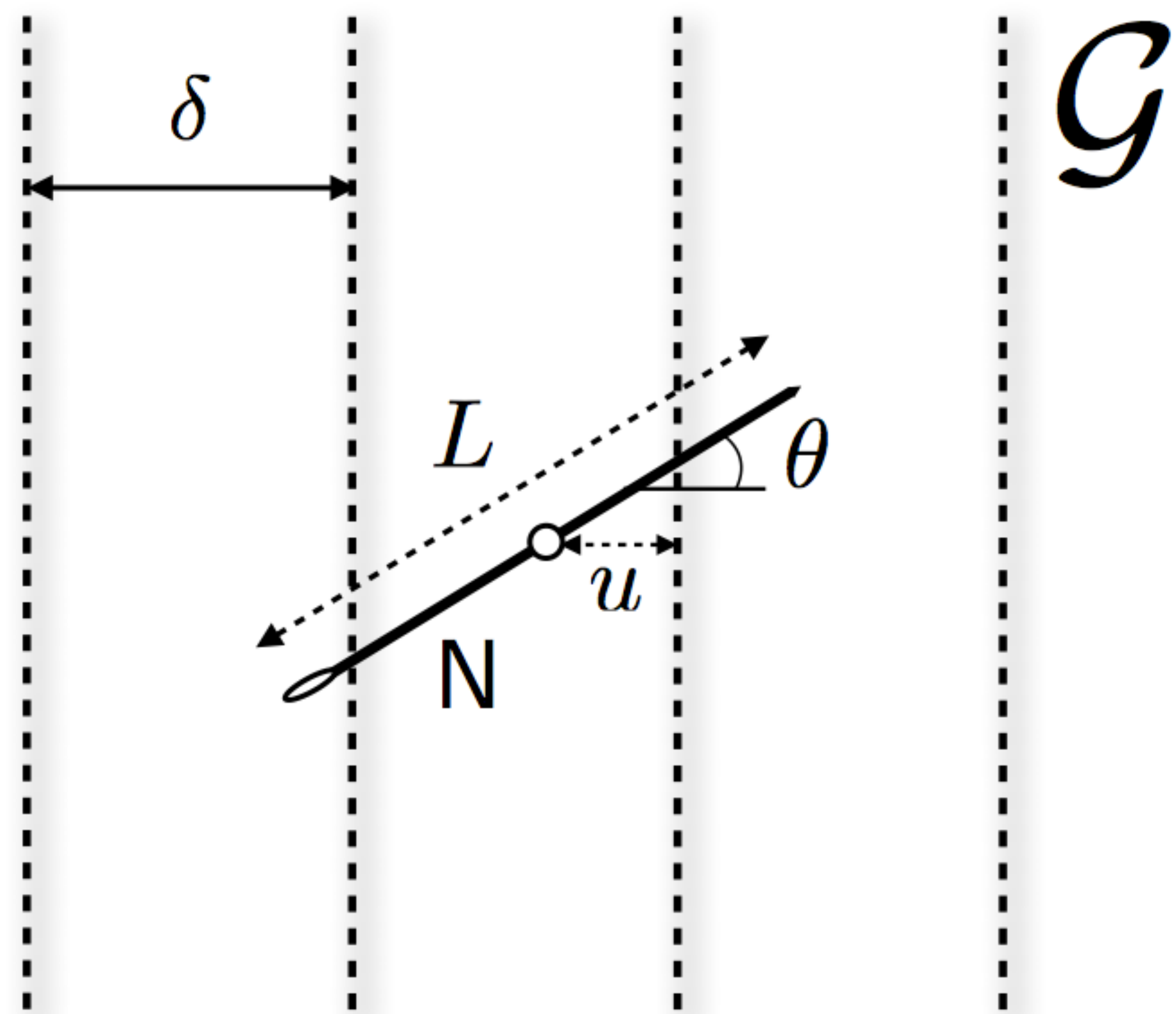
> *"I suppose that in a room where the floor is simply divided by parallel joints one throws a stick ("needle") in the air,*
>
> *and that one of the players bets that the stick will not cross any of the parallels on the floor,*
>
> *and that the other in contrast bets that the stick will cross some of these parallels;*
>
> ***one asks for the chances of these two players."***

(short diversion)

Buffon's needle problem

http://www.buffon.cnrs.fr

(Courtesy of E. Kowalski's blog)



$$\mathbb{P}[\,\mathsf{N}(u,\theta) \cap \mathcal{G} \neq \emptyset\,] \\ = ?$$

with $u \sim \mathcal{U}([0,\delta])$ and $\theta \sim \mathcal{U}([0,2\pi])$

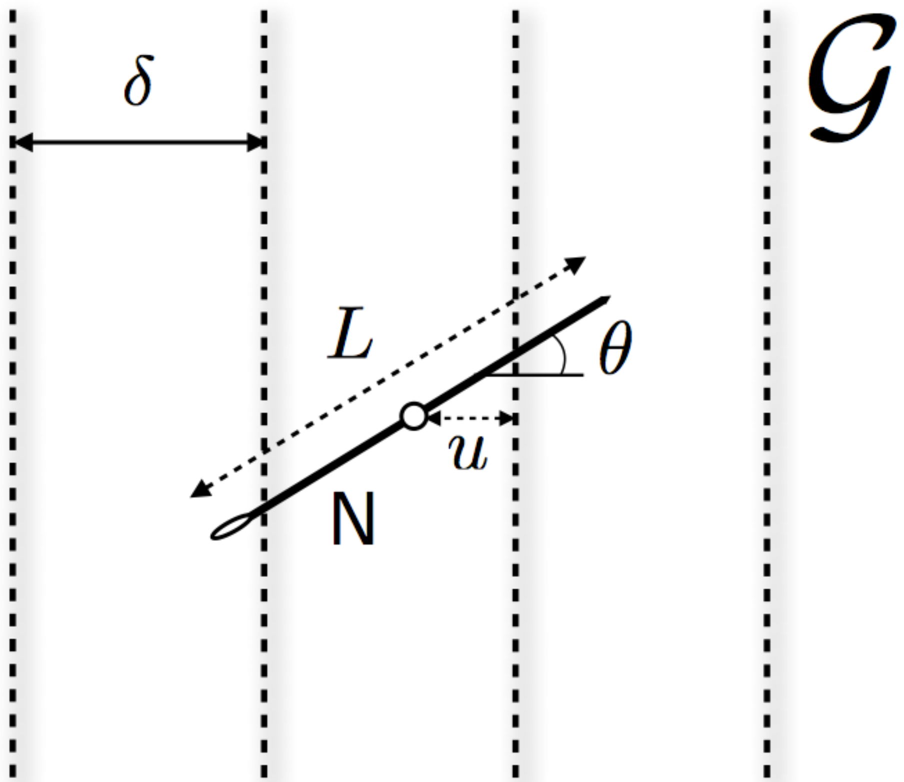
(short diversion)

Buffon's needle problem

http://www.buffon.cnrs.fr

Fact 1: if $L < \delta$, $\mathbb{P} = \frac{2}{\pi\delta} L$

(small integral to solve)



with $u \sim \mathcal{U}([0,\delta])$ and $\theta \sim \mathcal{U}([0,2\pi])$

(short diversion)

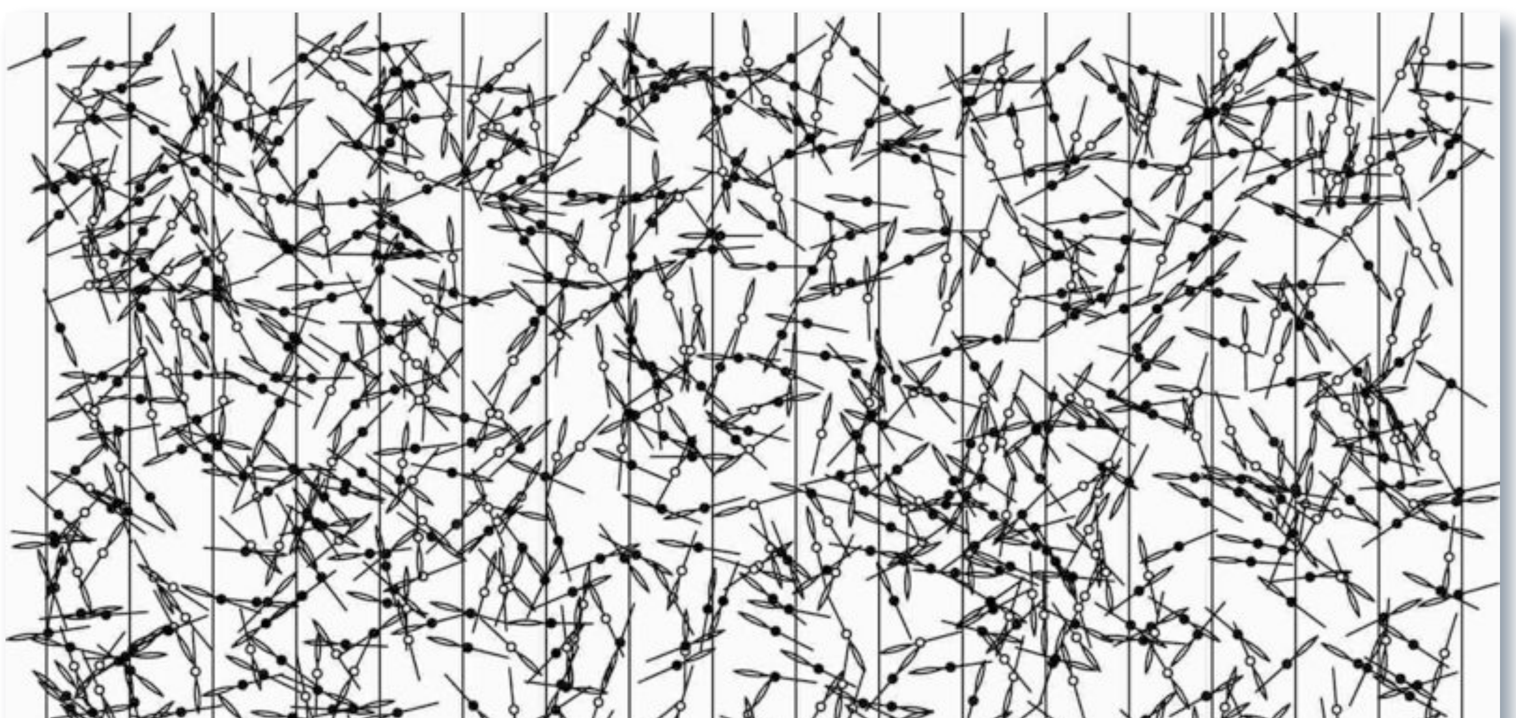
Buffon's needle problem

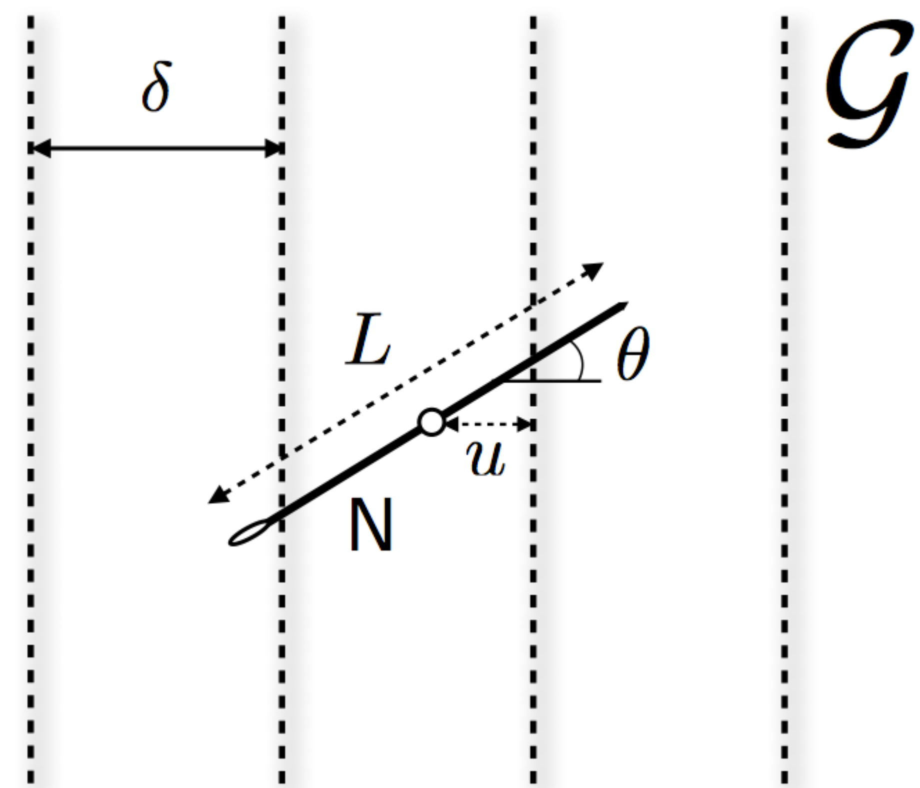
http://www.buffon.cnrs.fr

Fact 1: if $L < \delta$, $\mathbb{P} = \frac{2}{\pi\delta} L$

(small integral
to solve)

Has been used for estimating π !
(first "Monte Carlo" method)





with $u \sim \mathcal{U}([0, \delta])$ and $\theta \sim \mathcal{U}([0, 2\pi])$

(short diversion)

Buffon's needle problem

http://www.buffon.cnrs.fr

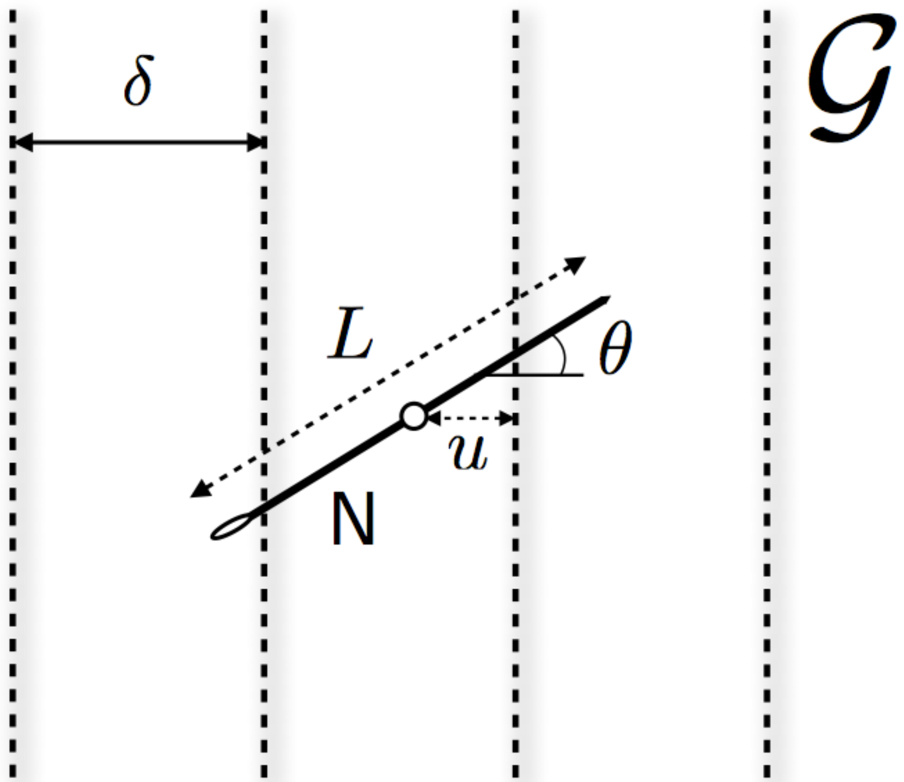
Fact 1: if $L < \delta$, $\mathbb{P} = \frac{2}{\pi\delta} L$

Fact 2: if $L \geqslant \delta$, $\mathbb{P} \neq \frac{2}{\pi\delta} L$ but

$$\mathbb{E}X = \frac{2}{\pi\delta} L,$$

with $X = \#\{\,\mathsf{N}(u,\theta) \cap \mathcal{G}\,\}$.

Proof: cut N in parts smaller than δ and sum expectations!



with $u \sim \mathcal{U}([0,\delta])$ and $\theta \sim \mathcal{U}([0,2\pi])$

(short diversion)

Buffon's needle problem

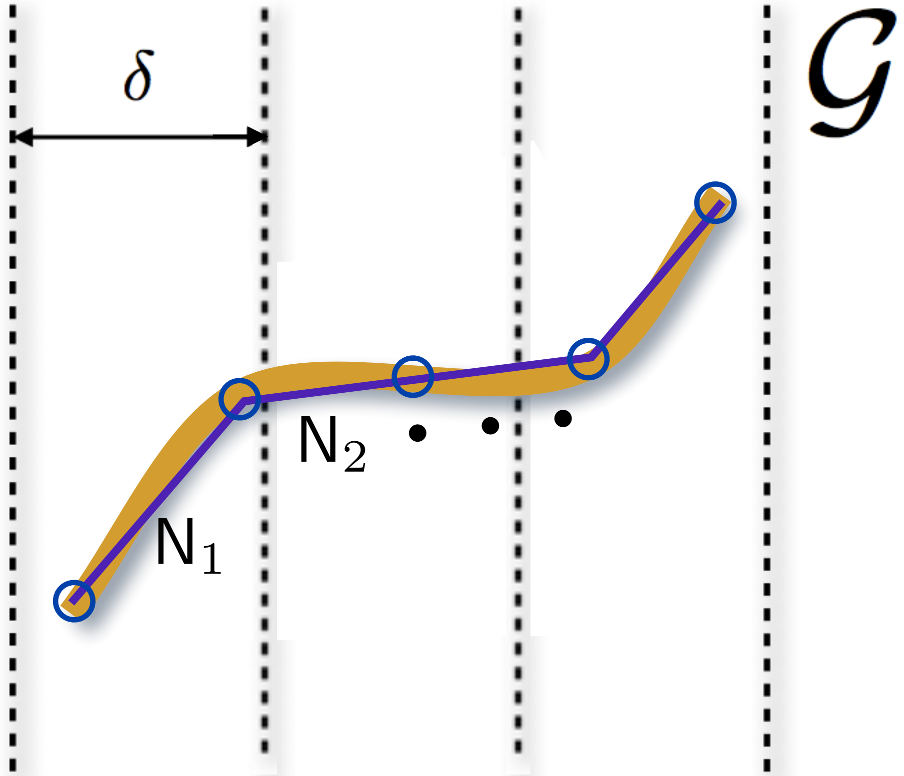
http://www.buffon.cnrs.fr

Fact 1: if $L < \delta$, $\mathbb{P} = \frac{2}{\pi\delta} L$

Fact 2: if $L \geqslant \delta$, $\mathbb{P} \neq \frac{2}{\pi\delta} L$ but

$$\mathbb{E}X = \frac{2}{\pi\delta} L,$$

with $X = \#\{\, \mathsf{N}(u,\theta) \cap \mathcal{G} \,\}$.



with $u \sim \mathcal{U}([0,\delta])$ and $\theta \sim \mathcal{U}([0,2\pi])$

Fact 3: It works for "noodles" (smooth curves)!

For information only.

(short diversion)

Buffon's needle problem

http://www.buffon.cnrs.fr

Fact 1: if $L < \delta$, $\mathbb{P} = \frac{2}{\pi\delta} L$

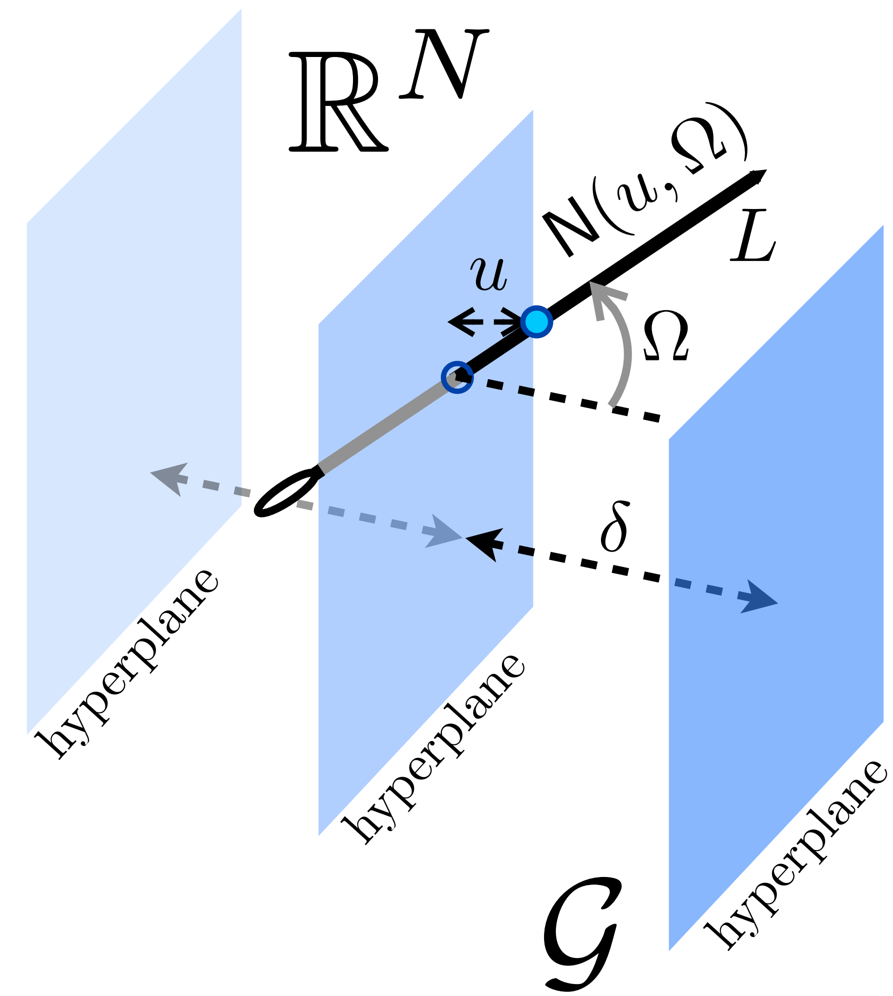
Fact 2: if $L \geqslant \delta$, $\mathbb{P} \neq \frac{2}{\pi\delta} L$ but

$$\mathbb{E}X = \frac{2}{\pi\delta} L,$$

with $X = \#\{\, \mathsf{N}(u,\theta) \cap \mathcal{G} \,\}$.

Fact 3: It works for "noodles" (smooth curves)!

Fact 4: It extends to N-dim.



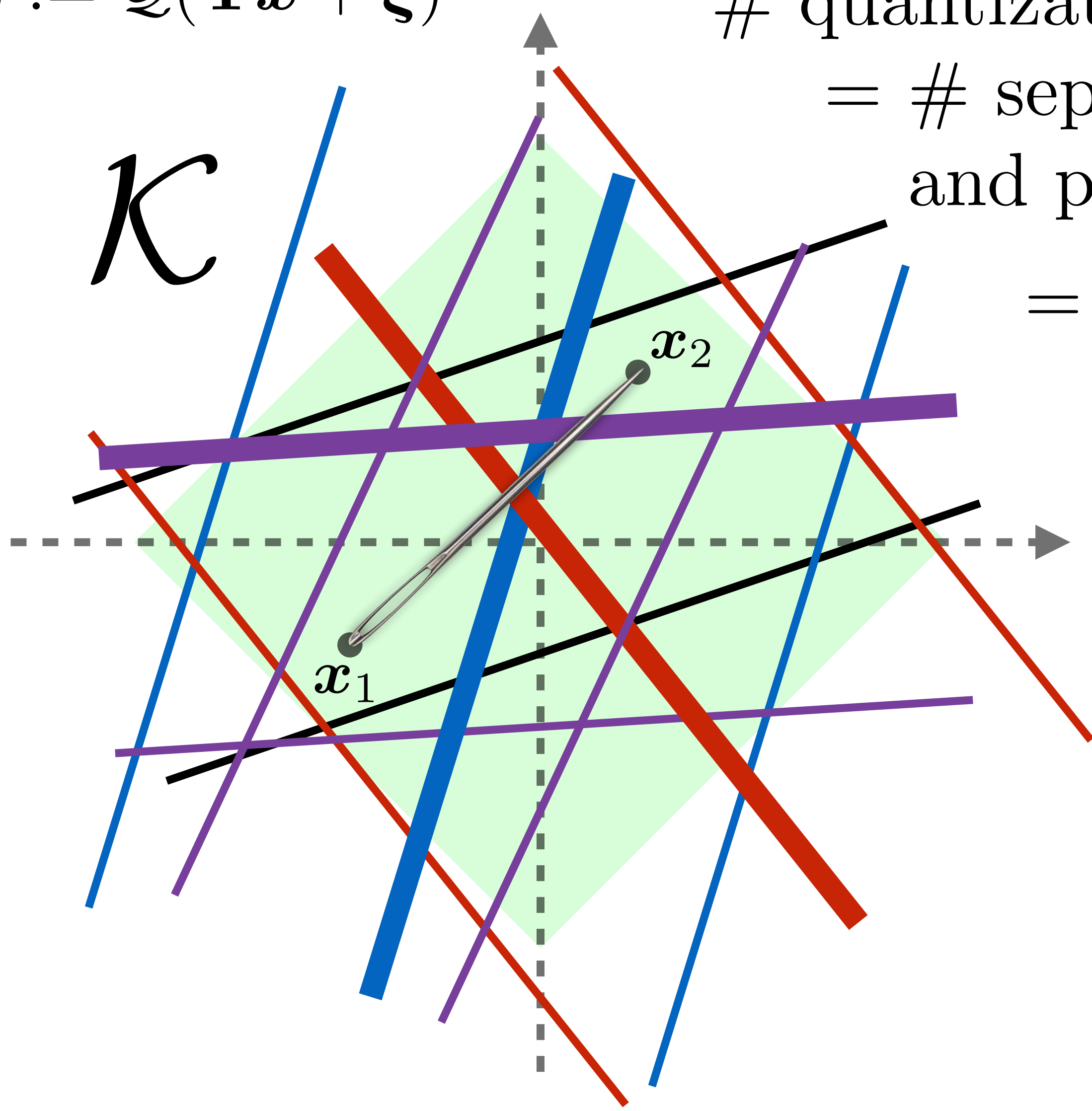
(short diversion)

(but not necessary)

(end of the short diversion)

Quantizing the RIP (approximate consistency)

$\mathsf{A}(\boldsymbol{x}) := \mathcal{Q}(\boldsymbol{\Phi x} + \boldsymbol{\xi})$

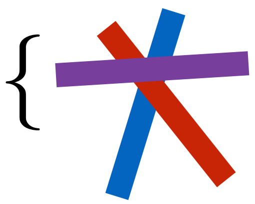


quantization frontiers separating $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$

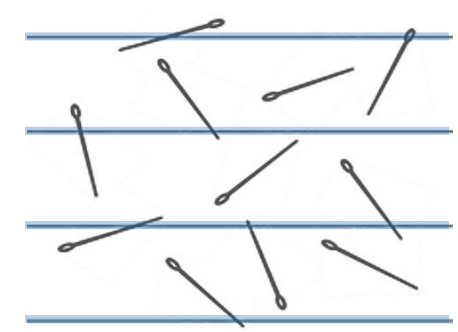
= # separating random hyperplanes oriented and positioned according to $(\boldsymbol{\Phi}, \boldsymbol{\xi})$

$$= \frac{1}{M\delta}\|\mathsf{A}(\boldsymbol{x}_1) - \mathsf{A}(\boldsymbol{x}_2)\|_1 \overset{??}{\approx} \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|$$

$|||$

#{  }

Buffon's needle problem



$\mathbb{E}$(intersections) $\propto$ length

http://www.buffon.cnrs.fr

(In 1733)

Hope: dithering sufficiently smooths discontinuities to allow for RIP matrices.

Quantizing the RIP (approximate consistency)

Let $\mathcal{K} \subset \mathbb{R}^N$ be a structured set ($e.g.$, sparse signals, low-rank matrices).

Let $\boldsymbol{\Phi}$ be a (ℓ_1, ℓ_2)-RIP$(\epsilon, \mathcal{K} - \mathcal{K})$ matrix, $i.e.$,

$$(1-\epsilon)\|\boldsymbol{x}\|^2 \leqslant \tfrac{c_{\boldsymbol{\Phi}}}{m}\|\boldsymbol{\Phi}\boldsymbol{x}\|_1^2 \leqslant (1+\epsilon)\|\boldsymbol{x}\|^2, \forall \boldsymbol{x} \in \mathcal{K} - \mathcal{K},$$

($e.g.$, Gaussian random matrix, circulant Gaussian random matrix for $\mathcal{K} = \Sigma_K$)

[Dirksen, Jung, Rauhut, 17]

Quantizing the RIP (approximate consistency)

Let $\mathcal{K} \subset \mathbb{R}^N$ be a structured set (*e.g.*, sparse signals, low-rank matrices).

Let $\boldsymbol{\Phi}$ be a (ℓ_1, ℓ_2)-RIP$(\epsilon, \mathcal{K} - \mathcal{K})$ matrix, *i.e.*,

$$(1-\epsilon)\|\boldsymbol{x}\|^2 \leqslant \tfrac{c_{\Phi}}{m}\|\boldsymbol{\Phi x}\|_1^2 \leqslant (1+\epsilon)\|\boldsymbol{x}\|^2, \forall \boldsymbol{x} \in \mathcal{K} - \mathcal{K},$$

(*e.g.*, Gaussian random matrix, circulant Gaussian random matrix for $\mathcal{K} = \Sigma_K$)

[Dirksen, Jung, Rauhut, 17]

Provided that $M \gtrsim \epsilon^{-2}\, C_{\mathcal{K}} \log(1 + \frac{1}{\delta\epsilon})$, (with $C_{\mathcal{K}} > 0$ an upper bound on $w(\mathcal{K})^2$)
with probability exceeding $1 - C\exp(-\epsilon^2 m)$,

$$(1-\epsilon)\|\boldsymbol{x}_1 - \boldsymbol{x}_2\| - c'\epsilon\delta \leqslant \tfrac{1}{m}\|\mathsf{A}(\boldsymbol{x}_1) - \mathsf{A}(\boldsymbol{x}_2)\|_1 \leqslant (1+\epsilon)\|\boldsymbol{x}_1 - \boldsymbol{x}_2\| + c'\epsilon\delta,$$

for all $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathcal{K} \cap \mathbb{B}^N$.

($\exists$ other variants with ℓ_2/ℓ_2 and standard RIP)

[LJ, Cambareri, 17]

Quantizing the RIP (approximate consistency)

Let $\mathcal{K} \subset \mathbb{R}^N$ be a structured set (*e.g.*, sparse signals, low-rank matrices).

Let $\boldsymbol{\Phi}$ be a (ℓ_1, ℓ_2)-RIP$(\epsilon, \mathcal{K} - \mathcal{K})$ matrix, *i.e.*,

$$(1-\epsilon)\|\boldsymbol{x}\|^2 \leqslant \tfrac{c_{\Phi}}{m}\|\boldsymbol{\Phi x}\|_1^2 \leqslant (1+\epsilon)\|\boldsymbol{x}\|^2, \forall \boldsymbol{x} \in \mathcal{K} - \mathcal{K},$$

(*e.g.*, Gaussian random matrix, circulant Gaussian random matrix for $\mathcal{K} = \Sigma_K$)

[Dirksen, Jung, Rauhut, 17]

Provided that $M \gtrsim \epsilon^{-2} C_{\mathcal{K}} \log(1 + \frac{1}{\delta\epsilon})$, (with $C_{\mathcal{K}} > 0$ an upper bound on $w(\mathcal{K})^2$)
with probability exceeding $1 - C\exp(-\epsilon^2 m)$,

$$(1-\epsilon)\|\boldsymbol{x}_1 - \boldsymbol{x}_2\| - c'\epsilon\delta \leqslant \tfrac{1}{m}\|\mathsf{A}(\boldsymbol{x}_1) - \mathsf{A}(\boldsymbol{x}_2)\|_1 \leqslant (1+\epsilon)\|\boldsymbol{x}_1 - \boldsymbol{x}_2\| + c'\epsilon\delta,$$

for all $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathcal{K} \cap \mathbb{B}^N$.

($\exists$ other variants with ℓ_2/ℓ_2 and standard RIP)

Decaying distortion:
$\epsilon = O(1/\sqrt{m})$

[LJ, Cambareri, 17]

Quantizing the RIP (approximate consistency)

Let $\mathcal{K} \subset \mathbb{R}^N$ be a structured set (*e.g.*, sparse signals, low-rank matrices).

Let $\boldsymbol{\Phi}$ be a (ℓ_1, ℓ_2)-RIP$(\epsilon, \mathcal{K} - \mathcal{K})$ matrix, *i.e.*,

$$(1-\epsilon)\|\boldsymbol{x}\|^2 \leqslant \tfrac{c_\Phi}{m}\|\boldsymbol{\Phi x}\|_1^2 \leqslant (1+\epsilon)\|\boldsymbol{x}\|^2, \forall \boldsymbol{x} \in \mathcal{K} - \mathcal{K},$$

(*e.g.*, Gaussian random matrix, circulant Gaussian random matrix for $\mathcal{K} = \Sigma_K$)

[Dirksen, Jung, Rauhut, 17]

Provided that $M \gtrsim \epsilon^{-2} C_{\mathcal{K}} \log(1 + \frac{1}{\delta\epsilon})$, (with $C_{\mathcal{K}} > 0$ an upper bound on $w(\mathcal{K})^2$)
with probability exceeding $1 - C\exp(-\epsilon^2 m)$,

$$(1-\epsilon)\|\boldsymbol{x}_1 - \boldsymbol{x}_2\| - c'\epsilon\delta \leqslant \tfrac{1}{m}\|\mathsf{A}(\boldsymbol{x}_1) - \mathsf{A}(\boldsymbol{x}_2)\|_1 \leqslant (1+\epsilon)\|\boldsymbol{x}_1 - \boldsymbol{x}_2\| + c'\epsilon\delta,$$

for all $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathcal{K} \cap \mathbb{B}^N$.

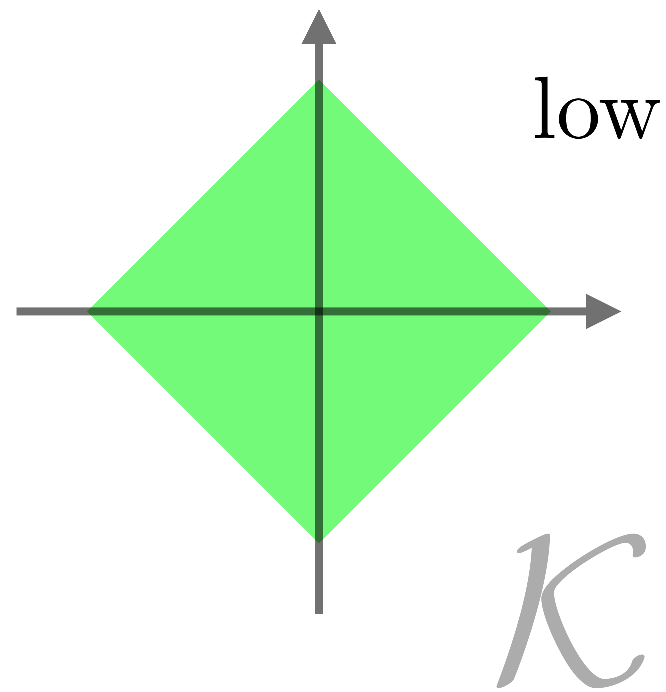
($\exists$ other variants with ℓ_2/ℓ_2 and standard RIP)

Dimensionality reduction!

Classification?

[LJ, Cambareri, 17]

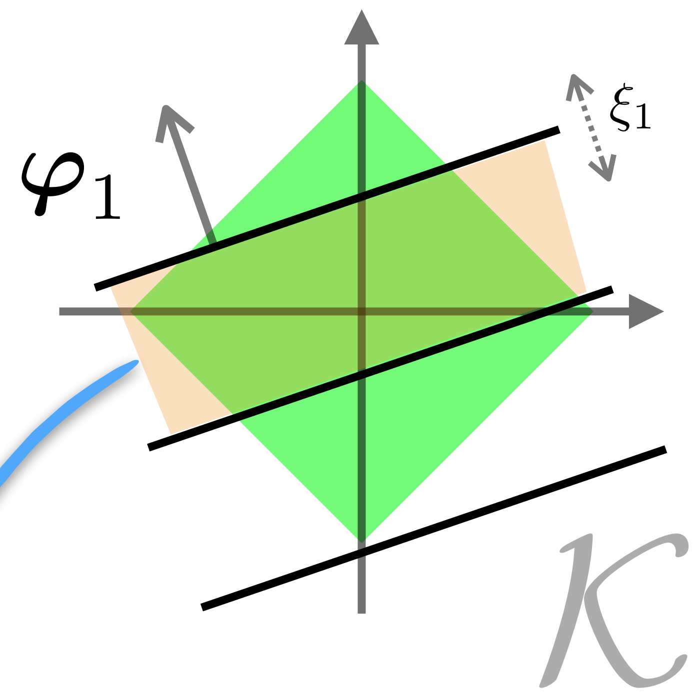
Control of the "consistency width"



low complexity set $\mathcal{K}$

(e.g., sparse signals,
low-rank matrix,
compressible signals, ...)

Control of the "consistency width"

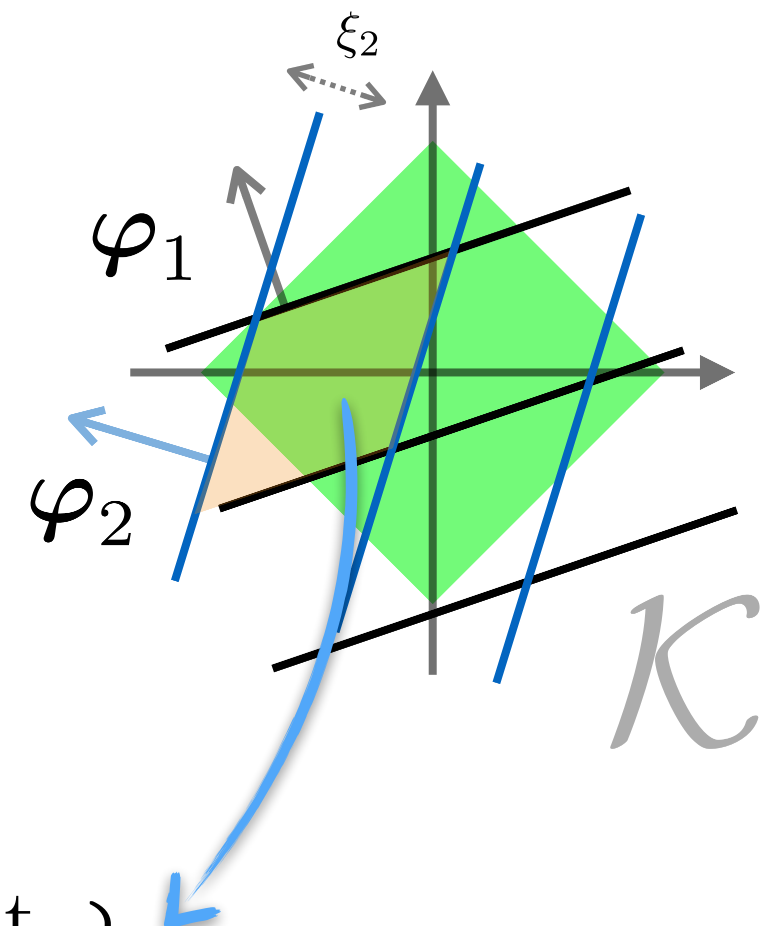


$$\boldsymbol{\Phi} = \begin{pmatrix} \boldsymbol{\varphi}_1^T \\ \vdots \\ \boldsymbol{\varphi}_M^T \end{pmatrix}$$

Signals $\boldsymbol{u} \in \mathcal{K}$ s.t.

$$\underbrace{\mathcal{Q}(\boldsymbol{\varphi}_1^\top \boldsymbol{u} + \xi_1)}_{\delta\lfloor(\boldsymbol{\varphi}_1^\top \boldsymbol{u} + \xi_1)/\delta\rfloor} = \text{cst.}$$

Control of the "consistency width"

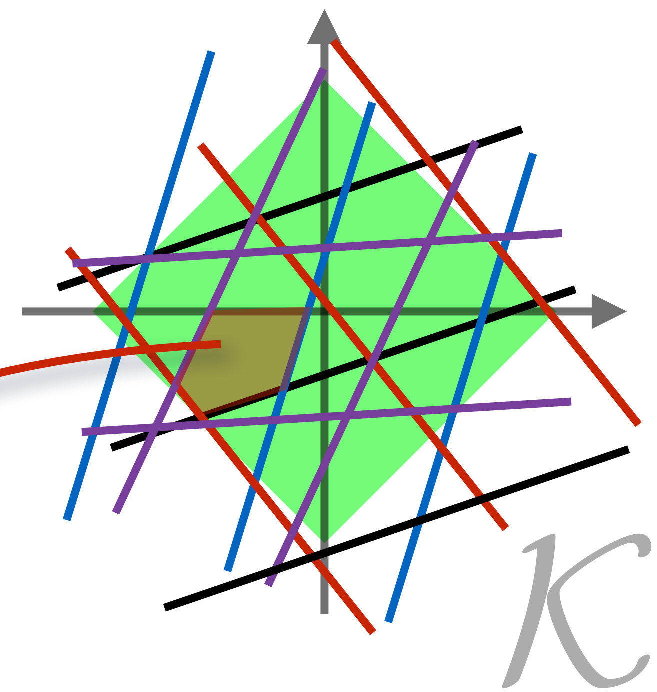


$$\boldsymbol{\Phi} = \begin{pmatrix} \boldsymbol{\varphi}_1^T \\ \vdots \\ \boldsymbol{\varphi}_M^T \end{pmatrix}$$

Signals $\boldsymbol{u} \in \mathcal{K}$ s.t.

$$\left.\begin{aligned} \mathcal{Q}(\boldsymbol{\varphi}_1^\top \boldsymbol{u} + \xi_1) &= \text{cst.} \\ \mathcal{Q}(\boldsymbol{\varphi}_2^\top \boldsymbol{u} + \xi_2) &= \text{cst.} \end{aligned}\right\}$$

Control of the "consistency width"



$$\boldsymbol{\Phi} = \begin{pmatrix} \boldsymbol{\varphi}_1^T \\ \vdots \\ \boldsymbol{\varphi}_M^T \end{pmatrix}$$

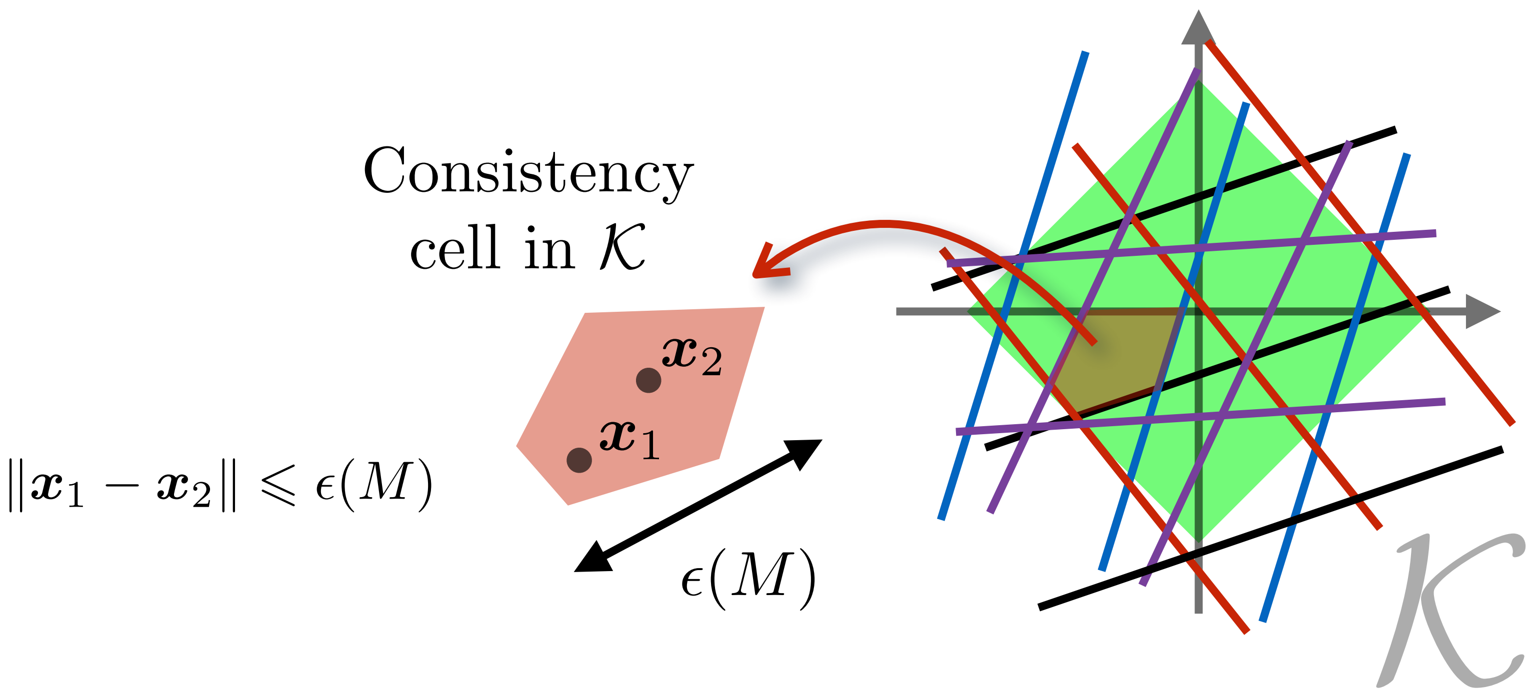
Signals $\boldsymbol{u} \in \mathcal{K}$ s.t.

$\mathsf{A}(\boldsymbol{u}) := \mathcal{Q}(\boldsymbol{\Phi}\boldsymbol{u} + \boldsymbol{\xi}) = \boldsymbol{y}$

for some $\boldsymbol{y} \in \delta\mathbb{Z}^M$

Consistency cell in $\mathcal{K}$

Control of the "consistency width"

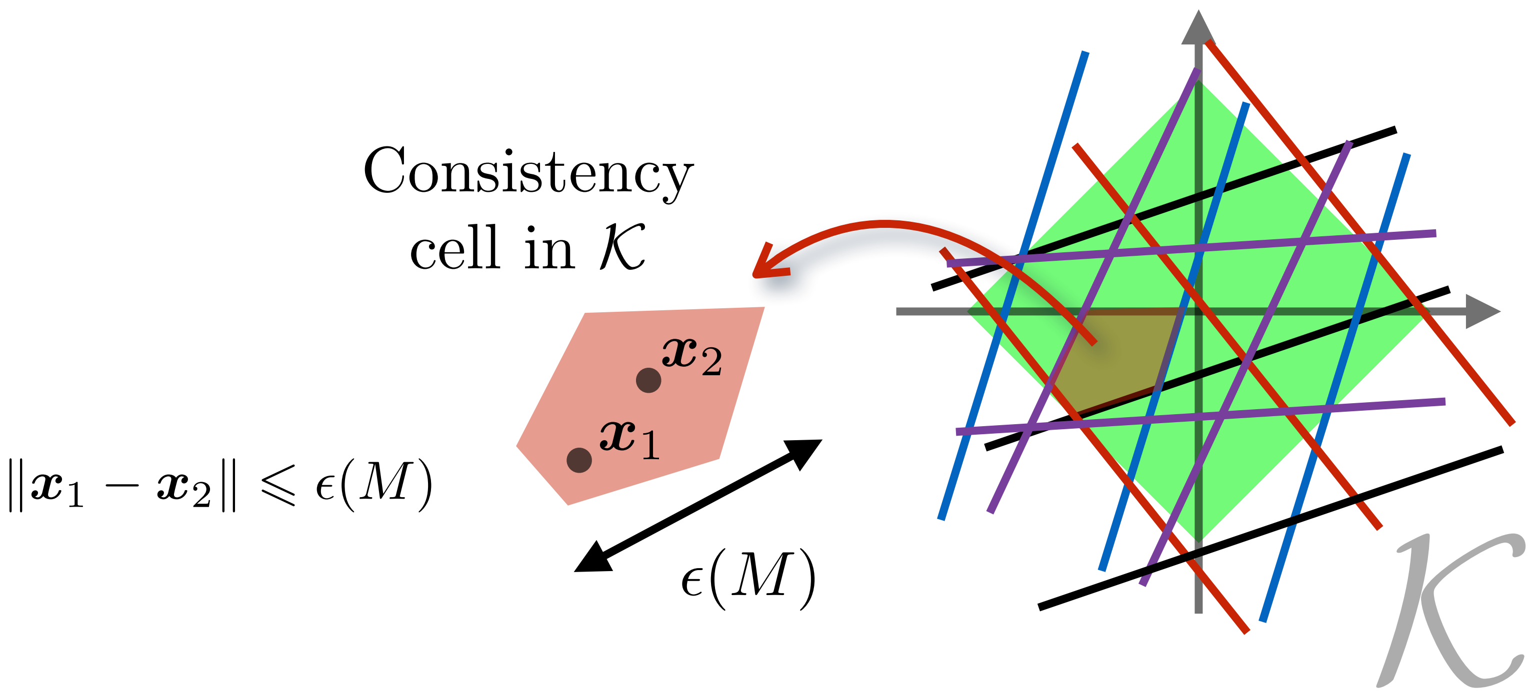


$$\boldsymbol{\Phi} = \begin{pmatrix} \boldsymbol{\varphi}_1^T \\ \vdots \\ \boldsymbol{\varphi}_M^T \end{pmatrix}$$

<u>Definition:</u> "Consistency width" $\epsilon(M) :=$
Largest distance between 2 points from any consistency cell.

($\equiv$ worst case error of algorithms with a consistent solution)

Control of the "consistency width"



$$\boldsymbol{\Phi} = \begin{pmatrix} \boldsymbol{\varphi}_1^T \\ \vdots \\ \boldsymbol{\varphi}_M^T \end{pmatrix}$$

For $\boldsymbol{\Phi}$ a random Gaussian matrix, with high probability,

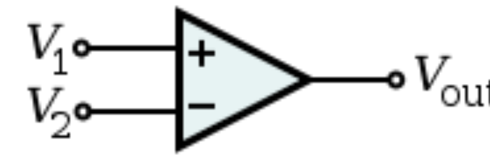
[LJ, 16], [LJ, 17]

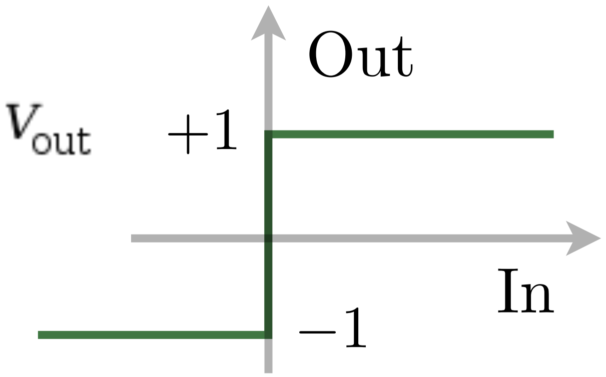
$$\epsilon(M) \leqslant C_{\mathcal{K},\delta}\, M^{-1/q}$$

with $q = 1$ (for, *e.g.*, sparse signals, low-rank matrices), or $q = 4$ for convex sets.

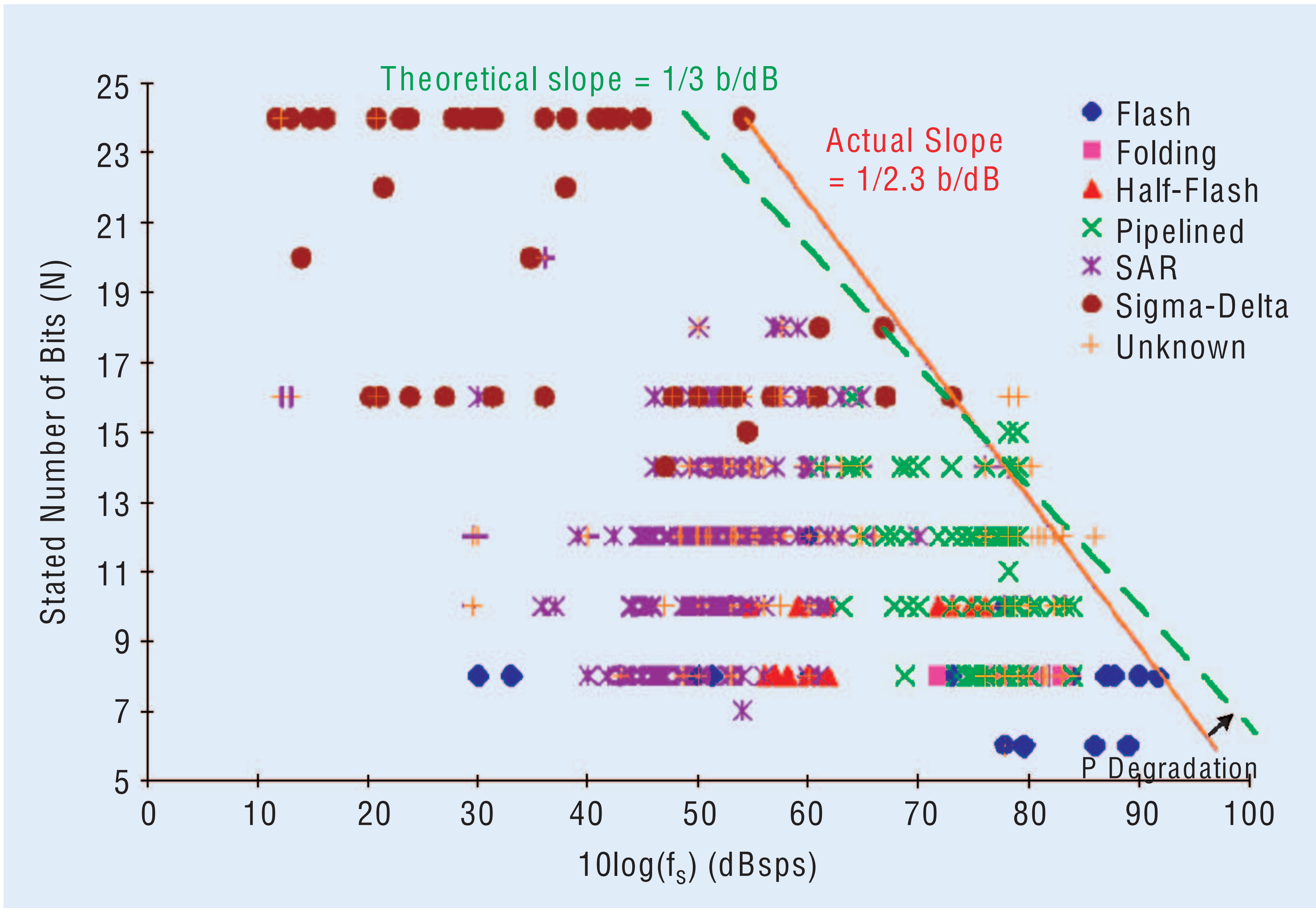
Open problem:
Extension to RIP matrices?

Binary embeddings





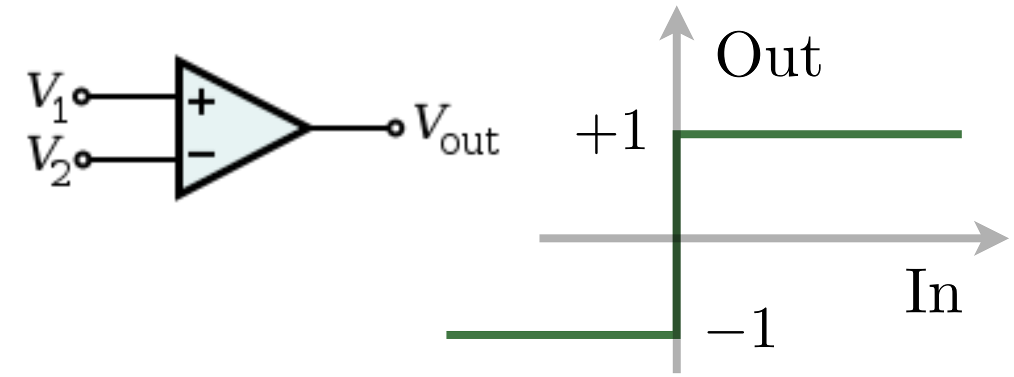
- Why 1-bit?



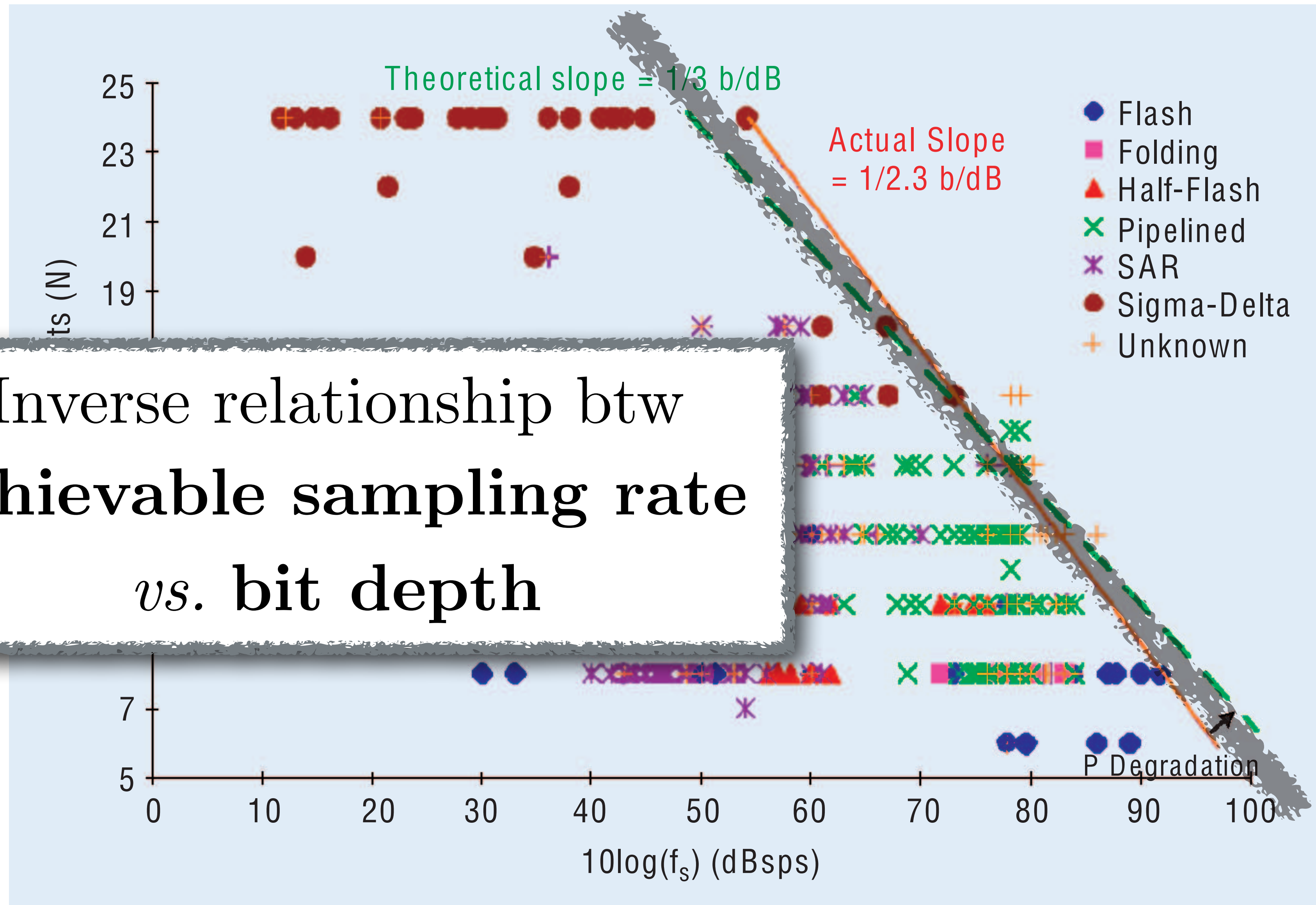
[FIG1] Stated number of bits versus sampling rate.

[From "Analog-to-digital converters" B. Le, T.W. Rondeau, J.H. Reed, and C.W.Bostian, IEEE Sig. Proc. Magazine, Nov 2005]

Binary embeddings



- Why 1-bit?



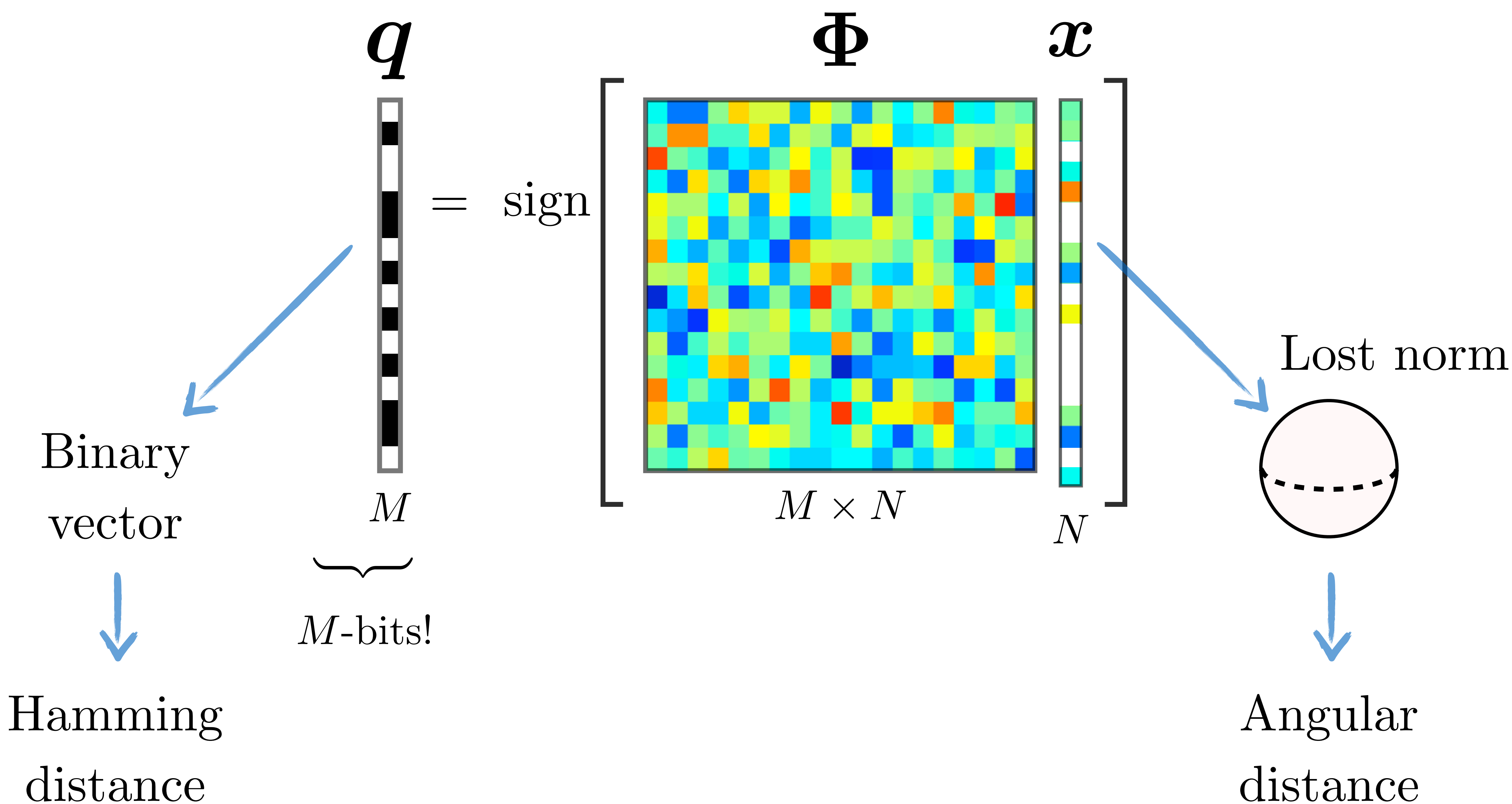
[FIG1] Stated number of bits versus sampling rate.

Inverse relationship btw **achievable sampling rate** *vs.* **bit depth**

[From “Analog-to-digital converters” B. Le, T.W. Rondeau, J.H. Reed, and C.W.Bostian, IEEE Sig. Proc. Magazine, Nov 2005]

Binary embeddings

- Why 1-bit?
- Embedding in which distances?



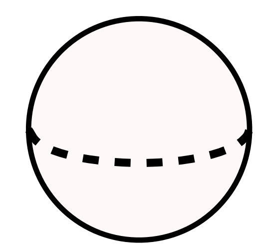
Binary embeddings

- Why 1-bit?
- Embedding in which distances?

$$d_H(\boldsymbol{u}, \boldsymbol{v}) = \tfrac{1}{M} \textstyle\sum_i (u_i \oplus v_i) \quad \text{(norm. Hamming)}$$
$$d_{\rm ang}(\boldsymbol{x}, \boldsymbol{s}) = \tfrac{1}{\pi} \arccos(\langle \boldsymbol{x}, \boldsymbol{s} \rangle) \quad \text{(norm. angle)}$$

Binary vector → Hamming distance

Lost norm



→ Angular distance

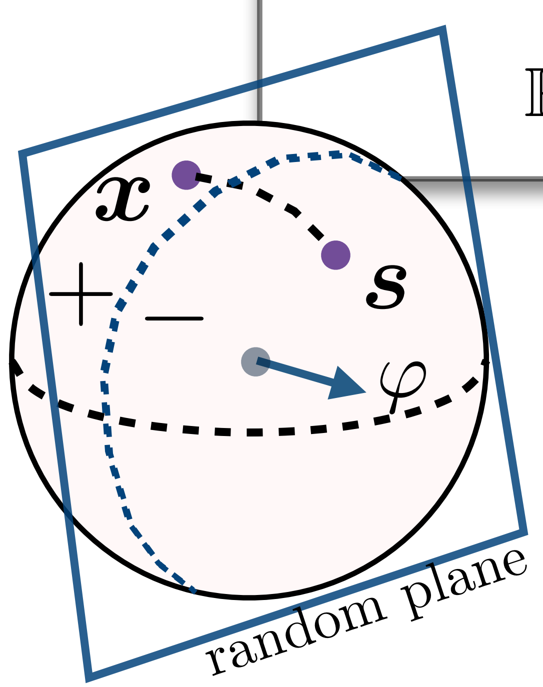
Binary embeddings

- Why 1-bit?
- Embedding in which distances?

$$d_H(\boldsymbol{u},\boldsymbol{v}) = \tfrac{1}{M}\textstyle\sum_i (u_i \oplus v_i) \quad \text{(norm. Hamming)}$$
$$d_{\rm ang}(\boldsymbol{x},\boldsymbol{s}) = \tfrac{1}{\pi}\arccos(\langle \boldsymbol{x},\boldsymbol{s}\rangle) \quad \text{(norm. angle)}$$

- Fact: **[e.g., Goemans, Williamson, '95]**

> Let $\boldsymbol{\Phi} \sim \mathcal{N}^{M\times N}(0,1)$, $A(\cdot) = \mathrm{sign}\,(\boldsymbol{\Phi}\,\cdot) \in \{-1,1\}^M$ and $\epsilon > 0$.
> For any $\boldsymbol{x},\boldsymbol{s} \in \mathbb{S}^{N-1}$, we have
>
> $$\mathbb{P}_{\boldsymbol{\Phi}}\big[\,\big|\,d_H\big(A(\boldsymbol{x}),A(\boldsymbol{s})\big) \;-\; d_{\rm ang}(\boldsymbol{x},\boldsymbol{s})\big|\;\leqslant\;\epsilon\,\big]\;\geqslant\;1-2\,e^{-2\epsilon^2 M}.$$



Thanks to $A(.)$, Hamming distance concentrates around vector angles!

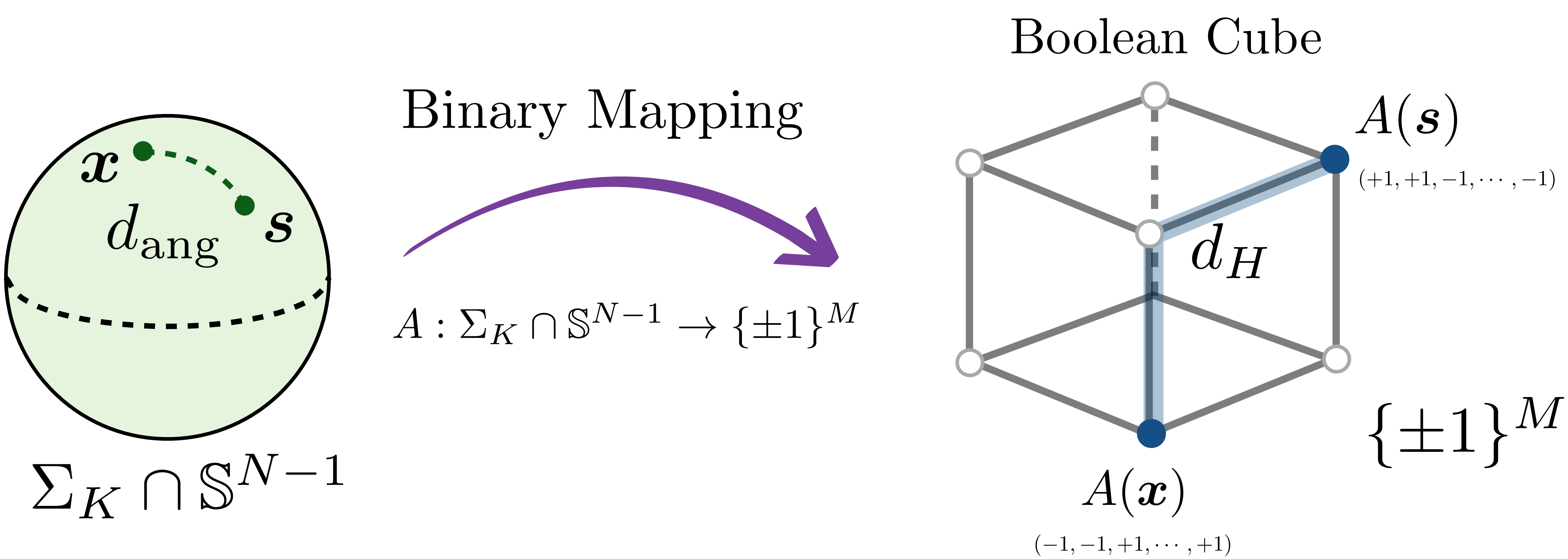
Binary ϵ-stable embedding

Kind of "binary restricted (quasi) isometry":

A mapping $A : \mathbb{R}^N \to \{\pm 1\}^M$ is a **binary ϵ-stable embedding** (BϵSE) of order K for sparse vectors if

$$d_{\rm ang}(\boldsymbol{x}, \boldsymbol{s}) - \epsilon \leqslant d_H(A(\boldsymbol{x}), A(\boldsymbol{s})) \leqslant d_{\rm ang}(\boldsymbol{x}, \boldsymbol{s}) + \epsilon$$

for all $\boldsymbol{x}, \boldsymbol{s} \in S^{N-1}$ with $\boldsymbol{x} \pm \boldsymbol{s}$ K-sparse.



Binary ϵ-stable embedding

Kind of "binary restricted (quasi) isometry":

A mapping $A : \mathbb{R}^N \to \{\pm 1\}^M$ is a **binary ϵ-stable embedding** (BϵSE) of order K for sparse vectors if

$$d_{\rm ang}(\boldsymbol{x}, \boldsymbol{s}) - \epsilon \leqslant d_H(A(\boldsymbol{x}), A(\boldsymbol{s})) \leqslant d_{\rm ang}(\boldsymbol{x}, \boldsymbol{s}) + \epsilon$$

for all $\boldsymbol{x}, \boldsymbol{s} \in S^{N-1}$ with $\boldsymbol{x} \pm \boldsymbol{s}$ K-sparse.

Binarized gaussian random projections

Let $\boldsymbol{\Phi} \sim \mathcal{N}^{M\times N}(0,1)$, fix $0 \leqslant \eta \leqslant 1$ and $\epsilon > 0$. If

$$M \ \geqslant \ \tfrac{4}{\epsilon^2}\big(K \log(N) + 2K \log(\tfrac{50}{\epsilon}) + \log(\tfrac{2}{\eta})\big),$$

then $\boldsymbol{\Phi}$ is a BϵSE with $\Pr > 1 - \eta$.

$$M = O(\epsilon^{-2} K \log N)$$

[LJ, J. Laska, PB, R. Baraniuk, '13]

Beyond strict sparsity ... [Plan, Vershynin, '13]

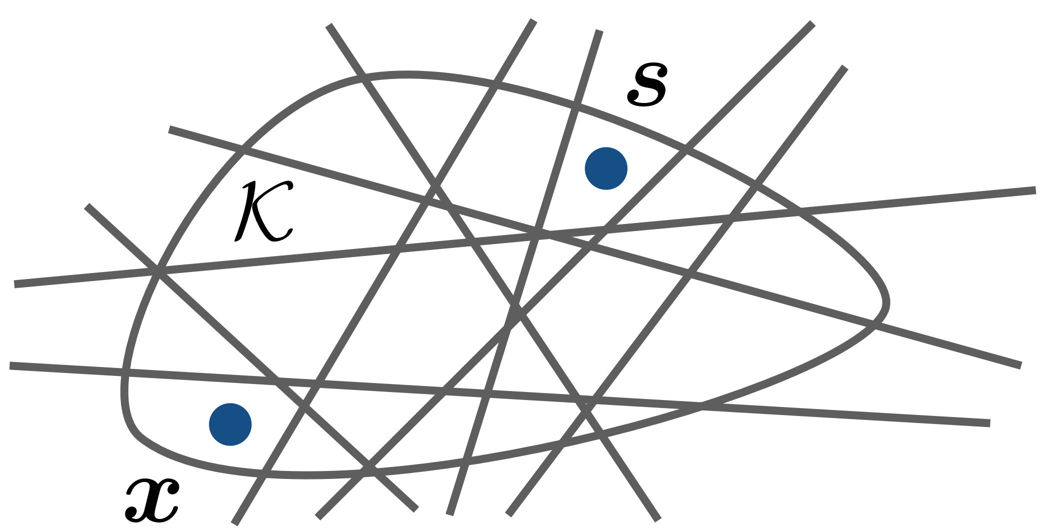
Proposition *Let* $\boldsymbol{\Phi} \sim \mathcal{N}^{M\times N}(0,1)$ *and* $\mathcal{K} \subset \mathbb{R}^N$. *Then, for some* $C, c > 0$, *if*

$$M \;\geqslant\; C\epsilon^{-6} w^2(\mathcal{K}),$$

then, with $Pr \geqslant 1 - e^{-c\epsilon^2 M}$, *we have*

$$d_{\rm ang}(\boldsymbol{x},\boldsymbol{s}) - \epsilon \;\leqslant\; d_H(A(\boldsymbol{x}), A(\boldsymbol{s})) \;\leqslant\; d_{\rm ang}(\boldsymbol{x},\boldsymbol{s}) - \epsilon, \quad \forall \boldsymbol{x},\boldsymbol{s} \in \mathcal{K}.$$

Random hyperplane tessellations



Beyond strict sparsity … [Plan, Vershynin, '13]

Proposition *Let* $\boldsymbol{\Phi} \sim \mathcal{N}^{M\times N}(0,1)$ *and* $\mathcal{K} \subset \mathbb{R}^N$*. Then, for some* $C, c > 0$*, if*

$$M \geqslant C\epsilon^{-6} w^2(\mathcal{K}),$$

then, with $Pr \geqslant 1 - e^{-c\epsilon^2 M}$*, we have*

$$d_{\rm ang}(\boldsymbol{x},\boldsymbol{s}) - \epsilon \leqslant d_H(A(\boldsymbol{x}), A(\boldsymbol{s})) \leqslant d_{\rm ang}(\boldsymbol{x},\boldsymbol{s}) - \epsilon, \quad \forall \boldsymbol{x},\boldsymbol{s} \in \mathcal{K}.$$

not as optimal but stronger result!

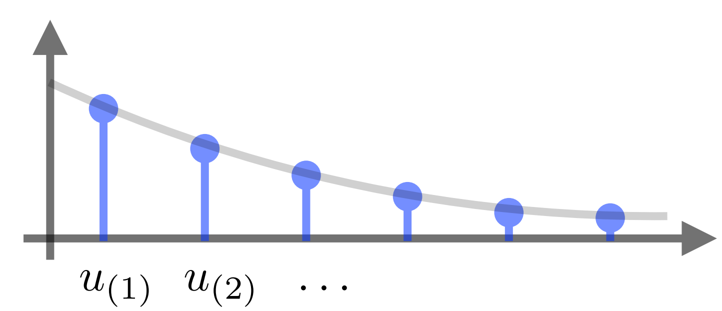
Generalize BϵSE to more general sets!

e.g., to non-conic sets such as:

Set of compressible signals:

$$\mathcal{C}_K = \{\boldsymbol{u} \in \mathbb{R}^N : \|\boldsymbol{u}\|_2/\|\boldsymbol{u}\|_1 \leqslant \sqrt{K}\} \supset \Sigma_K$$

with $w^2(\mathcal{C}_K) \leqslant cK \log N/K$.



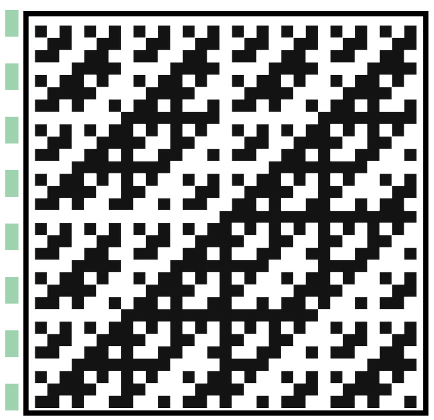
Beyond the Gaussian Domination

- Beware the counter example! (e.g., binary matrix)
- Several constructions for finite sets

e.g., **[F. Yu et al, '15][S. Oymak, '16][S. Dirksen, A. Stollenwerk, '16]**
[S. Dirksen, S. Mendelson, '18]

Selection matrix ($m \times n$), Circulant Gaussian ($n \times n$), $\mathrm{diag}(a_1, \cdots, a_n)$ ($n \times n$), Hadamard ($n \times n$), $\mathrm{diag}(b_1, \cdots, b_n)$ ($n \times n$):

$$\boldsymbol{\Phi} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \end{bmatrix} \begin{bmatrix} g_1 & g_2 & \cdots & g_n \\ g_2 & g_3 & \cdots & g_1 \\ \vdots & & \ddots & \vdots \\ g_n & g_1 & \cdots & g_{n-1} \end{bmatrix} \mathrm{diag}(a_1, \cdots, a_n) \; [\text{Hadamard}] \; \mathrm{diag}(b_1, \cdots, b_n)$$



"spreading" part

with $g_i, a_i, b_i \sim_{\mathrm{iid}} \mathcal{N}(0,1)$.

If $\log N \lesssim \epsilon^2 (\log n)^{-1} n^{1/3}$ and $m \gtrsim \epsilon^{-3} \log N$,
Then, $f(\cdot) = \mathrm{sign}\,(\boldsymbol{\Phi}\cdot)$ is a ϵ-binary embedding (*i.e.*, respect BϵSE)

Beyond the Gaussian Domination

- Beware the counter example! (e.g., binary matrix)
- Several constructions for finite sets

e.g., **[F. Yu et al, '15][S. Oymak, '16][S. Dirksen, A. Stollenwerk, '16]**
[S. Dirksen, S. Mendelson, '18]

+ other constructions (e.g., Fast JL transform with Gaussian)

For most, an upper bound on N or $\log N$ (with N the number of vectors)

Beyond the Gaussian Domination

- For low-complexity vectors: The mapping

$$f(\cdot) = \mathrm{sign}\,(\boldsymbol{\Phi}\,\cdot + \boldsymbol{\xi}), \text{ with } \xi_i \sim_{\mathrm{iid}} \mathcal{N}(0, R).$$

allows for dense non-Gaussian matrix (e.g., Bernoulli)

[S. Dirksen, S. Mendelson, '18]

$\forall \boldsymbol{x}, \boldsymbol{x}' \in \mathrm{conv}(\mathcal{K}), \|\boldsymbol{x} - \boldsymbol{x}'\| \geqslant \epsilon,$

$$c\,\frac{\|\boldsymbol{x} - \boldsymbol{x}'\|}{R} \leqslant d_H(f(\boldsymbol{x}), f(\boldsymbol{x}')) \leqslant c'\sqrt{\log(eR/\epsilon)}\,\frac{\|\boldsymbol{x} - \boldsymbol{x}'\|}{R},$$

w.h.p., provided $m \gtrsim R\epsilon^{-3}\log(R/\epsilon)w^2(\mathcal{K})$.

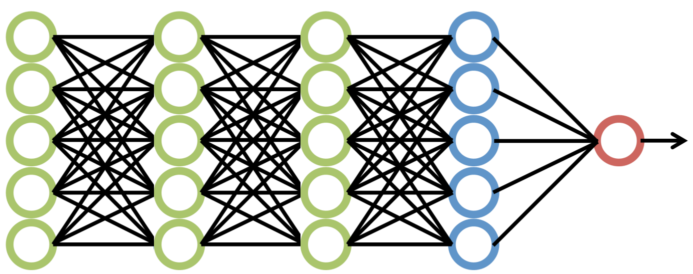
- Valid for any bounded, low-complexity set!
- asymmetric bounds
- restriction to well separated vectors

Binary Embedding for Deep Learning

R. Giryes, G. Sapiro and A.M. Bronstein,

"**Deep Neural Networks with Random Gaussian Weights: A Universal Classification Strategy?**"

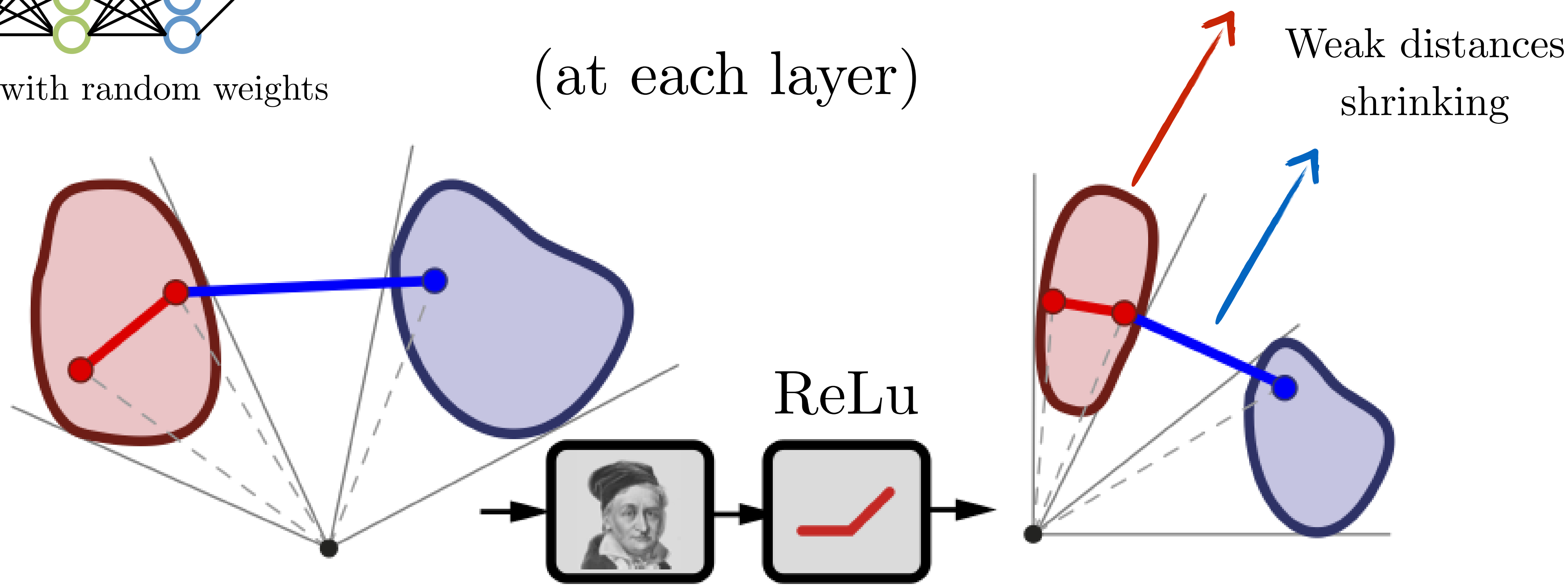
IEEE Transactions on Signal Processing, vol. 64, no. 13, pp. 3444-3457, Jul. 2016.



DNN/CNN with random weights

Strong distances
shrinking (small angle)

(at each layer)

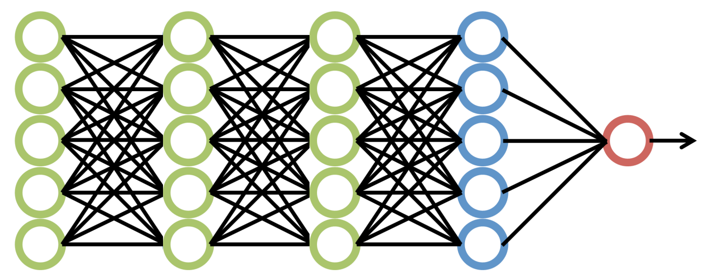


Binary Embedding for Deep Learning

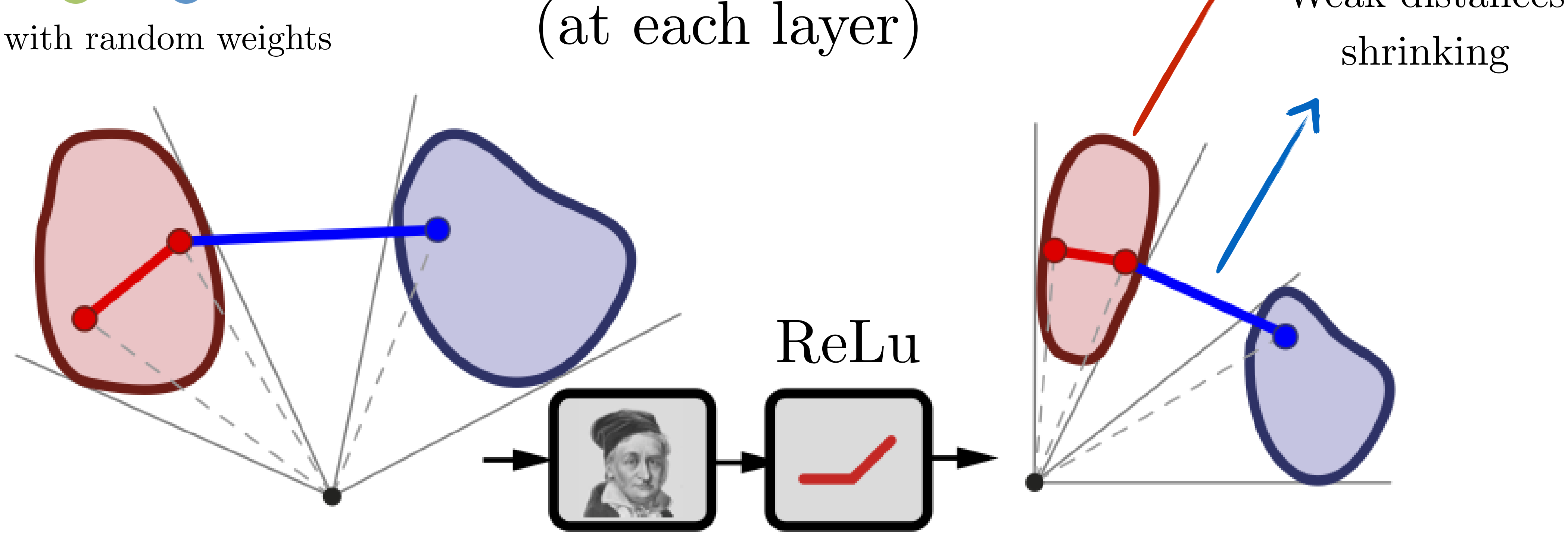
R. Giryes, G. Sapiro and A.M. Bronstein,

“**Deep Neural Networks with Random Gaussian Weights: A Universal Classification Strategy?**”

IEEE Transactions on Signal Processing, vol. 64, no. 13, pp. 3444-3457, Jul. 2016.



DNN/CNN with random weights



$\forall \boldsymbol{x}, \boldsymbol{s} \in \mathcal{K}$

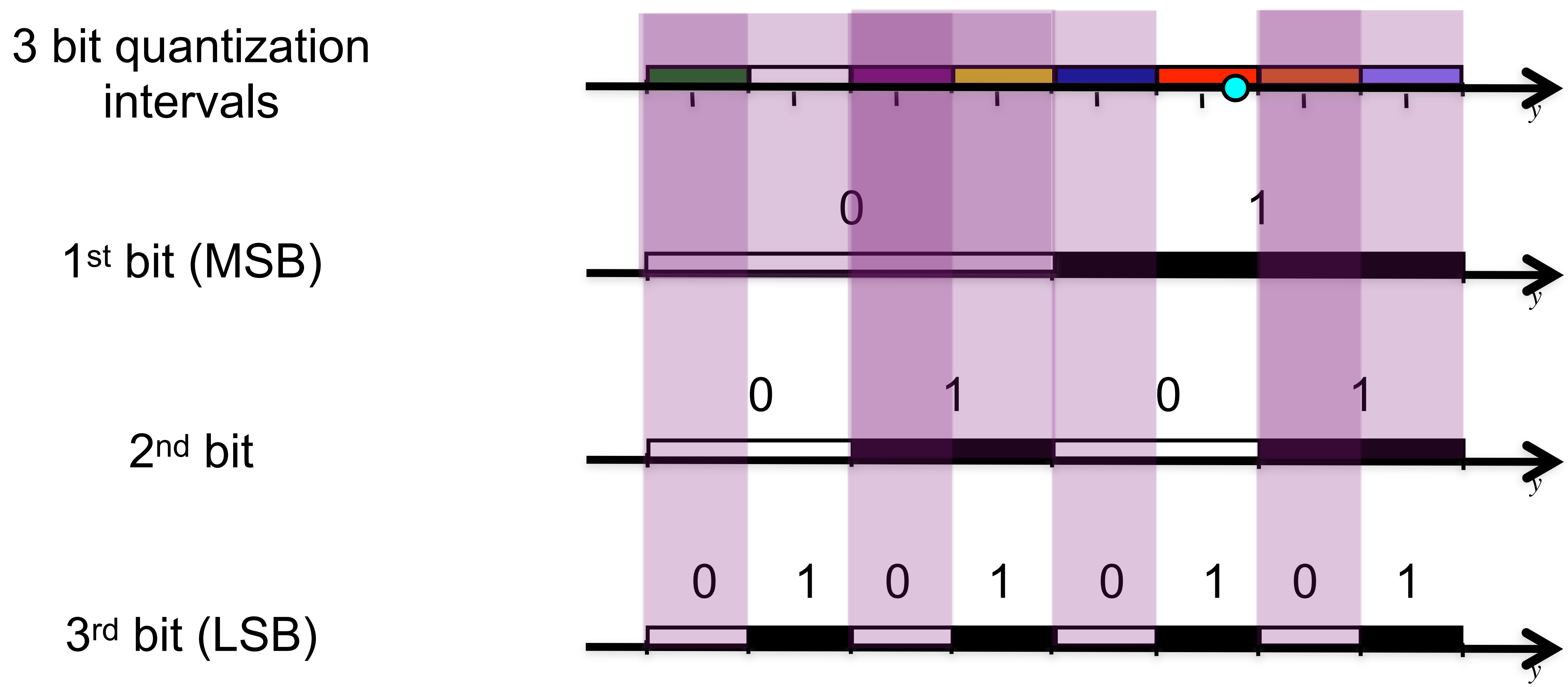
$$d_{\text{ang}}(\boldsymbol{x}, \boldsymbol{s}) - \epsilon \leqslant d_H(A(\boldsymbol{x}), A(\boldsymbol{s})) \leqslant d_{\text{ang}}(\boldsymbol{x}, \boldsymbol{s}) + \epsilon$$

$\approx \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{s}\|$ if $\angle(\boldsymbol{x}, \boldsymbol{s})$ small.

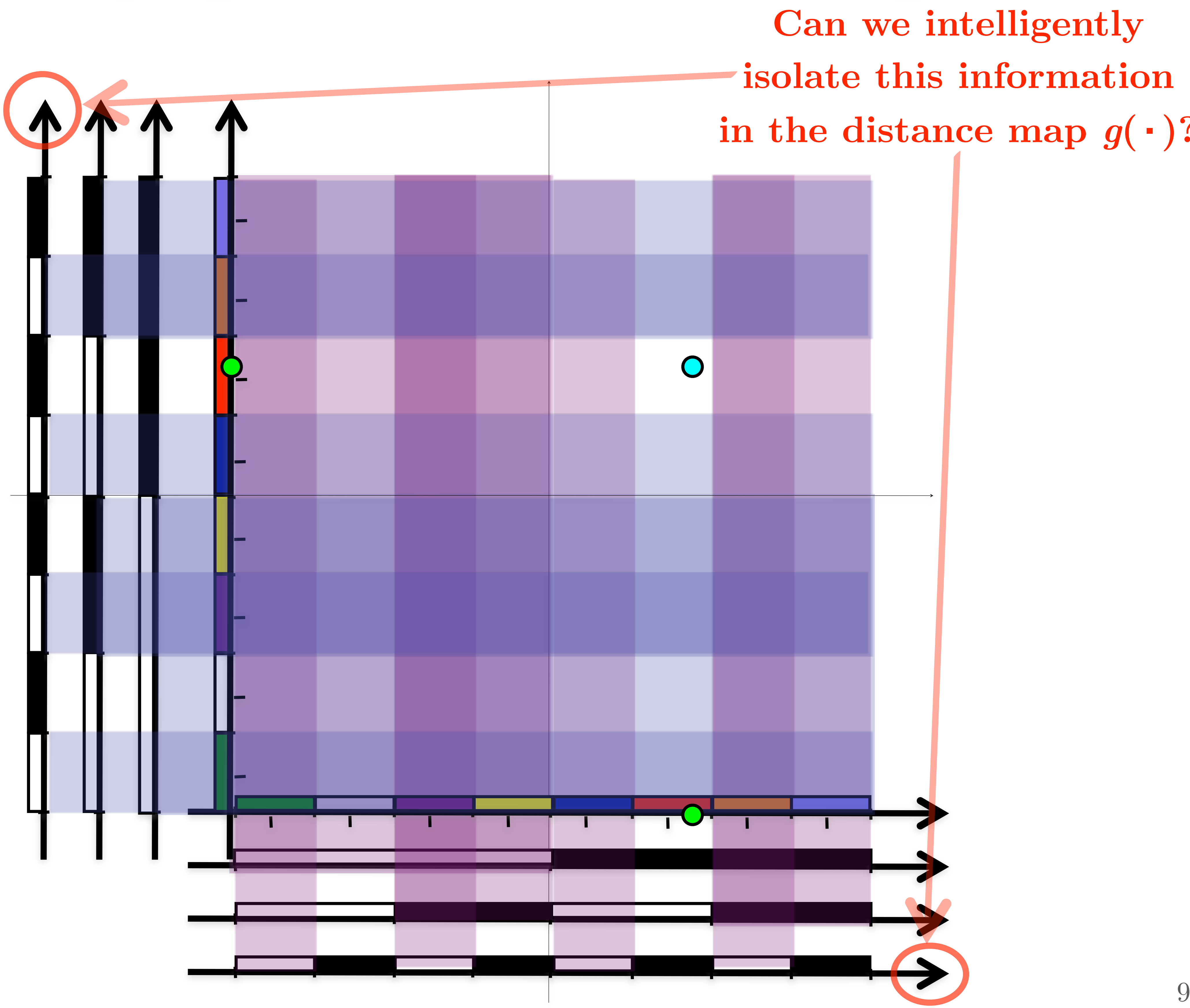
$$\Rightarrow \quad \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{s}\| - \epsilon \leqslant \frac{1}{\sqrt{m}}\|\rho(\boldsymbol{\Phi x}) - \rho(\boldsymbol{\Phi s})\| \leqslant \|\boldsymbol{x} - \boldsymbol{s}\| + \epsilon$$

Universally quantized embedding

What can a bit tell us?

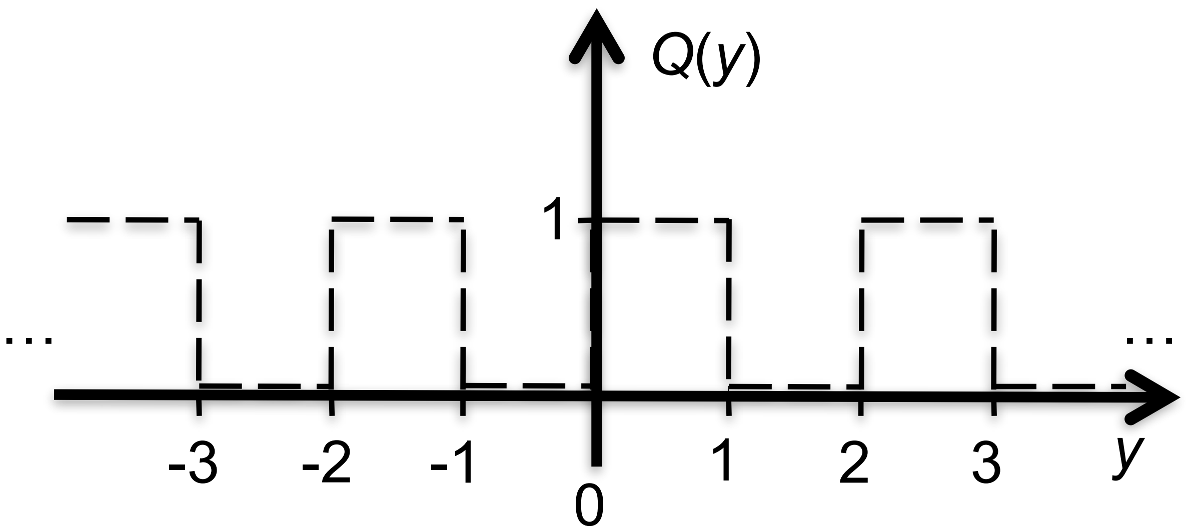


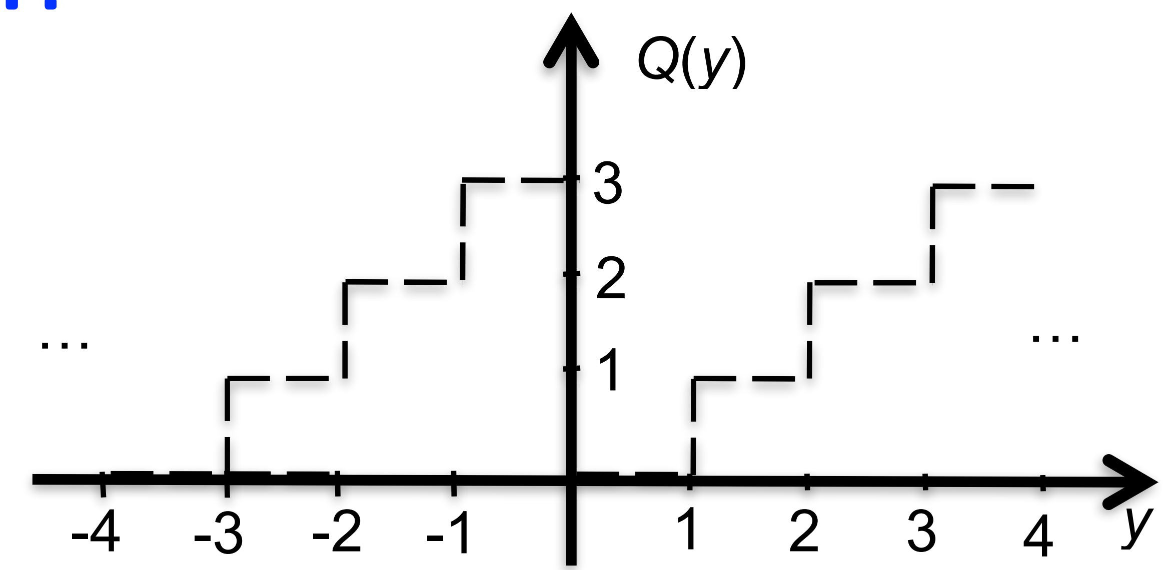
Universally quantized embedding



Rate-Efficient Scalar Quantization

Solution: **Modify the quantizer!**





Non-monotonic quantizer: Multiple intervals quantize to same value
(Focus on 1-bit quantizer today)

$A_{ij} \sim_{\text{i.i.d.}} \mathcal{N}(0, \sigma^2)$ $\quad$ $w_i \sim_{\text{i.i.d.}} \mathcal{U}([0, \Delta])$

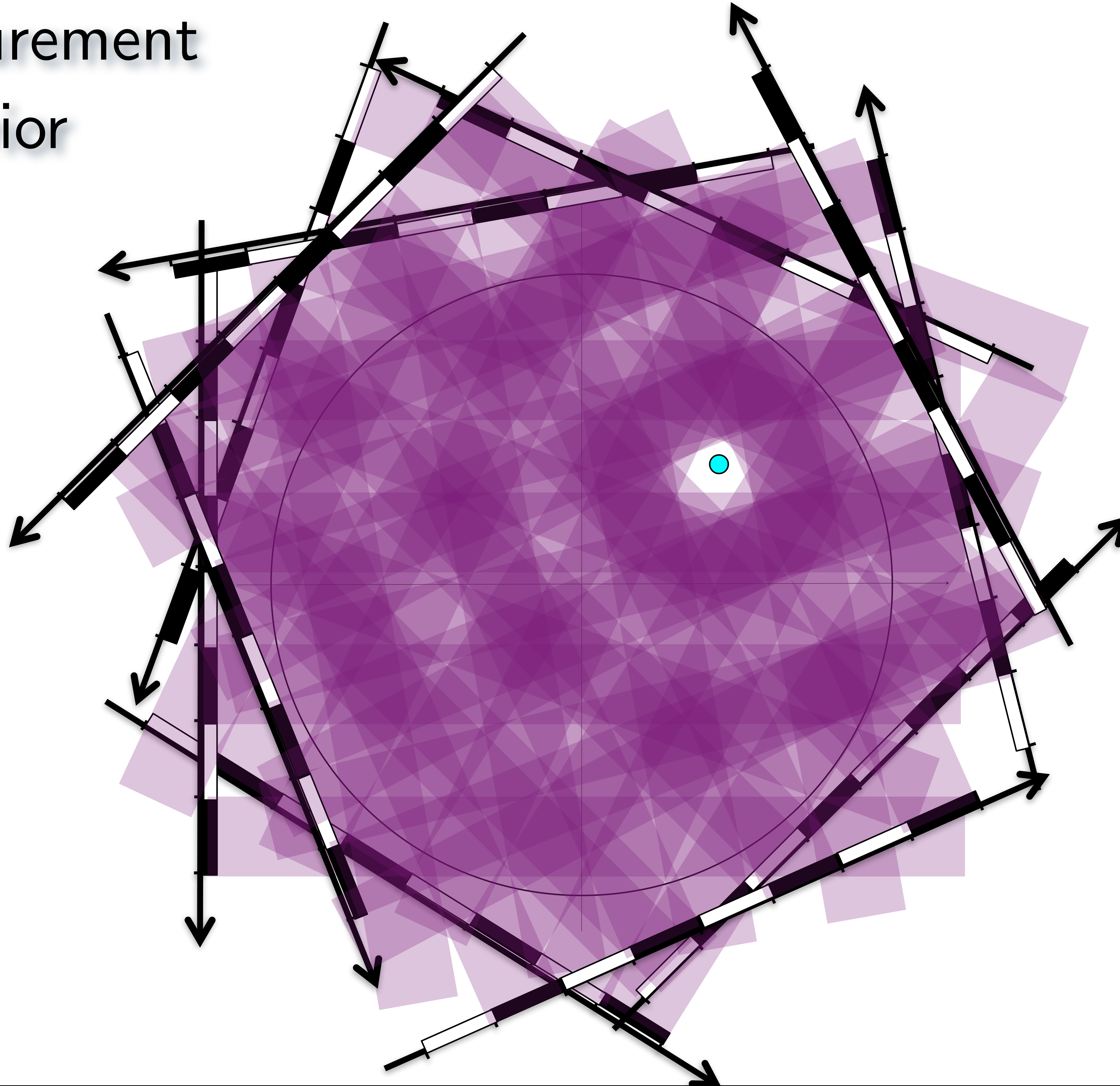
Measurements $\quad$ **Dither**

$$q_m = Q\left(\frac{\langle \mathbf{x}, \mathbf{a}_m \rangle + w_m}{\Delta_m}\right), \qquad \mathbf{q} = Q(\Delta^{-1}(\mathbf{A}\mathbf{x} + \mathbf{w}))$$

scalar quantizer
(non-monotonic)

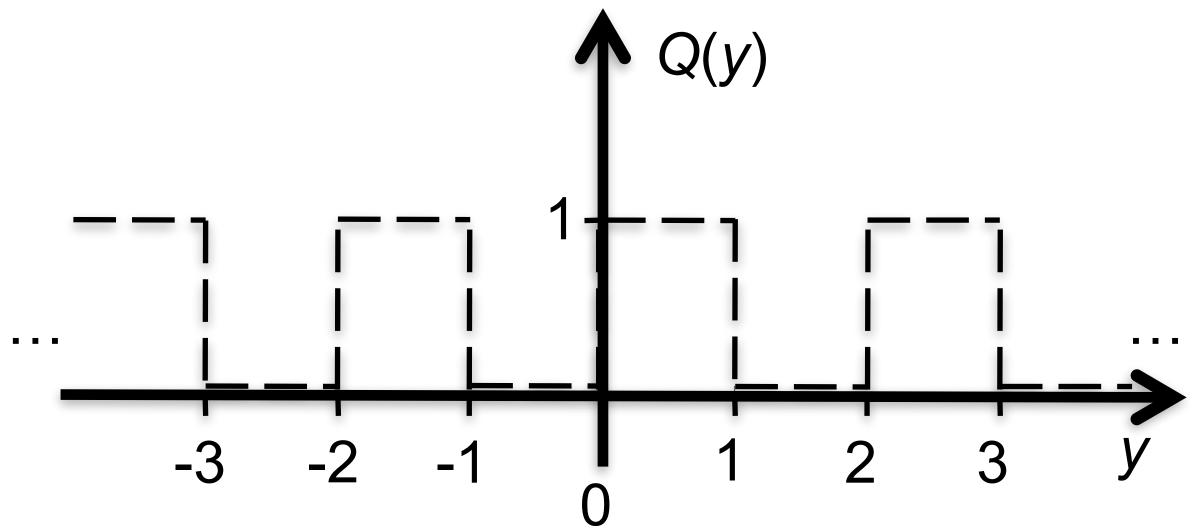
scaling/precision parameter
($\Delta_m = \Delta$, same for all measurements)

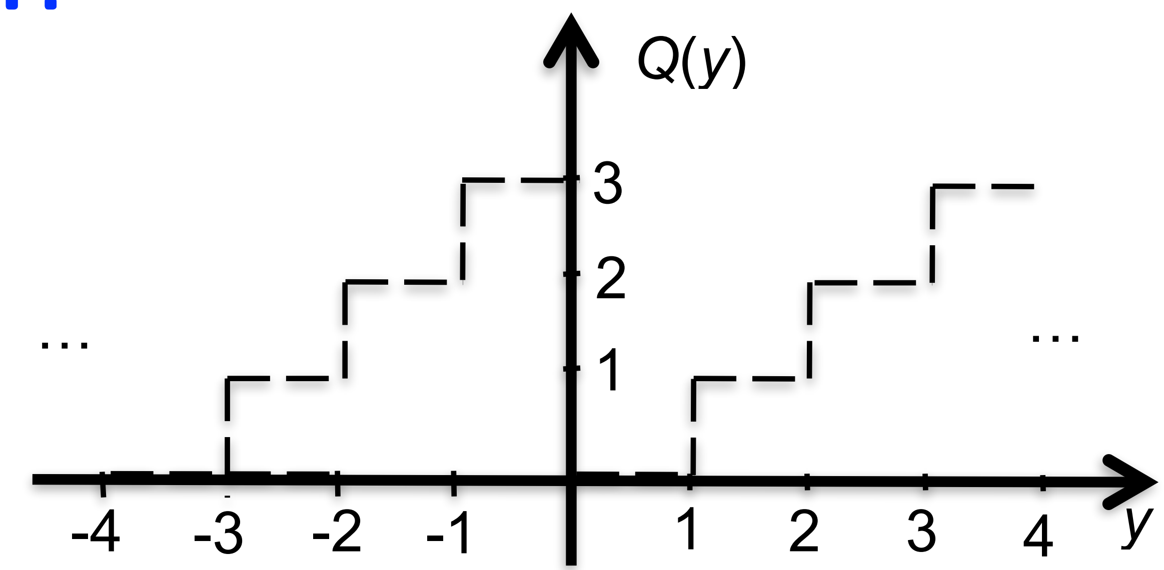
Measurement Behavior



Rate-Efficient Scalar Quantization

Solution: **Modify the quantizer!**





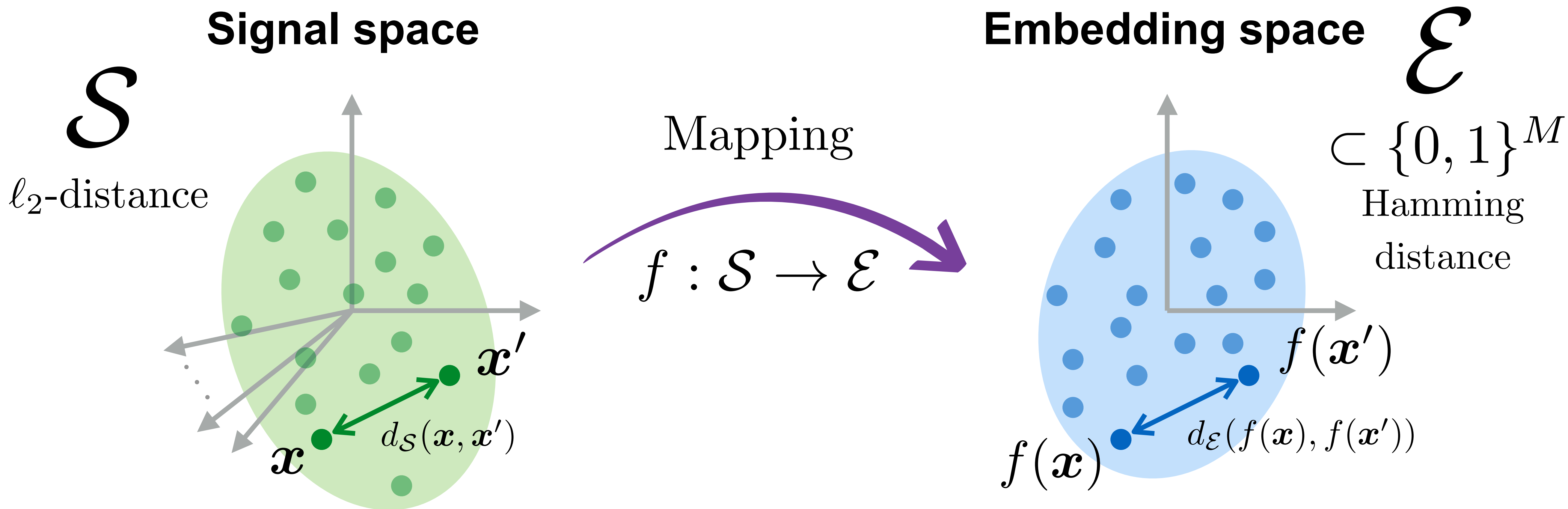
Non-monotonic quantizer: Multiple intervals quantize to same value
(Focus on 1-bit quantizer today)

Quantizer design fits the analysis framework

$$q_m = Q\left(\frac{\langle \mathbf{x}, \mathbf{a}_m \rangle + w_m}{\Delta_m}\right), \qquad \mathbf{q} = Q(\Delta^{-1}(\mathbf{A}\mathbf{x} + \mathbf{w}))$$

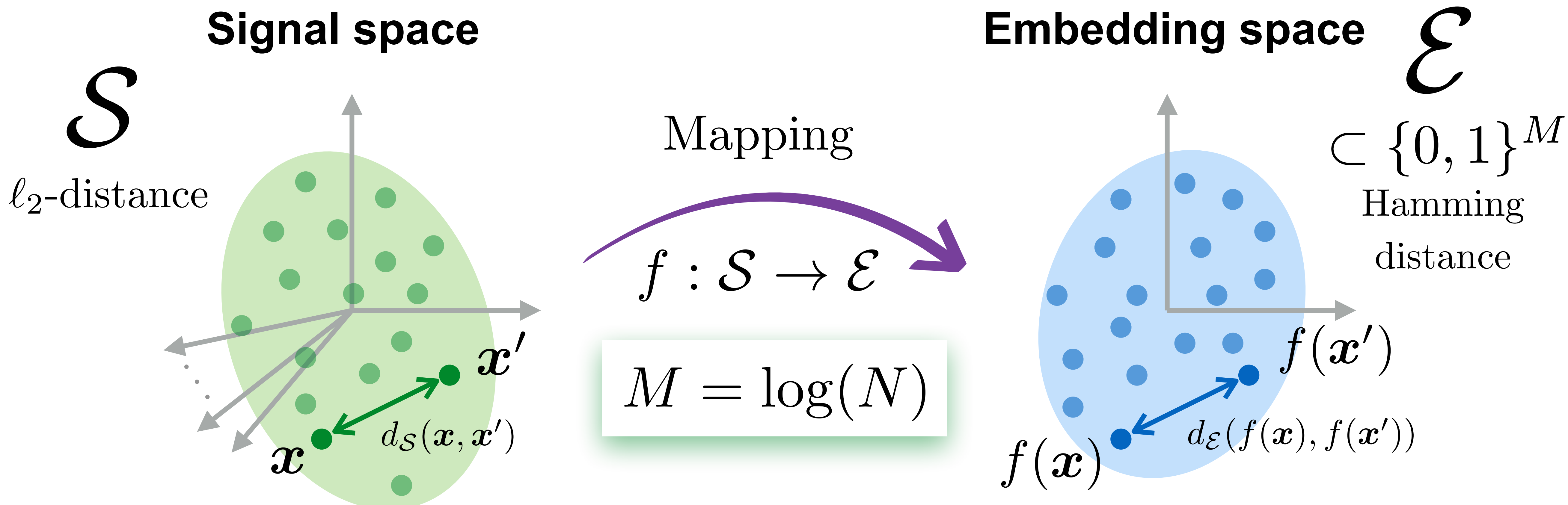
$$\mathbf{y} = h(\mathbf{A}\mathbf{x} + \mathbf{w})$$

Embedding Properties



$$f(\boldsymbol{x}) := \mathcal{Q}(\Delta^{-1}(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{w})),\ A_{ij} \sim_{\text{i.i.d.}} \mathcal{N}(0, \sigma^2),\ w_i \sim_{\text{i.i.d.}} \mathcal{U}([0, \Delta])$$

Embedding Properties



$$f(\boldsymbol{x}) := \mathcal{Q}(\Delta^{-1}(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{w})),\ A_{ij} \sim_{\text{i.i.d.}} \mathcal{N}(0, \sigma^2),\ w_i \sim_{\text{i.i.d.}} \mathcal{U}([0, \Delta])$$

For all $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{S} := \{\boldsymbol{x}_i : 1 \leqslant i \leqslant N\}$, with $d := \|\boldsymbol{x} - \boldsymbol{x}'\|$,

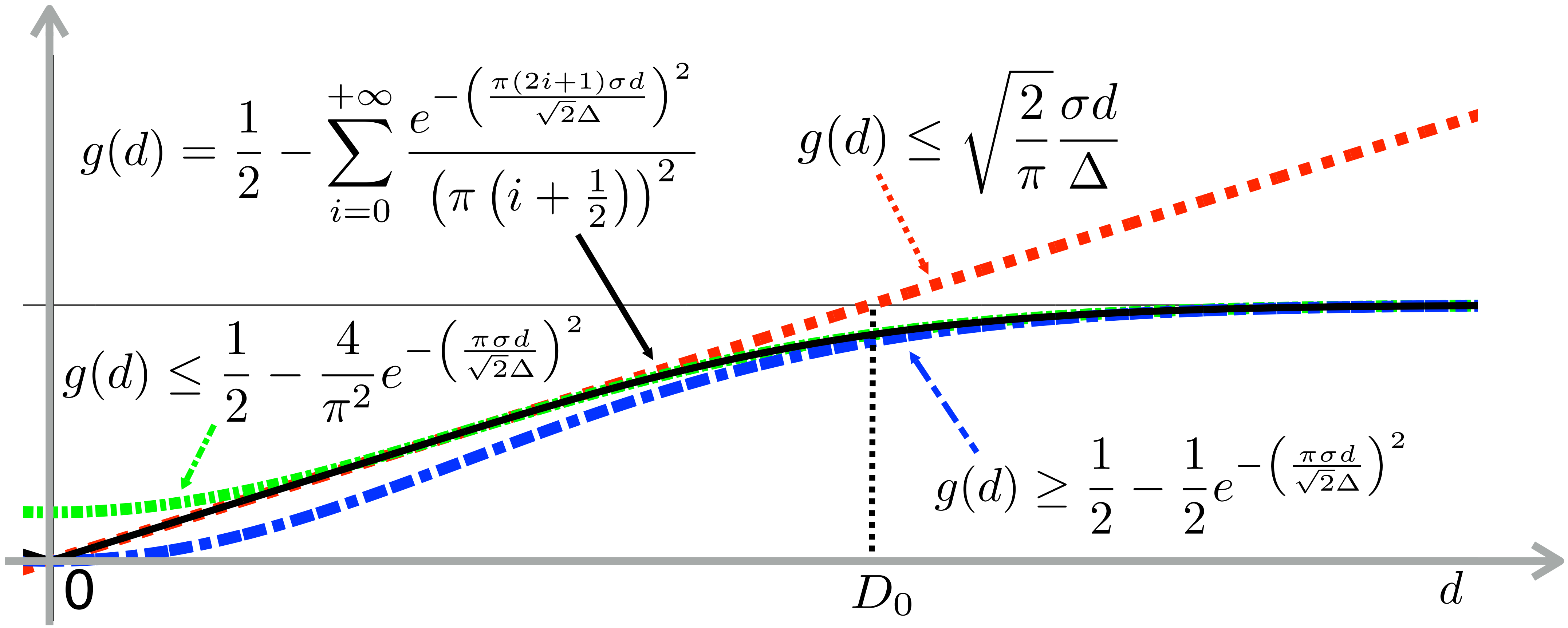
$$g(d) - \delta \leqslant d_H(f(\boldsymbol{x}), f(\boldsymbol{x}')) \leqslant g(d) + \delta, \qquad \text{w.h.p},$$

with

$$g(d) := \tfrac{1}{2} - \sum_{i=0}^{+\infty} \tfrac{4}{\pi^2(2i+1)^2} \exp\big(-\tfrac{\pi^2(2i+1)^2\sigma^2 d^2}{2\Delta^2}\big).$$

Error Behavior

$$g(d) - \delta \leqslant d_H(f(\boldsymbol{x}), f(\boldsymbol{x}')) \leqslant g(d) + \delta,$$



Distance estimate: $\tilde{d} = g^{-1}\left(d_H\left(f(\boldsymbol{x}), f(\boldsymbol{x}')\right)\right)$

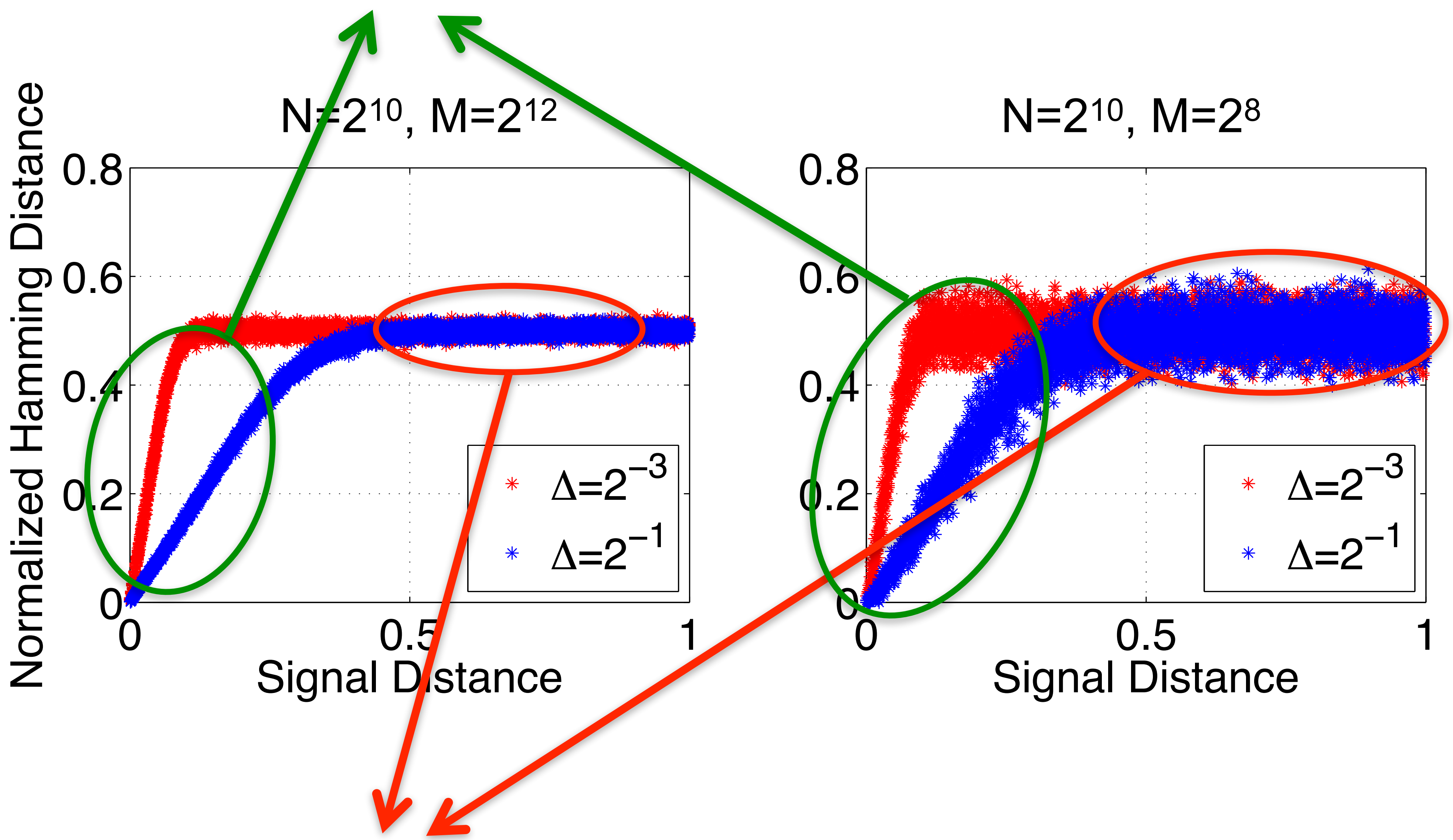
Estimate ambiguity: $\tilde{d} - \frac{\delta}{g'(\tilde{d})} \lesssim d \lesssim \tilde{d} + \frac{\delta}{g'(\tilde{d})}$

Properties (slope) controlled by choice of Δ

Error Behavior

$$g(d) - \delta \leqslant d_H(f(\boldsymbol{x}), f(\boldsymbol{x}')) \leqslant g(d) + \delta,$$

"Linear" region: $\ell_2 \propto d_H$, slope controlled by Δ

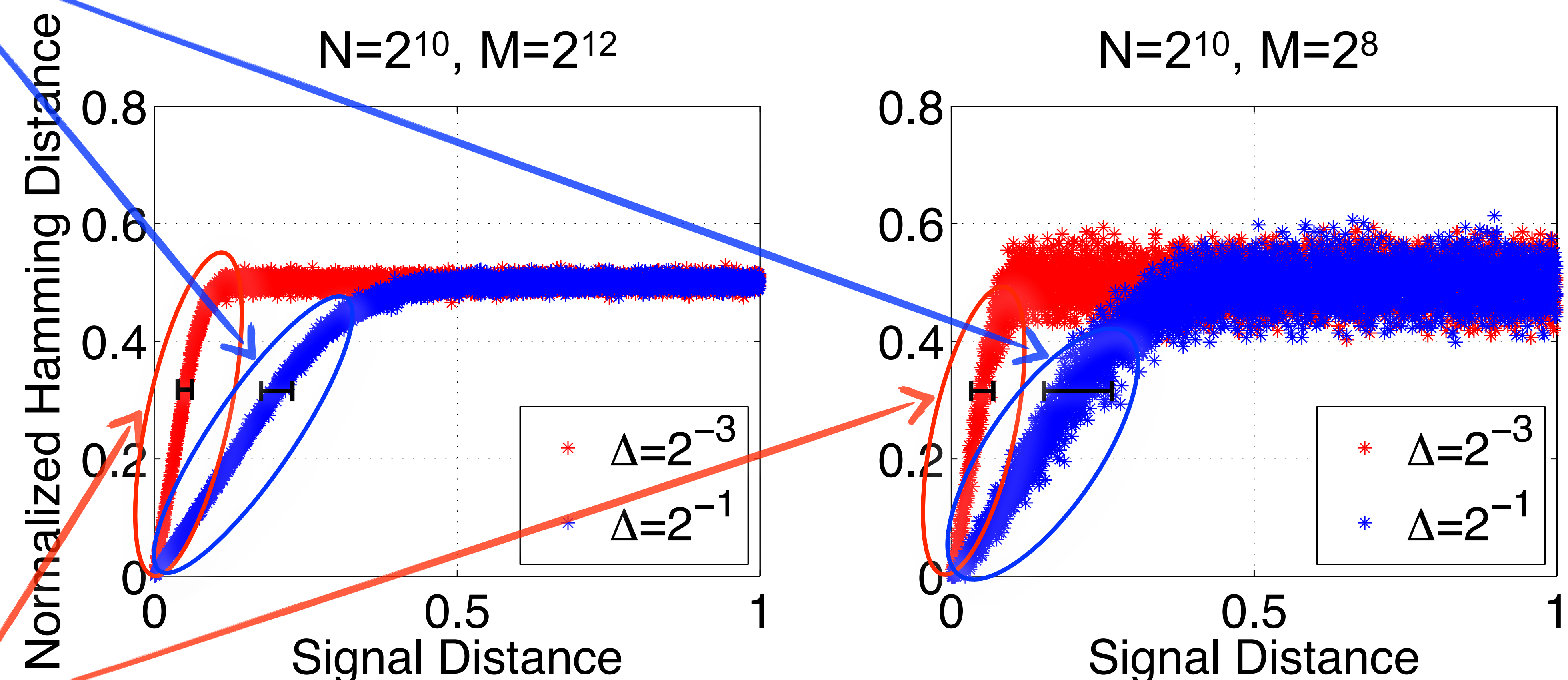


"Flat" region: no distance information

Error Behavior

$$g(d) - \delta \leqslant d_H(f(\boldsymbol{x}), f(\boldsymbol{x}')) \leqslant g(d) + \delta,$$

Large $\boldsymbol{\Delta}$: small slope, more ambiguity, preserves larger distances



Small $\boldsymbol{\Delta}$: large slope, less ambiguity, preserves smaller distances

Coffee/Tea break

Outline

1. Introduction
2. Fundamentals of embeddings and embedology
3. Quantized embeddings

   Coffee/Tea break ☕

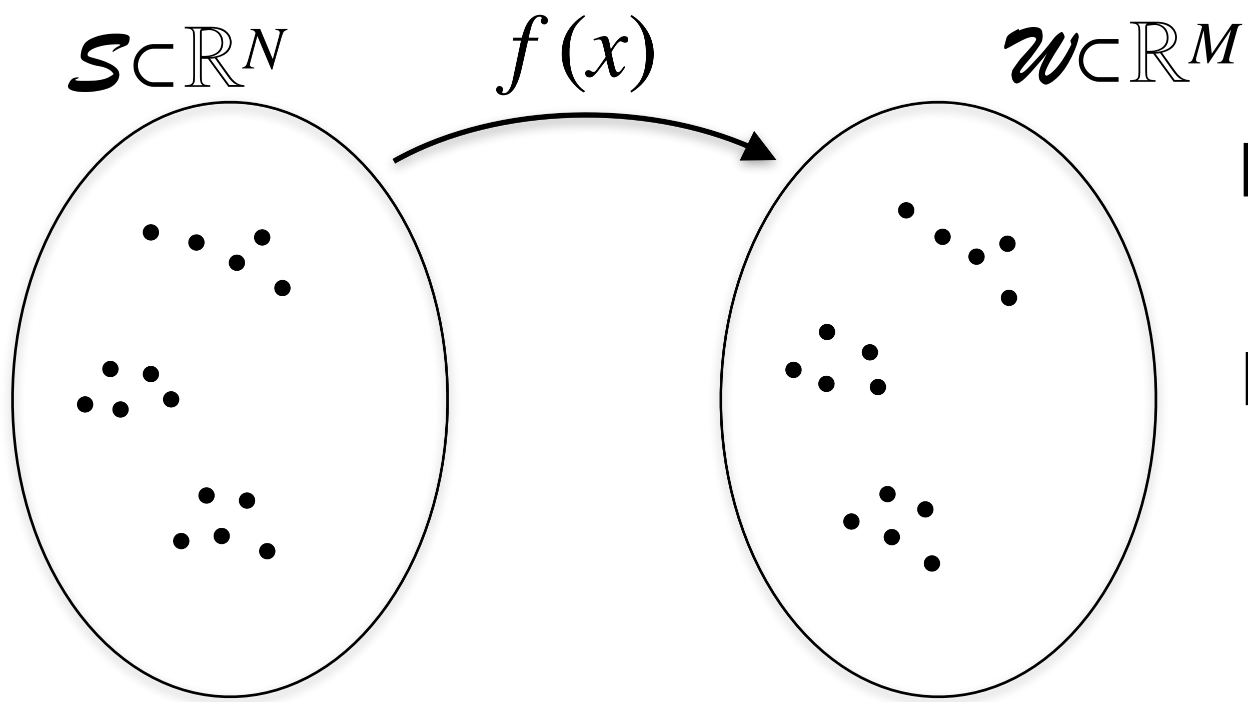
4. Embedding Design
5. Embeddings of Alternative Metrics
6. Learning Embeddings
7. Conclusions and open problems

GENERAL EMBEDDING DESIGN

Generalized Embedding Maps [B, Rane '13a]

Original space

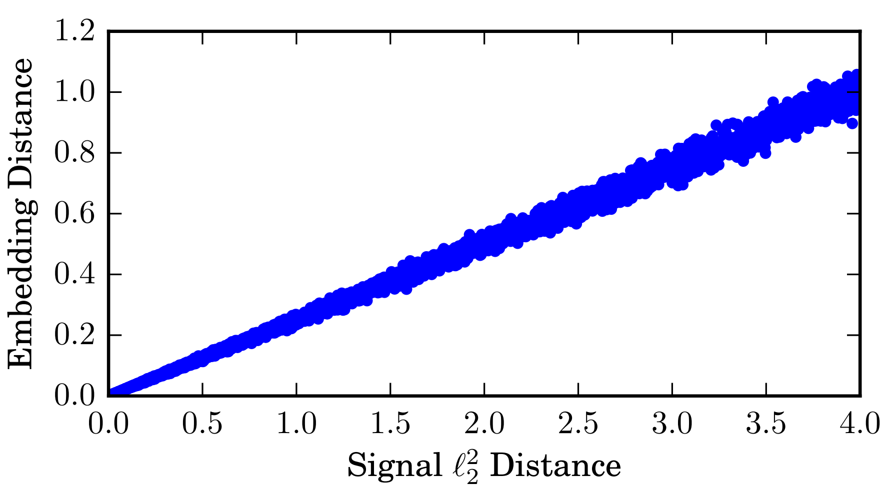
Distance metric: $d_{\mathcal{S}}$



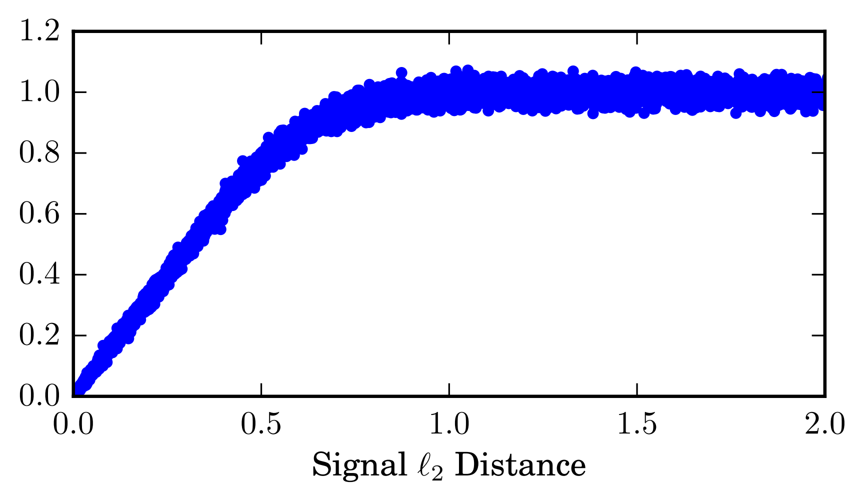
Embedding space

Embed in $\mathcal{W}$

Distance metric: $d_{\mathcal{W}}$



(a) Johnson-Lindenstrauss Embedding



(b) Universal Quantized Embedding

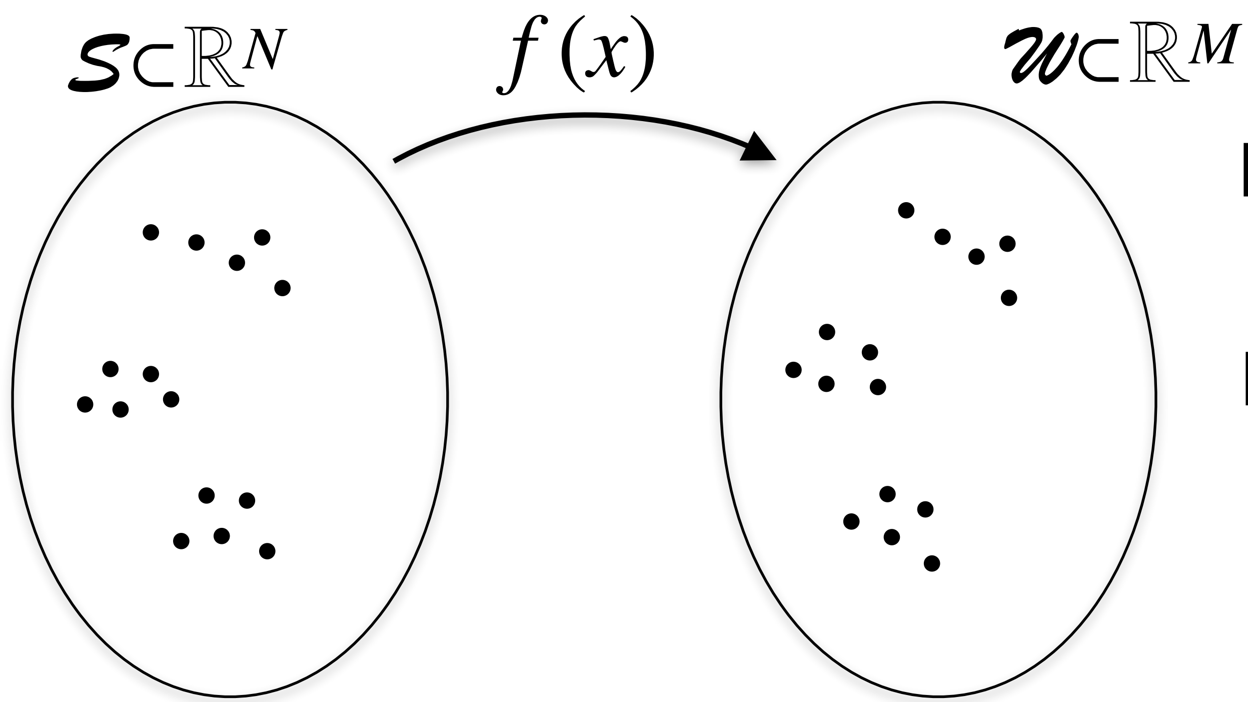
Can we construct a general **distance map?**

Can we characterize a general **distance map?**

Generalized Embedding Maps [B, Rane '13a]

Original space

Distance metric: $d_{\mathcal{S}}$



Embedding space

Embed in $\mathcal{W}$

Distance metric: $d_{\mathcal{W}}$

Assume we can construct a **distance map** $g(\cdot)$

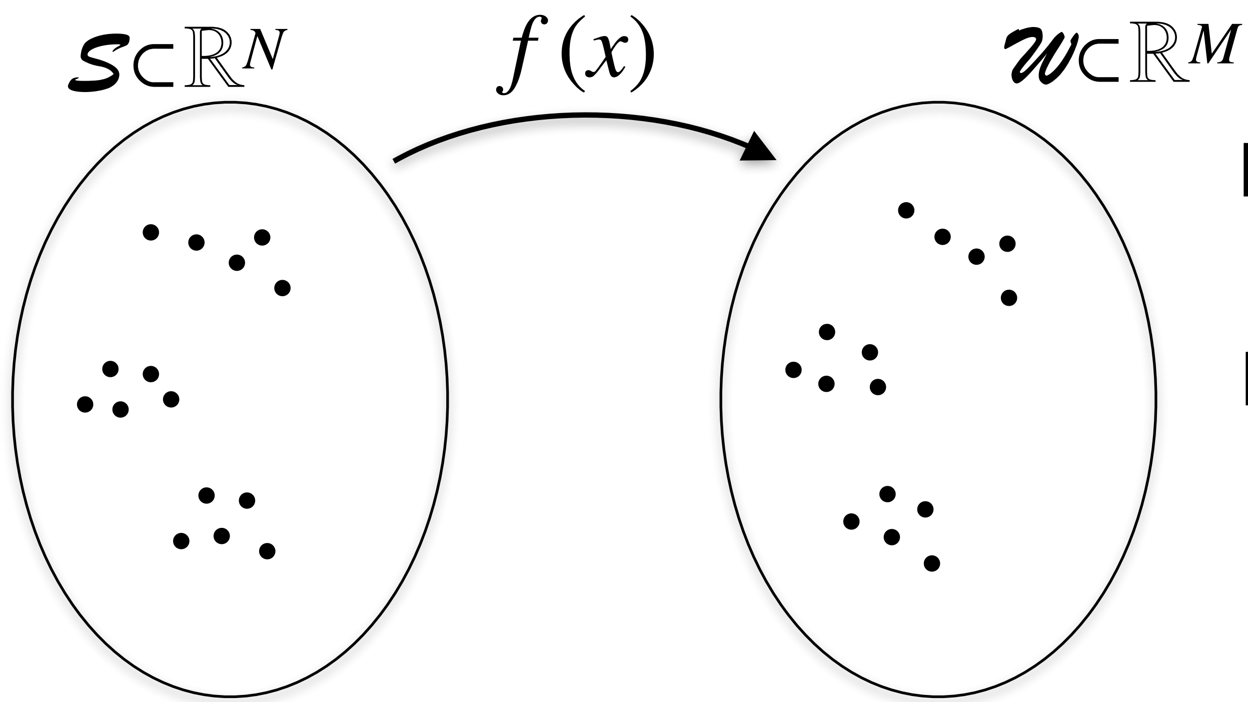
For all x,y in $\mathcal{S}$:

$$g(d_{\mathcal{S}}(\mathbf{x}, \mathbf{x}')) \approx d_{\mathcal{W}}(f(\mathbf{x}), f(\mathbf{x}'))$$

Generalized Embedding Maps [B, Rane '13a]

Original space

Distance metric: $d_{\mathcal{S}}$



Embedding space

Embed in $\mathcal{W}$

Distance metric: $d_{\mathcal{W}}$

Assume we can construct a **distance map** $g(\cdot)$

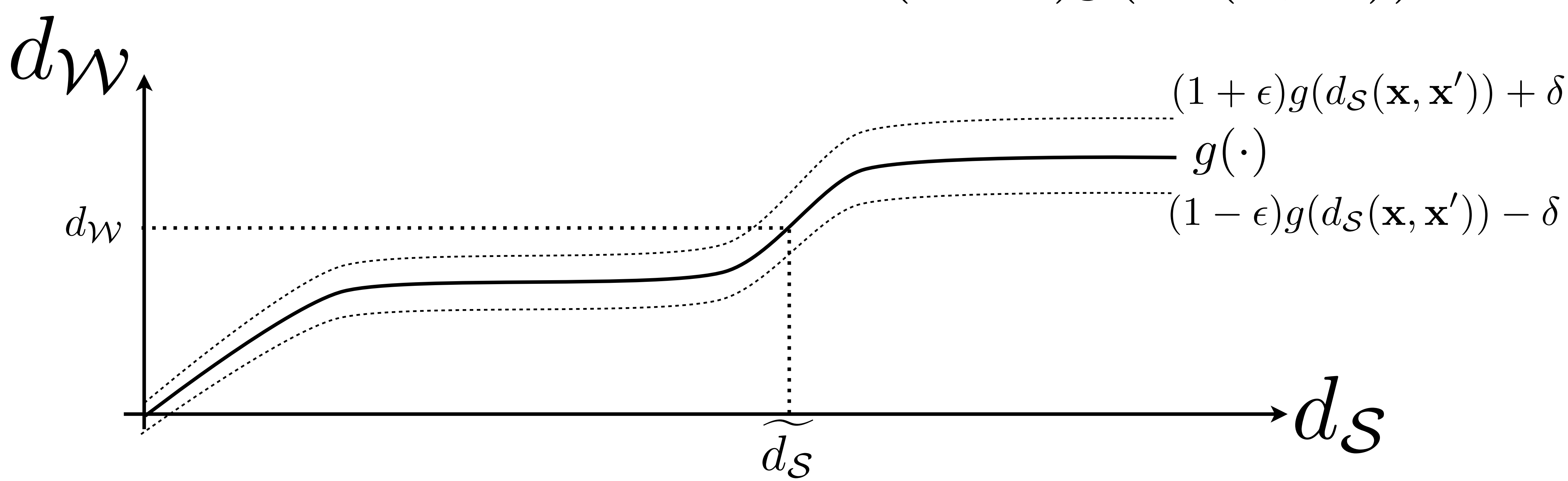
For all x,y in $\mathcal{S}$:

$$(1-\epsilon)g(d_{\mathcal{S}}(\mathbf{x},\mathbf{x}')) - \delta \leq d_{\mathcal{W}}(f(\mathbf{x}), f(\mathbf{x}')) \leq (1+\epsilon)g(d_{\mathcal{S}}(\mathbf{x},\mathbf{x}')) + \delta$$

Embedding Analysis [B, Rane '13a]

For all *x*,*y* in $\mathcal{S}$:

$$(1-\epsilon)g(d_{\mathcal{S}}(\mathbf{x},\mathbf{x}')) - \delta \leq d_{\mathcal{W}}(f(\mathbf{x}), f(\mathbf{x}')) \leq (1+\epsilon)g(d_{\mathcal{S}}(\mathbf{x},\mathbf{x}')) + \delta$$



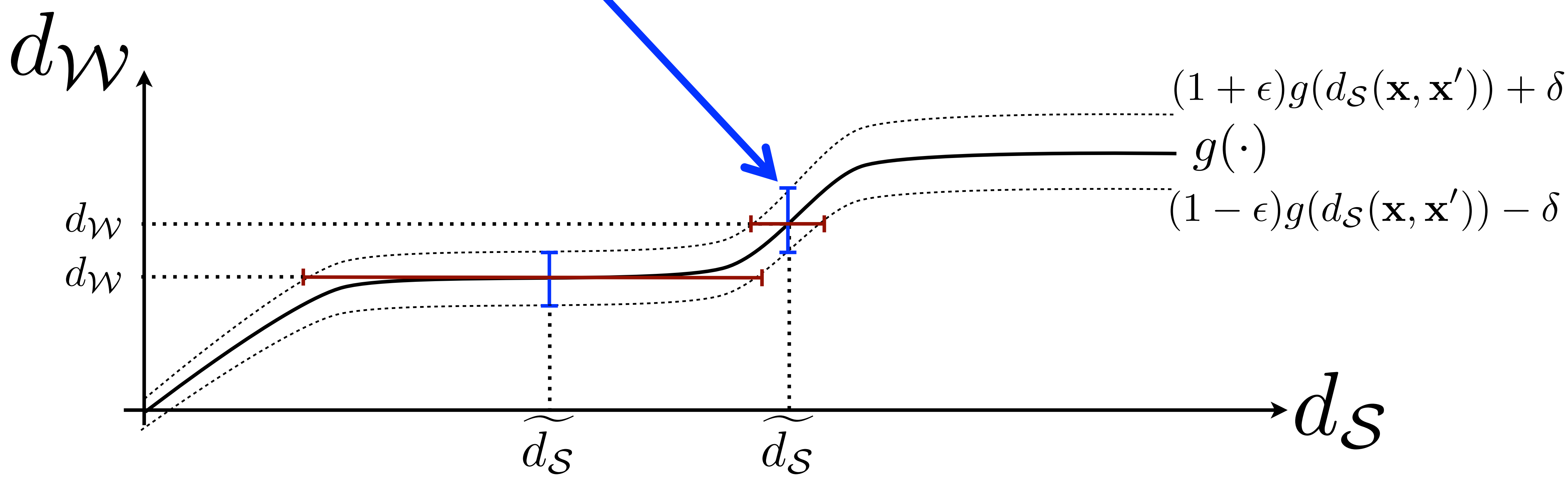
$$d_{\mathcal{W}}(f(\mathbf{x}), f(\mathbf{x}')) \approx g(d_{\mathcal{S}}(\mathbf{x},\mathbf{x}'))$$

$$\Rightarrow \widetilde{d_{\mathcal{S}}} = g^{-1}(d_{\mathcal{W}}(f(\mathbf{x}, f(\mathbf{x}'))$$

Embedding Analysis [B, Rane '13a]

For all *x*,*y* in $\mathcal{S}$:

$$(1-\epsilon)g(d_{\mathcal{S}}(\mathbf{x},\mathbf{x}')) - \delta \leq d_{\mathcal{W}}(f(\mathbf{x}), f(\mathbf{x}')) \leq (1+\epsilon)g(d_{\mathcal{S}}(\mathbf{x},\mathbf{x}')) + \delta$$



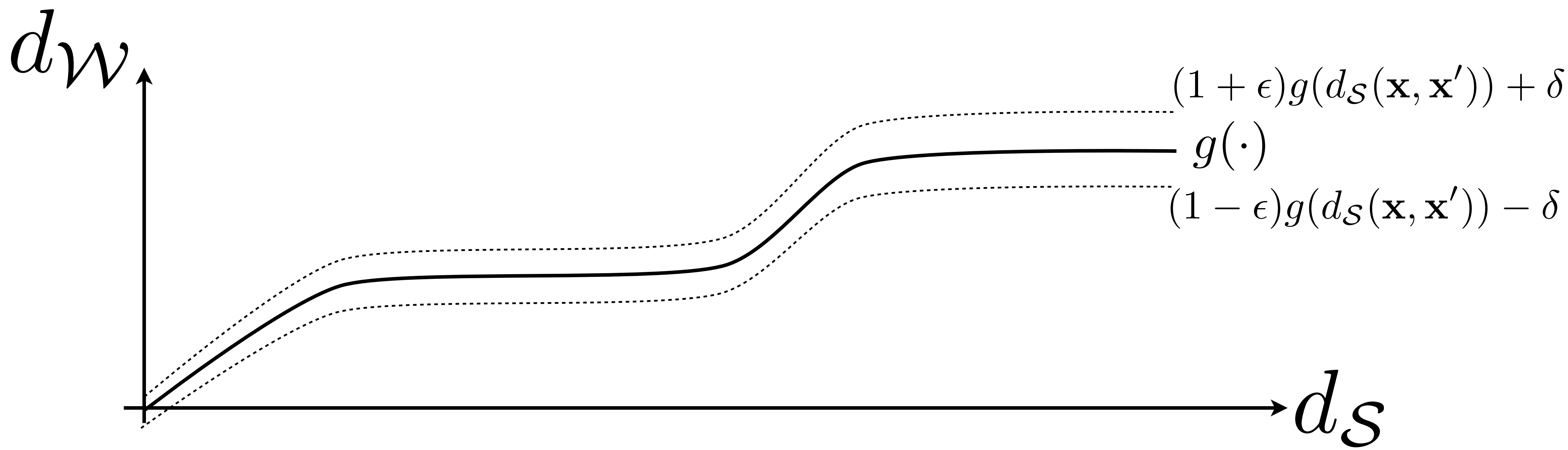
$$\left| d_{\mathcal{S}}(\mathbf{x},\mathbf{x}') - \widetilde{d_{\mathcal{S}}} \right| \lesssim \frac{\delta + \epsilon d_{\mathcal{W}}(f(\mathbf{x}), f(\mathbf{x}'))}{g'(\widetilde{d_{\mathcal{S}}})}$$

Accuracy depends on slope!

Embedding Analysis [B, Rane '13a]

For all x,y in $\mathcal{S}$:

$$(1-\epsilon)g(d_{\mathcal{S}}(\mathbf{x},\mathbf{x}')) - \delta \leq d_{\mathcal{W}}(f(\mathbf{x}), f(\mathbf{x}')) \leq (1+\epsilon)g(d_{\mathcal{S}}(\mathbf{x},\mathbf{x}')) + \delta$$



Can we achieve any distance map $g(\cdot)$?

No: $g(\cdot)$ must be ***sub-additive*** ($g(x+y) \lesssim g(x) + g(y)$)

Embedding Design [B, Rane '13b]

Q: Can we design embeddings?

A: Yes. We start with a **random matrix** $\mathbf{A} \in \mathbb{R}^{M \times N}$,

a **periodic function** $h(t) = h(t+1)$

and **random** i.i.d., uniform **dither** $\mathbf{w} \in [0, 1)$

$$\mathbf{y} = h(\mathbf{A}\mathbf{x} + \mathbf{w})$$

Fourier series coefficients of $h(\cdot)$: H_k

Also, assume bounded: $\bar{h} = \sup_t h(t) - \inf_t h(t)$

Distance Map

Can be relaxed...

$$\mathbf{A} \in \mathbb{R}^{M\times N}$$ i.i.d., Gaussian, variance σ^2

$$\mathbf{w} \in [0,1)$$ i.i.d, uniform

$$h(t) = h(t+1) \qquad \bar{h} = \sup_t h(t) - \inf_t h(t)$$

Fourier series coefficients of $h(\cdot)$: H_k

Resulting distance map:

$$g(d) = 2\sum_k |H_k|^2 \left(1 - e^{-\frac{1}{2}(\sigma d k)^2}\right)$$

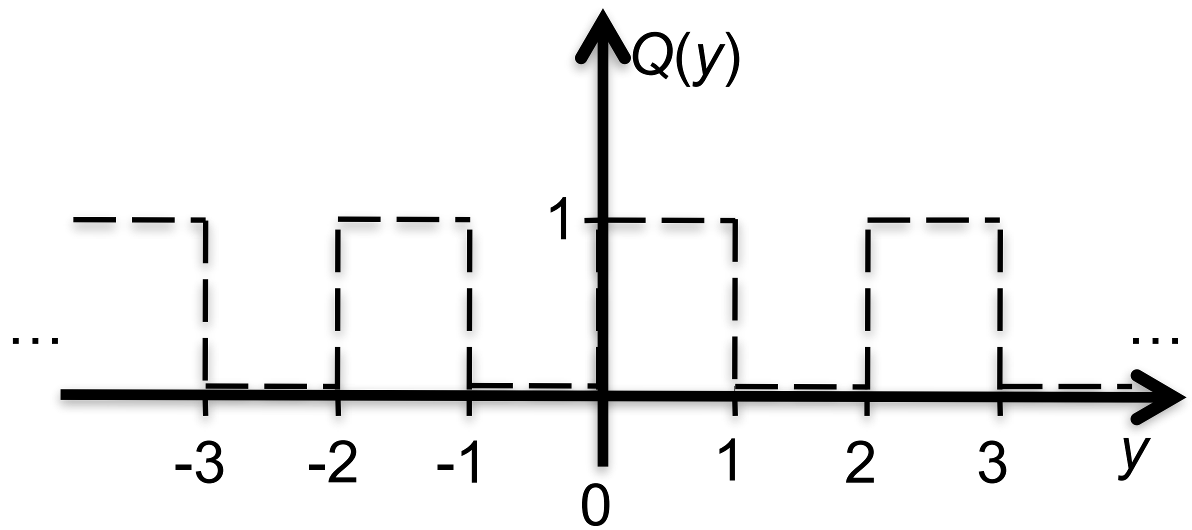
Theorem (Embedding Design) [B, Rane '13b]

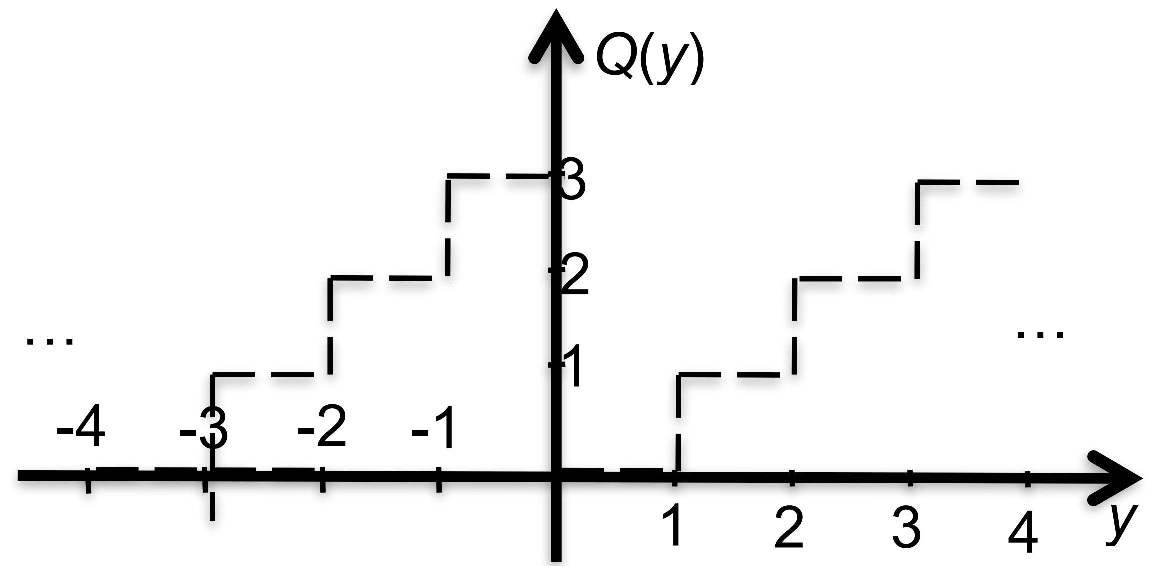
Consider a set $\mathcal{S}$ of Q points in $\mathbb{R}^N$, measured using $\mathbf{y} = h(\mathbf{A}\mathbf{x}+\mathbf{w})$, with $\mathbf{A}$, $\mathbf{w}$, and $h(t)$ as above. With failure probability $P_F \leq 2Q^2 e^{-2M\frac{\delta^2}{\bar{h}^4}}$ the following holds

$$g(\|\mathbf{x}-\mathbf{x}'\|_2) - \delta \leq \frac{1}{M}\|\mathbf{y}-\mathbf{y}'\|_2^2 \leq g(\|\mathbf{x}-\mathbf{x}'\|_2) + \delta$$

for all pairs $\mathbf{x},\mathbf{x}' \in \mathcal{S}$ and corresponding measurements $\mathbf{y},\mathbf{y}'$.

Example: Universal Quantization





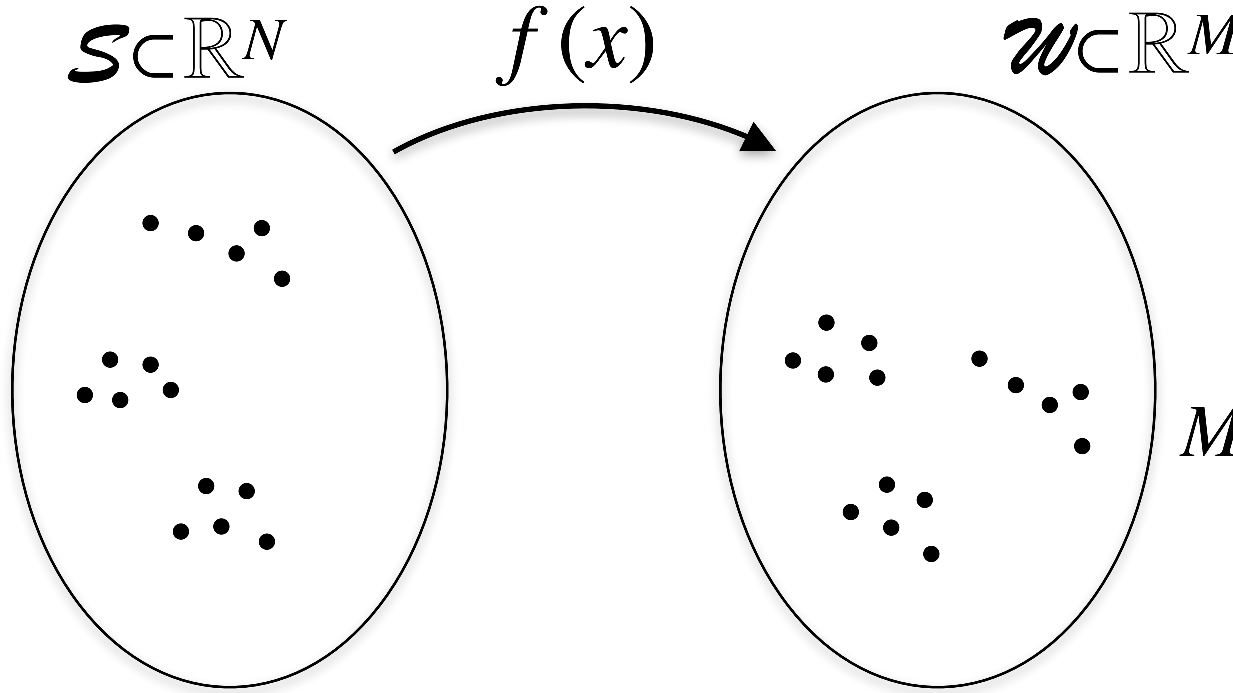
Quantizer design fits the analysis framework

$$q_m = Q\left(\frac{\langle \mathbf{x}, \mathbf{a}_m \rangle + w_m}{\Delta_m}\right), \qquad \mathbf{q} = Q(\Delta^{-1}(\mathbf{A}\mathbf{x} + \mathbf{w}))$$

$$\mathbf{y} = h(\mathbf{A}\mathbf{x} + \mathbf{w})$$

Embedding Properties [B, Rane '13a]

Original space
Distance metric: ℓ_2
P points in $\mathbb{R}^N$



Embedding space
Embed in $\{0,1\}^M$
Hamming distance
$M=O(\log P)$ dimensions

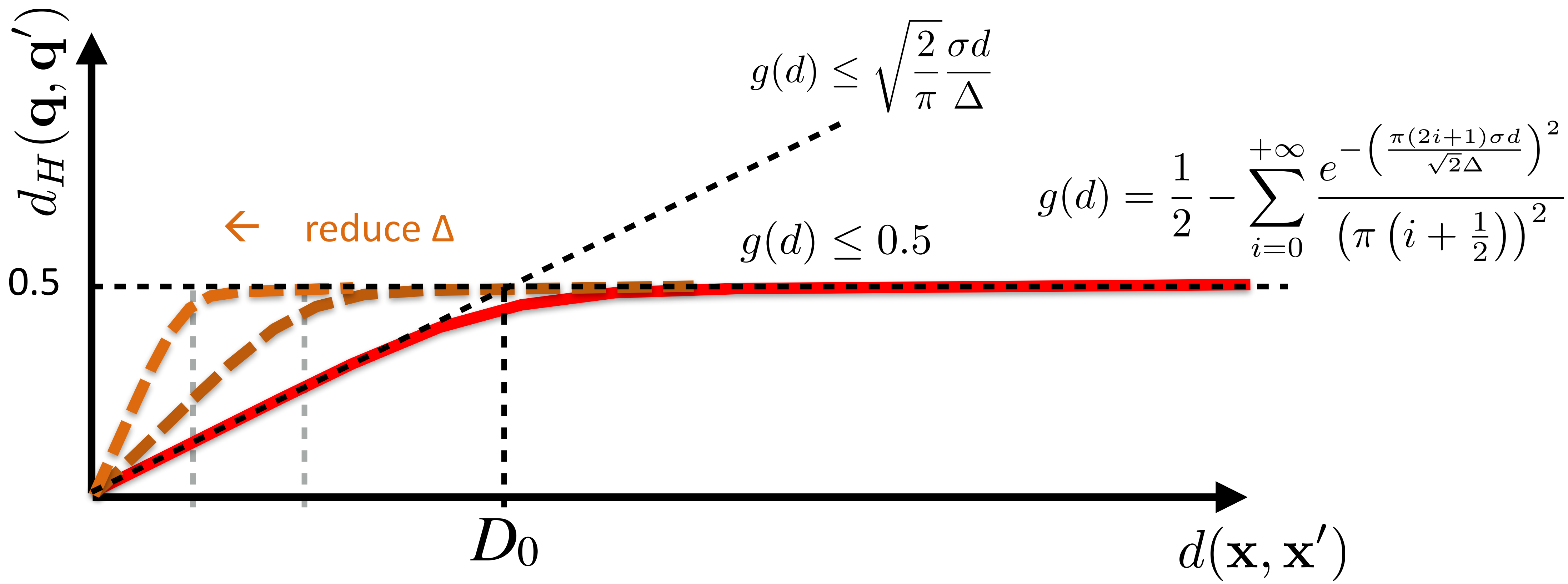
For all x,y in $\mathcal{S}$:

$$g(d)-\delta \leq d_H\left(f(x)-f(y)\right) \leq g(d)+\delta$$

$$g(d)=\frac{1}{2}-\sum_{i=0}^{+\infty}\frac{e^{-\left(\frac{\pi(2i+1)\sigma d}{\sqrt{2}\Delta}\right)^2}}{\left(\pi\left(i+\frac{1}{2}\right)\right)^2}$$

Error Behavior [B, Rane '13a]

$$g(d) - \delta \le d_H\left(f(x) - f(y)\right) \le g(d) + \delta$$

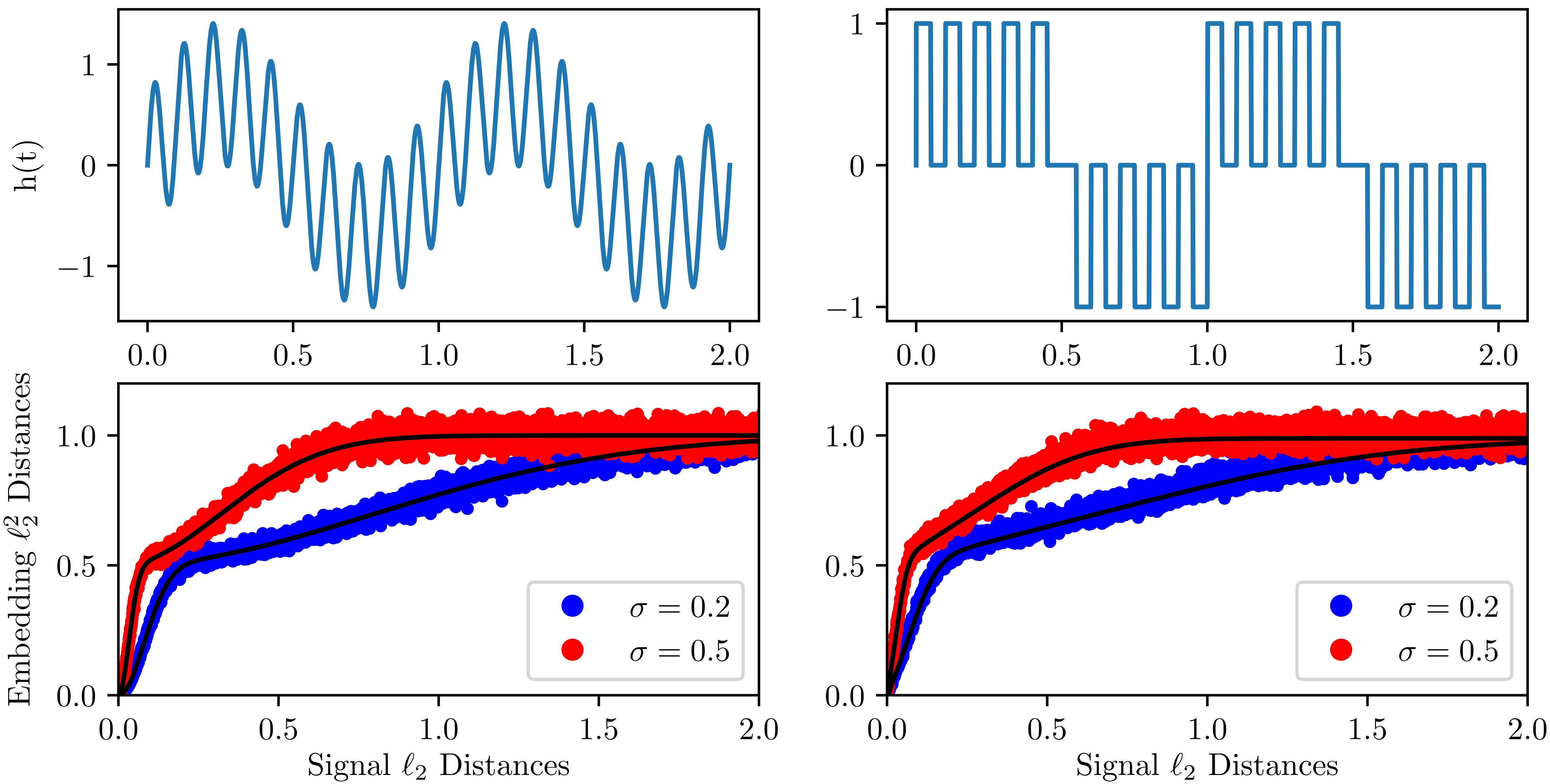


Distance estimate: $\widetilde{d} = g^{-1}\left(d_H\left(f(x), f(y)\right)\right)$

Estimate ambiguity: $\widetilde{d} - \frac{\delta}{g'(\widetilde{d})} \lesssim d \lesssim \widetilde{d} + \frac{\delta}{g'(\widetilde{d})}$

Properties (slope) controlled by choice of Δ

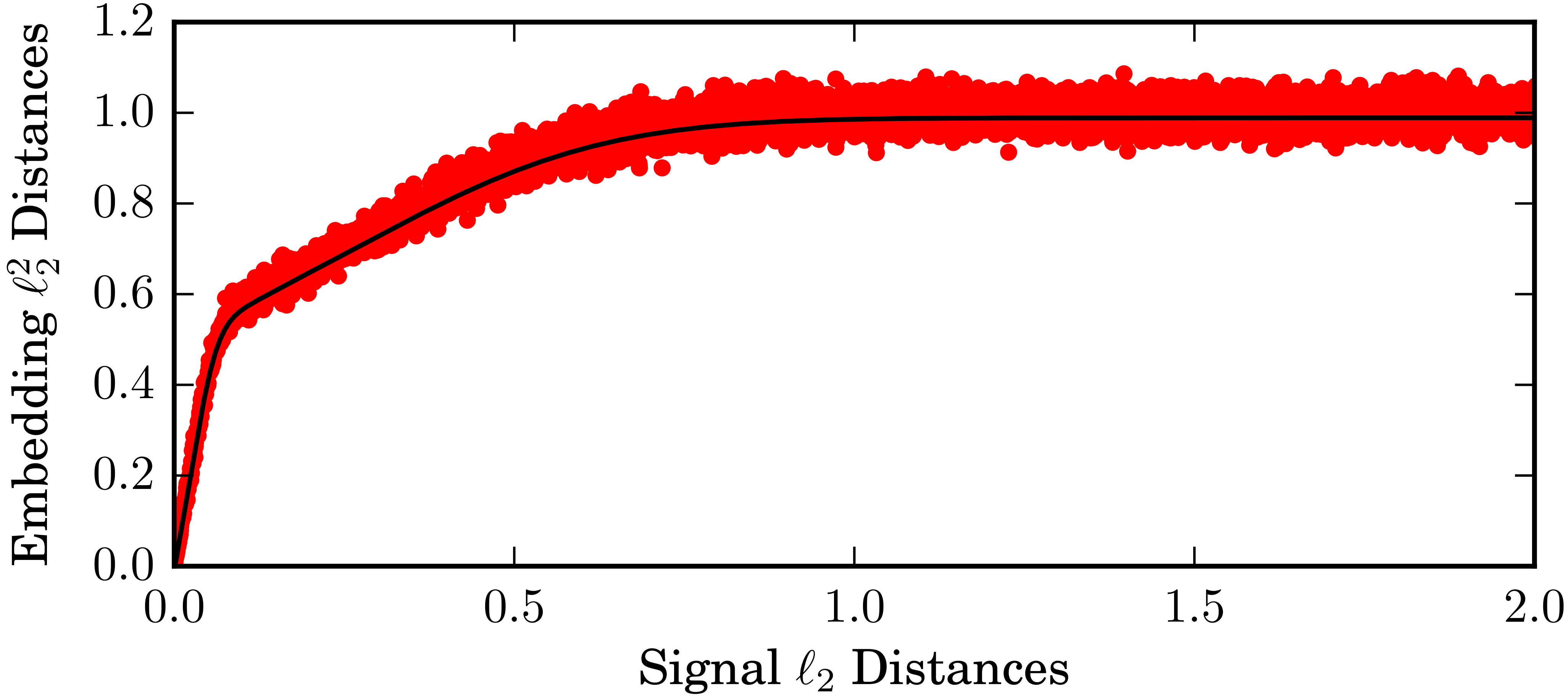
Other Examples

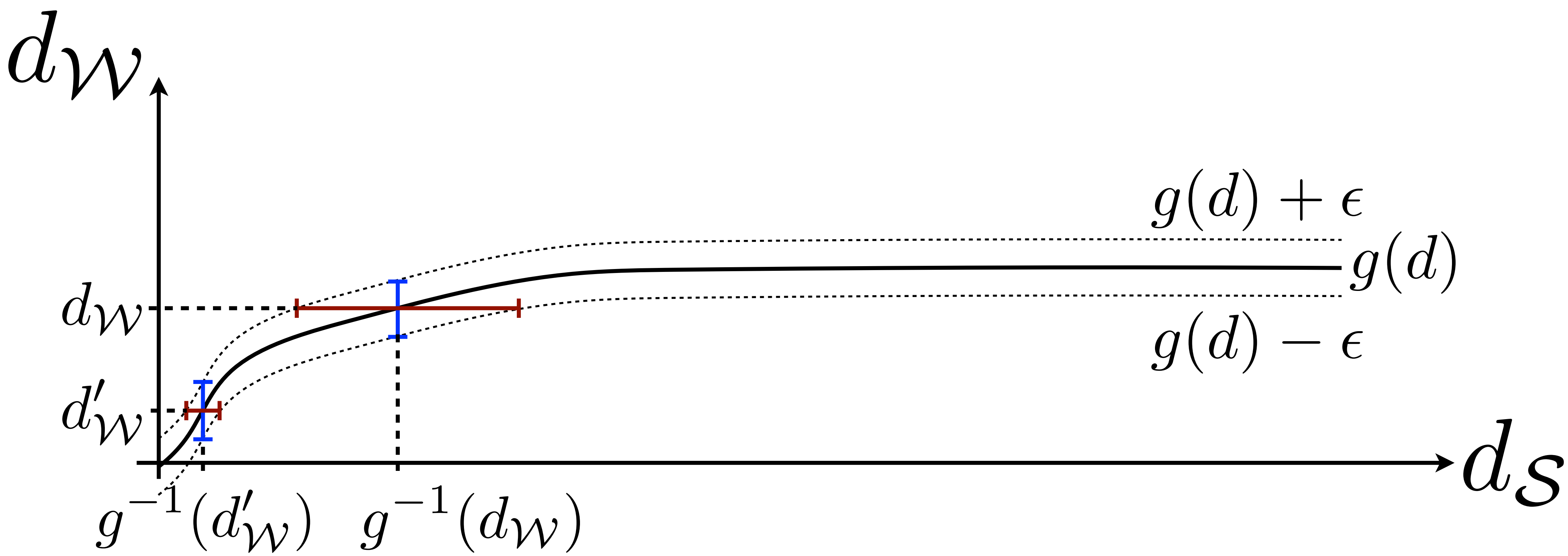


(a) Continuous Embedding
$h(t) = \frac{1}{\sqrt{2}}(\sin(2\pi t) + \sin(20\pi t))$

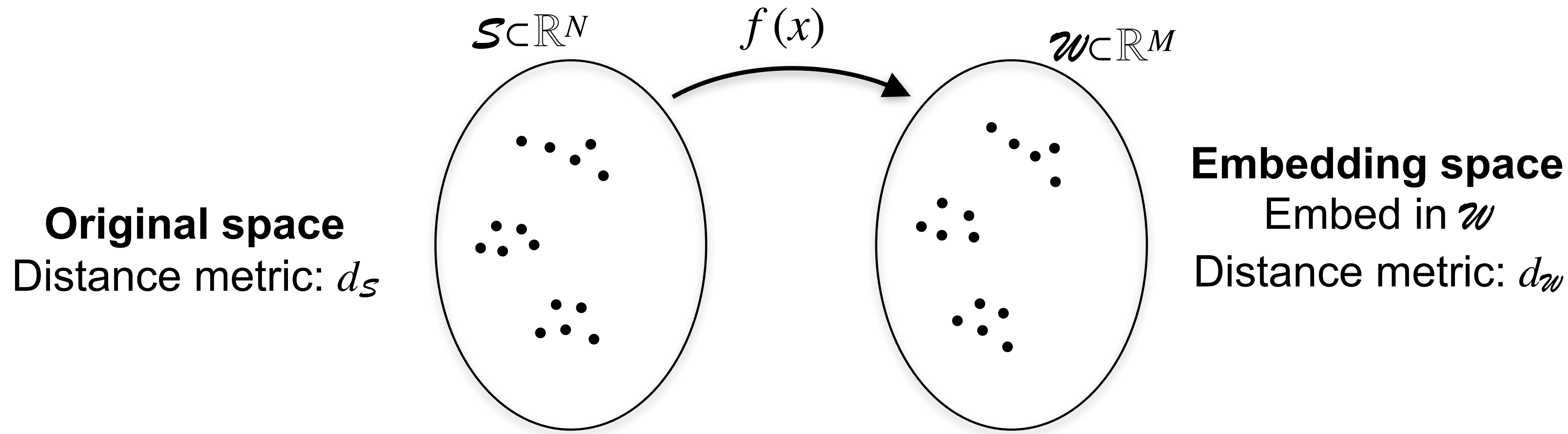
(b) Quantized Embedding
$h(t) = \frac{1}{2}(\text{sign}(\sin(2\pi t)) + \text{sign}(\sin(20\pi t)))$

Example Distance Ambiguity





Embedding Design: Comments



$$g(d_{\mathcal{S}}(\mathbf{x}, \mathbf{x}')) \approx d_{\mathcal{W}}(f(\mathbf{x}), f(\mathbf{x}'))$$

- Can this design achieve all possible $g()$?
  - Probably not! e.g., cannot use it for $g(d)=d$. General design still open problem.
- Quantization analysis from first part still applicable
  - In many cases, however, we can directly analyze a periodic quantized $h()$
- Theorem for embedding of point clouds; can easily extend to infinite *bounded* sets
  - E.g., manifolds, bounded sparse signals, etc.
- Using different pdf to generate A provides more flexibility
  - E.g., if drawn from Cauchy distribution, the embedding preserves ℓ_1 distance into ℓ_2
  - More generally, α-stable distributions can be used to embed arbitrary ℓ_p into ℓ_2

EMBEDDINGS AND ALTERNATIVE METRICS

- ℓ_1 Distances
- Angles/Inner Products
- Kernel Inner Products
- Lsh And Near Neighbors
- Classification

EMBEDDINGS AND ALTERNATIVE METRICS

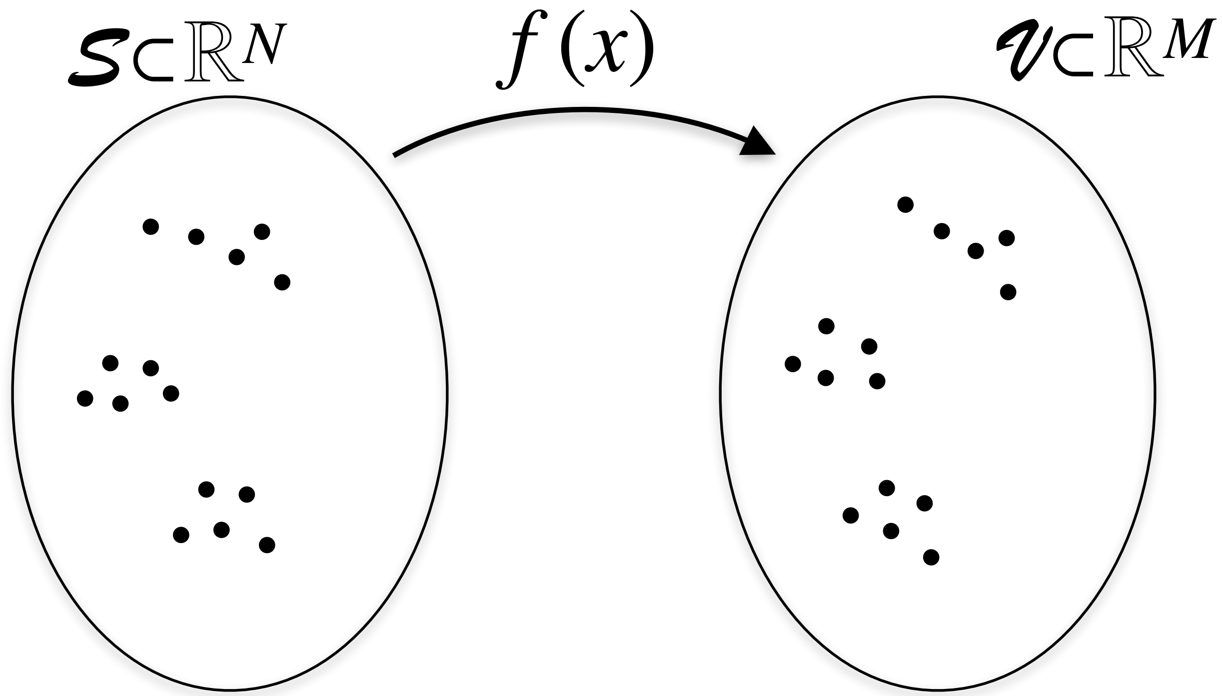
- ℓ_1 Distances
- Angles/Inner Products
- Kernel Inner Products
- Lsh And Near Neighbors
- Classification

ℓ_1 Distance Embedding

Original space

Distance metric: ℓ_1

P points in $\mathbb{R}^N$



Embedding space

Embed in $\mathbb{R}^M$

Distance metric: ℓ_1

Is a J-L style ℓ_1 embedding possible (i.e., $g(d)=d$)?

Generally NO! [Brinkman and Charikar '05]

Existing constructions:
looser guarantees on one side; error additive, not multiplicative [Indyk '00]

However, in some cases **we can trick it**

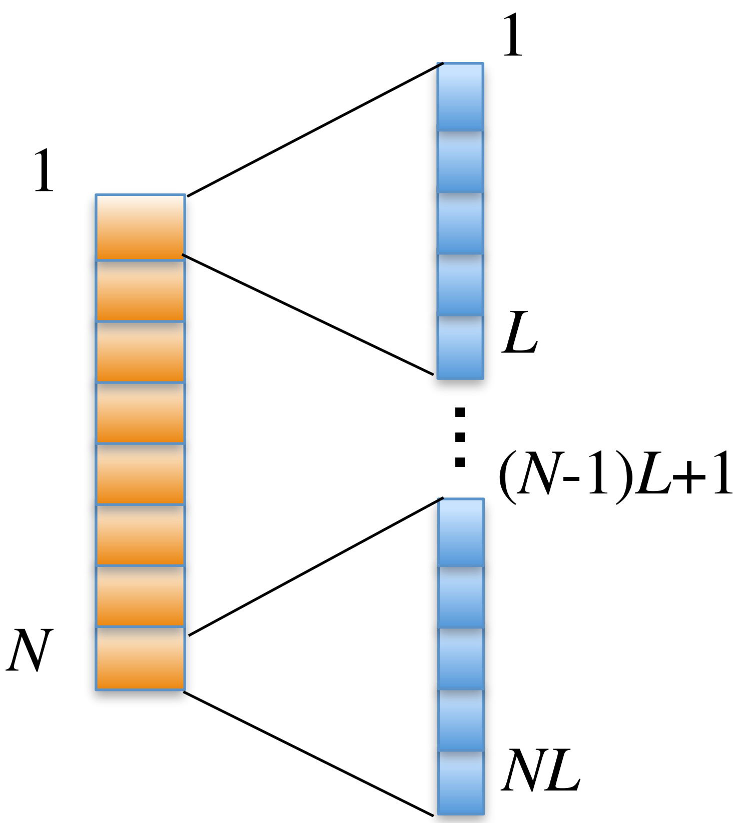
Approach: Map ℓ_1 to ℓ_2 and use ℓ_2 embeddings

*Note: embedding design from previous section can also be used to map ℓ_1 to ℓ_2, but cannot implement $g(d)=d$

ℓ_1 Distance Preservation

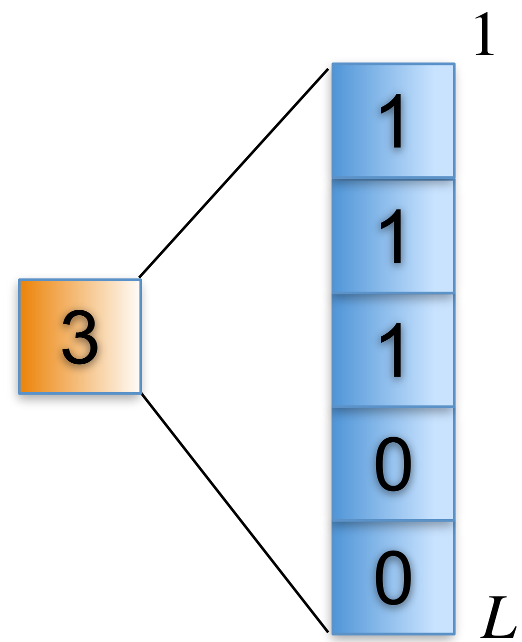
Assumption: integer (discrete) entries, bounded by L

Solution: perform L-times **dimension expansion**

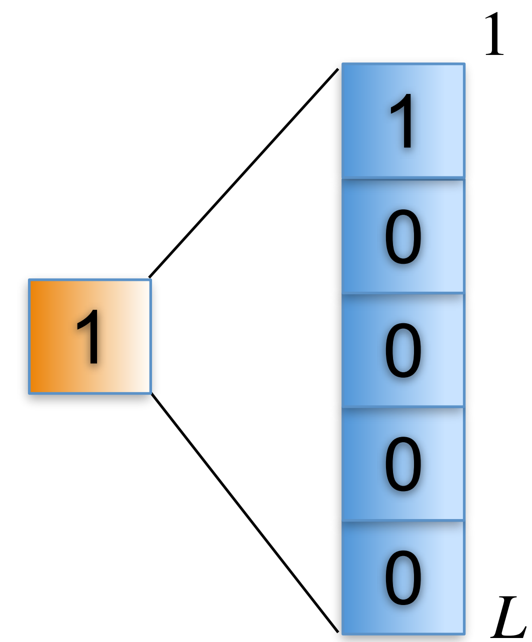


$$\mathbf{x} \longrightarrow \widetilde{\mathbf{x}}$$

Each coefficient x_n expanded to L dimensions: sequence of x_n ones followed by L-x_n zeros



$$x_n \longrightarrow \widetilde{\mathbf{x}}_n$$



$$x'_n \longrightarrow \widetilde{\mathbf{x}}'_n$$

$$|x_n - x'_n| = \|\widetilde{\mathbf{x}}_n - \widetilde{\mathbf{x}}'_n\|_2^2 \Rightarrow \|\mathbf{x} - \mathbf{x}'\|_1 = \|\widetilde{\mathbf{x}} - \widetilde{\mathbf{x}}'\|_2^2$$

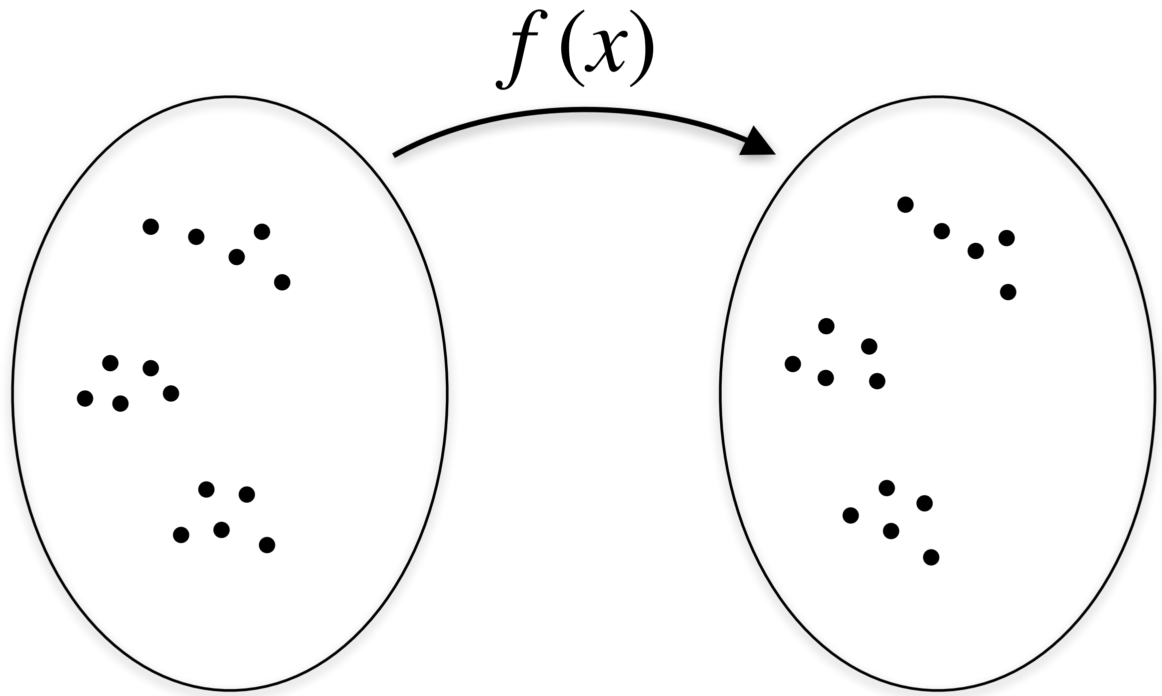
Is **dimensionality expansion a problem**?

No if J-L is used! M=O(log P), no dependence on N, or L

Other Distances

Original space

Distance metric: $d_{\mathcal{S}}$



Embedding space

Embed in $\mathcal{W}$

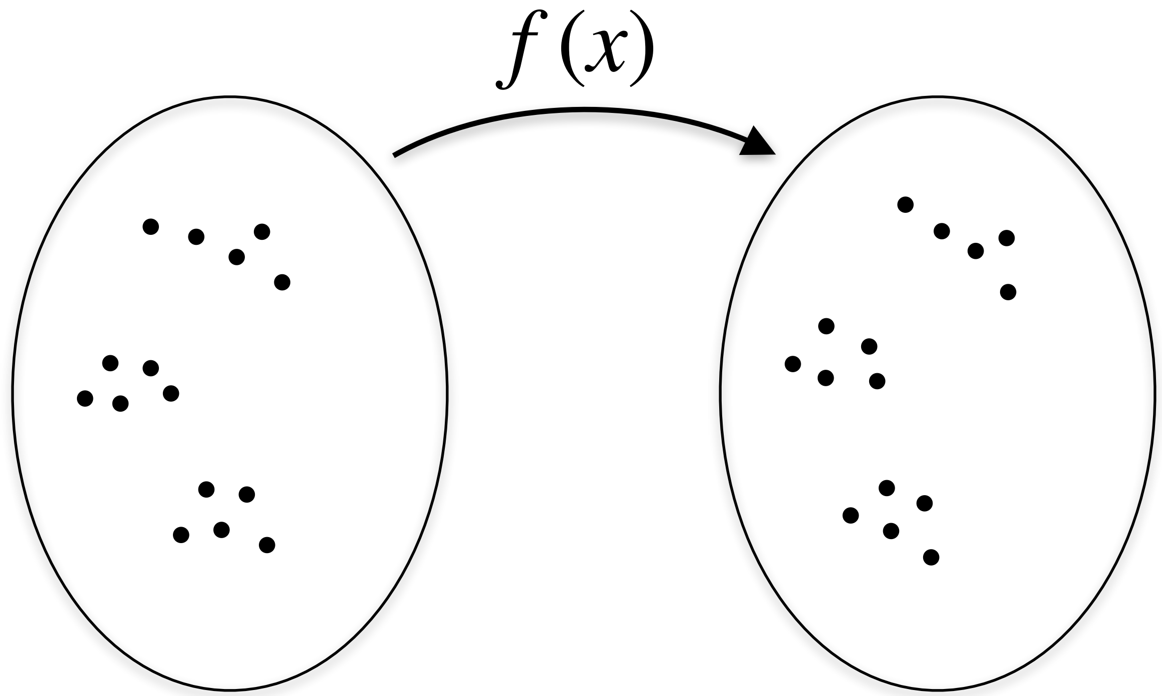
Distance metric: $d_{\mathcal{W}}$

- **General strategy**: map original distance to a space we know how to deal with
  - Often followed by a second (J-L style) dimensionality reduction in this space
- For the Edit Distance, Earth Mover's Distance (EMD), Shift metric:
  - Typical constructions map to ℓ_1 [Charikar et al. '02, '04,'06; Ostrovsky, Rabani '05; Cormode, Mutukrishnan '07; Andoni et al. '07;…]
  - May use $\ell_1 \rightarrow \ell_2$ mapping subsequently

Other Spaces And Functions

Original space

Distance metric: $d_{\mathcal{S}}$



Embedding space

Embed in $\mathcal{W}$

Distance metric: $d_{\mathcal{W}}$

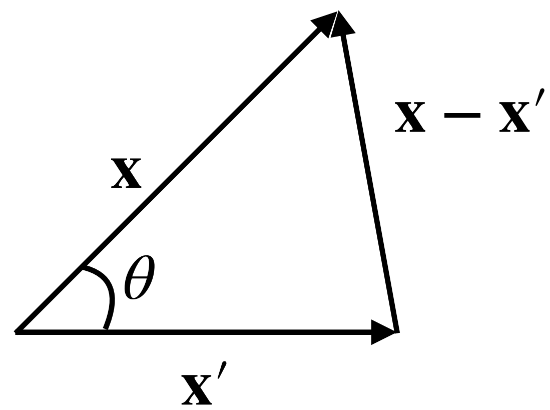
- Dynamical Systems and Tucken Embeddings [Eftekhari et al. ’17]
  - Embeddings that preserve information about the trajectory of a dynamical system
  - Embeddings preserve attractors of the dynamical system
  - Key result: delay-coordinate map (i.e., time samples of some states of the dynamical system for a fixed time window)

EMBEDDINGS AND ALTERNATIVE METRICS

- ℓ_1 Distances
- **Angles/Inner Products**
- **Kernel Inner Products**
- Lsh And Near Neighbors
- Classification

Angle/Inner Product Embeddings

If distances are preserved, we expect angles to be preserved as well!



$$\|\mathbf{x} - \mathbf{x}'\|_2^2 = \|\mathbf{x}\|_2^2 + \|\mathbf{x}'\|_2^2 - 2\langle \mathbf{x}, \mathbf{x}' \rangle$$

$$d_\angle(\mathbf{x}, \mathbf{x}') := \frac{1}{\pi}\theta = \frac{1}{\pi} \arccos \frac{\langle \mathbf{x}, \mathbf{x}' \rangle}{\|x\|_2 \|x'\|_2}$$

Given a J-L embedding, w/ ambiguity δ, can easily show

$$\left| \langle f(\mathbf{x}), f(\mathbf{x}') \rangle - \langle \mathbf{x}, \mathbf{x}' \rangle \right| \leq \delta \left(\|\mathbf{x}\|_2^2 + \|\mathbf{x}'\|_2^2 \right)$$

With a bit more care: $\left| \langle f(\mathbf{x}), f(\mathbf{x}') \rangle - \langle \mathbf{x}, \mathbf{x}' \rangle \right| \leq \delta \|\mathbf{x}\|_2 \|\mathbf{x}'\|_2$

[Davenport, B, Wakin, Baraniuk ’10]

If $\mathbf{x}, \mathbf{x}'$ are sparse and have the same support (JL→RIP):

[Haupt, Nowak ’07]

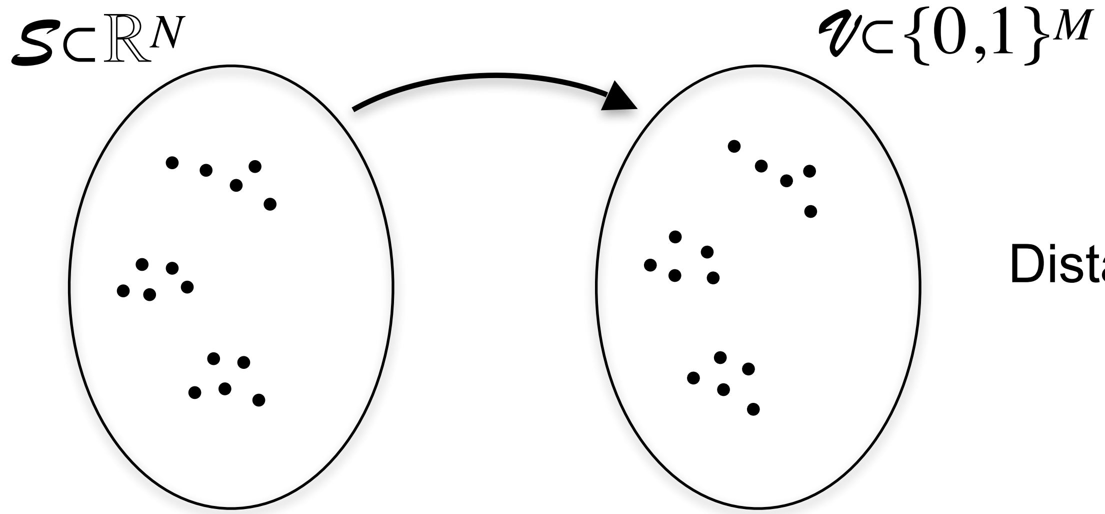
$$1 - \sqrt{3\delta} \leq \frac{d_\angle\left(f(\mathbf{x}), f(\mathbf{x}')\right)}{d_\angle(\mathbf{x}, \mathbf{x}')} \leq 1 + \sqrt{3\delta}$$

Recall: Binary Stable Embedding

$$f(\mathbf{x}) = \mathrm{sign}(\mathbf{A}\mathbf{x})$$

Original space
Distance metric: angle
(normalized inner product)

$$d_{\angle}(\mathbf{x}, \mathbf{x}') = \frac{1}{\pi} \arccos \frac{\langle \mathbf{x}, \mathbf{x}' \rangle}{\|x\|_2 \|x'\|_2}$$



Embedding space
Embed in $\{0,1\}^M$
Distance metric: normalized hamming distance

For all K-sparse $\mathbf{x}$, $\mathbf{x}$’ in $\mathbb{R}^N$:

$$d_{\angle}(\mathbf{x}, \mathbf{x}') - \delta \le d_H\left(f(\mathbf{x}), f(\mathbf{x}')\right) \le d_{\angle}(\mathbf{x}, \mathbf{x}') - \delta$$

using $M = O\left(\frac{1}{\delta^2}\left(K \log N + K \log \frac{1}{\delta}\right)\right)$ measurements

Binary Stable Embeddings are *angle* embeddings

Phase Instead of Sign [B '13]

Main idea: **phase** in $\mathbb{C}$ **generalizes sign** in $\mathbb{R}$

Q: How to obtain **phase from real signals?**

A: Measure with **complex measurement matrix**

$$\mathbf{A} \in \mathbb{C}^{M \times N},\ \mathbf{z} = \mathbf{Ax},\ \mathbf{y} = \angle(\mathbf{z}) = \angle\left(\mathbf{Ax}\right)$$

If $\mathbf{A}$ random, i.i.d. complex normal, **phase difference preserves angles**

$$E\left\{\left|\angle\left(\frac{z_m}{z'_m}\right)\right|\right\} = E\left\{\left|\angle\left(e^{i(y_m - y'_m)}\right)\right|\right\} = \pi d_\angle(\mathbf{x}, \mathbf{x}')$$

Resulting embedding guarantee:

$$\left|\frac{1}{M}\sum_m \left|\frac{1}{\pi}\angle\left(e^{1(y_m - y'_m)}\right)\right| - d_\angle(\mathbf{x}, \mathbf{x}')\right| \le \delta$$

using $\mathrm{O}(\log L)$ or $\mathrm{O}(K \log N/K)$ measurements

Bottom line: **Phase preserves angles, like signs do!**
(with similar additive ambiguity)

Quantization

$$\mathbf{A} \in \mathbb{C}^{M\times N},\ \mathbf{z} = \mathbf{A}\mathbf{x},\ \mathbf{y} = \angle(\mathbf{z}) = \angle\big(\mathbf{A}\mathbf{x}\big)$$

$\mathbf{A}$ i.i.d. Gausian ⇒ **phase unformly distributed:** $y_m \sim U(0, 2\pi)$

Optimal scalar quantizer **uniform, finite range:** $\Delta = \frac{\pi}{2^{B-1}}$

(using B bits per measurement)

$$\mathbf{A} \in \mathbb{C}^{M\times N},\ \mathbf{z} = \mathbf{A}\mathbf{x},\ \mathbf{y} = Q\left(\angle(\mathbf{z})\right) = Q\Big(\angle\big(\mathbf{A}\mathbf{x}\big)\Big)$$

$$\Downarrow$$

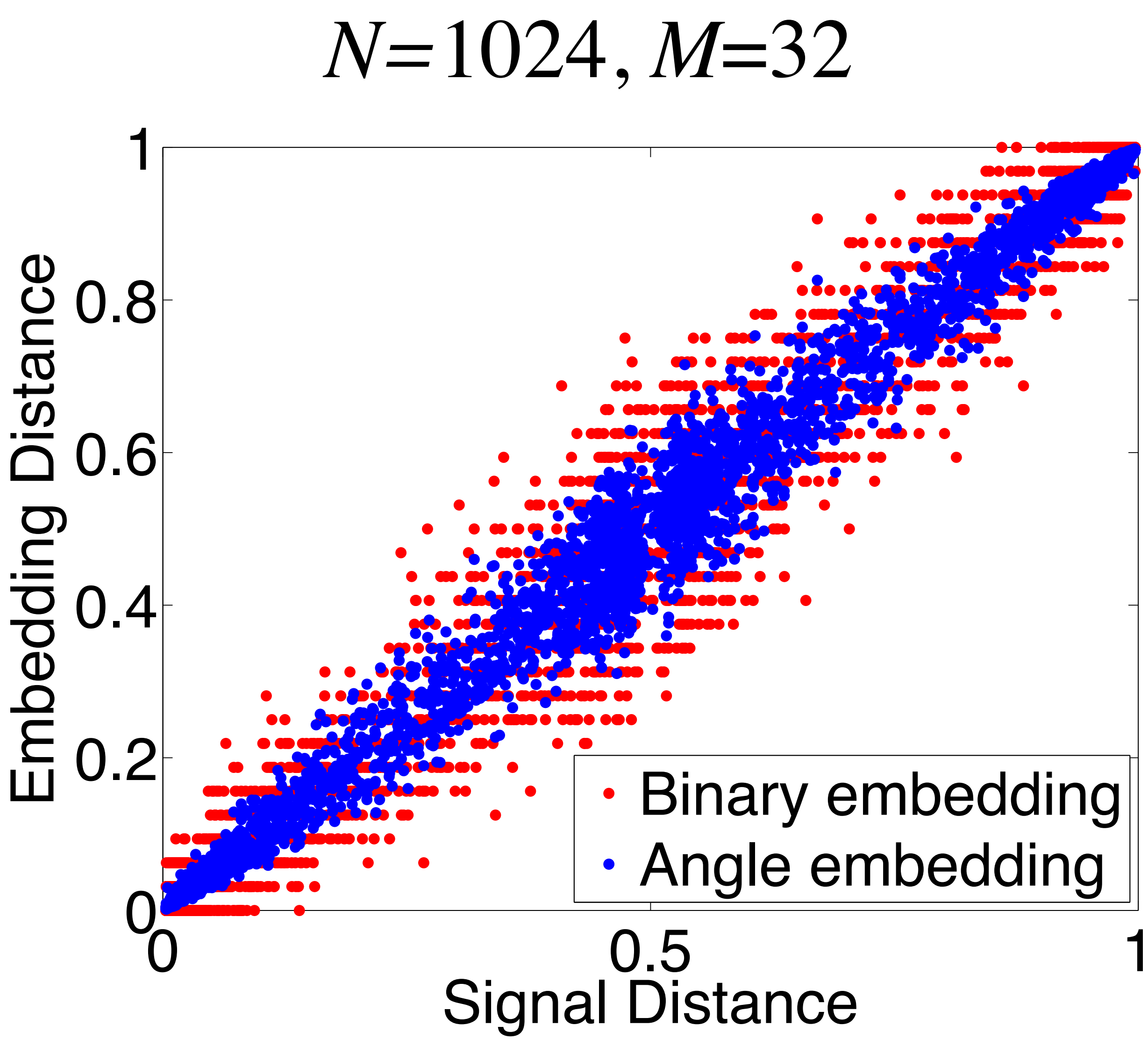
$$\left|\frac{1}{M}\sum_m \left|\frac{1}{\pi}\angle\left(e^{i(y_m - y'_m)}\right)\right| - d_\angle(\mathbf{x}, \mathbf{x}')\right| \leq \epsilon + 2^{-B+1}\pi$$

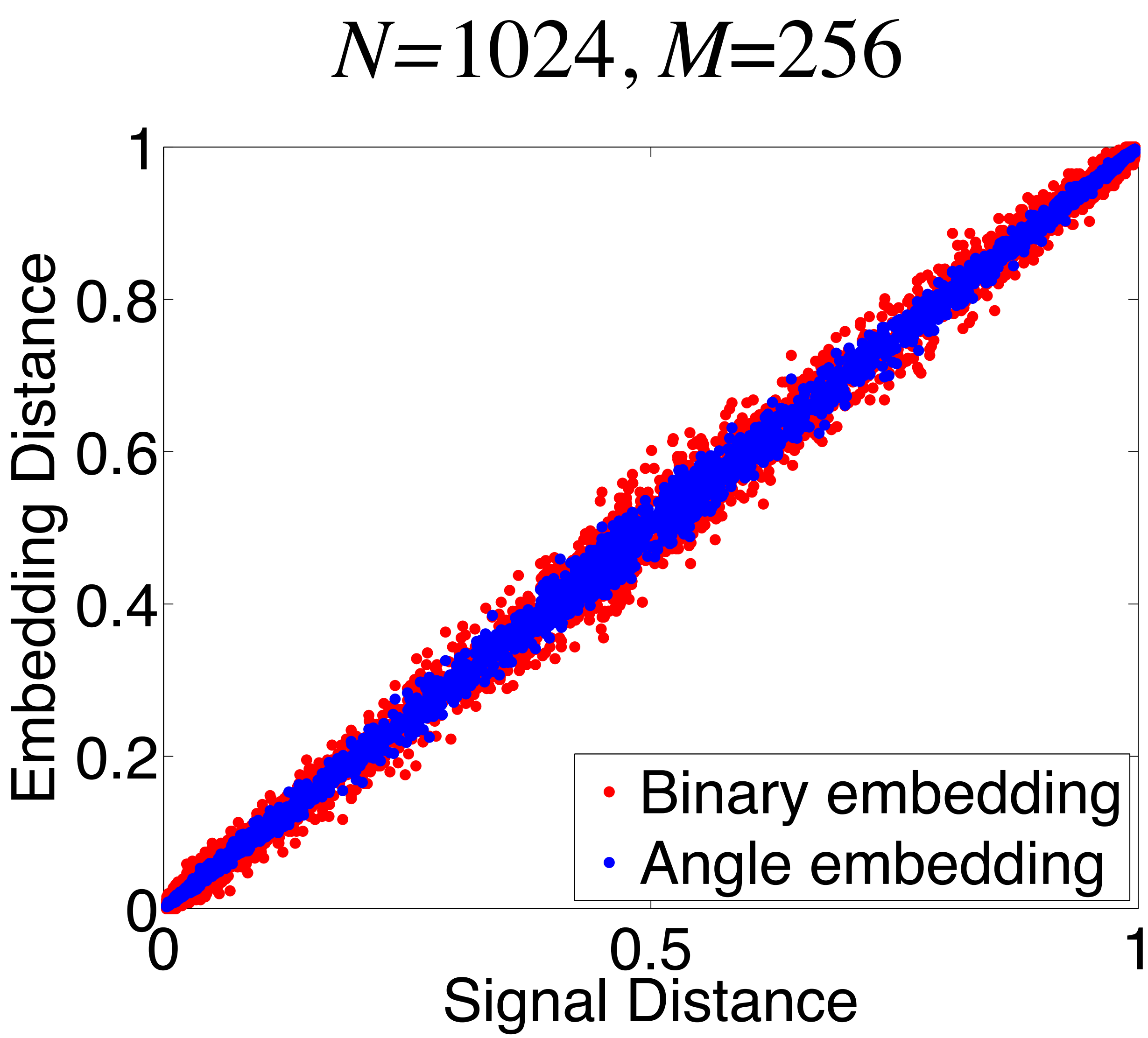
Total rate $R = MB$ using M measurements

Trade-off **embedding error** vs. **quantization error**

$$\epsilon = O\left(\frac{1}{\sqrt{M}}\right) \qquad \Delta = O\left(2^{-B}\right) = O\left(2^{-\frac{1}{M}}\right)$$

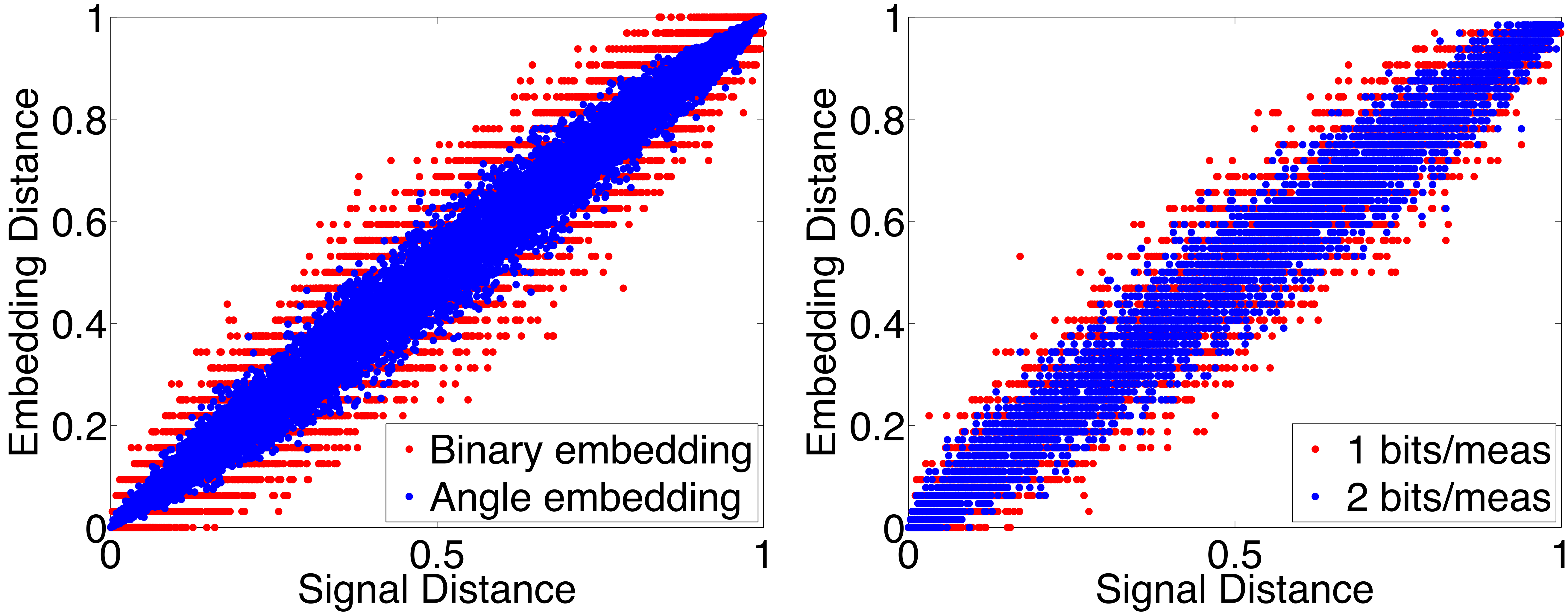
Comparison w/ BeSE



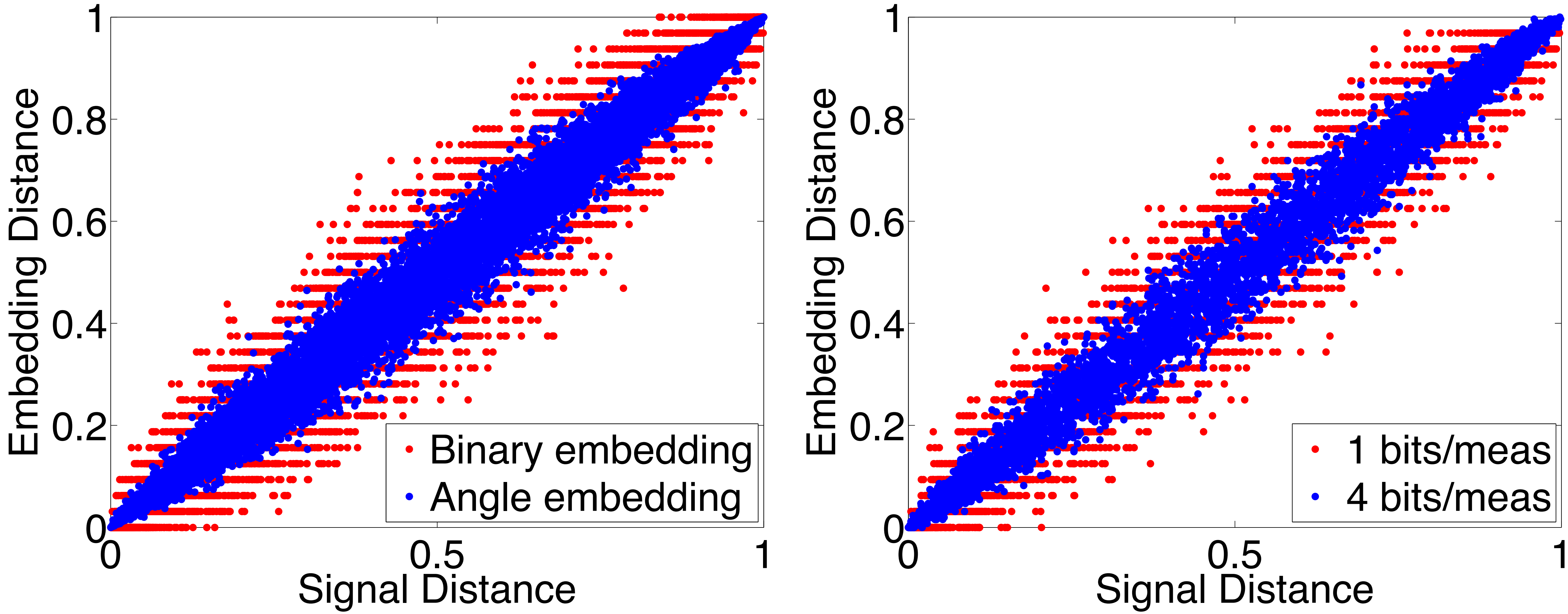


Phase **Angle Embedding is tighter** (as expected: it is analog)
For **smaller angles, tighter embedding** (suggests theory gap)

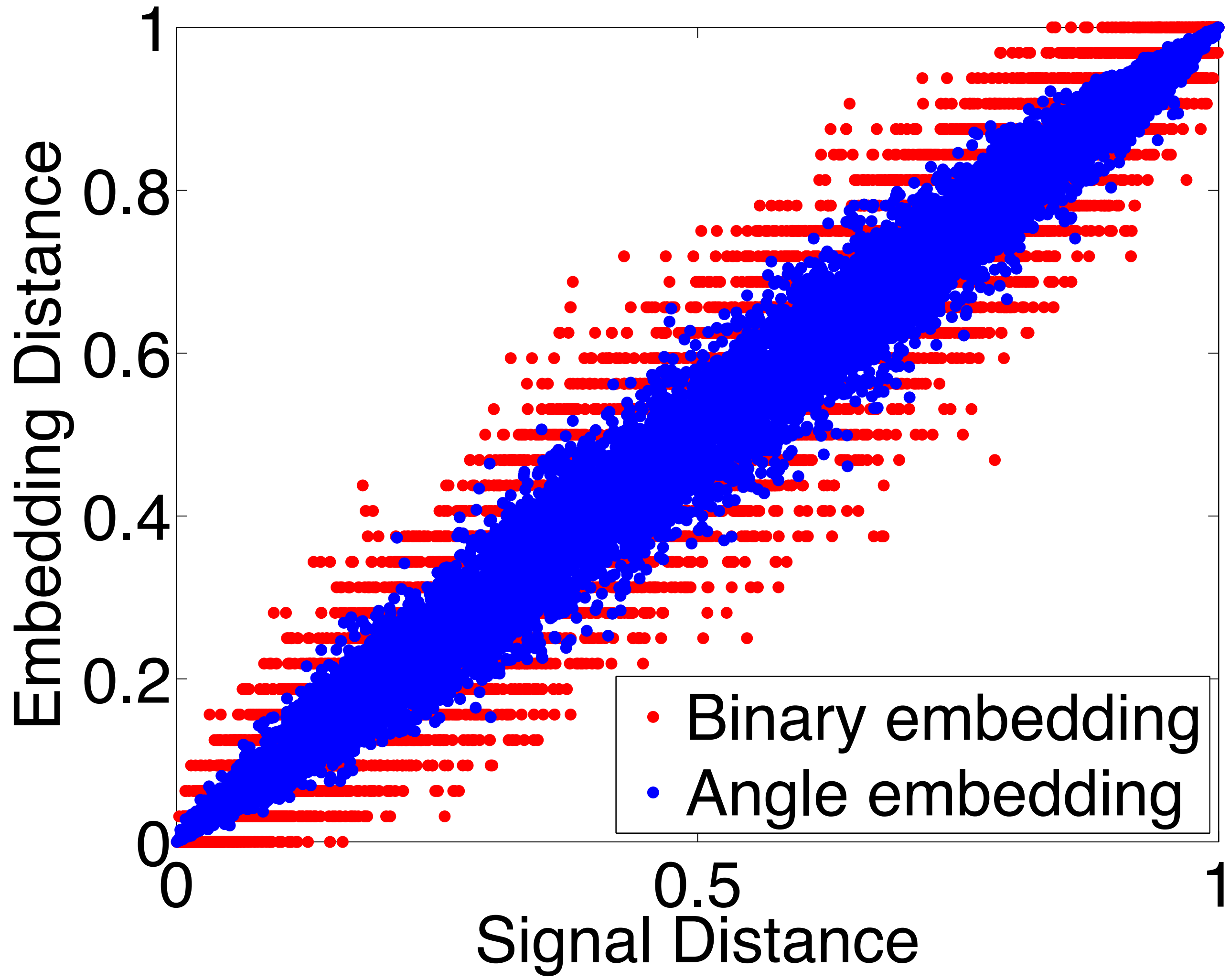
Quantization Effects

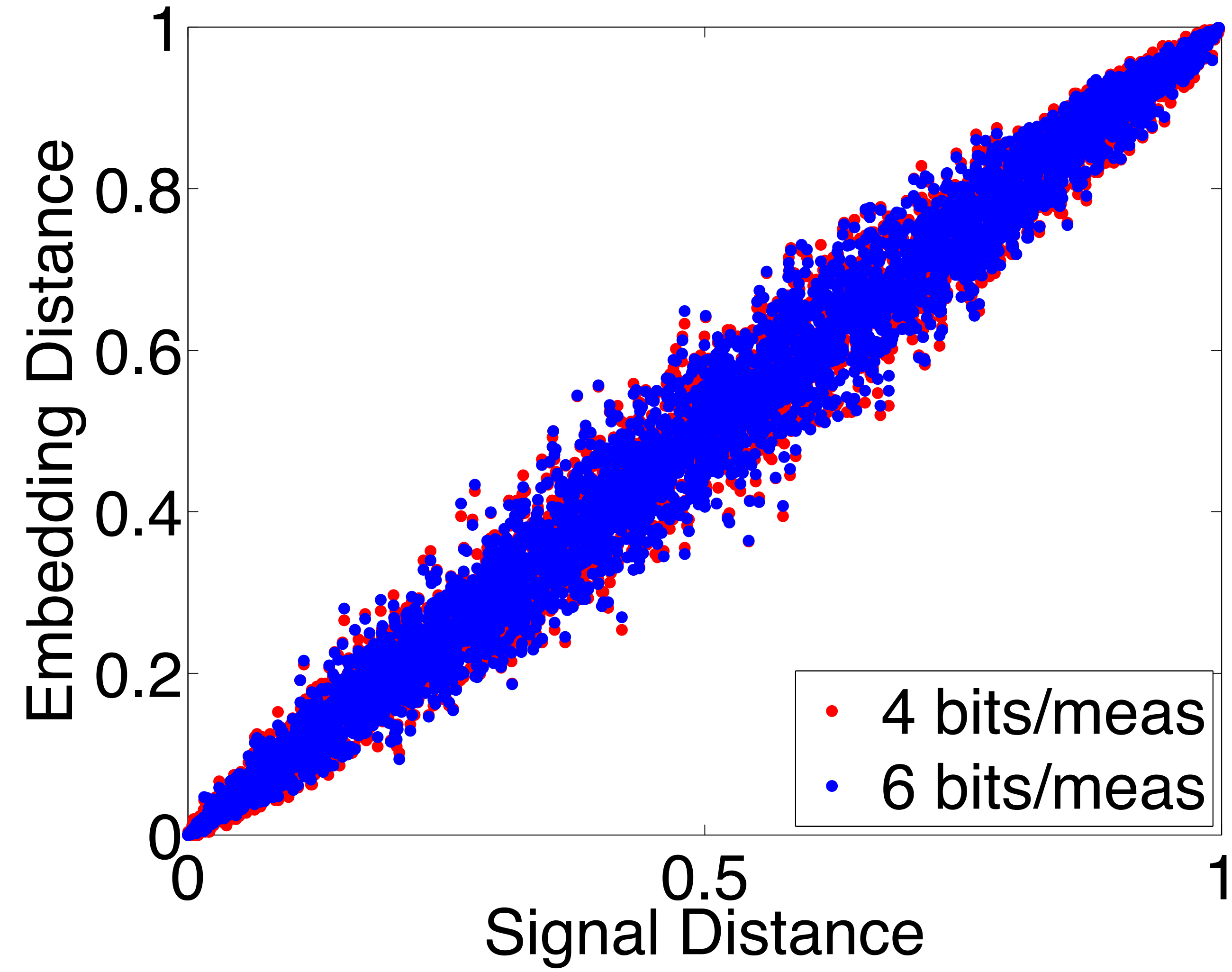


Quantization Effects



Quantization Effects





Benefit of increasing B is marginal

(Shift-Invariant) Inner Product Kernels [B, Mansour, Rane '16]

Can embeddings preserve kernel inner products? $K(\mathbf{x}, \mathbf{x}') = \langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle$

Yes. Using the same design as before

Let $f(\mathbf{x}) = h(\mathbf{A}\mathbf{x} + \mathbf{e})$ as before, with $\mathbf{y} = f(\mathbf{x})$. The kernel function

$$K(\mathbf{x}, \mathbf{x}') = \frac{1}{2M}\mathbf{y}^T\mathbf{y}' \qquad (1)$$

is shift invariant and approximates the radial basis function

$$K(\mathbf{x}, \mathbf{x}') \approx \frac{1}{2} - g\left(\|\mathbf{x} - \mathbf{x}'\|_2\right), \qquad (2)$$

with $g(d)$, as before.

Special case: $h(t)=\cos(t)$ → Random Fourier Features
(first instance of Kernel inner product embeddings)
[Rahimi, Recht '07]

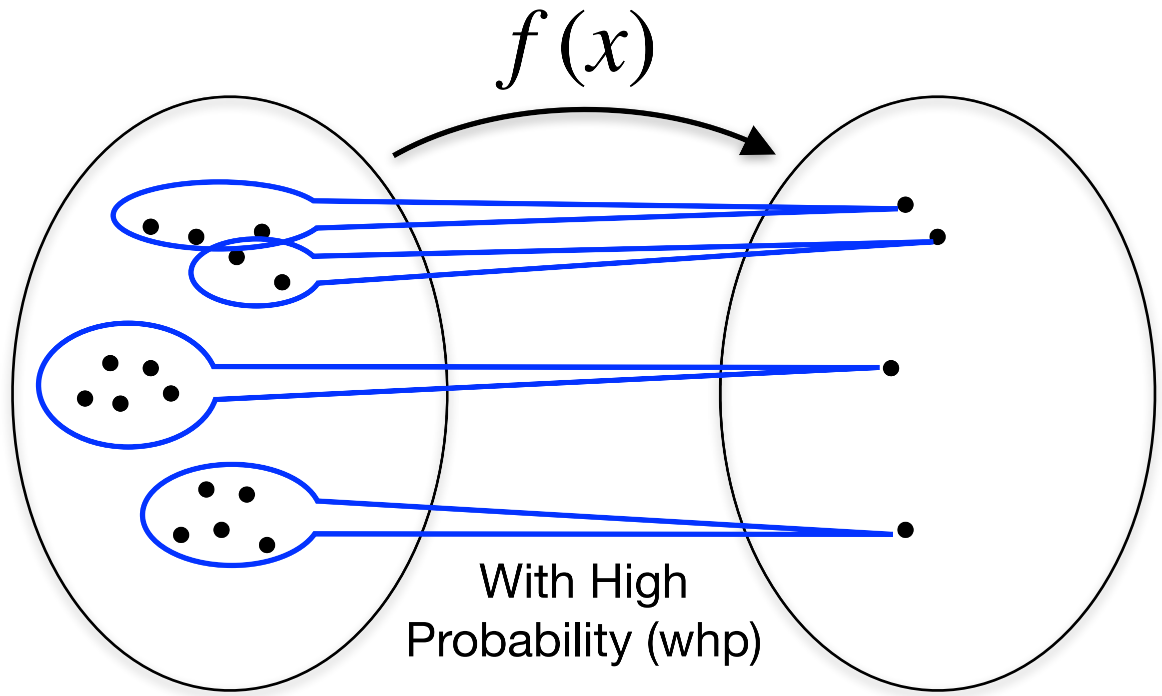
In other words: computing the standard inner product of the embedding is equivalent to computing the Kernel inner product on the data.

EMBEDDINGS AND ALTERNATIVE METRICS

- ℓ_1 Distances
- Angles/Inner Products
- Kernel Inner Products
- **Lsh And Near Neighbors**
- Classification

Locality Sensitive Hashing (LSH) [Indyk, Motwani '98]

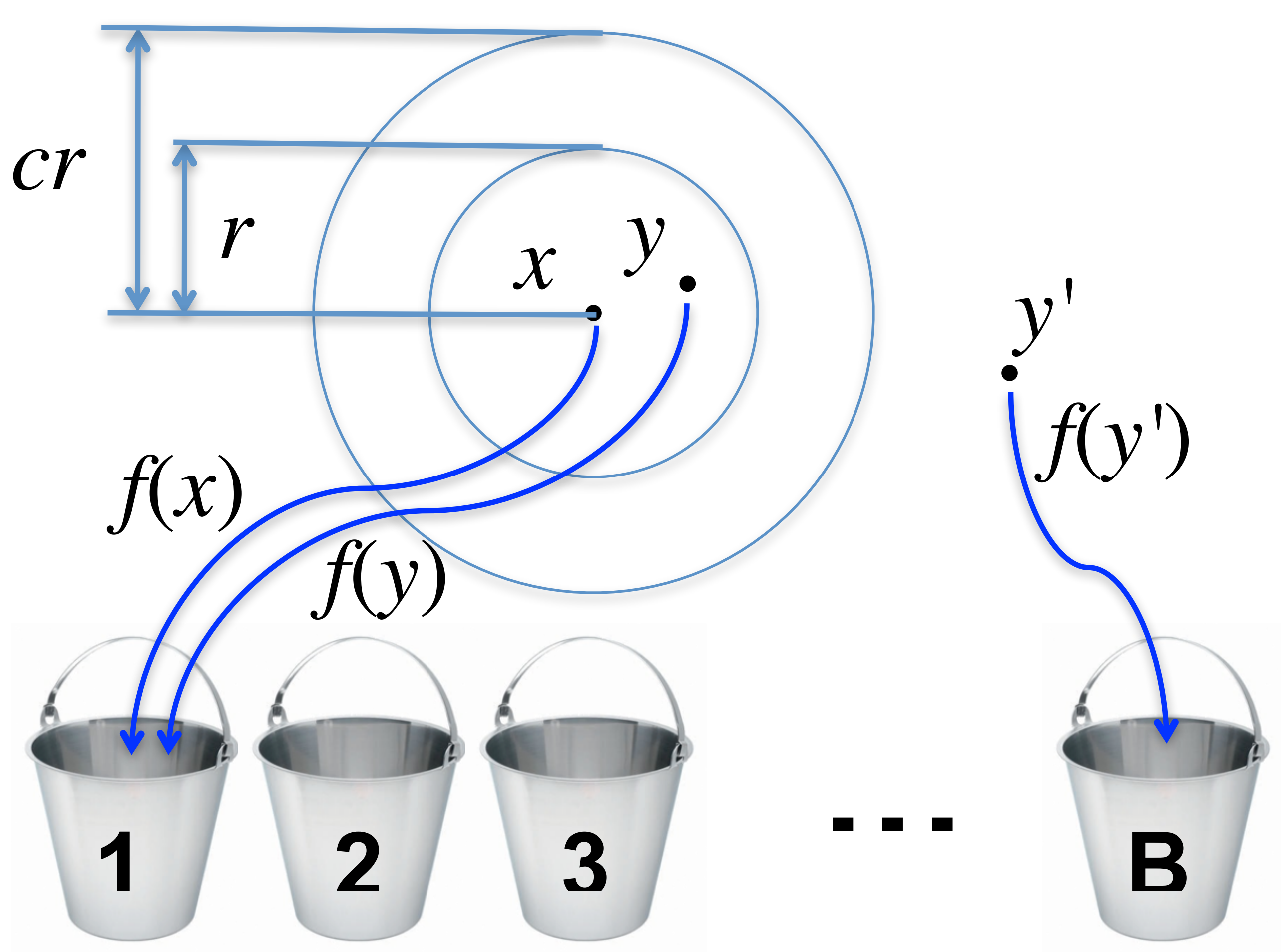
Original space
Distance metric: ℓ_2
P points in $\mathbb{R}^N$



Embedding space
Embed in $\{0,1\}^M$
Small M
Discrete space

- **Goal**: Speeding up Nearest Neighbor Search
- **Idea**: each signal in the space, compute a binary quantity with few bits, i.e., a "**hash**"
- Typical language in this literature: signals are **hashed** into "**buckets**"
  – When looking for near neighbors of a signal, compute it's hash and look only in that **bucket**
- If two signals have the same hash, then they are similar with high probability
  – The guarantee only goes one way: two signals might be similar but have very different hashes. On the other hand, if they have the same hash they probably they are similar.
  – Hash "distance" may not have any meaning
  – Might not find nearest neighbor, but will find a near one (*approximate* nearest neighbors)

Locality Sensitive Hashing [Indyk, Motwani, Andoni, et al]



Randomized signal hash $f: \mathbb{R}^N \to \mathbb{N}$ such that:

$d(x,y) \leq r \Rightarrow f(x) = f(y)$ with high probability
$d(x,y) \geq cr \Rightarrow f(x) \neq f(y)$ with high probability

No guarantee for $r \leq d(x,y) \leq cr$

Simplest (and most popular) **approach**

$$f(x) = \text{sign}(Ax)$$

- Not optimal hash, but simple to compute
  - Optimal LSH based on Leech lattice
- Assumes normalized signals
- Happens to also provide more information
  - Based on BeSE guarantees: embeds angles into Hamming distance
  - However, embedding not very accurate (very few measurements)

Algorithm:

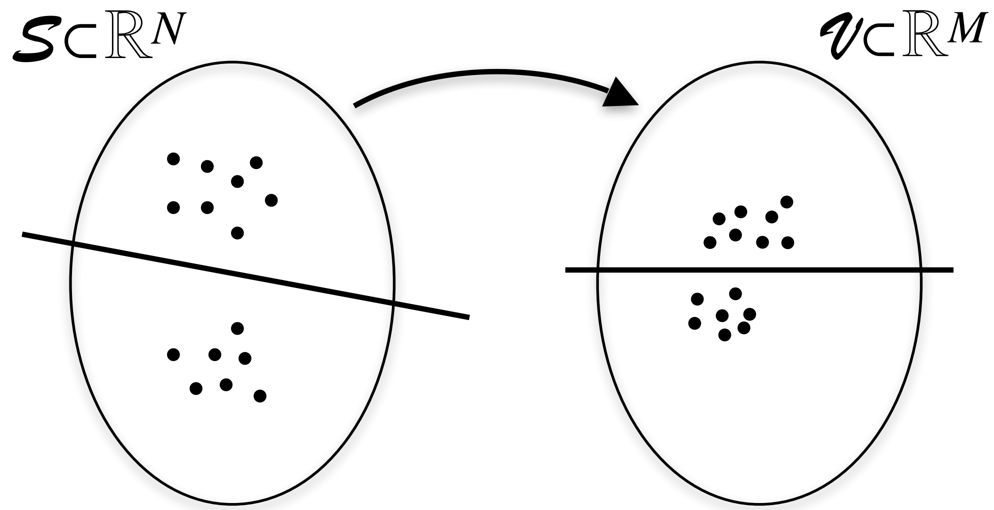
- **Preparation**: Hash all your signals into buckets. Each bucket has a list of signal with this hash
- **Execution**: Given signal x and it's hash, determine corresponding bucket. Signals in that bucket are approximate near neighbors of x.

EMBEDDINGS AND ALTERNATIVE METRICS

- ℓ_1 Distances
- Angles/Inner Products
- Kernel Inner Products
- Lsh And Near Neighbors
- **Classification**

Classification

Original space
Distance metric: ℓ_2
P points in $\mathbb{R}^N$
Linearly separable



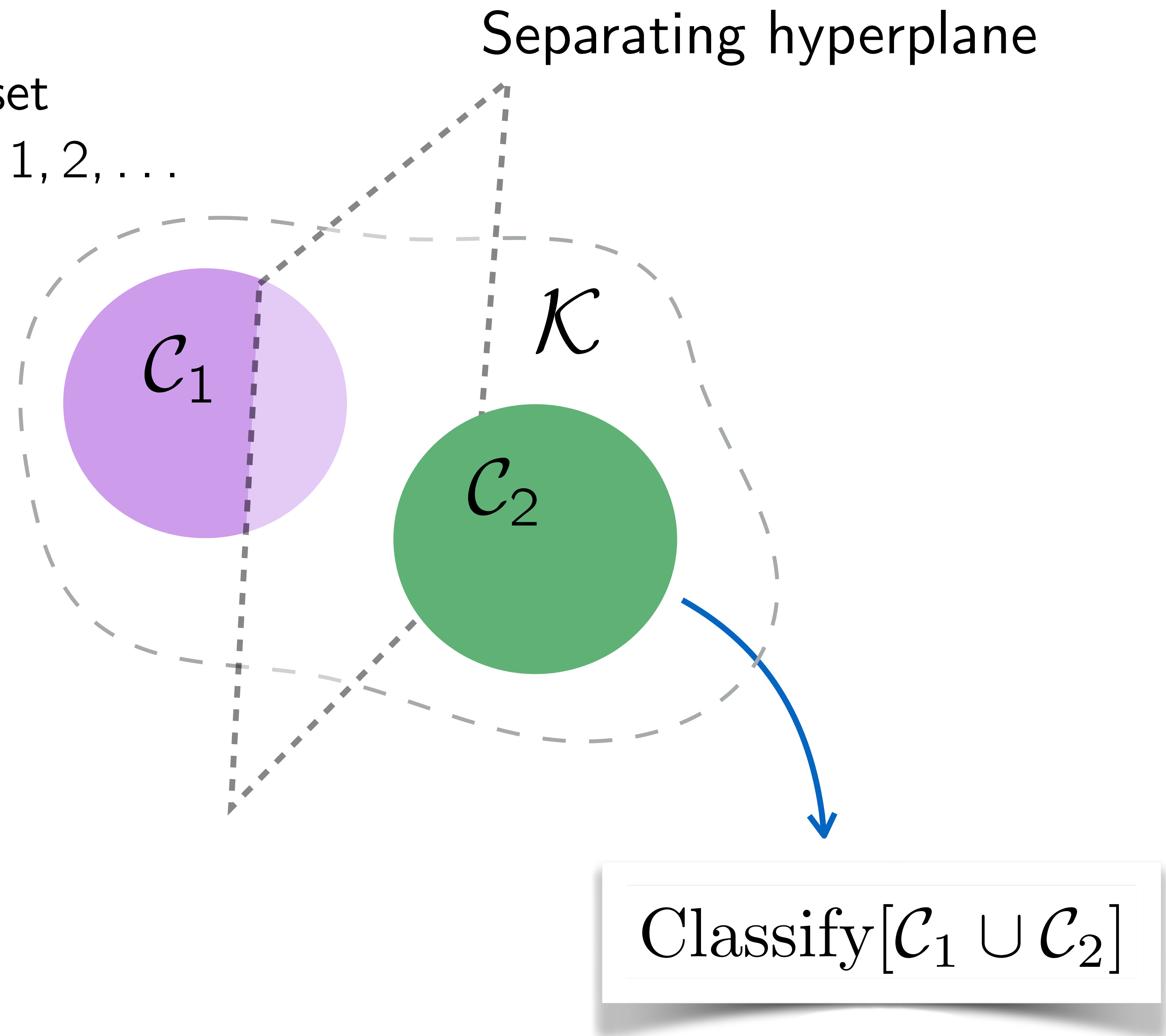
Embedding space
Distance metric: ℓ_2
Points still linearly separable

- **Goal**: Dimensionality reduction that respects linear boundaries/classification
- **Main approach**: Random projections, i.e., JL-style embeddings
- **Fundamental Question**: Given the geometry and separation of clusters, how much can we reduce dimension?
  - Secondary question: Can we quantize the projection and still preserve linear separability?

Linear Classification Big Picture

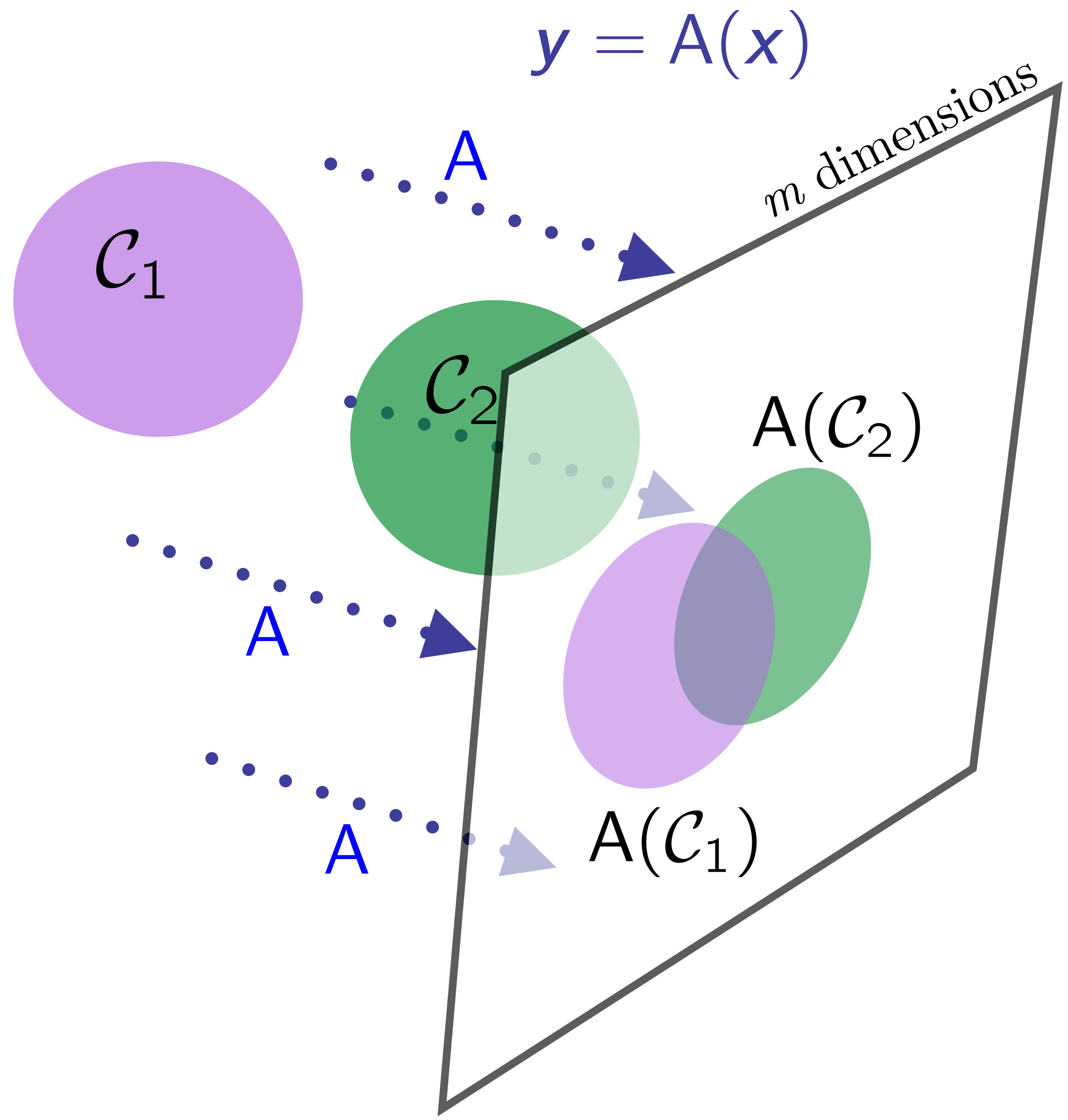
$\mathcal{K} \subset \mathbb{R}^n$ dataset

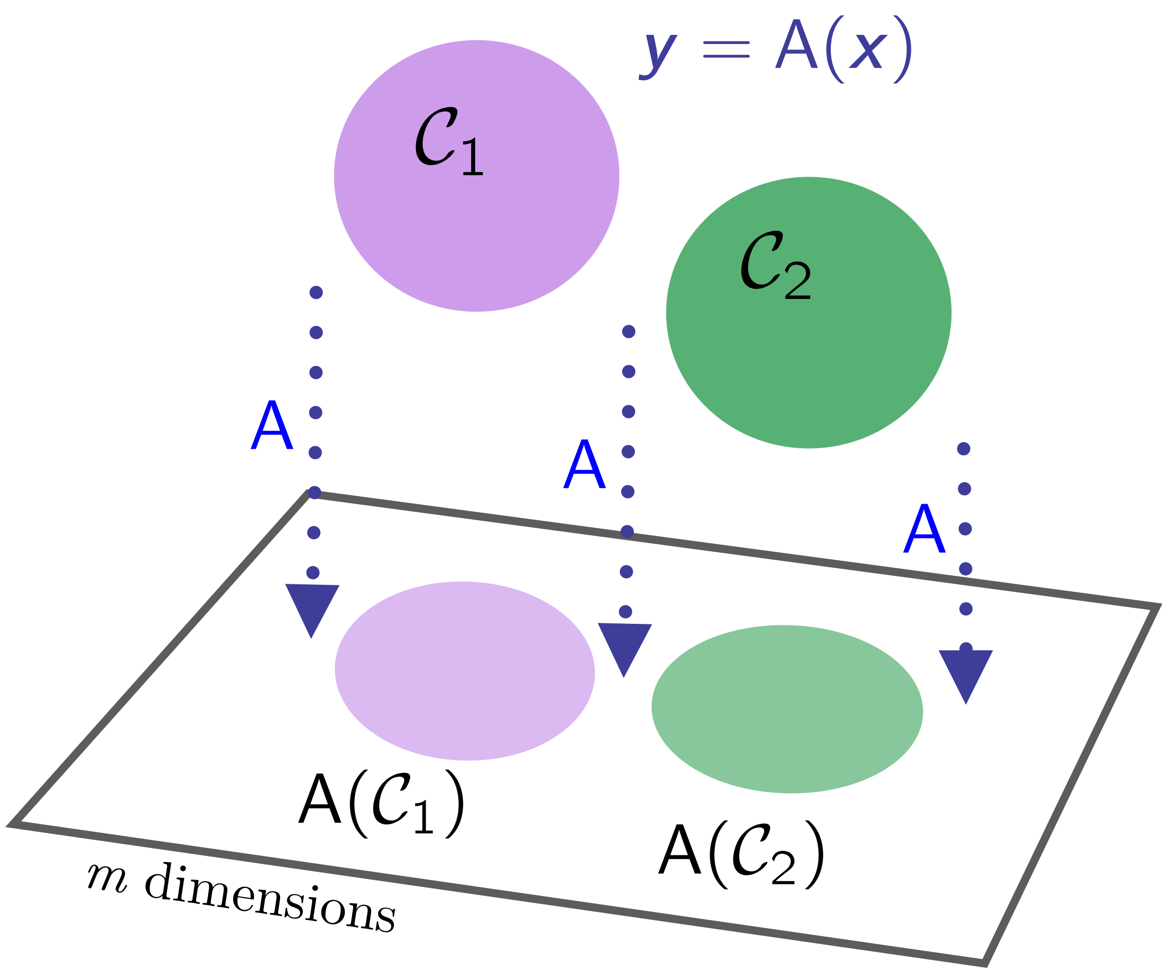
$\mathcal{C}_i \subset \mathcal{K}$ classes, $i = 1, 2, \ldots$



(e.g., LDA, SVM, PCA, K-Means, K-NN, ...)

Linear Separability After Embedding





The (Linear) Rare Eclipse Problem

Problem (Rare Eclipse Problem (Bandeira *et al.* '14)).

Let $\mathcal{C}_1, \mathcal{C}_2 \subset \mathbb{R}^n : \mathcal{C}_1 \cap \mathcal{C}_2 = \emptyset$ be closed convex sets, $\mathbf{\Phi} \sim \mathcal{N}^{m \times n}(0, 1)$. Given $\eta \in (0, 1)$, find the smallest m so that

$$p_0 := \mathbb{P}_{\mathbf{\Phi}}[\mathbf{\Phi}\mathcal{C}_1 \cap \mathbf{\Phi}\mathcal{C}_2 = \emptyset] \geq 1 - \eta.$$

Bandeira, Mixon, Recht '14 [BMR '14]

The (Linear) Rare Eclipse Problem

BMR '14: "*Gordon's escape through a mesh*" theorem

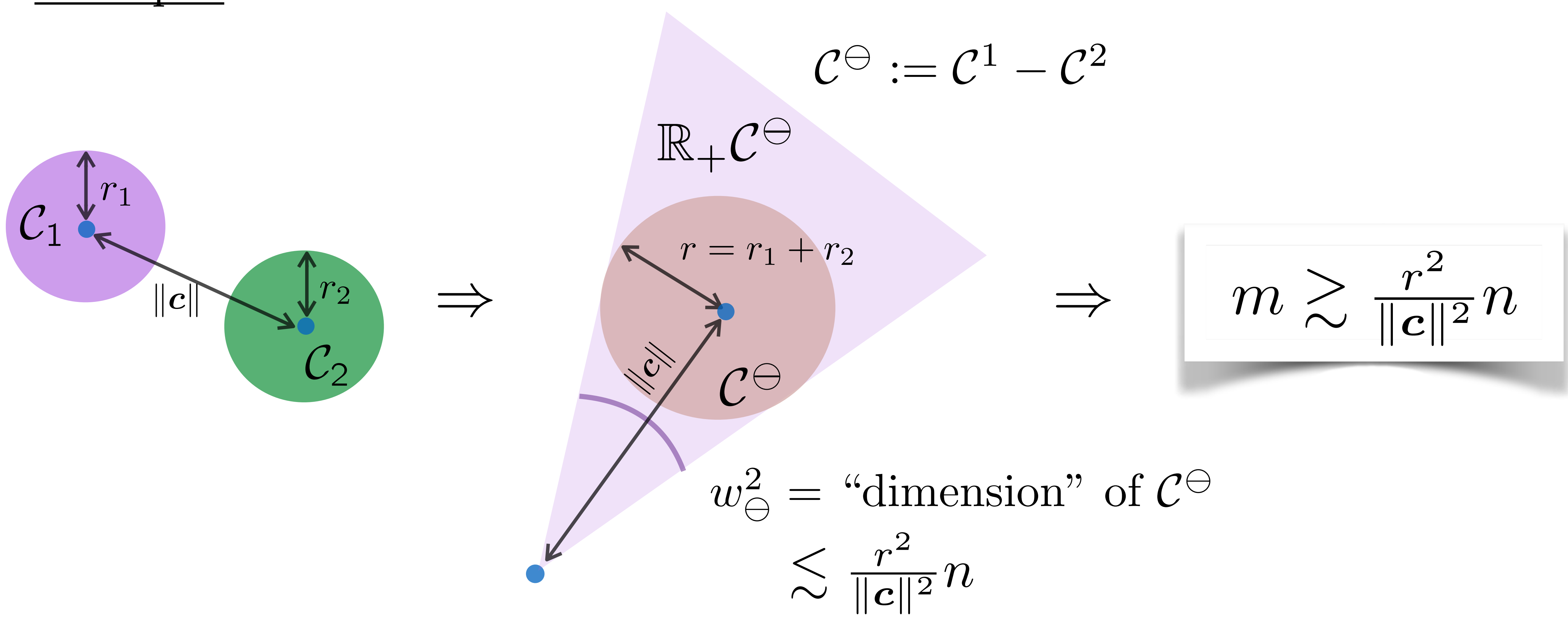
Proposition (Corollary 3.1 in BMR '14).

(& *really* tight [Amelunxen et al, 13])

Given $\eta \in (0,1)$, if $m > \left(w_{\ominus} + \sqrt{2 \log \frac{1}{\eta}}\right)^2 + 1$ then $p_0 \geq 1 - \eta$.

Bandeira, Mixon, Recht '14 [BMR '14]

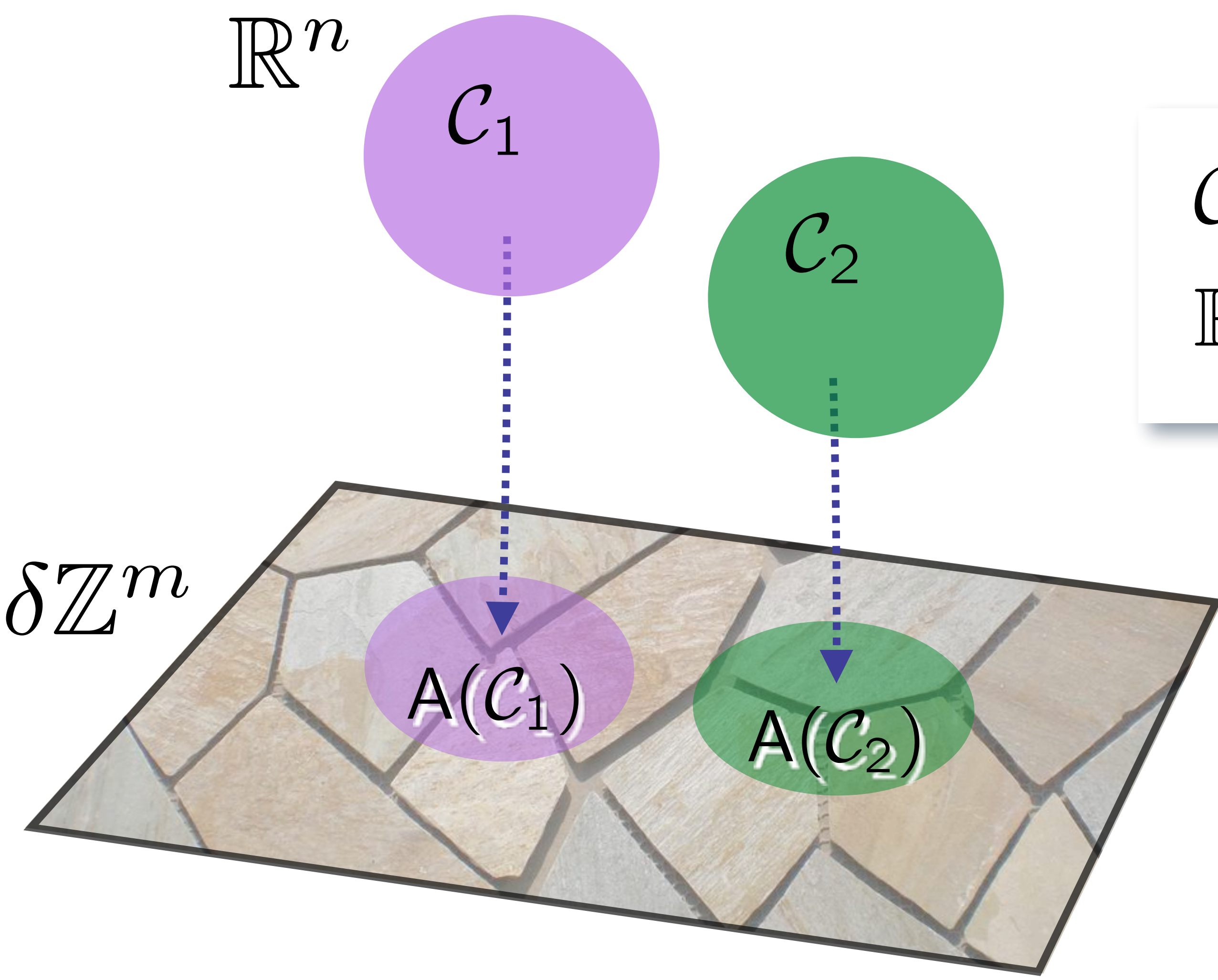
<u>Example:</u>



Quantization: The Rare Eclipse Problem "on Tiles"

$$\mathsf{A}(\boldsymbol{x}) := \mathcal{Q}(\boldsymbol{\Phi}\boldsymbol{x} + \boldsymbol{\xi})$$

with $\boldsymbol{\Phi}$ Gaussian random matrix,
$\mathcal{Q}(\lambda) = \delta\lfloor \frac{\lambda}{\delta} \rfloor$, $\xi_i \sim \mathcal{U}([0,\delta])$.



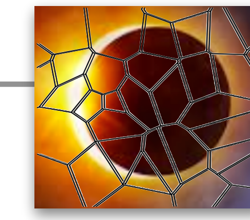
$\mathcal{C}_1$, $\mathcal{C}_2$, m and δ such that
$\mathbb{P}[\mathsf{A}(\mathcal{C}_1) \cap \mathsf{A}(\mathcal{C}_2) = \emptyset] \geqslant 1 - \eta$?

Idea: use the QRIP, i.e.,

$$\frac{1}{M\delta}\|\mathsf{A}(\boldsymbol{x}_1) - \mathsf{A}(\boldsymbol{x}_2)\|_1 \approx \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|$$

w.h.p.

The Rare Eclipse Problem "on Tiles"



- Combining (P1), (P2) and (P3) (+ massage) gives

Given $\sigma := \min_{\boldsymbol{z}\in\mathcal{C}^{\ominus}} \|\boldsymbol{z}\|$ and $w_{\cap} = w((\mathbb{R}_+\mathcal{C}^{\ominus}) \cap \mathbb{S}^{n-1})$.

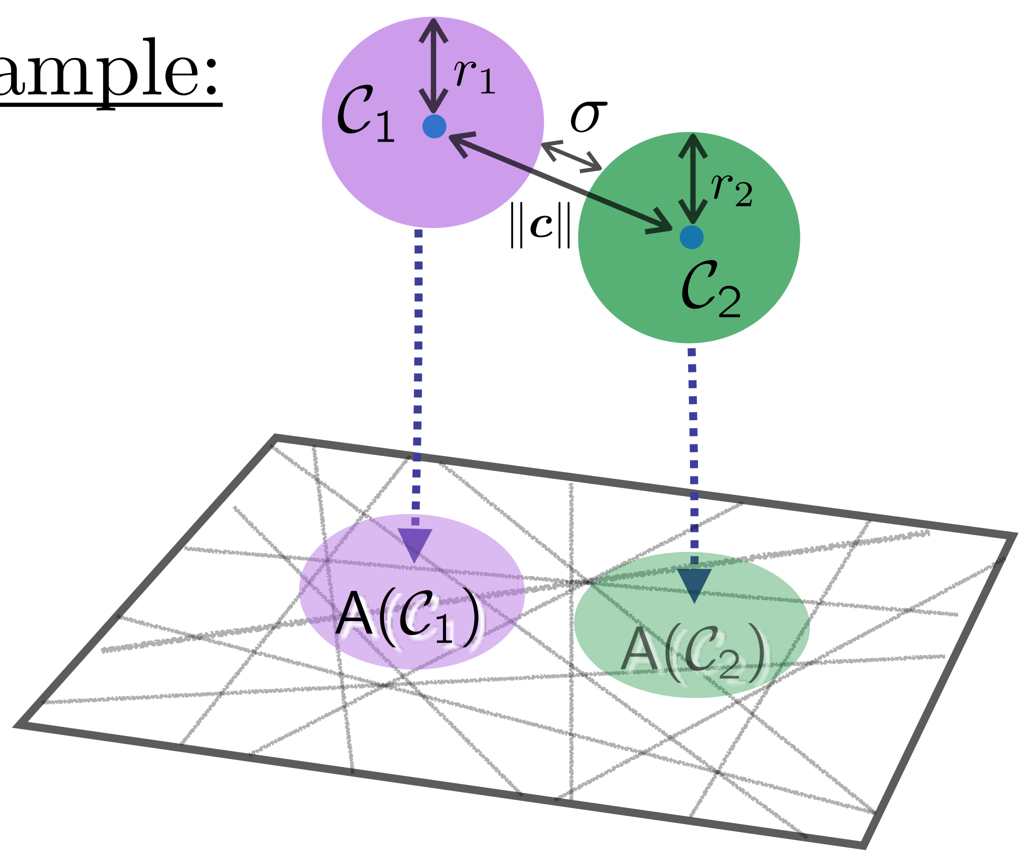
Provided

$$m \gtrsim \Big(\underbrace{w_{\ominus}^2}_{\text{linear}} + \underbrace{n\tfrac{\delta^2}{\sigma^2}}_{\text{quantiz.}}\Big)\ \Big(1 + \underbrace{\log\big(1+\tfrac{rm}{\delta n}\big)}_{\text{proof artifact?}}\Big) + \underbrace{w_{\ominus}^{-2}\log\tfrac{1}{\eta}}_{\text{linear}}\Big),$$

we have

$$\mathbb{P}[\mathsf{A}(\mathcal{C}_1) \cap \mathsf{A}(\mathcal{C}_2) = \emptyset] \geqslant 1 - \eta.$$

<u>Example:</u>



$$\Rightarrow \quad m \gtrsim \Big(\frac{r^2}{\|\boldsymbol{c}\|^2} + \frac{\delta^2}{(\|\boldsymbol{c}\|-r)^2}\Big)\, n$$

Note: $\delta > \sigma$ is allowed (dithering effect!)

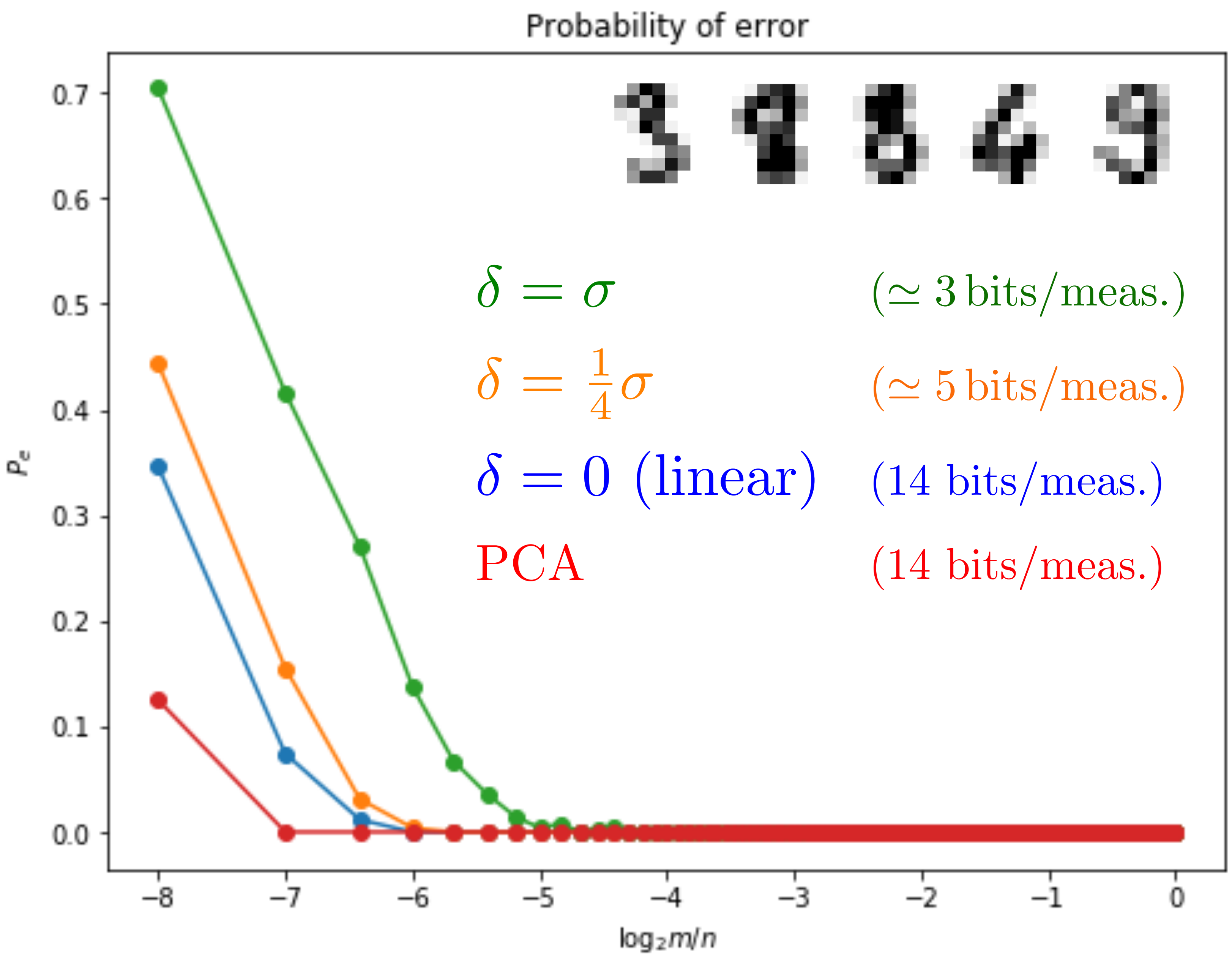
Note bis: $m > n$ not specially bad $(\delta\mathbb{Z}^m)$.

Simulations: Digit dataset (from scikit learn)

10 handwritten digits, 8x8 pixels (n=64), samples/class ≃ 12.

Training/Test sets = 50%/50%. $\sigma = \min_{i,j:i\neq j} \min_{\boldsymbol{u}\in\mathcal{C}_i,\boldsymbol{v}\in\mathcal{C}_j} \|\boldsymbol{u}-\boldsymbol{v}\|$

Classification: 5-NN Classifier.

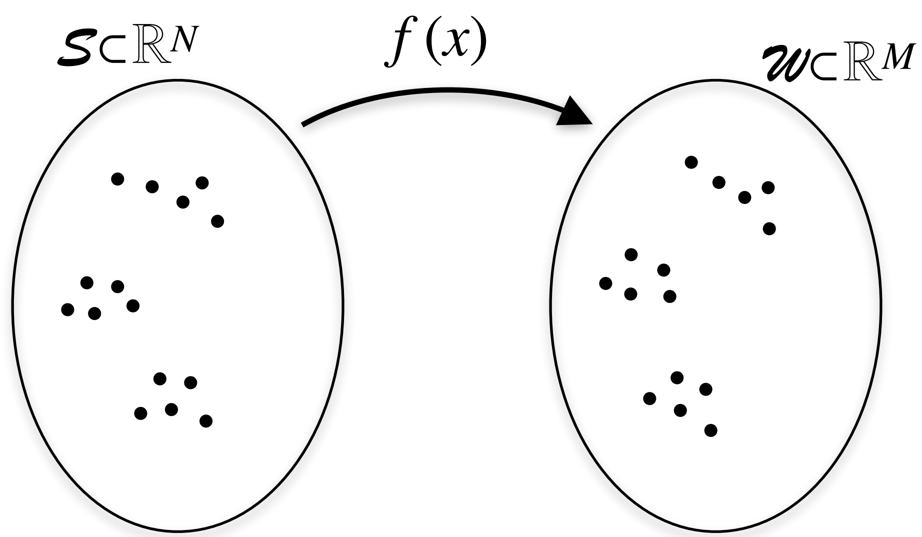


Try some code out here: **github.com/VC86/MLSPbox**

LEARNING EMBEDDINGS

Embedding Learning Objectives [Hegde et. al. '15], [Sadeghian et. al. '13]

Original space
Distance metric: $d_{\mathcal{S}}$

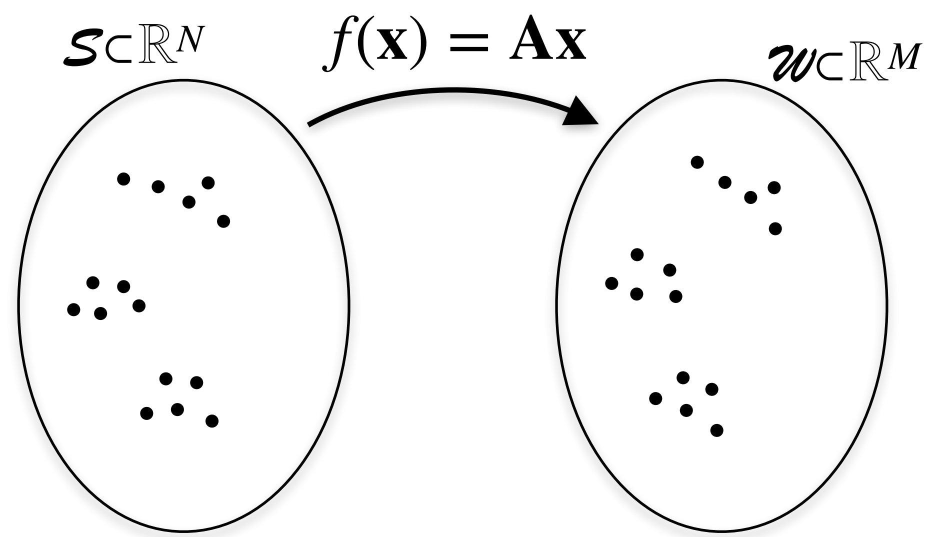


Embedding space
Embed in $\mathcal{W}$
Distance metric: $d_{\mathcal{W}}$

- **General objective**: learn $f(x)$ to optimize embedding aspects from sample data
  - Mostly to reduce the dimension M
- Very general problem
  - What distance metrics to consider?
  - What functions to restrict it to?
  - Is it possible to learn selective distortions?
- **Today**: J-L style embeddings
  - Linear embeddings (i.e., $f(x)=Ax$)
  - ℓ_2 distance metric
  - Some discussion on selective distortion
- **Note**: in the deep learning/ artificial neural networks literature the term "embedding learning" is very commonly used. This is a quite imprecise qualitative use of the term. To our knowledge there is no work establishing guarantees in preserving geometric aspects of the original space.

Embedding Learning Preparation

Original space
Distance metric: ℓ_2



Embedding space
Embed in $\mathcal{W}$
Distance metric: ℓ_2

Training set:

$$\mathscr{L} = \{\mathbf{x}_i \in \mathbb{R}^N, i = 1, \ldots, L\}$$

$\longrightarrow$

Secant set:

$$\mathcal{S} = \left\{ \mathbf{v}_{ij} = \frac{\mathbf{x}_i - \mathbf{x}_j}{\|\mathbf{x}_i - \mathbf{x}_j\|_2}, \mathbf{x}_i, \mathbf{x}_j \in \mathscr{L} \right\}$$

(normalized differences of all pairs)

Key realization: Preserving distances in training set is equivalent to preserving norms of the secant set:

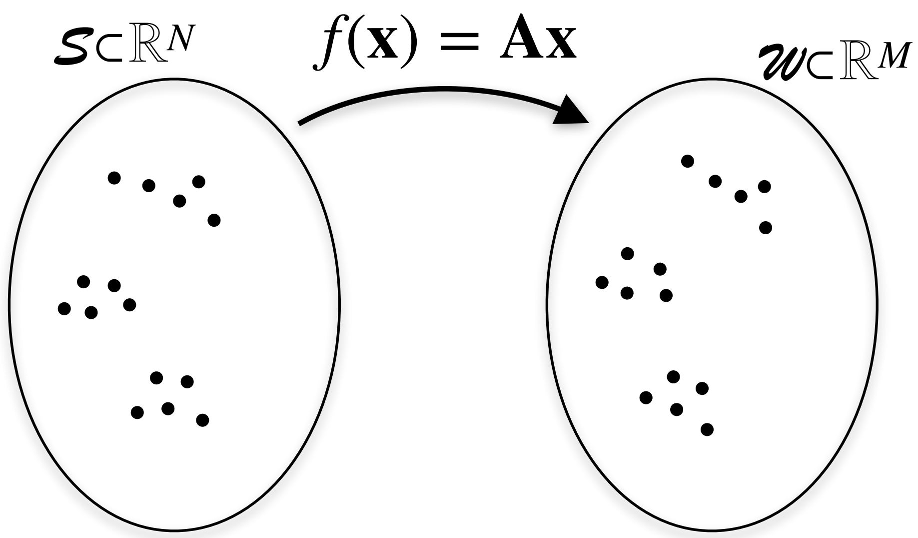
$$\left| \| \mathbf{A}\mathbf{x}_i - \mathbf{A}\mathbf{x}_j \|_2^2 - \| \mathbf{x}_i - \mathbf{x}_j \|_2^2 \right| \leq \delta \| \mathbf{x}_i - \mathbf{x}_j \|_2^2$$

$$\Longleftrightarrow$$

$$\left| \| \mathbf{A}\mathbf{v}_{ij} \|_2^2 - \| \mathbf{v}_{ij} \|_2^2 \right| \leq \delta$$

equal to 1 by construction

Embedding Learning Objectives



Training set:

$$\mathcal{L} = \{\mathbf{x}_i \in \mathbb{R}^N, i = 1,\ldots, L\}$$

Secant set:

$$\mathcal{S} = \left\{ \mathbf{v}_{ij} = \frac{\mathbf{x}_i - \mathbf{x}_j}{\|\mathbf{x}_i - \mathbf{x}_j\|_2}, \mathbf{x}_i, \mathbf{x}_j \in \mathcal{L} \right\}$$

(normalized differences of all pairs)

Learning: Find $\mathbf{A}$ that satisfies

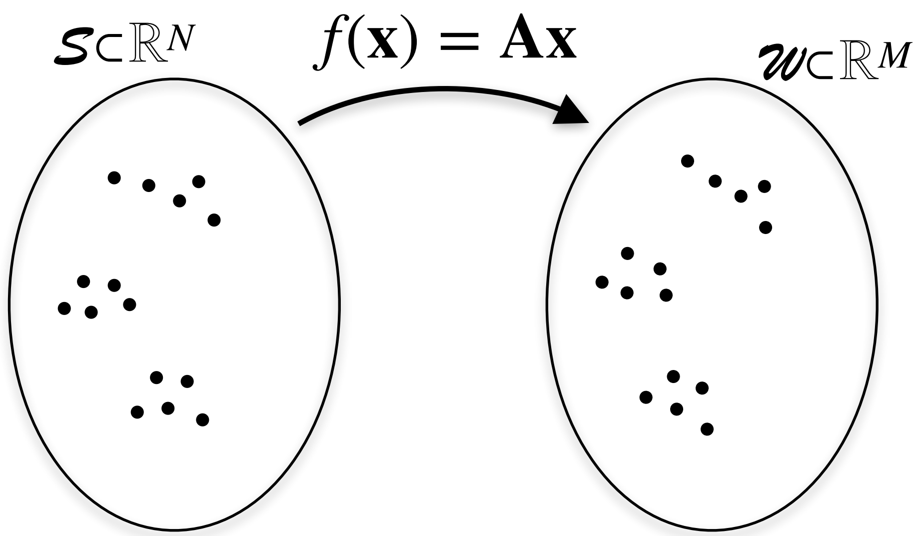
$$\left| \|\mathbf{A}\mathbf{v}_{ij}\|_2^2 - \|\mathbf{v}_{ij}\|_2^2 \right| \leq \delta \iff \left| \|\mathbf{A}\mathbf{v}_{ij}\|_2^2 - 1 \right| \leq \delta$$

Trick: $\|\mathbf{A}\mathbf{v}\|_2^2 = \mathbf{v}^T\mathbf{A}^T\mathbf{A}\mathbf{v} = \mathbf{v}^T\mathbf{P}\mathbf{v}$,
where $\mathbf{P} = \mathbf{A}^T\mathbf{A}$ is symmetric positive semi-definite, and $\mathrm{rank}(\mathbf{P}) = \mathrm{rank}(\mathbf{A})$.
Given $\mathbf{v}$, $\mathbf{v}^T\mathbf{P}\mathbf{v}$ is linear in $\mathbf{P}$

Optimization:

- Embedding accuracy δ (should be small)
- Dimension of $\mathbf{A} = \mathrm{rank}(\mathbf{A}) = \mathrm{rank}(\mathbf{P}) = M$ (should also be small)
- Different formulations lead to different optimization problems
  – Fix rank and optimize δ, or fix δ and optimize rank

Embedding Learning Objectives



Training set:

$$\mathscr{L} = \{\mathbf{x}_i \in \mathbb{R}^N, i = 1,\ldots,L\}$$

Secant set:

$$\mathscr{S} = \left\{ \mathbf{v}_{ij} = \frac{\mathbf{x}_i - \mathbf{x}_j}{\|\mathbf{x}_i - \mathbf{x}_j\|_2}, \mathbf{x}_i, \mathbf{x}_j \in \mathscr{L} \right\}$$

(normalized differences of all pairs)

$$\left| \, \| \mathbf{A}\mathbf{v} \|_2^2 - 1 \right| \leq \delta \iff \left| \mathbf{v}^T \mathbf{P} \mathbf{v} - 1 \right| \leq \delta, \ \mathbf{P} = \mathbf{P}^T \succcurlyeq 0$$

Ideal Optimization Problem

$$\widehat{\mathbf{P}} = \arg \min_{\mathbf{P}^T = \mathbf{P} \succeq 0} \operatorname{rank}(\mathbf{P})$$

$$\text{subject to } |\mathbf{v}_{ij}^T \mathbf{P} \mathbf{v}_{ij} - 1| \leq \delta \text{ for all } i \neq j.$$

Convex relaxation

$$\widehat{\mathbf{P}} = \arg \min_{\mathbf{P}^T = \mathbf{P} \succeq 0} \|\mathbf{P}\|_*$$

$$\text{subject to } |\mathbf{v}_{ij}^T \mathbf{P} \mathbf{v}_{ij} - 1| \leq \delta \text{ for all } i \neq j.$$

Alternative formulation

$$\widehat{\mathbf{P}} = \arg \min_{\mathbf{P}^T = \mathbf{P} \succeq 0} \max_{i \neq j} |\mathbf{v}_{ij}^T \mathbf{P} \mathbf{v}_{ij} - 1|$$

(δ)

$$\text{subject to } \operatorname{rank}(\mathbf{P}) \leq M \text{ and } \|\mathbf{P}\|_* \leq b$$

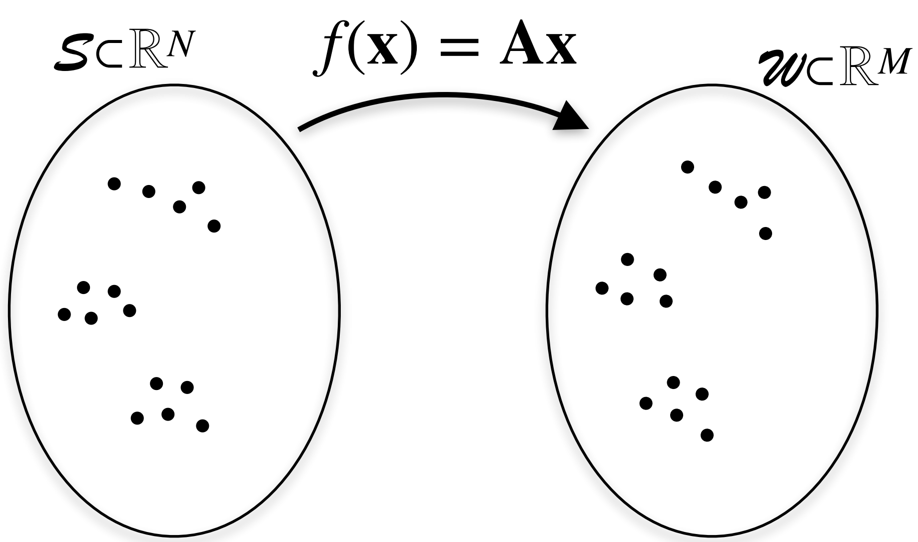
Final Step

Obtain $\mathbf{A}$ using the SVD of $\mathbf{P}$

$$\widehat{\mathbf{P}} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{U}^T$$

$$\Longrightarrow \mathbf{A} = \boldsymbol{\Sigma}^{1/2}\mathbf{U}^T$$

Embedding Learning Objectives



Training set:

$$\mathscr{L} = \{\mathbf{x}_i \in \mathbb{R}^N, i = 1, \dots, L\}$$

Secant set:

$$\mathscr{S} = \left\{ \mathbf{v}_{ij} = \frac{\mathbf{x}_i - \mathbf{x}_j}{\|\mathbf{x}_i - \mathbf{x}_j\|_2}, \mathbf{x}_i, \mathbf{x}_j \in \mathscr{L} \right\}$$

(normalized differences of all pairs)

$$\left| \| \mathbf{A}\mathbf{v} \|_2^2 - 1 \right| \leq \delta \Longleftrightarrow \left| \mathbf{v}^T \mathbf{P} \mathbf{v} - 1 \right| \leq \delta, \mathbf{P} = \mathbf{P}^T \succcurlyeq 0$$

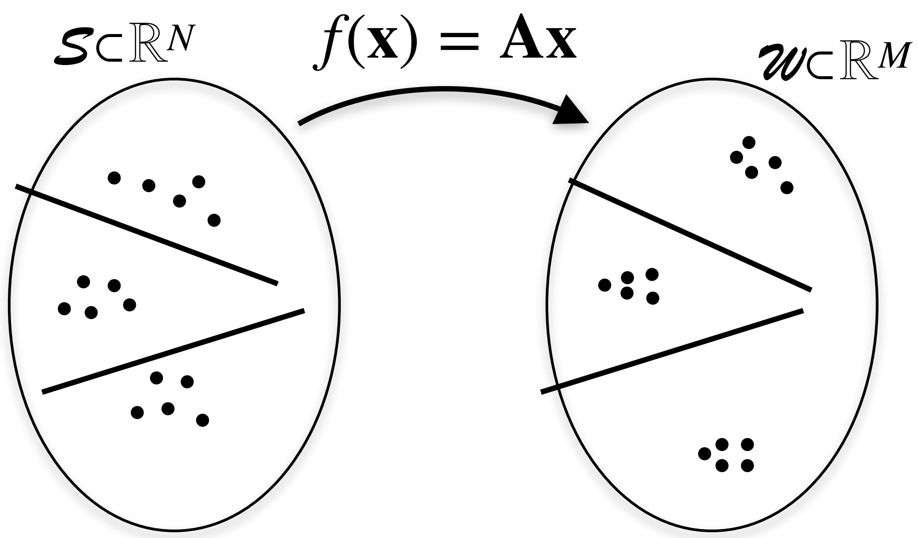
Generalization:

- Embedding accuracy δ holds only for training sample
- For signals similar to the ones in the training set, guarantee can be generalized
  - Exploits continuity of the linear embedding map

For any $\mathbf{z}$ s.t. $\|\mathbf{z} - \mathbf{x}\|_2 \leq \epsilon \|\mathbf{x}_2\|$ for all $\mathbf{x}$ in the training set, resulting isometry bound is [Sadeghian et. al. '13]

$$\bar{\delta} \leq \frac{\delta + \epsilon}{1 - \epsilon}$$

Learning Embeddings For Classification [Hegde et. al. '15]



Training set:

$$\mathcal{L} = \{\mathbf{x}_i \in \mathbb{R}^N, i = 1, \ldots, L\}$$

+ **class labels** for each $\mathbf{x}_i$

Secant set:

$$\mathcal{S} = \left\{ \mathbf{v}_{ij} = \frac{\mathbf{x}_i - \mathbf{x}_j}{\|\mathbf{x}_i - \mathbf{x}_j\|_2}, \mathbf{x}_i, \mathbf{x}_j \in \mathcal{L} \right\}$$

(normalized differences of all pairs)

Intuition:

If $\mathbf{x}_i$, $\mathbf{x}_j$ in the same class, we should not let their distance increase much (but *ok* if they come closer to each other)

If $\mathbf{x}_i$, $\mathbf{x}_j$ in different class, we should not let their distance decrease much (but *ok* if they go farther from each other)

Resulting Optimization:

$$\widehat{\mathbf{P}} = \arg \min_{\mathbf{P}^T = \mathbf{P} \succeq 0} \|\mathbf{P}\|_*$$

$$\text{subject to } \mathbf{v}_{ij}^T \mathbf{P} \mathbf{v}_{ij} \geq 1 - \delta \text{ for all } i \neq j \text{ in different classes.}$$

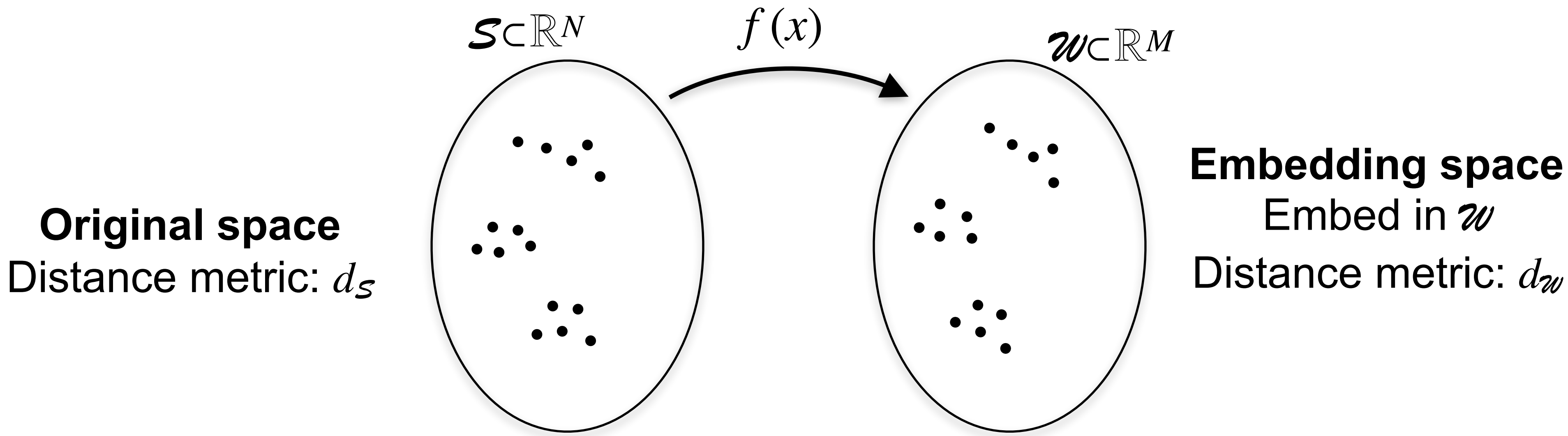
$$\mathbf{v}_{ij}^T \mathbf{P} \mathbf{v}_{ij} \leq 1 + \delta \text{ for all } i \neq j \text{ in the same class.}$$

SUMMARY AND CONCLUSIONS

Recap

2. Fundamentals of embeddings and embedology
   - Dimensionality reduction method
   - Main goal: preserves distances
   - Typical approach: randomization
3. Quantized embeddings
   - Quantization does not hurt that much
   - Careful quantization design can serve as a compression approach
4. Embedding Design
   - It is possible to design embeddings for selective distortions
   - Optimal embeddings are not known
5. Embeddings of Alternative Metrics
   - We can embed distances, angles, kernels, or anything else you might like
   - Hashing approaches can speed up computation (also have connections with quantized embeddings)
   - Embeddings can be designed to preserve some property (e.g. separation of classes)
6. Learning Embeddings
   - Of course it is possible to learn embeddings from data
   - Simple optimization problem
   - Carefully setting up the optimization allows for out-of-sample guarantees.

Conclusions



- Dimensionality reduction is a very rich a rich subject
  - Embeddings is just a small part

- Randomized embeddings provide “universal” approach to dimensionality reduction
  - Sometimes not the most efficient approach if the objective is strictly to reduce the number of dimensions
  - However, they provide computational advantages as they don’t depend on the data to design
  - A number of data-dependent dimensionality reduction techniques (e.g., PCA) make explicit or implicit assumptions on the data in order to provide guarantees
  - Randomization offers significant theoretical advantages and allows for strong guarantees, irrespective of the dataset

Toolkits?

- Most Embedding computation is trivial
  - Matlab:

```
A=randn[M,N];
y=A*x;
```

  - Python:

```
import numpy as np
import numpy.random as rnd
A = rnd.randn(M,N)
y = A.dot(x)
```

- For that reason, not many toolkits exist for randomized embeddings
  - Some
- LSH is a bit more intricate
  - Need to keep track of hashing tables efficiently
  - Most toolkits focus on that; the hashing part is also trivial to implement
    - There are several hashing approaches, with different advantages/ disadvantages
  - Several tools to keep track of (http://ann-benchmarks.com provides benchmarking and a list of a number of popular libraries)

Open Problems

- Overall, there has been a flurry of theory recently
  - Still there are quite a few research avenues
- Generally the question of embedding design is not well understood
  - What are desirable embedding properties for each application?
  - Is there a general embedding map that can implement arbitrary distortions?
  - What is the optimal embedding design given a desired distortion?
- Quantization and LSH
  - There is a strong connection between embedding quantization and LSH
  - This connection has not been explored
- Learning embeddings
  - Very little work in the area
  - Can we learn non-linear embeddings?
  - Can we learn quantized embeddings?
  - Can we learn embeddings for other distances and/or distance maps?
- Neural Networks/Deep Learning
  - What are Deep Networks embedding?
  - What kind of theoretical guarantees can we provide?

Thank you for your attention.
Questions?

More info and resources:
boufounos.com/embeddings

petros@boufounos.com *laurent.jacques@uclouvain.be*

(unsorted) Selected References (1/2) (for further reading)

- Johnson, William B., and Joram Lindenstrauss. "Extensions of Lipschitz mappings into a Hilbert space." Contemporary mathematics 26.189-206 (1984).
- Jacques, L., Laska, J. N., Boufounos, P. T., & Baraniuk, R. G. (2013). Robust 1-bit compressive sensing via binary stable embeddings of sparse vectors. IEEE Transactions on Information Theory, 59(4), 2082-2102.
- Boufounos, P.T., Rane, S. and Mansour, H., 2017. Representation and coding of signal geometry. Information and Inference: A Journal of the IMA, 6(4), pp.349-388.
- P. T. Boufounos, S. Rane, and H. Mansour, "Embedding-Based Representation of Signal Geometry," Excursions in Harmonic Analysis, Volume 5: The February Fourier Talks at the Norbert Wiener Center, pp. 155-178, 2017.
- Boufounos, Petros T. "Universal rate-efficient scalar quantization." IEEE transactions on information theory 58.3 (2012): 1861-1872.
- Plan, Yaniv, and Roman Vershynin. "Robust 1-bit compressed sensing and sparse logistic regression: A convex programming approach." IEEE Transactions on Information Theory 59.1 (2013): 482-494.
- Plan, Yaniv, and Roman Vershynin. "Dimension reduction by random hyperplane tessellations." Discrete & Computational Geometry 51.2 (2014): 438-461.
- Puy, Gilles, Mike E. Davies, and Rémi Gribonval. "Recipes for Stable Linear Embeddings From Hilbert Spaces to R^m" IEEE Transactions on Information Theory 63.4 (2017): 2171-2187.
- Baraniuk, Richard G., and Michael B. Wakin. "Random projections of smooth manifolds." Foundations of computational mathematics 9.1 (2009): 51-77.
- 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.
- Dirksen, Sjoerd, and Alexander Stollenwerk. "Fast binary embeddings with gaussian circulant matrices: improved bounds." Discrete & Computational Geometry (2016): 1-28.Y. Plan, R. Vershynin, "Dimension reduction by random hyperplane tessellations", 2013
- Sjoerd Dirksen, Shahar Mendelson, "Non-Gaussian Hyperplane Tessellations and Robust One-Bit Compressed Sensing", arXiv: 1805.09409
- Samet Oymak, "Near-optimal sample complexity bounds for circulant binary embedding", ArXiv:1603.03178 (2016)
- Indyk, P., Motwani, R.: Approximate nearest neighbors: towards removing the curse of di- mensionality. In: ACM Symposium on Theory of computing, pp. 604–613 (1998)

(unsorted) Selected References (2/2) (for further reading)

- F. Yu, et al. "On binary embedding using circulant matrices." The Journal of Machine Learning Research 18.1 (2017): 5507-5536.
- R. Giryes, G. Sapiro and A.M. Bronstein, "Deep Neural Networks with Random Gaussian Weights: A Universal Classification Strategy? ", IEEE Transactions on Signal Processing, vol. 64, no. 13, pp. 3444-3457, Jul. 2016.
- Achlioptas D. Database-friendly random projections: Johnson-Lindenstrauss with binary coins. Journal of computer and System Sciences. 2003 Jun 1;66(4):671-87.
- Ailon N, Chazelle B. The fast Johnson–Lindenstrauss transform and approximate nearest neighbors. SIAM Journal on computing. 2009 May 28;39(1):302-22.
- Foucart, Simon, and Holger Rauhut. A mathematical introduction to compressive sensing. Vol. 1. No. 3. Basel: Birkhäuser, 2013.
- Krahmer F, Ward R. New and improved Johnson–Lindenstrauss embeddings via the restricted isometry property. SIAM Journal on Mathematical Analysis. 2011 May 26;43(3):1269-81.
- Liutkus A, Martina D, Popoff S, Chardon G, Katz O, Lerosey G, Gigan S, Daudet L, Carron I. Imaging with nature: Compressive imaging using a multiply scattering medium. Scientific reports. 2014 Jul 9;4:5552.
- Jacques L. A quantized Johnson–Lindenstrauss lemma: The finding of Buffon's needle. IEEE Transactions on Information Theory. 2015 Sep;61(9):5012-27.
- G. C. Buffon, "Essai d'arithmetique morale," Supplément à l'histoire naturelle, vol. 4, 1777 (http://www.buffon.cnrs.fr/).
- Dirksen S, Jung HC, Rauhut H. One-bit compressed sensing with partial Gaussian circulant matrices. arXiv preprint arXiv:1710.03287. 2017 Oct 9.
- Jacques, Laurent, and Valerio Cambareri. "Time for dithering: fast and quantized random embeddings via the restricted isometry property." Information and Inference: A Journal of the IMA 6.4 (2017): 441-476.
- Xu C, Jacques L. Quantized compressive sensing with rip matrices: The benefit of dithering. arXiv preprint arXiv:1801.05870. 2018 Jan 17.
- R. Giryes, G. Sapiro and A.M. Bronstein, "Deep Neural Networks with Random Gaussian Weights: A Universal Classification Strategy?" , IEEE Transactions on Signal Processing, vol. 64, no. 13, pp. 3444-3457, Jul. 2016.
- Eftekhari A, Yap HL, Wakin MB, Rozell CJ. Stabilizing embedology: Geometry-preserving delay-coordinate maps. Physical Review E. 2018 Feb 26;97(2):022222.
- Ostrovsky, R., Rabani, Y.: Low distortion embeddings for edit distance. Journal of the ACM (JACM) 54(5), 23 (2007)