Harmonic Analysis on Graphs
Global vs. Multiscale Approaches

Boaz Nadler

Weizmann Institute of Science, Rehovot, Israel

July 2011

Joint work with
Matan Gavish (WIS/Stanford), Ronald Coifman (Yale), *ICML 10'*

Challenge: Organization / Understanding of Data

In many fields massive amounts of data collected or generated,

EXAMPLES:

financial data, multi-sensor data, simulations, documents, web-pages, images, video streams, medical data, astrophysical data, etc.

Challenge: Organization / Understanding of Data

In many fields massive amounts of data collected or generated,

EXAMPLES:

financial data, multi-sensor data, simulations, documents, web-pages, images, video streams, medical data, astrophysical data, etc.

Need to organize / understand their structure
Inference / Learning from data

Classical vs. Modern Data Analysis Setup and Tasks

Classical Setup:

- Data typically in a (low-dimensional) Euclidean Space.
- Small to medium sample sizes ($n < 1000$)
- Either all data unlabeled (unsupervised) or all labeled (supervised).

Classical vs. Modern Data Analysis Setup and Tasks

Classical Setup:

- Data typically in a (low-dimensional) Euclidean Space.
- Small to medium sample sizes ($n < 1000$)
- Either all data unlabeled (unsupervised) or all labeled (supervised).

Modern Setup:

- high-dimensional data, or data encoded as a *graph*.
- Huge datasets, with $n = 10^6$ samples or more. Few labeled data.

Classical vs. Modern Data Analysis Setup and Tasks

Classical Setup:

- Data typically in a (low-dimensional) Euclidean Space.
- Small to medium sample sizes ($n < 1000$)
- Either all data unlabeled (unsupervised) or all labeled (supervised).

Modern Setup:

- high-dimensional data, or data encoded as a *graph*.
- Huge datasets, with $n = 10^6$ samples or more. Few labeled data.

The well developed and understood standard tools of statistics not always applicable

Classical vs. Modern Data Analysis Setup and Tasks

Classical Setup:

- Data typically in a (low-dimensional) Euclidean Space.
- Small to medium sample sizes ($n < 1000$)
- Either all data unlabeled (unsupervised) or all labeled (supervised).

Modern Setup:

- high-dimensional data, or data encoded as a *graph*.
- Huge datasets, with $n = 10^6$ samples or more. Few labeled data.

The well developed and understood standard tools of statistics not always applicable

Question: Harmonic Analysis in such settings

Harmonic Analysis and Learning

In the past 20 years (multiscale) harmonic analysis had profound impact on statistics and signal processing.

Harmonic Analysis and Learning

In the past 20 years (multiscale) harmonic analysis had profound impact on statistics and signal processing.

Well developed theory, long tradition:

$$\textit{geometry of space } X \Rightarrow \textit{bases for } \{f : X \to \mathbb{R}\}$$

Harmonic Analysis and Learning

In the past 20 years (multiscale) harmonic analysis had profound impact on statistics and signal processing.

Well developed theory, long tradition:

$$\textit{geometry of space } X \Rightarrow \textit{bases for } \{f : X \to \mathbb{R}\}$$

Simplest Example: $X = [0, 1]$

Construct (multiscale) basis $\{\Psi_i\}$, that allows control of $|\langle f, \Psi_i\rangle|$ for some (smooth) class of functions f.

Theorem: ψ smooth wavelet, $\psi_{\ell,k}$ - wavelet basis, $f : [0, 1] \to \mathbb{R}$, then

$$\boxed{|f(x) - f(y)| < C|x - y|^{\alpha} \;\Leftrightarrow\; |\langle f, \psi_{\ell,k}\rangle| \leq C' 2^{-\ell(\alpha+1/2)}}$$

Setting

Harmonic analysis wisdom for $f : \mathbb{R}^d \to \mathbb{R}$

Setting

Harmonic analysis wisdom for $f : \mathbb{R}^d \to \mathbb{R}$

- Expand f in orthonormal basis $\{\psi_i\}$ (e.g. wavelet)

$$f = \sum_i \langle f, \psi_i \rangle \psi_i$$

Setting

Harmonic analysis wisdom for $f : \mathbb{R}^d \to \mathbb{R}$

- Expand f in orthonormal basis $\{\psi_i\}$ (e.g. wavelet)

$$f = \sum_i \langle f, \psi_i \rangle \, \psi_i$$

- Shrink / estimate coefficients

Setting

Harmonic analysis wisdom for $f : \mathbb{R}^d \to \mathbb{R}$

- Expand f in orthonormal basis $\{\psi_i\}$ (e.g. wavelet)

$$f = \sum_i \langle f, \psi_i \rangle \psi_i$$

- Shrink / estimate coefficients
- Useful when f well approximated by a few terms ("fast coefficent decay" / sparsity)

Setting

Harmonic analysis wisdom for $f : \mathbb{R}^d \to \mathbb{R}$

- Expand f in orthonormal basis $\{\psi_i\}$ (e.g. wavelet)

$$f = \sum_i \langle f, \psi_i \rangle \psi_i$$

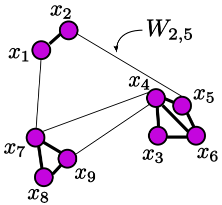
- Shrink / estimate coefficients
- Useful when f well approximated by a few terms ("fast coefficent decay" / sparsity)

Can this work on general datasets?

Need data-adaptive basis $\{\psi_i\}$ for space of functions $f : X \to \mathbb{R}$

Harmonic Analysis on Graphs

Setup: We are given dataset $X = \{x_1, \ldots, x_N\}$
with similarity / affinity matrix $W_{i,j}$



Goal: Statistical inference of *smooth* $f : X \to \mathbb{R}$

- Denoise f
- SSL / Regression / classification: extend f from $\tilde{X} \subset X$ to X

Semi-Supervised Learning

In many applications - easy to collect lots of unlabeled data, BUT labeling the data is expensive.

Question: Given (small) labeled set $\tilde{X} \subset X = \{x_i, y_i\}$ $(y = f(x))$, construct $\hat{f}$ to label rest of X.

Semi-Supervised Learning

In many applications - easy to collect lots of unlabeled data, BUT labeling the data is expensive.

Question: Given (small) labeled set $\tilde{X} \subset X = \{x_i, y_i\}$ $(y = f(x))$, construct $\hat{f}$ to label rest of X.

Key Assumption:

function $f(x)$ has some smoothness w.r.t graph affinities $W_{i,j}$.

otherwise - unlabeled data is useless or may even harm prediction.

The Graph Laplacian

In most previous approaches, key object is

GRAPH LAPLACIAN:

$$L = D - W$$

where $D_{ii} = \sum_j W_{ij}$,

Global Graph Laplacian Based SSL Methods

Zhu, Ghahramani, Lafferty, SSL using Gaussian fields and harmonic functions [ICML, 2003] , Azran [ICML 2007]

$$f(x) = \arg \min_{f(x_i)=y_i} \frac{1}{n^2} \sum_{i,j=1}^{n} W_{i,j}(f_i - f_j)^2 = \arg\min \frac{1}{n^2} \mathbf{f}^T L \mathbf{f}$$

Y. Bengio, O. Delalleau, N. Le Roux, [2006]
D. Zhou, O. Bousquet, T. Navin Lal, J. Weston, B. Scholkopf [NIPS 2004]

$$f(x) = \arg\min \left[\frac{1}{n^2} \sum_{i,j=1}^{n} W_{i,j}(f_i - f_j)^2 + \lambda \sum_{i=1}^{\ell} (f_i - y_i)^2 \right]$$

Global Graph-Laplacian Based SSL

Belkin & Niyogi (2003): Given similarity matrix W find first few eigenvectors of Graph Laplacian, $L\mathbf{e}_j = (D - W)\mathbf{e}_j = \lambda_j \mathbf{e}_j$.
Expand

$$\hat{f} = \sum_{j=1}^{p} a_j \mathbf{e}_j$$

Estimate coefficients a_j from labeled data.

Statistical Analysis of Laplacian Regularization

[N., Srebro, Zhou, NIPS 09']

Theorem: In the limit of large unlabeled data from Euclidean space, with $W_{i,j} = K(\|x_i - x_j\|)$, Graph Laplacian Regularizaton Methods

$$f(x) = \arg\min \left[\frac{1}{n^2} \sum_{i,j=1}^{n} W_{i,j}(f_i - f_j)^2 + \lambda \sum_{i=1}^{\ell} (f_i - y_i)^2 \right]$$

are weill posed for underlying data in 1-d, but *are not well posed* for data in dimension $d \geq 2$.
In particular, in limit of infinite unlabeled data, $f(x) \to const$ at all unlabeled x.

Eigenvector-Fourier Methods

Belkin, Niyogi [2003], suggested a different approach based on the Graph Laplacian $L = W - D$:

Given similarity W, find first few eigenvectors $(W - D)\mathbf{e}_j = \lambda_j \mathbf{e}_j$, Expand

$$\hat{y}(x) = \sum_{j=1}^{p} a_j \mathbf{e}_j$$

Find coefficients a_j by least squares,

$$(\hat{a}_1, \ldots, \hat{a}_p) = \arg\min \sum_{j=1}^{l} (y_j - \hat{y}(x_j))^2.$$

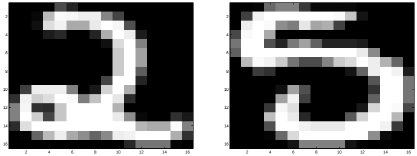
Toy example

Example: USPS benchmark

Toy example

Example: USPS benchmark

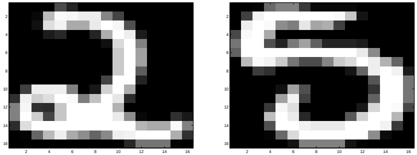
- X is USPS (ML benchmark) as 1500 vectors in $\mathbb{R}^{16\times 16} = \mathbb{R}^{256}$



Toy example

Example: USPS benchmark

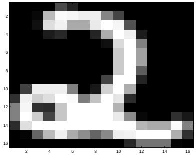
- X is USPS (ML benchmark) as 1500 vectors in $\mathbb{R}^{16\times 16} = \mathbb{R}^{256}$
- Affinity $W_{i,j} = \exp\left(-\|x_i - x_j\|^2\right)$

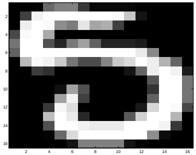


Toy example

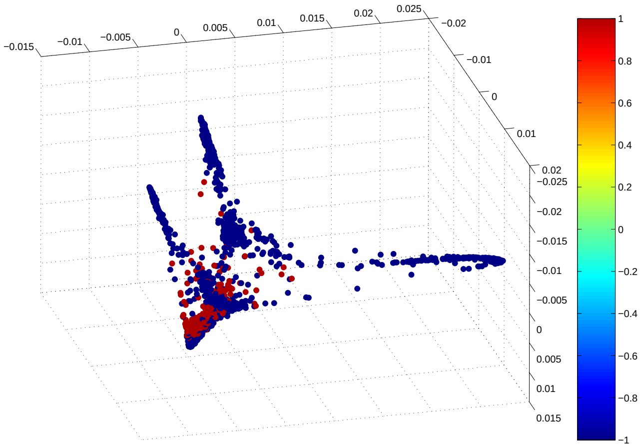
Example: USPS benchmark

- X is USPS (ML benchmark) as 1500 vectors in $\mathbb{R}^{16\times 16} = \mathbb{R}^{256}$
- Affinity $W_{i,j} = \exp\left(-\left\|x_i - x_j\right\|^2\right)$
- $f : X \to \{1, -1\}$ is the class label.





Toy example: visualization by kernel PCA



Toy example: Prior art?

Toy example: Prior art?

Generalizing Fourier: The Graph Laplacian eigenbasis

Toy example: Prior art?

Generalizing Fourier: The Graph Laplacian eigenbasis

- Take $(W - D)\psi_i = \lambda_i \psi_i$ where $D_{i,i} = \sum_j W_{i,j}$

Toy example: Prior art?

Generalizing Fourier: The Graph Laplacian eigenbasis

- Take $(W - D)\psi_i = \lambda_i \psi_i$ where $D_{i,i} = \sum_j W_{i,j}$
- *(Belkin, Niyogi 2004)* and others

Toy example: Prior art?

Generalizing Fourier: The Graph Laplacian eigenbasis

- Take $(W - D)\psi_i = \lambda_i \psi_i$ where $D_{i,i} = \sum_j W_{i,j}$
- *(Belkin, Niyogi 2004)* and others

In Euclidean setting

Toy example: Prior art?

Generalizing Fourier: The Graph Laplacian eigenbasis

- Take $(W - D)\psi_i = \lambda_i \psi_i$ where $D_{i,i} = \sum_j W_{i,j}$
- *(Belkin, Niyogi 2004)* and others

In Euclidean setting

- Typical coefficient decay rate in Fourier basis: **polynomial**

Toy example: Prior art?

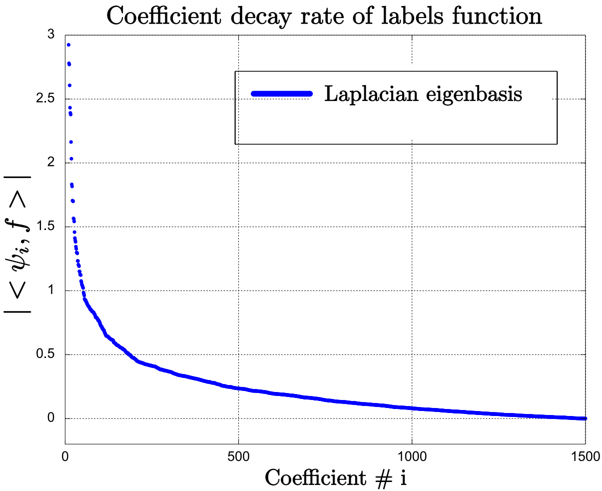
Generalizing Fourier: The Graph Laplacian eigenbasis

- Take $(W - D)\psi_i = \lambda_i \psi_i$ where $D_{i,i} = \sum_j W_{i,j}$
- *(Belkin, Niyogi 2004)* and others

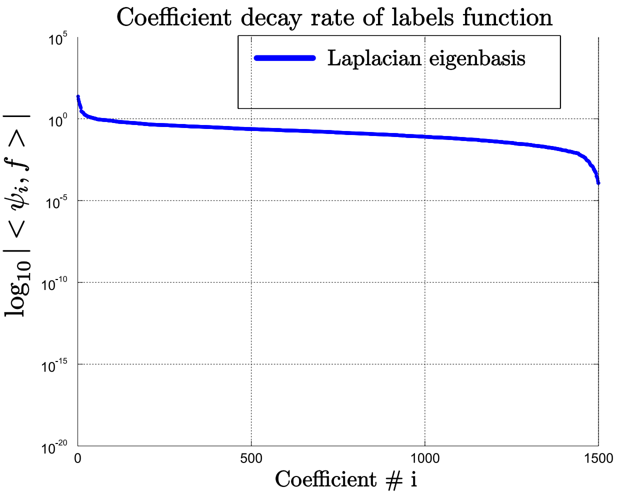
In Euclidean setting

- Typical coefficient decay rate in Fourier basis: **polynomial**
- **But** in wavelet bases: **exponential**

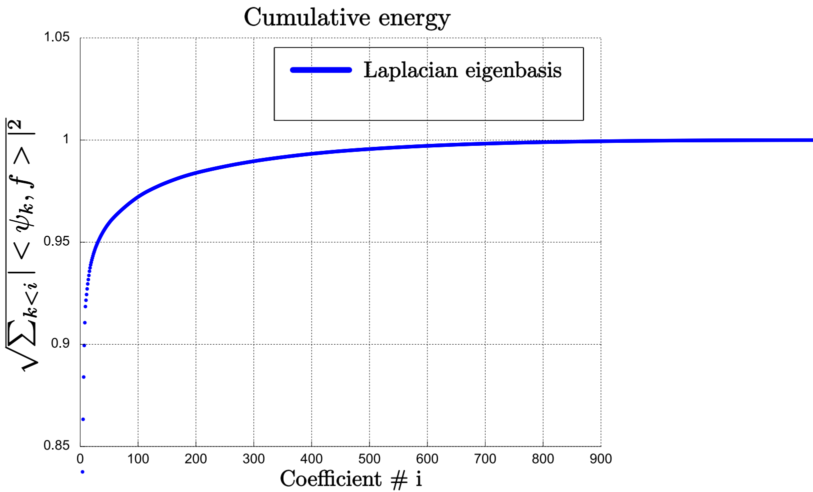
Toy example: Graph Laplacian Eigenbasis



Toy example: Graph Laplacian Eigenbasis



Toy example: Graph Laplacian Eigenbasis



Challenge

Challenge

Challenge: build multiscale "wavelet-like" bases on X

Challenge

Challenge: build multiscale "wavelet-like" bases on X

1. Construct adaptive multiscale basis on general dataset

Challenge

Challenge: build multiscale "wavelet-like" bases on X

1. Construct adaptive multiscale basis on general dataset
2. Investigate: coefficient $\langle f, \psi_i \rangle$ decay rate $\Leftrightarrow$ f smooth

Challenge

Challenge: build multiscale "wavelet-like" bases on X

1. Construct adaptive multiscale basis on general dataset
2. Investigate: coefficient $\langle f, \psi_i \rangle$ decay rate $\Leftrightarrow$ f smooth

Previous Works

- Difffusion Wavelets [Coifman and Maggioni]
- Multiscale Methods for Data on Graphs [Jansen, Nason, Silverman]
- Wavelets on Graphs via Spectral Graph Theory [Hammond et al.]

Challenge

Challenge: build multiscale "wavelet-like" bases on X

1. Construct adaptive multiscale basis on general dataset
2. Investigate: coefficient $\langle f, \psi_i \rangle$ decay rate $\Leftrightarrow$ f smooth

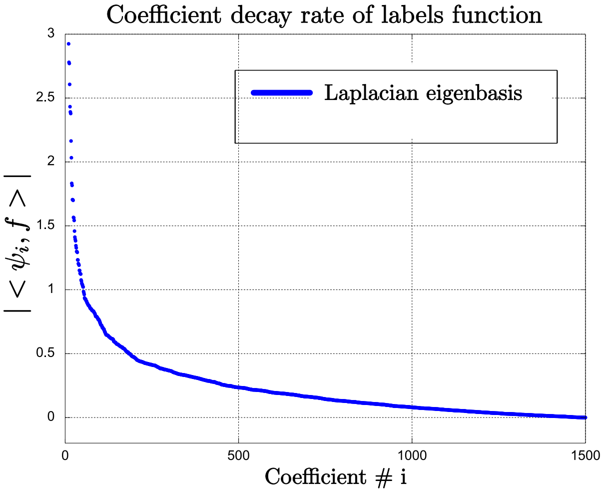
Previous Works

- Difffusion Wavelets [Coifman and Maggioni]
- Multiscale Methods for Data on Graphs [Jansen, Nason, Silverman]
- Wavelets on Graphs via Spectral Graph Theory [Hammond et al.]

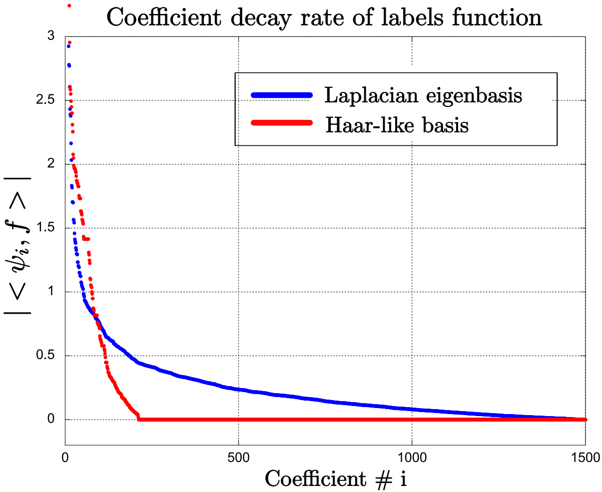
Our Work

- (Relavitely) Simple Construction
- Accompanying Theory.

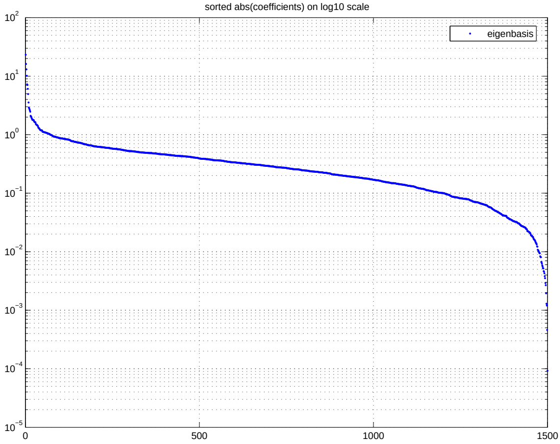
Toy example: intriguing experiment



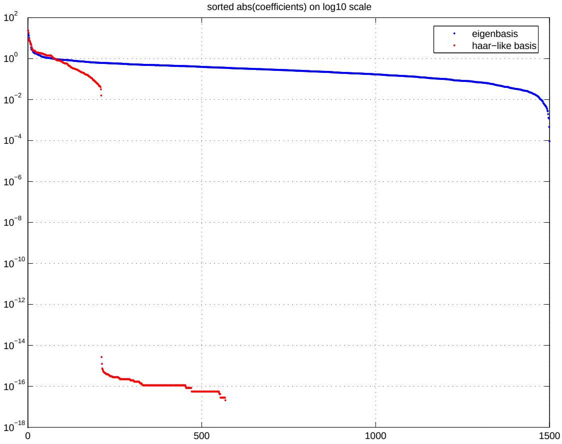
Toy example: intriguing experiment



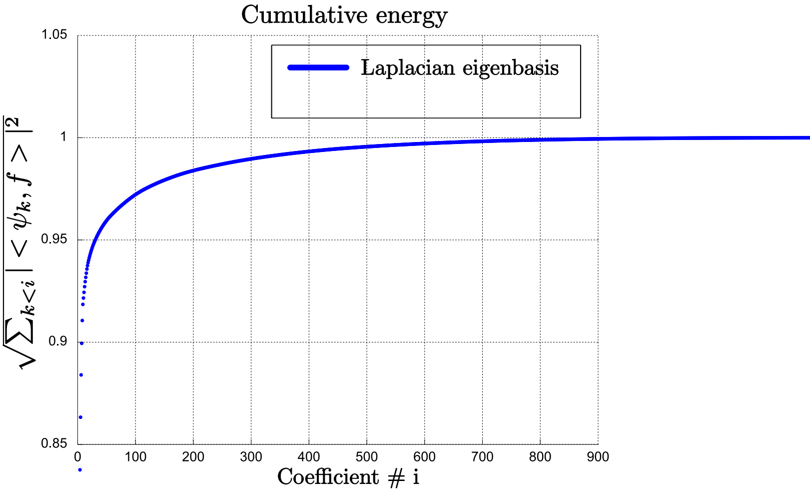
Toy example: intriguing experiment



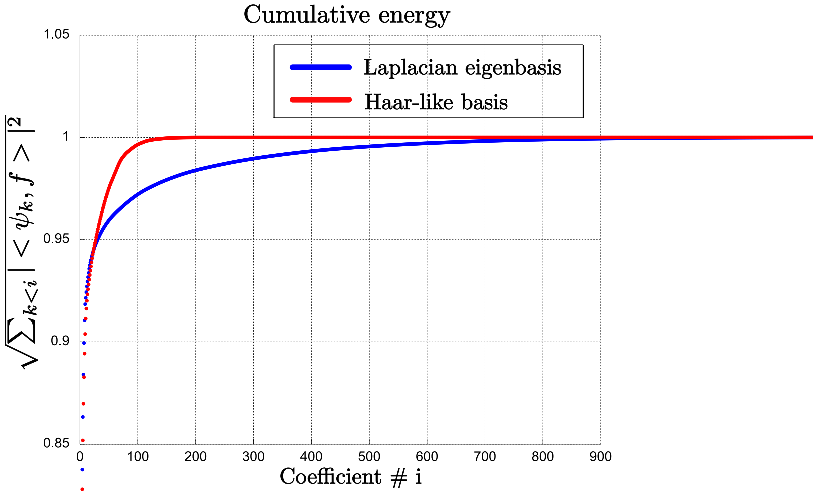
Toy example: intriguing experiment



Toy example: intriguing experiment



Toy example: intriguing experiment



Challenge and Result

Challenge and Result

Challenge: build "wavelet" bases on X

Challenge and Result

Challenge: build "wavelet" bases on X

1. Construct adaptive multiscale basis on general dataset

Challenge and Result

Challenge: build "wavelet" bases on X

1. Construct adaptive multiscale basis on general dataset
2. Investigate: coefficient $\langle f, \psi_i \rangle$ decay rate $\Leftrightarrow$ f smooth

Challenge and Result

Challenge: build "wavelet" bases on X

1. Construct adaptive multiscale basis on general dataset
2. Investigate: coefficient $\langle f, \psi_i \rangle$ decay rate $\Leftrightarrow$ f smooth

Key Results

Challenge and Result

Challenge: build "wavelet" bases on X

1. Construct adaptive multiscale basis on general dataset
2. Investigate: coefficient $\langle f, \psi_i \rangle$ decay rate $\Leftrightarrow$ f smooth

Key Results

1. Partition tree on X induces “wavelet” *Haar-like bases*

Challenge and Result

Challenge: build "wavelet" bases on X

1. Construct adaptive multiscale basis on general dataset
2. Investigate: coefficient $\langle f, \psi_i \rangle$ decay rate $\Leftrightarrow$ f smooth

Key Results

1. Partition tree on X induces "wavelet" *Haar-like bases*
2. Assuming "Balanced" tree

$$\boxed{f \text{smooth} \iff \text{fast coefficient decay}}$$

Challenge and Result

Challenge: build "wavelet" bases on X

1. Construct adaptive multiscale basis on general dataset
2. Investigate: coefficient $\langle f, \psi_i \rangle$ decay rate $\Leftrightarrow$ f smooth

Key Results

1. Partition tree on X induces "wavelet" *Haar-like bases*
2. Assuming "Balanced" tree

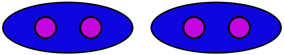
$$\boxed{f\text{smooth} \iff \text{fast coefficient decay}}$$

3. Novel SSL scheme with learning guarantees assuming smooth functions on tree.

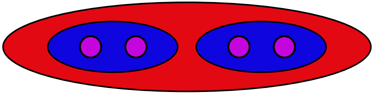
The Haar Basis on $[0, 1]$

The Haar Basis on $[0,1]$

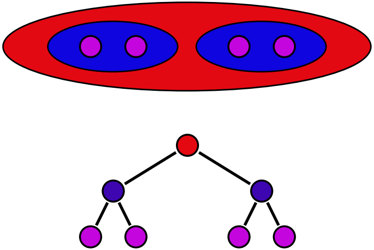
The Haar Basis on $[0, 1]$



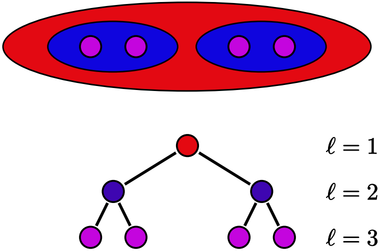
The Haar Basis on $[0, 1]$



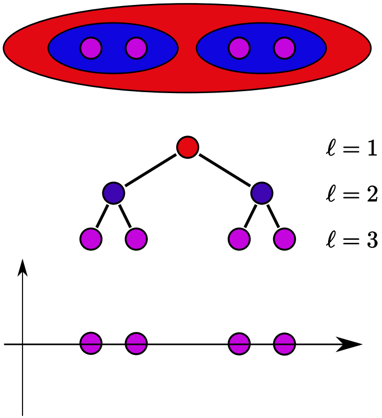
The Haar Basis on $[0, 1]$



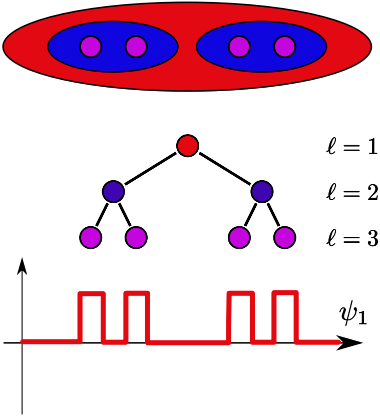
The Haar Basis on $[0,1]$



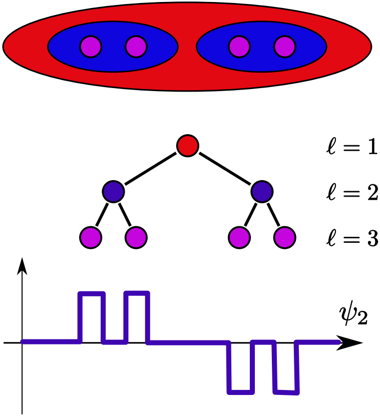
The Haar Basis on $[0, 1]$



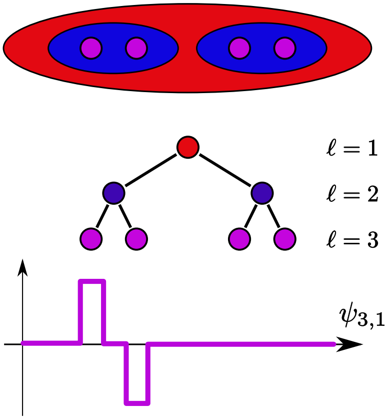
The Haar Basis on [0, 1]



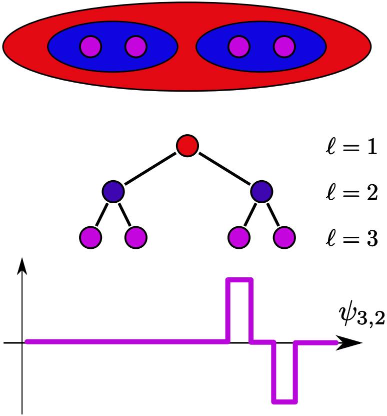
The Haar Basis on $[0, 1]$



The Haar Basis on $[0, 1]$

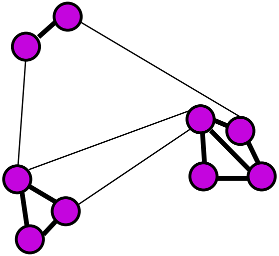


The Haar Basis on $[0, 1]$

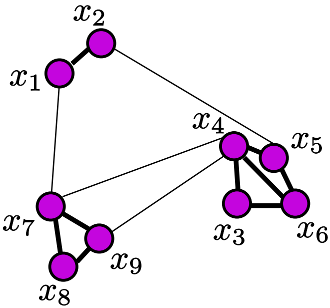


Hierarchical partition of X

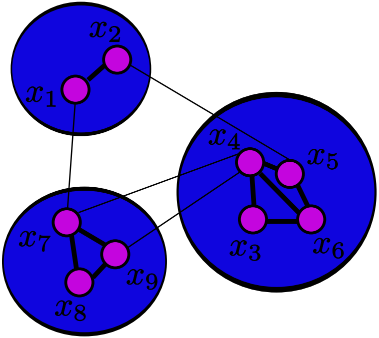
Hierarchical partition of X



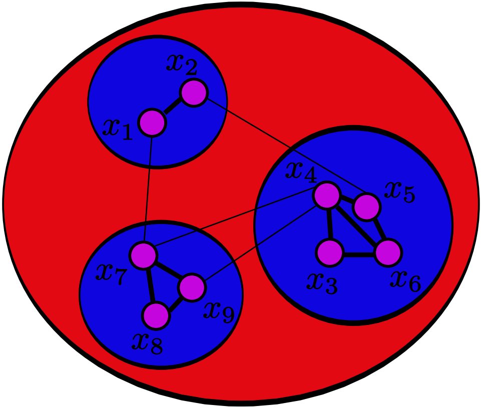
Hierarchical partition of X



Hierarchical partition of X

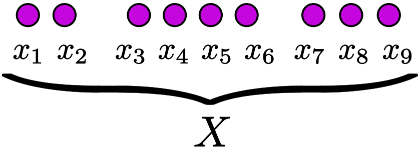


Hierarchical partition of X



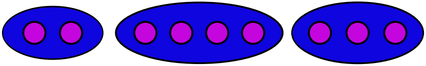
Partition Tree (Dendrogram)

Partition Tree (Dendrogram)

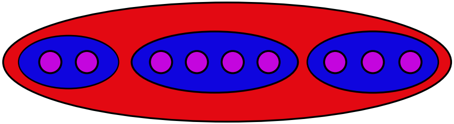


Partition Tree (Dendrogram)

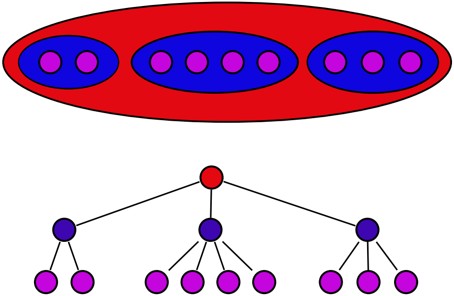
Partition Tree (Dendrogram)



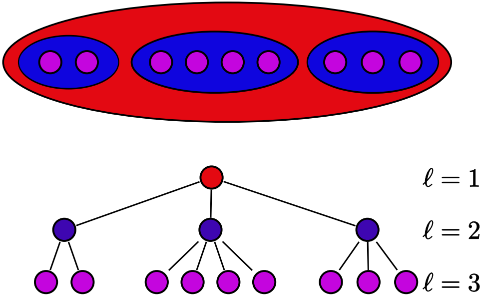
Partition Tree (Dendrogram)



Partition Tree (Dendrogram)



Partition Tree (Dendrogram)



Partition Tree $\Rightarrow$ Haar-like basis

[Lee, N., Wasserman, AOAS 08']

Simple Observation:

Hierarchical Tree $\rightarrow$ Mutli-Resolution Analysis of Space of Functions
$\rightarrow$ Haar-like multiscale basis

Partition Tree $\Rightarrow$ Haar-like basis

[Lee, N., Wasserman, AOAS 08']

Simple Observation:

Hierarchical Tree $\rightarrow$ Mutli-Resolution Analysis of Space of Functions
$\rightarrow$ Haar-like multiscale basis

$V = \{f \,|\, f : X \rightarrow \mathbb{R}\}$
$V^\ell = \{f \,|\, f$ constant at partitions at level $\ell\}$

Partition Tree $\Rightarrow$ Haar-like basis

[Lee, N., Wasserman, AOAS 08']

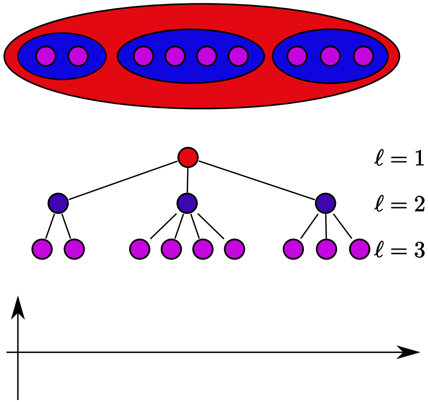
Simple Observation:

Hierarchical Tree $\rightarrow$ Mutli-Resolution Analysis of Space of Functions
$\rightarrow$ Haar-like multiscale basis

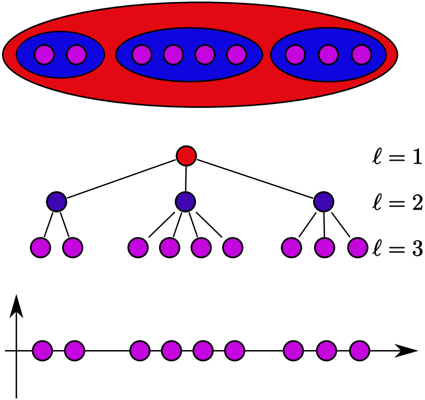
$V = \{f \,|\, f : X \rightarrow \mathbb{R}\}$
$V^\ell = \{f \,|\, f \text{ constant at partitions at level } \ell\}$

$$V^1 \subset V^2 \subset \ldots V^L = V$$

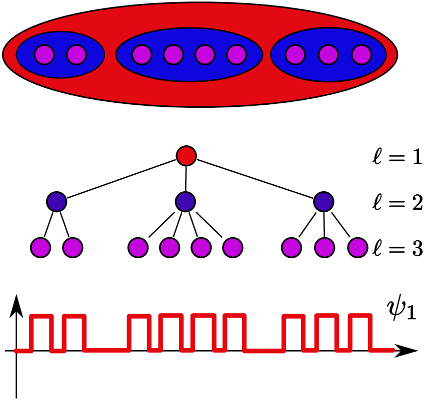
Partition Tree ⇒ Haar-like basis



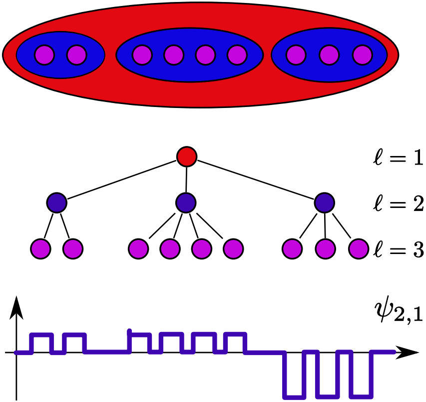
Partition Tree $\Rightarrow$ Haar-like basis



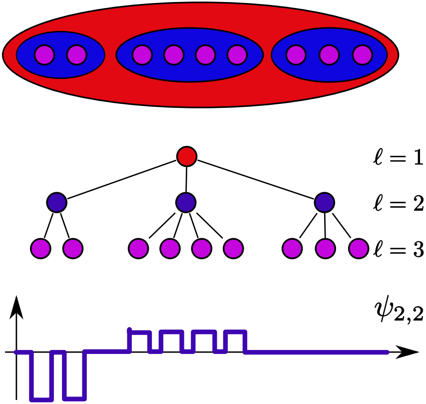
Partition Tree ⇒ Haar-like basis



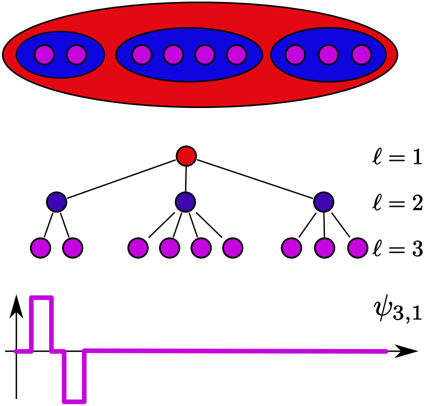
Partition Tree $\Rightarrow$ Haar-like basis



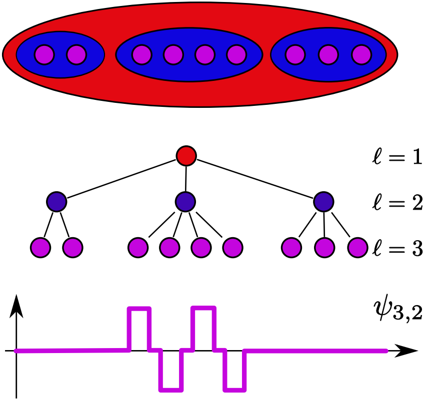
Partition Tree $\Rightarrow$ Haar-like basis



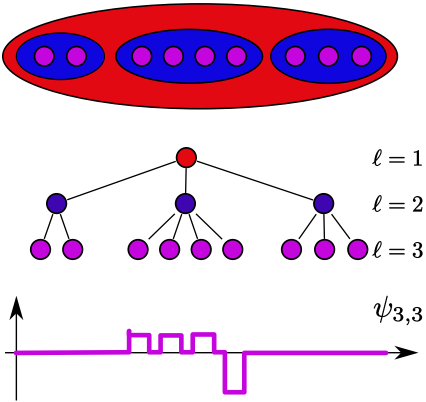
Partition Tree $\Rightarrow$ Haar-like basis



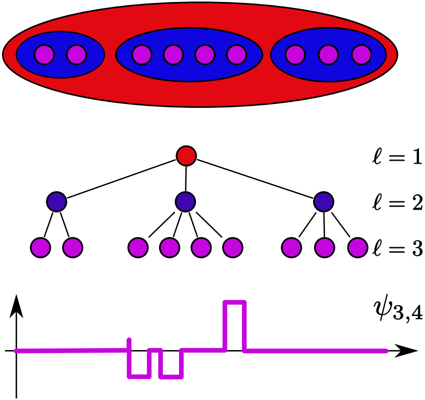
Partition Tree $\Rightarrow$ Haar-like basis



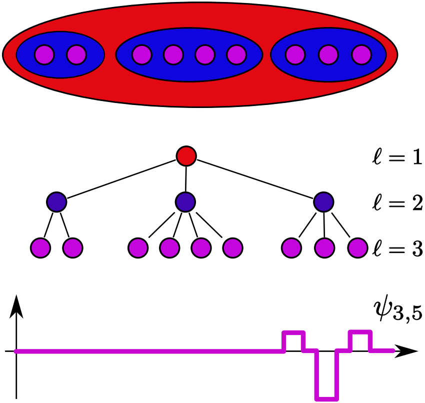
Partition Tree $\Rightarrow$ Haar-like basis



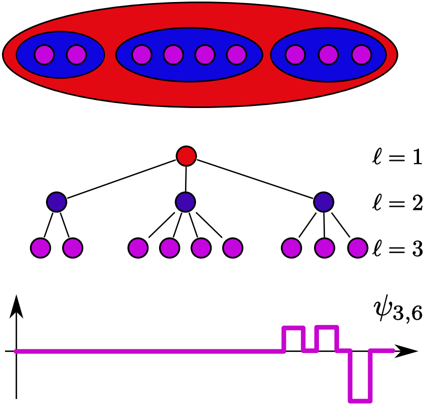
Partition Tree ⇒ Haar-like basis



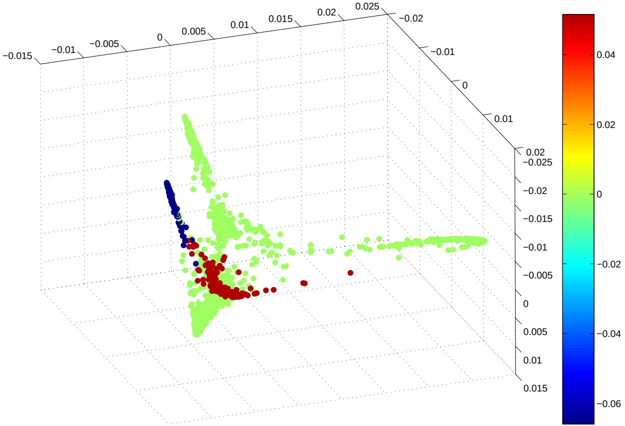
Partition Tree ⇒ Haar-like basis



Partition Tree ⇒ Haar-like basis



Toy example: Haar-like basis function



Smoothness $\Longleftrightarrow$ Coefficient decay

How to define smoothness ?

Smoothness $\Longleftrightarrow$ Coefficient decay

How to define smoothness ?

- Partition tree induces tree metric $d(x_i, x_j)$ [ultrametric]

Particular example of Space of Homogeneous Type [Coifman & Weiss]

Smoothness $\Longleftrightarrow$ Coefficient decay

How to define smoothness ?

- Partition tree induces tree metric $d(x_i, x_j)$ [ultrametric]
- Measure smoothness of $f : X \to \mathbb{R}$ in tree metric

Particular example of Space of Homogeneous Type [Coifman & Weiss]

Smoothness $\Longleftrightarrow$ Coefficient decay

How to define smoothness ?

- Partition tree induces tree metric $d(x_i, x_j)$ [ultrametric]
- Measure smoothness of $f : X \to \mathbb{R}$ in tree metric

Theorem:

Let $f : X \to \mathbb{R}$. Then

$$|f(x_i) - f(x_j)| \leq C d(x_i, x_j)^\alpha \;\Leftrightarrow\; |\langle f\,, \psi_{\ell,i}\rangle| \leq C' q^{-\ell(\alpha+1/2)}\,.$$

where

- q measures the **tree balance**
- α measures function smoothness w.r.t. tree

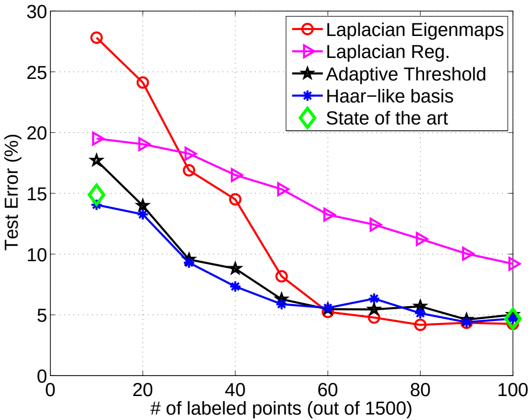
Particular example of Space of Homogeneous Type [Coifman & Weiss]

Application: SSL

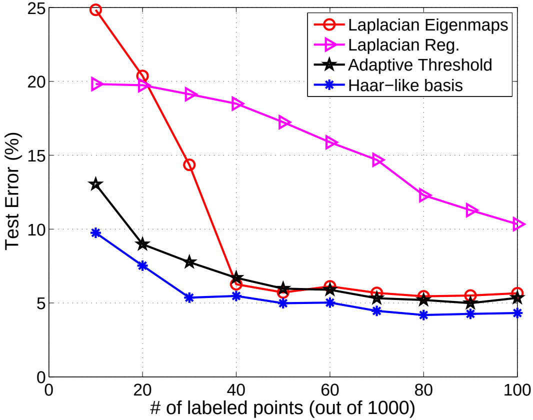
Given dataset X with weighted graph G, similarity matrix W, labeled points $\{x_i, y_i\}$,

1. Construct a balanced hierarchical tree of graph
2. Construct corresponding Haar-like basis
3. Estimate coefficients from labeled points

Toy Example: Benchmarks



Toy Example: MNIST 8 vs. {3,4,5,7}



Summary

Summary

1. A balanced partition tree induces a useful “wavelet” basis

Summary

1. A balanced partition tree induces a useful “wavelet” basis
2. *Designing* the basis allows a theory connecting Coefficients decay, smoothness and learnability

Summary

1. A balanced partition tree induces a useful "wavelet" basis
2. *Designing* the basis allows a theory connecting Coefficients decay, smoothness and learnability
3. "Wavelet arsenal" becomes available: wavelet shrinkage, etc

Summary

1. A balanced partition tree induces a useful “wavelet” basis
2. *Designing* the basis allows a theory connecting
   Coefficients decay, smoothness and learnability
3. ”Wavelet arsenal” becomes available: wavelet shrinkage, etc
4. Interesting harmonic analysis. Promising for data analysis?

Summary

1. A balanced partition tree induces a useful “wavelet” basis
2. *Designing* the basis allows a theory connecting Coefficients decay, smoothness and learnability
3. ”Wavelet arsenal” becomes available: wavelet shrinkage, etc
4. Interesting harmonic analysis. Promising for data analysis?
5. *Computational experiments motivate theory and vice-versa*

Summary

1. A balanced partition tree induces a useful “wavelet” basis
2. *Designing* the basis allows a theory connecting Coefficients decay, smoothness and learnability
3. ”Wavelet arsenal” becomes available: wavelet shrinkage, etc
4. Interesting harmonic analysis. Promising for data analysis?
5. *Computational experiments motivate theory and vice-versa*

The End

www.wisdom.weizmann.ac.il/∼nadler