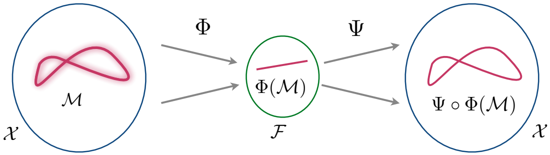
MIT 9.520/6.860, Fall 2017
Statistical Learning Theory and Applications

Class 19: Data Representation by Design

What is data representation?

Let $\mathcal{X}$ be a data-space



A **data representation** is a map

$$\Phi : \mathcal{X} \to \mathcal{F},$$

from the data space to a **representation space** $\mathcal{F}$.

A **data reconstruction** is a map

$$\Psi : \mathcal{F} \to \mathcal{X}.$$

Name game

$$\Phi : \mathcal{X} \to \mathcal{F}, \qquad \Psi : \mathcal{F} \to \mathcal{X}$$

Different names in different fields:

- **learning**: feature map/pre-image
- **signal processing**: analysis/synthesis
- **information theory**: encoder/decoder
- **computational geometry**: representation=embedding

Learning and data representation

$$f(x) = \langle w, \Phi(x) \rangle_{\mathcal{F}}, \quad \forall x \in \mathcal{X}$$

Two-step learning scheme:

- *Data representation*: $\Phi : \mathcal{X} \to \mathcal{F}, \quad x \mapsto \Phi(x)$
- *Supervised learning* of w in $\mathcal{F}$

Representation examples:

- By design: Fourier, Frames, Random projections, Kernels
- Unsupervised: VQ, K-means/K-flats, Sparse Coding, Dictionary Learning, PCA, Autoencoders, NMF, RBF networks
- Supervised: Neural Networks, ConvNets, Supervised DL

Road map

- Prologue/summary: **Learning theory** and data representation
- Part I: Data representations by **design**
- Part II: Data representations by **learning**
- Part III: **Deep** data representations

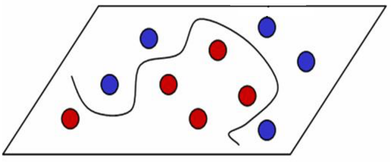
Data representation & learning theory

Supervised learning is the most mature and well understood form of machine learning.

Foundational results in learning theory establish when learning is possible & show the importance of data representation.

keywords: sample complexity, no free lunch theorem, reproducing kernel Hilbert space

Key theorem in supervised learning



- Supervised learning: find unknown function

$$f: \mathcal{X} \to \mathcal{Y}$$

given examples $S_n = \{(x_i, y_i)\}_{i=1}^{n} \in (\mathcal{X}, \mathcal{Y})$.

- Key theorem: *finite sample complexity*[1] only possible within a suitable space of hypothesis space $\mathcal{H} \subset \{f | f : \mathcal{X} \to \mathcal{Y}\}$.

---

[1] Number of samples required to achieve an accuracy with a given confidence.

More formally...

- **Data space** $\mathcal{X} \times \mathcal{Y}$ with probability distribution ρ
- **Loss function** $V : \mathcal{Y} \times \mathcal{Y} \to [0, \infty)$,

Problem: Solve

$$\inf_{f \in \mathcal{F}} \mathcal{E}(f), \quad \mathcal{E}(f) = \int_{\mathcal{X} \times \mathcal{Y}} V(f(x), y) d\rho(x, y).$$

given a **training set** $S_n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ sampled *identically and independently* with respect to ρ.

Note:
- ρ fixed but **unknown**
- $\mathcal{F}$ space of **all** (measurable) functions

Learning algorithms & hypothesis space

$$\inf_{f \in \mathcal{F}} \mathcal{E}(f),$$

- **Learning algorithm**: procedure providing an approximate solution $\hat{f}$ given a training set S_n.
- **Hypothesis space**: space of all possible solutions $\mathcal{H}$ that can be returned by a learning algorithm.

Examples: Regularization Nets, Kernel Machines/SVM, Neural Networks, Nearest Neighbors ...

Sample complexity

The quality of a learning algorithm is captured by the sample complexity.

Definition (Sample Complexity)

For all $\epsilon \in [0, \infty)$, $\delta \in [0, 1]$, *an algorithm has sample complexity* $n_{\mathcal{H}}(\epsilon, \delta, \mathcal{H}) \in \mathbb{N}$ *if*

$$\forall n \geq n_{\mathcal{H}}(\epsilon, \delta, \rho), \qquad \mathbb{P}\left(\mathcal{E}(\hat{f}) - \inf_{f \in \mathcal{H}} \mathcal{E}(f) \geq \epsilon\right) \leq \delta$$

Note:

- Space of all functions $\mathcal{F}$ is replaced by the **hypothesis space** $\mathcal{H}$.
- **Probably approximately correct (PAC)** solution, with $n_{\mathcal{H}}(\epsilon, \delta, \mathcal{H})$ samples achieves accuracy ϵ with confidence $1 - \delta$.

Key theorem: No free lunch!

The sample complexity of an algorithm can be infinite if $\mathcal{H}$ is too big (e.g. space of all possible function $\mathcal{F}$)

$$\sup_{\mathcal{F}} \sup_{\rho} n_{\mathcal{F}}(\varepsilon, \delta, \rho) = \infty$$

$$\inf_{\hat{f}} \sup_{\rho} n_{\mathcal{F}}(\epsilon, \delta, \rho) = \infty$$

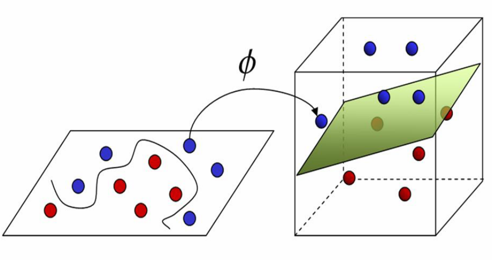
Take home message (1):
Learning with finite samples is possible only if an algorithm operates in a constrained hypothesis space.

Hypothesis space & data representations

Under weak assumptions:

$$\text{hypothesis space } \mathcal{H} \Leftrightarrow \text{data representation } \Phi : \mathcal{X} \to \mathcal{F}$$

$$f(x) = \langle w, \Phi(x) \rangle_{\mathcal{F}}$$



Hypothesis space & data representation

Requirements on the hypothesis space $\mathcal{H}$:

- statistical arguments (e.g., sample complexity)
- **computational considerations**

A function space suitable for

- **efficient computations**,
- defining empirical quantities (e.g. empirical data error)

$\Rightarrow$ **reproducing kernel Hilbert spaces (RKHS)**.

RKHS

Definition (RKHS)

Hilbert space of functions for which evaluation functionals are continuous, i.e. for all $x \in \mathcal{X}$

$$|f(x)| \leq C_x \|f\|_{\mathcal{H}}.$$

Recall that aside from other technical aspects a Hilbert space is:

- a (possibly) infinite dimensional **linear space**[2]
- endowed with an **inner product** (hence, norm, distance, notion of orthogonality etc)

[2] closed with respect to sum and multiplication by scalars

RKHS and data representation

Theorem

If $\mathcal{H}$ is a RKHS there exists a representation (feature) space $\mathcal{F}$ and a data representation $\Phi : \mathcal{X} \to \mathcal{F}$, such that for all $f \in \mathcal{H}$ there exists w satisfying

$$f(x) = \langle w, \Phi(x) \rangle_{\mathcal{F}}, \quad \forall x \in \mathcal{X}.$$

- $\mathcal{H}$ is equivalent to feature map $\Phi : \mathcal{X} \to \mathcal{F}$

$$\mathcal{H} = \{f : \mathcal{X} \to \mathcal{Y} : \exists w \in \mathcal{F}, f(x) = \langle w, \Phi(x) \rangle_{\mathcal{F}}, \forall x \in \mathcal{X}\}$$

- Feature space $\mathcal{F}$ (Hilbert space isometric to $\mathcal{H}$):

$$\|f\|_{\mathcal{H}} = \inf\{\|w\|_{\mathcal{F}}, w \in \mathcal{F}\}$$

Take home message 2:
Under (relatively) mild assumptions the choice of a hypothesis space and a data representation are *equivalent*.

End of prologue

$$f(x) = \langle w, \Phi(x) \rangle_{\mathcal{F}}$$

Currently: theory and algorithms to **provably** learn w from data with Φ assumed to be given...

although in practice the data representation Φ is known to often make the **biggest difference**.

Road map

- Prologue: **Learning theory** and data representation
- Part I: Data representations by **design**
- Part II: Data representations by **learning**
- Part III: **Deep** data representations
- Epilogue: What's next?

Plan

Data representations that are **designed**:

1. Classic representations in **Signal Processing**
   - unitary, basis, Fourier
   - frames
   - dictionaries
   - randomized
2. Representations for **Machine Learning**
   - feature maps to kernels

Notation

$\mathcal{X}$: data space

- $\mathcal{X} = \mathbb{R}^d$ or $\mathcal{X} = \mathbb{C}^d$ (also more general later).
- $x \in \mathcal{X}$

Data representation: $\Phi : \mathcal{X} \to \mathcal{F}$.

$$\forall x \in \mathcal{X}, \exists z \in \mathcal{F} : \Phi(x) = z \in \mathcal{F}$$

$\mathcal{F}$: representation space

- $\mathcal{F} = \mathbb{R}^p$ or $\mathcal{F} = \mathbb{C}^p$
- $z \in \mathcal{F}$

Data reconstruction: $\Psi : \mathcal{F} \to \mathcal{X}$.

$$\forall z \in \mathcal{F}, \exists x \in \mathcal{X} : \Psi(z) = x \in \mathcal{X}$$

Unitary data representations

Let $\mathcal{X} = \mathcal{F} = \mathbb{C}^d$ and $\{a_1, \ldots, a_d\}$ an orthonormal basis in $\mathbb{C}^d$.

Consider $\Phi : \mathcal{X} \to \mathcal{F}$ such that for all $x \in \mathcal{X}$

$$\Phi(x) = (\langle x, a_1 \rangle, \ldots, \langle x, a_d \rangle)$$

Remarks on Φ

- can be identified with $d \times d$ **matrix** U with rows given by the atoms $a_1, \ldots, a_d$,
- is a **linear** map, $\Phi(x) = Ux$,
- is a **unitary** transformation: $U^* U = I$.

Unitary transformations

$$U^* U = UU^* = I$$

Isomorphism between two Hilbert spaces

$$\Phi : \mathcal{X} \to \mathcal{F}$$

Bijective function that preserves the inner product

$$\langle \Phi(x), \Phi(x') \rangle_{\mathcal{F}} = \langle x, U^* U x' \rangle_{\mathcal{X}} = \langle x, x' \rangle_{\mathcal{X}}, \quad \forall x, x' \in \mathcal{X}$$

Reconstruction for unitary data representation

Consider $\Psi : \mathcal{F} \to \mathcal{X}$ such that,

$$\Psi(z) = \sum_{k=1}^{d} a_k z^k, \quad \forall z \in \mathcal{F}$$

Reconstruction:

$$x = \sum_{k=1}^{d} a_k(\langle a_k, x \rangle) = \sum_{k=1}^{d} a_k z^k, \quad \forall x \in \mathcal{X}$$

Remarks on Ψ

- can be identified with the $d \times d$ **matrix** U^* with columns given by the atoms,
- is a **linear** map $\Psi(z) = U^* z$,
- is **exact**, in the sense $\Psi \circ \Phi = U^* U = I$.

Metric properties of unitary representations

Satisfy **Parseval's identity** (norm preservation)

$$\|\Phi(x)\|^2 = \sum_{k=1}^{d} |\langle x, a_k \rangle|^2 = \|x\|^2, \qquad \forall x \in \mathcal{X}.$$

Representation is an **isometry** (distance preservation)

$$\|\Phi(x) - \Phi(x')\| = \|x - x'\|, \qquad \forall x, x' \in \mathcal{X},$$

Example: Fourier representation (DFT)

Fourier basis: orthonormal basis of $\mathbb{C}^d$ formed by the atoms:

$$\{a_k\}_{k=1}^d = \left\{ \frac{1}{\sqrt{d}}, \frac{1}{\sqrt{d}} e^{-2\pi i k \frac{1}{d}}, \frac{1}{\sqrt{d}} e^{-2\pi i k \frac{2}{d}}, \dots, \frac{1}{\sqrt{d}} e^{-2\pi i k \frac{(d-1)}{d}} \right\}$$

Representation (**discrete Fourier transform (DFT)**):

$$\Phi(x) = Ux = z, \quad z^k = \frac{1}{\sqrt{d}} \sum_{j=0}^{d-1} x^j e^{-2\pi i k \frac{j}{d}}, \quad k = 0, \dots, d-1,$$

Reconstruction (**inverse DFT**):

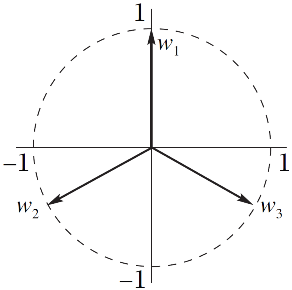
$$\Psi(z) = U^* z = x, \quad x^j = \frac{1}{\sqrt{d}} \sum_{k=0}^{d-1} z^k e^{2\pi i j \frac{k}{d}}, \quad j = 0, \dots, d-1.$$

The pursuit of the right basis

Choice of the basis U or *dictionary of atoms* $\{a_k\}_{k=1}^{d}$ reflects **prior information** about the data or the problem, e.g.

- physical system (frequencies)
- intepretability (spectral content)
- ...

Can this be extended to **more general dictionaries** *than orthonormal bases?* [3]



[3] Image credit: C. Hale, "What is a Frame?", 2003

Frames

Generalization of a basis: a weaker form of Parseval's identity.

Definition (Frame)

A finite set of atoms $\{a_1, \ldots, a_p\}, a_k \in \mathbb{R}^d$ *for which there exists* $0 < A \leq B < \infty$ *such that for all* $x \in \mathcal{X}$

$$A \|x\|^2 \leq \sum_{k=1}^{p} |\langle x, a_k \rangle|^2 \leq B \|x\|^2 .$$

Remarks:

- **Tight** frame: $A = B$.
- **Parseval** frame: $A = B = 1$.
- **Union of orthonormal bases** (renormalized) is a tight frame.

Frame examples

1. $\{a_k\} = \{e_1, e_1, e_2, e_2, \dots\}, \quad \{e_i\}_{i=1}^d \in \mathbb{R}^d$

$$\sum_{k=1}^{p} |\langle x, a_k \rangle|^2 = \sum_{k=1}^{d} |\langle x, e_k \rangle|^2 + \sum_{k=1}^{d} |\langle x, e_k \rangle|^2 = 2||x||^2$$

tight frame for $\mathbb{R}^d$ with $A = B = 2$

2. $\{a_k\} = \{e_1, \frac{1}{\sqrt{2}}e_2, \frac{1}{\sqrt{2}}e_2, \frac{1}{\sqrt{3}}e_3, \frac{1}{\sqrt{3}}e_3, \frac{1}{\sqrt{3}}e_3 \dots\}, \quad \{e_i\}_{i=1}^d \in \mathbb{R}^d$

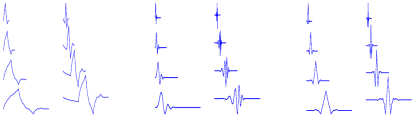
$$\sum_{k=1}^{p} |\langle x, a_k \rangle|^2 = \sum_{k=1}^{d} k \left| \left\langle x, \frac{1}{\sqrt{k}} e_k \right\rangle \right|^2 = \sum_{k=1}^{d} |\langle x, e_k \rangle|^2 = ||x||^2$$

Parseval frame for $\mathbb{R}^d$ with $A = B = 1$

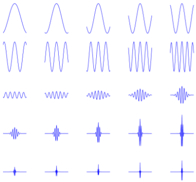
Frame examples (cont.)

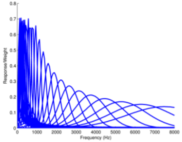
***Many* other useful examples** of frames:

- wavelets: dyadic scaling and translations [4]



- Gabor frames





- curvelets: scale, rotation, translation (tight frame)
- shearlets
- ...

[4] Image credit: L. Jacques, et. al., 2013

Frame data representation

Let $\mathcal{X} = \mathbb{R}^d, \mathcal{F} = \mathbb{R}^p$ and consider the representation

$$\Phi : \mathcal{X} \to \mathcal{F}, \quad \Phi(x) = (\langle x, a_1 \rangle, \dots, \langle x, a_p \rangle), \quad \forall x \in \mathcal{X}.$$

Remarks:

- **linear** map,
- can be identified with a $p \times d$ **rectangular matrix** F,

$$\Phi(x) = Fx, \quad \forall x \in \mathcal{X}$$

Metric properties of frame representations

Relaxed Parseval's identity

$$A\|x\|^2 \leq \|\Phi(x)\|^2 \leq B\|x\|^2, \qquad \forall x \in \mathcal{X}.$$

Stable representation/embedding

$$A\|x - x'\|^2 \leq \|\Phi(x) - \Phi(x')\|^2 \leq B\|x - x'\|^2, \qquad \forall x, x' \in \mathcal{X}.$$

Stable isometries: preserve distances (up to distortions), not isometries

Non-unitary frame operator

Remarks (cont.):

- linear map $\Phi(x) = Fx, \quad \forall x \in \mathcal{X}$
- F is **not unitary** $F^*F \neq I$

Note that

$$\langle Fx, z \rangle_{\mathcal{F}} = \sum_{k=1}^{p} \langle a_k, x \rangle z^k = \left\langle \sum_{k=1}^{p} a_k z^k, x \right\rangle,$$

then

$$F^* z = \sum_{k=1}^{p} a_k z^k, \quad \forall z \in \mathcal{F}$$

Frame operator

$$T = F^*F : \mathcal{X} \to \mathcal{X}, \quad Tx = \sum_{k=1}^{p} a_k \langle a_k, x \rangle, \quad \forall x \in \mathcal{X}.$$

Frame operator invertibility

Remarks (cont.):

- F is **not unitary** $T = F^*F \neq I$ ...
- ... however $T = F^*F$ is **invertible**.

$$T = F^*F : \mathcal{X} \to \mathcal{X}, \quad Tx = \sum_{k=1}^{p} a_k \langle a_k, x \rangle, \quad \forall x \in \mathcal{X}.$$

Proof.

1. $F^*z = \sum_{k=1}^{d} a_k z^k, \quad \forall z \in \mathcal{F}$
2. using linearity $\sum_{k=1}^{p} |\langle x, a_k \rangle|^2 = \|Fx\|_{\mathcal{F}}^2 = \langle Fx, Fx \rangle = \langle Tx, x \rangle, \ \forall x \in \mathcal{X}$.
3. rewrite frame bound

$$A \leq \frac{\langle Tx, x \rangle}{\|x\|^2} \leq B, \quad \forall x \in \mathcal{X}.$$

4. $\frac{\langle Tx, x \rangle}{\|x\|^2}$ is the Rayleigh quotient of T: *minimized by its smallest eigenvalue.*

□

Frame data reconstruction

Consider $\Psi : \mathcal{F} \to \mathcal{X}, \quad \mathcal{F} = \mathbb{R}^p, \quad \mathcal{X} = \mathbb{R}^d$

$$\Psi(z) = \sum_{k=1}^{p} \tilde{a}_k z^k, \quad \forall w \in \mathcal{F},$$

where

$$\tilde{a}_k = T^{-1} a_k, \quad k = 1, \dots, p, \quad T = F^* F$$

Remarks on Ψ

- **linear**,
- also as **rectangular matrix** $\tilde{F}$ (with suitable atoms as columns)

$$\Psi(z) = \tilde{F} z = \left(\langle z, \tilde{a}_1 \rangle, \dots, \langle z, \tilde{a}_p \rangle \right), \quad \forall z \in \mathcal{F}.$$

- **well defined** and **exact**, $\Psi \circ \Phi = I$.

Exact reconstruction

Remarks (cont.)

- Ψ is **well defined** and
- reconstruction is **exact**, $\Psi \circ \Phi = I$.

Proof.

For all $x \in \mathcal{X}$ with $z = Fx \in \mathcal{F}$, then

$$\Psi(z) = \sum_{k=1}^{p} \tilde{a}_k z^k = T^{-1} \sum_{k=1}^{p} a_k \langle x, a_k \rangle = T^{-1} Tx = x.$$

$\square$

Note:

It is also easy to check this by writing

$$\Psi(z) = \tilde{F} z = \left(\langle z, \tilde{a}_1 \rangle, \ldots, \langle z, \tilde{a}_p \rangle \right), \quad \forall z \in \mathcal{F}.$$

$$\Psi(z) = \Psi(\Phi(x) = \Psi(Fx) = \tilde{F} Fx = T^{-1} F^* F = T^{-1} Tx = x$$

Linear representation given a dictionary

Consider a **general (redundant) dictionary**

$$\{a_1, \ldots, a_p\}, \quad a_k \in \mathbb{R}^d, \quad p > d,$$

spanning a space of dimension smaller than d.

Linear representation letting $\mathcal{F} = \mathbb{R}^p$

$$\Phi : \mathcal{X} \to \mathcal{F}, \quad \Phi(x) = (\langle x, a_1 \rangle, \ldots, \langle x, a_p \rangle) = w, \quad \forall x \in \mathcal{X}.$$

- $\Phi(x)$ identified by $p \times d$ matrix $Cx = w$

Linear reconstruction given a dictionary

Reconstruction problem is **ill-posed**:

$$\textit{find } x \in \mathcal{X} \textit{ by solving } \Phi(x) = Cx = w.$$

Define reconstruction by the **minimization problem**

$$\Psi(w) = \arg\min_{x \in \mathcal{X}} \|x\|_2, \quad \text{subject to} \quad \Phi(x) = w,$$

or using the linear maps

$$Dw = \arg\min_{x \in \mathcal{X}} \|x\|_2, \quad \text{subject to} \quad Cx = w,$$

Given the pseudoinverse of the representation,

$$D = C^{\dagger} = (C^* C)^{-1} C^*$$

.

Representation and reconstruction given a dictionary II

Complementary point of view

Consider the reconstruction (de-coding)

$$\Psi : \mathcal{F} \to \mathcal{X}, \quad x = Dw = \sum_{k=1}^{d} a_k w^k, \quad \forall w \in \mathcal{F},$$

...and then an associate representation (coding)

$$\Phi(x) = \arg\min_{w \in \mathcal{F}} \|w\|_2 \,, \quad \text{subject to} \quad Dw' = x,$$

so that

$$C = D^\dagger$$

.

Non-linear reconstruction given a dictionary

Representation and reconstruction from **regularizers** other than the square norm.

$$\Phi(x) = \arg\min_{w \in \mathcal{F}} R(w), \quad \text{subject to} \quad Dw = x.$$

e.g., sparsity:

$$\Phi(x) = \arg\min_{w \in \mathcal{F}} \|w\|_1, \quad \text{subject to} \quad Dw = x.$$

Remarks:

- **sparsity**: characterize data by few atoms.
- redundant (overcomplete) dictionaries.
- solution cannot be computed in closed form:
  - involves solving a **convex, non-smooth problem**,
  - e.g. *splitting methods*.

Noisy data

$$\Phi(x) = \underset{w \in \mathcal{F}}{\arg\min}\, R(w), \quad \text{subject to} \quad \|Dw - x\|^2 \leq \delta, \;\; \delta > 0$$

where δ is a precision related to the noise level.

Alternative formulations:

- Constrained:

$$\Phi(x) = \underset{w \in \mathcal{F}}{\arg\min}\, \|Dw - x\|^2, \quad \text{subject to} \quad R(w) \leq r, \quad r > 0$$

- Penalized:

$$\Phi(x) = \underset{w \in \mathcal{F}}{\arg\min}\, \|Dw - x\|^2 + \lambda R(w), \quad \lambda > 0$$

Remarks

- Suitable choice of a **dictionary** allows to define a representation and robust reconstruction.
- Reconstruction/representation can become **harder** as more general dictionaries are considered.
- **Redundancy** allows for flexibility possibly at the expense of representation dimensionality.

Q:Is it is possible to work with more **compact** representations?

Randomized linear representation

Consider a set of **random atoms** of size smaller then data dimension:

$$\{a_1, \dots, a_k\}, \quad k < d.$$

where the atoms are, for example, vectors with i.i.d. normal entries.

Randomized representation (of reduced dimensionality):

$$\Phi : \mathcal{X} \to \mathcal{F} = \mathbb{R}^k, \quad w = \Phi(x) = (\langle x, a_1 \rangle, \dots, \langle x, a_k \rangle), \quad \forall x \in \mathcal{X}.$$

Johnson-Lindenstrauss Lemma

Randomized representation defines a **stable embedding** (ϵ-isometry), i.e.

$$(1-\epsilon)\left\|x-x'\right\|_2^2 \leq \left\|\Phi(x)-\Phi(x')\right\|_2^2 \leq (1+\epsilon)\left\|x-x'\right\|_2^2$$

for given $\epsilon \in (0,1)$, with probability $1-\delta$ and for all $x, x' \in Q \subset \mathcal{X}$, if the number of random projections is

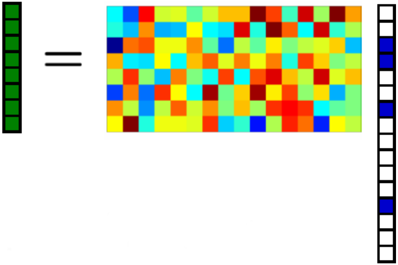
$$k = O\left(\frac{\log(|Q|/\delta)}{\epsilon^2}\right)$$

Random matrix design

Restricted Isometry Property (RIP)

$$(1-\delta_s)\,\|x\|_2^2 \leq \|CX\|_2^2 \leq (1+\delta_s)\,\|x\|_2^2$$

for matrix C, x being $s-$**sparse** with $0 < \delta_s < 1$.



Random matrices have shown to have bounded δ_s.

- Gaussian, Bernoulli, and partial Fourier satisfy RIP with $k \approx s$.

Compressed Sensing

Exact reconstruction is possible provided:

- class of data Q is sufficiently **"nice"** (e.g., sparse vectors)
- number of projections is sufficiently **large**,
- projection matrix is nearly orthonormal (RIP).

Example:
If $\mathcal{C}$ is the of $s-$**sparse** vectors and $k \sim s \log \frac{d}{s}$, then exact reconstruction is possible with high probability, considering

$$\Psi(w) = \arg\min_{x \in \mathcal{X}} \|x\|_1, \quad \text{subject to} \quad \Phi(x) = Cx = w,$$

Randomized representation beyond linearity

CS extensions consider **non-linear** randomized representations

$$\Phi : \mathcal{X} \to \mathcal{F}, \quad w = \Phi(x) = (\sigma(\langle x, a_1 \rangle), \dots, \sigma(\langle x, a_k \rangle)), \quad \forall x \in \mathcal{X}$$

for some non-linear function $\sigma : \mathbb{R} \to \mathbb{R}$.

From signal processing to kernel machines

So far:

- unitary, frames & dictionary representations
- randomized representations & compressed sensing

Note: interplay between distance preservation and reconstruction.

Such methods:

- lead to **parametric** supervised learning models

$$f(x) = \langle w, \Phi(x) \rangle_{\mathcal{F}}, \quad \Phi : \mathcal{X} \to \mathbb{R}^p, \ p < \infty,$$

- mostly restricted to **vector data**.

Recall: **Kernel methods** provide a way to tackle both issues.

Few remarks on kernels

- Computational complexity is independent of feature space dimension... but becomes **prohibitive for large scale** learning

  $\Rightarrow$ subsampling/randomized approximations.

- While flexible, kernel methods rely on the choice of the kernel...

  can it be **learned**?

  $\Rightarrow$ supervised multiple kernel learning.

Wrap-up

This class: Data representations by design

- orthonormal basis,
- frames,
- dictionaries,
- random projections,
- kernels.

...based on **prior assumptions** about the problem or data.

Next class: Can they be **learned from data**?

- Part II: Data representations by **learning**
- Part III: **Deep** data representations