02471 Machine Learning for Signal Processing

Sparse analysis models and time-frequency analysis

Tommy Sonne Alstrøm

Cognitive Systems section

DTU Compute
Department of Applied Mathematics and Computer Science

Outline

- Course admin
- Last week review
- Sparsity promoting algorithms
- Signal representation
- Time-frequency analysis
- Next Week

Material: ML 10.1, 10.2–10.2.1 (until p.476), 10.2.2 (until p.482), 10.5–10.6.

Feedback

- Feedback from last week:
- Feedback from you is a critical component for improving both the course and my teaching.
- Type of feedback
  - Mention one thing that worked?
  - Mention one thing should be improved (both in current lecture and last weeks exercise)?
  - Mention one thing you would change if you gave the lecture.

Problem sets

- Problem set 2 due Sunday 27/10 at 23.59.
- Problem set 2 re-submissions due Sunday 10/11 at 23.59.
- Problem set 3 is available from next week, and is due 15/12 23.59. Counts 20% towards the final grade.

Last week review

What is sparsity-aware learning

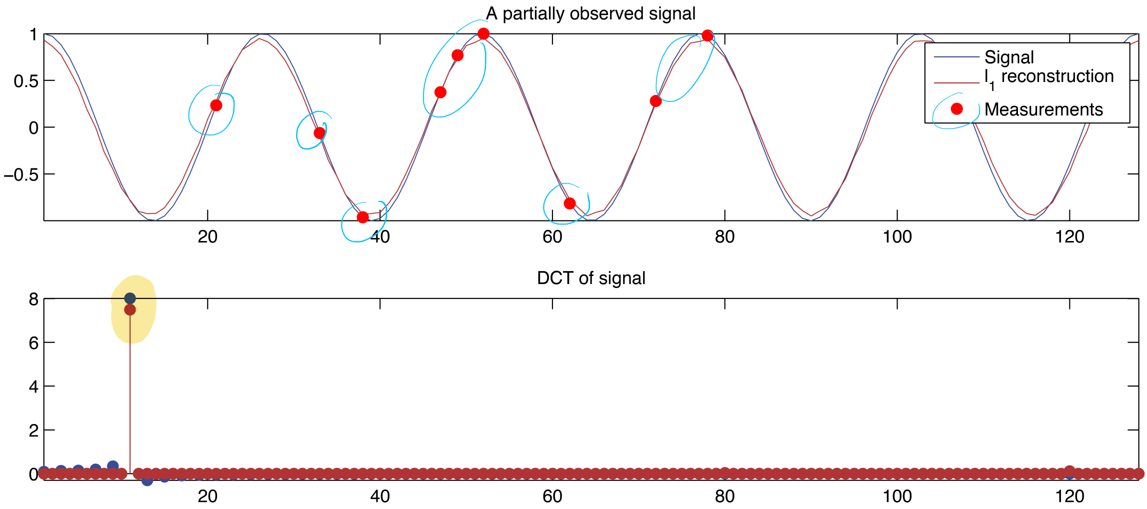
In a number of practical problems, it is known that either the underlying model is sparse or it is sparse in a transform domain, e.g., in the Fourier transform domain.

By sparse models is meant that most of the unknown coordinates in the model are zero.

Leads to compressed sensing, where the goal is to directly acquire as few samples as possible that encode the minimum information, which is needed to obtain a compressed signal representation.

The solution is often found using ℓ_1 norm regularization, and the corresponding solution is called LASSO (least absolute shrinkage and selection operator).

Signal reconstruction using ℓ_1



The ℓ_p norm

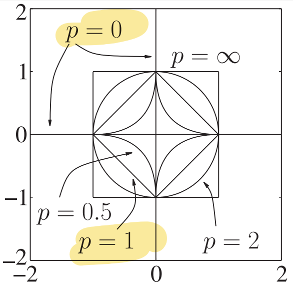
The ℓ_p norm

$$\|\boldsymbol{\theta}\|_p := \left(\sum_{i=1}^{l} |\theta_i|^p\right)^{1/p}, \quad 0 < p < \infty$$

$$\|\boldsymbol{\theta}\|_\infty := \arg\max_i |\theta_i| \quad \text{(max element)}$$

$$\|\boldsymbol{\theta}\|_0 := |\{i \,|\, \theta_i \neq 0, i = 1, \cdots, l\}| \quad \text{(number of nonzeros)}$$

l is the size of vector $\boldsymbol{\theta}$.



Notation remarks:
$|x|$ is the numerical value when $x \in \mathbb{R}$.
$|\mathcal{X}|$ is the cardinality (size) of the set $\mathcal{X}$.

LASSO minimization problem

LASSO minimization

LS Rego

$$\hat{\boldsymbol{\theta}}_1 = \arg\min_{\boldsymbol{\theta}} J(\boldsymbol{\theta}, \lambda) = \sum_{i=0}^{n} (y_i - \boldsymbol{\theta}^T \boldsymbol{x}_i)^2 + \lambda \|\boldsymbol{\theta}\|_1$$

Nomenclature

- $\hat{\boldsymbol{\theta}}_{LS}$ denotes the least squares solution
- $\hat{\boldsymbol{\theta}}_R$ denotes the least squares solution with ℓ_2 regularization (Ridge regression)
- $\hat{\boldsymbol{\theta}}_1$ denotes the least squares solution with ℓ_1 regularization (LASSO)

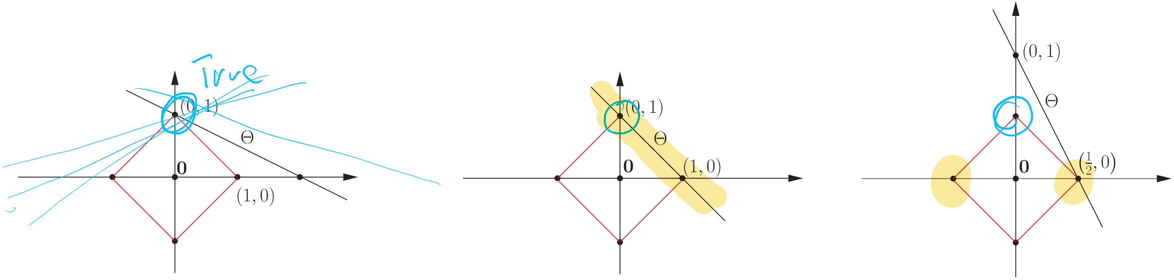
When is ℓ_1 unique? Example of recovery I

- Consider an unknown system.
- The system only has two parameters, and you would like to recover those parameters.
- You know the system is on the form $f(x_1, x_2) = \theta_1 x_1 + \theta_2 x_2$.
- You can only measure once, that is, you can only select one pair of (x_1, x_2).
- To simulate this environment, we create the θ's for the unknown system, and select those as $(\theta_1, \theta_2) = (0, 1)$ (abitrary choice).
- We will now see what happens if we select 3 possible measurements, $\boldsymbol{x}_a = (1/2, 1)$, $\boldsymbol{x}_b = (1, 1)$, $\boldsymbol{x}_c = (2, 1)$.
- These three measurements will all give a response of $f(x_1, x_2) = 1$ for the unknown system.

When is ℓ_1 unique? Example of recovery II

True $\boldsymbol{\theta} = (0, 1)$. Searching for Θ

Test measurements: $\boldsymbol{x}_a = (1/2, 1)$, $\boldsymbol{x}_b = (1, 1)T$, $\boldsymbol{x}_c = (2, 1)$ (all result in $f(\boldsymbol{x}) = 1$).
Let us draw possible solutions to the linear system, $f(\boldsymbol{x}) = \theta_1 x_1 + \theta_2 x_2 = 1$, and select the solution with the lowest ℓ_1 norm.



Summary

Sparsity-aware learning have many cool applications, in statistics, signal processing, machine learning.

- In the pursuit of sparse solution, we arrived at using the ℓ_1 norm as the computationally most efficient norm.
  - The norm is convex.
  - The ℓ_0 "norm" leads to the sparsest solution, but is not convex.
- Under special circumstances, the ℓ_1 regularization will find the sparsest solution.
- LASSO solves the ℓ_1 norm regularization problem.

Sparsity promoting algorithms

A greedy approach

Solving $\|\theta\|_0$

The OMP algorithm – algorithm 10.1 in the book

k < N < l

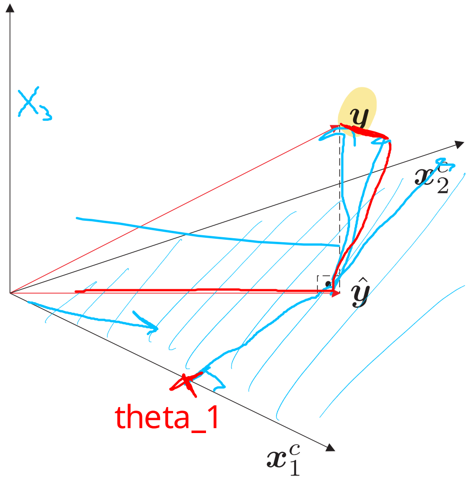
- **Initialize**
  - $\boldsymbol{\theta}^{(0)} = \mathbf{0} \in \mathbb{R}^l$.
  - $S^{(0)} = \emptyset$. Solution set
  - $\boldsymbol{e}^{(0)} = \boldsymbol{y}$. error vector
- **For** $i = 1, \cdots, k$ **Do**
  - Select the column in X that forms the smallest angle with the error. x^c
  - Update the indices of active vectors, $S^{(i)}$. s={3, 5}
  - Update the parameter vector $\boldsymbol{\theta}^{(i)}$ using least squares using the columns in X indexed by $S^{(i)}$.
  - Update the error vector. y - \hat{y} $\mid \theta^{(i)} = (X^TX)^{-1}X^Ty$ X'=[x^c_3, X^c_5]
- **End For**

Parameters:

k is the number of non–zero components (must be smaller than the number of observations)

Why does the OMP work?

Remember, the inner product actually carries out a projection



Iterative Shrinkage/thresholding (IST)

FISTA

solving l_1

The naive IST formula (10.3)–(10.7) in the book (estimates the LASSO solution)

- **Initialize**
  - $\boldsymbol{\theta}^{(0)} = \mathbf{0} \in \mathbb{R}^l$.
  - Select the value of μ stepsize
  - Select the value of λ reg.
- **For** $i = 1, \cdots$ **Do**
  - $\boldsymbol{e}^{(i-1)} = \boldsymbol{y} - X\boldsymbol{\theta}^{(i-1)}$ LMS, week 3-4
  - $\tilde{\boldsymbol{\theta}} = \boldsymbol{\theta}^{(i-1)} + \mu X^T \boldsymbol{e}^{(i-1)}$
  - $\boldsymbol{\theta}^{(i)} = \text{sign}\left(\tilde{\boldsymbol{\theta}}\right) \max\left(\left|\tilde{\boldsymbol{\theta}}\right| - \lambda\mu, 0\right)$ week 6
    1/2 lambda
- **End For**

X^TX = I

Parameters:

μ is still the step size, but also affects the shrinkage.

λ is the regularization parameter.

Route to IST

ex 7.2

If we have the cost function,

LS

$$J(\boldsymbol{\theta}) = \frac{1}{2}\|\boldsymbol{y} - X\boldsymbol{\theta}\|_2^2$$

We know from earlier weeks that the gradient descent update is

$$\boldsymbol{\theta}^{(i)} = \boldsymbol{\theta}^{(i-1)} + \mu X^T \boldsymbol{e}^{(i-1)}$$

GD

However, it turns out that this is also the solution to the following optimization problem

LS

$$\boldsymbol{\theta}^{(i)} = \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^l} \left\{ J\left(\boldsymbol{\theta}^{(i-1)}\right) + \left(\boldsymbol{\theta} - \boldsymbol{\theta}^{(i-1)}\right)^T \frac{\partial J\left(\boldsymbol{\theta}^{(i-1)}\right)}{\partial \boldsymbol{\theta}} + \frac{1}{2\mu}\left\|\boldsymbol{\theta} - \boldsymbol{\theta}^{(i-1)}\right\|_2^2 \right\}$$

This is shown by taking the derivative w.r.t $\boldsymbol{\theta}$ and set to zero.

Route to IST step 2

From previous slides:

$$\boldsymbol{\theta}^{(i)} = \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^l} \left\{ J\left(\boldsymbol{\theta}^{(i-1)}\right) + \left(\boldsymbol{\theta} - \boldsymbol{\theta}^{(i-1)}\right)^T \frac{\partial J\left(\boldsymbol{\theta}^{(i-1)}\right)}{\partial \boldsymbol{\theta}} + \frac{1}{2\mu}\left\|\boldsymbol{\theta} - \boldsymbol{\theta}^{(i-1)}\right\|_2^2 \right\}$$

From prev Slide

LASSO minimizes

LS

reg

$$\boldsymbol{\theta}^{(i)} = \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^l} \left\{ \frac{1}{2}\|\boldsymbol{y} - X\boldsymbol{\theta}\|_2^2 + \lambda\|\boldsymbol{\theta}\|_1 \right\}$$

Combining these gives

LS

$$\boldsymbol{\theta}^{(i)} = \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^l} \left\{ J\left(\boldsymbol{\theta}^{(i-1)}\right) + \left(\boldsymbol{\theta} - \boldsymbol{\theta}^{(i-1)}\right)^T \frac{\partial \boldsymbol{J}\left(\boldsymbol{\theta}^{(i-1)}\right)}{\partial \boldsymbol{\theta}} + \frac{1}{2\mu}\left\|\boldsymbol{\theta} - \boldsymbol{\theta}^{(i-1)}\right\|_2^2 + \lambda\|\boldsymbol{\theta}\|_1 \right\}$$

Solving this minimization problem (taking the derivative, set to 0, and solve for $\boldsymbol{\theta}$), will yield the IST update (and avoids the problem we had last week where we assumed $X^T X = I$).

Summary

- We presented two algorithms,
- OMP is a heuristic greedy approach, that simply builds up a k–sparse solution vector in k steps. Hence, it solves the ℓ_0 solution.
  - There is no garantee that OMP finds the optimal ℓ_0 solution.
  - LARS and LARS-LASSO are extensions of OMP.
- IST is an iterative shrinkage/thresholding (IST) type algorithm, and estimates the ℓ_1 solution.
  - It is a naive implementation we derived, and in practice one should use e.g. FISTA.
- Under certain circumstances (theory not part of curriculum, but listed in 9.6–9.7), the ℓ_0 and ℓ_1 minimizer has the same solution.
- You will implement both in the exercise today.

Signal representation

Discrete Fourier transform (DFT)

Freq

X DFT X IDFT $\tilde{X}$

Analysis

Freq $$X(k) = \sum_{n=0}^{N-1} x(n) e^{-j\frac{2\pi}{N}kn}$$
signal

Synthesis

signal $$x(n) = \frac{1}{N}\sum_{k=0}^{N-1} X(k) e^{j\frac{2\pi}{N}kn}$$
Freq

k corresponds to frequency kF_s/N

ML 10.5

Linear signal representations

A linear signal representation model can be thought of as

Linear signal representations

$$\tilde{s} = \Phi^H s \qquad \text{analysis}$$

$$s = \Phi \tilde{s} \qquad \text{synthetis}$$

Learn it?

- s is the vector of raw samples.
- $\tilde{s}$ is the transformed vector.
- Φ is the unitary transformation matrix, $\Phi\Phi^H = I$.

There are many choices of matrices, we can choose Φ^H as a matrix of fourier coefficients (complex matrix).
Other choices could be the DCT matrix (which is real), or wavelet matrix.

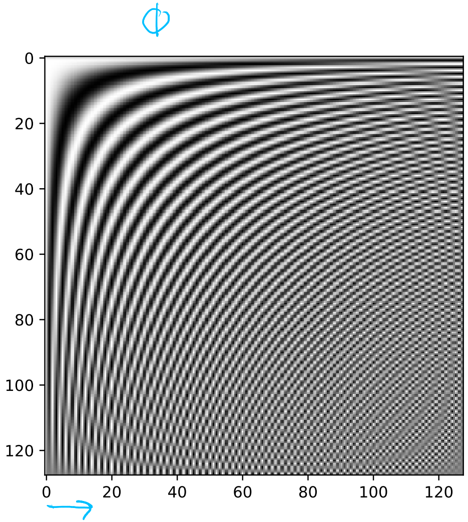
The DCT case

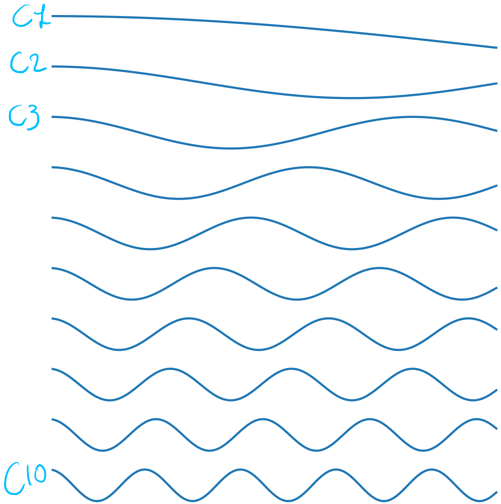
$$\text{DCT-II (dct)} : X_k = \sum_{n=0}^{N-1} x_n \cos\left[\frac{\pi}{N}\left(n+\frac{1}{2}\right)k\right] \quad k = 0,\dots,N-1$$

$$\text{DCT-III (idct)} : X_k = \frac{1}{2}x_0 + \sum_{n=1}^{N-1} x_n \cos\left[\frac{\pi}{N}n\left(k+\frac{1}{2}\right)\right] \quad k = 0,\dots,N-1$$

These transforms can be organized as real matrices, such that Φ^H corresponds to DCT-II and Φ corresponds to DCT-III.

The DCT matrix



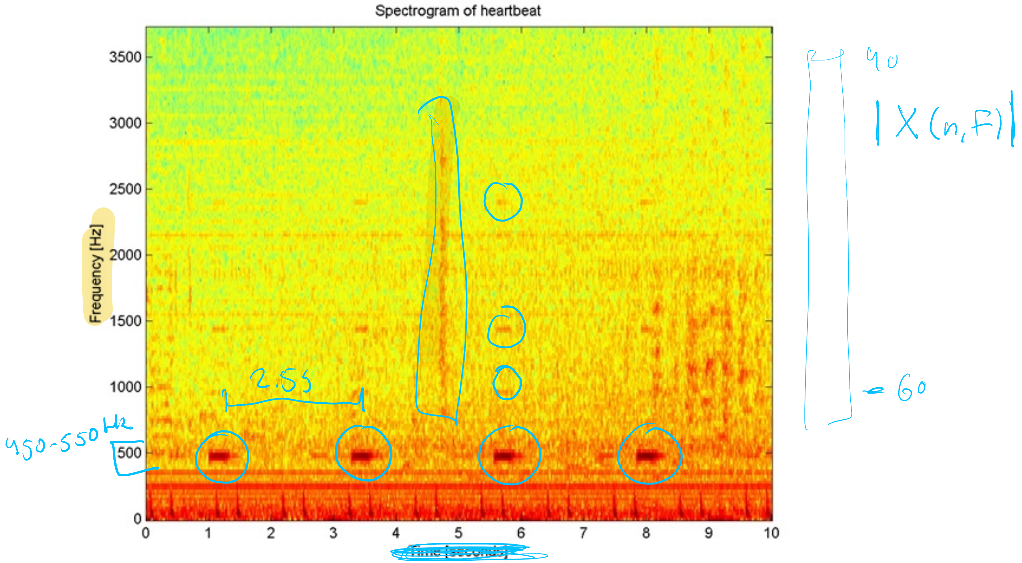


Summary

- We operate with linear signal representations.
- The Fourier transform, the DCT transform (and more) can be recast as linear projections.
- The model is closely related to dictionary learning, where Φ is estimated from data instead of predefined.

Time-frequency analysis

The spectrogram of a heartbeat



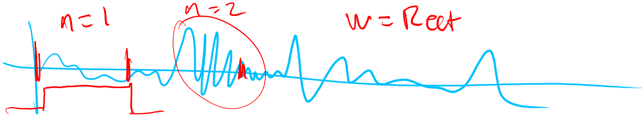
Short-time Fourier transform (STFT)

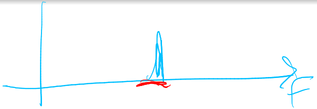
DFT

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{-j\frac{2\pi}{N}kn}$$

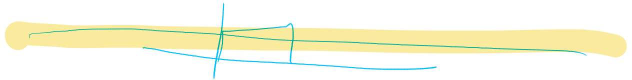
STFT

$$X(n,k) = \sum_{m=-\infty}^{\infty} \underbrace{x(m)w(m-n)}_{\text{new signal: } \hat{x}(n)} e^{-j\frac{2\pi}{N}kn}$$





Windows



Without loss of generality, let us define

old window

$$F[\hat{x}(n)] = x(n) \cdot w(n)$$

From properties of Fourier transform, we know multiplication in the time domain is convolution in the frequency domain

org window

$$\hat{X}(f) = X(f) * W(f) = \int_{-1/2}^{1/2} X(s) W(f-s) ds$$

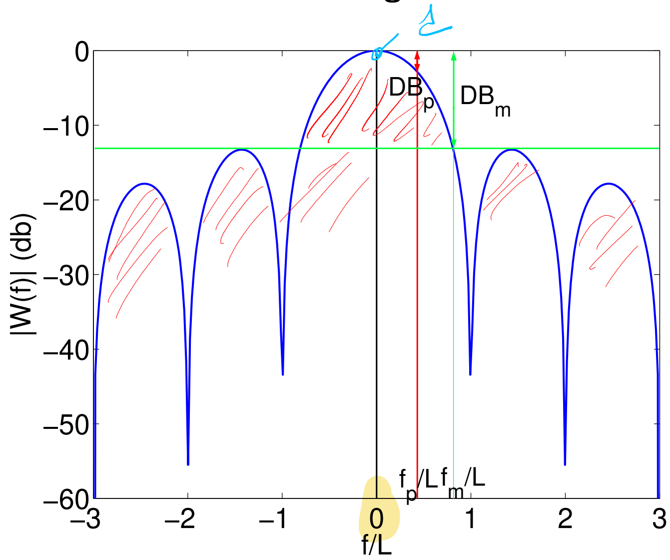
How would the ideal $W(f)$ look like?

$\hat{X}(F) \approx X(f)$

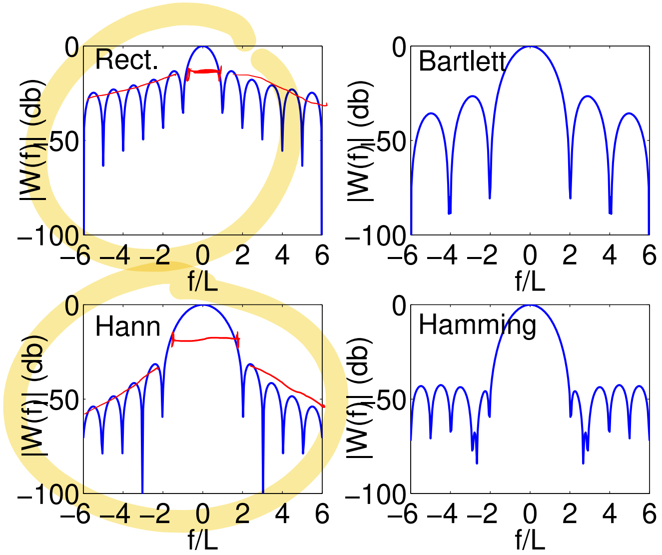
$W(f$

https://en.wikipedia.org/wiki/Window_function

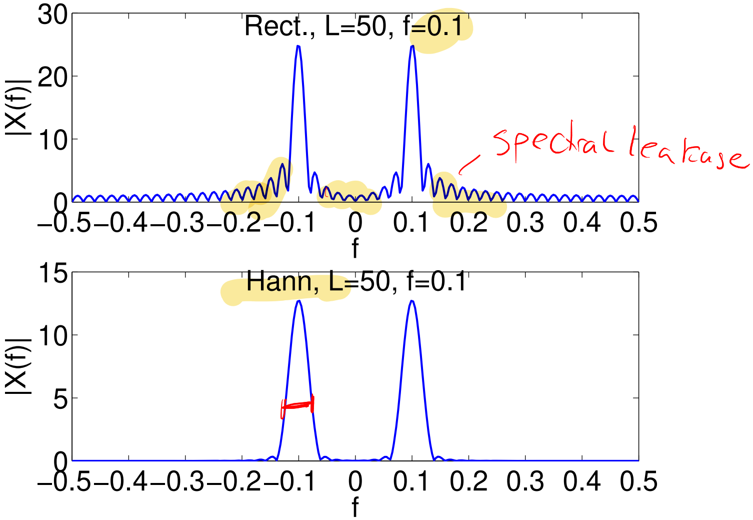
On window functions: resolution and leakage



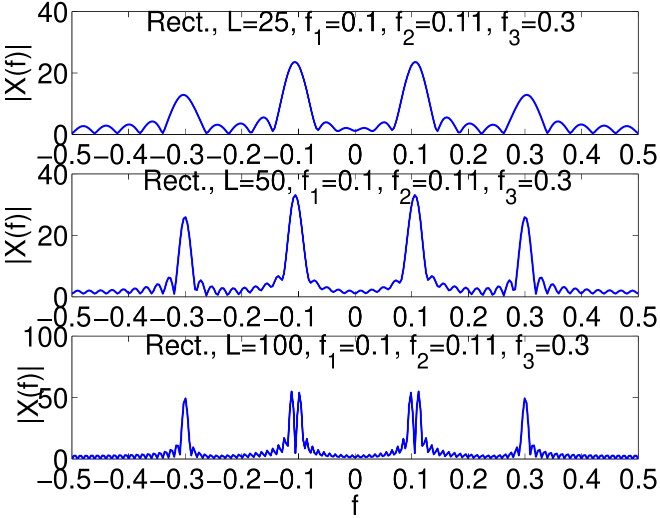
Spectral shape of windows



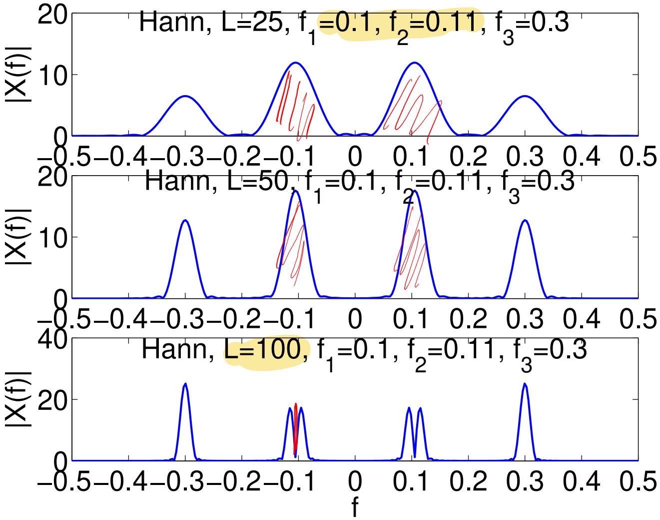
Example: sinusoid signal



Example: rectangular on 3 sinusoid



Example: Hanning on 3 sinusoids



The spectrogram

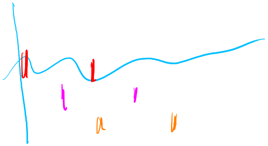
STFT

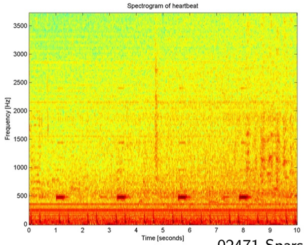
$$X(n,k) = \sum_{m=-\infty}^{\infty} x(m)w(m-n)e^{-j\frac{2\pi}{N}kn}$$

The spectrogram

The magnitude spectrum computed using STFT, ie $|X(n,k)|$.

Two important parameters; the block size B (window size), and the hop size S (stride, or window overlap).

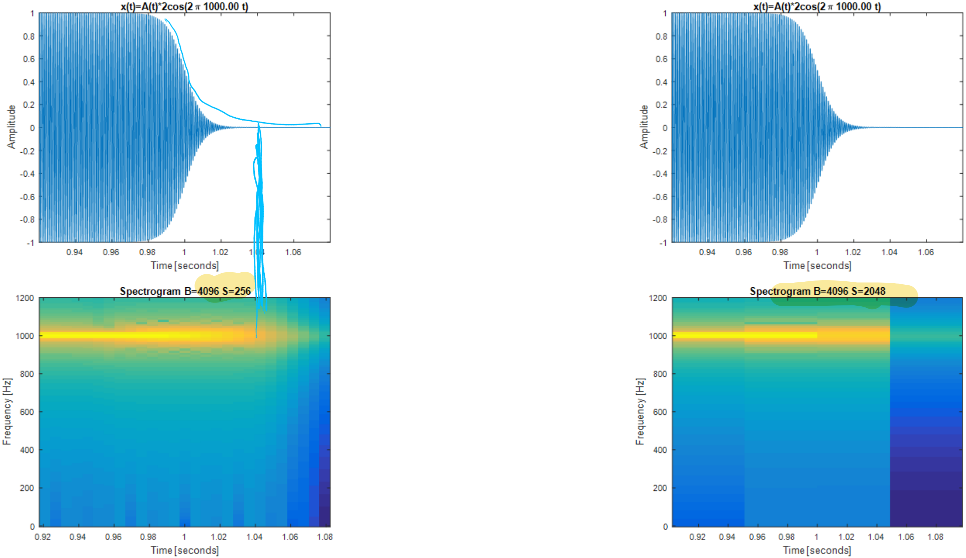




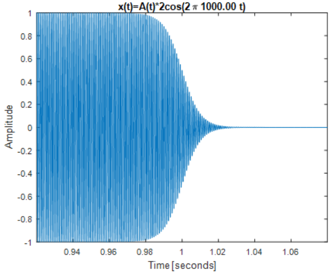
STFT of a cosine

Time

STFT



STFT of a cosine



Considerations for the STFT

- Time overlap of windows: smooth transitions little perceptual artifacts, however, should be done so "perfect" that synthesis is possible.
- The shape and length of the window: slow variation of amplitude within a window and little leakage.
- The number of points N (frequency bins) in the DFT: only one partial in each band.

Lecture summary

- Two algorithms were presented, OMP and IST. And they solve ℓ_0 and ℓ_1 respectively.
- This leads to linear signal representation models.
- If the signal is not stationary, the representation models can be applied on smaller chunks of the signal. This approach is called "time-frequency analysis":
  - This can be e.g. the Fourier transform, which lead to the short-time Fourier transform (STFT).
  - Other window choices leads to the Gabor transform.
  - Other base function choices leads to wavelets.
  - The T–F approach can be used to extract features from signals (the time series gets a vector-space representation), that can then be applied to a machine learning classifier.

Next week

Material: ML 2.5, 19.1–19.3, 19.5–19.7.

2022 p 5.4

- PCA brief primer (very brief, leads to ICA).
- Independent component analysis (ICA).
- Dictionary learning (NMF, k–SVD).